89-0076 251

Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition

RLE Technical Report No. 536

December 1988

James Robert Glass

19960614 008

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

DTIC QUALITY INSPECTED 1

DO NOT REMOVE
SECTOR TECHNOLOGY

Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition

RLE Technical Report No. 536

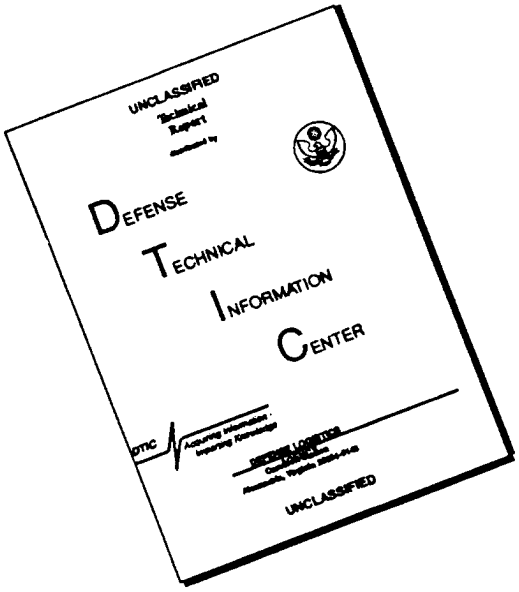
December 1988

James Robert Glass

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

This work has been supported by the Natural Sciences and Engineering Research Council of Canada and the Defence Advanced Research Projects Agency.

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

OSP 95294

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARPA Order No. 4585
6a. NAME OF PERFORMING ORGANIZATION Research Laboratory of Electronics Massachusetts Institute of Technology 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Advanced Research Projects Agency
6c. ADDRESS (City, State, and ZIP Code) 77 Massachusetts Avenue Cambridge, MA 02139	7b. ADDRESS (City, State, and ZIP Code) 1400 Wilson Blvd. Arlington, VA 22217
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Res., Math. & Physical Sciences Res. Program 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-82-K-0727
8c. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, VA 22217	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: PROJECT NO.: NR-049-542 TASK NO.: WORK UNIT ACCESSION NO.:

11. TITLE (Include Security Classification)
Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition

12. PERSONAL AUTHOR(S)
James R. Glass

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Report	FROM _____ TO _____	December 1988	157

16. SUPPLEMENTARY NOTATION
Technical Report 536, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1988.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Please see next page

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Elisabeth Colford - RLE Contract Reports	22b. TELEPHONE (Include Area Code) (617)258-5871 22c. OFFICE SYMBOL

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

19. ABSTRACT

Phonetic recognition can be viewed as a process through which the acoustic signal is mapped to a set of phonological units used to represent a lexicon. Traditionally, researchers often prescribe an intermediate, phonetic description to account for coarticulation. This thesis presents an alternative approach whereby this phonetic-level description is bypassed in favor of directly relating the acoustic realizations to the underlying phonemic forms. In this approach, the speech signal is transformed into a set of segments which are described completely in acoustic terms. Next, these acoustic segments are related to the phonemes by a grammar which is determined using automated procedures operating on a set of training data. Thus important acoustic regularities that describe contextual variations are discovered without the need to specify a set of preconceived units such as allophones.

The viability of this approach depends critically on the ability to detect important acoustic landmarks in the speech signal, and to describe these events in terms of an inventory of labels that captures the regularity of phonetic variations. In the present implementation, the signal is first transformed into a representation based on an auditory model developed by Seneff. Next, important acoustic landmarks are located, and embedded in a multi-level structure called a dendrogram, in which information ranging from coarse to fine is represented in a unified framework. Each acoustic region in the dendrogram is then described by a set of acoustic labels determined through a hierarchical clustering algorithm.

An analysis of the multi-level structure on a set of 500 utterances recorded from 100 different talkers indicates that over 96% of important acoustic-phonetic events are located, with an insertion rate of less than 5%. Differences between the dendrogram structure and hand-marked boundaries appear to be quite systematic. An analysis of the clustering algorithm on data from all phonemes indicate that it is possible to assign an acoustic segment to one of a small set of acoustic categories, each having a meaningful phonetic interpretation. An examination of the realizations of weak voiced fricatives, and velar stop consonants indicate that it is possible to determine a finite number of regular acoustic forms which capture consistent contextual dependencies. Additionally, there is evidence that these regularities can generalize across sets of phonemes with similar phonetic features.

Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition

by
James Robert Glass

Submitted to the Department of Electrical Engineering
and Computer Science on May 10, 1988 in partial fulfillment
of the requirements for the degree of Doctor of Philosophy

Abstract

Phonetic recognition can be viewed as a process through which the acoustic signal is mapped to a set of phonological units used to represent a lexicon. Traditionally, researchers often prescribe an intermediate, phonetic description to account for coarticulation. This thesis presents an alternative approach whereby this phonetic-level description is bypassed in favor of directly relating the acoustic realizations to the underlying phonemic forms. In this approach, the speech signal is transformed into a set of segments which are described completely in acoustic terms. Next, these acoustic segments are related to the phonemes by a grammar which is determined using automated procedures operating on a set of training data. Thus important acoustic regularities that describe contextual variations are discovered without the need to specify a set of preconceived units such as allophones.

The viability of this approach depends critically on the ability to detect important acoustic landmarks in the speech signal, and to describe these events in terms of an inventory of labels that captures the regularity of phonetic variations. In the present implementation, the signal is first transformed into a representation based on an auditory model developed by Seneff. Next, important acoustic landmarks are located, and embedded in a multi-level structure called a dendrogram, in which information ranging from coarse to fine is represented in a unified framework. Each acoustic region in the dendrogram is then described by a set of acoustic labels determined through a hierarchical clustering algorithm.

An analysis of the multi-level structure on a set of 500 utterances recorded from 100 different talkers indicates that over 96% of important acoustic-phonetic events are located, with an insertion rate of less than 5%. Differences between the dendrogram structure and hand-marked boundaries appear to be quite systematic. An analysis of the clustering algorithm on data from all phonemes indicate that it is possible to assign an acoustic segment to one of a small set of acoustic categories, each having a meaningful phonetic interpretation. An examination of the realizations of weak voiced fricatives, and velar stop consonants indicate that it is possible to determine a finite number of regular acoustic forms which capture consistent contextual dependencies. Additionally, there is evidence that these regularities can generalize across sets of phonemes with similar phonetic features.

Thesis Supervisor: Dr. Victor W. Zue
Title: Principal Research Scientist

Acknowledgments

I extend my deepest gratitude to Victor Zue, my thesis advisor, for giving me the opportunity to learn from his expertise in speech and acoustic-phonetics. I thank him for his continuing guidance, and thoughtful advice, as well as for his constant support, encouragement, and friendship during the course of this research. Working with Victor has been both challenging and rewarding, an experience for which I will always be grateful.

I would like to thank the members of my thesis committee, Dennis Klatt, Campbell Searle, Stephanie Seneff, and Ken Stevens, for their interest in this work, and for their many helpful insights, and advice, and for their critiques of versions of this thesis.

I also thank Corine Bickley, Nancy Daly, Carol Espy-Wilson, John Pitrelli, and Stefanie Shattuck-Hufnagel for reading various drafts of this thesis, and for their comments, and suggestions.

I thank the members of the Speech Communications Group for many helpful suggestions and discussions, and for creating a stimulating and enjoyable environment in which to conduct research. In particular, I would like to thank Rob Kassel, Hong Leung, Mike Phillips, Mark Randolph, Stephanie, and Stefanie for discussions which helped clarify issues related to this work.

I would like to thank Scott Cyphers, David Kaufman, Katy Isaacs, and Keith North, for keeping all the equipment running smoothly. I thank Scott, David, Hong, Rob, and Mark for writing software packages which made the Lisp Machines such a powerful research facility. I would also like to thank Rob and Dave Whitney for their efforts in getting the laser printers to work. I am grateful to Rob for the many hours he invested in creating a wonderful environment for making documents.

I would like to thank Lori Lamel for finishing her thesis at the same time as me. Her determination, and moral support helped make the task of completing this document a little more bearable.

I thank Lori Sorayama for her constant support and encouragement, and for raising my spirits whenever they started to sag.

Finally, I thank my parents, and my brother and sisters, for their continuing love, and for believing that one day I would actually finish.

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and by DARPA.

Contents

1 Introduction 8
- 1.1 The Discrete Nature of Speech . . . 9
- 1.2 The Continuous Nature of Speech . . . 9
  - 1.2.1 Describing Speech Sounds . . . 15
  - 1.2.2 Sources of Variability in Speech . . . 15
- 1.3 Decoding the Speech Signal . . . 16
  - 1.3.1 Whole Word Units . . . 19
  - 1.3.2 Phonological Units . . . 20
  - 1.3.3 Acoustic Units . . . 24
- 1.4 Thesis Scope . . . 28

2 Signal Representation and Database 30
- 2.1 Signal Representation of Speech . . . 30
- 2.2 Database Description . . . 35

3 Acoustic Segmentation 36
- 3.1 Finding Acoustic Landmarks . . . 36
  - 3.1.1 Locating Points of Maximum Change . . . 37
  - 3.1.2 Associations . . . 41
- 3.2 Multi-level Acoustic Description of Speech . . . 43
  - 3.2.1 Scale-Space Filtering . . . 43
  - 3.2.2 Hierarchical Structuring . . . 46
- 3.3 Performance Evaluation . . . 53
  - 3.3.1 Evaluation Criterion . . . 53
  - 3.3.2 Evaluation Results . . . 54
- 3.4 Chapter Summary . . . 61

4 Acoustic Classification 62
- 4.1 Hierarchical Clustering . . . 63
- 4.2 Coarse Acoustic Classification . . . 66
  - 4.2.1 Signal Representation . . . 66
  - 4.2.2 Cluster Evaluation . . . 68
- 4.3 Finding Regularities in the Realization of Phonemes . . . 79
  - 4.3.1 Weak Voiced Fricatives . . . 79
  - 4.3.2 Velar Stop Consonants . . . 83
  - 4.3.3 Modeling Time-Varying Changes . . . 86
- 4.4 Chapter Summary . . . 91

5 Signal Representation Refinements **92**
5.1 Limitations of the Mean-Rate Response 92
5.2 Synchronous Analysis of Speech 95
5.2.1 The Channel Sum Waveform 95
5.2.2 Determining Event Locations 97
5.2.3 Producing the Synchronous Response 97
5.3 Extracting the Fundamental Frequency 104
5.4 Dimensional Analysis of the Mean-rate Response 107
5.4.1 Statistical Analysis 109
5.4.2 Stability of Correlation Estimates 112
5.4.3 Principal Component Analysis 113
5.5 Chapter Summary 117

6 Summary and Future Work **118**
6.1 Signal Representation 119
6.2 Acoustic Segmentation 120
6.3 Acoustic Classification 121
6.4 Applications to Machine Recognition of Speech 123
6.4.1 Motivation for Segmentation 124
6.4.2 Explicit vs Implicit Segmentation 125
6.4.3 A Probabilistic Framework 126

Bibliography **129**

A Dendrograms **138**

B Stochastic Segmentation Computation **148**

List of Figures

1.1 The continuous nature of speech. 10
1.2 Examples of palatalization in American English. 12
1.3 Common acoustic realizations of the phoneme /u/. 13
1.4 Word boundary effects. 14
1.5 Alternative strategies for decoding the speech signal. 18
1.6 Alternative perspectives of speech. 21
1.7 Spectrograms of the words 'butter,' and 'away.' 22

2.1 Block diagram of parts of Seneff's auditory model. 32
2.2 Critical-band filter frequency response (after Seneff). 33
2.3 Comparison of two spectral representations. 34

3.1 Partial trajectory of the word 'international.' 38
3.2 Different measures of change. 40
3.3 Multi-level description using scale-space filtering. 45
3.4 Multi-level description using a hierarchical structure. 49
3.5 A non-monotonically increasing dendrogram. 50
3.6 A dendrogram computed with a Euclidean distance metric. 52
3.7 Dendrogram alignment procedure. 55
3.8 An aligned dendrogram. 56
3.9 Distribution of dendrogram boundary heights. 60

4.1 Seed clusters membership. 67
4.2 Rate of change of distortion versus number of clusters. 69
4.3 Spectra of ten clusters. 72
4.4 Phonetic hierarchical structure with ten clusters. 75
4.5 Spectra of twelve clusters. 77
4.6 Spectra of /ð/ clusters. 81
4.7 Spectra of /v/ clusters. 82
4.8 Segment duration comparison for /k/. 84
4.9 Spectra of /k/ clusters. 85
4.10 Spectra of /g/ clusters. 88
4.11 Spectra of /æ/ clusters. 90

5.1 Channel sum waveforms. 94
5.2 Impulse response of the auditory model. 96
5.3 Computing event locations. 98
5.4 Determining event locations. 99
5.5 Computation involved in computing a synchronous response. 101

5.6 Comparison of mean-rate spectra for the word 'time.' 102
5.7 Stop release deletion rate. 103
5.8 Determining periodicities with an AMDF. 105
5.9 Fundamental frequency extraction in the presence of noise. 106
5.10 Distribution of speech in the auditory channels. 110
5.11 Histogram of the correlation among different channels. 111
5.12 Average correlation versus channel separation. 112
5.13 Stability of the correlation estimate. 113
5.14 Fraction of total variance explained by components. 114
5.15 First ten speaker-independent components. 115
5.16 Off-diagonal Correlations 116

6.1 Voice onset time of velar stop consonants. 123
6.2 A probabilistic network. 128

A.1 Dendrogram of 'A muscular abdomen is good for your back.' 139
A.2 Dendrogram of 'Any contributions will be greatly appreciated.' 140
A.3 Dendrogram of 'A doctor was in the ambulance with the patient.' . . 141
A.4 Dendrogram of 'Rob sat by the pond and sketched the stray geese.' . 142
A.5 Dendrogram of 'Bagpipes and bongos are musical instruments.' . . . 143
A.6 Dendrogram of 'Even a simple vocabulary contains symbols.' 144
A.7 Dendrogram of 'Ambidextrous pickpockets accomplish more.' 145
A.8 Dendrogram of 'My ideal morning begins with hot coffee.' 146
A.9 Dendrogram of 'The small boy put the worm on the hook.' 147

B.1 Stochastic segmentation computation. 149

List of Tables

2.1 Typical sentences from the TIMIT database. 35

3.1 Hierarchical structuring of acoustic landmarks. 47
3.2 Top ten sources of deletions. 57
3.3 Top five sources of insertions. 59

4.1 Pre-clustering procedure for data reduction. 65
4.2 Distributions for ten clusters (percent normalized by phone). 70
4.3 Distributions for ten clusters (percent normalized by cluster). 73
4.4 Distributions for twelve clusters (percent normalized by phone). . . . 78
4.5 Summary of distributions of preceding contexts for /ð/. 80
4.6 Distributions of some following contexts for /k/. 87
4.7 Summary of distributions of following contexts for /k/. 89
4.8 Summary of distributions of following contexts for /æ/. 89

Chapter 1

Introduction

The human ability to perceive speech is often taken for granted. After all, this is something we all learn to do as infants and use throughout our daily lives, usually without conscious effort. Ironically, this process is actually quite complicated. Despite several decades of vigorous research, there are no definitive explanations for many of the fundamental issues involved in the perception of speech [52,58,75,88]. Scientists cannot yet explain precisely how the human listener converts the speech signal into linguistic units, and how these units are employed to extract the message intended by the talker. To date, speech remains a paradox. At an abstract level, speech is encoded in discrete and invariant units, such as phonemes, distinctive features, or words. At a physical level, however, the speech signal is continuous, and appears to be quite variable. Researchers have long been challenged to understand the relationship between these two contrasting levels of representation.

This thesis reports an investigation into the acoustic nature of the speech signal. Specifically, it explores the utility of a discrete, acoustic level of representation whose units can be automatically determined from the speech signal. The motivation for this approach will be discussed in the remaining sections of this chapter.

1.1 The Discrete Nature of Speech

All languages appear to make use of a finite number of distinguishable, mutually exclusive sounds which are concatenated together in time to produce speech. These basic linguistic units are called *phonemes*, and possess unique articulatory configurations [28]. The word 'mitt' for example consists of the three-phoneme sequence /m/, /ɪ/, and /t/, where the /m/ is produced by closing the mouth, lowering the velum, and vibrating the vocal folds while exhaling. Words that differ in their phonemic form have different linguistic meanings. Replacing the phoneme /m/ by /n/ changes the word 'mitt' to 'knit.' The English language has an inventory of some 40 phonemes containing vowels, semivowels, nasals, aspirants, fricatives, and stops [67].

It is widely accepted that a phoneme may be characterized by a small set of distinctive *features*, where a feature is a minimal unit which distinguishes two maximally-close phonemes [12,53]. For instance, the phonemes /d/ and /t/ are distinguished by the feature *voice*. Phonemes which share a given feature form natural classes with common characteristics. In English for example, the feature *nasal* is shared by the phonemes /m/, /n/, and /ŋ/. It is believed that around fifteen to twenty distinctive features are necessary to account for phonemic contrasts in all languages of the world, although a given language will typically require only ten to fifteen of these features [112]. Distinctive features are powerful descriptors, since they describe speech with a minimum amount of redundancy and can be used to characterize significant linguistic generalizations [6,30].

1.2 The Continuous Nature of Speech

In stark contrast to a phonemic level of representation, which is discrete and invariant, the speech signal is continuous and exhibits a considerable amount of variability. One of the most notable sources of variability stems from the production of the phonemes themselves. Speech is generated by moving the articulators through

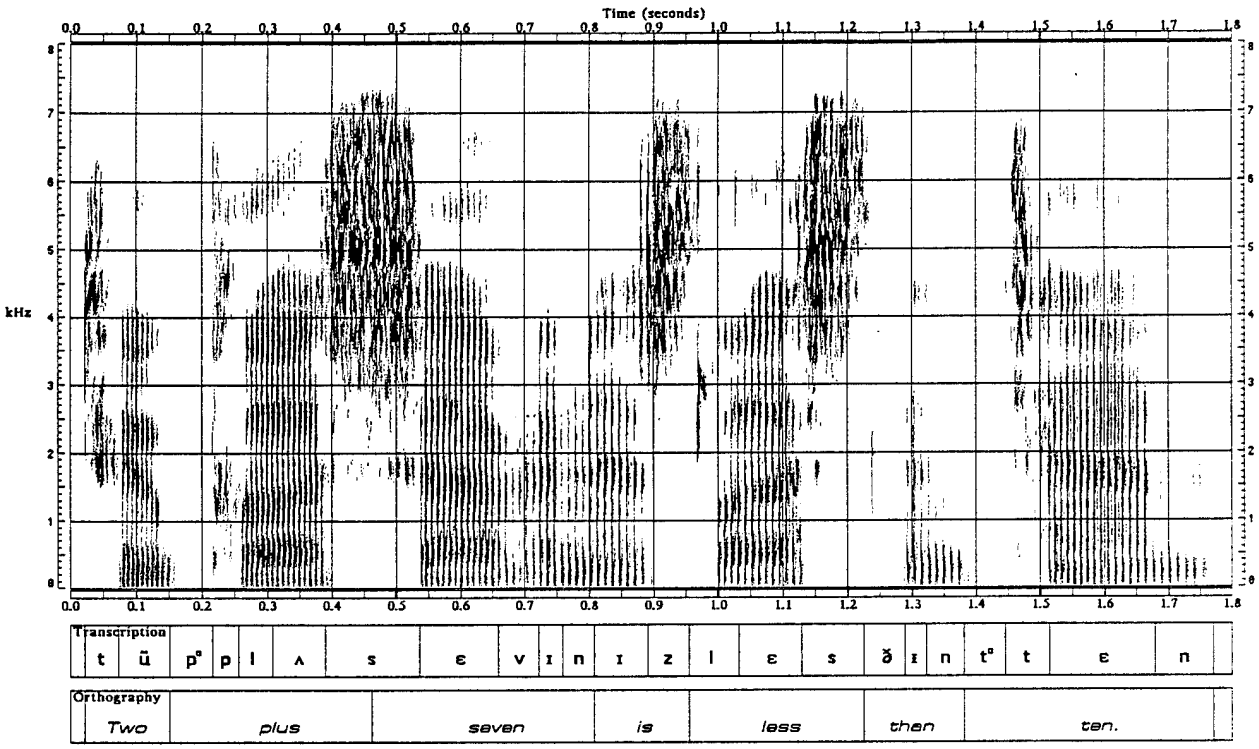


Figure 1.1: The continuous nature of speech.

Digital spectrogram of the sentence 'Two plus seven is less than ten,' spoken by a male talker. The utterance illustrates some common kinds of coarticulation found in continuous speech [123]. A time-aligned phonetic and orthographic transcription of the sentence are shown below the spectrogram.

a coordinated series of gestures guided by the target articulations of the individual phonemes [28,36]. Due to inertia and the fact that the articulators are controlled separately, the boundaries between the realizations of adjacent phonemes are blurred, so that it is not possible to identify precisely where the realization of one phoneme ends and the next one begins. As a result, the acoustic realization of a phoneme will depend on the immediate phonemic environment. This contextual influence is known as *coarticulation.*

A concrete illustration of how phonemic information manifests itself in the acoustic signal is shown in Figure 1.1 which contains a spectrogram of the sentence 'Two

plus seven is less than ten,' spoken by a male speaker. As Zue has discussed [123], this sentence contains a number of common examples of the influence of local context on the acoustic properties of underlying phonemes. Consider for example, the acoustic characteristics of the phoneme /t/, which starts the first and last words of the sentence. In both cases the phoneme is realized as a closure followed by a burst. Close examination of the spectrogram reveals that the burst characteristics are slightly different. In particular, the burst frequency is lower for the first /t/ than for the second, a direct consequence of anticipatory coarticulation caused by the rounded vowel /u/ in the word 'two.'

Another example of coarticulation may be found in the acoustic realization of the phoneme /ɛ/ in the three words 'seven,' 'less,' and 'ten.' The second /ɛ/ is influenced by the adjacent /l/, such that the second formant shows articulatory undershoot, while the third /ɛ/ is heavily nasalized, as evidenced by the smearing of the first formant. Other examples of coarticulation may be found in the production of the strident fricatives. The spectra corresponding to the /z/ in 'is,' and the /s/ in 'less,' show an increase in the lower-cutoff frequency near the end of the segment. In the first case, the increase is due to the following lateral consonant, while in the second case the increase is due to the following dental fricative, which is more anterior in its place of articulation than the /s/.

In addition to coarticulation, the acoustic realization of a phoneme is dependent on the phonology incorporated by the speaker. Although the distinction between these two phenomena is often difficult to delineate clearly, it is evident that regular changes take place in sounds when they occur in different relationships with other sounds. The resulting transformations can often dramatically affect the acoustic realization of a phoneme. An example of such a transformation would be the /s š/ sequence being realized as a long [š] in the word pair 'gas shortage,' or the /d y/ sequence begin realized as a [ǰ] in the word pair 'did you.' These effects, which are both illustrated in Figure 1.2, are instances of palatalization, which is a common phonological transformation in American English. These examples are not solely

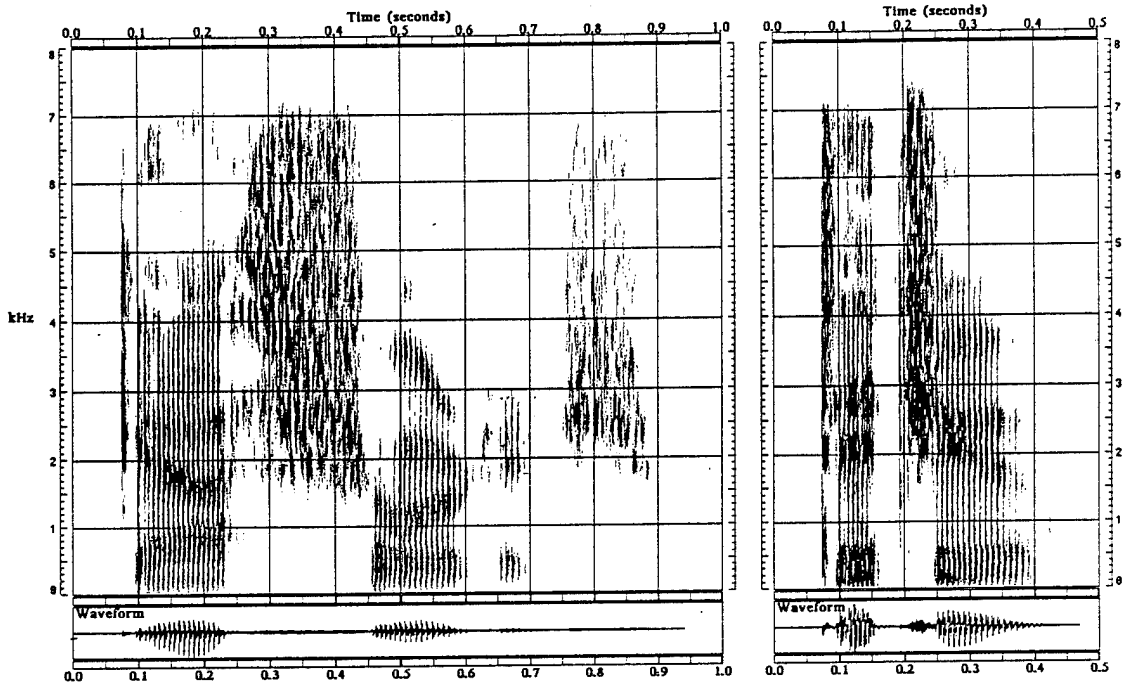


Figure 1.2: Examples of palatalization in American English.

Digital spectrogram of the word pairs 'gas shortage,' and 'did you,' spoken by a male talker. The utterances illustrate a common phonological transformation in American English.

caused by coarticulation, since the effect is not generally found in word pairs such as 'wash socks,' or 'I did,' where these phoneme sequences are reversed. The degree to which these effects are found in the speech signal depends on the speaking style and speaking rate of the talker, and also depends on the dialect.

Although coarticulatory and phonological effects are often viewed as complicating the decoding process, they can provide redundant sources of information about the nature of adjacent sounds, and can serve as valuable sources of suprasegmental information. For example, the phoneme /u/, which is considered a back vowel, is regularly fronted in an alveolar environment as in the word 'dune,' illustrated in Figure 1.3 [67]. Knowledge of this phenomenon should make it easier to distinguish a fronted /u/ from other high, front vowels in the same context, such as /i/ in 'dean.'

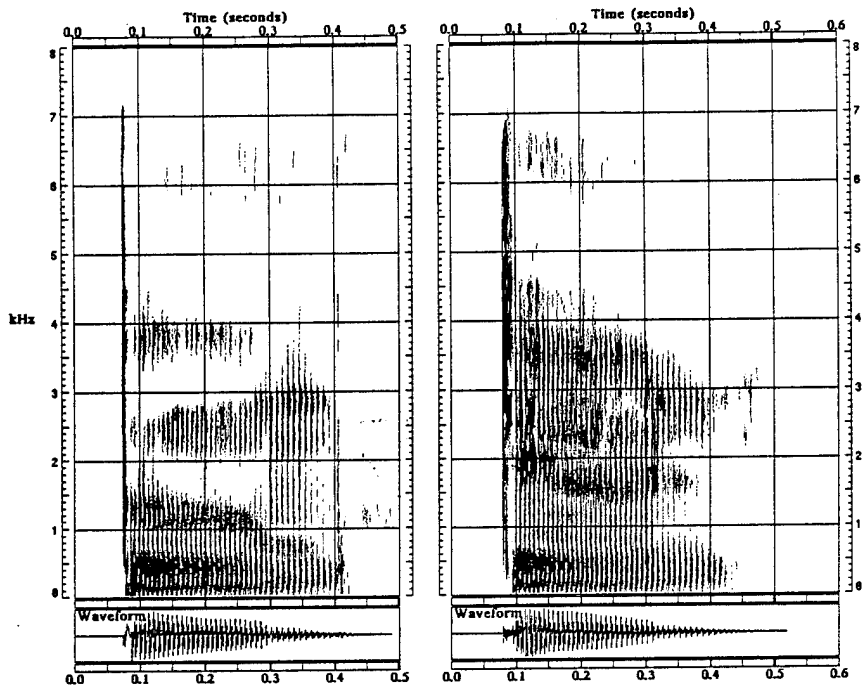


Figure 1.3: Common acoustic realizations of the phoneme /u/.

Digital spectrogram of the words 'boom,' and 'dune,' spoken by a male talker. The utterances illustrate a regular variation of the realization of the phoneme /u/ in American English.

In cases where the underlying phoneme sequences are identical, the phonology of a language will often provide suprasegmental information about syllable or word boundaries. In the word pairs 'great rain' and 'grey train,' or 'at ease' and 'a tease,' for instance, contrasting realizations of the phoneme /t/ can help suggest the location of the syllable boundary. In these examples, which are illustrated in Figure 1.4, the /t/ will typically be realized as an unreleased, retroflexed, flapped, or aspirated stop, respectively [15,68,83,95]. Thus, a knowledge of phonology and coarticulation combined with an appropriate level of detail in the acoustic-phonetic description can help decode the speech signal.

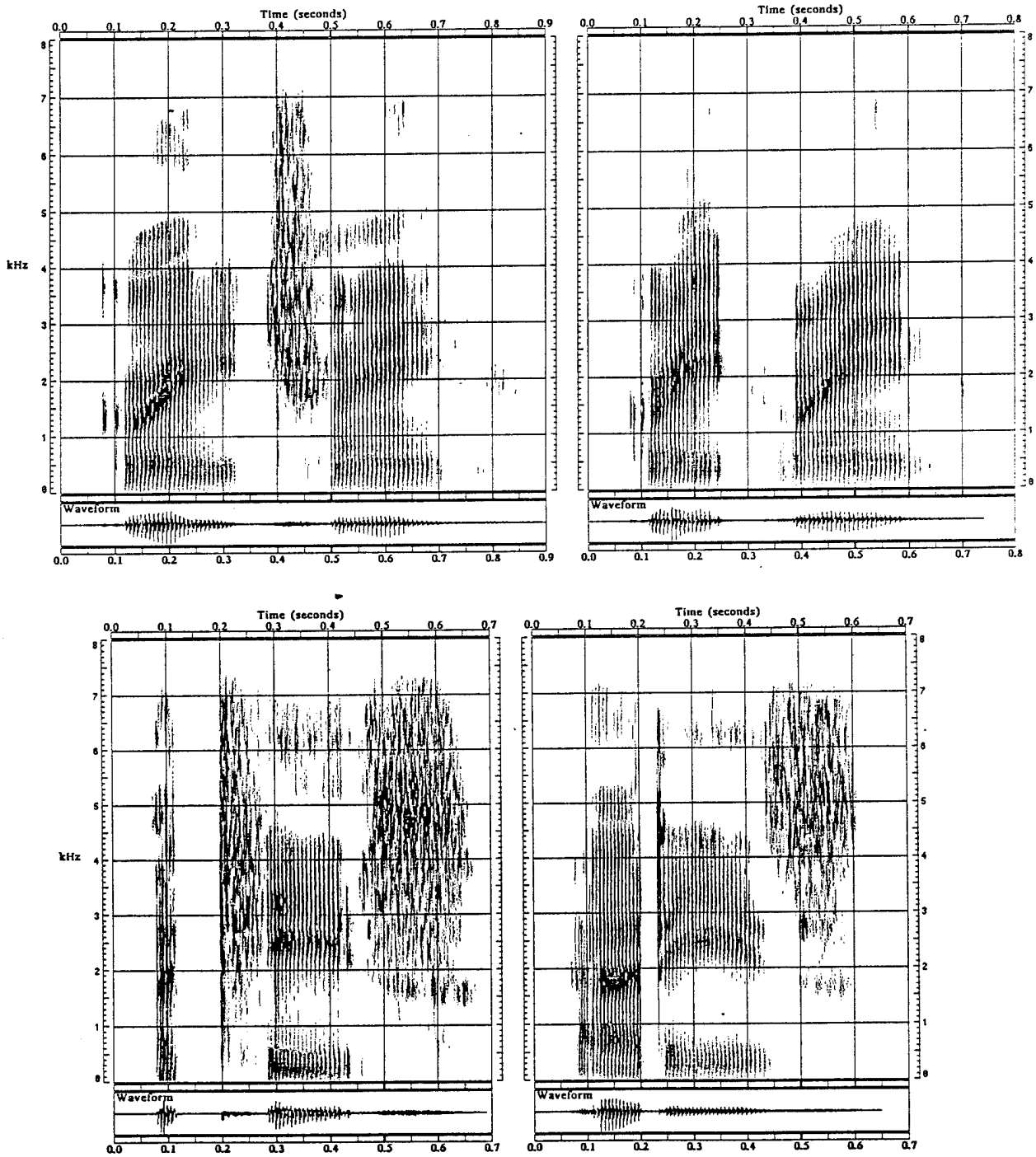


Figure 1.4: Word boundary effects.

Digital spectrogram of the word pairs 'grey train,' 'great rain,' 'a tease,' and 'at ease,' spoken by a male talker. The utterances illustrate a regular variation of the realization of the phoneme /t/ at word boundaries in American English.

1.2.1 Describing Speech Sounds

Since coarticulation causes the acoustic realization of a phoneme to vary as a function of context, scientists often find it useful to describe the sequences of speech sounds with a more precise label than the phoneme. The units of choice are usually *phones*, which have a rich inventory of symbols with which to describe sound segments in the speech signal [67]. Thus, the rounded /t/ and nasalized /ɛ/ in Figure 1.1 could be described by the phones [t̞], and [ɛ̃], respectively. The amount of detail that is useful depends on the purpose of the transcription. An underlying phoneme is typically described by a single *allophone*, although the inventory of units is often enhanced to include significant contextual variations in the realization of a phoneme [67]. For instance, a flap is often used to describe the production of the phoneme /t/ in a word such as 'butter' in American English.

Speech researchers often align a phonetic transcription with the acoustic signal in order to be able to measure properties of the individual phones [26,71]. Boundaries are typically located at places where significant feature changes take place, such as the onset or offset of voicing. Often, however, it is difficult to know where to place such boundaries [71]. This is especially true for the liquids /l/ and /r/, and glides /w/ and /y/, since their acoustic realizations often vary slowly in time [26]. An example of a time-aligned phonetic transcription is shown at the bottom of Figure 1.1. This figure illustrates that an aligned phonetic transcription essentially marks significant acoustic landmarks in the speech signal.

1.2.2 Sources of Variability in Speech

In addition to the sources of variation described previously, there are many extra-linguistic factors which also affect the acoustic properties of the speech signal. Some of these factors include:

- *Environment.* The acoustic signal often contains information about the physical environment which can complicate the decoding process [37]. If speech is recorded with a microphone, the properties of the signal depend on the recording apparatus and the nature of the recording conditions.

- *Inter-speaker differences.* The acoustic characteristics of speech sounds depend upon the physiological structure of the vocal apparatus. In particular, there can be large acoustical differences in the speech of men, women, and children. In addition, the speaking rate can vary from one speaker to another.

- *Intra-speaker differences.* The same speaker can pronounce an utterance differently on separate occasions for many reasons including sickness, mood, audience, and stress patterns on the word or phrase.

Finally, it is also possible for a speaker to occasionally distort the acoustic realization of a phoneme so severely that it cannot be identified, despite a knowledge of the phonetic environment [123]. These distortions are tolerable because, in addition to acoustic-phonetic knowledge, listeners are able to apply phonotactic, phonological, syntactic, prosodic, semantic, and pragmatic constraints to help recognize an utterance.

1.3 Decoding the Speech Signal

For many decades scientists have been seeking to understand the decoding process which maps the speech signal to words in a lexicon. An indication of the current state of knowledge may be obtained by measuring the degree to which this process can be emulated by a machine. A survey of available literature indicates that the current performance of automatic speech recognition systems falls well below human capabilities [23,58]. In order to obtain acceptable performance levels, recognition tasks are typically limited along one or more dimensions by restricting the vocabulary size, applying syntactic constraints, training the system to a single speaker and/or requiring

that words be spoken in isolation [52,75,123]. For many years in fact, the seemingly overwhelming amount of variation present in the signal combined with the slow rate of progress in quantifying acoustic-phonetic knowledge caused many researchers to speculate that there was a limited amount of phonetic information which could be recovered from the speech signal, and that the answer to speech recognition lay in incorporating constraint provided by sources such as syntax, semantics, and prosody [19,21,96]. In the following analogy, Hockett presents a common sentiment about the nature of the acoustic-phonetic information encoded in the speech signal [48]:

> "Imagine a row of Easter eggs carried along a moving belt; the eggs are of various sizes, and variously colored, but not boiled. At a certain point, the belt carries the row of eggs between the two rollers of a wringer, which quite effectively smash them and rub them more or less into each other. The flow of eggs before the wringer represents the series of impulses from the phoneme source; the mess that emerges from the wringer represents the output of the speech transmitter. At a subsequent point, we have an inspector whose task it is to examine the passing mess and decide, on the basis of the broken and unbroken yolks, the variously spread out albumin, and the variously colored bits of shell, the nature of the flow of eggs which previously arrived at the wringer."

A study by Cole and Zue nearly ten years ago clearly showed that the acoustic signal, even as displayed in the form of a spectrogram, is rich in phonetic information, and is not nearly as barren and impoverished as many had claimed [18,122]. However, phonetic recognition accuracies on the order of human levels of performance have yet to be attained by machine [52,58]. Although scientists are aware of many kinds of variability that affect the speech signal, a rigorous understanding of the relationship between the acoustic signal and items in the lexicon remains elusive. As illustrated in Figure 1.5, there are several ways to model the relationship between the speech signal and the lexicon. The following sections describe each of these approaches in more detail.

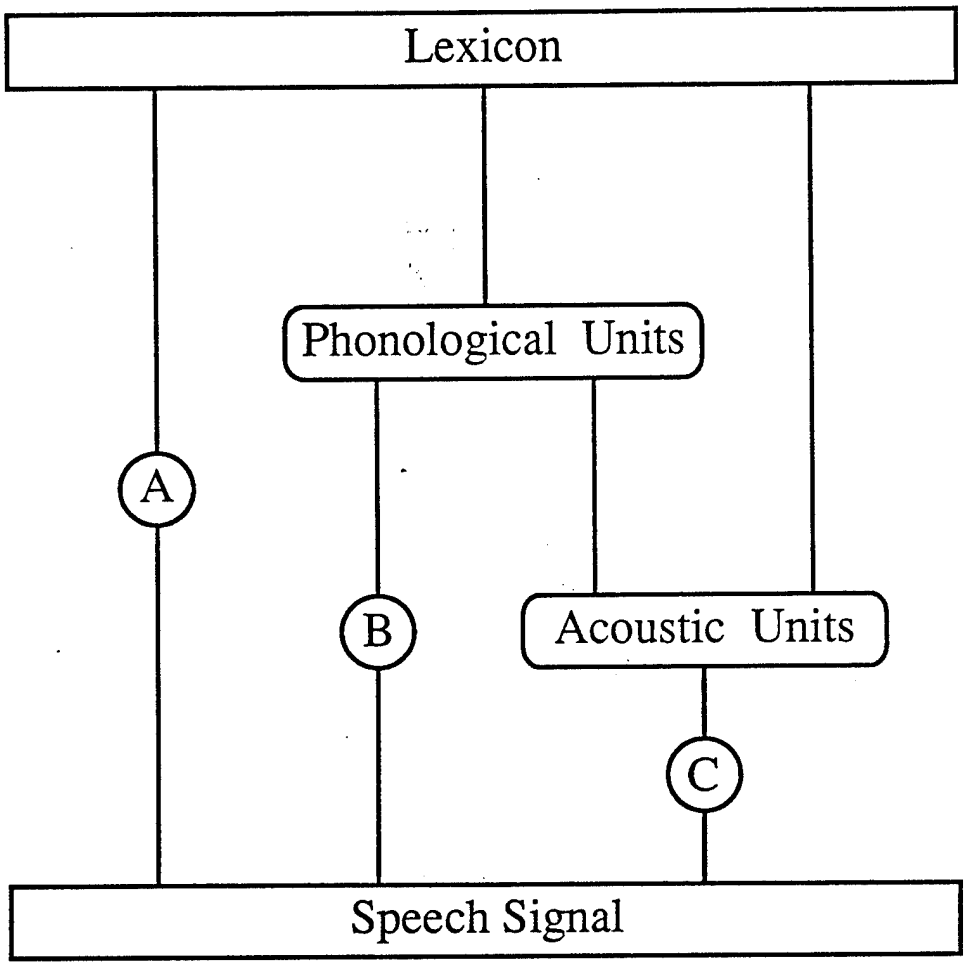


Figure 1.5: Alternative strategies for decoding the speech signal.

In path 'A' the signal is mapped directly to words in the lexicon. In path 'B' the signal is transformed into a set of phonological units which represent words in the lexicon. In path 'C' the signal is transformed into a set of acoustic units which can subsequently map to either phonological units or words in the lexicon.

1.3.1 Whole Word Units

The simplest model of the decoding process is one which assumes that there is no intermediate level of description of the speech signal. Each entry in the lexicon has a speech model which is independent from all other models. This approach corresponds to path 'A' in Figure 1.5. From an implementation perspective, this approach is highly attractive because of its simplicity. Many successful speech recognition systems have embraced this approach, and avoid any intermediate representation of the speech signal by performing whole-word recognition [23]. The advantage of this approach appears to be its ability to model word-internal variabilities with greater accuracy than approaches based on smaller sized units [13]. Using word-sized units usually limits the capabilities of these approaches to small-vocabulary tasks, due to the limited amount of training data available for individual words. For larger tasks, it appears that some form of sub-word unit is necessary [73,117].

In principle, approaches which advocate an intermediate linguistic level of representation should theoretically be able to handle more challenging tasks. This type of approach, which corresponds to path 'B' in Figure 1.5, is also supported by compelling evidence for the existence of an intermediate level of description of items in the lexicon [88]. For instance, it is difficult to explain common speech production errors such as substitutions and exchanges, without assuming some kind of segmental organization of words [39,105]. There is also evidence from studies of perception and short-term memory errors that words are represented in the lexicon in terms of segments [57,81,115]. In fact, one of the fundamental assumptions of linguistic analysis is that words can be represented as a sequence of discrete units [6]. For these reasons, whole-word matches have been compared to approaches more likely utilized by infants in the early learning stages of speech perception [56], or by animals responding to simple oral commands [34].

1.3.2 Phonological Units

As Fant pointed out over 25 years ago, there are at least three ways to describe the speech signal with an intermediate level of description [30]. First, speech may be considered to be a set of overlapping importance functions, each describing the strength of a particular phoneme at any given time. Alternatively, speech can be viewed as a series of asynchronous features with binary, or continuous degrees of strength. Finally, speech may be seen as a sequence of minimal sound segments, the boundaries of which are defined by distinct changes in the signal. These three perspectives are illustrated in Figure 1.6. Although these interpretations of speech might at first appear somewhat different, they are in fact quite compatible with each other.

Fant's first view of speech corresponds to how a phoneme is realized in the acoustic signal. Although every phoneme has a canonical articulatory configuration, the production of a sequence of phonemes does not produce a corresponding sequence of distinct acoustic segments. As was described previously, the articulators spend a finite amount of time moving from one articulatory configuration to another. Instead of a sequence of concatenated static articulations then, it is more appropriate to view the realization of phonemes as a continuous sequence of articulatory gestures. This is true for all sounds, but is especially applicable to time-varying sounds where there is often little or no stationary acoustic interval in the signal. Often, for example, the phoneme /t/ in a word such as 'butter' is realized as a flap, which is a quick, continuous gesture of the tongue tip moving to and from the roof of the mouth. Another example of a gesture is the realization of the phoneme /w/ in a word such as 'away,' where the lips are typically moving continuously to and from an extreme rounded position. These two effects are illustrated in Figure 1.7. The concept of such gestures, and the reality that there are no clear phonemic boundaries in the speech signal, are captured in terms of Fant's time-varying importance functions. For the most part, such functions are usually considered abstract, although there have been attempts to extract such functions from the acoustic signal [79]. Such a task would seem to be

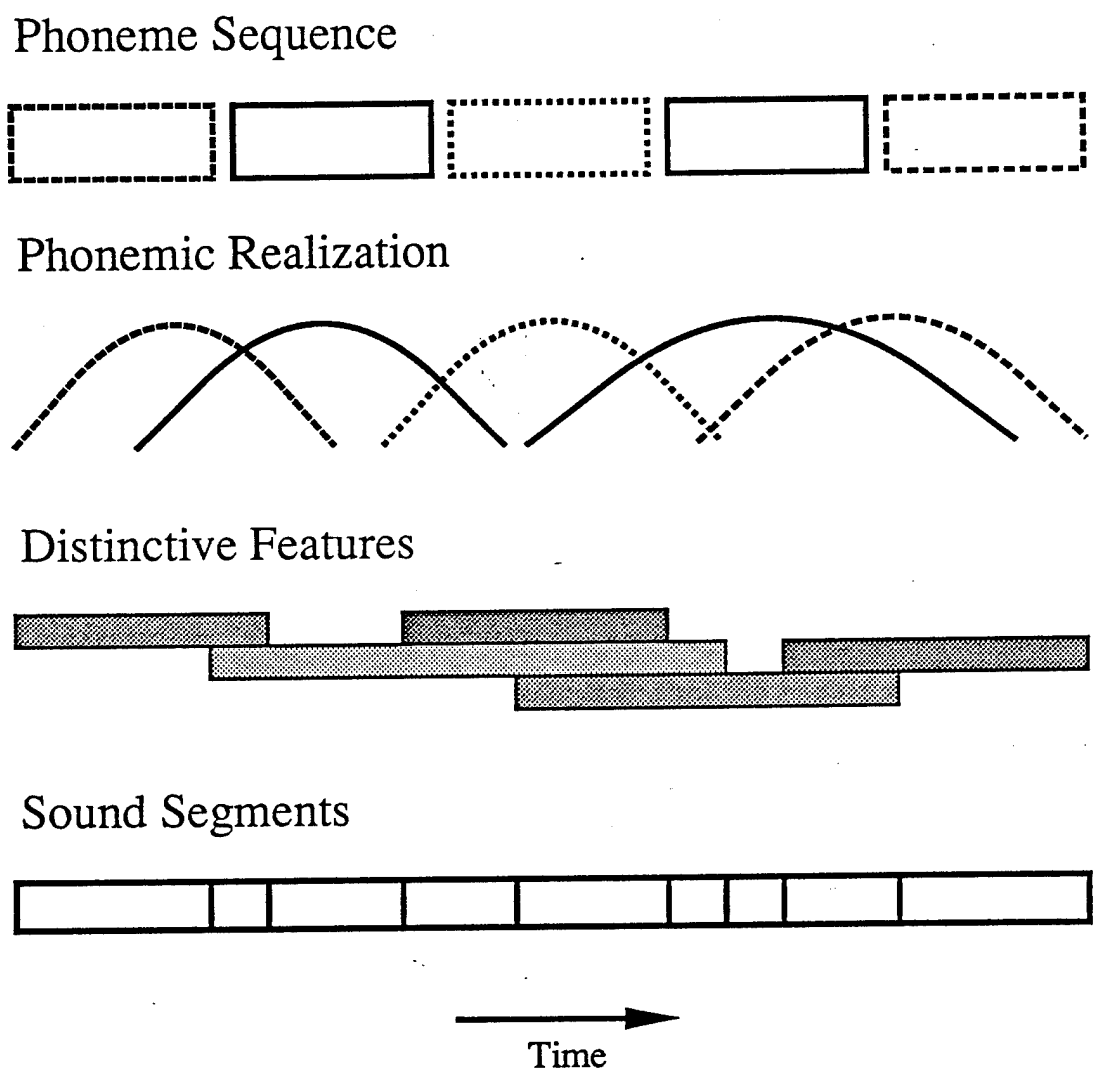


Figure 1.6: Alternative perspectives of speech.

Fant's perspectives of the speech signal [30]. From top to bottom they correspond to: (1) a sequence of ideal non-overlapping phonemes, (2) a sequence of continuously varying importance functions relating the influence of a phoneme on the speech waveform, (3) an overlapping sequence of features properties which may extend over several speech segments, and (4) a sequence of minimal sound segments, the boundaries of which are defined by distinct changes in the speech signal.

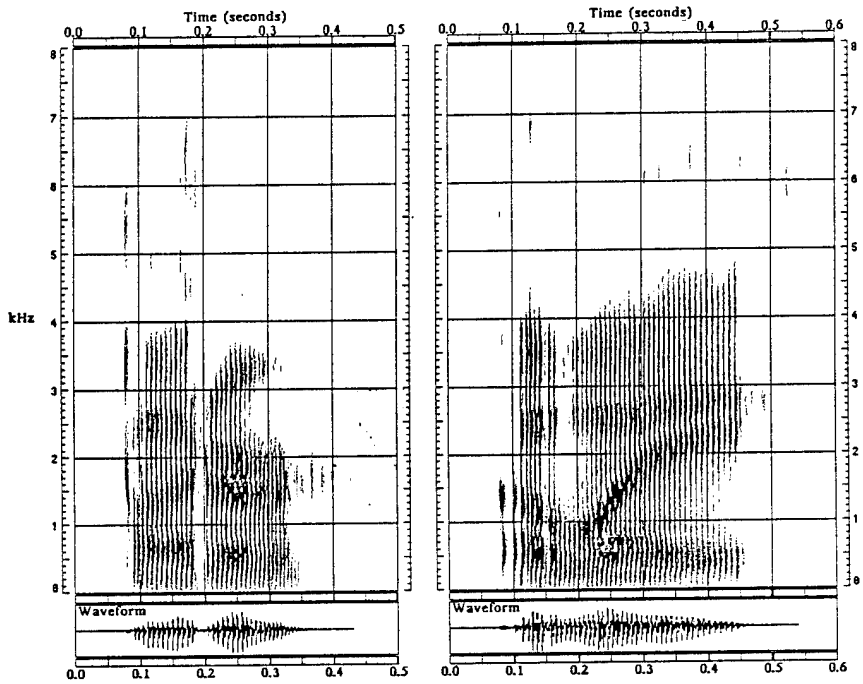


Figure 1.7: Spectrograms of the words 'butter,' and 'away.'

Digital spectrogram of the words 'butter,' and 'away,' spoken by a male talker.

extremely difficult, since scientists do not fully understand the complex relationship between articulator configurations and the acoustic signal.

Fant's view of speech as asynchronous feature functions corresponds directly to distinctive feature theory discussed previously. The uncertainty about the continuous or discrete nature of the features reflects the fact that scientists do not yet fully understand the acoustic correlates of these features [18,26,32,33,123]. Although theoretically such features properties are either present or absent, it is conceivable that in practice it might prove necessary to represent a property with some form of probability or strength [61].

Fant's final view of speech as a sequence of minimal sound units is also based on distinctive features. Instead of overlapping segments, however, the speech signal is divided into a sequence of minimal sound segments, where a segment boundary

is associated with the absolute or relative appearance or discontinuation along the time scale of one or more feature properties. In principle, this perspective is similar to that provided by a narrow phonetic transcription of the speech signal. In practice, however, the inventory of phonetic units used to describe the speech signal is usually only slightly larger than the number of phonemes found in the language [67]. This discrepancy points out a fundamental problem with the use of the phone as a phonological unit for decoding the speech signal.

Unlike phonemes or distinctive features, there are no constraints which limit the number of phones that can be used to describe the speech signal. Phones are descriptive units, and are intended to be used as a tool for describing the speech signal [67]. In this respect they are an extremely useful representation for scientists studying the languages of the world. However, there is no criterion defining how many phones are necessary or sufficient to describe the speech signal. It is therefore difficult to determine precisely how many phones are required to provide accurate coverage of the speech signal, and it is difficult to justify using one particular inventory of phones instead of another. The net result is that a selected inventory of phones can be somewhat arbitrary. This point can be verified by observing the various numbers of allophones incorporated by speech recognition systems which use the allophone as the basic phonological unit. In American English for instance, the number of allophones incorporated into these systems can range from 40 to 60 [54,59,96].

The issue of accountability is also faced by all variants of the allophone, such as diphones, tri-phones, or phonemes-in-context [14,59,86,116]. Typically, such approaches expand the inventory of phones by incorporating context, since contextual factors are known to have an important influence on the realization of a phone. Once again however, there is no rigorous criterion for determining when the inventory of units is complete. To illustrate this point, one need only consider the fact that the acoustic realization of a voiceless stop consonant will often depend on the stress of the local environment [76]. In a phrase such as 'a papa,' for instance, the first /p/ will usually be aspirated while the second /p/ could easily be unaspirated, even though the local

contexts of the two consonants are virtually identical.

1.3.3 Acoustic Units

All phonological approaches are similar in that they apply a preconceived notion of the nature of the intermediate units of representation of speech to the decoding process. While these representations are powerful linguistic descriptions of speech, it is reasonable to investigate whether there are others, given that there appear to be shortcomings associated with the representations described previously. Specifically, it would be interesting to know if there were a set of descriptive units motivated from an *acoustic* perspective. Such units would be compatible with Fant's view of speech as a sequence of minimal sound segments, but would derive their origin from acoustic, rather than linguistic, criteria.

An inventory of acoustically motivated units could be used as a criterion for justifying a set of allophones to represent the speech signal. Such a set of units might suggest additional allophones which might have been otherwise overlooked. By searching for acoustic units of speech, researchers would be provided with a more rigorous criterion for selecting an inventory of allophones.

Alternatively however, it is reasonable to question whether a phonetic level of description could be bypassed altogether. If the task of phonetic decoding is viewed as a general process whereby the acoustic signal is mapped to a set of phonological units representing a lexicon, it is possible to consider an alternative approach to decoding the speech signal. As illustrated by path 'C' in Figure 1.5, the speech signal could be first transformed into a sequence of acoustic units before being subsequently mapped to underlying phonological units such as phonemes. From Figure 1.5 it is also clear that it is possible to avoid a phonological representation entirely by directly mapping the acoustic units to words in the lexicon [73,117]. At this point, it is perhaps too soon to tell which strategy would be more appropriate. It would seem however that

a direct mapping between the lexicon and acoustic units would sacrifice some of the flexibility afforded by a phonological level of representation.

Before discussing the role of acoustic units in the decoding process any further however, it is necessary to consider what criterion may be used to define such units. As mentioned previously, Fant's criterion of an acoustic segment is essentially based on distinctive features [28,29,30]. Boundaries are associated with the absolute or relative appearance or discontinuation along the time scale of one or more feature properties. As Fant pointed out, there are some practical problems in implementing this scheme because it is difficult to tell acoustically where transitions occur in some features. His definition of a boundary is therefore a subjective one, and so depends to some extent on the investigator if segment boundaries are marked manually. The lack of a clear acoustic criterion for defining a boundary eliminates any possibility of automatically determining these sound segments. A more serious problem with this definition of a sound segment is that it does not gracefully describe momentary or time-varying articulations, such as a flap. Fant attempts to remedy this problem by including additional feature properties identifying transitions. No objective criterion for defining these properties was proposed.

Catford also discussed the issue of segmentation, and suggested that the speech signal may be delineated into a sequence of sound segments [11]. His definition of a segment was tied to articulatory features. In a fashion similar to Fant, he first tried to define segment boundaries as corresponding to some change in the underlying features. Likewise, he also noted the difficulties in describing momentary or continuous gestures. In order to provide a single general definition for a segmental unit, he proposed that a segment be a quasi-stationary articulatory posture flanked by an approaching onset and a departing offset. This definition held even when the intervening steady period was reduced to zero.

Catford's definition of a sound segment provides the basis for an objective criterion for defining an *acoustic* segment. In this view speech may be described as a sequence of acoustic events corresponding to onsets and offsets in some representation of the

speech signal. An acoustic segment corresponds to the interval of speech spanning successive events. Note that this definition does not require the intervening interval to be a steady-state sound. In the acoustic realization of the /w/ in a word such as 'away', an onset and offset would delineate an acoustic segment, even though the intervening interval is unlikely to be a steady-state sound. This definition of an acoustic segment does not rule out the importance of alternative acoustic landmarks. For instance, there are many sounds for which some extreme point of articulation would be an appropriate point to measure, such as the point of maximum closure in the /w/ in the word 'away' [26]. However not all speech sounds exhibit such extrema, and the precise nature of the extrema depends on the particular sound segment. Thus, it is difficult to propose a general definition for an acoustic segment in terms of extrema in the signal representation. For this reason, an acoustic segment is defined in terms of an onset and offset, since they are related to the notion of an articulatory gesture, and would seem to capture a basic property of all speech sounds.

The important result of viewing speech as consisting of acoustic onsets and offsets, is that it provides a mechanism for deriving a purely acoustic description of speech. The first step in the decoding process then consists of determining the location of onsets and offsets in the acoustic signal. The next step is to map each segment to one of a finite set of acoustic units. This in turn precipitates the need for a mechanism for determining an appropriate set of acoustic units.

In Fant's original description of speech as a sequence of minimal sound segments, each acoustic segment could be described by a bundle of phonetic features. If each feature could be mapped to one or more acoustic correlates, then there would be a one-to-one mapping between points in the phonetic feature space and points in the space represented by all corresponding acoustic correlates. Determining the inventory of acoustic units would be straightforward, since they would correspond to the points which inhabited this acoustic space.

As was pointed out previously, this perspective of speech is idealized, since scientists do not yet fully understand the acoustic correlates of the distinctive features.

It remains to be determined if correlates can be assigned binary values. Presently, a point in the phonetic feature space is more likely to map to a region, or a cloud of points, in the acoustic space, rather than a single point, since the acoustic correlates have varying degrees of strength. Depending on the compactness of the clouds, it may still be possible to determine an inventory of units to describe the sound segments. Each unit would correspond to a cloud of points, and could be represented by a mean or centroid. This can be viewed as a clustering problem, since it is necessary to group all similar points into the same class, while distinguishing among sounds which are dissimilar.

Since a sound segment is represented by a bundle of phonetic features, it is not necessary to maintain a direct relationship between a phonetic feature and one or more dimensions of the acoustic space. Any acoustic representation of the speech signal could be used, although it is clear that some representations would be better than others. A desirable representation is one in which sound segments having the same set of phonetic features are tightly clustered in the acoustic space, and well separated from other such clusters. However, most acoustic representations of the speech signal fall within the general framework of defining acoustic units in terms of acoustic clusters.

One of the requirements for clustering approaches is for an extensive amount of data. The more points which can be gathered, the more robustly clusters will be able to be observed. With the availability of a large body of data [69], it is now feasible to perform such an investigation. The objective of this thesis is, in fact, to develop a methodology with which it is possible to search for meaningful acoustic regularities in the speech signal. More specifically, it investigates if there is evidence for acoustic segments in speech, and if there is evidence for regular patterns of behavior in these acoustic segments. The following section describes the scope of the thesis in more detail.

1.4 Thesis Scope

Since a framework for automatically determining acoustic regularities does not yet exist, the major undertaking of this thesis is to generate a purely acoustic description of the speech signal, and to show how it is possible to capture significant and meaningful regularities on the basis of acoustic information alone.

The prime objective of this work is to automate all aspects of the proposed framework. There are three reasons for this approach. First, using well defined criteria for description and organization reduces the number of subjective decisions involved. Second, automating these processes allows for a much more extensive amount of data to be observed than would otherwise be possible. Presumably these two factors would improve the robustness and reliability of any conclusions drawn from a study of the data. Finally, precisely because of its objective nature, this machinery can act as an independent source of evidence for any hypotheses about the nature of speech. Thus, it would appear to be a useful aid in speech analysis since it serves as a mechanism to gain insight into the relationship between underlying phonological units and their acoustic realizations.

Any attempt to automatically determine acoustic regularities in the speech signal must consider three important areas concerned with representing, describing, and organizing the speech signal. In the following chapter, the signal representation used for all subsequent acoustic analysis will be described. Specifically, the representation is based on the mean-rate response outputs of an auditory model developed by Seneff [103]. This chapter also describes the database used for all subsequent speech analysis.

In Chapter 3, a procedure is described for automatically locating acoustic landmarks in the speech signal. The particular approach which is investigated is based on the belief that many different kinds of landmarks are important for speech analysis, and that it is extremely difficult to capture all of these landmarks with a single level of description. Thus, in the approach taken, an attempt is made to develop a procedure which provides a multi-level description of the speech signal. Such a description is

motivated more fully in the chapter, and is described and evaluated using the dataset described previously.

In Chapter 4, a procedure is described for automatically finding acoustic regularities in the speech signal. The goal of the algorithm is to produce an accurate description of acoustic events in the speech signal. Thus, realizations of phonemes that are acoustically similar should fall into the same class. If, on the other hand, a phoneme falls into more than one class, then the different acoustic realizations should suggest the presence of important contextual variations. After developing the clustering procedure, it is applied to two different tasks. First, an examination of all speech sounds is made, without assuming any number of units *a priori*. The goal in this investigation iss to determine what major acoustic regularities could be observed using spectral cross-sections of the mean-rate response outputs as the source of acoustic information. The second investigation examines several individual phonemes in order to illustrate how an acoustic clustering of data can help to suggest major sources of variability in the realization of a phoneme.

In Chapter 5, some further work is described which attempts to improve the temporal and spectral properties of the mean-rate response. Furthermore, a study is reported which attempts to measure the statistical distribution of the mean-rate response outputs in order to determine if a more compact representation of the outputs can be motivated.

The final chapter summarizes the results of this work, and discusses the application of the ideas presented in this thesis as a viable approach to decoding the speech signal.

Chapter 2

Signal Representation and Database

2.1 Signal Representation of Speech

There are at least two important factors to consider when selecting a representation of the speech signal. Certainly the most critical requirement of any representation is that it preserve the important information-bearing elements of the speech signal. Another desirable property of a representation is that it suppress irrelevant details in the speech signal. Thus, the choice of a representation often depends on the particular task. A useful representation for speaker identification might not be applicable for a word recognition task for example, because individual speaker characteristics are a source of extra-linguistic information in the speech signal.

Historically, the short-time spectrum has played a major role in speech analysis. In this representation, the speech signal is described as a temporal sequence of spectra. Transitions in the underlying acoustic signal show up as differences between adjacent spectra. The use of this type of representation is supported by speech production theory which indicates that the natural resonances of the vocal tract provide a concise description of most speech sounds [28]. There is also a substantial amount of physiological evidence suggesting that the ear performs a form of spectral analysis at the early processing stage [36]. This indicates that properties relevant to the perception of speech can be contained in a spectral representation of the speech signal. Most forms of spectral analysis, such as those based on the Discrete Fourier Transform or

Linear Prediction, do not incorporate other known properties of the human auditory system [85,94,99]. Although the importance of such properties remains unanswered, it is clear that the evolution of speech was influenced by the constraints of both the production and the perception mechanisms [110]. Any attempt at finding acoustic regularities in the speech signal would probably be well served by paying attention to the natural constraints of the auditory system, since it will serve to focus the attention on details which are likely to be perceptually important.

Although there are many aspects of human audition that are as yet unexplained, scientists have begun to build systems which model the auditory process up to the level of the auditory nerve [1,16,40,77,103]. In particular, the model which formed the basis for all of the spectral representations used in this thesis was one developed by Seneff which incorporates several known properties of the human auditory system, such as critical-band filtering, half-wave rectification, adaptation, saturation, forward masking, and spontaneous response [103].

Seneff's model of the auditory system has several stages. Those stages of her model that are relevant to the current research are illustrated in Figure 2.1. The first stage is a bank of 40 linear filters equally spaced on a Bark frequency scale, with center frequencies spanning a range from 130 to 6,400 Hz. The frequency response of these critical-band filters is shown in Figure 2.2. The envelope outputs of this stage are known as the critical band envelopes. The second stage of Seneff's auditory model attempts to model the transformation from basilar membrane vibration to nerve fiber response. This part of the model incorporates such non-linearities as dynamic range compression, half-wave rectification, short-term and rapid adaptation, and forward masking. The outputs of this stage correspond to a probability of firing on the auditory nerve. The envelope outputs of this stage correspond to a short-term average, or mean probability of firing, and are therefore called the mean rate response. A final stage of Seneff's auditory, not shown in Figure 2.1, determines the synchronous response of each filter by measuring the extent of dominance of information at the filter's characteristic frequency. The synchrony response, however, is not considered

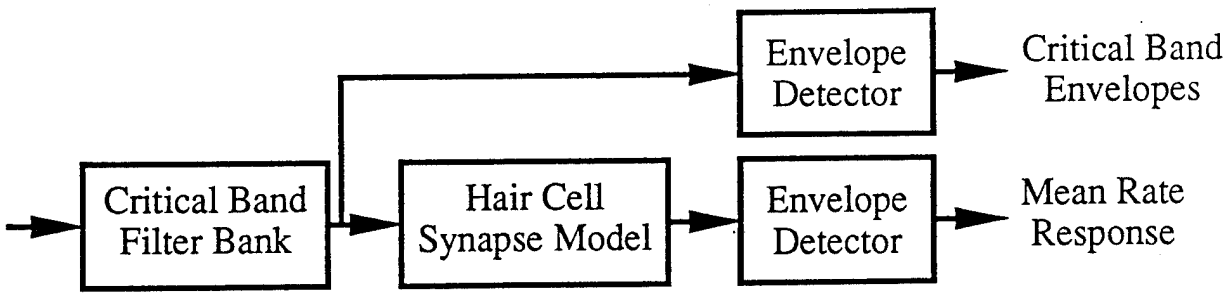


Figure 2.1: Block diagram of parts of Seneff's auditory model.

for use in this thesis.

The mean-rate response outputs appear to enhance important acoustic landmarks in the speech signal. There are two aspects to this enhancement. As illustrated in Figure 2.3, the onsets and offsets from one sound to another appear to be sharper than is the case for the critical-band filter outputs. Second, forward masking appears to greatly attenuate many low-amplitude sounds because the output falls below the spontaneous firing rate of the nerve fibers. These two effects combine to sharpen acoustic boundaries in the speech signal. Since the characteristics of the mean-rate response seemed particularly well suited to the goals of this thesis, these outputs were used for all aspects of this work.

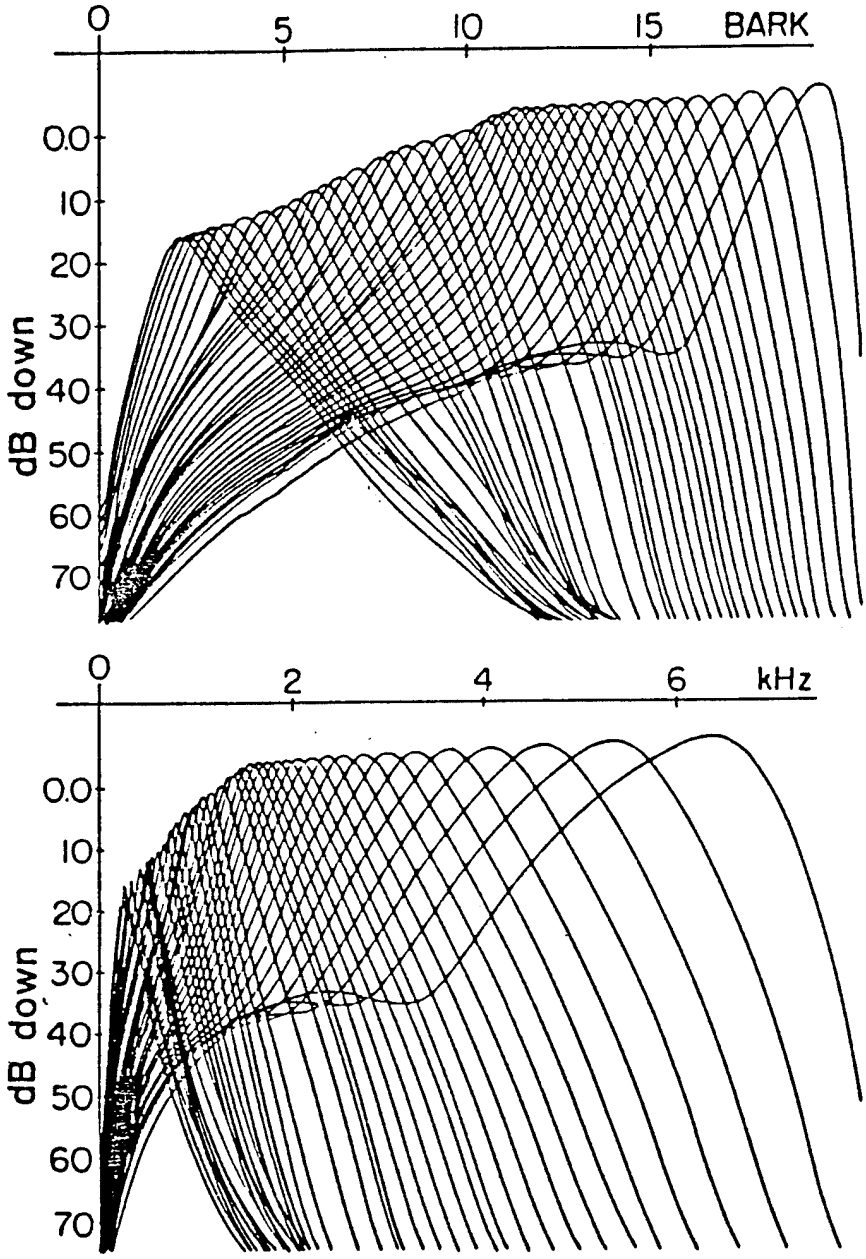


Figure 2.2: Critical-band filter frequency response (after Seneff).

This figure illustrates the frequency response of the critical-band filters in Seneff's auditory model [103]. The displays shows the response characteristics plotted along a Bark scale (top) [125], and a linear frequency scale (bottom).

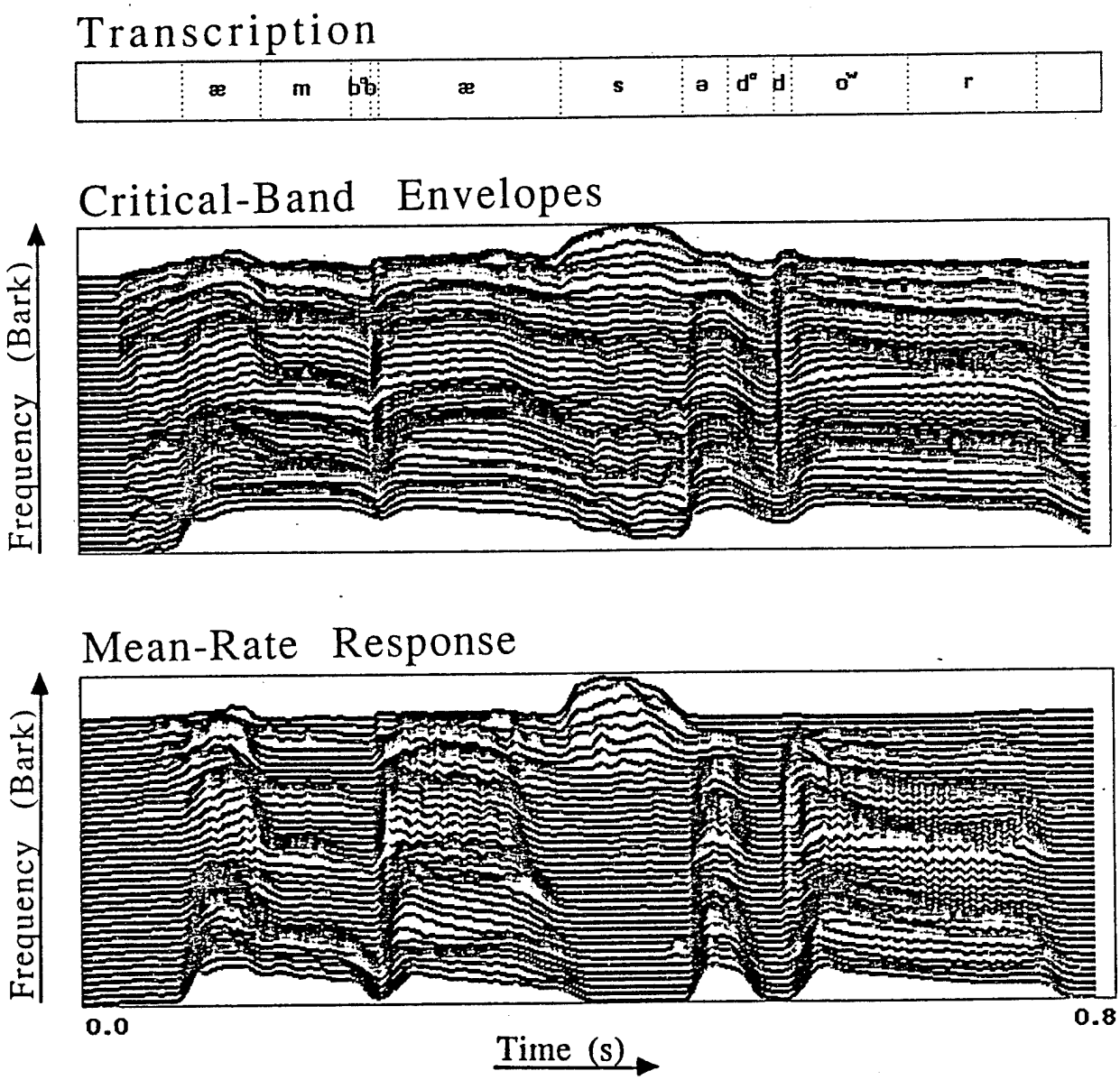


Figure 2.3: Comparison of two spectral representations.

A comparison of critical-band envelopes (top), and mean-rate response (bottom) for the word 'ambassador.'

2.2 Database Description

Before proceeding to the following chapters, it is appropriate to describe the database used for all speech analysis in slightly more detail. The data used for this work are a subset of the TIMIT acoustic-phonetic database, which was recorded at Texas Instruments, and phonetically transcribed at MIT [35,69]. The entire database consists of 10 sentences recorded from each of 630 talkers. Two of the sentences were calibration sentences; the same two were spoken by all talkers. These sentences are useful for dialectical studies of American English [17]. Five of the sentences were drawn from a set of 450 phonetically compact sentences hand-designed at MIT with emphasis on as complete a coverage of phonetic pairs as is practical [69]. The remaining three sentences were randomly selected sentences from the Brown corpus, and were intended to provide examples of typical American English sentences [65,35].

All work reported in this thesis used the phonetically compact sentences of the first 200 talkers, representing a total of 1000 sentences. Typical examples of these sentences may be found in Table 2.1.

Table 2.1: Typical sentences from the TIMIT database.

"A muscular abdomen is good for your back."
"Any contributions will be greatly appreciated."
"A doctor was in the ambulance with the patient."
"Rob sat by the pond and sketched the stray geese."
"Bagpipes and bongos are musical instruments."
"Even a simple vocabulary contains symbols."
"Ambidextrous pickpockets accomplish more."
"Coconut cream pie makes a nice dessert."
"My ideal morning begins with hot coffee."
"The small boy put the worm on the hook."

Chapter 3

Acoustic Segmentation

In Chapter 1, it was proposed that speech may be delineated into a sequence of acoustic segments, where a segment corresponds to the interval of speech spanning an onset and offset in some representation of the speech signal. In this chapter, a procedure is described for automatically locating acoustic landmarks in the speech signal. In addition to establishing a criterion for locating acoustic landmarks, an algorithm is developed which provides a multi-level description of these landmarks. This approach is motivated by the observation that it is extremely difficult to capture all important events with a single level of description. Finally, an analysis of the multi-level description is made by comparing landmarks in the acoustic structure to major acoustic-phonetic boundaries marked in a set of 500 manually transcribed utterances.

3.1 Finding Acoustic Landmarks

Given that an acoustic segment is delineated by an onset and offset, it is possible to construct procedures which attempt to automatically determine the location of these events. There are, in fact, at least two different perspectives from which to view this problem. One viewpoint is that an onset or offset represents a local maximum in the rate of change in some parameter representing the speech signal, since at these points the signal is undergoing significantly more change than in the neighboring environment. This phenomenon is illustrated in Figure 3.1 for part of the word 'international.' By observing the distance between adjacent points in this figure, it

appears that the signal changes more rapidly at the boundaries between phones, and is more stable within phones themselves.

Figure 3.1 suggests another view, which is that speech consists of a temporal sequence of quasi-stationary acoustic segments. From this perspective, the points within an acoustic segment are more similar to each other than to the points in adjacent segments. This criterion for an acoustic segment can be seen simply as a local clustering problem whereby it must be decided if any particular frame is more similar to the sounds immediately preceding or following it. Viewing the problem of finding acoustic segments from this perspective offers the advantage that the clustering procedure can be quite sophisticated if desired, and can be made to be adaptable to both short and long duration segments.

Both views of finding acoustic events offer the advantage that only relative measures are made, so that neither thresholds nor training are involved. Another potential advantage of making relative local measures is that such procedures may be insensitive to background noise or long term changes in the signal. Furthermore, such algorithms do not need to be tuned to the individual characteristics of a speaker. Finally, this procedure does not make use of the entire utterance, so it is capable of analyzing the speech signal virtually instantaneously. The following sections discuss these approaches in more detail.

3.1.1 Locating Points of Maximum Change

The first view of acoustic events corresponds to finding local maxima in the rate of change of some multi-dimensional representation of the speech signal, $\vec{S}_t$. In this thesis, rate of change is defined to be the magnitude of the first derivative, $\dot{\vec{S}}_t$. Since most machine analysis of speech is performed in a discrete manner, the derivative operation needs to be approximated by a discrete operator. If the backward difference were used [85], a discrete rate of change function would be given by

$$\|\vec{S}_n - \vec{S}_{n-1}\| \tag{3.1}$$

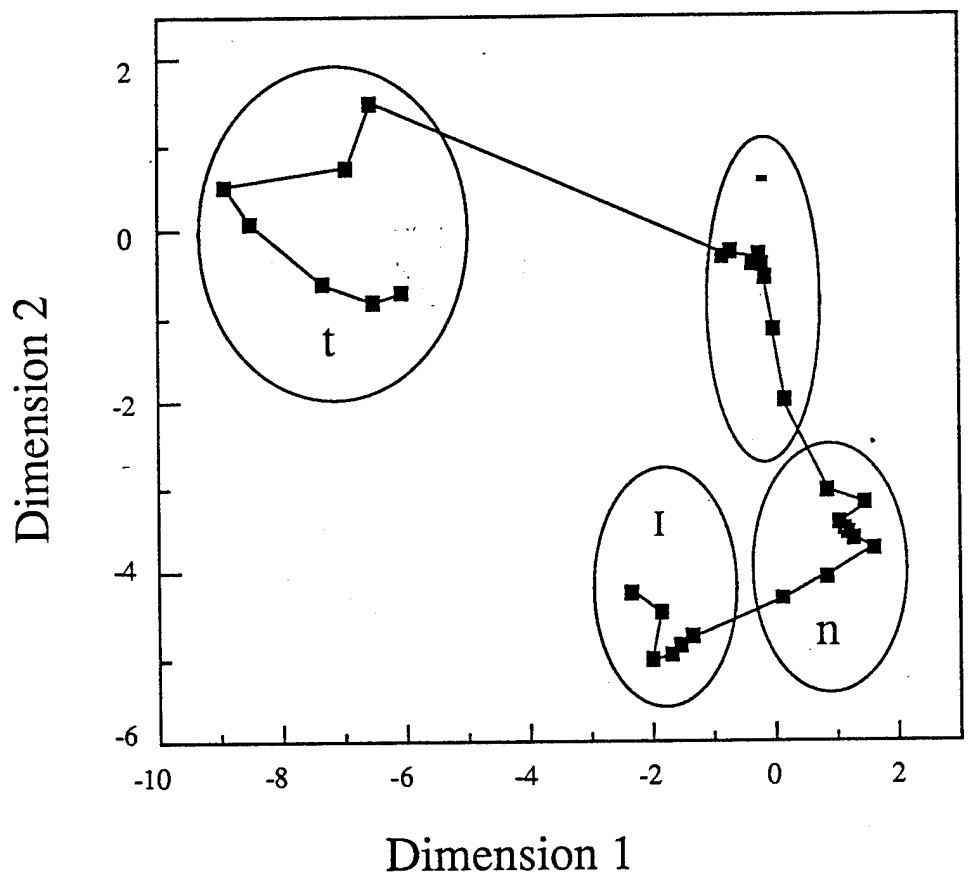


Figure 3.1: Partial trajectory of the word 'international.'

This figure illustrates a trajectory of the first four phones of the word 'international,' spoken by a male talker. Each point in the plot corresponds to a output of the mean-rate response sampled every 5 ms. The dimensions of the plot correspond to the values of the first and second principal components which are derived in Chapter 5. Points which form a subset of transcribed phonetic units have been circled. The third phone corresponds to the period of complete closure before the release of the /t/.

This function computes a distance between consecutive points in the signal representation, which matches the earlier qualitative analysis of Figure 3.1. In practice, such a measure of change is quite sensitive to small fluctuations in the signal representation, $\vec{S}_n$, so the representation is often smoothed before being differenced.

An alternative approach is to use an operator such as a derivative of a Gaussian to perform both functions at the same time. The Gaussian function is given by

$$g_t(\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{t^2}{2\sigma^2}}$$

and has a derivative with a maximum at time $-\sigma$, and a minimum at time σ. For a discrete approximation to a derivative, the negative of the Gaussian derivative is often sampled directly

$$d_n(\sigma) = -\frac{d}{dt} g_t(\sigma)\Big|_{t=nT}$$

where T is the sampling period. A new function for rate of change is then given by

$$c_n(\sigma) = \|\vec{S}_n * d_n(\sigma)\| \tag{3.2}$$

where each dimension of the $\vec{S}_n$ is convolved with $d_n(\sigma)$. Note that this equation is a simple measure of change, where the sensitivity is controlled by the σ of the derivative of the Gaussian filter. In essence, $c_n(\sigma)$ compares portions of the representation, $\vec{S}_n$, separated by 2σ. For $\sigma = \frac{T}{2}$ for example, the function $d_n(\sigma)$ could be approximated by the sequence $\{-1, 1\}$, which reduces Equation 3.2 to Equation 3.1.

In this thesis, the signal representation used was the mean-rate response described in the last chapter. Again it is important to point out that many alternative representations of the speech signal could be used, since $c_n(\sigma)$ is a general definition of change. For a representation such as the mean-rate response, this function captures energy changes in each channel. Onsets and offsets will show up as large values of $c_n(\sigma)$. More subtle changes in the representation, such as formant transitions, will also show up as changes, since energy is moving from one channel to another. Figure 3.2 illustrates several outputs of $c_n(\sigma)$ for values of sigma of 5, 10, and 15 ms. All operations were computed with a sampling period of 5 ms.

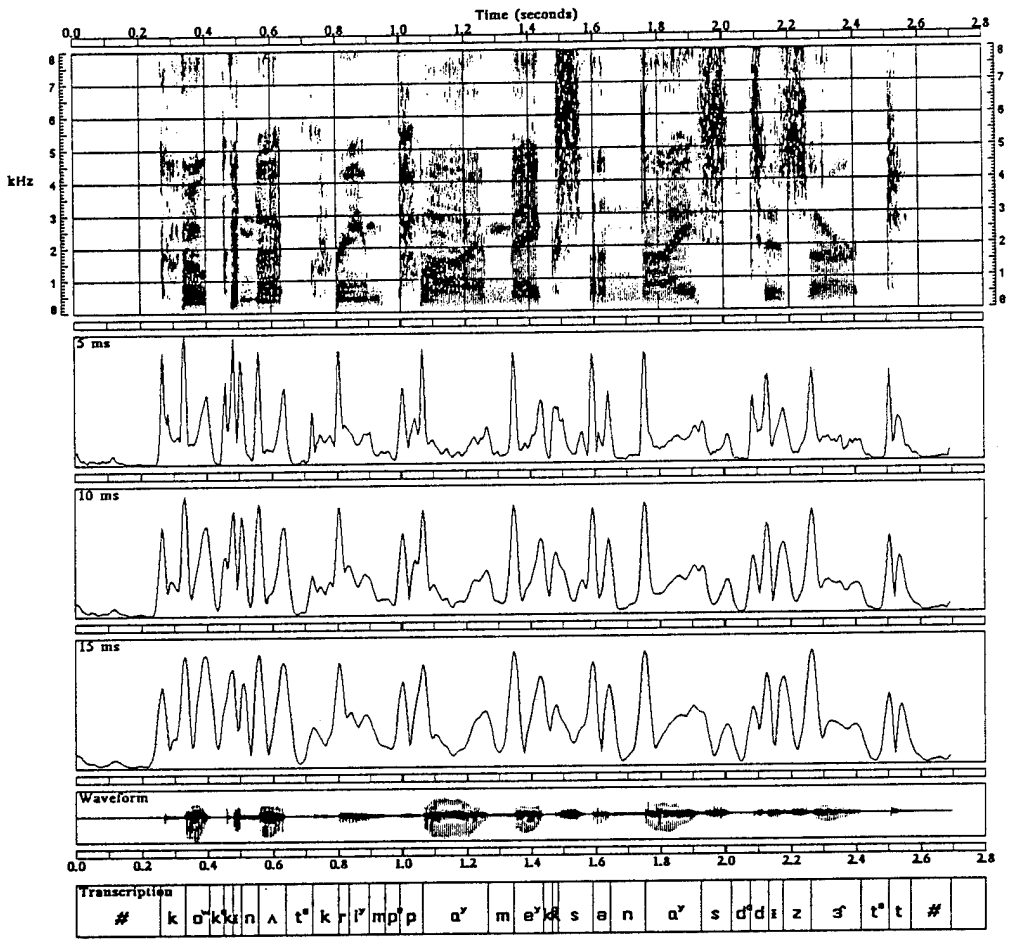


Figure 3.2: Different measures of change.

This figure illustrates computation of change on an utterance 'Coconut cream pie makes a nice dessert,' spoken by a female talker. A spectrogram of the utterance is shown at the top. The middle three displays correspond to the outputs of the function $c_n(\sigma)$ defined in Equation 3.2 for three different sensitivities, $\sigma = 5$, 10, and 15 ms. The bottom two panels display the speech waveform and an aligned phonetic transcription.

Figure 3.2 illustrates quite clearly the difficulty involved with automatically locating meaningful acoustic landmarks in the speech signal. In some instances landmarks are extremely abrupt, such as at the release of a stop. Often, such events are quite short in duration, as is the case for the second /k/ in the word 'coconut' which is approximately located at time 0.5 s. In this example, the release of the /k/ is followed shortly thereafter by the onset of voicing in the following vowel. In order to be able to distinguish these two events, and to be able to accurately determine their location in the speech signal, a fine level of sensitivity in the rate of change function, $c_n(\sigma)$, is necessary. Note that two distinct maxima may be seen in $c_n(5)$, but only a single maxima is found in $c_n(15)$. Conversely however, a narrow perspective of the speech signal can make it difficult to distinguish gradual changes in the speech signal, such as transitions between vowels and semivowels, from short-term perturbations. As a result, a function such as $c_n(5)$ will tend to be quite noisy, and will introduce many spurious events. This phenomena is illustrated in the final vowel /ɝ/, which has a gradual offset which is difficult for the $c_n(5)$ to distinguish from perturbations occurring within the vowel itself.

3.1.2 Associations

The second perspective for finding acoustic landmarks introduced earlier consists of performing a local clustering, where it is necessary to determine if a point in the multi-dimensional space is more similar to the points which immediately precede or follow it. At a simple level, this strategy is similar to the approach developed in the previous section. Consider, for example, the case where $\sigma = T$. In this case, the derivative of Gaussian operator, $d_n(\sigma)$, may be approximated by a $\{-1, 0, 1\}$ sequence so that the rate of change function is given by

$$c_n(\sigma) = \|\vec{S}_{n+1} - \vec{S}_{n-1}\|$$

Local maxima are located in this function by looking for positive-to-negative zero-crossings in the first derivative of this signal. If the same derivative of a Gaussian

operator is used to perform this computation, then the process reduces to looking for positive-to-negative zero-crossings in

$$a_n = \|\vec{S}_{n+2} - \vec{S}_n\| - \|\vec{S}_n - \vec{S}_{n-2}\| \tag{3.3}$$

Alternatively, if forward and backward distances are defined respectively as

$$f_n = \|\vec{S}_{n+2} - \vec{S}_n\|$$

and

$$b_n = \|\vec{S}_n - \vec{S}_{n-2}\|,$$

then Equation 3.3 can be viewed as a comparison between two distances. When a_n is positive, the forward distance, f_n, is greater than the backward distance, b_n. Thus, the point $\vec{S}_n$ is more similar to its immediate past than its immediate future. When a_n is negative, the opposite is true. A positive-to-negative transition in the value of a_n therefore corresponds to a switching in association from the past to the future.

In the previous example, the forward and backward distances were computed by comparing the point $\vec{S}_n$ to points two frames away. The distance criterion may be generalized however so that the forward distance, f_n, compares the point $\vec{S}_n$ to some set of points immediately following it. Conversely, the backward distance, b_n, compares the point $\vec{S}_n$ to some set of points immediately preceding it. A more general definition allows the possibility of adapting the sensitivity of the associations to the local environment, and thus alleviating some of the problems described in the previous section which are inherent with a fixed level of sensitivity in the rate of change function, $c_n(\sigma)$.

One such algorithm was developed for the task of detecting nasal consonants in continuous speech [41]. An indication of the use of this algorithm for analysis of all sounds is provided by a study which compared acoustic landmarks determined by the associations algorithm with boundaries provided by an aligned phonetic transcription [42]. A quantitative measure of the match between the two descriptions was obtained by tabulating boundary insertions and deletions on a set of 500 utterances over a wide

range of sensitivity of the associations algorithm. The minimum total insertion and deletion error observed was 25% of the actual number of phonetic boundaries. All analyses indicated that it was not possible to describe all acoustic events of interest with a single level of description. If the sensitivity was set too high then too many events would be postulated. If the sensitivity was set too low however, then significant events would be missed.

One way around this dilemma is to incorporate all events into a single multi-level structure where less significant events tend to inhabit the lower levels of the structure, while more significant events propagate to the higher levels in the structure. Such a structure would allow for a comprehensive acoustic description of the speech signal in a single coherent framework. The advantage of this type of representation is that it captures both coarse and fine acoustic information in one uniform structure. Acoustic-phonetic analysis is then a path-finding problem in a highly constrained search space. The following section describes this approach in more detail.

3.2 Multi-level Acoustic Description of Speech

3.2.1 Scale-Space Filtering

In the signal processing literature, one of the more interesting procedures for generating a multi-level description of a signal is scale-space filtering, which was first proposed by Witkin [120]. This procedure determines the structure of one-dimensional signals by looking for inflection points at successively greater amounts of smoothing. In practice, this is usually achieved by convolving the signal with a second derivative of a Gaussian, and looking for zero crossings in the resulting output. The amount of smoothing is controlled by the σ of the Gaussian. One of the important properties of a Gaussian filter is that new zero crossings will not be introduced as the amount of smoothing increases [4]. This property enables the scale-space structure to be interpreted more easily, since the location of the zero crossings can be tracked as a function of σ.

For an analysis of the change function, $c_n(\sigma)$, the important locations in the signal correspond to local maxima. Thus, it is more appropriate to convolve this signal with a first derivative of a Gaussian, and subsequently to look for zero crossings. An example of this particular scale-space procedure is shown in Figure 3.3. The filtering operation was performed using the change function $c_n(5)$, which is the top change display in Figure 3.2. This input was used since it was able to capture most events of interest. In Figure 3.3, the dark lines correspond to local maxima in the spectral change waveform, while the dashed lines correspond to local minima. This figure clearly illustrates that, as the amount of smoothing increases, the locations of the zero crossings of an adjacent maximum and minimum will converge until at a certain level they cancel each other out. At the lowest, and most sensitive, level of the scale-space structure, acoustic landmarks correspond to all local maxima in the change function $c_n(5)$. At higher, and less sensitive levels, however, only the more robust local maxima are preserved.

The scale-space procedure appears to be a useful mechanism for describing the structure of the acoustic signal since it incorporates all acoustic events into a single framework [78,119]. However, there are at least two shortcomings of this procedure for the analysis of speech. First, there are instances where insignificant events propagate to extremely high levels in the scale-space structure, solely because they are slightly more significant than the surrounding environment. This situation is illustrated in Figure 3.3 in the silence portion at the beginning of the utterance, where an acoustic event propagates to a very high level even though the acoustic environment is very homogeneous. More importantly however, landmarks which are quite distinct but which are near more important landmarks will disappear at a very low level in the scale-space description. This problem is also illustrated in Figure 3.3 in the second syllable of the word 'coconut,' which has a sequence of three closely spaced events corresponding to the release of the stop, the onset of voicing, and the point of oral closure. Although these events are all acoustically significant, two are eliminated at very low smoothing levels in the scale-space structure. With a minimal amount of

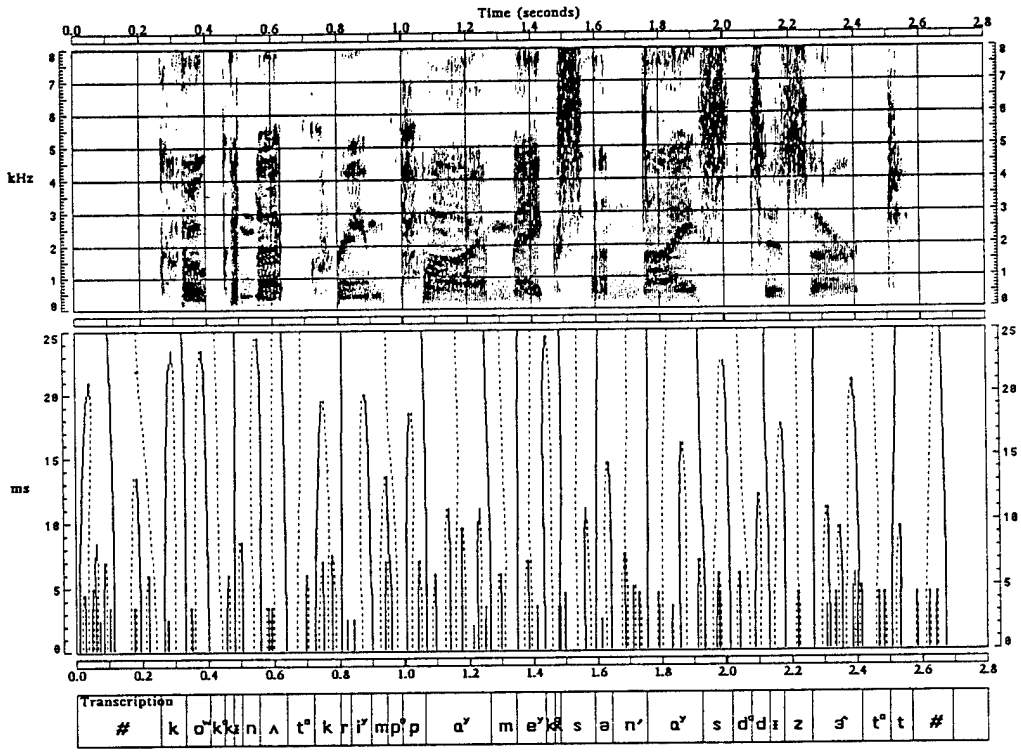


Figure 3.3: Multi-level description using scale-space filtering.

This figure illustrates computation of scale space filtering on a change function ($\sigma = 5$ ms) for the utterance 'Coconut cream pie makes a nice dessert,' spoken by a female talker. The top display is a spectrogram of the utterance. The middle display contains the scale space structure. The vertical dimension corresponds to the σ of the Gaussian derivative. Solid lines correspond to local maxima, dashed lines correspond to local minima. An aligned phonetic transcription is shown at the bottom of the figure.

smoothing the three narrow peaks are quickly turned into a single wide peak. This phenomenon would hold for any sequence of closely spaced events, and is an inherent part of the scale-space procedure since it is intimately tied to increasing amounts of smoothing in the time domain. For speech analysis, this is very unsatisfactory, since short events are often acoustically quite significant. This also means that the scale-space structure will vary as a function of the duration of a segment, which means that it is not very stable.

3.2.2 Hierarchical Structuring

The fact that there are short events that are often quite distinct from their local environment suggests that a hierarchical clustering procedure, which incorporates some kind of temporal constraint, might be appropriate for ranking the significance of an acoustic event [8,24]. Such a procedure was in fact developed for this thesis, and appears to avoid some of the problems associated with scale-space filtering.

The clustering algorithm which produces a multi-level description of the speech signal is similar to the concept used for locating acoustic events with the associations framework. First, the algorithm uses all events found by some procedure to define 'seed regions.' Next, each region is associated with either its left or right neighbor using a similarity measure. Similarity is computed with a distance measure applied to the average spectral vectors of each region. When two adjacent regions associate with each other, they are merged together to form a single region. This new region subsequently associates itself with one of its neighbors. The merging process continues until the entire utterance is described by a single acoustic event. The process is described more formally in Table 3.1.

This procedure produces a tree structure whose terminal nodes correspond to the N seed regions determined by the original acoustic landmarks. Nodes in the tree successively collapse until there is but one single region spanning the entire utterance. Each node in the tree splits into two subgroups. By keeping track of the distance

Table 3.1: Hierarchical structuring of acoustic landmarks.

1. Find boundaries $\{b_i, 0 \leq i \leq N\}, \quad t_i < t_j, \quad \forall i < j$.
2. Create initial region set $R_0 = \{r_0(i), 0 \leq i < N\}$.
   $r_0(i) \equiv r(i, i+1)$.
3. Create initial distance set $D_0 = \{d_0(i), 0 \leq i < N\}$.
   $d_0(i) \equiv d(r_0(i), r_0(i+1))$.
4. Until $R_N = \{r_N(0)\} \equiv r(0, N)$
   For any k such that $d_j(k-1) > d_j(k) < d_j(k+1)$
   (a) $r_{j+1}(i) = r_j(i), \quad 0 \leq i < k$
   (b) $r_{j+1}(k) = merge(r_j(k), r_j(k+1))$
   (c) $r_{j+1}(i) = r_j(i+1), \quad k < i < N-j-1$
   (d) $R_{j+1} = \{r_{j+1}(i), \quad 0 \leq i < N-j-1\}$
   (e) $d_{j+1}(i) = d_j(i), \quad 0 \leq i < k-1$
   (f) $d_{j+1}(k-1) = max(d_j(k-1), d(r_j(k-1), r_{j+1}(k)))$
   (g) $d_{j+1}(k) = max(d_j(k+1), d(r_{j+1}(k), r_j(k+1)))$
   (h) $d_{j+1}(i) = d_j(i+1), \quad k < i < N-j-1$
   (i) $D_{j+1} = \{d_{j+1}(i), \quad 0 \leq i < N-j-1\}$

Definitions:

- b_i is a boundary occurring at time t_i.
- $r(i,j)$ is a region spanning times t_i to t_j.
- $r_j(i)$ is the ith region of the jth iteration.
- $d(i,j)$ is the distance between regions i and j.
- $d_j(i)$ is the ith distance of the jth iteration.
- $merge(r(i,j), r(j,k))$ combines two adjacent regions to produce a region $r(i,k)$ spanning times t_i to t_k.
- The distances $d_j(-1)$ and $d_j(N-j)$ are infinite.

at which two regions merge into one, the multi-level description can be displayed in a tree-like fashion as a *dendrogram*, as illustrated in Figure 3.4 for the utterance 'Coconut cream pie makes a nice dessert.'[1] From the bottom towards the top of the dendrogram the acoustic description varies from fine to coarse. The release of the initial /k/, for example, may be considered to be a single acoustic event or a combination of two events (release plus aspiration) depending on the level of detail desired. Note that there is no single level in the dendrogram which contains all acoustic events of interest. For some landmarks, such as the /sd/ sequence in the word pair 'nice dessert,' a more sensitive description is necessary to locate an acoustic landmark. In other situations, such as for the diphthong /a^{y}/, in the word 'nice,' a more robust description is necessary in order to avoid internal landmarks.

The resulting structure is a mechanism for describing all acoustic events which occur in the signal in a single coherent framework. The procedure for creating this structure is attractive because it uses only relative measures of acoustic similarity during construction, and as such is likely to be largely independent of the speaker, vocabulary, and background noise, and as a consequence is quite robust. Further, there are no thresholds, or other heuristics involved in constructing the structure.

An examination of the procedure shows that for merging order to be unimportant, the distance function for a given region and its environment must increase monotonically. This behavior is achieved by defining the distance between regions $r(i,j)$ and $r(j,k)$ as the maximum of the distance between these regions or their subregions having boundary b_j. An example of the structure which would result if this condition were not maintained is shown in Figure 3.5. From this figure it is clear that the dendrogram structure is not monotonically increasing. A typical example of why this is so may be found in the word 'coconut.' At a level of 22.5, the acoustic segment corresponding to initial silence and /k/ release merges with the acoustic segment corresponding to the vowel /o^{w}/. The resulting average spectra of the new region

[1]The term dendrogram refers to the diagrammatic representation of any hierarchical grouping of objects [24,55].

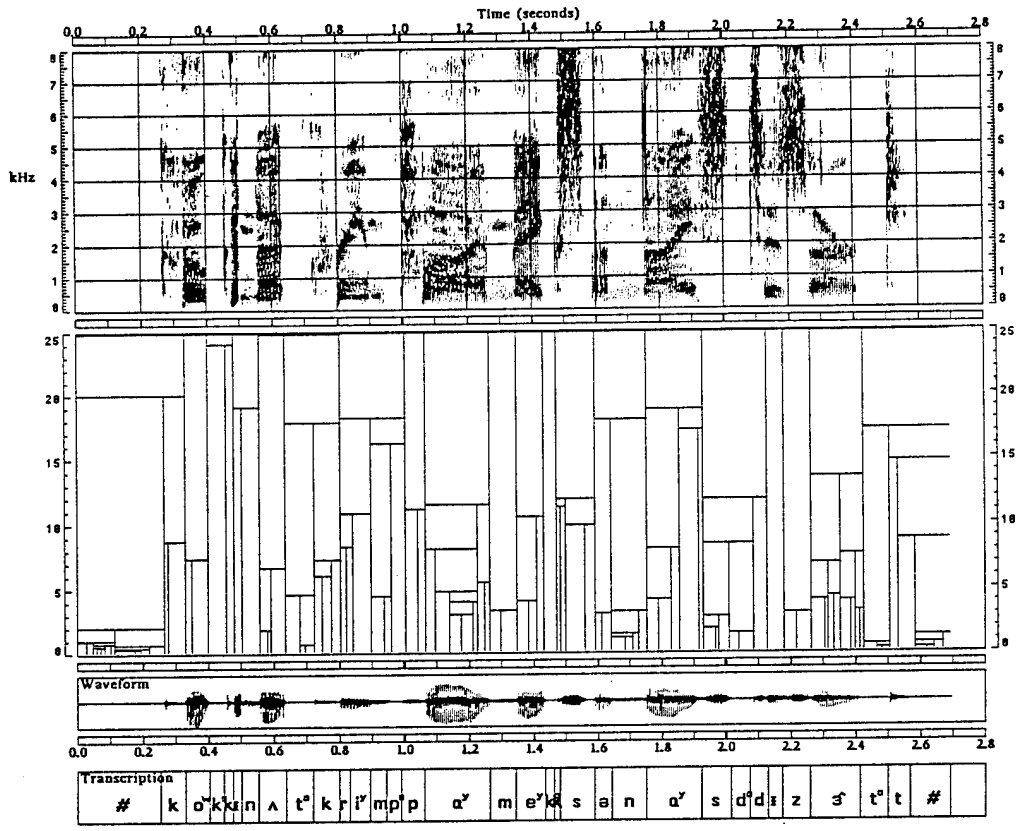


Figure 3.4: Multi-level description using a hierarchical structure.

This figure illustrates computation of the dendrogram structure for the utterance 'Coconut cream pie makes a nice dessert,' spoken by a female talker. The top display corresponds to a spectrogram of the utterance. The middle display contains the dendrogram structure. The top few nodes of the dendrogram structure are not shown. The speech waveform and an aligned phonetic transcription are shown at the bottom of the figure.

RESEARCH LABORATORY *of* ELECTRONICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139

Publications Office
Room 36-464

April 5, 1989

Dr. Tice F. DeYoung
SPAWAR-SYSCOM
Code 3241
Washington, D.C. 20363-5100

In accordance with the terms of the Navy Contract N00039-85-C-0254 we are sending you the following material:

Type of Material :	Technical Report No. 536
Title:	Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition
Written by:	J.R. Glass
Published:	December 1988
Number of Copies:	One

Elisabeth O. Colford
RLE Contract Reports

OSP 95826

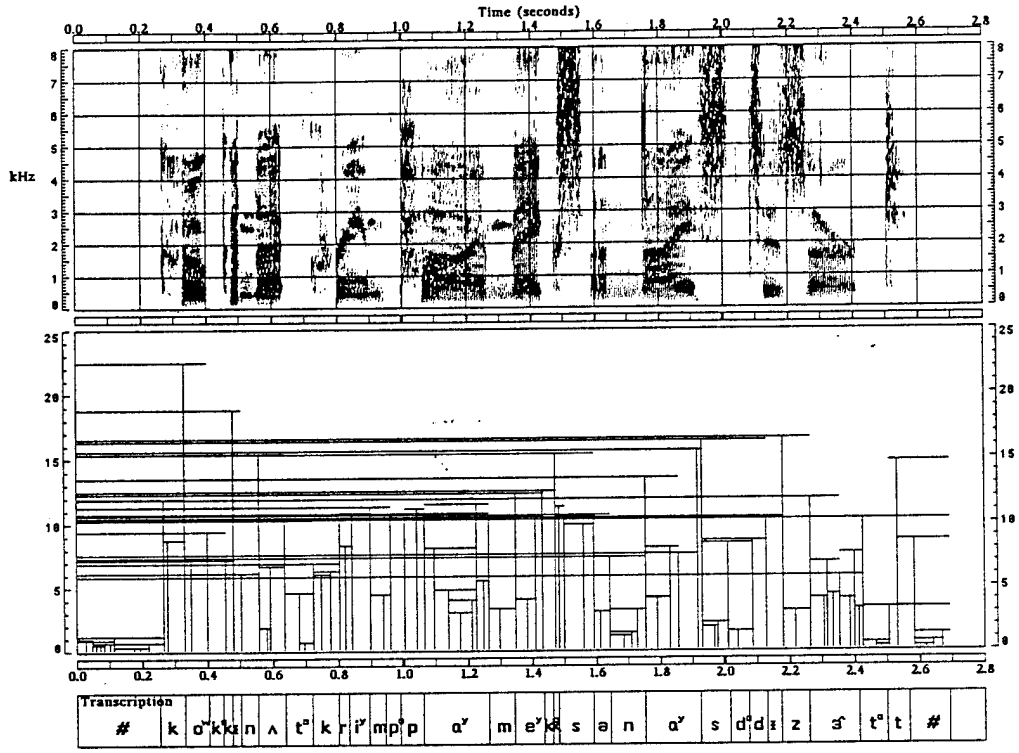


Figure 3.5: A non-monotonically increasing dendrogram.

This figure illustrates computation of the dendrogram structure which results when distances are not propagated correctly. The top display corresponds to a spectrogram of the utterance. The middle display contains the dendrogram structure. The top few nodes of the dendrogram structure are not shown. An aligned phonetic transcription is shown at the bottom of the figure.

subsequently looks much more similar to the spectra of the following /k/ closure than did the spectra of the vowel itself. As a result, the new region merges with the closure region at a level of 9.5 in the dendrogram.

In addition to producing a structure which is difficult to interpret in some instances, the lack of a monotonically increasing structure means that the order of construction is important. A structure which is built in a left-to-right manner cannot be guaranteed to be identical to one built right-to-left, or in any other fashion. By ensuring that distances always increase, both problems are successfully avoided.

The particular distance function applied to the acoustic regions is not crucial to the success of the dendrogram, as long as a reasonable metric is used [24]. Currently, every region $r(i,j)$ is represented by an average of the spectral frames from b_i to b_j. The distance between two adjacent regions is a distance between their two average spectra. If the distance used is an average of all the distances between two regions, the result is virtually the same, although significantly more computation is involved. Both of these metrics are more stable than one using a maximum or a minimum distance.

The criteria used to determine an acceptable distance function between two acoustic vectors were

$$d(\vec{x},\vec{x}) = 0$$

and

$$d(\vec{x},\vec{y}) = d(\vec{y},\vec{x})$$

The first distance between acoustic vectors which was implemented was a Euclidean distance, $\|\vec{x}-\vec{y}\|$. This metric is reasonable since the variance in each dimension is approximately the same. However, the Euclidean metric over-emphasizes the total gain in the region, minimizing the importance of spectral shape. This effect is illustrated in Figure 3.6, which shows that the subsegments of the vowels /ɑʸ/ and /ɝ/ do not successfully cluster together. By including a measure of the correlation of the spectra, the dendrogram structure was improved in cases where the gains were significantly different. This was done by dividing the Euclidean distance by the normalized dot product, or cosine of the angle between $\vec{x}$ and $\vec{y}$, $\frac{\vec{x}\cdot\vec{y}}{\|\vec{x}\|\|\vec{y}\|}$. The distance between regions was essentially Euclidean when the spectral shapes were similar, but was increased significantly when the spectral shapes were dissimilar. Distance metrics are discussed in more detail in the final chapter.

Because each region is represented by only one spectral average, this structuring process is much faster than the scale-space representation which was first investigated. Also, because the structure depends on the relative acoustic difference of adjacent

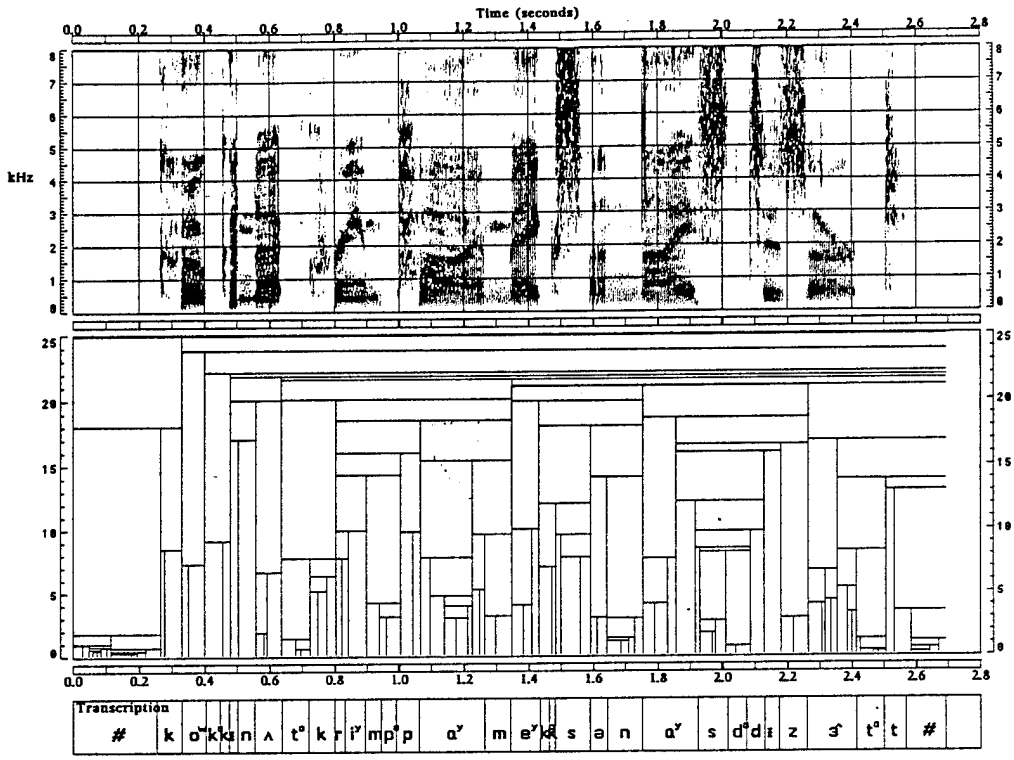


Figure 3.6: A dendrogram computed with a Euclidean distance metric.

This figure illustrates computation of the dendrogram structure when a Euclidean distance metric is used. The top display corresponds to a spectrogram of the utterance. The middle display contains the dendrogram structure. The top few nodes of the dendrogram structure are not shown. An aligned phonetic transcription is shown at the bottom of the figure.

regions, short events such as stop bursts, flaps, or schwas, which are acoustically quite different from their local environment are maintained at a very high level in the dendrogram, unlike their behavior in the scale-space representation. This effect can be clearly seen by comparing Figures 3.3 and 3.4. Additional examples of dendrograms may be found in Appendix A.

3.3 Performance Evaluation

3.3.1 Evaluation Criterion

An important question to ask about the dendrogram, or about any acoustic description of speech, is whether or not it is able to capture relevant acoustic-phonetic landmarks which will be useful for decoding the speech signal. In the case of a multi-level description such as the dendrogram, it is not sufficient to merely mark these locations; it is also desirable to organize the events in a meaningful fashion. Thus, for instance, it might be desirable that important acoustic-phonetic landmarks propagate to higher levels in the structure than less significant landmarks. A criterion such as this would explain why dendrograms in Figures 3.5 and 3.6 were considered to be inferior to the one displayed in Figure 3.4.

Since the dendrogram structure provides an acoustic description of the speech signal, it is difficult to define precisely what is, and what is not, an important acoustic landmark. One could argue, for instance, that any structure is a reasonable description of the acoustic signal, and that no multi-level description is any better or worse than another. In this work, a phonetic transcription was used to determine the identity of important acoustic-phonetic landmarks. The phonetic transcription essentially marks major acoustic landmarks in the speech signal, such as the release of stop consonants, the onset of voicing, or points of oral closure [104]. Although there are many cases where it is difficult to precisely determine the location of a boundary between adjacent phones, the boundaries in the phonetic transcription usually correspond to important landmarks which would be useful for subsequent acoustic-phonetic analyses of the speech signal. As such, they define a reasonable criterion for judging the structure of the dendrogram. Clearly however, differences between the dendrogram and the phonetic transcription are not necessarily 'errors,' especially if the differences occur in a systematic manner. These differences might point out inconsistencies in the nature of the phonetic transcription for instance.

If the dendrogram structure were able to organize landmarks in the speech signal in a meaningful fashion according to the phonetic transcription, then landmarks which matched boundaries in the transcription should propagate to higher dendrogram levels than landmarks which did not match any boundary. At some level in the dendrogram, it should therefore be possible to find a region, $r(i,j)$, whose boundaries, b_i, b_j, match those of some phone in the phonetic transcription. In other words, it should be possible to find a path of contiguous acoustic segments in the dendrogram corresponding to the phones in the phonetic transcription.

This concept was implemented by finding the sequence of acoustic segments whose boundaries minimized the error with phone boundaries. The criterion for determining errors is illustrated in Figure 3.7. In cases where the dendrogram deleted an important acoustic landmark, or organized it in such a fashion that it was not possible to obtain a match with the phonetic transcription without inserting many extra segments, a phonetic boundary could be deleted by the alignment procedure. Alternatively, if the dendrogram had too many landmarks, such that it was not possible to avoid extra landmarks without deleting many phonetic boundaries, insertions were permitted. An example of an aligned path is shown in Figure 3.8. Note that in general there is a one-to-one match between acoustic segments and phones, and that boundaries are very close to each other.

Once an algorithm for aligning a phonetic transcription with the dendrogram had been developed, it was possible to gather statistics on insertions and deletions which provided a reasonable indication of how well the dendrogram was organizing acoustic-phonetic information in the speech signal. The following section describes the results of the evaluation in more detail.

3.3.2 Evaluation Results

The evaluation was based on a set of 500 utterances from 100 different talkers of the TIMIT database [69]. These sentences contain nearly 19,000 phones. The 10 ms

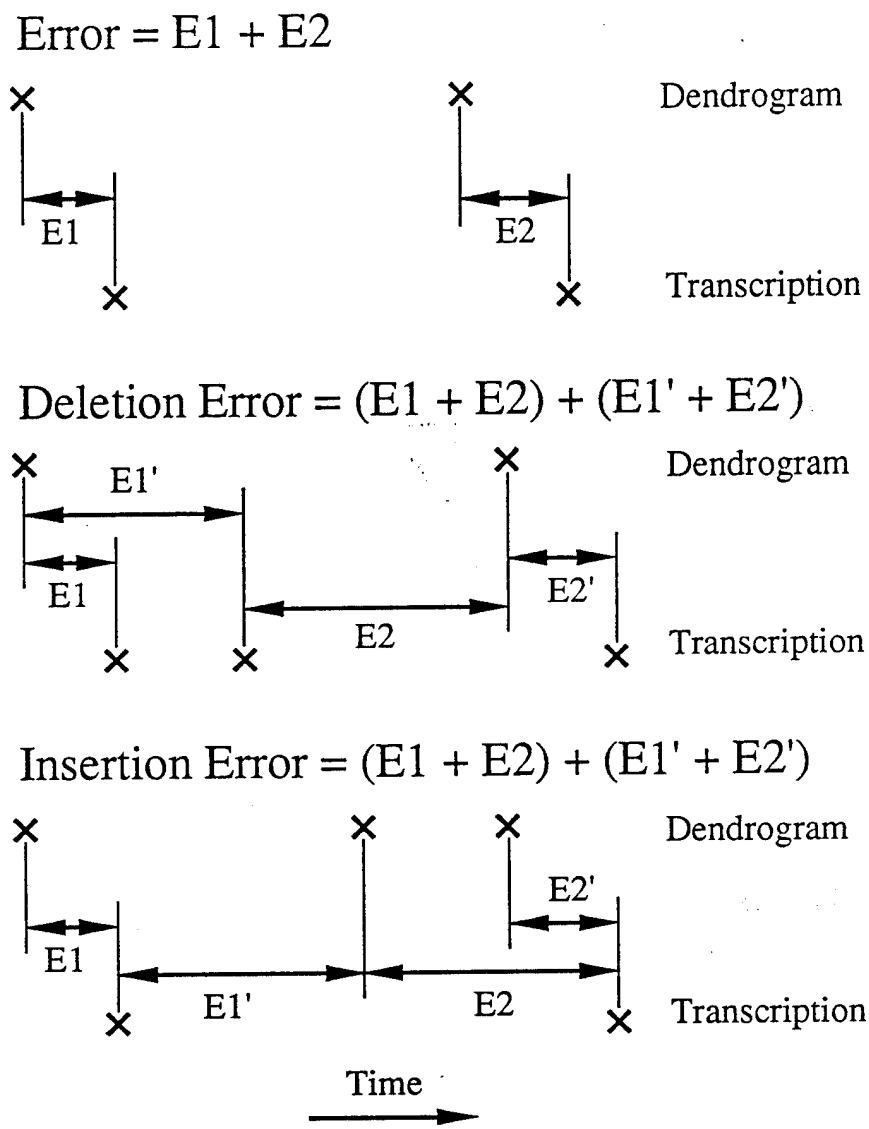


Figure 3.7: Dendrogram alignment procedure.

This figure illustrates the criterion used to determine the differences of a match between boundaries in the dendrogram and boundaries in the phonetic transcription. The top display shows that when there is a one-to-one match between a pair of boundaries, the error is the sum of the time differences between individual boundaries. The middle display shows that when an extra boundary is present in the phonetic transcription, the error is the total error of mapping each pair of boundaries in the phonetic transcription to the pair of dendrogram boundaries. This corresponds to a deletion of a phonetic boundary. The bottom display shows that when an extra boundary is present in the dendrogram, the error is the total error of mapping each pair of boundaries in the dendrogram to the pair of boundaries in the phonetic transcription. This corresponds to an insertion of extra acoustic boundaries.

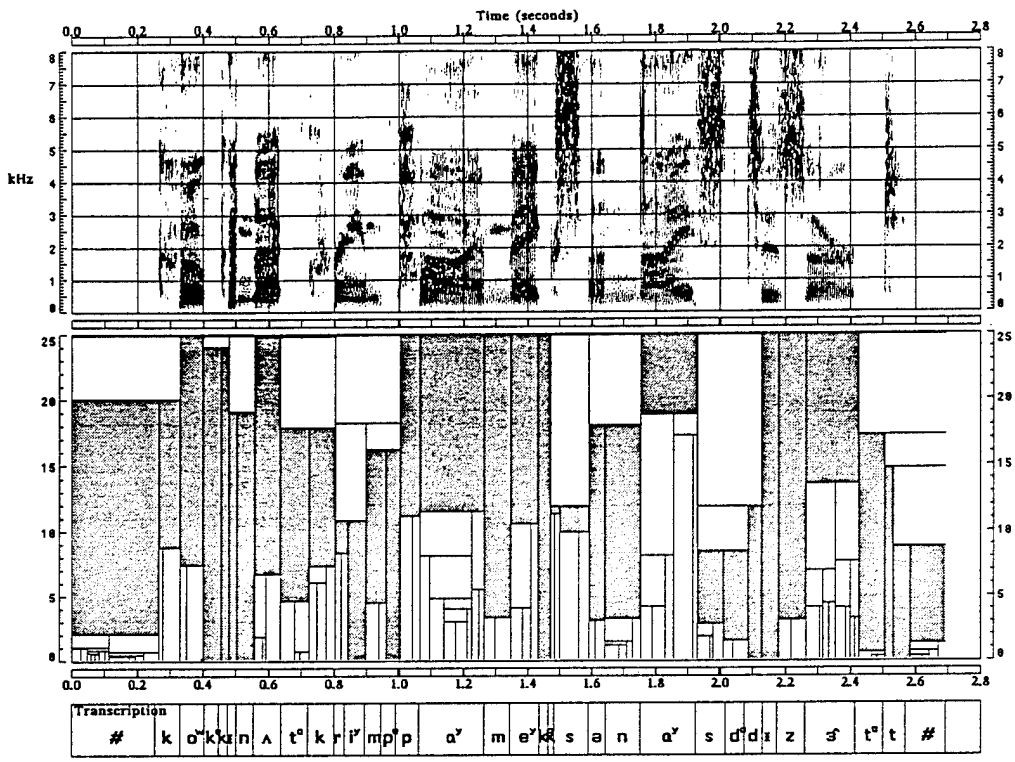


Figure 3.8: An aligned dendrogram.

This figure illustrates the results of automatically aligning the phonetic transcription with the dendrogram. Aligned regions have been shaded in the dendrogram.

association algorithm described earlier was used to locate acoustic landmarks. The signal representation, $\vec{S}_n$, was the set of outputs of the mean-rate response, sampled every 5 ms. Dendrograms were constructed from the acoustic landmarks, and were aligned with the phonetic transcription. An evaluation of the alignments indicated that 3.5% of the phone boundaries were deleted while 5.0% of the phones required an inserted acoustic segment. An analysis of the duration differences of boundaries in the phonetic transcription with those found automatically indicated that 70% of the boundaries were within 10 ms of each other, and that over 90% were within 20 ms.

An examination of the context of the deletion errors indicated that they were highly concentrated in a small number of contexts. Of the over 1750 actual phonetic

Table 3.2: Top ten sources of deletions.

context	% deleted	total (#)
/b/ release	38	310
/d/ release	20	330
nasal + closure	17	406
(nasal or closure) + /ð/	14	129
/g/ release	10	174
/p/ release	10	347
/r/ vowel	9	432
/l/ vowel	6	422
/t/ release	4	500
/k/ release	3	520

contexts which occurred, which was less than half of the 61^2 theoretically possible combinations, deletions occurred in 221, or under 13%, of these contexts. On average, there were 2.9 occurrences of each deletion. The top ten sources of deletion error have been summarized in Table 3.2. For example, the first row of this table shows that 38% of the 310 releases associated with the phoneme /b/ were not found in the aligned dendrogram path. These ten sources of error account for over 75% of all deletions. Specifically, over half the errors could be associated with the release of the six stop consonants. Fully one third could be accounted for by /b/ and /d/ alone. Over 15% of the errors could be accounted for by /ð/ when preceded by a nasal or by silence and by nasal closure deletions. Nearly 10% could be accounted for by deleted boundaries between /r/ and /l/ and a following vowel. Some of the remaining significant statistics were for the /ar/, /al/, /yü/, and /ɔr/ combinations, which accounted for an additional 5% of the deletions.

The data on the insertions are slightly more difficult to analyze. One way to view an insertion is as splitting a phonetic segment into more than one acoustic segment. A diphthong such as /a^y/ might be represented by two acoustic segments, for instance. Often however, an insertion corresponds to an acoustic segment which falls somewhere between two adjacent phonemes, such as a period of silence between

a fricative and a following vowel. In this case, it would be of interest to know the identities of the two phones on either side of an inserted acoustic segment. In fact, insertions were analyzed in this manner. If a phone was split into two or more acoustic regions, the acoustic segment which overlapped the most with the phone boundaries was assigned the phone label, while the extra segments were labeled insertions. An inserted segment could then be assigned a context based on the label of the adjacent segments. A sequence of two insertions was never observed.

An analysis of the insertions showed that they were more diffuse than deletion errors. Insertions occurred in 481, or just over 27%, of the total number of actual contexts observed. This is more than twice the number of contexts which were observed to have deletions. On average, there were 1.8 occurrences of each inserted segment. In general, there appeared to be two major sources of insertions. First, there was often a distinct acoustic segment between two phones which was not noted in the phonetic transcription. A good example of this may be found in Figure 3.8 between the /s/ in 'makes,' and the schwa in 'a.' From an acoustic perspective, there is a clear acoustic segment which is different from the fricative and the vowel. The phonetic transcription makes no note of this fact however, and marks only the onset of voicing. If, for some reason, the dendrogram had not organized the silence with the fricative, an insertion would likely have been required. This type of phenomenon occurred frequently when there was a change in the source of excitation.

Another source of insertion occurred when a phonetic segment was subject to a significant amount of change over the course of the segment. A good example of this is the diphthong /aʸ/ in Figure 3.8 in the word 'nice.' Apart from a short region between the vowel and the following /s/, there are two robust acoustic segments which describe this vowel, one which captures the low, back portion of the vowel, and the other which describes the high, front portion of the vowel. Although in this example these regions merge to form a single acoustic segment, this is not always the case. A model of a diphthong might therefore need to capture the temporal acoustic structure in order to account for these changes. In addition to diphthongs, these types

Table 3.3: Top five sources of insertions.

left context	right context	% error	total #
[eʸɑʸɛæ]	[s]	30	80
[ɚɝ]	[z]	27	56
[w l]	[eʸiʸ]	19	161
[f]	[ɔɑr ɚ]	18	114
[ɪɪ]	[ŋ]	14	81

of errors also occurred in vowel-semivowel sequences where there was an extra segment which modeled the dynamic portion of the transition between the two phones. For example, the most common insertion error was in /liʸ/ sequences, which contained 18 insertions out of a total of 91 occurrences (20%). The distribution of the five most common contexts of insertion error is shown in Table 3.3. The contexts were loosely grouped according to similarity, and were chosen based on the number of errors which occurred. Although the data in this table tend to confirm the previous discussion, there were many cases where the dendrogram organized the acoustic landmarks in a manner which made it difficult to find a good match with the phonetic transcription.

In addition to the analysis of differences between dendrogram boundaries and those of the phonetic transcription, a comparison was made between the dendrogram heights of landmarks which were aligned to phonetic boundaries and the heights of those landmarks which were not aligned to phonetic boundaries. A distribution comparing these two is shown in Figure 3.9. By conditioning this distribution on the local acoustic context, further separation between these two groups is possible. Note that this type of information lends itself naturally to a probabilistic framework for finding the best path through the dendrogram. This point will be discussed in more detail in the final chapter.

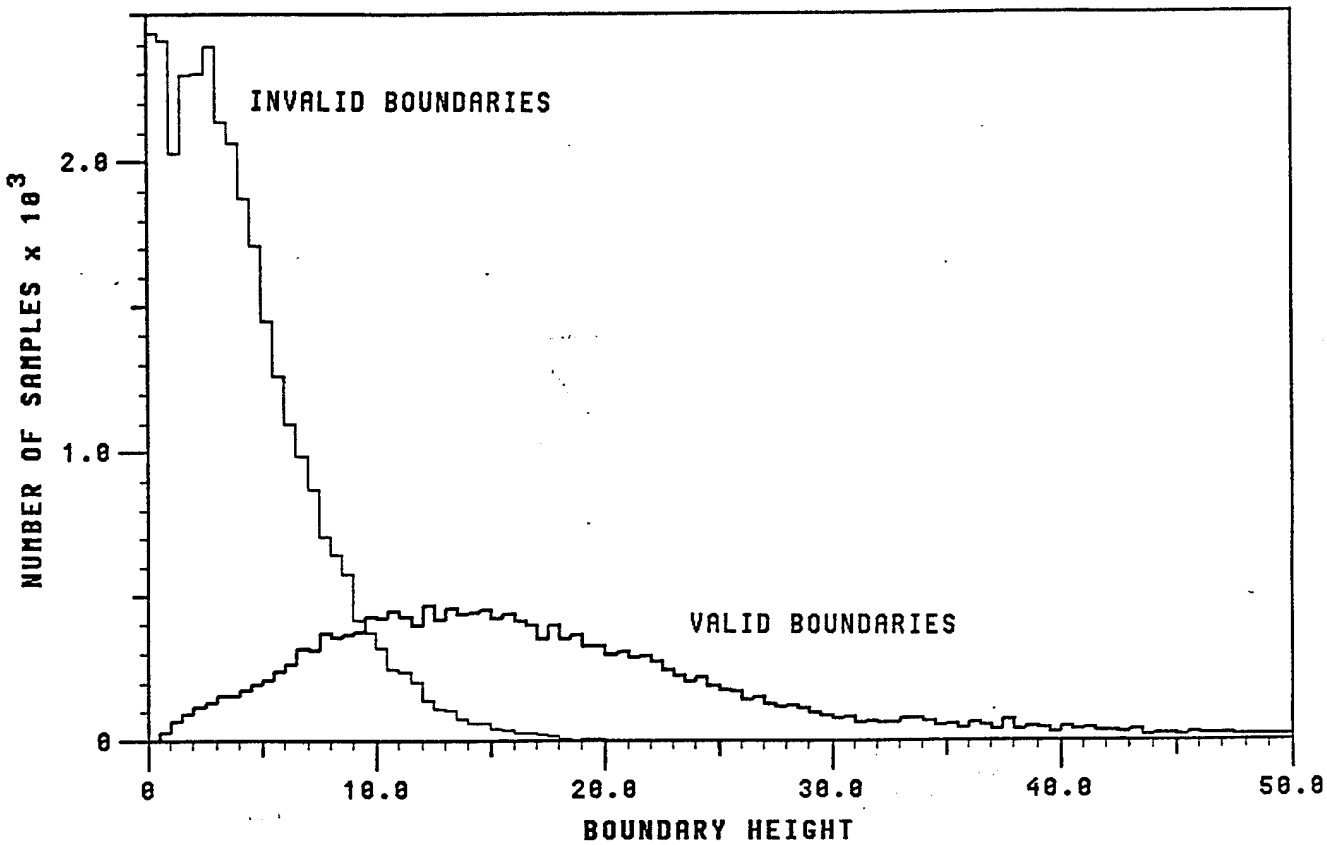


Figure 3.9: Distribution of dendrogram boundary heights.

This figure compares the distribution of the heights of those dendrogram boundaries which were aligned with a phonetic boundary to those which were not aligned with a phonetic boundary. Height is defined by the vertical dimension on the dendrogram display.

3.4 Chapter Summary

This chapter investigated procedures to automatically locate acoustic landmarks in the speech signal. A change function, $c_n(\sigma)$, was motivated as a means for locating local maxima in the rate of change of the signal representation of the speech signal. This function was shown to be related to an associations procedure, whereby similarities are computed between a given point in time and the local environments immediately preceding and following. Acoustic landmarks are located whenever the association switches from past to future.

An analysis of these algorithms indicated that it was difficult to determine a single level of sensitivity capable of capturing all acoustic events of interest with a single level of description. Thus, a procedure was developed which provided a multi-level description of the speech signal by performing a time-ordered hierarchical clustering. An analysis of the multi-level structure on a set of 500 utterances indicated that over 96% of the acoustic-phonetic events of interest were located, with an insertion rate of less than 5%. The fact that both the deletion and insertion rate were quite small indicated that acoustic landmarks were being located and organized by the dendrogram in a meaningful way.

Chapter 4

Acoustic Classification

In Chapter 1, a method that directly maps an acoustic description of the speech signal to underlying phonological units used to represent lexical entries was proposed. The viability of such an approach depends critically on the ability to detect important acoustic landmarks in the speech signal. It further depends on the ability to describe these events in terms of an inventory of labels that captures the regularity of phonetic variations. In the previous chapter, a criterion for defining an acoustic landmark was proposed, and a procedure was described which automatically located these events. This chapter discusses how the resulting segments can be classified into a set of acoustic labels, and develops a procedure to automatically determine these labels from a large body of training data.

There are at least three important criteria which need to be fulfilled by any set of acoustic labels. First, a set of labels must be able to capture important phonetic regularities. Sounds with similar phonetic features should be described by the same set of labels. Thus, it is reasonable for phones such as [ɪ] or [i^y] to be described by the same acoustic label, but much less acceptable for [s] and [ɛ] to have the same acoustic label. The second important characteristic of a set of acoustic labels is that they must be able to distinguish between acoustic differences due to regular variations in the realization of a phoneme. As was discussed in Chapter 1, capturing this systematic behavior should provide additional sources of constraint in the decoding process, and is therefore a desirable property of a set of acoustic units. Finally, it is clear that a

set of acoustic labels should be robust across speakers and datasets.

The ultimate goal of any approach advocating a set of acoustic units is to fully determine a set of labels that captures all three of these conditions. However, the ability to achieve this goal depends critically on the correct representation of the signal for *all* fine phonetic distinctions, and the availability of a tremendous amount of training data, in order to achieve robust results. These two requirements make this goal well beyond the scope of this thesis, and a subject for continuing research. Although these factors limit the scope of the investigation, they do not preclude developing a methodology with which it is possible to search for acoustic regularities in the speech signal. In addition, it is still possible to perform some preliminary investigations to provide an indication of the viability of this approach.

This chapter attempts to address two of the issues related to acoustic classification. First, an investigation is made which attempts to determine if it is possible to uncover major sources of acoustic regularity that capture basic properties of all phonemes. This is accomplished by applying a hierarchical clustering procedure to data from all phonemes, using the mean-rate response as the basis of the signal representation.

Second, an investigation is made which attempts to demonstrate that it is possible to capture important context dependencies of individual phonemes. This is done by applying a hierarchical clustering procedure to a small number of phonemes. The goal of this study is to determine if there appear to be a finite number of regular acoustic forms which have consistent contextual dependencies. Additionally, an attempt is made to illustrate that these regularities can generalize across sets of phonemes with similar features.

4.1 Hierarchical Clustering

The procedure used to organize all acoustic data was a stepwise-optimal agglomerative hierarchical clustering procedure [24]. This procedure was selected because it

produces an explicit hierarchical structure of acoustic clusters. At the highest level in the structure, all data are represented in a single acoustic class. At the lowest level in the structure, all data are represented by the original seed clusters. The objective is to find a level of description which can explain much of the acoustic variance in the data, while having a meaningful phonological interpretation. The advantage of this approach is that it allows for a large amount of data to be observed without requiring any preconceived notion about the number of acoustic classes to expect.

In the interest of reducing the amount of computation involved in the hierarchical clustering, a pre-clustering data reduction was achieved by merging similar data vectors with an iterative, nearest-neighbor procedure described in Table 4.1. The objective of this procedure was to reduce the number of clusters while maintaining adequate coverage of the data. In this procedure a vector is merged into an existing cluster if the distance between it and the cluster mean falls below a threshold, D_{th}. Otherwise, a new cluster is formed with this vector, and the procedure repeats. In the end, all clusters with membership of two or less are discarded, and the data are resorted into the remaining clusters. This last step ensures that each vector is truly sorted to its nearest cluster. This step was also taken during the analysis of any set of clusters of the hierarchical structure. The value of D_{th} was determined experimentally from a subset of the training data, and was set to maximize the number of clusters with more than two members. All distances were Euclidean distances between a data point and the cluster mean.

The hierarchical clustering procedure used a stepwise-optimal approach which on each iteration merged the two clusters which minimized the increase of the total distortion, D_T [24]. Total distortion was defined as the total square error between a cluster mean, $\vec{m}$, and its constituents,

$$D_T = \sum_{\forall c_i} \sum_{x \in c_i} \|\vec{x} - \vec{m}_i\|^2$$

where c_i is the ith cluster. Merging clusters c_i and c_j produces an increase in the total

Table 4.1: Pre-clustering procedure for data reduction.

1. Select one random data point to form the first cluster.
2. Until all data have been selected,
   (a) Select a data point at random from the remaining data,
   (b) Compute distance D_{min} to closest existing cluster,
   (c) If $D_{min} < D_{th}$
   Merge data point with closest existing cluster,
   Otherwise form a new cluster with data point.
3. Prune all clusters with one or two members.
4. Resort all data to their closest cluster.

distortion ΔD_T of

$$\sum_{x \in c_i} \|\vec{x} - \vec{m}_{ij}\|^2 - \|\vec{x} - \vec{m}_i\|^2 + \sum_{x \in c_j} \|\vec{x} - \vec{m}_{ij}\|^2 - \|\vec{x} - \vec{m}_j\|^2$$

where

$$\vec{m}_{ij} = \frac{1}{n_i + n_j}(n_i \vec{m}_i + n_j \vec{m}_j)$$

where n_i is the number of elements in the ith cluster. Merging the two clusters which minimizes ΔD_T at each step results in a distance metric between clusters which is a Euclidean distance weighted by the number of elements in each cluster,

$$d(c_i, c_j) = \sqrt{\frac{n_i n_j}{n_i + n_j}} \|\vec{m}_i - \vec{m}_j\|$$

This procedure was used because it tended to favor growth by adding small clusters to large clusters rather than merging medium-sized clusters. The net effect was to produce a more stable and well-behaved structure. Note that this procedure does not restrict the distance metric to be Euclidean. In fact a subsequent experiment, to be discussed later in this chapter, found that a zero-mean distance metric produced superior results.

4.2 Coarse Acoustic Classification

4.2.1 Signal Representation

The first attempt at finding acoustic regularities investigated the entire inventory of speech sounds. For simplicity, each dendrogram segment was represented by a spectral average of the mean-rate response over the entire segment. As stated previously, this particular choice determined the nature of the acoustic clusters which could be observed in the training data. Although the precise nature of any observed regularities could not be known ahead of time, it was believed that they would cluster along dimensions similar to manner of articulation.

In order to assist the phonological interpretation of the acoustic organization, only regions which aligned with the phonetic transcription were used as training data. In Figure 3.8 for example, only the shaded regions were used for that utterance. This approach was clearly a simplification, since it assumed that a phonetic unit mapped to a single acoustic segment. When combined with the fact that each acoustic segment was represented solely by an average spectral vector, it was possible that many dynamic events, such as diphthongs, were excessively blurred. A more sophisticated model which accounts for some of the temporal variation in the realization of a phone is discussed later in this chapter. By accepting only regions which matched the phonetic transcription, it was possible to assign to each region a phonetic label which could be used later to help interpret the clustering results. In cases where an acoustic boundary had been deleted, the phone which overlapped the most with the acoustic region was used. In cases where an acoustic region was inserted, the same criterion was used. Thus, in an insertion, a phone label was usually associated with two acoustic segments.

The combination of using an average spectral vector for each acoustic segment and a single acoustic segment for each phone, introduced another side effect in the clustering results. By clustering with segment averages instead of individual spec-

tral frames, it is likely that many spurious frames were eliminated, thus reducing the amount of noise in the subsequent phonetic analysis of the acoustic classes. In addition, the use of a single vector per phone eliminated phoneme duration as a possible factor in the clustering results. Thus, the data from phones which were inherently longer in duration would not overwhelm the data from shorter phones.

A set of 500 TIMIT sentences, recorded from 100 different speakers [69], was used to train the classifier. These data comprised over 24 minutes of speech and contained over 290,000 spectral frames. Restriction to the time-aligned dendrogram regions reduced the data to just under 19,000 regions. The pre-clustering procedure on these regions produced 560 clusters covering nearly 96% of the original data. All of the data were then resorted into these 560 seed clusters. The distribution of the number of tokens found in each cluster is illustrated in Figure 4.1. The mode of this distribution was 6, the median 16, and the average was 34. The hierarchical clustering was then performed on these seed clusters.

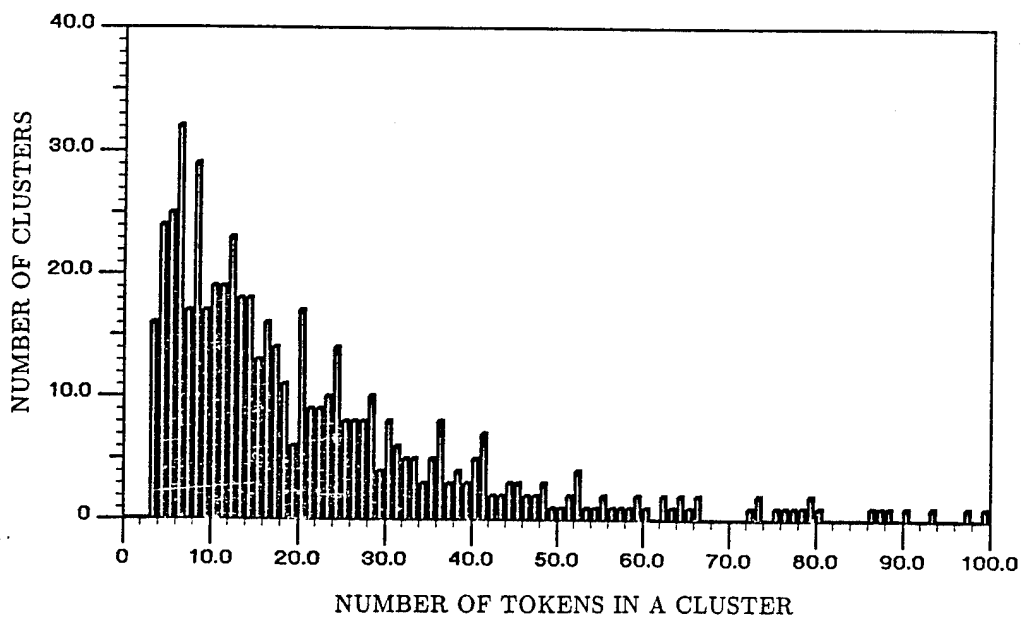


Figure 4.1: Seed clusters membership.

This figure illustrates the distribution of the number of tokens found in each of the 560 seed clusters determined with the pre-clustering algorithm.

4.2.2 Cluster Evaluation

The hierarchical clustering algorithm arranges the clusters in a tree-like structure in which each node bifurcates at a different level. The experimenter thus has the freedom to select the number of clusters and the associated spectral vectors for pattern classification. Several types of analysis were performed to help decide which clusters were most meaningful.

In order to determine the stability of the clustering procedure, the clustering experiment was repeated on several different databases. An examination of the phonetic contents of the clusters revealed that the top three or four levels of the tree structure are quite stable. For instance, the top few clusters essentially separate all consonants from vowels. The vowel clusters subsequently divide into groups corresponding to different extremes of the vowel space, while the obstruent clusters divide into subgroups such as silence, nasals, and fricatives. After the first few levels of the hierarchical structure however, the splits appear to become more data-dependent. From these observations it appeared that the number of clusters for reliable pattern classification should not exceed twenty.

As an attempt at a more quantitative analysis of the hierarchical structure, the amount of distortion involved in sorting the training set into a given set of clusters was measured. Distortion was again defined as the total square error between a cluster mean and its constituents. For a given number of clusters, the set with the minimum total distortion was designated as the best representation of the data. Figure 4.2 illustrates the rate of decrease in the distortion as the number of clusters increases from one to twenty. From this plot it may be seen that the most significant reductions in the distortion occur within approximately the first ten clusters. Afterwards the rate of decrease levels off to around 1%.

The relative merit of a set of clusters was also judged by examining the distribution of phonetic information within each set. This was done by observing the distribution of the phone labels associated with each spectral vector in the manner described

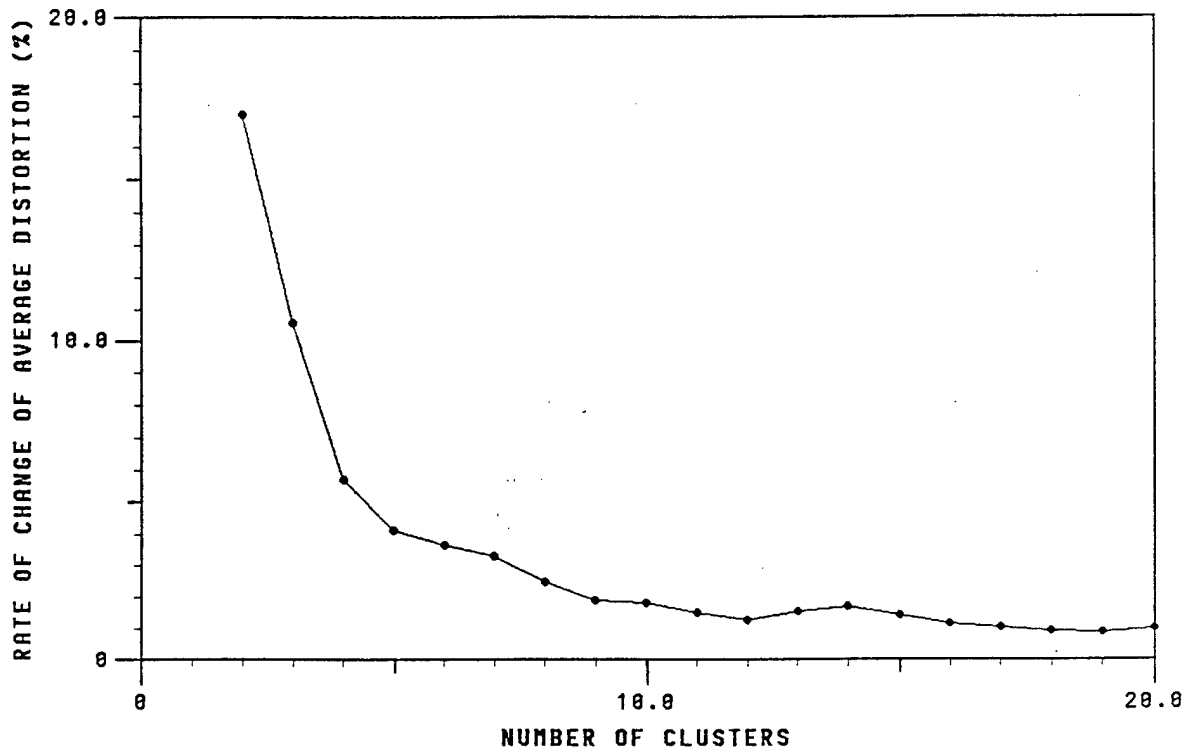


Figure 4.2: Rate of change of distortion versus number of clusters.

previously. The distributions were computed by resorting all of the original training data into a given set of clusters. An example of the distributions of the phone labels into a set of 10 clusters may be found in Table 4.2. The distributions in this table have been normalized for each phone, and were rounded to the nearest percent. In general, a given phone tends to fall into one or two acoustic classes. The few exceptions to this case will be discussed shortly. For example, of the 663 [iʸ]'s which were observed in the training data, 57% fell into cluster 1, while 18% fell into cluster 2. Of the 828 [s]'s which were observed, 69% fell into cluster 10, while 23% fell into cluster 9.

The average spectra associated with these 10 clusters are plotted in Figure 4.3. These average spectra show that the first five clusters capture information common to most vowels and semivowels. The average spectrum of cluster 1 shows resonances at low and high frequencies, which are typical of high front vowels, whereas the average spectrum of cluster 5 contains two low resonances, which are associated more with

Table 4.2: Distributions for ten clusters (percent normalized by phone).

Phone	1	2	3	4	5	6	7	8	9	10	Total #
iʸ	57	18	4	1	1	[illegible]	3	8	1	1	663
y	43	13	6	2	1	8	12	14	1	1	131
ʉ	37	17	14	3	4	9	2	15			176
eʸ	25	45	7	9	2	5	2	5			334
ɪ	20	39	16	17	1	2	1	4			500
ɨ	20	14	22	4	5	13	3	18			893
ɾ	1	4	41	28	14	7	1	6			559
ɚ	4	3	46	6	16	7	4	12			401
ɜ		3	60	23	6	3	1	2			230
ɛ	5	16	21	48	4	2	1	3			408
æ	6	10	28	46	3	1	1	5			334
ɑʸ	9	6	19	54	5	3	1	3			326
ʌ	1	1	23	58	14		1	2			284
ɑʷ	1	1	14	58	16	4	2	4			105
ɑ			13	64	18	1	2	2			371
ɔʸ	7	4	24	37	24	1	1	1			68
ʊ	2	20	20	33	21	3		2			61
ə	5	4	24	18	35	7	1	6			502
u	5	6	19	9	31	21	1	8			112
oʷ		2	13	26	49	6	2	2			250
ɔ			5	36	53	2	1	1			322
l̩	1		1	3	75	17	2	1			133
l	2	1	3	11	55	22	4	2			586
w	1		3	9	42	40	6				265
m̩					4	62	33				24
m	1				4	47	41	7			424
n̩	4		1		4	44	41	6			123
ŋ̩						50	50				4
ɹ̩	5		8			35	23	29			62
ŋ	2				2	39	51	5	1		132
n	1				1	35	58	4			735
v					1	17	72	7		4	246
[illegible]						9	84	7			94
[illegible]						6	92	2			97
dˤ						3	95	1			505
gˤ						3	96	2			185
tˤ						1	96	1		2	712
kˤ						2	97				555
[illegible]						2	97	1			1066
pˤ						2	98				372
bˤ						2	98				306
ʔ	2					14	48	26	2	7	42
ʔ	3	1	5	3	10	13	41	25			358
ð	4	4	2		1	12	40	32	1	6	250
θ						1	66	16	4	13	119
f						3	55	26	4	12	347
ɾ	6	1	3			16	49	24	1	1	190
b	4	2	7		4	7	31	41	2	2	123
p	2	3	4	1	3	5	26	51	4	1	308
h	3		4		1	6	26	52	4	5	140
ɦ	3		3	3	3	6	19	61			31
k	1	2	4		4	4	19	50	10	5	514
g	3	3	8	1	4	5	15	51	7	3	151
d	2	4	2	1		1	15	39	22	13	249
t	1	2	1			1	8	44	32	10	509
ʒ							5	10	48	38	21
ʧ		2			1	3	3	6	74	12	153
ʃ						2	7	6	71	13	193
ʤ							4	7	68	20	168
s						1	5	3	23	69	828
z							9	4	13	74	541
Total #	1094	917	1634	1837	1575	1605	5949	2006	975	1299	18891

back vowels or the semivowels, /l/ or /w/. The spectra of the remaining five clusters appear to be more associated with consonants. The peak distribution of energy in these clusters ranges from low frequency, common to the nasal consonants, to high frequency, common to the strident fricatives.

The fact that the distributions in each cluster contain a significant amount of phonetic regularity is illustrated in Table 4.3, where the cluster distributions have been normalized relative to the number of tokens in each cluster. For example, this table indicates that 35% of the tokens in cluster 1 are associated with [iʸ], while 16% may be associated with [ɪ]. Close examination of this table reveals that there is a large amount of phonetic regularity captured by these particular clusters. The first five acoustic clusters capture information about general vowel classes. For instance, the first cluster tends to contain high front sonorants such as [iʸ], [y], and [ü]. The fifth cluster tends to contain back sonorants such as [oʷ], and [ɔ]. This cluster also contains the majority of the lateral consonants as well.

The remaining five clusters in this example appear to cluster the remainder of the consonants. For instance, cluster six is dominated by sonorant consonants such as the nasal consonants. Cluster seven appears to be dominated by weak fricatives, and silence. Cluster eight contains phones which are associated with a predominance of high-frequency energy, such as stop-consonant releases and the aspirants. Finally, clusters nine and ten are dominated by the palatal and alveolar strident fricatives, respectively.

In order to help visualize the phonetic structure captured by a particular set of acoustic clusters, the phone distributions were themselves used as the basis of a hierarchical clustering. For example, the phone distributions for ten clusters are found in Table 4.2. The distance between two clusters of phones was defined as the average distance between each phone distribution in one cluster and each phone distribution in the other cluster. The distance metric used was a Euclidean distance between the distributions. The hierarchical structure produced by using the distributions in

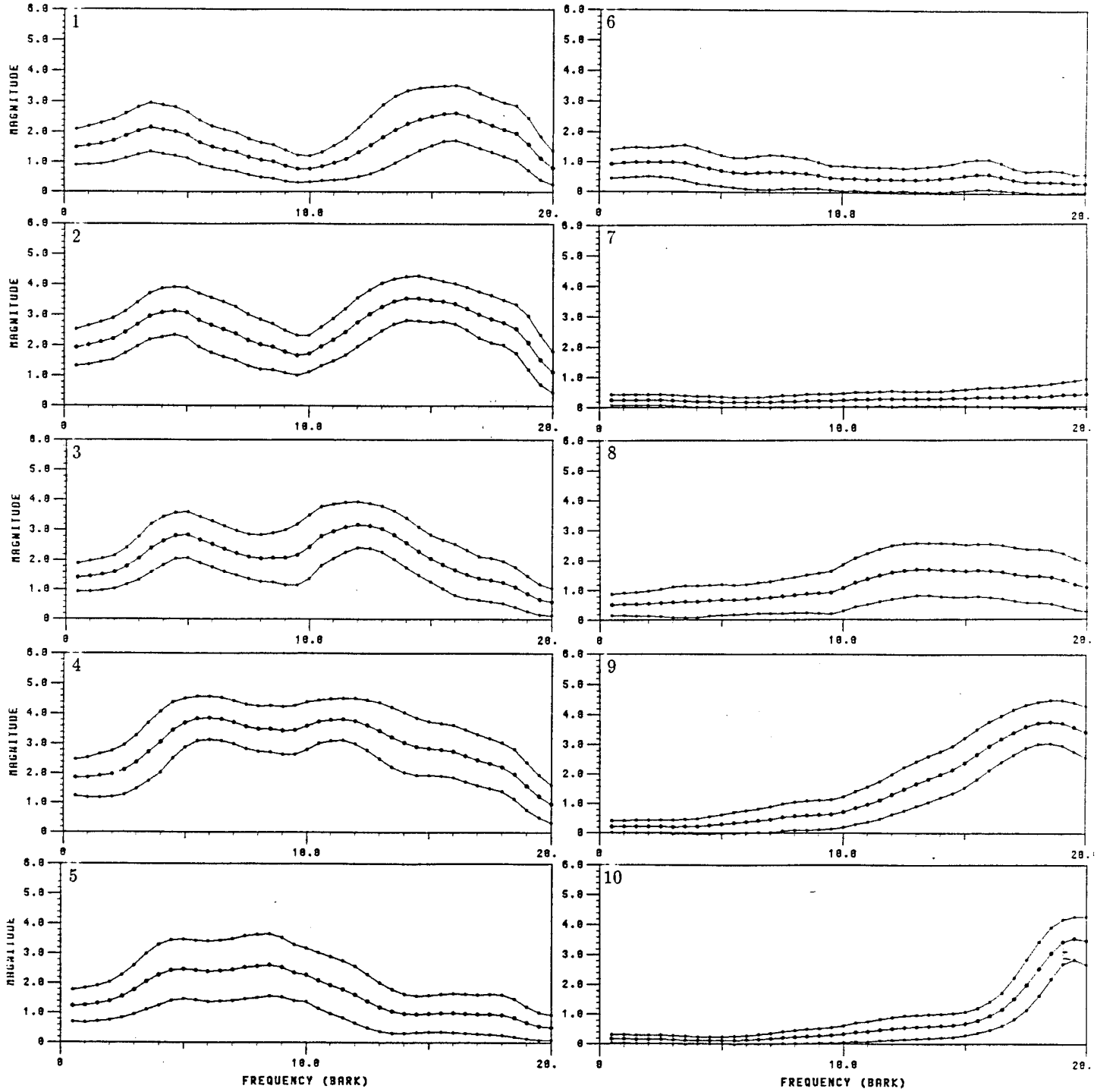


Figure 4.3: Spectra of ten clusters.

This figure illustrates the spectra of ten acoustic classes determined with the hierarchical clustering procedure. Each spectrum is plotted along a Bark frequency scale. The mean of each cluster is plotted between curves that are one standard deviation away from the mean. The label on each cluster corresponds to the labels in Tables 4.2 and 4.3.

Table 4.3: Distributions for ten clusters (percent normalized by cluster).

Phone	1	2	3	4	5	6	7	8	9	10	Total #
y̤	5	2				1		1			131
ṳ	6	3	1			1		1			176
iʸ	35	13	2			3		3			663
ɨ	16	14	12	2	3	7	1	8			893
ɪ	9	21	5	5		1		1			500
eʸ	8	16	1	2		1		1			334
ɝ		1	8	3	1						230
ɚ	1	1	11	1	4	2		2			401
ɾ		2	14	9	5	2		2			559
ɛ	2	7	5	11	1	1		1			408
æ	2	4	6	8	1			1			334
ɑʸ	3	2	4	10	1	1					326
ɑ			3	13	4						371
ʌ			4	9	2						284
ɑʷ			1	3	1						105
ɔʸ			1	1	1						68
u	1	1	1	1	2	1					112
ʊ		1	1	1	1						61
oʷ		1	2	4	8	1					250
ə	2	2	7	5	11	2		2			502
ɔ			1	6	11						322
l	1	1	1	4	20	8		1			586
w			1	1	7	7					265
l̩					6	1					133
ŋ̩											4
m̩						1					24
ɾ̩						1		1			62
n̩						3	1				123
ŋ̩						3	1				132
m					1	12	3	1			424
n	1					16	7	1			735
v						3	3	1		1	246
ɫ							1				97
ɹ							1				94
ə̤								1			42
gᵘ							3				185
bᵘ							5				306
pᵘ							6				372
dᵘ						1	8				505
kᵘ						1	9				555
tᵘ							12			1	712
ɹ						1	17	1			1066
ʔ	1		1	1	2	3	2	4			358
b			1				1	2			123
ð	1	1				2	2	4		1	250
f						1	3	5	2	3	347
ɾ	1					2	2	2			190
θ							1	1	1	1	119
ɦ								1			31
h							1	4	1	1	140
g			1					4	1		151
p	1	1	1		1	1	1	8	1		308
k		1	1		1	1	2	13	5	2	514
t		1					1	11	17	4	509
d		1					1	5	6	3	249
ʒ									1	1	21
ç									12	1	153
j								1	12	3	168
ʃ								1	14	2	193
z							1	1	7	31	541
s						1	1	1	19	44	828
Total #	1094	917	1634	1837	1575	1605	5949	2006	975	1299	18891

Table 4.2 is shown in Figure 4.4. From this figure it is also apparent that the major manner groups are being captured by the set of 10 clusters described previously. Thus, at low levels in the phonetic structure (which correspond to higher points in Figure 4.4) phones with similar features tend to group together, giving rise to classes such as high front sonorants, silence, nasal consonants, etc. Note for instance, that the alveolar and palatal strident fricatives each form distinct acoustic classes before subsequently merging into a single acoustic class. The retroflexed phones merge together to form a single class at a relatively low level in the phonetic structure. The semivowels /w/ and /l/, as well as syllabic /l̩/, also form a common group. The stop consonants tend to group by manner of articulation. The only exception is the /p/ which first groups with the aspirants. This grouping is reasonable since the representation used was an average of the entire segment. Thus, the release of the /p/ would tend to be dominated by the aspiration. A qualitative analysis of these structures, as the number of acoustic clusters increased, showed that after ten clusters the hierarchical organization did not change significantly.

Finally, the phonetic distributions for the clusters obtained from the training data were compared to those from a new set of 500 sentences spoken by 100 different talkers. The percentage difference for a given cluster and phoneme was, on the average, around 1%, suggesting that the results did not change significantly. Closer examination revealed that the larger differences were mostly due to sparse data.

On the basis of these investigations, it appeared that by representing each acoustic segment with an average cross-section of the mean-rate response, and by clustering with a Euclidean distance metric, it was possible to determine a relatively small number of acoustic classes which appear to be capturing general manner properties. These results provided solid evidence that it was possible to determine a robust acoustic description of phonetic events through the use clustering procedures.

An examination of the distributions in Tables 4.2 and 4.3 indicated that although the realization of a phone tended to be distributed into clusters which were acoustically similar, this was not always the case. There appeared to be three reasons for

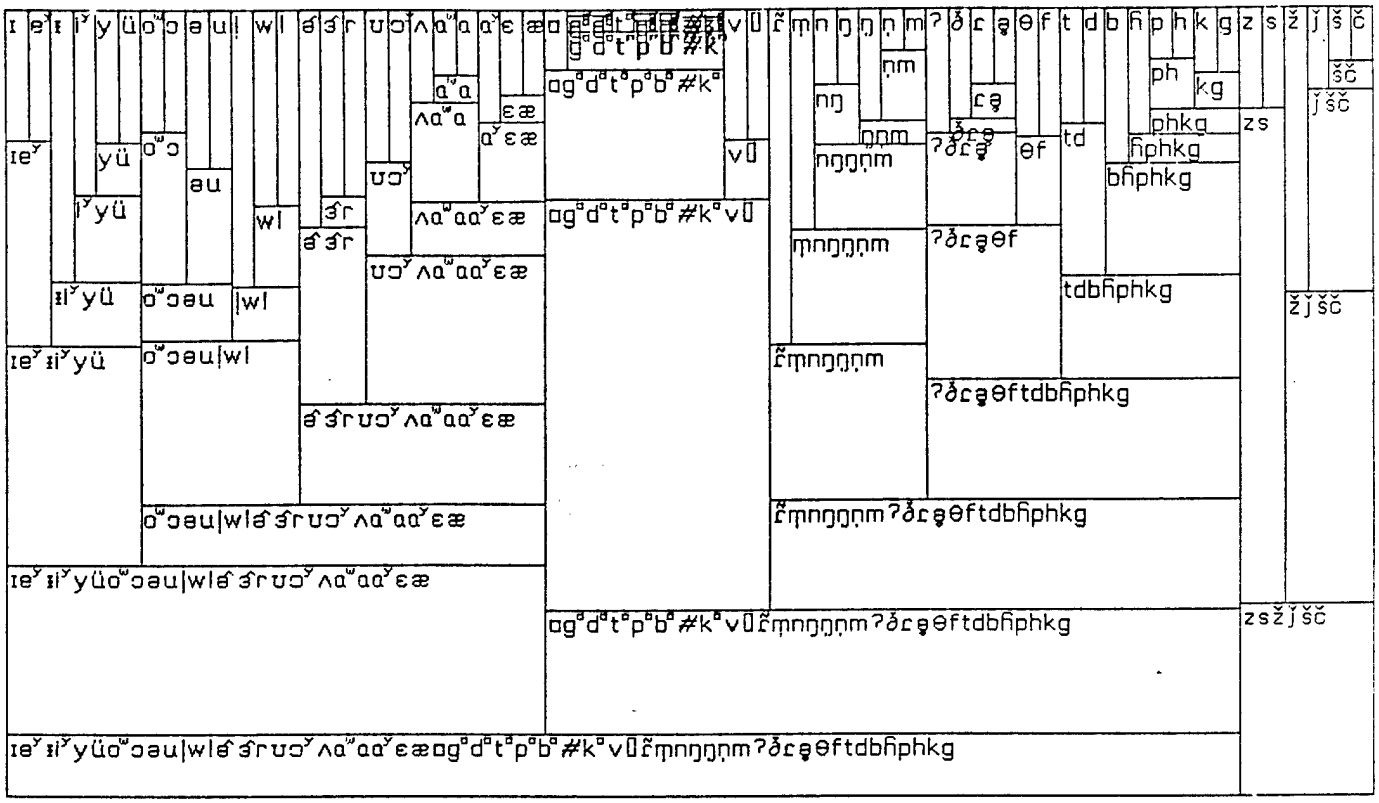


Figure 4.4: Phonetic hierarchical structure with ten clusters.

This figure illustrates the hierarchical structure which results from using the distributions in Tables 4.2 as an observation vector for each phone.

this behavior. First, it became apparent that in some cases, the signal representation was not adequate. For instance, an examination of the distributions indicated that many of the nasal consonants fell into both an acoustic cluster with a low-frequency murmur, as well as a cluster which was predominantly silence. An examination of the hierarchical structure indicated that those nasals which fell into the silence class were often heavily attenuated so that they were virtually flat in the mean-rate response. The utterances used for this work were recorded with a noise-canceling microphone. One of the properties of this microphone is to cancel out sounds which do not emanate directly from the mouth. Thus, sounds primarily transmitted through the nostrils or through tissue vibration are heavily attenuated. The result is that there

is significantly less information in the low-frequency channels than would be present with different recording mechanisms. In terms of the acoustic clustering, this phenomenon suggested that an alternative signal representation, with larger gains in the low-frequency channels, might improve the clustering results in these cases.

A second reason for differences in the distributions could be attributed to the Euclidean distance metric which was used to perform the clustering. An example of this type of behavior is found in the distributions of the [y], or [ɨ], where a consistent fraction fell into acoustic clusters dominated by silence or consonants. An examination of the hierarchical structure indicated that in these cases, the gain was either very large or very small. Since this type of effect would reduce the purity of the mapping between the underlying phonemes and the acoustic classes, an alternative distance metric was explored which subtracted the mean value of each spectrum before computing the Euclidean distance. This distance metric can still be shown to be stepwise optimal, so that the hierarchical clustering procedure remained the same.

When the zero-mean distance metric was substituted for the Euclidean metric it was found that the acoustic classes had a smaller amount of distortion. When combined with the increased gain in the low-frequency channels of the auditory model, an analysis of the acoustic structure showed that there were two additional classes which seemed quite robust. The first additional cluster contained strong low-frequency energy, while the second contained energy in the mid-frequency range. These spectra are illustrated in Figure 4.5. In comparing these clusters to those determined with the Euclidean distance metric, which were shown in Figure 4.3, the standard deviation appears to be consistently smaller for a majority of clusters, even though the average spectral shapes are often very similar. The normalized distributions of these twelve clusters are shown in Table 4.4. By comparing these distributions with those of Table 4.2 it can be seen that the phone distributions are often more tightly centered around acoustically similar clusters. For instance, 74% of the [iʸ] tokens fell into cluster 1, and 12% fell into cluster 2, using the zero-mean distance metric, compared to 57% and 18% respectively for the Euclidean distance metric.

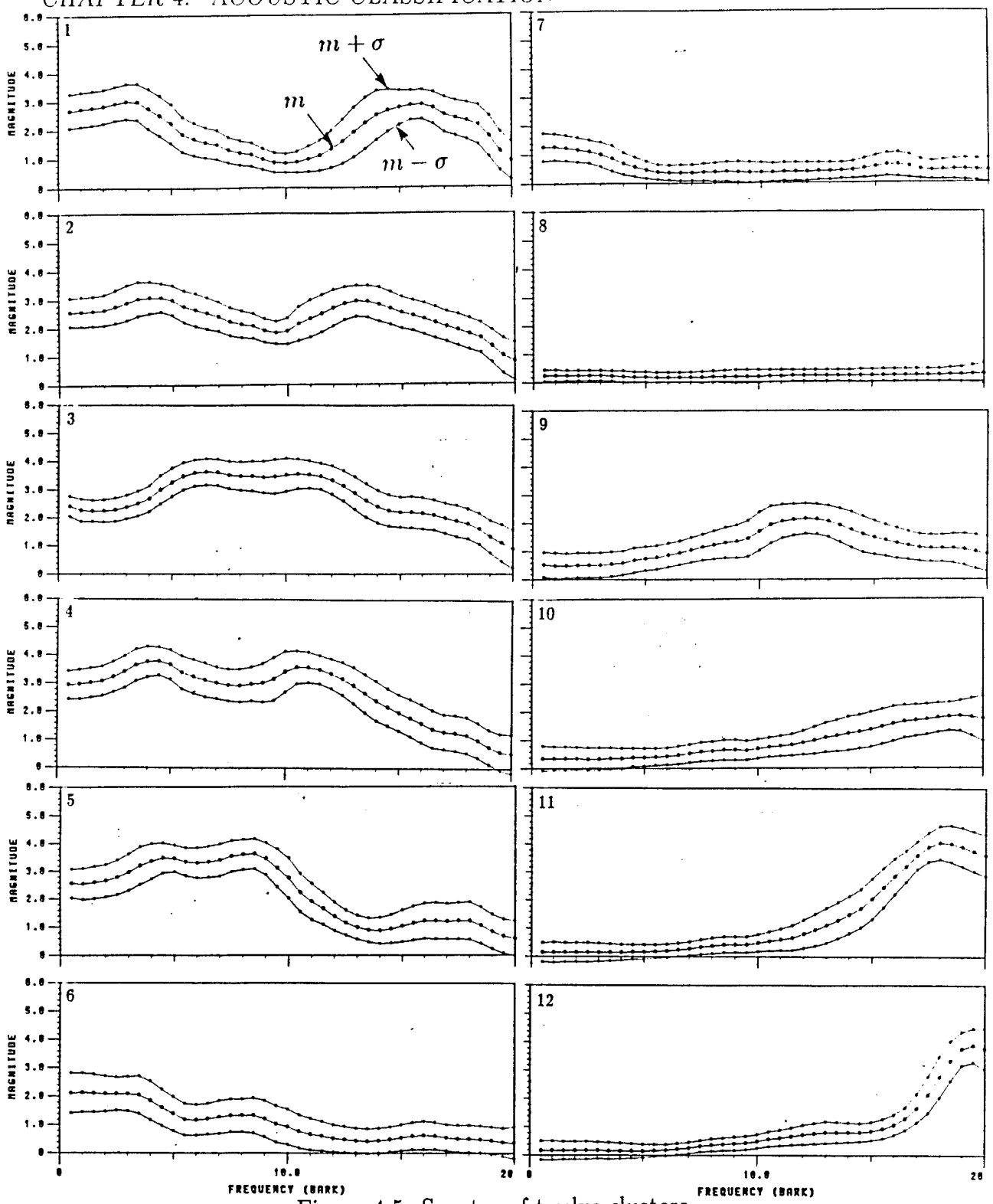


Figure 4.5: Spectra of twelve clusters.

This figure illustrates the spectra of twelve acoustic classes determined with the hierarchical clustering procedure and the zero-mean distance metric. Each spectrum is plotted along a Bark frequency scale. The mean of each cluster is surrounded by one standard deviation. The label on each cluster corresponds to the labels in Table 4.4.

Table 4.4: Distributions for twelve clusters (percent normalized by phone).

Phone	1	2	3	4	5	6	7	8	9	10	11	12	Total #
iʸ	74	12		1		4	4	2		1			670
y̞	53	20		3		6	14	1		2	1		137
ü	35	34		10	1	12	6	3					178
eʸ	35	43	3	1	1	6	4	3	2	1			334
ɪ	24	46	4	12	2	7	2	2	1				501
ɨ	19	39	4	12	1	12	6	6	2	1			896
ɛ	3	42	21	17	3	3	1	6	2	1			412
æ	1	33	44	4	1	1	2	7	5	1			336
ʌ		6	46	27	12	3	1	3	1				284
ɑʸ	9	14	62	4	2	2	2	2	4				323
ɑʷ		1	74	6	10	5		3	2				109
ɑ		1	69	13	12	1	1	2	1				365
ɚ	2	24	7	51	3	3	3	3	2				410
ɹ	1	14	10	59	6	5	1	1	3				568
ɝ	1	13	6	75	2	1	2						225
ə	1	12	16	30	22	10	3	4	1				508
ɔʸ	8	17	17	22	25	9	2			2			65
u	3	14	1	26	21	28	5	2					110
ʊ	3	30	7	21	25	11		3					61
oʷ		5	16	15	52	6	3	2					248
ɔ			34	6	56	1	1	1	1	1			321
l̩			2	1	83	11	1	3	1				132
l	1	2	5	6	51	27	4	2					582
w		1	1	6	49	34	6	3					270
m̩			4			46	38	12					24
n̩	2	4	1	2	2	34	34	21	1				123
m	1	1	1	1	1	30	49	15	1				429
ŋ	3		1		2	21	51	21		1			140
n	1	1				18	54	24	1				731
ɾ̃		13	5	2		18	35	19	6	2			62
ɲ							67	33					3
[illegible]						1	3	78	16	1			93
ɡᵃ	1	1				1	17	79	1	1			183
ɢᵃ							13	83		3			490
tᵃ							4	89	1	5		1	707
dᵃ						1	9	90	1				289
kᵃ							6	92	1	1			534
[illegible]						1	2	94	2	1			101
pᵃ							2	98					356
[illegible]							1	98	1				1060
v		2	1	1	1	9	14	59	3	8			237
ð	1	2				3	11	59	8	15			246
θ							4	54	4	34	2	3	114
ʔ	4	11	13	1	4	2	7	45	9	5			357
ɚ	9	2				9	7	43	14	14		2	44
ɾ	8	10		2		2	22	42	10	4		1	189
f							1	42	14	40	1	1	339
ɦ	6	3	6		3		12	22	25	22			32
b	3	2	6	1	2	1	2	34	30	19		1	112
p		1	8					28	43	19	1		307
h	1	1	5				2	27	40	24			140
k			6		1			18	43	27	4	2	513
g	1	1	11	1		1	1	12	43	28		1	149
d	1	1					2	14	17	54	5	6	242
t			1				1	10	18	53	11	6	494
š		1	1				1	7	2	16	68	6	186
ǰ								2	2	18	69	9	171
č		1						5	4	16	69	5	150
ž								5	5	24	52	14	21
z							1	6		19	13	61	535
s							1	5	1	12	16	64	840
Total#	1186	1700	1500	1483	1230	1084	1375	5257	978	1351	665	979	18788

A third reason for smearing in the distributions was due to blurring of dynamic behavior in an acoustic segment caused by representing each segment by a single average spectrum. This effect can be seen in Table 4.2 by comparing the diphthongs /ɑʸ/ or /ɔʸ/ to vowels such as /ɑ/ or /ɔ/. The latter two vowels have more tokens centered in a single acoustic cluster than do the diphthongs. The diphthongs tend to have more tokens in clusters 1 or 2, which tend to contain more of the front vowels. This point can be seen very clearly by noting that the vowels /ɑ/ and /ɔ/ have no tokens in clusters 1 or 2, whereas the two diphthongs have 15% and 11% of their tokens in these two clusters. This trend is even more pronounced in Table 4.4. This observation would seem to indicate that the information in the diphthongs was being blurred by the spectral averaging. As was mentioned earlier, this phenomenon motivated a more sophisticated clustering procedure which will be described later in the following section.

The final source of change in the distributions was not due to the signal representation, nor the distance metric, but appeared to be caused by basic regularities in the acoustic realization of a phone. The phone [ð] for example, appeared to fall into several classes, one of which was dominated by a low-frequency energy, while another was dominated by a high-frequency energy. This observation motivated an investigation into determining acoustic regularities of individual phonemes. The results of these investigations are reported in the following section.

4.3 Finding Regularities in the Realization of Phonemes

4.3.1 Weak Voiced Fricatives

For reasons just mentioned, the first phoneme to be examined in more detail was /ð/. In order to obtain as much as data as possible, the two 100 talker datasets were combined to form a single dataset. There were 473 instances of [ð] in this larger dataset. The clustering procedures incorporated in the last section were essentially

Table 4.5: Summary of distributions of preceding contexts for /ð/.

Preceding Context	ð-1 (% of total)	ð-2 (% of total)	Total (# of tokens)
+obstruent	81	19	286
+sonorant	25	75	187

duplicated on these data. An average mean-rate response of aligned dendrogram segments was used for the acoustic representation. The pre-clustering algorithm generated a set of 34 seed clusters with more than 2 members. The hierarchical clustering procedure was then applied to these seed clusters.

An analysis of the results indicated that there were two types of /ð/'s, which were acoustically quite distinct, as illustrated in Figure 4.6 along with spectrograms of prototypical utterances. An examination of the contexts in each cluster indicated that the following context did not significantly affect the nature of the phoneme. In American English, the phoneme /ð/ is nearly always syllable initial, and is therefore nearly always followed by a vowel. The nature of the vowel did not appear to affect the realization of the phoneme. The preceding context appeared to play a more important role however. In particular, when the phoneme was preceded by silence or an obstruent, it was classified into cluster 'ð-1' 81% of the time. However, if the phoneme was preceded by a sonorant, it was classified into cluster 'ð-2' 75% of the time. A more detailed summary is found in Table 4.5.

When this process was repeated on 437 /v/'s using an average spectral shape, two distinct acoustic classes were observed. As shown in Figure 4.7, the top two clusters had similar spectral characteristics as those found for /ð/. Thus, it would appear that the two weak voiced fricatives have similar systematic patterns of behavior.

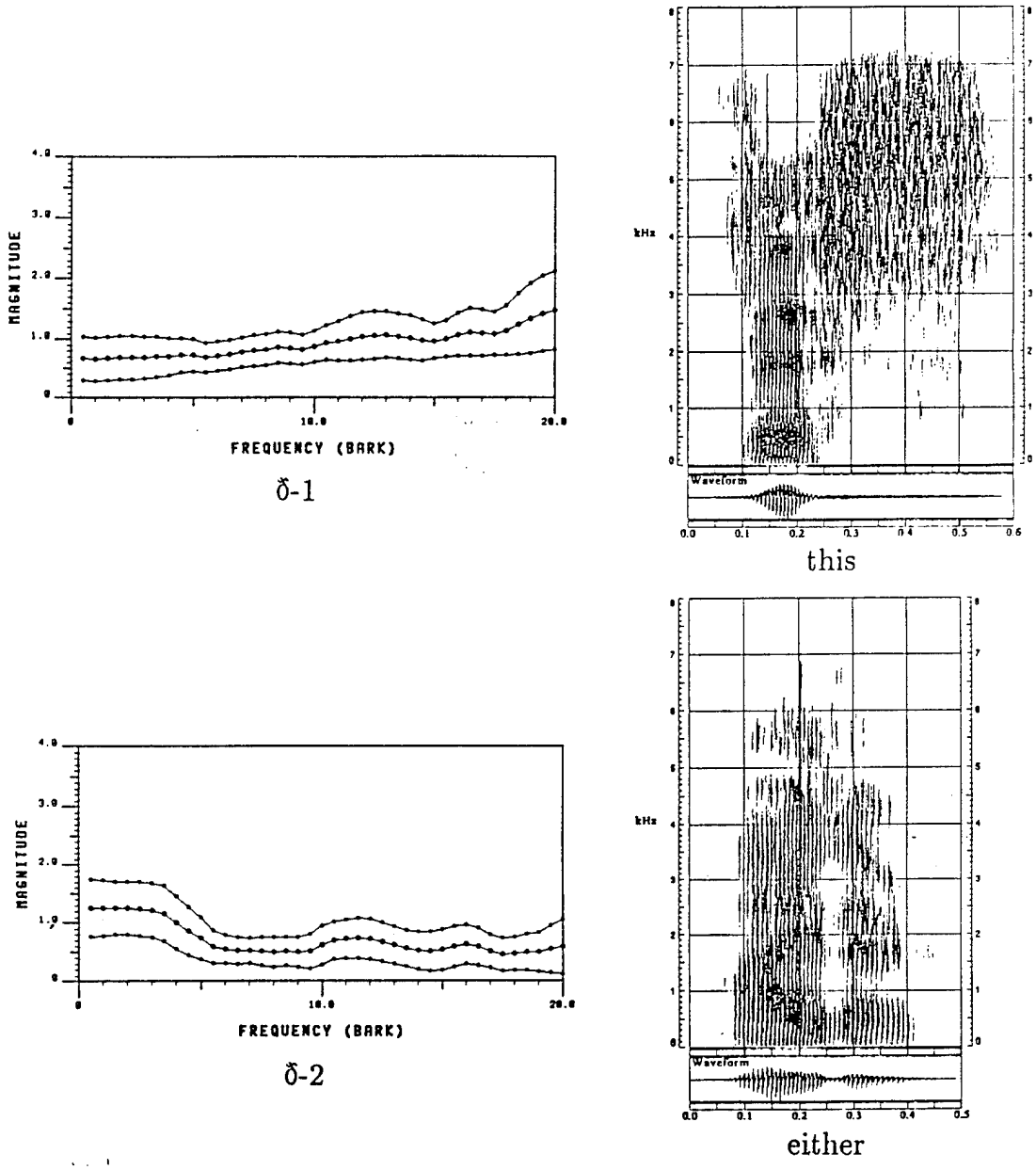


Figure 4.6: Spectra of /ð/ clusters.

This figure illustrates the top two acoustic clusters associated with the /ð/. Each spectrum is plotted along a Bark frequency scale. The mean of each cluster is surrounded by one standard deviation. The label on each cluster corresponds to the labels in Table 4.5. Spectrograms of prototypical contexts are illustrated on the right.

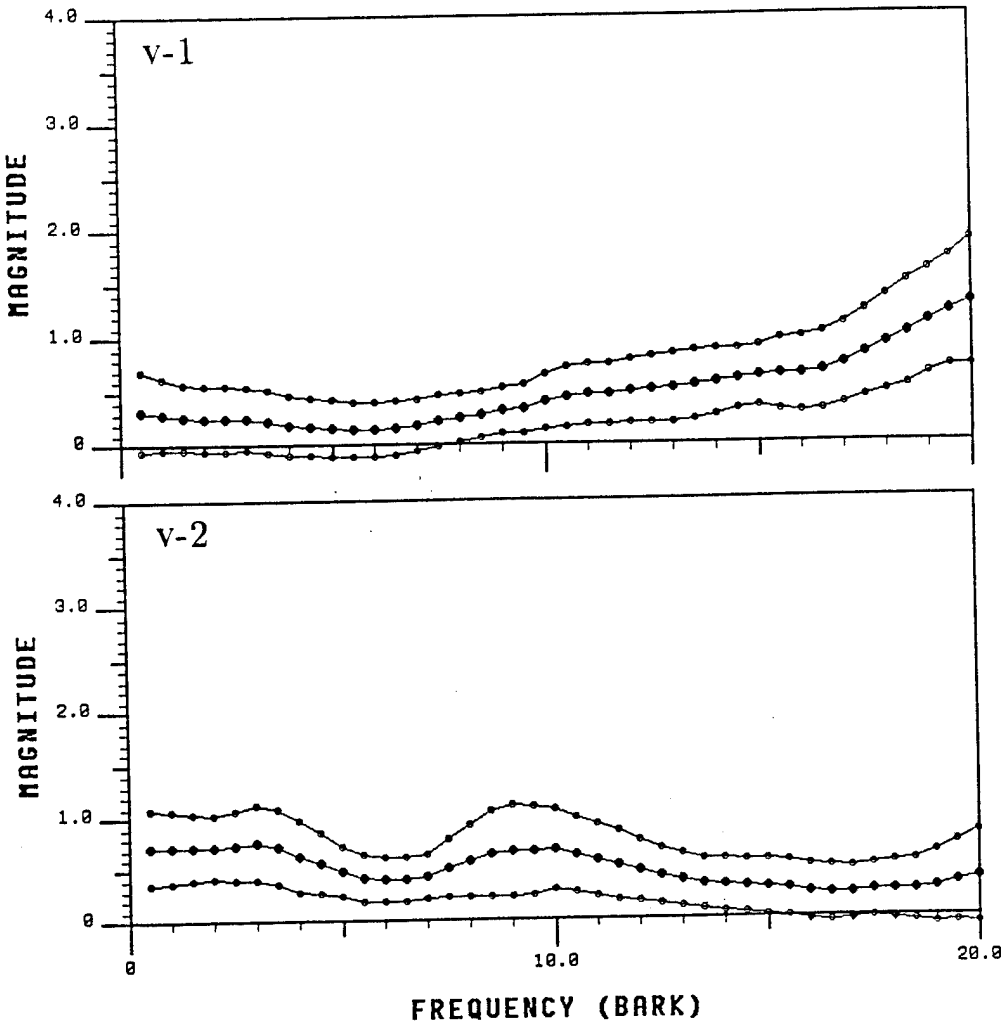


Figure 4.7: Spectra of /v/ clusters.

This figure illustrates the top two acoustic clusters associated with /v/. Each spectrum is plotted along a Bark frequency scale. The mean of each cluster is surrounded by one standard deviation. These clusters have shapes that are very similar to the two clusters found for /ð/, shown in Figure 4.6.

4.3.2 Velar Stop Consonants

The next phoneme to be examined in more detail was /k/. This phoneme was investigated because of a belief held by many phoneticians that there are two allophones of this consonant, one corresponding to a 'front' /k/, the other corresponding to a 'back' /k/. The goal of the study was determine if it was possible to motivate these two allophones from an analysis of acoustic regularity of the release of the stop consonant. In fact, the results of this study indicated that these two phenomena were readily found in the data. In addition, there was evidence for a third systematic regularity as well. The investigation is described in more detail in the remainder of this section.

In order to eliminate the influence of the following vowel during the aspiration portion of the stop release, the analysis was focused around the release of the stop. This was achieved by computing the average mean-rate response from the shortest acoustic segment located at the release of the stop. The average duration of such segments was less than 24 ms, compared to an average voice onset time of over 57 ms for /k/. Figure 4.8 illustrates the fraction of time taken by the acoustic sub-segments compared to the actual voice onset time of the phonetic token. Note that the set of points forming a diagonal in the scatter plot are cases where there was but a single acoustic segment corresponding to the release of the /k/.

In the 1000-utterance dataset, there were 952 [k] tokens which had a release. In the same manner described previously, the pre-clustering algorithm was used to select a set of seed clusters. In this case, when a threshold was selected to maximize the number of clusters with more than 2 members, 77 seed clusters were generated. The hierarchical clustering procedure was then run on these clusters.

From these classifications it became apparent that the spectral cross-section sampled at the release varied substantially, and that there were at least three robust acoustic clusters present in the data. These clusters are illustrated in Figure 4.9, along with spectrograms of prototypical utterances. An analysis of the immediate

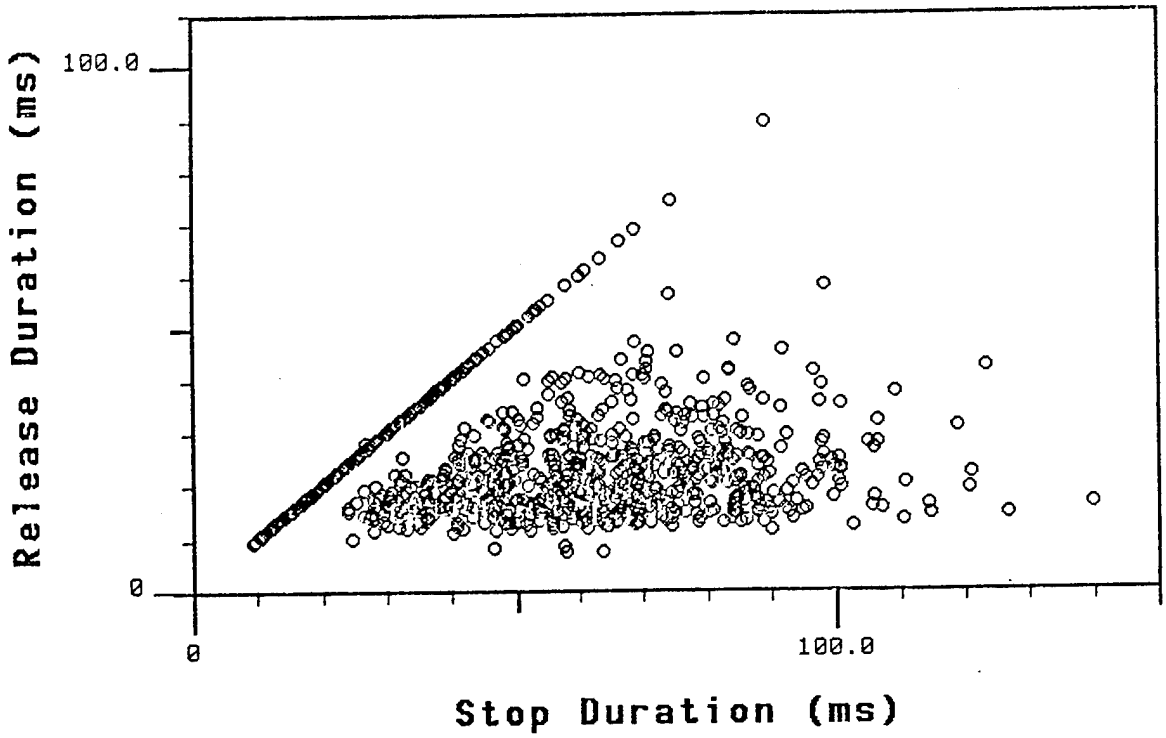


Figure 4.8: Segment duration comparison for /k/.

This figure presents a scatter plot comparing the voice onset time (horizontal axis) to the duration of the acoustic segment chosen to represent the burst release. All durations are in ms.

phonetic context indicated that the preceding context was nearly always either a closure or a silence (there were 3 instances of a preceding /ŋ/, and one instance of a preceding /ɑ/). Thus, the investigation of the role of context focused on the following phone. The distributions of the following contexts have been summarized in Table 4.6 for the three clusters illustrated in Figure 4.9. For example, when a /k/ was followed by the vowel /iʸ/, all of the spectral cross-sections taken from the acoustic segment nearest the release fell into class 'k-1.'

The distributions in Table 4.6 indicate that the following context played a significant role in the acoustic realization of the release. In particular, as in previous cases, there appeared to be major groups with similar distributional properties. The major classes tended to cluster together in terms of front vowels, unrounded back vowels, and rounded back vowels, although /u/ and /ʊ/ do not fit this last trend.

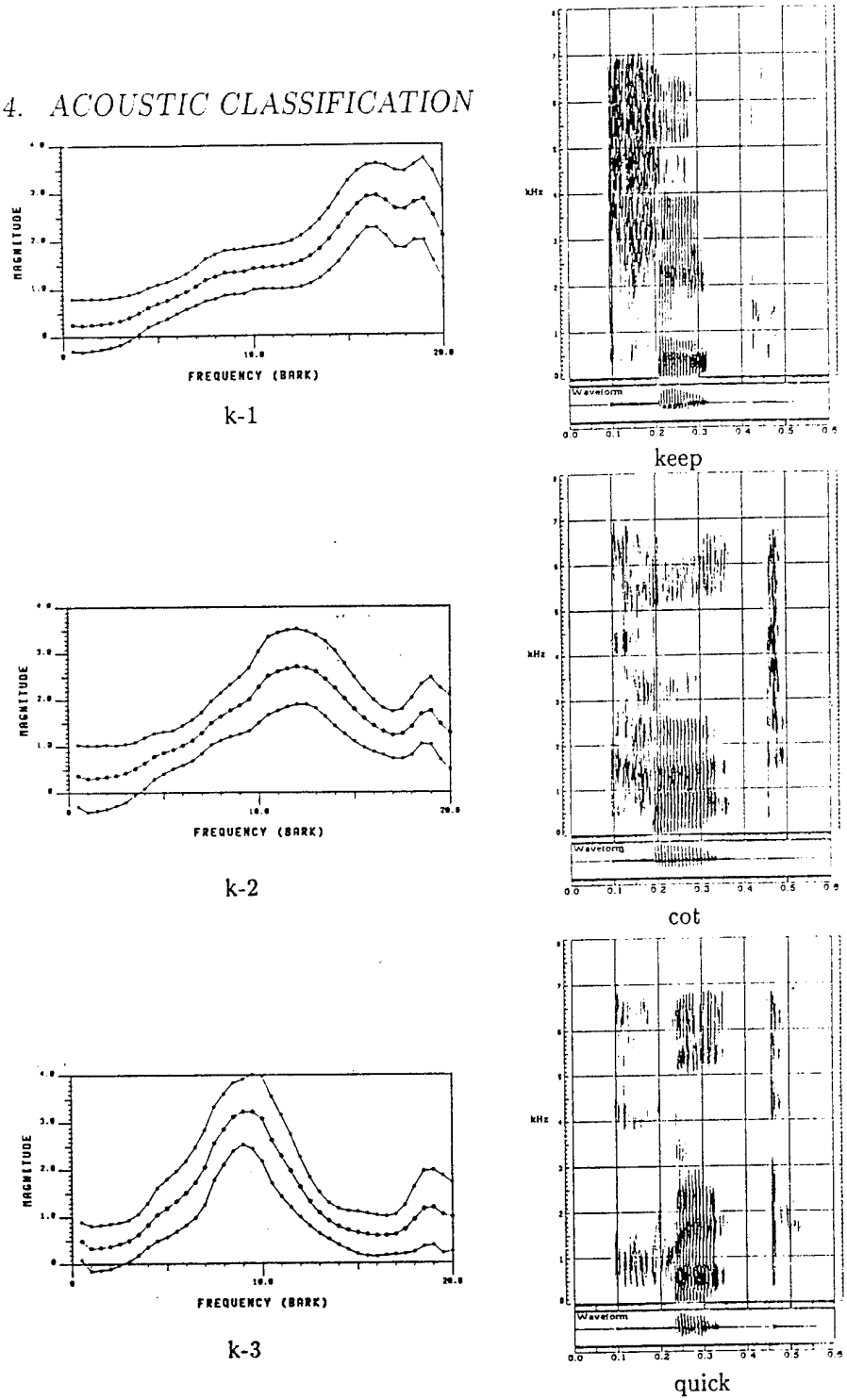


Figure 4.9: Spectra of /k/ clusters.

This figure illustrates the top three acoustic clusters associated with the release of the /k/. Each spectrum is plotted along a Bark frequency scale. The mean of each cluster is surrounded by one standard deviation. The label on each cluster corresponds to the labels in Table 4.6. Spectrograms of prototypical contexts are illustrated on the right.

However, there is not much data available for these two vowels, and either of them could have been partially fronted. A summary of distributions grouped according to major phonetic features may be found in Table 4.7. For example, this table illustrates that when a /k/ was followed by a front vowel the acoustic segment associated with the release of the stop fell into class 'k-1' 88% of the time. Similarly, when a /k/ was followed by a retroflexed sonorant, the acoustic segment was classified into cluster 'k-2' 67% of the time.

As a means of checking that the spectral shape was not being influenced by the presence of aspiration in the release of the /k/, and to verify that the number of clusters was stable, a similar clustering experiment was performed on 294 /g/ tokens from the same set of data. The results of the hierarchical clustering experiment indicated that there were also three robust acoustic clusters with important contextual dependencies. The three top /g/ spectral clusters are shown in Figure 4.10. These spectra are extremely similar to those determined for /k/. Thus, in a fashion similar to the weak voiced fricatives, it would appear that the velar stop consonants have similar systematic patterns of behavior.

4.3.3 Modeling Time-Varying Changes

As was mentioned previously, one of the limitations of the acoustic representation used for the classification analysis is that a phone is mapped to a single acoustic segment. As a result, the time-varying characteristics cannot be adequately captured. In light of these limitations, some preliminary studies of the vowel /æ/ were made, except that the vowel was represented as a sequence of two acoustic segments, each represented by an average spectral vector. A distance between two sounds was the combined distance between each of the respective subsegments. In the 1000 utterance dataset described previously, there were 571 examples of [æ]. The pre-clustering procedure generated a set of 48 seed clusters. The hierarchical structure produced two clusters which appeared to be significantly different in the preliminary portion of

Table 4.6: Distributions of some following contexts for /k/.

Following Context	k-1 (%)	k-2 (%)	k-3 (%)	Total (#)
iʸ	100			24
y	100			22
æ	98	2		55
eʸ	96	4		26
ɪ	88	12		17
ɛ	88	12		25
ɑʷ	87	13		23
ɨ	72	18		58
ü	67	33		24
ɑʸ	33	66		6
ɑ	23	61	16	51
ʌ	22	76	2	58
ɝ	19	78	3	27
ə	14	79	7	29
ɚ	8	85	7	13
u	17	66	17	6
oʷ	12	65	23	26
ʊ	27	64	9	11
r		58	42	55
l̩	2	48	50	54
ɔ	3	44	53	45
l	9	38	53	70
w		25	75	53
ɔʸ		22	78	9

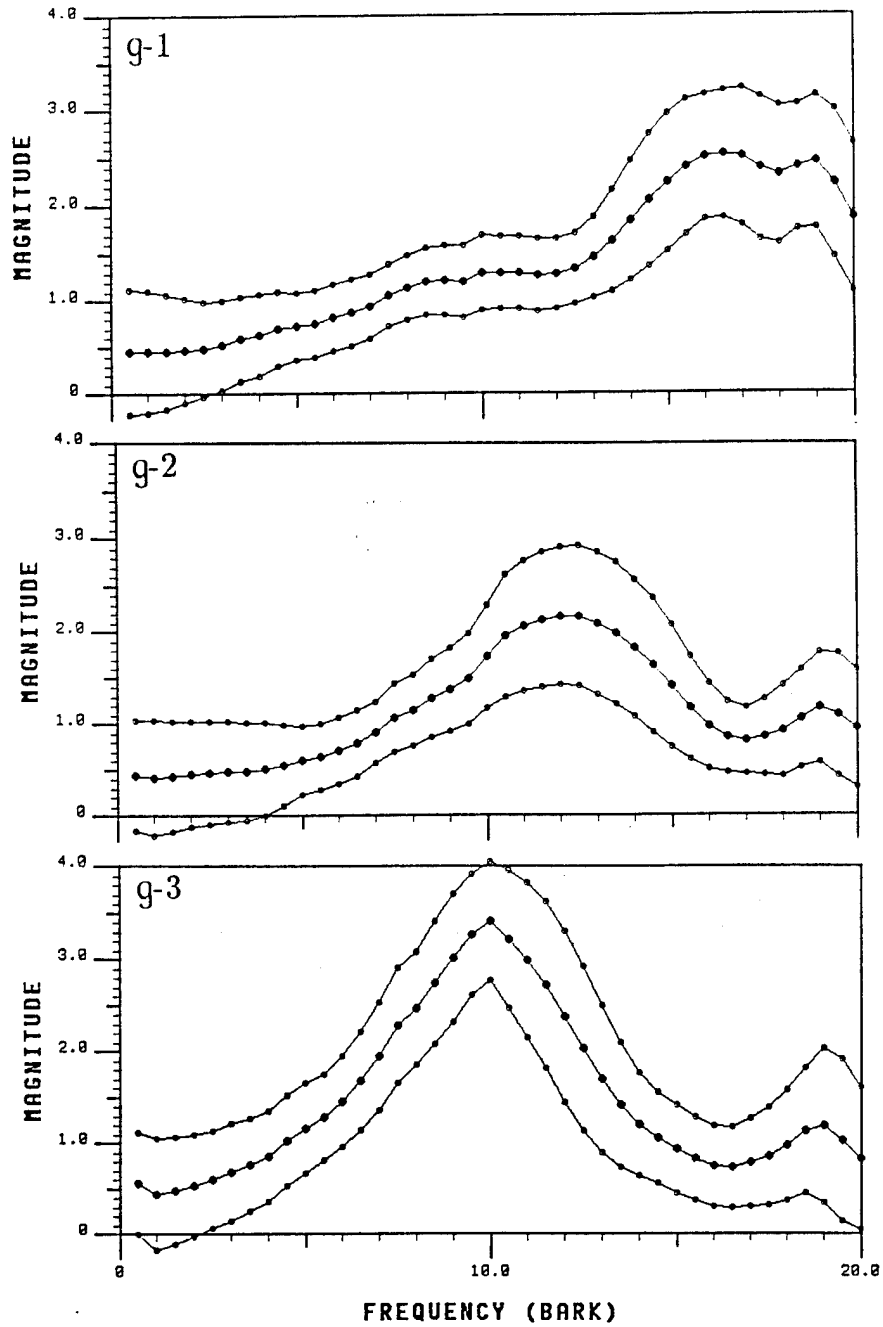


Figure 4.10: Spectra of /g/ clusters.

This figure illustrates the top three acoustic clusters associated with the release of the /g/. Each spectrum is plotted along a Bark frequency scale. The mean of each cluster is surrounded by one standard deviation. These clusters have very similar shapes to the three clusters found for /k/, shown in Figure 4.9.

Table 4.7: Summary of distributions of following contexts for /k/.

Following Context	k-1 (% of total)	k-2 (% of total)	k-3 (% of total)	Total (# of tokens)
+front	88	12		274
+back, −round	21	71	8	154
+back, +round	5	42	53	150
+retroflex	6	67	26	95
+lateral	6	43	52	124
+fricative	27	69	4	97

Table 4.8: Summary of distributions of following contexts for /æ/.

Following Context	æ-1 (% of total)	æ-2 (% of total)	Total (# of tokens)
+nasal	30	70	205
−nasal	69	31	351

the vowel. A plot of the two distributions is shown in Figure 4.11. From this figure it is apparent that one of the most noticeable differences between the two clusters is that the first portion of 'æ-2' has a higher second formant than the first portion of 'æ-1.' An examination of the immediate phonetic contexts of each cluster indicated a strong tendency of the 'æ-2' to precede a nasal consonant. As shown in Table 4.8, when a vowel was followed by a nasal, it fell into the 'æ-2' category 70% of the time, while when a vowel was not followed by a nasal, it fell into the 'æ-1' category 69% of the time. This result seems to provide additional acoustic support for the raised /æ/ phenomenon discussed in the literature [66].

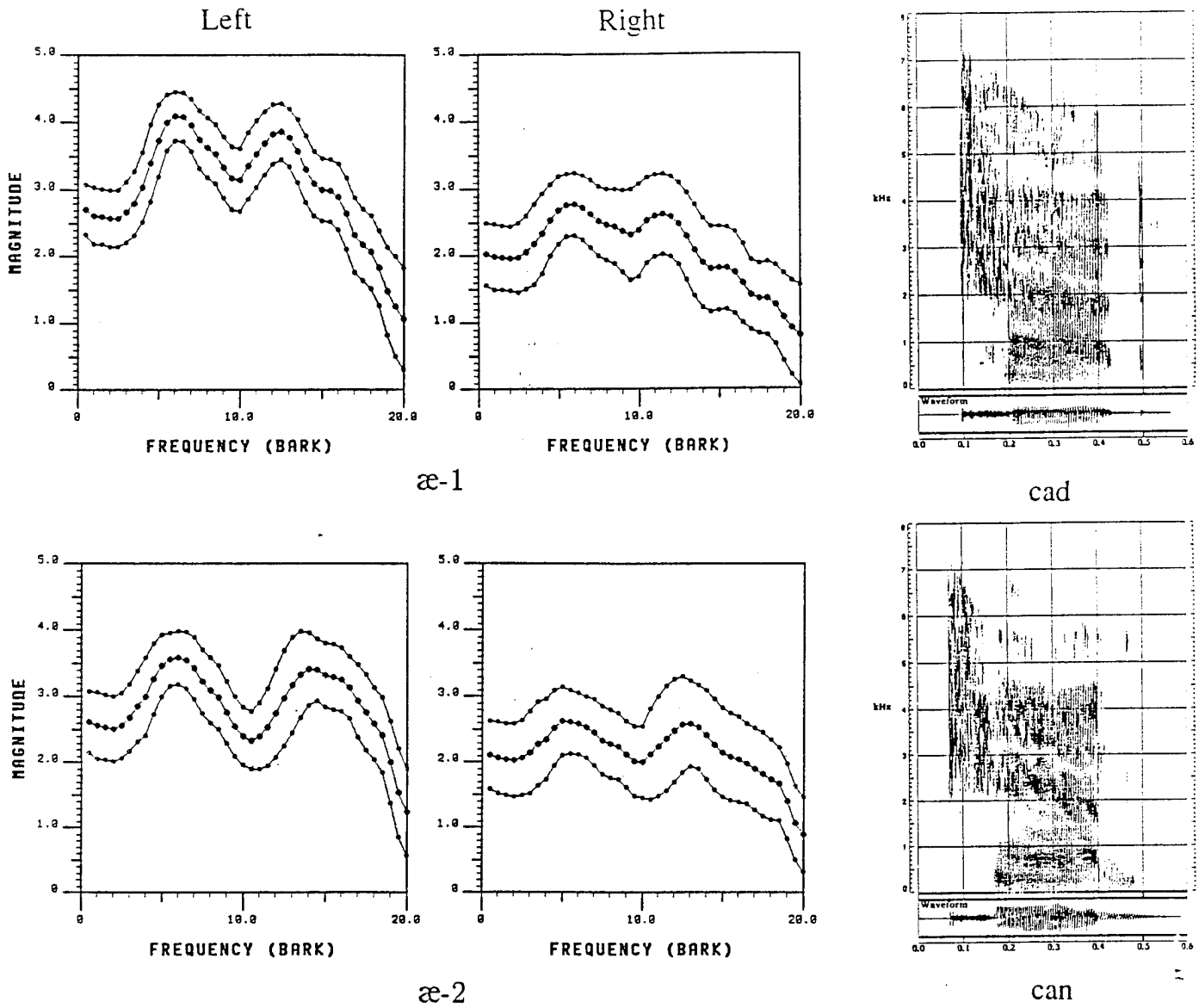


Figure 4.11: Spectra of /æ/ clusters.

This figure illustrates the top two acoustic clusters associated with the vowel /æ/. Each cluster corresponds to a left spectra and a right spectra. Each spectrum is plotted along a Bark frequency scale. The mean of each cluster is surrounded by one standard deviation. The label on each cluster corresponds to the labels in Table 4.8. Spectrograms of prototypical contexts are illustrated on the right.

4.4 Chapter Summary

This chapter investigated procedures to automatically determine acoustic labels from a large body of training data. Since it was not feasible to attempt an exhaustive study due to limitations of the signal representation and training data, two smaller studies were made. The first study attempted to determine if it was possible to uncover major sources of acoustic regularity that capture basic properties of all phonemes. This was accomplished by applying a hierarchical clustering procedure to data from all phonemes, using the mean-rate response as the basis of the signal representation. The results of this study indicate that it is possible to assign an acoustic segment to one of a small set of acoustic categories, each having a meaningful phonetic distribution.

The second study attempted to demonstrate that it was possible to capture important context dependencies of individual phonemes. This was done by applying a hierarchical clustering procedure to the weak voiced fricatives, /ð/ and /v/, and the velar stop consonants, /k/ and /g/. A slightly modified version of this procedure was also applied to /æ/. The results of this study indicate that it is possible to determine a finite number of regular acoustic forms which capture consistent contextual dependencies. Additionally, there is evidence that these regularities can generalize across sets of phonemes with similar phonetic features.

Chapter 5

Signal Representation Refinements

This chapter presents the results of two investigations which attempted to improve the characteristics of the mean-rate response. The first study attempted to enhance the temporal and spectral properties of the mean-rate response outputs by sampling the outputs of the auditory model pulse-synchronously. The second study performed a dimensional analysis of the mean-rate response outputs in order to determine if a more compact representation of the outputs could be found. These studies were motivated by observed inadequacies of the mean-rate response for the tasks of acoustic segmentation and classification, which were reported previously.

5.1 Limitations of the Mean-Rate Response

The mean-rate response values depend upon the amount of post smoothing or averaging performed in each channel of the auditory model. Unfortunately there are practical reasons why no single amount of smoothing is satisfactory for all speech sounds. If the channel outputs are not smoothed enough then a significant amount of ripple will be present during voiced periods. If a spectral representation is created by downsampling these smoothed outputs at fixed time intervals, then the resulting spectrum will also exhibit these temporal fluctuations. This phenomenon is known as 'pitch ripple,' and can be clearly seen in the last syllable in Figure 2.3. Pitch ripple causes difficulty for algorithms attempting to segment the speech signal, because there can be substantial differences in the values of nearby spectra. If the channel outputs

are smoothed excessively however, the temporal resolution necessary to resolve short-duration events will be greatly reduced, thus making it difficult to automatically locate many short-duration events, such as stop bursts.

The trade-off between too much and too little temporal resolution is not unique to the mean-rate response, but is a classic problem faced by most spectral representations. Typically, the parameters of most representations are designed to minimize the amount of pitch ripple present in the spectrum [23,94,98]. This resolution of the problem seems rather unsatisfactory however, especially since it appears that the unsmoothed auditory outputs contain a tremendous amount of information about the temporal structure of the speech signal. As illustrated in Figure 5.1, all channels respond to sudden events, such as the release of a stop consonant, or the onset of a glottal pulse. If the locations of these events were known, it would be possible to sample the outputs synchronous to these events, instead of the fixed-rate sampling procedure normally used. Good temporal resolution could be maintained, since less smoothing would be required, yet the problem with pitch ripple would be avoided since spectra could be sampled at a consistent point in each pitch period. Different variations of this idea (commonly known as pitch-synchronous analysis) have been explored by many investigators [47,50,80,87]. The approach described here is slightly more general in that a pulse-synchronous analysis is advocated, which would be synchronous to all events in the speech waveform. This representation would also be computed solely from information provided by an auditory model.

The following section presents a three stage procedure which was developed for pulse-synchronous sampling of the second stage outputs. The first part of this procedure produces a signal containing useful information about the excitation in the speech waveform. The method for obtaining this signal is a slightly modified version of one proposed by Seneff for extracting periodic information from speech [103]. The second stage of the procedure derives sampling points from the excitation signal, which are used in the third and final stage to produce a pulse-synchronous spectral representation.

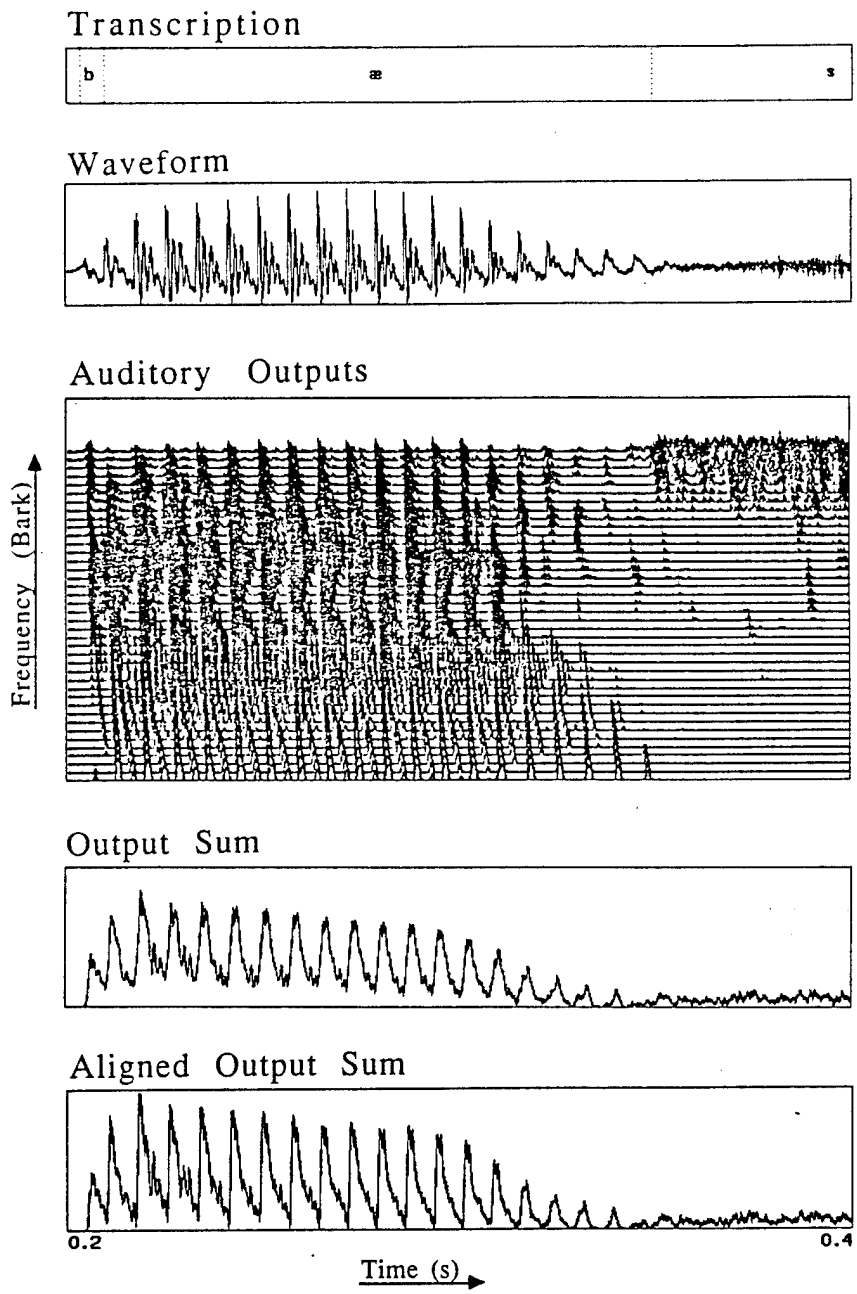


Figure 5.1: Channel sum waveforms.

This figure provides a magnified view of the second stage outputs for the second syllable of the word 'ambassador.' From the top the displays are (1) aligned phonetic transcription, (2) speech waveform, (3) second stage outputs, (4) a sum across all channels, and (5) a sum across all aligned channels (see text).

5.2 Synchronous Analysis of Speech

5.2.1 The Channel Sum Waveform

The middle display of Figure 5.1 illustrates the fact that all channels respond to sudden events occurring in the speech signal. In this figure, there is clear response to the release of the stop consonant and to all points of the onset of glottal closure. In voiced regions the responses to excitation are most clearly located in channels which are not centered over any vocal-tract resonance. It is possible to reduce the effect of the temporal fluctuations due to excitation by smoothing the outputs in each channel in time. Alternatively, it is possible to *enhance* these variations by summing the outputs across all channels. This effect has been noted previously by other researchers such as Searle *et al.* [100], and Seneff [103]. An example of the channel sum waveform is shown in the display immediately below the auditory outputs in Figure 5.1. This waveform exhibits strong peaks in response to sudden events occurring in the speech waveform.

Due to the different filtering performed on each channel there is a small delay between their respective response times. This variation in delay, which is on the order of 1.2 ms over the course of the 40 channels, is illustrated in the top panel of Figure 5.2, which shows the response of the auditory model to an impulse. By assuming a simple linear relationship among the delays of the different channels, it is possible to produce a reasonable alignment of channel outputs. As shown in the bottom panel of Figure 5.1, this alignment results in a channel sum waveform with sharper onsets, and larger peak-to-valley ratios than a sum of the unaligned waveforms. The net result is that the aligned sum is capable of distinguishing events in the auditory outputs with higher resolution than the sum waveform produced from unaligned outputs.

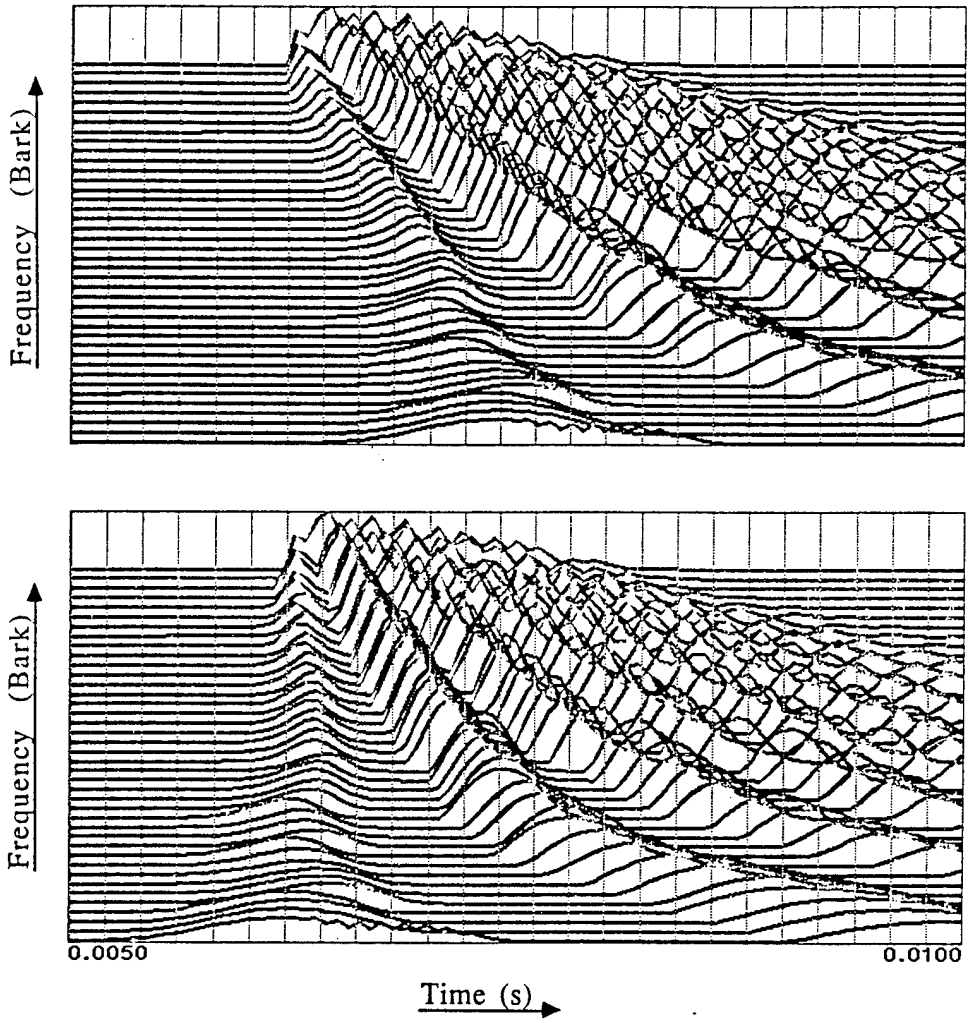


Figure 5.2: Impulse response of the auditory model.

This figure shows the response of the second stage outputs to an impulse. The top display shows the delay among the different channels when there is no alignment. The bottom display shows the delay when a linear offset of 0.03 ms/channel is incorporated.

5.2.2 Determining Event Locations

One of the most critical steps in the synchronous analysis procedure involves determining where significant events occur in the speech waveform. Currently, events are determined from the channel sum waveform as illustrated in Figure 5.3. First, the channel sum waveform is differentiated and smoothed by convolution with a derivative of a Gaussian filter ($\sigma = 3$ ms). The amount of smoothing determines the upper resolution of a fundamental frequency, and determines how sensitive the algorithm is to spurious events.

Local maxima in the derivative of the channel sum waveform correspond to points of maximal onset. As illustrated in Figure 5.4, the locations of these onsets are not useful places to sample the auditory outputs since all channels tend to be responding strongly at these points. If there is no vocal-tract resonance in a given channel, however, the channel output will decay according to the bandwidth of the channel itself. Conversely, if the channel is centered over a vocal-tract resonance having a smaller bandwidth than the channel, then the channel outputs will decay at a slower rate. By delaying the sampling, the resulting spectral shape will be superior than that obtained by sampling at the point of maximum onset. Currently, the outputs are sampled at the point of maximum rate of offset of the channel outputs. This location appears to correspond to a point where weaker channel outputs have decayed substantially, whereas channels centered over significant resonances have not yet decayed. This point serves to illustrate another property of the synchronous response: since it is sampled in a consistent manner, it produces a more consistent spectral representation.

5.2.3 Producing the Synchronous Response

A fixed-rate analysis is computed by low-pass filtering the auditory outputs and sampling the filtered outputs at fixed time intervals. A benefit of this procedure is

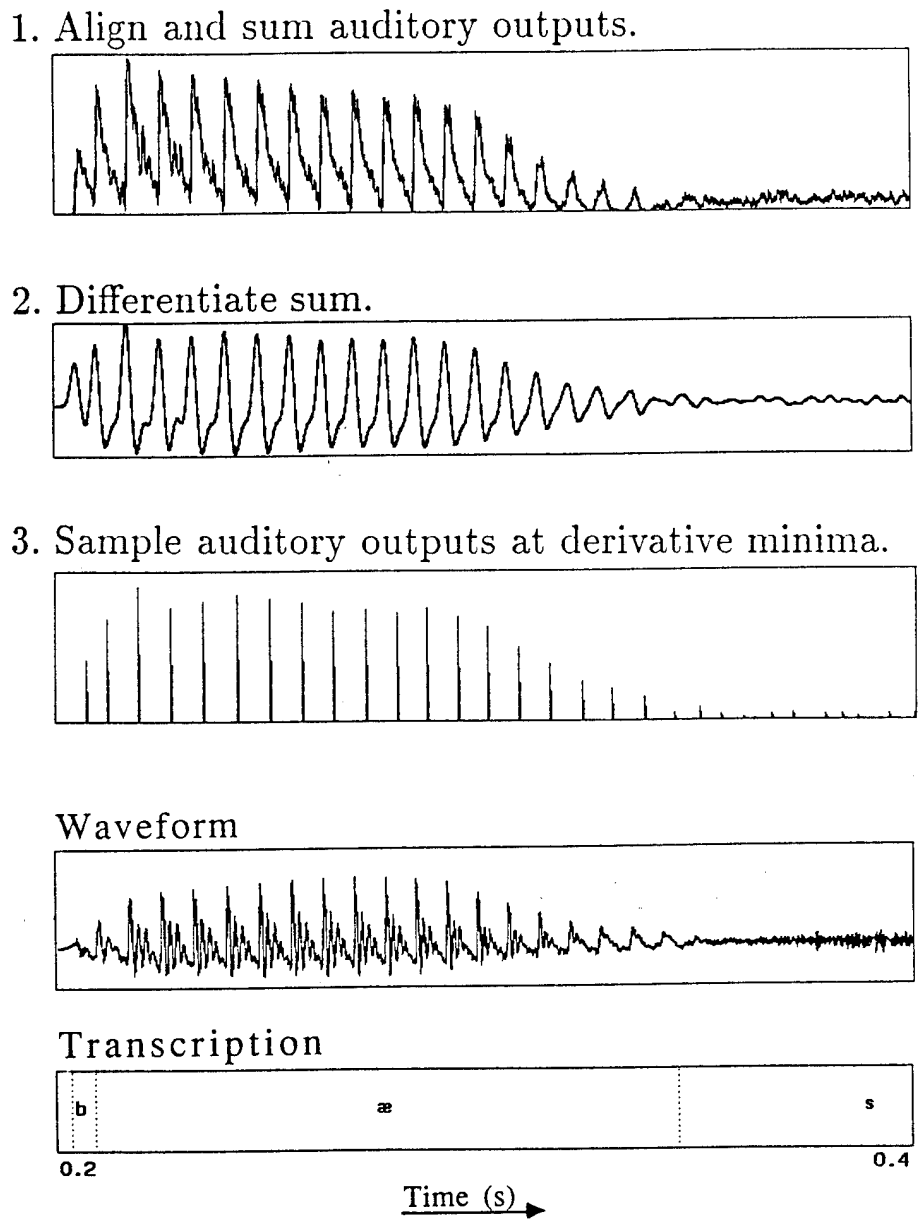


Figure 5.3: Computing event locations.

This figure illustrates the three steps necessary to compute the location of important sampling points in the auditory outputs. These data are taken from the middle syllable of the word 'ambassador.' First, the aligned channel sum waveform is computed. Second, the derivative of the sum waveform is computed. Third, sample points are located at local minima in the derivative.

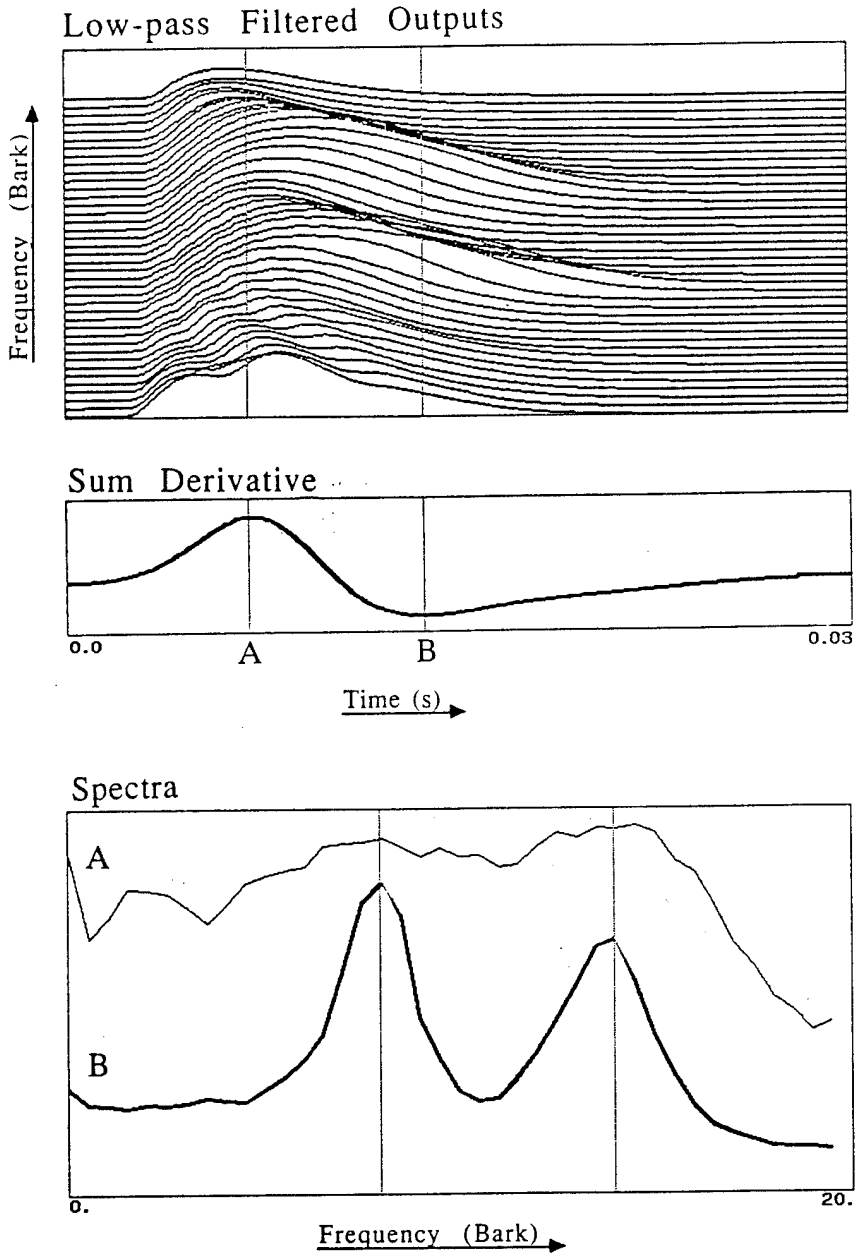


Figure 5.4: Determining event locations.

This figure illustrates the importance of sampling the filtered auditory outputs at the correct location. The example is taken from a synthetic stimuli made up of two decaying tones (resonances at 1 and 2 kHz, bandwidth of 70 Hz). If the outputs are sampled at the point of maximum onset in the channel sum waveform (A), the resulting spectral cross section does not adequately resolve the two resonance frequencies. If the outputs are sampled at a later time, such as at the point of maximum offset (B), then the spectral resolution is significantly better.

that it is conceptually straightforward; however, it has the inherent temporal resolution problem which was discussed previously. This section describes an alternative procedure which samples the second stage outputs of the auditory model synchronous to landmarks found in the channel sum waveform. Since there are many sounds, such as silence, frication, or aspiration, for which no salient events occur, the goal of the synchronous analysis is to sample the outputs at regular intervals (every 5 ms), focusing on important landmarks when they exist in the speech signal.

The pulse-synchronous procedure is illustrated in Figure 5.5. In this approach, the auditory outputs are again smoothed and sampled at a fixed analysis rate. However, within each analysis frame, the sampling is made at the location of the most significant event which has occurred in the channel sum waveform during the frame interval. In voiced regions, the event locations should correspond to onsets of glottal closure. Events should also be located at the release of bursts. In voiceless regions however, they occur somewhat randomly. As shown in Figure 5.5, since the sampling is made at events, significantly less smoothing of the original auditory outputs is required. This will produce a spectral representation with temporal properties superior to those found in the mean-rate response. Comparing Figures 2.3 and 5.5, it is clear that the pulse-synchronous response has sharper onsets and offsets than the mean-rate response, which should make it easier to locate significant acoustic events in the speech signal. Additionally, there is much less pitch ripple in the voiced regions of the utterance in the synchronous response. Further, by sampling at a consistent point in the speech signal, the spectral characteristics are more consistent than for the mean-rate response. The superior characteristics of the synchronous response are clearly illustrated in Figure 5.6 which shows the response to an utterance with a fundamental frequency as low as 25 Hz.

Since the pulse-synchronous representation of the speech signal appears to be potentially superior to the mean-rate response outputs, a more quantitative comparison was made. While a detailed investigation is beyond the scope of this thesis, some preliminary experiments have been made for the task of acoustic segmentation. As

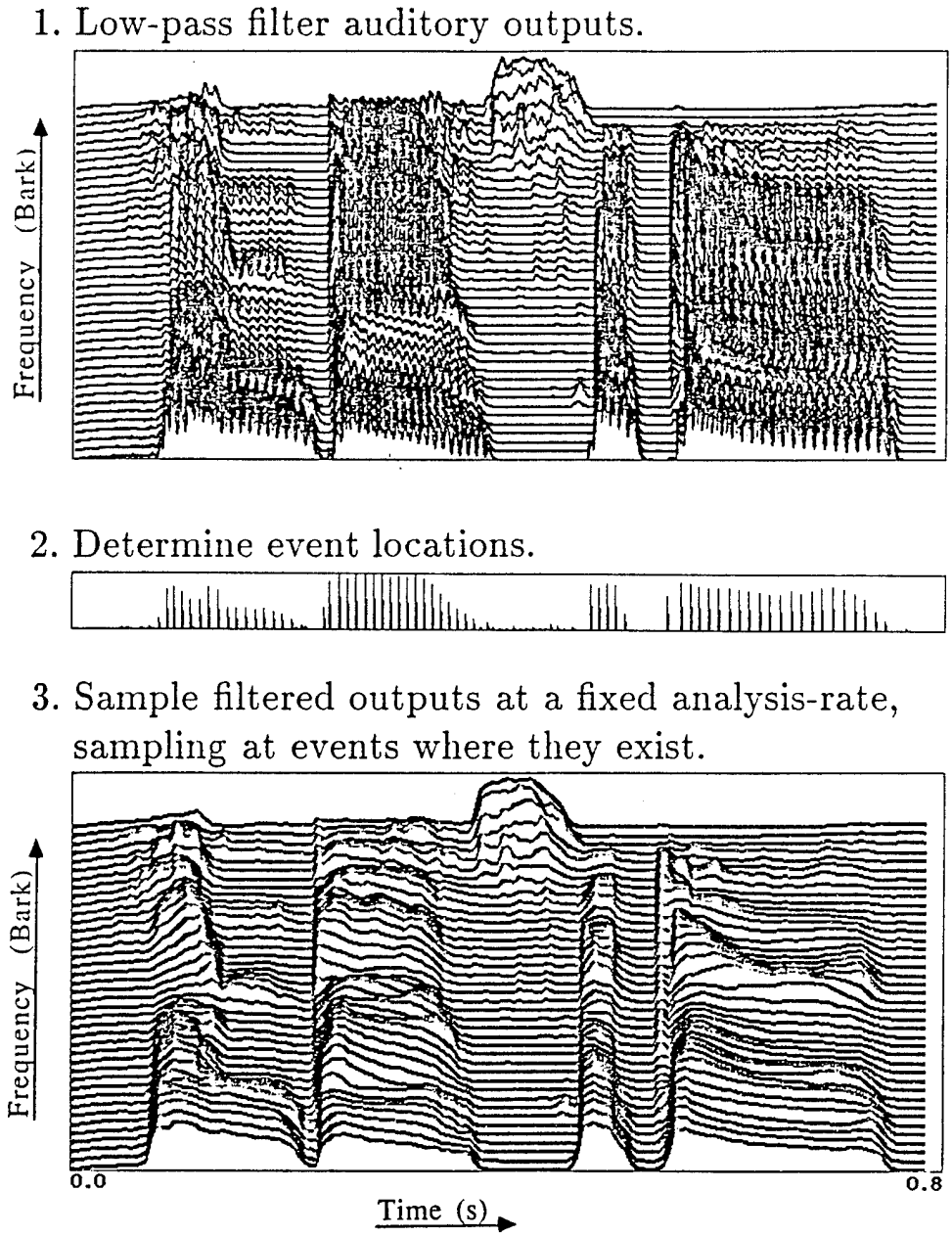


Figure 5.5: Computation involved in computing a synchronous response.

The three stages necessary in computing a pulse-synchronous response: (1) low-pass filter auditory outputs, (2) determine sampling points, and (3) sample outputs at a fixed analysis-rate, sampling at events wherever they exist.

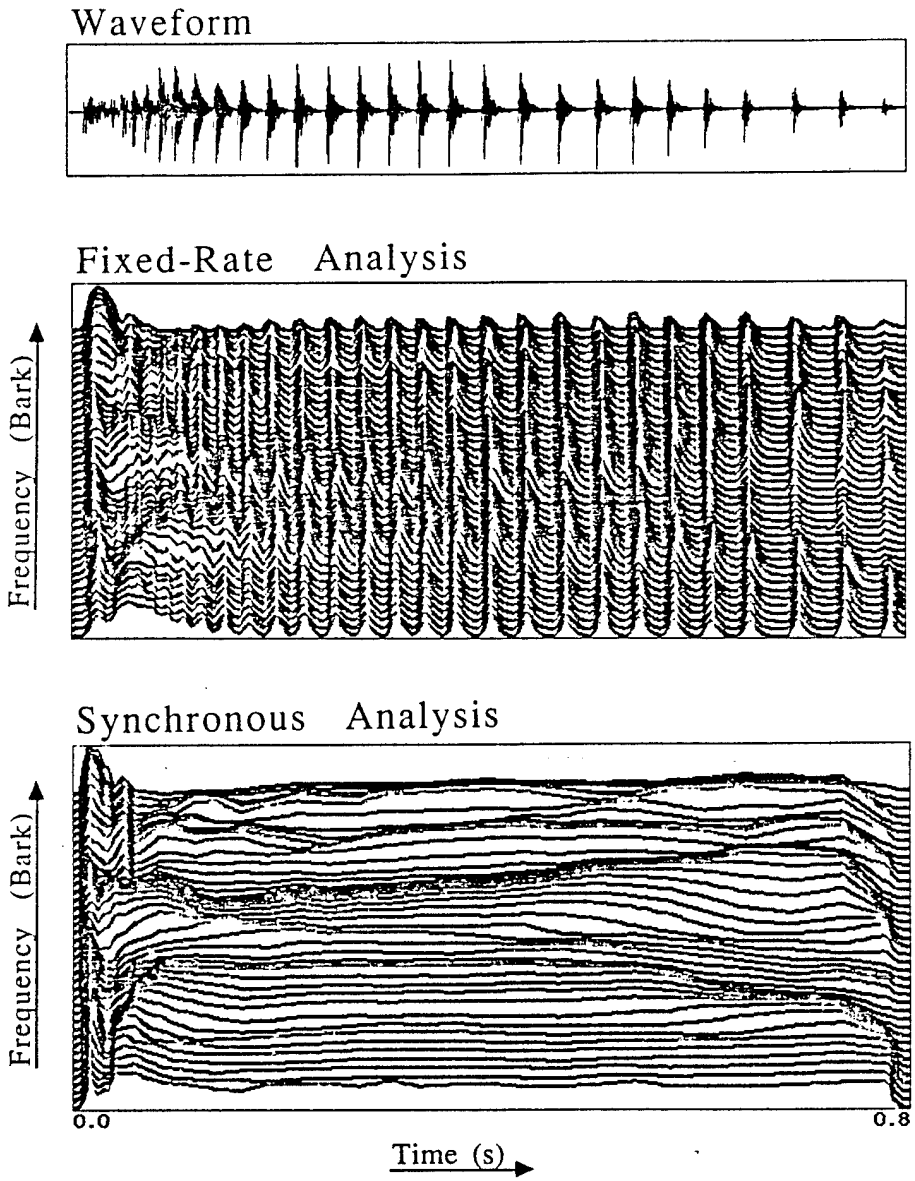


Figure 5.6: Comparison of mean-rate spectra for the word 'time.'

The top display contains the speech waveform, the middle display contains the mean-rate response, and the bottom display contains the pulse-synchronous response.

was reported in Chapter 3, an analysis of the dendrogram indicated that many times there was no acoustic event associated with the release of a stop consonant. Figure 5.7 shows the fraction of time that there was no event corresponding to a release of a stop consonant on the aligned dendrogram path. In many cases, the lack of an event was due to a lack of temporal resolution in the mean-rate response representation. As may be seen from the figure, substituting the pulse-synchronous response more than halved the number of cases where no release was found. This result provides evidence that the pulse-synchronous representation might be useful for the task of acoustic segmentation.

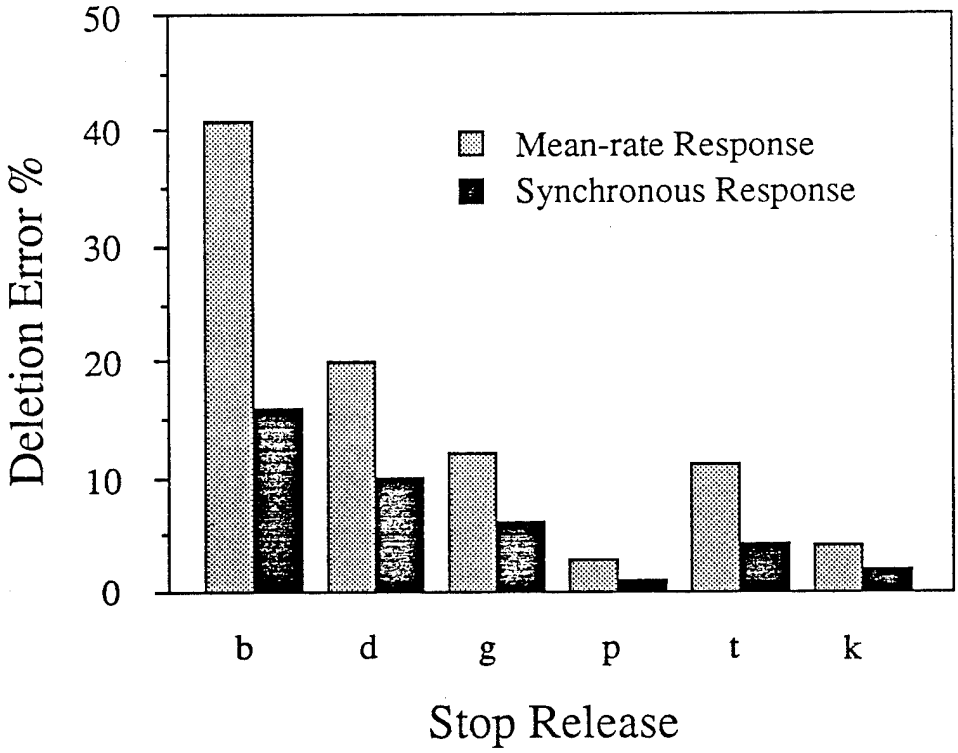


Figure 5.7: Stop release deletion rate.

This figure compares the dendrogram deletion error of the release of the six stop consonants using two different spectral representation. The first was the mean-rate response, while the second was the pulse-synchronous response.

5.3 Extracting the Fundamental Frequency

Since the channel sum waveform enhances the excitation characteristics of the speech signal relative to vocal-tract specific information, it would appear to be a useful representation for extracting the fundamental frequency of voicing. Such a measure of information would complement any spectral representation of the speech signal, and could prove useful for prosodic analyses of speech. Knowledge of the fundamental frequency would also be a valuable source of constraint in the synchronous spectral analysis procedure since it provides a mechanism for eliminating spurious events. The goal of this section is to illustrate how the fundamental frequency could be extracted from the channel sum waveform, and to discuss some interesting characteristics of the channel sum waveform.

Several different procedures have been developed for extracting information about the fundamental frequency [25,43,82,84,93,101,108]. Two of the more common time-domain representations are the autocorrelation function and average magnitude difference function (AMDF). Both of these functions compare a windowed portion of the waveform with another windowed portion delayed by a time period τ. The period of the fundamental frequency is determined by finding the τ with the maximum amount of similarity between the two waveform portions.

In the AMDF computation,

$$AMDF(\tau) = \Sigma \left| x(t) - x(t-\tau) \right|, \qquad T_1 \leq t \leq T_2$$

Periodic sounds will produce a null in the AMDF function at the appropriate delay as illustrated in Figure 5.8 for a pure tone. In practice, the fundamental period usually corresponds to the deepest null in the AMDF function.

Figure 5.9 is an example of the fundamental frequency determined from the channel sum waveform for the word 'ambassador.' In this case, the fundamental frequency was determined based on information provided by an AMDF function, although the

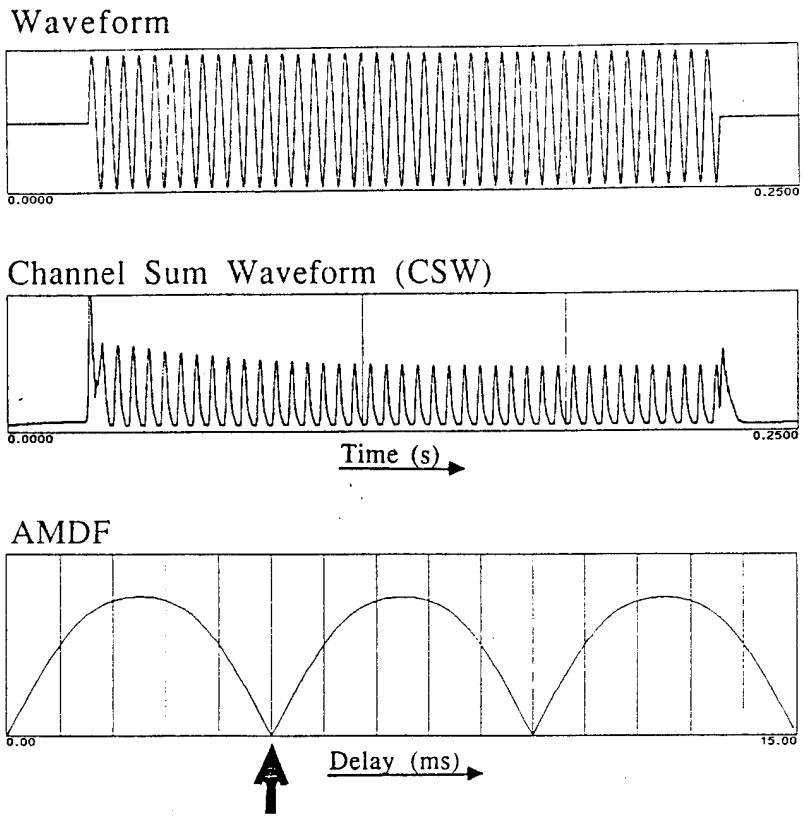


Figure 5.8: Determining periodicities with an AMDF.

This figure illustrates the use of the AMDF function for extracting the fundamental frequency. The top display contains the waveform corresponding to a 20 ms tone with a frequency of 200 Hz. The middle display shows the corresponding channel sum waveform. The bottom display shows the AMDF function computed between the two markers shown in the middle display. This function shows nulls at multiples of 5 ms.

channel sum waveform could be used as the input to any pitch-detection algorithm. It is interesting to note that the channel sum waveform appears to be quite robust in the presence of noise. Figure 5.9 shows how fundamental frequency information in the channel sum waveform is degraded in the presence of a 0 dB signal-to-noise ratio. The fundamental frequency measure appears to be quite robust, especially when compared to the Gold-Rabiner algorithm [43].

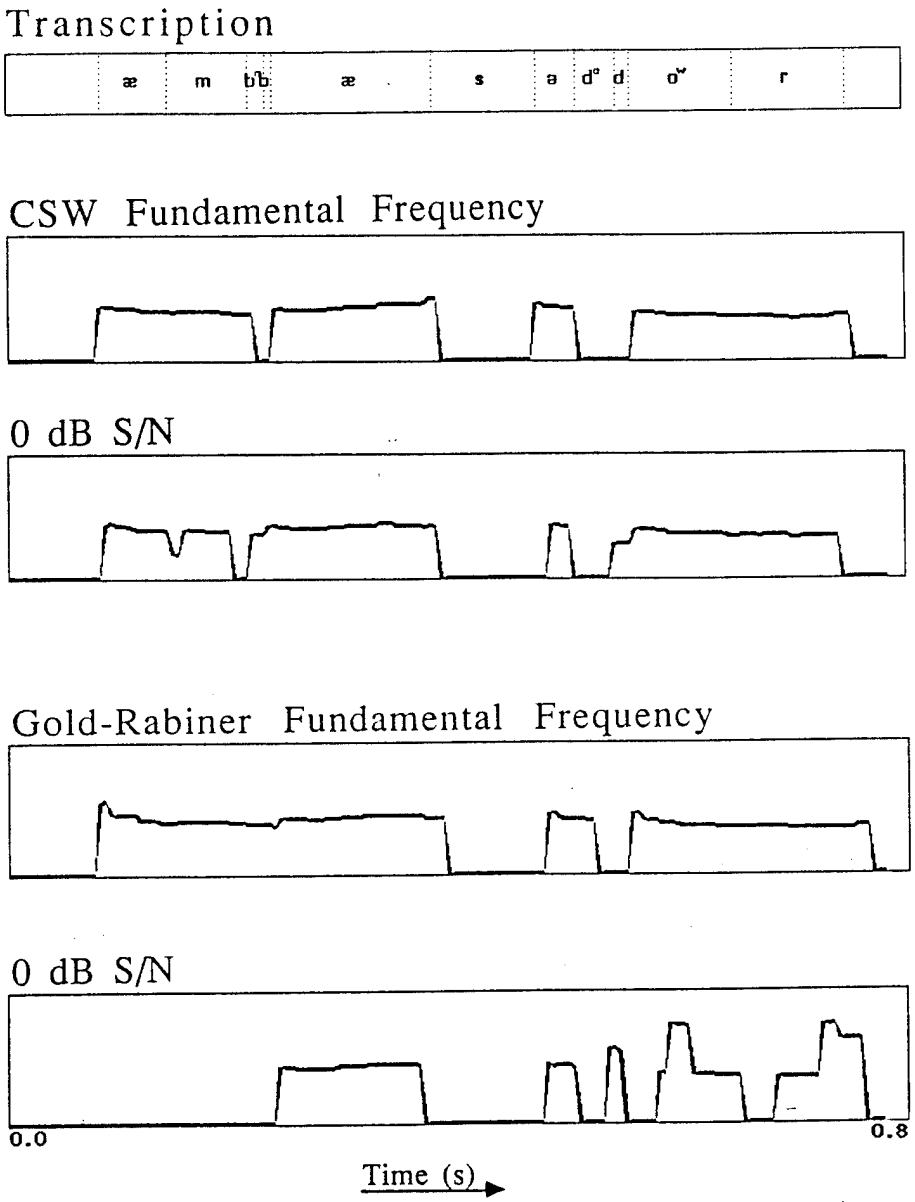


Figure 5.9: Fundamental frequency extraction in the presence of noise.

This figure illustrates the potential robustness of information in the channel sum waveform in the presence of noise for the word 'ambassador,' spoken by a male talker. From the top the displays are (1) the aligned phonetic transcription, (2) the fundamental frequency extracted from the channel sum waveform using an AMDF function, (3) the same procedure with 0 dB S/N, (4) the fundamental frequency extracted by the Gold-Rabiner procedure [43], and (5) with 0 dB S/N.

5.4 Dimensional Analysis of the Mean-rate Response

If the instantaneous outputs of the auditory model are considered to be a point in a multi-dimensional space, where the dimensionality is equal to the number of channels, then a time sequence of such points corresponds to a trajectory through this space. Previous studies have shown that the space traversed by speech spectra is highly correlated in both frequency and time, and that it is possible to reduce the the number of dimensions used to represent the spectra through the application of principal component analysis [74,90,121].

Principal component analysis is a statistical procedure which finds an efficient representation of a set of correlated data [55]. Geometrically, this procedure may be viewed as a rotation of the original coordinate system to one where the axes represent dimensions with maximum variability. This representation usually allows for a simpler description of the data covariance structure.

The principal components themselves are orthogonal linear combinations of the original dimensions computed by solving for the eigenvalues of the covariance matrix. The first m components are the eigenvectors corresponding to the first m eigenvalues. The components are optimal in the sense that a fixed number of dimensions will explain a maximum amount of the total variance in the original space, and will regenerate the original data with minimum least-square error [55].

Principal component analysis is relevant to speech analysis for several reasons. By reducing the number of dimensions needed to represent the speech signal, speech can be stored and transmitted more efficiently. This concept has been the subject of several investigations of speech coding [63,121]. If we assume that noise is equally distributed throughout the original subspace, then this kind of analysis procedure will also enhance the signal-to-noise ratio since the majority of the variance is captured in a fewer number of dimensions.

Another benefit of using a subset of the principal components, as was alluded to

earlier, is to get a better understanding of the underlying structure of the speech signal. On a study of Dutch vowels, Klein *et al.* found that there was a good match between the first few principal components and the perceptual space [89,90,92].

The most relevant aspect of the use of principal component analysis for this work concerns the issue of modeling the covariance structure of speech sounds. In the previous chapters, procedures were described for segmenting the speech signal, and for organizing these acoustic segments into acoustic classes. In all of these procedures, the concept of acoustic distance is extremely important. If the speech signal can be considered to traverse a space defined by the outputs of the auditory model, then it will be necessary to compute distances between points within this space. Any distance metric would benefit from incorporating knowledge about the covariance structure of such a space. Principal component analysis is important because it considerably simplifies the amount of effort needed to model this covariance structure. There are two reasons for this simplification. First, by reducing the number of dimensions required to represent the signal, fewer data are required to model the covariance structure. Second, because the eigenvectors are orthogonal, the new dimensions are much less correlated than the original dimensions. Thus, it might prove reasonable to ignore inter-dimensional correlations and assume that each dimension is statistically independent from all other dimensions. This is the assumption of any metric which uses a simple Euclidean, or weighted Euclidean, computation [113]. Clearly, the smaller the correlation between the different dimensions, the more accurate such assumptions should be.

The following sections analyze the characteristics of the outputs of the auditory model used for all speech analysis. Subsequently, a principal component analysis is performed in order to rotate as much of the information as possible onto a set of orthogonal coordinates.

5.4.1 Statistical Analysis

The investigation into the statistical properties of the mean-rate response was based on a study of 500 sentences of the TIMIT database, totalling approximately 25 minutes of speech. Figure 5.10 shows a summary of the amplitude distributions across all 40 channels. From this figure it is clear that the distributions are highly skewed towards a value of zero, irrespective of whether or not regions which correspond to silence are included in the statistics. This property is probably a result of the nonlinearities present in the transduction stage of the auditory model which attenuate low amplitude sounds in any channel. This effect is seen clearly in the lower frequency channels during the /s/ in Figure 2.3. Figure 5.10 also illustrates that data are distributed in a similar manner in the different channels.

In addition to individual channel statistics, general measures of correlation were also made by computing a 40x40 dimensional correlation matrix for each of the 100 speakers in the dataset. The distributions of the off-diagonal correlations from the 100 matrices are shown in Figure 5.11. The average amount of off-diagonal correlation among the channel outputs is 0.6, which indicated that many of the channels were highly correlated. A closer examination of the correlation distributions was performed by measuring the average correlation, C, between channels separated by a constant, k. Specifically,

$$C(k) = E(\rho_{ij}\delta(|i-j-k|)$$

where ρ_{ij} is the correlation coefficient between the ith and j^{th} channels and,

$$\delta(i) = \begin{cases} 1 & i = 0 \\ 0 & i \neq 0 \end{cases}$$

The average was computed using coefficients from all 100 matrices. A plot of C, shown in Figure 5.12, indicates that the auditory outputs are most correlated in adjacent channels. This result was also established by Li *et al.* using contour plots [74].

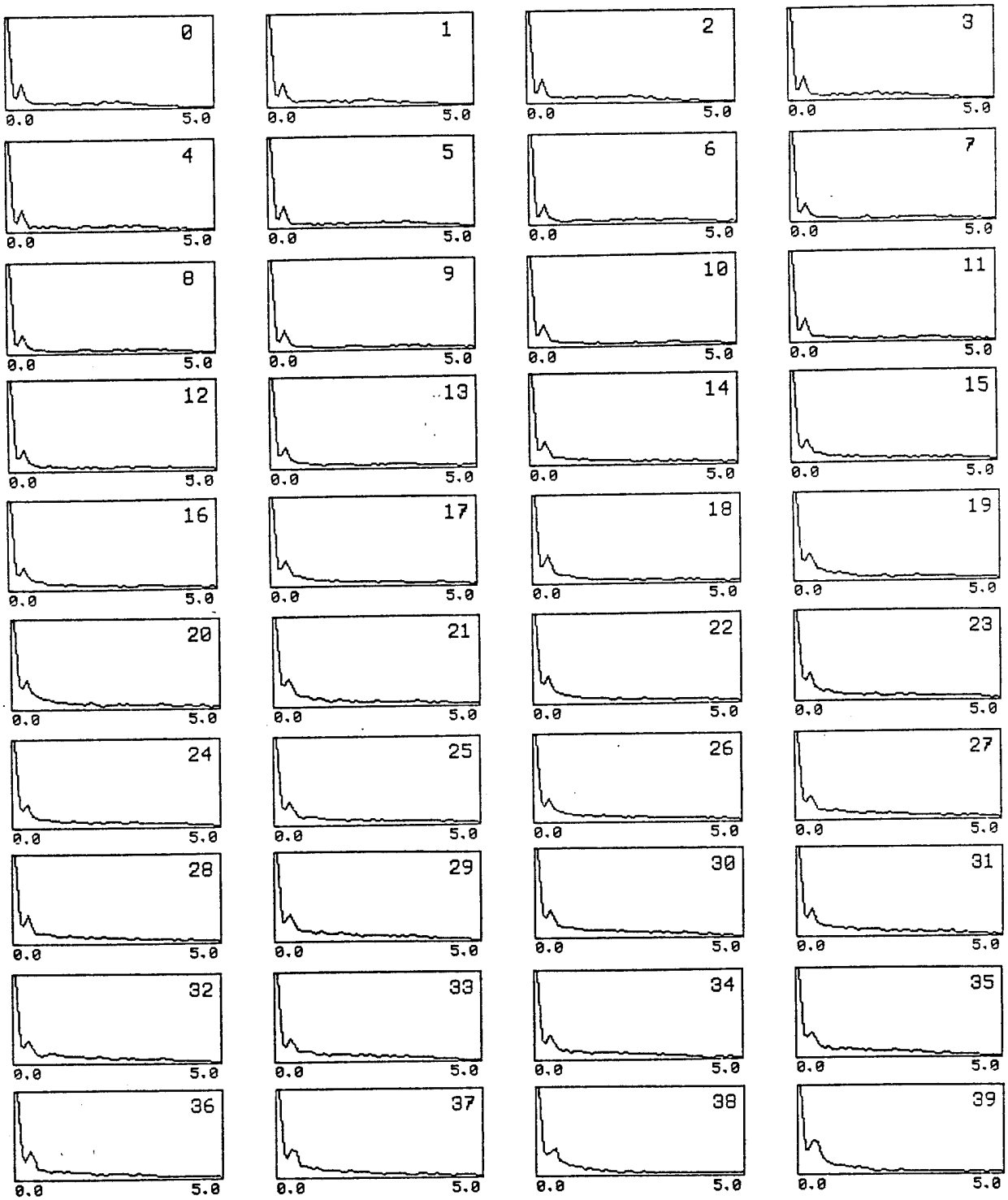


Figure 5.10: Distribution of speech in the auditory channels.

This figure shows the distribution of the mean-rate response outputs in the 40 channels of the auditory model. Data were collected from 500 sentences from the TIMIT database. Channel 0 corresponds to the lowest frequency channel.

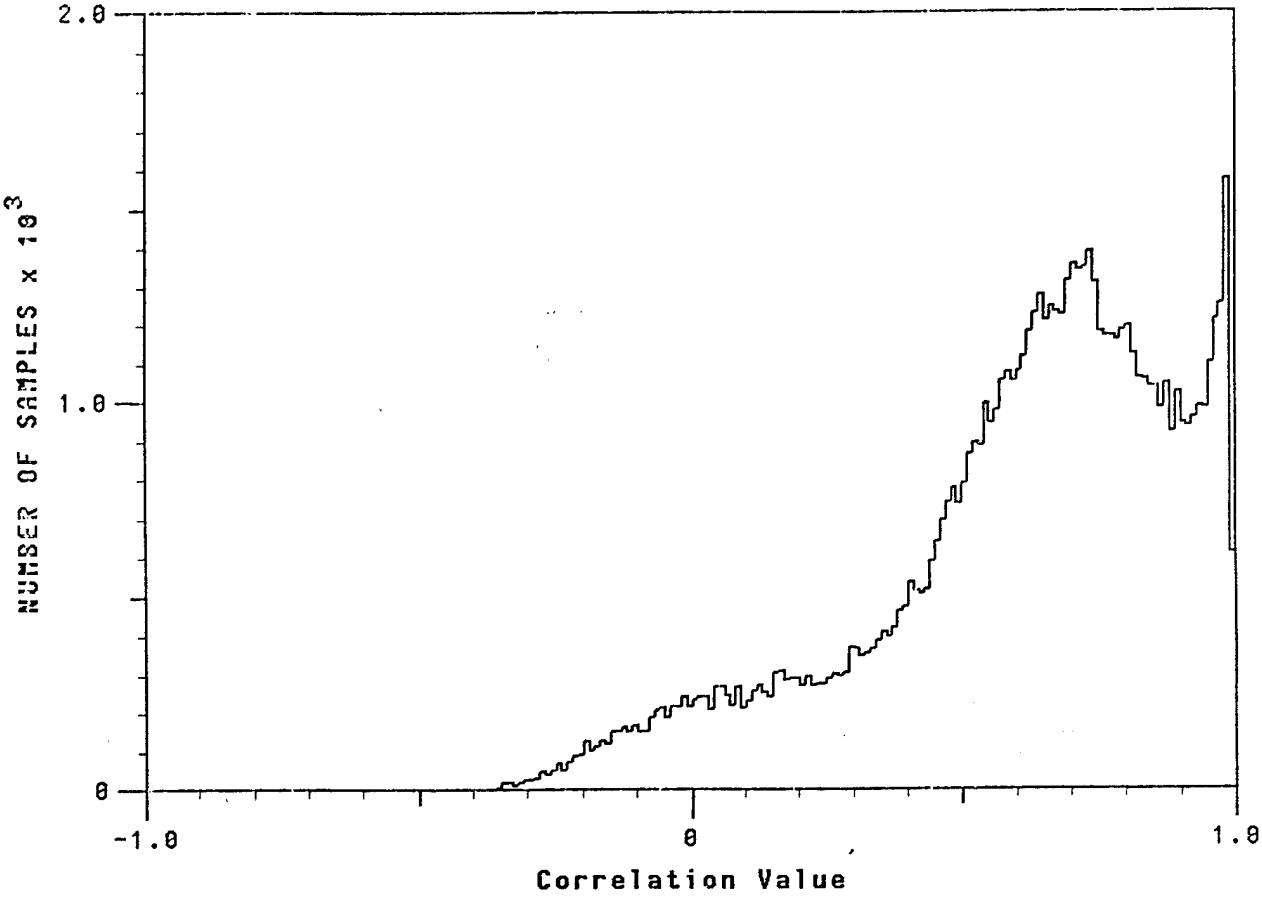


Figure 5.11: Histogram of the correlation among different channels.

This figure shows a histogram of the off-diagonal correlation coefficients from the 100 speaker correlation matrices.

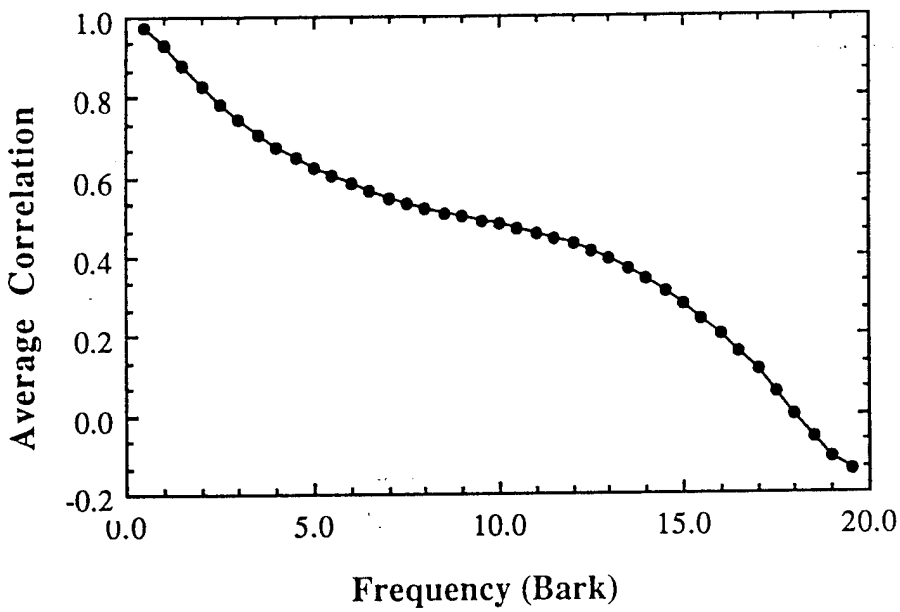


Figure 5.12: Average correlation versus channel separation.

This figure shows the average amount of correlation between channels plotted as a function of the frequency difference between channels.

5.4.2 Stability of Correlation Estimates

Since neighboring channels were highly correlated, on average, some form of data reduction was possible by applying principal component analysis. Before performing principal component analysis on the speaker covariance data, however, it was important to be aware of how stable the covariance estimates were, given the finite amount of speech data available for each speaker. In a previous study, Li *et al.* showed that a 35 dimensional correlation matrix stabilized after approximately 3000 frames of data, sampled every 10 ms. A similar approach to determine the stability was undertaken in this study. Using a speaker for whom a large quantity of speech was available, a correlation matrix was computed from approximately 100 seconds of speech which, when sampled every 5 ms, resulted in nearly 20,000 sample points. Using a different set of data from the same speaker, correlations were incrementally estimated by gradually increasing the amount of speech involved in each computation. A plot of the average magnitude difference between the off-diagonal correlation coefficients of the fixed correlation matrix and those of the incremental matrices is shown in Fig-

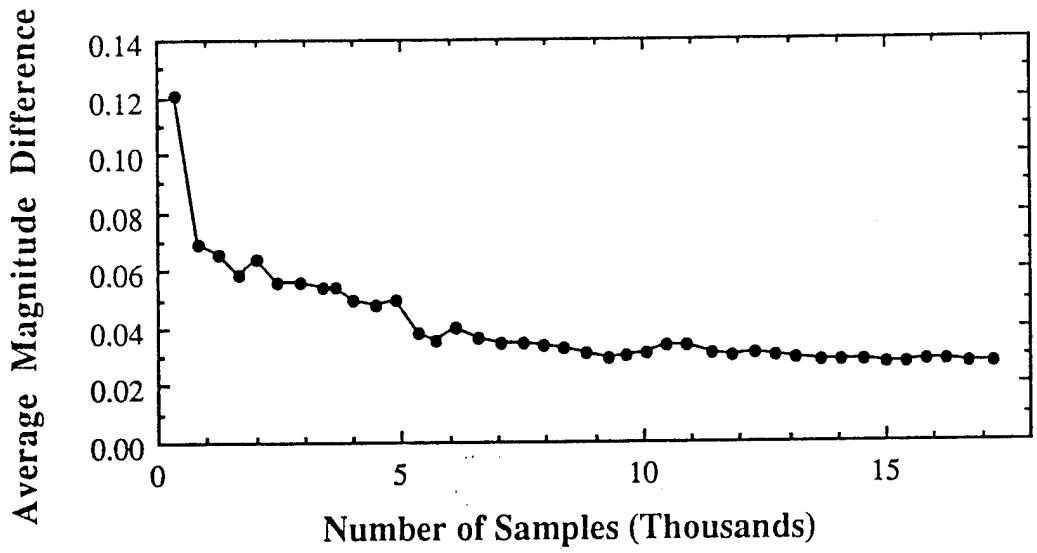


Figure 5.13: Stability of the correlation estimate.

This figure shows the average magnitude difference between the coefficients of a fixed correlation matrix, and those of a matrix computed from increasing amounts of data.

ure 5.13. From these data it would appear that after approximately 6000 frames the correlation stability has settled to near its minimum value. This result matches well with the result of Li *et al.* since the sampling rate used in this study was twice that in their study. Both results indicate that after approximately 30 seconds of speech, the correlation matrix stabilize. Additionally, it was clear that 15 seconds of speech, which was the average amount of speech available for each speaker in the TIMIT dataset, could produce a reasonable estimate of the amount of correlation among the auditory outputs.

5.4.3 Principal Component Analysis

The first investigation using principal component analysis studied individual speaker characteristics. A summary of the properties of the resulting components is shown in Figure 5.14, which plots the fraction of total variance explained as the number of dimensions increases. Averaged over 100 speakers, over 98% of the variance could be

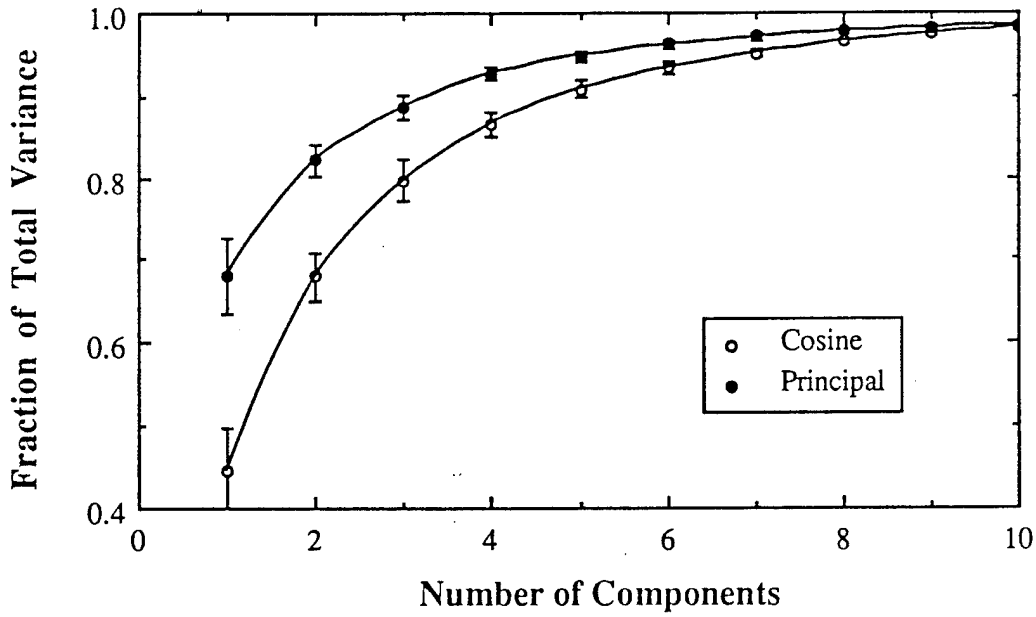


Figure 5.14: Fraction of total variance explained by components.

This figure illustrates the fraction of total variance explained by a set of components as the number of components increases. Each point represents the average fraction across all 100 speakers. The error bars represent one standard deviation. The top curve corresponds to the speaker-dependent principal components. The lower curve corresponds to half-cosine components.

explained by the first 10 dimensions.

The second analysis made with the principal component technique was to combine all data together to perform a single speaker-independent analysis of the total covariance structure. Interestingly, the resulting components explained almost as much of the total variance as did the speaker-dependent components. This would seem to indicate that the components were capturing only the general properties of speech, and were not overly sensitive to individual speaker characteristics. A plot of the first ten speaker-independent components is shown in Figure 5.15.

From this figure, it is apparent that the components look very sinusoidal. This observation has been made previously by others to suggest the use of a cosine weighting function [64]. As shown in Figure 5.14, the first 10 cosine components can also explain a significant fraction of the variance in the speech data. From this plot, it is

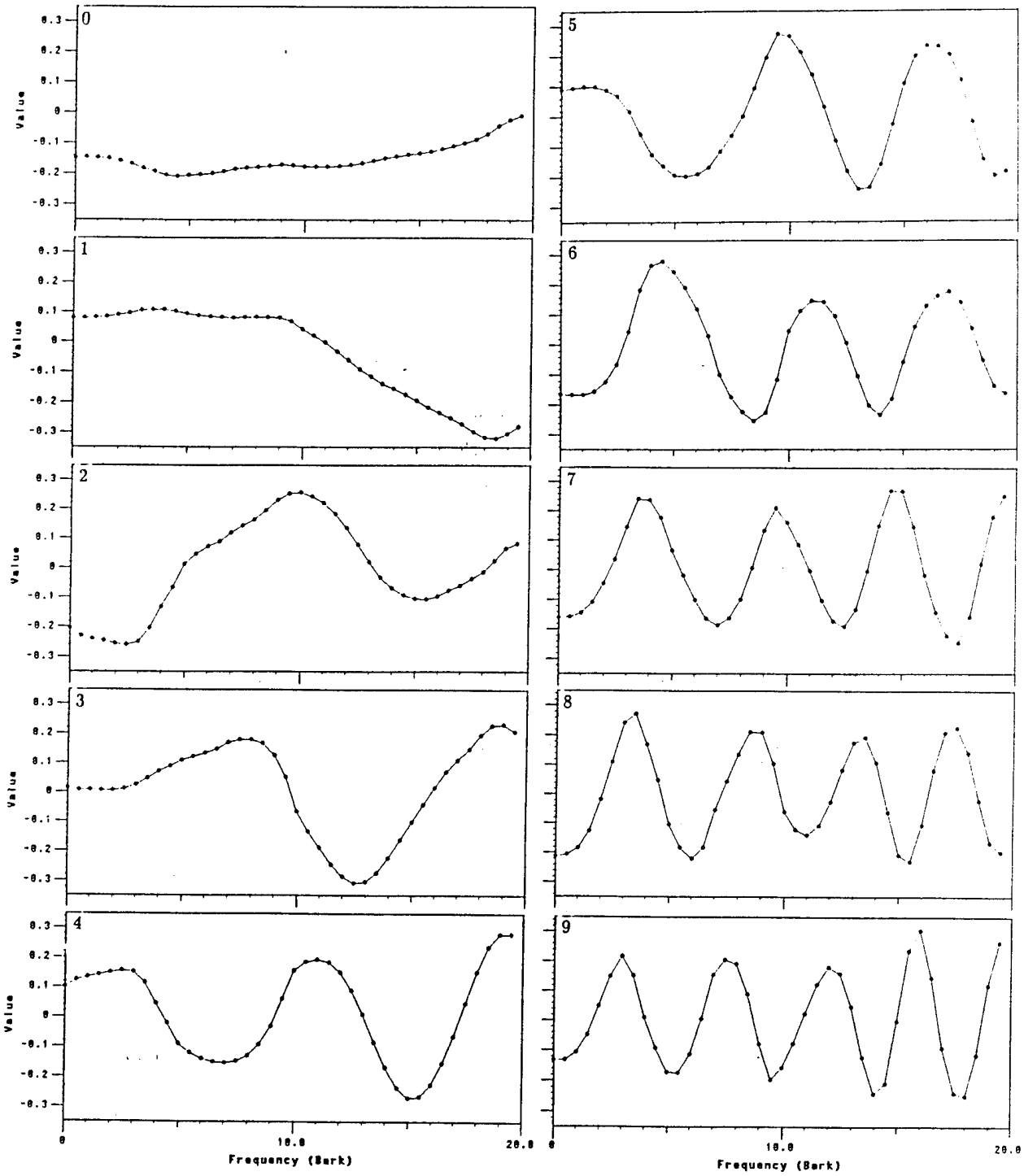


Figure 5.15: First ten speaker-independent components.

This figure presents the values of the first ten speaker-independent principal components.

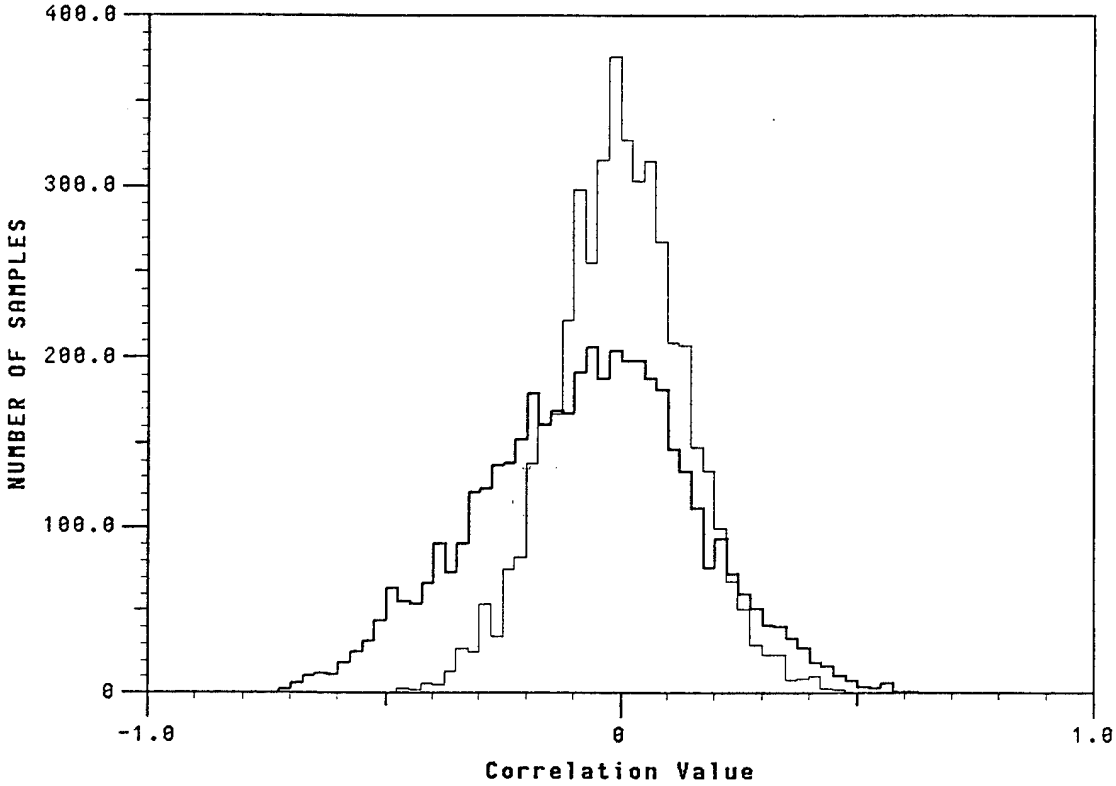


Figure 5.16: Off-diagonal Correlations

This figure illustrates the distribution of the off-diagonal coefficients from the 100 speaker correlation matrices when computed in the speaker-independent dimensions (thin line), and the half-cosine dimensions (thick line). This plot is computed the same way as Figure 5.11.

not clear if there is in fact any advantage to using speaker-independent components over the cosine components since both representations explain a large fraction of the total variance in the speech signal with a small number of dimensions. However, as illustrated in Figure 5.16, the off-diagonal correlations are consistently smaller for the speaker-independent components than for the cosine components. An examination of the correlation as a function of the off-diagonal also indicates that the average correlation of the principal component dimensions is virtually zero, while the average value of the first off-diagonal correlation in the cosine components is –0.159.

These analyses indicate that it is quite reasonable to reduce the dimensionality of the auditory outputs by a factor of four. In fact, it appears that there might be a considerable advantage in doing so because the resulting dimensions appear to be

largely uncorrelated. Thus, a diagonal distance metric would be better motivated in these dimensions than in the original representation. This observation perhaps explains why a diagonal distance metric in the cepstral domain is superior to diagonal metrics in the spectral domain since cepstra correspond to a cosine transformation which we have seen to be a highly uncorrelated set of dimensions [113].

5.5 Chapter Summary

The work in this chapter was motivated by previously reported studies of acoustic segmentation and classification. After evaluation of the segmentation algorithms, an attempt was made to develop a representation with temporal properties superior to those of the mean-rate response. As was discussed in this chapter, the approach which was taken attempted to sample the outputs of the second stage of the auditory model pulse-synchronously, rather than at fixed time intervals. The resulting representation appears to have improved temporal and spectral characteristics compared to the mean-rate response.

After studies on acoustic classification, it became apparent that the distance metric was important in determining the nature of the units. Investigations of the properties of the mean-rate response indicated that there was a significant amount of correlation among the channels. A principal component analysis of this data indicated that a set of speaker-independent components could explain a significant amount of the covariance structure of the data with approximately one quarter of the original dimensions.

Chapter 6

Summary and Future Work

In this thesis a framework for decoding the speech signal based upon an acoustic description is advocated. By developing a framework that provides a rigorous segmental acoustic description of the speech signal, it is hoped that researchers will be offered a useful mechanism for uncovering regularities in the acoustic realization of underlying phonological units in the speech signal. At the very least, such a mechanism should be able to verify hypotheses about sources of allophonic variation. At best, it provides a mechanism to help to discover additional sources of regularity. For these two reasons, an acoustic description is believed to be a valuable approach to speech analysis.

In Chapter 1 it was suggested that a segmental acoustic description of speech could form part of an overall strategy for relating the speech signal to lexical entries. In this approach, the acoustic signal would be transformed into a sequence of acoustic events, and the relationship between these entities and the underlying phonemic forms would be determined by searching for regularities in a large set of training data. In effect, this approach bypasses a standard phonetic level of description, searching for units of speech which are distinguishable by their acoustic characteristics.

In the following chapters, three fundamental issues concerning an acoustic description of speech were explored. First, the outputs of an auditory model were examined for use as inputs for a segmental description of speech. Second, a mechanism was developed for describing the speech signal as a sequence of segments, delineated by

landmarks. Finally, a procedure was developed for finding regular sources of behavior in the segmental description of speech. In the following sections, some of the issues raised by these explorations will be discussed.

6.1 Signal Representation

The representation used for all aspects of the work in this thesis was based on the mean-rate response outputs of an auditory model developed by Seneff [103]. The use of an auditory model as the foundation for a signal representation was motivated by the belief that it is important to incorporate the constraints provided by the human auditory system into the representation of the speech signal. This is because speech is a purely human form of communication and has evolved as a function of the constraints of both the production and the perception systems. Although this argument is reasonable, it is important to justify these claims with experimental evidence. In a study reported elsewhere, an experiment was conducted which examined the ability of different spectral representations to be used for acoustic segmentation [42]. The results of this study found that acoustic segmentation could consistently be performed more reliably using critical-band filter outputs or mean-rate response outputs than more standard DFT or LPC-based spectral representations. This type of experiment could be extended to include the pulse-synchronous representation described in Chapter 5. Similar comparisons should also be made for the task of acoustic classification, and should be extended to recognition tasks as well [5,16,49].

Another dimension in which the auditory representations could be evaluated is by their ability to handle noise. The results of several studies have suggested that the auditory representations are robust under the presence of noise [40,50]. In addition to examining the behavior of Seneff's auditory model in this respect, it would also be worthwhile to understand the response to other factors which contribute to variability in the speech signal, but which are irrelevant to phonetic distinctions [10,60].

An alternative way of examining the usefulness of auditory representations is

to perform resynthesis and intelligibility experiments [102]. Experimental evidence demonstrating that important information was being captured by these representations would strengthen the argument for their use.

One of the current limitations of the auditory model described here is that it cannot adapt to long term changes in the speech signal. This is because the filter gains are fixed. If the speech signal becomes too loud or too soft, the auditory outputs become saturated, or fall below the spontaneous firing rate. Currently, this problem is avoided by pre-recording the entire utterance, and then normalizing all values so that the maximum value is always the same. For speech, this typically corresponds to a low vowel, since these sounds tend to have the loudest output [27].

A related problem involves adapting to changes in spectral tilt, due to changes in the recording conditions, or environmental changes. As was mentioned previously, the TIMIT utterances used for this work were recorded with a noise-canceling microphone, which results in significantly less energy in low-frequency channels than would be present with different recording mechanism. As was pointed out in Chapter 4, this result made it difficult to distinguish between sounds such as nasal consonants and silence, since the low-frequency murmur was greatly attenuated. The severity of the problem was reduced by increasing the gain in the low-frequency channels. However, this issue points out the need for work in the area of automatic long-term adaptation.

6.2 Acoustic Segmentation

In Chapter 3, a procedure was developed which attempted to automatically locate important acoustic landmarks in the speech signal. Since it is difficult to capture all acoustic events with a single level of description, a procedure was developed which provided a multi-level description of the speech signal by performing a time-ordered hierarchical clustering. An analysis of the resulting structure indicated that the majority of acoustic landmarks were being located and organized by the dendrogram in a meaningful way.

The procedure for locating acoustic events was motivated by the concept of an acoustic segment being delineated by an onset and an offset in the speech signal. As was mentioned previously, this idea is difficult to implement in practice because the nature of important landmarks in the speech signal can be quite variable and can be extremely difficult to distinguish from variations in the speech signal which have no linguistic significance. As a means of reducing the difficulties associated with this problem, the notion of a multi-level description was introduced, and an algorithm was developed which attempted to produce such a structure. Although preliminary investigations appear encouraging, it is important to point out that the associations and dendrogram algorithms are implementations of a general idea and are not the most important issues themselves.

Given that the algorithms reported here do appear to organize information in a meaningful fashion, there are many ways in which further work may be pursued. The construction of the dendrogram is extremely simple, since each acoustic segment is represented by an average spectral vector. It is possible that a more sophisticated representation, such as one which included time-varying information, might improve the dendrogram structure in cases where changes were quite gradual, as in vowel-semivowel sequences. In addition, the distance metric itself is quite simplistic. It would therefore be worthwhile to explore the use of other metrics, such as those with a stronger perceptual motivation [3,46,60,107]. Finally, it would be worthwhile to investigate the behavior of the dendrogram structure under degraded conditions.

6.3 Acoustic Classification

In Chapter 4, a procedure was developed for automatically clustering a set of sound segments into groups based on acoustic criteria. This procedure was applied to two different areas; one which investigated the coarse acoustic classes of all speech sounds, another which examined the context-dependent realizations of a few individual phonemes. These studies attempted to demonstrate two necessary requirements

of any approach advocating a set of acoustic units to describe the speech signal: (1) that it is possible to uncover major sources of regularity that capture basic properties of all phonemes and, (2) that it is possible to capture important context-dependencies of individual phonemes.

Throughout these experiments, the representation of the data, and the distance metric used to measure the similarity between data, were quite simple. In the first investigation for instance, a Euclidean distance metric was used to measure the distance between acoustic segments. As was pointed out previously, the mean value of a spectral vector was often detrimental in determining the closest acoustic cluster. When the zero-mean distance metric was substituted for the Euclidean distance metric slightly improved results were obtained. However, the zero-mean distance metric is only a single example of an alternative. It would of great interest to examine metrics incorporating the principal components since these elements appear to capture more of the covariance structure amongst the different spectral channels. Alternatively, other metrics, such as those based on more perceptually important factors could be explored [3,60].

In addition to exploring alternative distance metrics, it is also possible to make the representation of the signal more sophisticated. For instance, observations could be made at specific locations in the speech signal, or dynamic measures could be made over the course of the segment. Ultimately it will be necessary to describe the realization of a phone in terms of its dynamic, as well as its static characteristics.

The results of the studies of individual phonemes indicated that it was possible to automatically capture important context-dependencies based on acoustic information. There was also evidence suggesting that these regularities can generalize across sets of phonemes with similar phonetic features. In the study of the velar consonants for instance, it was found that the top three acoustic clusters look strikingly similar. Since it will be ultimately necessary to distinguish between these sounds, it will be necessary to incorporate other forms of representation. In the case of the velar consonants for instance, the duration of the voice onset time is a well known factor for distinguishing

between these two consonants. Figure 6.1 illustrates the differences between the voice onset time (VOT) of the velar stops in a set of 500 TIMIT utterances. Note that the VOT of /g/ is rarely longer than 50 ms. The overlap between these two distributions is partially caused by mixing data from both stressed and unstressed environments.

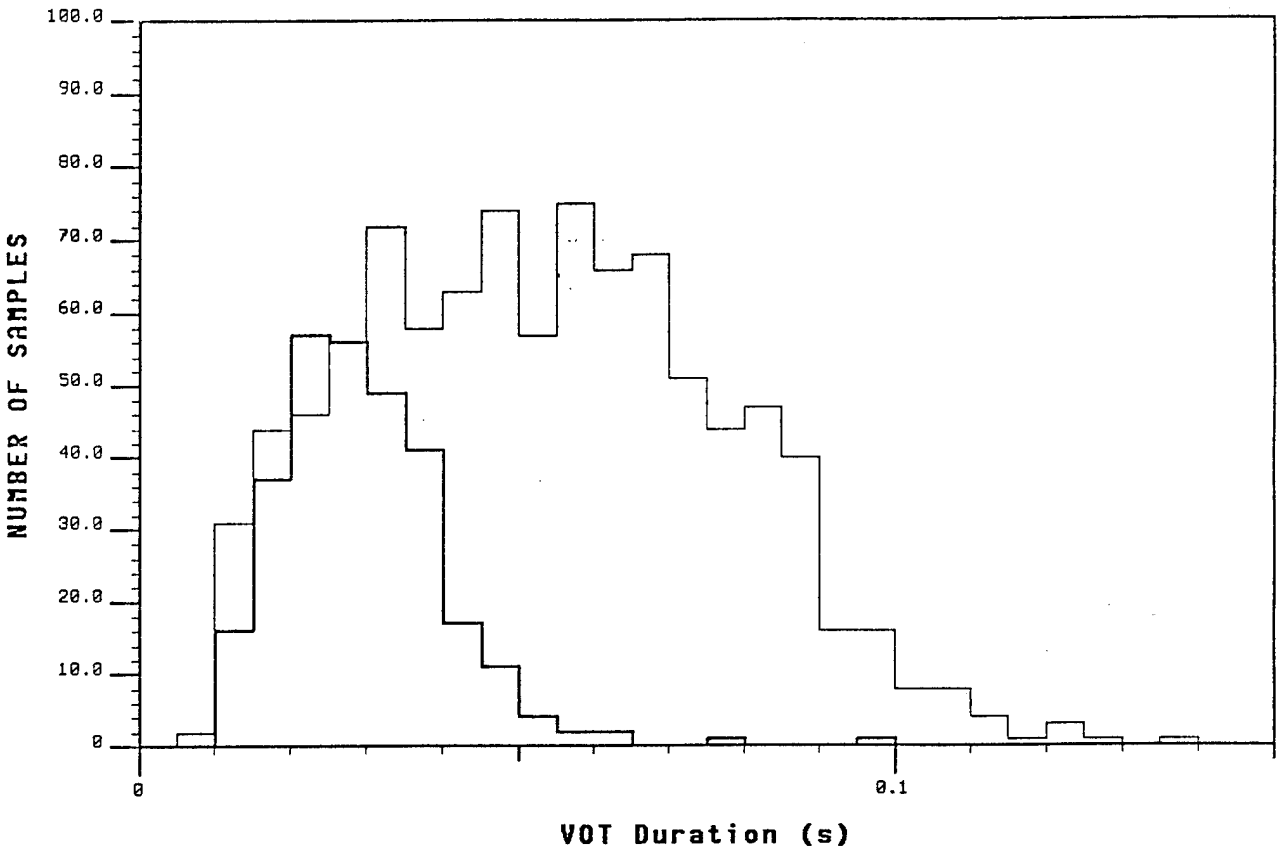


Figure 6.1: Voice onset time of velar stop consonants.

This figure compares the voice onset times of /g/, shown in dark lines, versus /k/. Durations are in seconds.

6.4 Applications to Machine Recognition of Speech

In this final section, some of the more practical implications of a segmental description of speech are discussed by comparing this approach to those which do not embrace the concept of a segment, but instead process the speech signal on a frame-

by-frame basis [94,98] . For the purposes of the remainder of this discussion it will be assumed that some intermediate level of representation between the speech signal and the lexicon is hypothesized. The desirability of such a level has been discussed in Chapter 1.

6.4.1 Motivation for Segmentation

The first observation to be made about an acoustic segment is that it is more a natural unit of speech than a frame. A unit such as a frame is usually motivated by the minimum resolution needed to resolve short-time events, and has no general motivation for all speech sounds. While fixed-rate analysis procedures are reasonable representations of the speech signal, it is less reasonable to restrict the window of observation to a frame. By doing so, it becomes much more difficult to incorporate acoustic-phonetic information into the decoding process. A frame is only a single observation of the speech signal which forms part of a larger trajectory. A point-by-point examination of this trajectory limits the ways in which the signal may be analyzed. A segmental level of description of speech is a more powerful framework for applying acoustic-phonetic information since it provides a broader perspective of the local environment. Essentially, a frame-based analysis is but one of the many alternative strategies available to a segment based analysis.

One of the kinds of acoustic-phonetic information that can be easily incorporated into a segment based framework is information about the inherent duration of sounds. For example, duration is well known to be a significant factor in determining voicing in many consonants, and is important in contrasting between tense and lax vowels such as /æ/ and /ɛ/ [59]. It is more difficult to gracefully incorporate durational information into a frame-based analysis of speech since the duration of a segment is not known until the next segment has begun [72].

In addition to providing a graceful framework for incorporating explicit durational information, a segmental framework can easily avoid the implicit duration weighting

which is inherent in many frame-based systems [34,52]. Such weighting emphasizes the importance of longer segments relative to shorter segments. Since this particular weighting is likely to be undesirable, it is important that a framework be flexible enough to easily incorporate different weighting mechanisms.

Another advantage of a segmental framework is that it is easier to incorporate relational temporal measurements than would be possible in a frame-based approach. In addition, a segmental framework easily allows a focus of attention on particular points in the speech signal, rather than uniformly weighting each observation point.

In summary, a frame-based approach is a subset of a segmental description of speech, and has less flexibility to capture certain kinds of acoustic-phonetic information. In the following sections the various approaches to a segmental framework will be examined.

6.4.2 Explicit vs Implicit Segmentation

To date, the most common form of segmental description of speech has been one where landmarks are delineated explicitly in the speech signal. This is typically followed by acoustic-phonetic analysis of the resulting segments. There is a substantial amount of literature testifying to the difficulty of the task of segmentation [44,52]. Typically, a segmentation of the signal will either delete important acoustic-phonetic landmarks, or will insert irrelevant landmarks. Thus far, this type of single level of description has not proved to be very effective.

An alternative to an explicit segmentation of the speech signal is an implicit, or stochastic segmentation which considers all possible alternatives. The obvious advantage of this kind of approach is that it is not prone to deleting or inserting segments as are explicit segmenters. The obvious disadvantage of these approaches are their enormous search space, with the associated computation costs. As discussed in Appendix B, an approach which considers all possible ways to segment n frames

into subsegments will have a total of 2^{n-1} possible segmentations. For example, a 2 second utterance analyzed with a frame-rate of 10 ms would have well over 10^{60} possible segmentations. Search strategies which perform a best first search with dynamic programming would reduce the search space to slightly more than an order of magnitude more than would be involved with a frame-based analysis.

The multi-level framework developed in this thesis appears to be a compromise between a single level of description and a fully stochastic model. Many alternative segmentations are considered, but the number of alternatives are substantially reduced by paying attention to landmarks found in the speech signal. Shrinking the size of the search space not only reduces the amount of computation, it also reduces the number of opportunities to make a mistake. A stochastic approach considers all possible segments everywhere, which is effect a 'blind' search strategy.

Another advantage of the multi-level approach is that the resulting structure is explicit, and may be displayed in the form of a dendrogram. One of the disadvantages of a fully stochastic segmentation is that it is difficult to understand the competing alternatives, since the segmental description is implicit. Although the dendrogram will have 'errors,' the number of errors will be substantially smaller than is possible with a single level of description. In Chapter 3 it was shown that these errors are quite systematic, which suggests that they could be modeled.

6.4.3 A Probabilistic Framework

In all analyses of the dendrogram presented in this thesis, a path through the dendrogram was always located with a knowledge of the phonetic transcription. In speech recognition applications, it will be necessary to specify a path through the dendrogram that is the most probable, given the acoustic evidence. This could be achieved by assigning a likelihood to each landmark in the dendrogram. For example, distributions such as the one illustrated in Figure 3.9, which plots the heights of valid boundaries and the heights of invalid ones, can be used to derive a likelihood

measure. By assigning a probability to each boundary, it is subsequently possible to turn the dendrogram into a probabilistic network. An example of such a network is shown in Figure 6.2. In the figure, the most probable path is the single path of sequential segments. Less likely segments with fewer boundaries are drawn on top of the most probable path, less likely segments with more boundaries are drawn below the most probable path. This approach is simplistic and should become more tightly coupled with the labeling process. Note, however, that this framework is not sensitive to duration. Thus, short but acoustically distinct segments can be considered quite probable. This is illustrated for the [k], and [ɪ] in the second syllable of the work 'coconut.' In addition, this mechanism normalizes all paths, so that two paths which span the same interval of time have both accounted for all intervening boundaries. Thus, it is possible to prune the search space if desired.

In summary, there are several reasons why a segmental description could prove to be a useful representation for automatically decoding the speech signal. A multi-level structure such as the dendrogram can potentially make use of many of the advantages of an explicit description of speech, as well as those provided by purely stochastic approaches which consider many alternative segmentations of the speech signal. For this reason, the application of these structures to the task of speech recognition is a tantalizing area of further investigation.

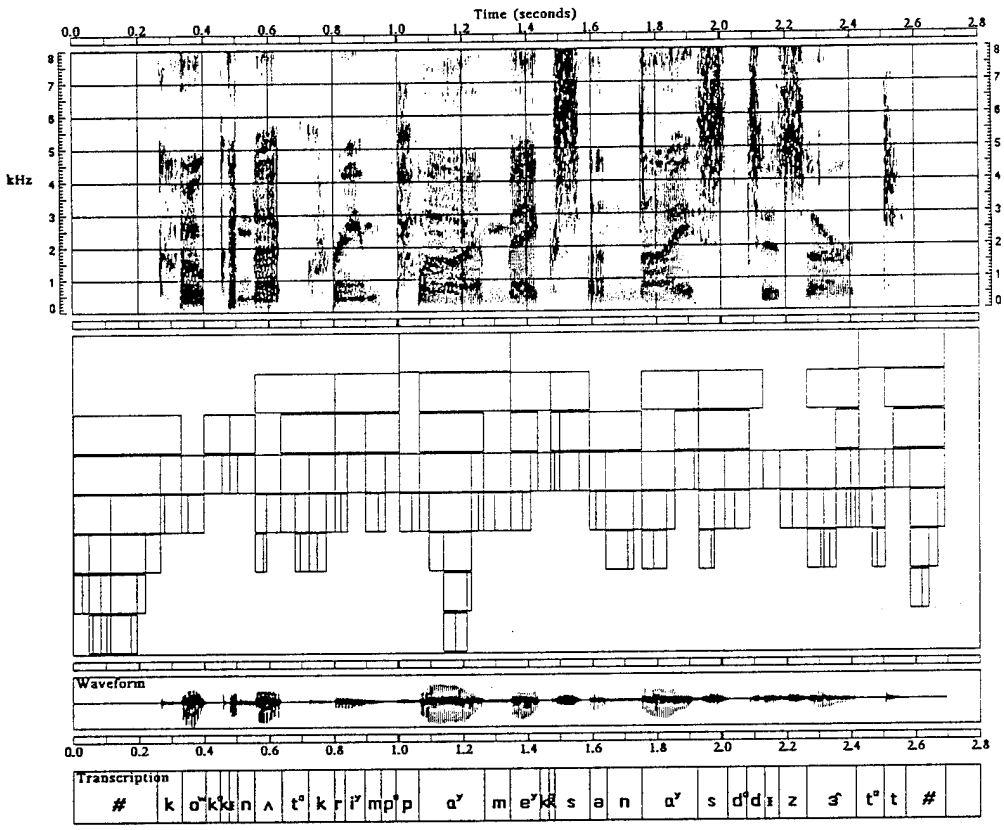


Figure 6.2: A probabilistic network.

This figure illustrates an example of an acoustic network for the utterance 'Coconut cream pie makes a nice dessert,' spoken by a female talker. Below the spectrogram is an acoustic network which is a probabilistic version of the dendrogram structure shown in Figure 3.4. The most probable path is single path of sequential segments. Less likely segments with fewer boundaries are drawn on top of the most probable path, while less likely alternatives with more boundaries are drawn below the path.

Bibliography

[1] J.B. Allen, "Cochlear Modeling," *IEEE ASSP Magazine,* vol. 2, no. 1, pp. 3-29, 1985.

[2] R. Andre-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. ASSP,* vol. 36, no. 1, pp. 29-40, 1988.

[3] C. Aoki, "Comparison of selected distance metrics on several phonetic distinctions," S.M Thesis, MIT, 1985.

[4] J.Babaud, A.P. Witkin, M. Baudin, and R.O. Duda, "Uniqueness of the Gaussian kernel for scale-space filtering," *IEEE Trans. on Pattern Anal. Mach. Intell.,* vol. 8, pp. 26-33, 1986.

[5] M. Blomberg, R. Carlson, K. Elenius, and B. Granström, "Auditory models as front ends in speech recognition systems," in *Invariance and Variability in Speech Processes,* J.S. Perkell and D.H. Klatt, eds., Lawrence Erlbaum Ass., New Jersey, 1986.

[6] S.E. Blumstein and K.N. Stevens, "Phonetic features and acoustic invariance in speech," *Cognition,* vol. 10, pp. 25-32, 1981.

[7] Z.S. Bond and S. Garnes, "Misperceptions of fluent speech," In *Perception and production of fluent speech,* R.A. Cole, ed., Erlbaum, Hillsdale, N.J., 1980.

[8] J.S. Bridle and N.C. Sedgwick, "A method of segmenting acoustic patterns, with applications to automatic speech recognition," *Proc. ICASSP,* pp. 656-659, 1977.

[9] M.A. Bush and G.E. Kopec, "Network-based connected digit recognition," *IEEE Trans. ASSP,* vol. 35, no. 10, pp. 1401-1413, 1987.

[10] R. Carlson, B. Grandström, and D.H. Klatt, "Vowel perception: The relative perceptual salience of selected acoustic manipulations," *STL-QPSR,* vol. 3-4, pp. 73-83, 1979.

[11] J.C. Catford, *Fundamental Problems in Phonetics,* University Press, Indiana, 1977.

[12] N. Chomsky and M. Halle, *The Sound Patterns of English,* Harper and Row, New York, 1968.

[13] Y.L. Chow, R.M. Schwartz, S. Roucos, O.A. Kimball, P.J. Price, G.F. Kubala, M.O. Dunham, M.A. Krasner, and J. Makhoul, "The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system," *Proc. ICASSP,* pp. 1593-1596, 1986.

[14] Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, P.J. Price, S. Roucos, and R.M. Schwartz, "BYBLOS: The BBN continuous speech recognition system," *Proc. ICASSP,* pp. 89-92, 1987.

[15] K.W. Church, "Phonological parsing and lexical retrieval," *Cognition,* vol. 25, pp. 53-70, 1987.

[16] J.R. Cohen, "Application of an adaptive auditory model to speech recognition," *Proc. Symp. on speech recognition,* Montreal, pp. 8-9, July, 1986.

[17] R. Cohen, G. Baldwin, J. Bernstein, H. Murveit, and M. Weintraub, "Studies for an adaptive recognition lexicon," *Proc. DARPA Speech Recognition Workshop,* Report no. SAIC-87/1644, 1987.

[18] R.A. Cole, A.I. Rudnicky, V.W. Zue, D.R. Reddy, "Speech as patterns on paper," in *Perception and Production of Fluent Speech,* R.A. Cole, ed., Lawrence Erlbaum Associates, Hillsdale, N.J., 1980.

[19] R.A. Cole and J. Jakimik, "A model of speech perception," in *Perception and Production of Fluent Speech,* R.A. Cole, ed., Lawrence Erlbaum Associates, Hillsdale, N.J., 1980.

[20] T.H. Crystal and A.S. House, "Characterization and modelling of speech-segment durations," *Proc. ICASSP,* pp. 2791-2794, 1986.

[21] P.B. Denes, "On the statistics of spoken English," *J. Acoust. Soc. Amer.,* vol. 35, no. 6, pp. 892-904.

[22] P.B. Denes and E.N. Pinson, *The Speech Chain,* Anchor Books, Garden City, N.J., 1972.

[23] G.R. Doddington and T.B. Schalk, "Speech recognition: Turning theory to practice," *IEEE Spectrum,* vol. 18, no. 9, pp. 26-32, 1981.

[24] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis,* John Wiley and Sons, New York, 1973.

[25] J.J. Dubnowski, R.W. Schafer, and L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust. Speech and Signal Proc.,* vol. 24, pp. 2-8, 1976.

[26] C.Y. Espy-Wilson, "An acoustic-phonetic approach to speech recognition: application to the semivowels," Ph.D. Thesis, MIT, 1987.

[27] G. Fairbanks, A.S. House, and A.L. Stevens, "An experimental study of vowel intensities," *J. Acoust. Soc. Amer.,* vol. 22, no. 4, pp. 457-459, 1950.

[28] G. Fant, *Acoustic Theory of Speech Production*, Mouton and Co., 's-Gravenhage, Netherlands, 1960.

[29] G. Fant, "Studies of minimal speech sound units," *STL-QPSR,* vol. 2, pp. 1-11,1961.

[30] G. Fant, "Descriptive analysis of the acoustic aspects of speech," *Logos*, vol. 5, no. 1, pp. 3-17, 1962.

[31] G. Fant, "Auditory patterns of speech," *STL-QPSR,* vol. 3, pp. 16-20, 1964.

[32] G. Fant, "The nature of distinctive features," *STL-QPSR,* vol. 4, pp. 1-15, 1966.

[33] G. Fant, "Distinctive features and phonetic dimensions," *STL-QPR,* vol. 2-3, pp. 1-18, 1969.

[34] G. Fant, "Automatic recognition and speech research," *STL-QPSR,* vol. 1, pp. 16-31, 1970.

[35] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," *Proc. DARPA Speech Recognition Workshop,* Report no. SAIC-86/1546, 1986.

[36] J.L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972.

[37] J.L. Flanagan et al., "Automatic speech recognition in severe environments," National Research Council report, National Acadamy Press, Washington, 1984.

[38] G.D. Forney Jr., "The Viterbi algorithm," *Proc. IEEE,* vol. 61, pp. 268-278, 1978.

[39] V. Fromkin, *Speech Errors as Linguistic Evidence,* The Hague, Mouton, 1973.

[40] O. Ghitza, "Robustness against noise: The role of timing-synchrony measurement," *Proc. ICASSP,* pp. 2372-2375, 1987.

[41] J.R. Glass and V.W. Zue, "Recognition of nasal consonants in American English," *Proc. DARPA Speech Recognition Workshop,* Report no. SAIC-86/1546, pp. 25-29, 1986.

[42] J.R. Glass and V.W. Zue, "Signal representation for acoustic segmentation," *Proc.* 1st *Australian Conf. on Speech Science and Tech.,* pp. 124-129, 1986.

[43] B. Gold and L.R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.,* vol. 46, pp. 442-448, 1962.

[44] H.R. Goldberg, D.R. Reddy, and R. Suslick, "Parameter independent machine segmentation and labeling," *IEEE Symp. on speech recognition,* Pittsburgh, pp. 106-111, 1974.

[45] R. Goldhor, "Representation of consonants in the peripheral auditory system: A modeling study of the correspondence between response properties and phonetic features," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1985.

[46] A.H. Gray Jr., J.D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust. Speech Sig. Proc.,* vol. 24, no. 5, pp. 380-391, 1976.

[47] Hess, W.J., "A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech", *IEEE Transactions ASSP*, Vol. 24, pp 14-25, 1976.

[48] C.F. Hockett, *Manual of Phonology,* Indiana University Publications in Anthropology and Linguistics, no. 1, Bloomington, 1955.

[49] M.J. Hunt and C. Lefèbvre, "Speech recognition using a cochlear model," *Proc. ICASSP*, pp. 1979-1982, 1986.

[50] M.J. Hunt and C. Lefèbvre, "Speech recognition using an auditory model with pitch-synchronous analysis," *Proc. ICASSP,* pp. 813-816, 1987.

[51] D.P. Huttenlocher and V.W. Zue, "A model of lexical access from partial phonetic information," *Proc. ICASSP,* pp. 26.4.1-26.4.4, 1984.

[52] S.R. Hyde, "Automatic speech recognition: A critical survey and discussion of the literature," in *Human Communication: A Unified View,* E.E. David and P.B. Denes, eds., McGraw-Hill, New York, 1972.

[53] R. Jakobson, C.G.M. Fant, and M. Halle, *Preliminaries to Speech Analysis,* MIT Press, Cambridge, 1963.

[54] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE,* vol. 64, pp. 532-556, 1976.

[55] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1982.

[56] P.W. Jusczyk, "A model of the development of speech perception," in *Invariance and Variability in Speech Processes,* J.S. Perkell and D.H. Klatt, eds., Lawrence Erlbaum Ass., New Jersey, 1986.

[57] D.H. Klatt, "Structure of confusions in short-term memory between English consonants," *J. Acoust. Soc. Amer.,* vol. 44, no. 2, pp. 401-407, 1968.

[58] D.H. Klatt, "Review of the ARPA speech understanding project, *J. Acoust. Soc. Amer.,* vol. 62, no. 6, pp. 1345-1366, 1977.

[59] D.H. Klatt, "Speech perception: A model of acoustic-phonetic analysis and lexical access," *J. Phonetics,* vol. 7, no. 3, pp. 279-312, 1979.

[60] D.H. Klatt, "Prediction of preceived phonetic distance from critical-band spectra: A first step," *Proc. ICASSP,* pp. 1278-1281, 1982.

[61] D.H. Klatt, "Models of phonetic recognition I: Issues that arise in attempting to specify a feature-based strategy for speech recognition," *Proc. Symp. on speech recognition,* Montreal, pp. 63-66, July, 1986.

[62] W. Klein, R. Plomp, and L.C.W. Pols, "Vowel spectra, vowel spaces, and vowel identification," *J. Acoust. Soc. Amer.*, vol. 48, no. 4, pp. 999-1009, 1970.

[63] H.P. Kramer and M.V. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE Trans. Inform. Theory,* vol. 2, pp. 41-46, 1956.

[64] G.F. Kubala, "A feature space for acoustic-phonetic decoding of speech," S.M. Thesis, MIT, 1984.

[65] H. Kuchera and W.N. Francis, "Computational analysis of present-day American English," *Brown University Press,* Providence, R.I., 1967.

[66] W. Labov, "Sources of inherent variation in the speech process," in *Invariance and Variability in Speech Processes,* J.S. Perkell and D.H. Klatt, eds., Lawrence Erlbaum Ass., New Jersey, 1986.

[67] P. Ladefoged, *A Course in Phonetics,* Harcourt Brace Jovanovich, New York, 1982.

[68] L.F. Lamel and V.W. Zue, "Properties of consonant sequences within words and across word boundaries," *Proc. ICASSP,* San Diego, 1984.

[69] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop,* Report no. SAIC-86/1546, 1986.

[70] W.A. Lea, "Trends in Speech Recognition," Prentice Hall Inc., Englewood Cliffs, New Jersey, 1980.

[71] H.C. Leung, "A procedure for automatic alignment of phonetic transcriptions with continuous speech," S.M. Thesis, MIT, 1985.

[72] S.E. Levinson, "Structural methods in automatic speech recognition," *Proc. IEEE,* vol. 73, no. 11, pp. 1625-1650, 1985.

[73] S.E. Levinson, "Continuous speech recognition by means of acoustic/phonetic classification from a hidden Markov model," *Proc. ICASSP,* pp. 93-96, Dallas, 1987.

[74] K.P. Li, G.W. Hughes, and A.S. House, "Correlation characteristics and dimensionality of speech spectra," *J. Acoust. Soc. Amer.,* vol. 46, no. 4, pp. 1019-1025, 1969.

[75] N. Lindgren, "Machine recognition of human language," *IEEE Spectrum,* vol. 2, no. 4, pp. 44-59, 1965.

[76] L. Lisker and A.S. Abramson, "Some effects of context on voice onset time in English stops," *Language and Speech,* vol. 10, pp. 1-28, 1967.

[77] R.F. Lyon, "Computational models of neural auditory processing," *Proc. ICASSP-84,* 1984.

[78] R.F. Lyon, "Speech recognition in scale space," *Proc. ICASSP,* pp. 1265-1268, 1987.

[79] S.M. Marcus and R.A.J.M. van Lieshout, "Temporal decomposition of speech," *IPO Annual Progress Report,* vol. 19, pp. 25-31, Institute for Perception Research, Eindhoven, The Netherlands, 1984.

[80] M.V. Mathews, J.E. Miller, and E.E. David Jr., "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Amer.,* vol. 33, no. 2, pp. 179-186, 1961.

[81] G.A. Miller and P.E. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.,* vol. 27, no. 2, pp. 338-352, 1955.

[82] J.A. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Trans. Acoust. Speech and Signal Proc.,* vol. 22, pp. 330-338, 1974.

[83] L. Nakatani and K.D. Dukes, "Locus of segmental cues for word juncture," *J. Acoust. Soc. Amer.,* vol. 62, no. 3, pp. 714-719, 1977.

[84] A.M. Noll, "Cepstrum Pitch Determination", *J. Acoust. Soc. Amer.,* vol. 41, no. 2, pp. 293-309, 1967.

[85] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1975.

[86] G.E Peterson, W. Wang, and E. Silvertsen, "Segmentation techniques for speech synthesis," *J. Acoust. Soc. Amer.,* vol. 30, pp. 739-742, 1958.

[87] E.N. Pinson, "Pitch Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths", *J. Acoust. Soc. Amer.,* vol. 35, no. 8, pp. 1264-1273, 1963.

[88] D.B. Pisoni and P.A. Luce, "Acoustic-phonetic representations in word recognition," *Cognition,* vol. 25, pp. 21-52, 1987.

[89] R. Plomp, L.C.W. Pols, and J.P. v.d.Geer, "Dimensional analysis of vowel spectra," *J. Acoust. Soc. Amer.,* vol. 41, no. 3, pp. 707-712, 1967.

[90] L.C.W. Pols, L.J.Th. v.d.Kamp, and R. Plomp, "Perceptual and physical space of vowel sounds," *J. Acoust. Soc. Amer.,* vol. 46, no. 2, pp. 458-467, 1969.

[91] L.C.W. Pols, "Real-time recognition of spoken words," *IEEE Trans. Comput.,* vol. 20, pp. 972-978, 1971.

[92] L.C.W. Pols, H.R.C. Tromp, and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," *J. Acoust. Soc. Amer.,* vol. 53, no. 4, pp. 1093-1101, 1973.

[93] 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 and Signal Proc.,* vol. 24, pp. 399-418, 1976.

[94] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978.

[95] M.A. Randolph and V.W. Zue, "The role of syllable structure in the acoustic realization of stops," *Proc.* 11th *Int. Congress of Phonetic Sciences,* Tallinn, Estonia, 1987.

[96] R.A. Reddy, "Speech recognition by machine: A review," *Proc. IEEE,* vol. 64, pp. 501-531, 1976.

[97] S. Roucos and M.O. Dunham, "A stochastic segment model for phoneme-based continuous speech recognition," *Proc. ICASSP-87,* pp. 73-76, 1987.

[98] R.M. Schwartz and V.W. Zue, "Acoustic-phonetic recognition in BBN SPEECHLIS," *Proc. ICASSP,* pp. 21-24, 1976.

[99] C.L. Searle, J.Z. Jacobson, and S.G. Rayment, "Stop consonant discrimination based on human audition," *J. Acoust. Soc. Amer.,* vol. 65, no. 3, pp. 799-809, 1979.

[100] C.L. Searle, J.Z. Jacobson, and B.P. Kimberley, "Speech as patterns in the 3-space of time and frequency," in *Perception and Production of Fluent Speech,* R.A. Cole, ed., Lawrence Erlbaum Associates, Hillsdale, N.J., 1980.

[101] S. Seneff, "Real-time harmonic pitch detector," *IEEE Trans. Acoust. Speech and Signal Proc.,* vol. 26, pp. 385-365, 1978.

[102] S. Seneff, D.H. Klatt, and V.W. Zue, "Design considerations for optimizing the intelligiblity of DFT-based, pitched-excited, critical-band spectrum speech analysis/resynthesis system", *Speech Communication Group Working Papers,* No. 1, pp. 31-46, 1982.

[103] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics,* vol. 16, no. 1, pp. 55-76, 1988.

[104] S. Seneff and V.W. Zue, "Transcription and alignment of the TIMIT database," report distributed with the TIMIT database by NBS.

[105] S. Shattuck-Hufnagel and D.H. Klatt, "The limited use of distinctive features and markedness in speech production: Evidence from speech error data," *J. Verbal Learning and Verbal Behavior,* vol. 18, pp. 41-45.

[106] D.W. Shipman and V.W. Zue, "Properties of large lexicons: Implications for advanced isolated word recognition systems," *Proc. ICASSP,* pp. 546-549, 1982.

[107] H.F. Silverman and N.R. Dixon, "A comparison of several speech-spectra classification methods," *IEEE Trans. Acoust. Speech Sig. Proc.,* vol. 24, no. 4, pp. 289-295, 1976.

[108] M.M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio and Electroacoustics*, vol. 16, pp. 262-266, 1968.

[109] K.N. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View,* E.E. David and P.B. Denes, eds., McGraw-Hill, New York, 1972.

[110] K.N. Stevens, "Acoustic correlates of some phonetic categories," *J. Acoust. Soc. Amer.,* vol. 63, no. 3, pp. 836-842, 1980.

[111] K.N. Stevens and S.E. Blumstein, "The search for invariant acoustic correlates of phonetic features," In *Perspectives on the study of speech,* P.D. Eimas and J.L. Miller, eds., Erlbaum, Hillsdale, N.J., 1981.

[112] K.N. Stevens, "Models of phonetic recognition II: An approach to feature-based recognition," *Proc. Symposium on Speech Recognition,* Montreal, pp. 67-68, July, 1986.

[113] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. ASSP,* vol. 35, no. 10, pp. 1414-1422, 1987.

[114] N. Umeda, "Consonant duration in American English," *J. Acoust. Soc. Amer.,* vol. 61, no. 3, pp. 846-858, 1977.

[115] W.A. Wickelgren, "Distinctive features and errors in short-term memory for English vowels," *J. Acoust. Soc. Amer.,* vol. 38, pp. 583-588, 1965.

[116] W.A. Wickelgren, "Context-sensitive coding, associative memory and serial order in speech behaviour, *Psychological Review,* vol. 76, pp. 1-15, 1969.

[117] 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. ICASSP,* pp. 821-824, 1987.

[118] M. Withgott and M.A. Bush, "On the robustness of phonetic information in short-time speech spectra," *Proc. Symposium on Speech Recognition,* Montreal, pp. 101-102, July, 1986.

[119] M. Withgott, S.C. Bagley, R.F. Lyon, and M.A. Bush, "Acoustic-phonetic segment classification and scale-space filtering," *Proc. ICASSP,* pp. 860-863, 1987.

[120] A.P. Witkin, "Scale-space filtering: A new approach to multi-scale description," *Proc. ICASSP,* 1984.

[121] S.A. Zahorian and M. Rothenberg, "Principal-components analysis for low-redundancy encoding of speech spectra," *J. Acoust. Soc. Amer.,* vol. 69, no. 3, pp. 832-845, 1981.

[122] V.W. Zue and R.A. Cole, "Experiments on spectrogram reading," in *Proc. ICASSP*, pp. 116-119, 1979.

[123] V.W. Zue, "The use of speech knowledge in automatic speech recognition," *Proc. IEEE,* vol. 73, no. 11, pp. 1602-1615, 1985.

[124] V.W. Zue, "Models of phonetic recognition III: The role of analysis by synthesis in phonetic recognition," *Proc. Symposium on Speech Recognition,* Montreal, pp. 69-70, July, 1986.

[125] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Amer.,* vol. 33, pp. 248-249.

Appendix A

Dendrograms

This appendix contains examples of dendrograms of typical utterances from the TIMIT database. Each figure contains four displays of: (1) a wide-band spectrogram. (2) a dendrogram, (3) the speech waveform and, (4) the aligned phonetic transcription. The shaded regions in the dendrograms correspond to the sequence of acoustic segments which best aligned with the hand-marked phonetic transcription, based on the procedure described in Chapter 3.

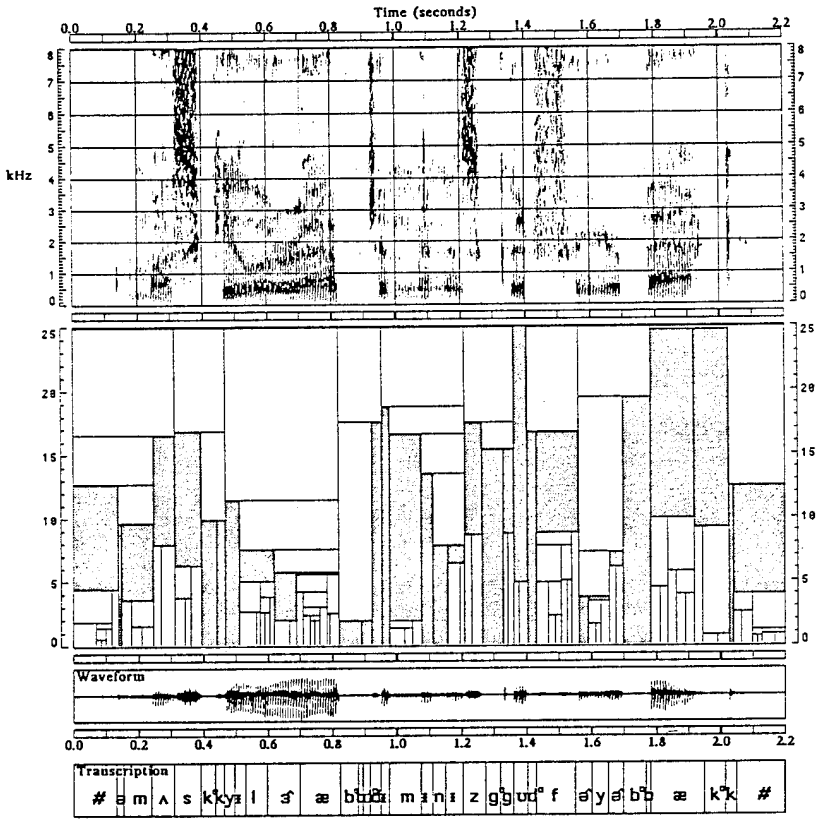


Figure A.1: Dendrogram of 'A muscular abdomen is good for your back.'

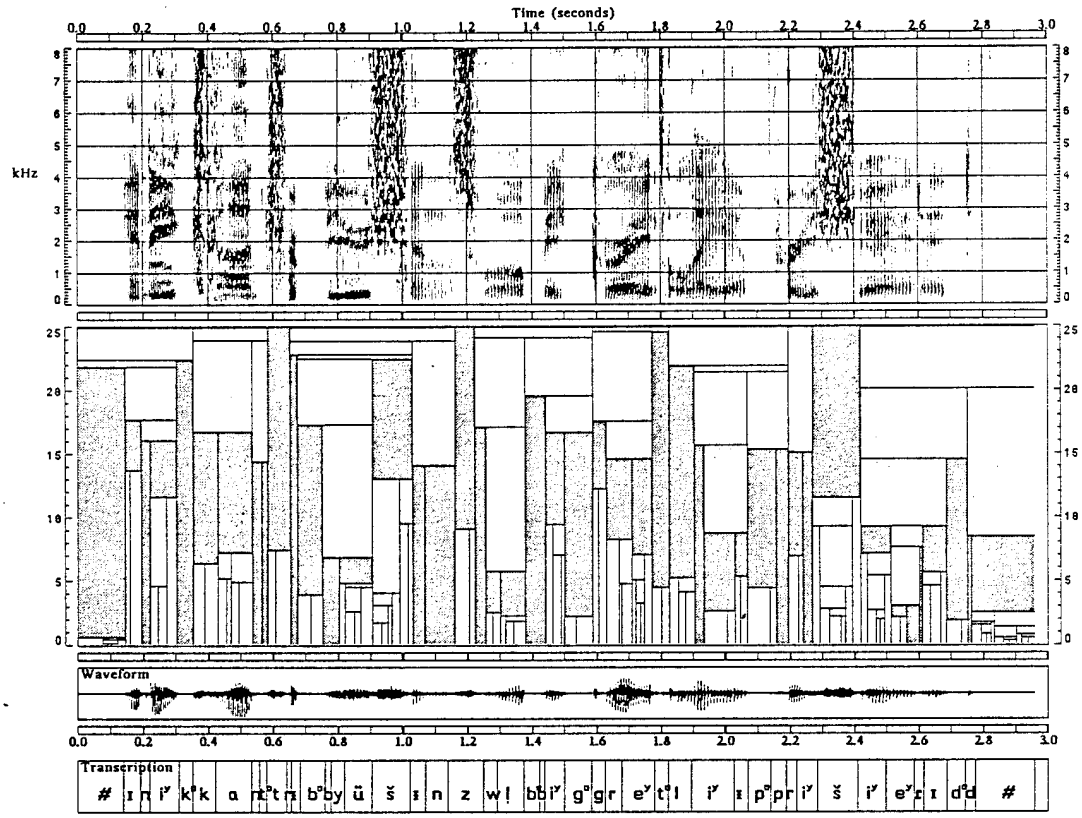


Figure A.2: Dendrogram of 'Any contributions will be greatly appreciated.'

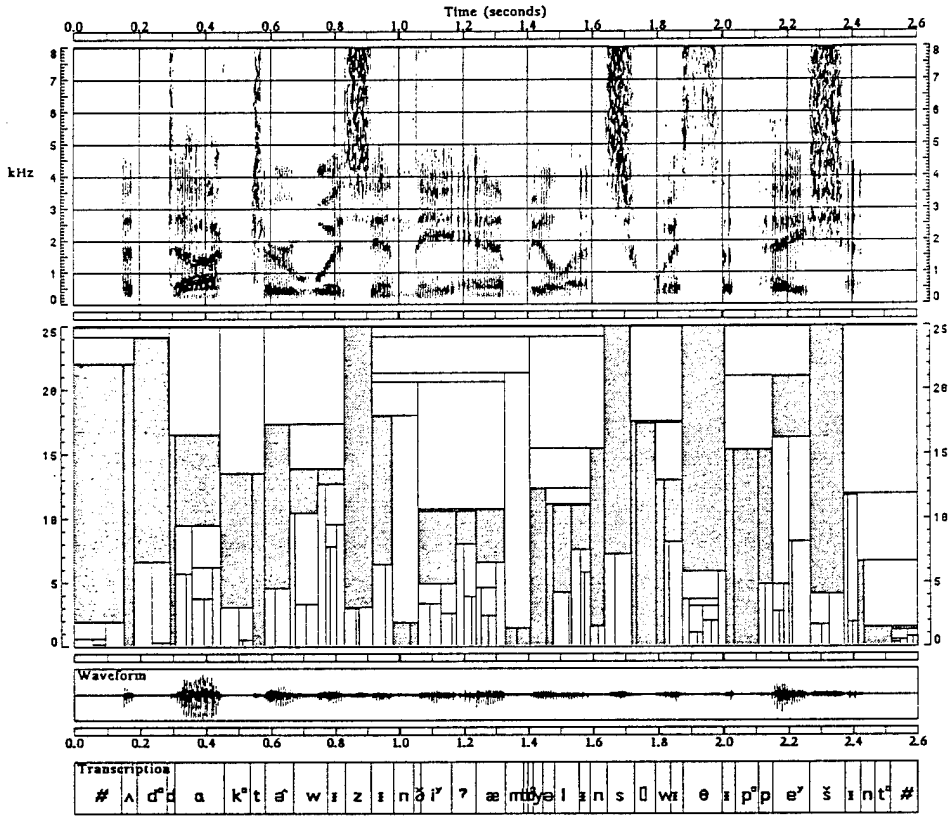


Figure A.3: Dendrogram of 'A doctor was in the ambulance with the patient.'

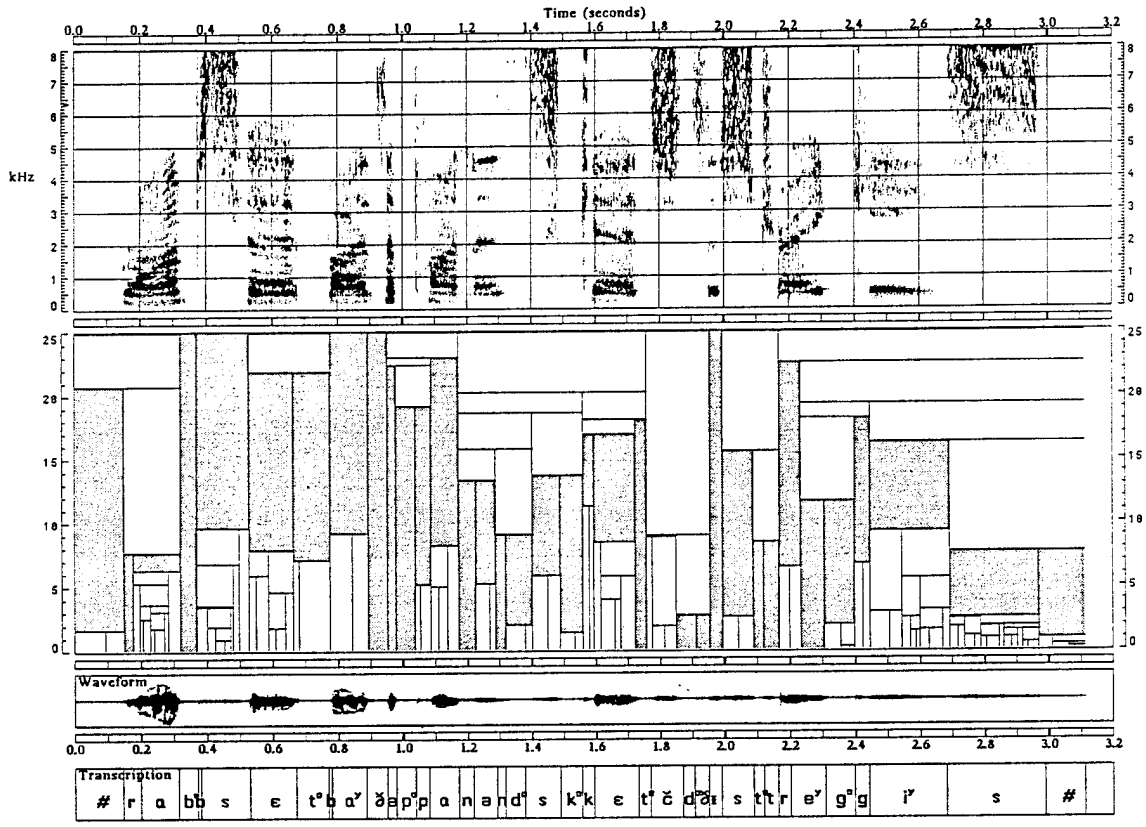


Figure A.4: Dendrogram of 'Rob sat by the pond and sketched the stray geese.'

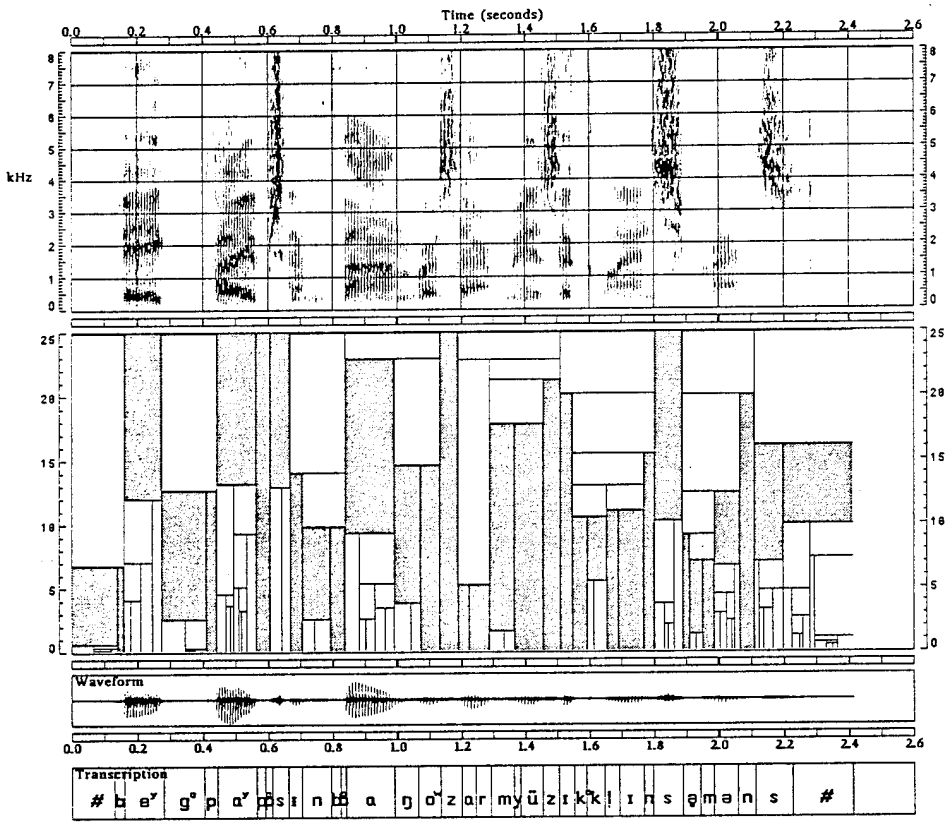


Figure A.5: Dendrogram of 'Bagpipes and bongos are musical instruments.'

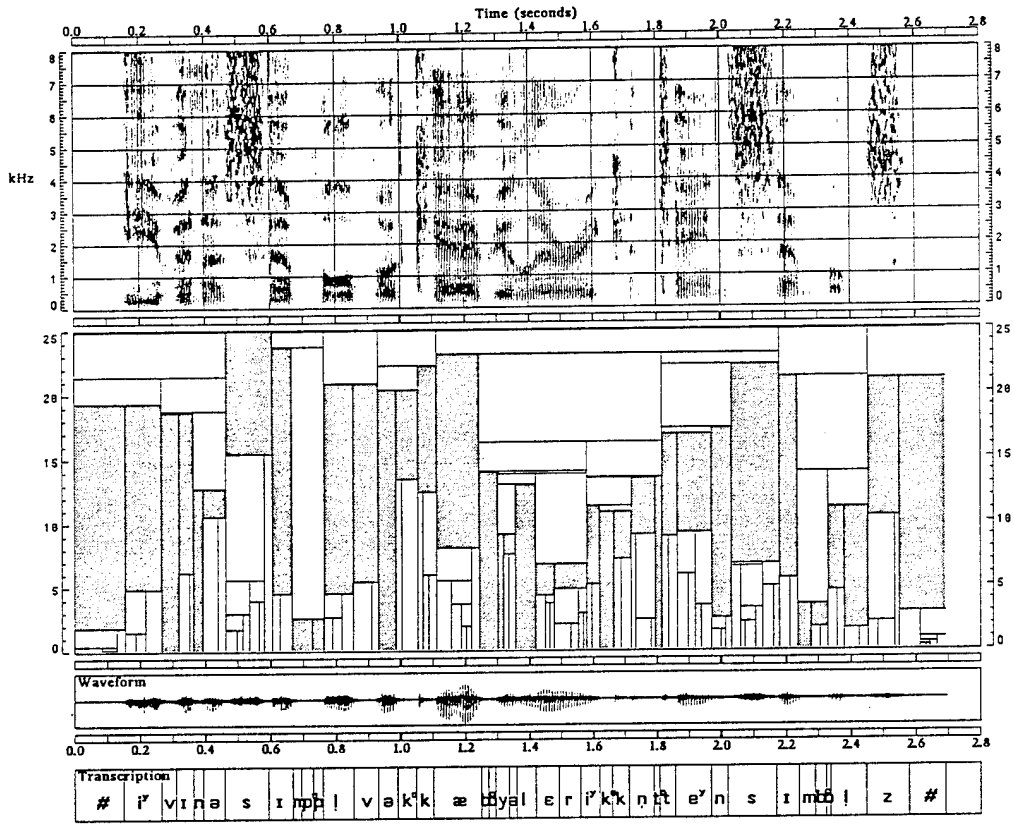


Figure A.6: Dendrogram of 'Even a simple vocabulary contains symbols.'

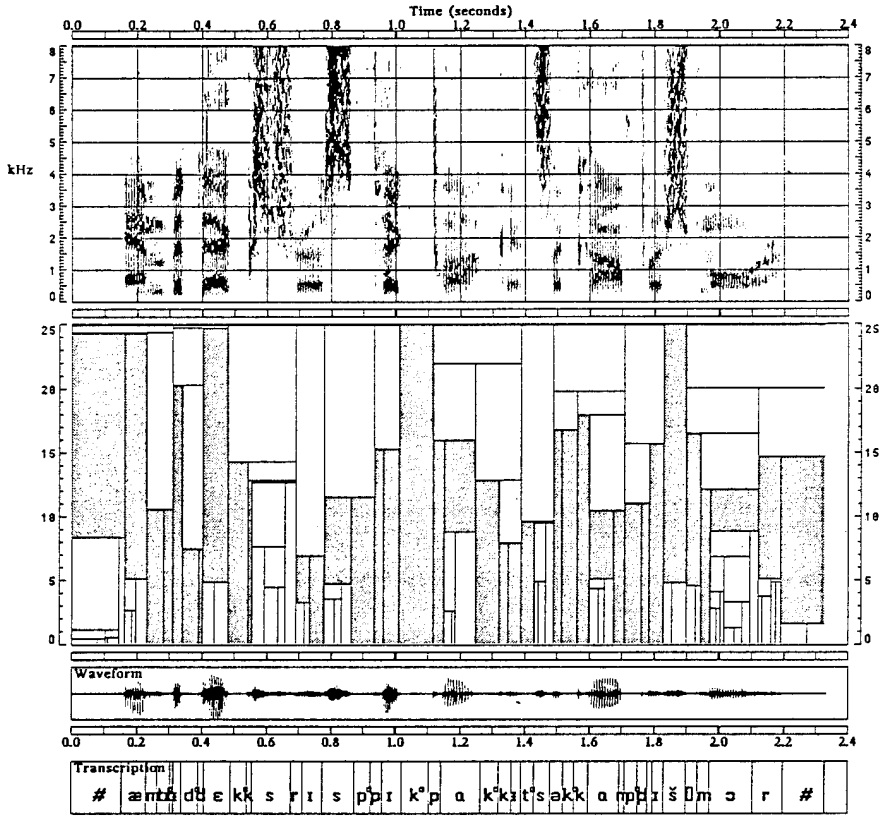


Figure A.7: Dendrogram of 'Ambidextrous pickpockets accomplish more.'

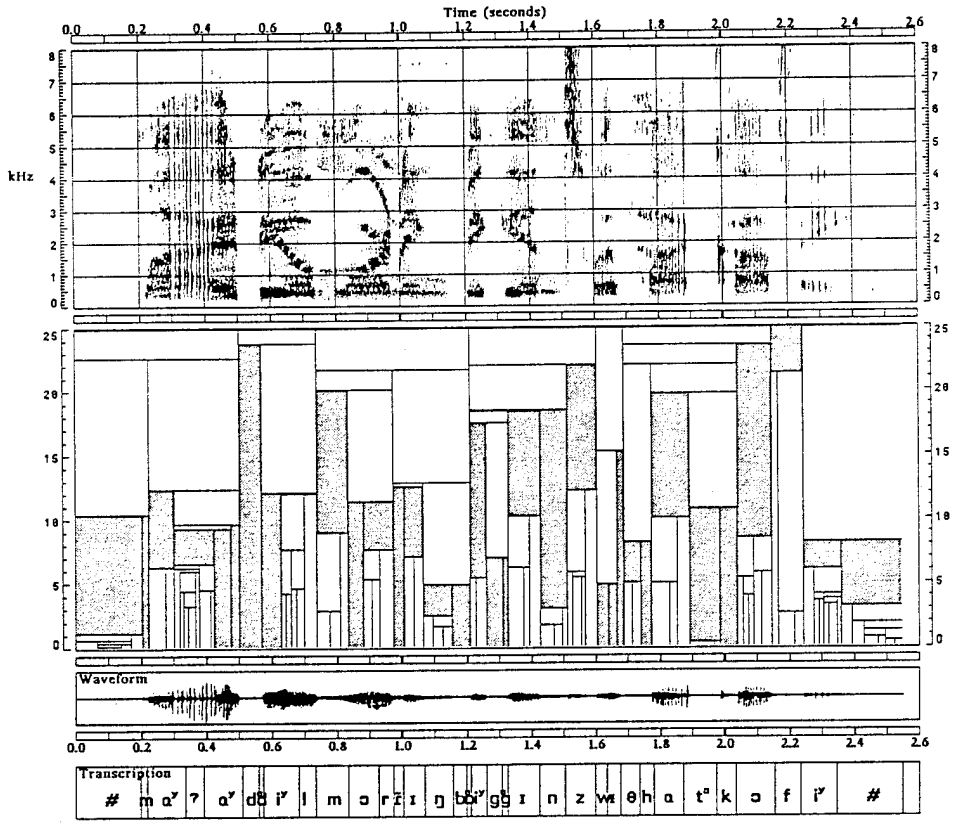


Figure A.8: Dendrogram of 'My ideal morning begins with hot coffee.'

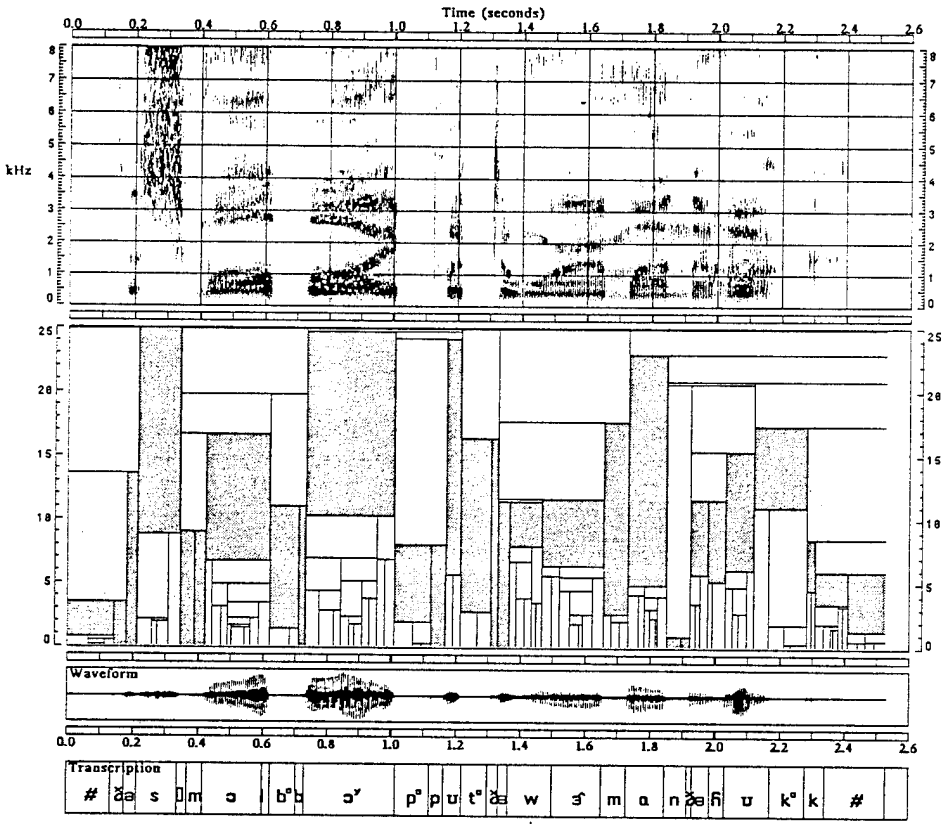


Figure A.9: Dendrogram of 'The small boy put the worm on the hook.'

Appendix B

Stochastic Segmentation Computation

A stochastic segmentation considers all possible segmentations of a set of n frames. This appendix illustrates the amount of computation that is involved with such an approach. First, consider the number of possible ways there are to divide n frames of speech into m possible segments, $s(n,m)$. This is a recursive computation since we must consider all possible lengths of the first segment along with all possible segmentations of the remaining frames into $m-1$ segments. Specifically,

$$s(n,m) = \sum_{k=1}^{n-m+1} s(n-k, m-1) \qquad \text{where} \quad n \geq m, \quad \text{and} \quad s(n,1) = 1, \forall n \qquad \text{(B.1)}$$

The upper limit on the length of a segment is to ensure that all remaining segments have a length of at least one frame. This computation is illustrated in Figure B.1. Note that this figure suggests another computation for

$$s(n,m) = s(n-1,m) + s(n-1,m-1)$$

or that there are $\binom{n-1}{m-1}$ ways to segment n frames into m segments.

Since the value of m is usually unknown, the total number of possible segmentations is the sum of the number of segmentations for any value of m,

$$S_T(n) = \sum_{m=1}^{n} s(n,m)$$

In Figure B.1 this corresponds to summing along the vertical dimension. From the figure we can see that there are 2^{n-1} possible segmentations of n frames. A simpler

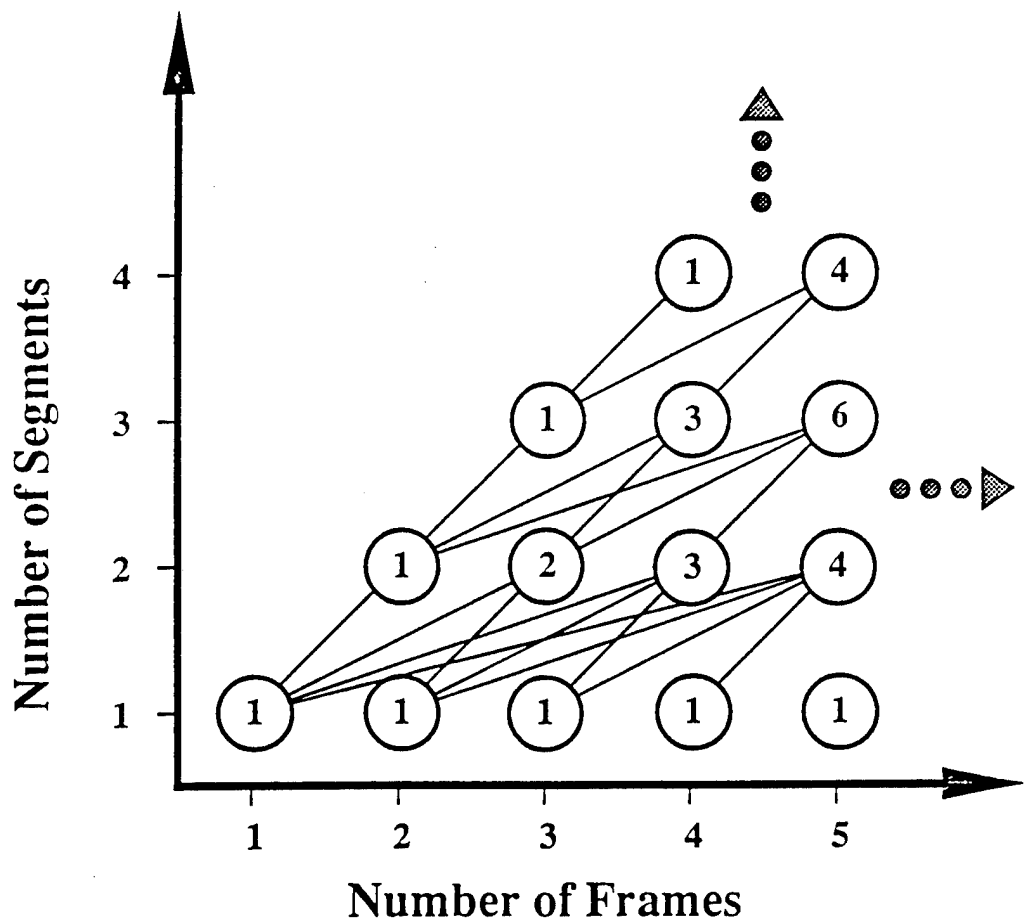


Figure B.1: Stochastic segmentation computation.

This figure illustrates the amount of computation involved in a stochastic segmentation. The value in each circle corresponds to the number of ways there are to segment n frames into m segments, $s(n, m)$. This figure also illustrates the recursive nature of the computation since, as shown in equation B.1, $s(n, m)$ may be computed from $\{s(n-1, m-1), s(n-2, m-1)...\}$.

way of arriving at this number would have been to consider that there are $n-1$ possible frames which can start (end) a segment since the first (last) frame must start (end) a segment. Given that any of the $n-1$ frames has a choice of starting (ending) a segment, there are 2^{n-1} possible segmentations.

No matter how the analysis is made, the number of possible segmentations is overwhelming for any reasonable amount of speech. A 2 second utterance analysed with a frame-rate of 10 ms would have well over 10^{60} possible segmentations for instance. One way to reduce the space is to restrict the maximum length of a given segment to some number of frames, L. The computation is then modified to

$$s_L(n,m) \quad = \quad \sum_{k=1}^{min(L,n-m+1)} s_L(n-k,m) \quad = \quad s(n,m) - s(n-L,m)$$

since we can consider this as adding up only a fraction of possible segmentations. From this computation we can see that there are $2^{n-1} - 2^{n-L-1} = 2^{n-L-1}(2^{L-1}-1)$ possible segmentations. Reasonable values of L can be determined from studies of duration of sounds in continuous speech. A reasonable average phone duration is approximately 80 ms [20]. Using a frame-rate of 5 ms, and assuming that a phone is never larger than twice its average, will lead to a value of L on the order of 30. Note however, that the total number of possible segmentations remains virtually the same as before. Thus, some other form of reducing the search space is necessary.

With search strategies which perform best-first search with dynamic-programming, the amount of search involved in a stochastic segmentation is much more reasonable. Consider for example the case where M objects are to be recognized. This is a search problem, and it is desired to find the most probable sequence of the M objects. An algorithm which is commonly used to perform the decoding is the Viterbi search [38], which basically expands all possible paths uniformly in time, but uses dynamic programming to prune less probable paths. In this example, we will consider two strategies for representing the M objects. First, each object will be represented as a single state in a finite state network, where each state may connect to any other state. From a computational perspective, at the nth observation, the ith object can

either connect to itself, or it can connect to one of the other $M-1$ objects. All together, there are M^2 possible connections at the nth observation. If there are a total of N observations, then there will be a total of M^2N combinations. The amount of computation therefore varies linearly with time.

Consider now a segmental framework where each object must hypthesize a precise segment duration, unlike the previous framework. In this case, the ith state must consider ML_i possible connections, where L_i is the number of possible segment lengths the ith state may have. The total number of combinations at the nth observation is therefore M^2L_{ave} where L_{ave} is the average length of the M segments. Previously it was pointed out that a reasonable value of L_{ave} would be around 16. Thus we can see that a stochastic segment based approach would in fact increase the total number of computations by a little more than an order of magnitude. While this is substantial, it is far less computation than would be involved in an exhaustive search.

Apart from the increase in computation, there is an additional drawback to a stochastic segment based model. With a simple first-order model, such as was used in the preceding example, and is common in the literature in the form of hidden Markov models, the Viterbi algorithm may be made quite efficient in memory storage, since connections were only concerned with the previous observation. In a segmental model, however, this efficiency is lost, and storage on the order of ML_{ave} would be necessary.