

BOSTON UNIVERSITY
COLLEGE OF ENGINEERING

DISSERTATION

ACOUSTIC ANALYSIS AND MODELING OF SPEECH BASED
ON
PHONETIC FEATURES

BY

NABIL N. BITAR

B.S. (Hons.), Electrical Engineering, Boston University, 1989

M.S., Electrical Engineering, Boston University, 1992

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

1998

Approved by

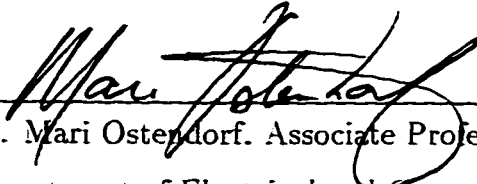
First Reader



Dr. Carol Espy-Wilson, Assistant Professor.

Department of Electrical and Computer Engineering
Boston University

Second Reader



Dr. Mari Osterdorf, Associate Professor.

Department of Electrical and Computer Engineering
Boston University

Third Reader



Dr. John Makhoul, Chief Scientist.

BBN Technologies, GTE Internetworking
Adjunct Professor.

Department of Electrical Engineering,
Northeastern University

Fourth Reader



Dr. William L. Oliver, Associate Professor.

Department of Electrical and Computer Engineering
Boston University

Acknowledgments

In many ways, this thesis was the work of several people to whom I feel obliged to express my sincere gratitude. Among these people is my academic advisor professor Carol Espy-Wilson. Professor Espy-Wilson provided me with well appreciated technical guidance throughout the different phases of this thesis. She taught me what I know about different aspects of speech and she financially supported my research through her research grants. I thank Carol for the different types of support she provided me.

I am also grateful to my readers: professor Mari Ostendorf, Dr. John Makhoul, professor William Oliver and Dean Peter Levin who took the time to read the thesis and provide valuable comments. Especially, I would like to acknowledge professor Ostendorf's thorough reading of this thesis and her valuable comments that added to its clarity. Professor Ostendorf was also my teacher in two graduate classes that I very much enjoyed: pattern recognition and stochastic processes.

I am also grateful to professor Hamid Nawab who was my academic and research advisor during my master studies. Professor Nawab taught me the basics of digital signal processing and research.

My friends and colleagues in graduate school made it all worthwhile and enjoyable. Among these friends are: Ashvin Kannan, Ramamurthy Mani, Demetrios Paneras, Venkatesh Chari, Joseph Winograd, Rukmini Iyer, Suzanne Boyce, Caroline Huang, Ana Solar, Marina Pilar-Santamarina, Jamil Sobh, Ibrahim Bechwati, Ziad El-Ghezzawi and Robert Habib. To all of them, thanks for the wonderful times talking about technical issues, chatting in hallways or engaging in discussions that

spanned politics, philosophy and the social plights of the universe.

Many thanks are also due to my friends outside of school who made our stay away from home feels like home. Among these friends are the Ghannam's, the Rajeh's, Nasser Abou Khozam, Assef Zobian and Nibal Harati. Sharing a social, lunch or dinner with them was always the perfect break. My friend and x-roommate Ali Salem was also always there to discuss things with me. To all of them, many thanks.

Many thanks to the National Sciences Foundation (NSF) and to the Hariri Foundation for the financial support that they provided me at different stages of my education.

I would also like to thank my wife Ghada and my daughter Lama who were "most often" patient with me and very understanding. They took the usual answers to their requests: "No I don't have time now", "I have to prepare for an exam" or "I have to stay late in the lab tonight", among many other "justified excuses", reasonably well. My wife endured with me seven years of graduate school. My daughter opened her eyes to this world when I was a graduate student. To Those of you who went through graduate school, you know what that means.

Last but not least, I would like to thank my parents for the many sacrifices they made for me and for the values they instilled in me. I only hope that I can be as good to my children. Their encouragement and support have been my strength.

I finally dedicate this thesis to my friend and father Najib Bitar. To my regret, my father did not live to cherish this moment with me. He passed away during the first year of my PhD, but he has always been with me, in my heart.

ACOUSTIC ANALYSIS AND MODELING OF SPEECH BASED ON PHONETIC FEATURES

(Order No.)

NABIL N. BITAR

Boston University, College of Engineering, 1998

Major Professor: Carol Espy-Wilson Assistant Professor of Electrical Engineering

ABSTRACT

Acoustic modeling and analysis of speech based on phonetic features is explored in the current research for speaker-independent speech recognition. Phonetic features are minimal speech units that describe the manner and place of articulation of the sounds of a language. In this research, it is shown that phonetic features have acoustic signatures in the speech signal that can be reliably extracted in a manner that reduces the effects of speaker-differences. Moreover, it is postulated based on the conducted experiments that using phonetic features as the basic speech units allows for the modeling of contextual variability in a general and natural way.

A major thrust of this thesis is in the development of algorithms that extract the acoustic properties of the phonetic features. These algorithms make measurements on the speech signal that are motivated by acoustic phonetics and spectrographic analysis. A measurement is made at a time-instant relative to its value at another instant and/or is made in a frequency band relative to another. Such relative measurements focus on the linguistic content of the speech signal reducing the effects of interspeaker variability.

In one part of this thesis, acoustic measurements were developed based on sub-

jective acoustic analysis. An event-based recognition system that uses these measurements, combined by “fuzzy” rules, was developed and compared to a Hidden Markov Model (HMM) system using (1) the same measurements but modified to fit the frame-based HMM system and (2) Mel-cepstral parameters. The results show that the event-based approach produces comparable results to the HMM frame-based system for the undertaken task of broad-class speech recognition. In addition, it is shown that the developed measurements perform better than the cepstral parameters in this task.

An automatic optimization procedure based on the Fisher criterion and classification trees was developed to automate the derivation of acoustic measurements. Using this procedure, manner and place-of-articulation acoustic measurements were developed. These measurements were evaluated in phonetic-feature classification tasks and in a 10-class recognition task using an HMM system. Recognition results compared favorably to those obtained with Mel-cepstral parameters. The results show that the developed measurements target the intended linguistic information and are robust to speaker differences.

Contents

1	Introduction	1
2	Acoustic Modeling of Speech: Background	5
2.1	Issues in Acoustic Modeling	5
2.2	Acoustic Modeling Methodologies	9
2.2.1	Mathematical Models	10
2.2.2	The Acoustic-Phonetic Approach	14
3	Phonetic Features and Acoustic Modeling	18
3.1	Phonetic Features: an overview	19
3.2	Acoustic Properties of Phonetic Features	20
3.3	Philosophy in Designing Acoustic Parameters: Relative Measures	23
3.4	Uncertainty Modeling	28
3.4.1	Sources of Uncertainty in the Speech Signal	28
3.4.2	Fuzzy Logic Framework	29
3.4.3	Probabilistic Approach	32
4	Database	33

5	Manner-Class Recognition Based on Phonetic Features	37
5.1	Acoustic Parameters	39
5.2	APs and EBS	41
5.2.1	Mapping the APs to the phonetic-feature space	43
5.2.2	EBS	48
5.2.3	Error Analysis	53
5.3	HMM Recognition System and APs	55
5.3.1	Modification of the APs for the HMM framework	55
5.3.2	Recognition Experiments	59
5.3.3	Results and Discussion	61
5.4	Concluding Remarks	63
6	Parameter Optimization	65
6.1	Motivation	65
6.2	Procedure	66
6.3	Fisher Criterion: First Stage	70
6.4	Classification Trees: Second Stage	72
6.4.1	Node-Splitting Criterion:	74
6.4.2	Determining Tree Size and Important Parameters	76
7	Optimized Acoustic Parameters	79
7.1	Sonorancy	79
7.1.1	Acoustic Parameters that target the sonorancy feature	81
7.1.2	Parameter Optimization	82
7.1.3	Classification Results	95

7.2	Anterior Place-of-Articulation for Stridents	98
7.2.1	Acoustic Parameters to Identify the Anterior Place-of-Articulation for Stridents	99
7.2.2	Optimized Parameters	102
7.2.3	Classification Results	113
7.3	Labial, Alveolar and Velar Place-of-Articulation Parameters for the Stop Consonants	114
7.3.1	Acoustic Parameters for Identifying the Stop Place of Articulation	119
7.3.2	Optimized Parameters	123
7.3.3	Classification Results	125
7.4	Syllabicity	130
7.4.1	Algorithm for Detecting a Syllabic/Nonsyllabic Acoustic Event	132
7.4.2	Optimized Parameters	134
7.4.3	Classification Results	139
7.5	Stridency	141
7.5.1	Acoustic Parameters for Stridency: strident obstruents vs. weak fricatives	141
7.5.2	Acoustic Parameters for Stridency: Affricates vs. Stops	144
8	Speech Manner and Obstruent Place-of-Articulation Recognition	148
8.1	Experimental Objectives	150
8.2	Signal Representation	150
8.3	Acoustic Models	152
8.4	Acoustic Model Training	153

8.5	Recognition Experiments	154
8.5.1	Baseline Experiments	154
8.5.2	Performance of APs	155
8.5.3	APs vs. Cepstra: A Performance Comparison	162
8.5.4	Gender Experiments: APs vs. Cepstra	164
8.6	Concluding Remarks	165
9	Discussion and Conclusions	168
9.1	The Feature-Based Approach to Speech Analysis and Recognition as a learning Tool	168
9.2	Summary	169
9.3	Future Work	175
A	Fuzzy Evaluation Index	180
	Bibliography	183

List of Figures

2.1	A diagram depicting the different knowledge sources contributing to the recognition of a spoken utterance.	6
3.1	The hierarchy of phonetic feature organization.	20
5.1	The hierarchy of manner feature organization adopted in parameter development.	44
5.2	(a) parameter: normalized energy (100-400 Hz) (b) membership function for that parameter being strong (c) Membership values assigned to normalized energy.	45
5.3	An example of a representation in the phonetic-feature space for the word "amorist". The items displayed in this figure from top to bottom are: (1) spectrogram, (2) degree of silence, (3) degree of sonorancy, (4) degree of a syllabic peak, (5) degree of a nonsyllabic dip, (6) degree of noncontinuancy (7) degree of frication and (8) phonetic transcription where a phone label appears at the beginning of the associated time-segment. Note that (4) and (5) mark particular time-instants.	47

5.4	This figure illustrates the set of parameters listed in Table 5.7. These parameters are: (a) abrupt onset, (b) E3-6. (c) E0.2-3. (d) dtp ₂₋₃ . (e) dtp _{0.64-2.8} . (f) ptd ₂₋₃ . (g) ptd _{0.64-2.8} . (h) dtp _{RI} . (i) RI. (j) zcr. (k) voicing-probability. (l) E0-2-2-8 and (m) E0.1-0.4.	60
6.1	Fisher criterion for the parameter which computes the energy between f_{st} and f_{end} relative to the overall energy at a given frame instant. The origin is $F3 - 1000$ (Hz).	70
6.2	Distribution of best parameter computed using the anterior samples for males (m) and females (f). (a) Parameter was defined relative to third formant (F3) location. (b) Parameter was independent of F3.	71
6.3	Separation between three classes using the Fisher criterion is based on maximizing the distances among the class centroids and the centroid of the pooled data.	73
7.1	Voicing probability distribution for (a) vowels and (b) syllabic consonants. The TIMIT symbols for phones are used along the horizontal axis.	84
7.2	Voicing probability distribution for (a) sonorant consonants and (b) obstruents. The TIMIT symbols for phones are used along the horizontal axis.	85
7.3	Histograms of voicing probability computed over the training data (a) sonorant samples and (b) nonsonorant samples. Histograms of the peak cross correlation coefficient computed over the training data (c) sonorant samples and (d) nonsonorant samples.	91

7.4	Histograms of $E[0:688]/E[4000:8000]$ computed over (a) the sonorant samples and (b) the nonsonorant samples in the training data. Histograms of $E[0:375]/\text{average}(E[0:375])$ computed over (c) the sonorant samples and (d) the nonsonorant samples.	92
7.5	The distribution of voicing probability for males and females is shown in (a) for the sonorant samples and in (b) for the nonsonorant samples. The distribution of cross-correlation peak for males and females is shown in (c) for the sonorant samples and in (d) for the nonsonorant samples.	93
7.6	The distribution of $E[0 : 688]/E[4000 : 8000]$ for males and females is shown in (a) for the sonorant samples and in (b) for the nonsonorant samples. The distribution of $E[0 : 375]/\text{average}(E[0 : 375])$ for males and females is shown in (c) for the sonorant samples and in (d) for the nonsonorant samples.	94
7.7	This spectrogram shows a sonorant “v” and a sonorant /ð/ between 0.4 and 0.5 seconds. The sentence is “I gave them several choices and let them set the priorities” spoken by a male speaker.	96
7.8	This spectrogram shows a canonical fricated “v” at about 1.2 seconds. The sentence is “But it did print good verse and good fiction” spoken by a female speaker.	97
7.9	This figure shows the constriction formed by the tongue in front of the alveolar ridge during the production of the anterior sound /s/. (Taken from Kent [63])	100

7.10	This figure shows the constriction formed by raising the tongue dorsum against the palate, behind the alveolar ridge, during the production of the nonanterior sound /ʒ/. (Taken from Kent [63])	101
7.11	Spectrogram of the utterance “Approach your interview with statuesque composure” spoken by a female speaker. The alveolar /s/ fricatives at about 1.5 and 2.1 seconds have strong energy starting at about 4000 Hz, whereas the palatal /ç/ affricates at 0.55 and 1.82 seconds and the palatal fricative /ʒ/ at 2.77 seconds have strong energy starting at about 2000 Hz and 2300 Hz, respectively.	102
7.12	Histogram of $E[F3 - 187 : F3 + 594]/E[0 : 8000]$ for (a) the anterior samples and (b) the nonanterior samples	111
7.13	The distributions of $E[F3 - 187 : F3 + 594]/E[0 : 8000]$ for the anterior strident phones /s/ and /z/ in (a) and (b), respectively and for the nonanterior strident phones /ʒ/, /ç/, /j/ and /ʒ/ in (c), (d), (e) and (f), respectively.	112
7.14	The shape of the vocal tract during the production of the labial stop /b/. Labial stop consonants are produced by forming a complete closure at the lips.	115
7.15	The shape of the vocal tract during the production of the alveolar stop /d/. Alveolar stop consonants are produced by forming a complete closure at the alveolar ridge with the tongue tip.	116
7.16	The shape of the vocal tract during the production of the velar stop /g/. Velar stop consonants are produced by forming a complete closure in the velum area with the tongue dorsum.	117

7.17	Histograms showing the distribution of $E[F3 + 31 : F3 + 3250]/E[0 : F3 + 31]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.	126
7.18	Histograms showing the distribution of $E[F3 - 1750 : F3]/E[0 : 8000]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.	127
7.19	Histograms showing the distribution of $E[F3 + 281 : F3 + 1187]/E[0 : F3 + 281]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.	128
7.20	Histograms showing the distribution of $E[F3+750 : F3+1050]/E[F3+1050 : 8000]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.	129
7.21	Energy profile typical of intervocalic sonorant consonants. The minimum energy value at point C is measured relative to the smaller of the two surrounding maxima at points A and B.	134
7.22	Energy profile typical of prevocalic sonorant consonants. The minimum energy value at point B is measured relative to the energy maximum at point A.	134
7.23	Energy profile typical of postvocalic sonorant consonants. The minimum energy value at point B is measured relative to the energy maximum at point A.	135
7.24	(a) Within-vowel energy minimum relative to the smaller of the two surrounding maxima within the same vowel. (b) Sonorant-consonant energy minimum relative to the smaller of the two energy maxima in the left and right-context vowels.	138

7.25	(a) Within-vowel energy minimum relative to the smaller of the two surrounding maxima within the same vowel. (b) Sonorant-consonant energy minimum relative to the smaller of the two energy maxima in the left and right-context vowels.	139
7.26	Histograms showing $E[F3 + 94 : 8000]/average(E[F3 + 94 : 8000])$ distributions for: (a) strident obstruents and (b) weak fricatives. . . .	145
7.27	Probability densities of $E[F3 + 94 : 8000]/average(E[F3 + 94 : 8000])$ for: (a) strident obstruents and (b) weak fricatives as a function of gender (females (f) and males (m))	146
A.1	An $S - type$ membership function.	182
A.2	The function $K(x)$ involved in the computation of the entropy.	182

List of Tables

3.1	Phonetic features and articulatory correlates based on Chomsky and Halle [7], and Ladefoged [28].	21
3.2	Phonetic features and articulatory correlates based on Chomsky and Halle [7], and Ladefoged [28] (cont.).	22
3.3	Phonetic features and their acoustic correlates.	24
3.4	Phonetic features and their acoustic correlates (cont.).	25
5.1	The features, their acoustic correlates and the corresponding acoustic parameters.	42
5.2	Mapping between phonetic features and manner classes.	49
5.3	Recognition results comparing EBS with APs to the HMM system using <i>MFCC_E</i> . In scoring, the splits, merges and synonyms in Table 5.5 were counted as correct.	50
5.4	Recognition results comparing EBS with APs (13 parameters) to the HMM system using <i>MFCC_E_δ1_δ2</i> (39 parameters) as the front-end. In scoring, the splits, merges and synonyms in Table 5.5 were counted as correct.	51

5.5	Splits, merges and synonyms that were scored as correct. Category 1 + category 2 means a sequence of category1 and category2.	52
5.6	Error analysis. The errors listed here were deduced from the manner-class recognition results obtained using the EBS. These errors may be explained by speech variability well documented in literature.	56
5.7	The phonetic features, their acoustic correlates and the corresponding acoustic parameters.	59
5.8	Recognition results. <i>MFCC_E</i> refers to Mel-frequency cepstral coefficients & normalized energy. <i>MFCC_Eδ1δ2</i> refers to <i>MFCC_E</i> & their 1st and 2nd derivatives. <i>AP</i> refers to acoustic parameters. <i>APδ1δ2</i> refers to <i>AP</i> and their 1st and 2nd derivatives. Each entry contains % correct/% accuracy. No splitting or merging was allowed in scoring.	62
5.9	Recognition results using 8 mixtures. Training done with speech produced by females. Recognition done with speech produced by males. .	64
5.10	Recognition results obtained when HMM was used as the recognition framework. In scoring, the splittings, mergings and substitutions listed in Table 5.5 were allowed.	64
7.1	Parameters selected in the Fisher-Criterion stage of the parameter optimization process.	86
7.2	These parameters were selected based on the Fisher-Criterion stage of the parameter optimization process.	88

7.3	Parameters selected from the two-stage optimization process to distinguish between sonorant and nonsonorant sounds.	90
7.4	Sonorant-Feature Classification Results	95
7.5	Confusion Results among sonorant and nonsonorant sounds	95
7.6	Parameters that were derived from center-of-gravity and spectral peak.	105
7.7	Energy-Based Parameters for the anterior/nonanterior feature.	107
7.8	The parameters selected by the optimization process to distinguish among the anterior and nonanterior strident sounds. The % correct in each row is the correct classification rate obtained by adding the parameter in that row to the parameters(s) in the previous row(s).	110
7.9	Classification results on the training and test sets for the anterior and nonanterior strident sounds. The classifier was the classification tree obtained in development. The parameters in Table 7.8 were the only used in the classification tree.	114
7.10	Generic energy-ratio parameters for identifying the English stop place of articulation	122
7.11	The acoustic parameters selected to distinguish among the stop consonants.	124
7.12	Classification results on the training and test sets for the stop place of articulation. The classifier was the classification tree obtained in development. The parameters in Table 7.11 were the only ones used in this classification tree.	130
7.13	Energy parameters that are used to detect intervocalic nonsyllabic events.	137

7.14	Energy parameters that are used to detect prevocalic and postvocalic nonsyllabic events.	140
7.15	Classification results on the training and test sets for syllabicity/nonsyllabicity.	140
7.16	Generic acoustic parameters to distinguish between the strident obstruents and the weak fricatives.	142
7.17	Selected acoustic parameters to distinguish between the strident obstruents and the weak fricatives.	143
7.18	Classification results for strident obstruents vs. weak fricatives using the classification tree built in the development stage and the parameters in Table 7.17.	144
7.19	Selected acoustic parameters to distinguish between the affricates and the stop consonants.	147
7.20	Classification results for affricates vs. stop consonants using the classification tree built in the development stage and the parameters in Table 7.19.	147
8.1	This table shows the mapping between the TIMIT labels (represented by IPA symbols) and the speech classes used in this chapter.	151

8.2	Phonetic features, acoustic correlates, and APs used in the HMM recognition system. A <i>dip_to_peak</i> energy parameter is computed by first locating dips and peaks and then computing, in each frame between the peak and the adjacent dip, the difference in energy between the energy at the peak location and the energy in each frame. A <i>peak_to_dip</i> parameter is computed similarly, but relative to the energy at the dip location instead of the energy at the peak location.	156
8.3	Recognition results. <i>MFCC_E</i> refers to 12 Mel-cepstral coefficients normalized & log energy. <i>MFCC_E_δ1_δ2</i> refers to <i>MFCC_E</i> & their 1st and 2nd derivatives. <i>AP</i> refers to 20 acoustic parameters. <i>AP_δ1</i> refers to <i>AP</i> and their 1st derivative. Each entry contains % correct/% accuracy.	157
8.4	The confusion matrix when the signal representation consisted of <i>AP</i> and the observation distribution given an HMM state was Gaussian. All numbers are in percentage.	158
8.5	The confusion matrix when the signal representation consisted of <i>AP_δ1</i> and the observation distribution given an HMM state was Gaussian. All number are in percentage.	159
8.6	The confusion matrix when the signal representation consisted of <i>AP</i> and the observation distribution given an HMM state was a mixture of 8 Gaussians. All numbers are in percentage.	160
8.7	The confusion matrix when the signal representation consisted of <i>AP_δ1</i> and the observation distribution given an HMM state was a mixture of 8 Gaussians. All numbers are in percentage.	161

8.8 Recognition results using 8 mixtures. First column, training done with speech produced by males and recognition done with speech produced by females. Second column, training done with speech produced by females and recognition done with speech produced by males. . . . 165

Chapter 1

Introduction

Speech is the natural means of communication among humans. However, the medium of communication between machines and humans has been largely limited to keyboards and CRT displays. Thus, there is a great desire for enabling machines to recognize human voice, to understand it and, whenever appropriate, to respond back using speech. Embedding these capabilities in machines has been the center of research for many decades. This research can be divided into four main areas: (1) speech synthesis, (2) speech recognition, (3) speech understanding and (4) speech generation. The area of speech synthesis has progressed much faster than the other three. Speech synthesizers of high intelligibility such as Dectalk and Klattalk [1] have appeared in the market place since the mid 80's. In contrast to speech synthesis, speech recognition products have not appeared in the market until recently. Examples of these products are the Dragon Dictate which is a discrete word recognition system, customer designed applications such as fill-in-blank reports by voice, discrete digit recognition over the telephone line implemented by AT&T and oth-

ers as well as the voice-operated Apple Macintosh. However, these applications are very limited and speech recognition technology is still far from allowing a human to converse freely with a machine. This is not to undermine the fact that this area has evolved tremendously in the last decade. Today, speaker-independent continuous speech recognition systems can achieve word recognition accuracy near 95% on read and quiet speech involving a 5 k-word closed vocabulary (results reported on 5-K word WSJ task [2]). The trend is moving towards more challenging tasks involving larger and open vocabulary as well as casual speech, as opposed to read speech, in addition to dealing with speech in noise. These advances in speech recognition are mainly due to a better understanding of speech and, more importantly, to improved statistical modeling methodologies coupled by the availability of large training corpora. The best performing systems today utilize the Hidden Markov Model (HMM) as the basic structure and in some cases they include the Stochastic Segment Model (SSM) [3] in multistage recognition and Neural Networks (NN) in a hybrid approach (c.f. [4] [5]). These structures are used to model the acoustic manifestation of English sounds and/or to model the language using statistical training methods. It is notable, however, that the performance of these systems tends to sharply degrade on tasks that involve casually articulated speech such as the switchboard database where reported word recognition rates are about 60 – 70%. Furthermore, moving across databases usually requires a lot of tuning and training to achieve good performance.

Despite the advances of the last decade, there are still outstanding challenges that call for further research at many levels of the speech recognition process. There is now an effort to evaluate the contributions of different acoustic modeling methodologies in order to gain an understanding of their advantages and disadvantages. Further-

more. there is general understanding that dealing with more challenging tasks such as casual non-read speech requires a better understanding of speech variability and coarticulation phenomena [6]. Perhaps, one of the main disadvantages of the approaches employed by current systems is that little can be gained in terms of our understanding of the speech process since the relationship between a phoneme and its acoustic manifestation is not represented explicitly. In this thesis, we deal with acoustic modeling of speech for speech recognition. However, we undertake the acoustic phonetic approach whereby speech units are modeled explicitly. The objective of this research is threefold: (1) develop a signal representation that consists of acoustic parameters that target linguistic information represented by phonetic features in the speech signal. (2) explore the viability of a new speech recognition paradigm based on phonetic features and acoustic events. (3) use this paradigm as a tool for speech analysis so that a better understanding of the acoustic manifestation of sounds as well as a better understanding of coarticulation can be gained, and (4) integrate and test the developed signal representation in the HMM framework.

In Chapter 2, the different methodologies used in acoustic modeling are briefly discussed in order to motivate the research presented in this thesis. The phonetic feature theory that forms the backbone of this approach is briefly reviewed in Chapter 3. In Chapter 3, the undertaken approach to acoustic modeling is also motivated and discussed. In Chapter 4, the database used in this thesis is discussed and justified. In Chapter 5, acoustic parameters that target manner-of articulation phonetic features are presented. The performance of these parameters in an event-based paradigm for manner-class speech recognition is presented, analyzed and compared to the performance of a slightly adapted form of the same parameters in an HMM framework and

to Mel-cepstral parameters in the HMM framework. In Chapter 6, a procedure based on objective criteria is defined to automate the derivation of acoustic parameters that target phonetic features. This procedure was used in Chapter 7 to derive acoustic parameters for the phonetic features: sonorant, syllabic and strident in addition to the anterior phonetic feature that distinguishes among strident fricatives and the labial, velar and alveolar places of articulation that distinguish among the stop consonants. The performance of these parameters was evaluated in classification tasks. A subset of the derived parameters was used in a manner-place recognition task within the HMM framework. The performance of these parameters was compared to that of Mel-cepstra using the HMM recognition paradigm. The place-manner recognition experiments are presented in Chapter 8. Finally, the main conclusions from this research are summarized in Chapter 9 and directions for future research are suggested.

Chapter 2

Acoustic Modeling of Speech:

Background

This chapter is a review of the main issues encountered in the acoustic modeling of speech and of the approaches that have been taken in addressing this problem: the Hidden Markov Model (HMM), the Stochastic Segment Model (SSM) and the acoustic-phonetic approach. The assumptions made in each approach as well as the advantages and the disadvantages of these methodologies are discussed.

2.1 Issues in Acoustic Modeling

A complete speech recognition system is shown in Figure 2.1. Many knowledge sources contribute to the recognition process of a received speech utterance. One of these knowledge sources is the acoustic model. The acoustic model is an integral part of a recognition system since it provides the link between the acoustic signal and the recognition lexicon represented in terms of linguistic units. Acoustic modeling of

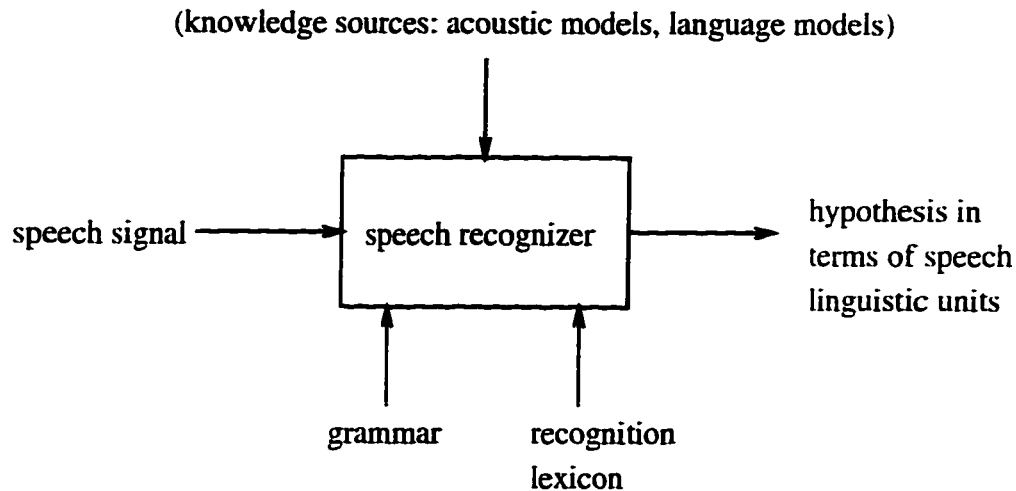


Figure 2.1: A diagram depicting the different knowledge sources contributing to the recognition of a spoken utterance.

speech is mainly concerned with the following three issues:

- Choice of speech units to be modeled.
- A signal representation that appropriately captures speech acoustic properties. Such a representation is traditionally referred to as the feature space.
- A modeling methodology or structure that allows the mapping from the signal representation to the selected speech units.

Choice of speech units: Speech can be described as a hierarchy of different linguistic units. For instance, phonetic features form phonemes (sounds of the language) which combine to form syllables, then words, phrases and finally a discourse. The question of what constitutes minimal speech units is a theoretical one that puzzled linguists as well cognitive scientists for decades. Minimal speech units ranging from phonetic features up to words have been considered. However, there is a general con-

sensus in speech recognition research that phonemes are the speech building blocks although several others have argued for the phonetic features [7] [8] and syllables[9]. The choice of minimal speech units for the purpose of speech recognition by machine is limited for three reasons. First, the minimal speech units must be chosen so that they can uniquely describe lexical items which are usually words. Second, the chosen speech units must allow for the generalization of the different phonological processes that take place in fluent speech. Finally, it must be possible to develop reliable acoustic models that sufficiently describe the acoustic manifestations of these speech units from a limited amount of data. In this regard, words best fit the first constraint but fail the second and the third. On the other hand, phonemes may satisfy the first and third requirements but fail the second one¹. However, we observe that phonetic features satisfy all three requirements equally well.

Signal representation: There are a variety of acoustic properties that characterize the different sounds of the language. Some of these properties are more apparent in the time domain while others are more evident in the frequency domain or in the evolution of the frequency content as a function of time. Therefore, time-frequency distributions that simultaneously describe the time-frequency structure of speech signals have been used. However, the acoustic structure of the different sounds impose different requirements, that are often conflicting, on the time-frequency representa-

¹Triphone models try to capture the effect of phonological processes on the acoustic realization of phonemes. However, two limitations of such models are that (1) they only capture the effect of the left and right phonemes on the modeled phoneme although a phonological process may spread over more than one phone and (2) they do not model or gain advantage of the fact that phonological processes act at a sub-phonemic level rather than at the phoneme level.

tion. For instance, the stops and affricates are realized with a fast transient that occurs over a short time interval. Thus, the detection of such an event requires time resolution that is finer than a pitch period (8 to 10 ms for a typical male). On the other hand, vowel sounds (e.g., /a/) are quasistationary sounds which share the periodicity attribute. Detecting periodicity requires analysis over a time interval that spans two or more pitch periods so that the harmonics can be resolved. The requirements of fine time-resolution and fine frequency-resolution cannot be simultaneously satisfied as asserted by the Heizenberg's uncertainty principle. This uncertainty principle states that the frequency bandwidth of a signal is inversely proportional to its duration. Therefore, two frequency components of a signal can be resolved if the durations of both components are sufficiently large so that their bandwidths do not smear. Different time-frequency representations that provide different tradeoffs in time-frequency resolution have been proposed [10]. Among the time-frequency representations are the short-time Fourier transform (STFT) [11], Wavelet transform [12] and Wigner distribution [13]. Moreover, models of the peripheral auditory system that take into account speech characteristics have been developed and applied in speech research [14]. The most widely accepted signal representation for speech recognition today consists of Mel-warped cepstral coefficients. This signal representation has produced the best recognition results so far. The Mel-warped cepstra are obtained by first computing the STFT of the signal and then computing the energy within Mel-warped frequency bands at each time-frame [15]. The cepstral coefficients at each time-frame are obtained from the discrete-cosine-transform (e.g., [16]) of these energy parameters. Mel-warping accounts for the human auditory system that has high frequency resolution in the low frequency bands and poor frequency

resolution in the upper frequency bands. The cepstrum transformation results in a signal representation with independent components or no redundancy.

Spectrogram reading experiments coupled with the waveform suggest that the STFT is sufficient to extract the acoustic properties of the different speech sounds ². Thus, the STFT is the transformation of choice in the current research. Acoustic properties that correspond to the different sounds are extracted from the STFT and/or the time-waveform.

Model Structure: The model structure is a link between the lexical-representation component and the signal-representation component of a speech recognition system. The most popular structures employed in today's speech recognition systems are the HMM, SSM and Neural Networks. The advantages and disadvantages of these structures are discussed in Section 2.2.1. In this thesis, the HMM is utilized in some experiments. In addition, rule-based models of phonemes in terms of phonetic features are developed. These models make explicit use of speech knowledge as discussed in Section 2.2.2.

2.2 Acoustic Modeling Methodologies

In acoustic modeling, two approaches have been undertaken: the mathematical modeling approach and the acoustic-phonetic approach. In this section, these two approaches are briefly reviewed illustrating their advantages and disadvantages.

²A spectrogram is a two-dimensional representation of the STFT with frequency on the vertical axis, time on the horizontal axis and amplitude is represented by a grey scale or a color scale.

2.2.1 Mathematical Models

The most popular and widely adopted structure in acoustic modeling is that of the Hidden Markov Model (HMM) [17]. In this structure, a speech unit is represented by an automatum (a sequence of states). Transitions between states are governed by transition probabilities and are only allowed left-to-right. The mapping between the acoustic representation of the speech signal and the HMM is done through observation probabilities conditioned on the state only. That is, the acoustic model assumes that the consecutive speech frames are conditionally independent given the state sequence and therefore uncorrelated. This unrealistic assumption represents a major drawback in the HMM framework since the correlation between adjacent time instants is evident by the continuous motion of the articulators. Thus, the HMM does not model this aspect of the speech process despite the fact that it is the most successful in speech recognition applications by machines. To have a model for a speech unit, the different probabilities involved have to be obtained. In this regard, issues that relate to the type of distributions (parametric versus non-parametric, Gaussian or Laplacian, etc.) and the estimation of these distributions have been addressed. Due to mathematical ease, Gaussian distributions have been used. One of the major research issues has been whether unimodal multivariate Gaussian is better than multimodal (mixture) Gaussian. Related to this decision is the amount of data needed to train these models (i.e., estimate the different parameters). It has been found that when enough data is available, the latter results in better acoustic models which yield higher recognition performance. To deal with the problem of data requirement, tied mixtures have been used.

The Stochastic Segment Model (SSM) was proposed in [18] to model feature trajectories that persist throughout a speech unit such as a phone. The advantage of the SSM is that, in contrast to the HMM framework, it allows capturing the correlation between consecutive speech frames. In the SSM framework, there are issues common with the HMM that are related to the type of observation distributions and their estimations. However, there are also issues specific to the SSM such as model duration and the correlation between the consecutive micro-segments that constitute the SSM³. Since phonemes have variable durations, corresponding models may have variable durations, or number of microsegments, proportional to phone durations. Alternatively, SSM's of all phones may have the same number of microsegments with linear-time warping used in mapping the speech frames to the microsegments. The correlation between consecutive micro-segments can be captured with a probability distribution of the SSM that has a non-diagonal covariance matrix. Such a probability distribution, however, requires a lot of training data and is rarely used in practice. A disadvantage of the SSM is that it is computationally expensive. This computational expense prohibits the use of the SSM as a framework for word recognition. In word-recognition applications, the SSM has successfully been used as a second-pass within the N-best rescoring paradigm [3]. In this paradigm, the top N candidate sentences that correspond to a spoken utterance are obtained using an HMM system. Each of these sentences is then rescored using SSM's of the constituent phones. The sentence that receives the highest score is chosen as the hypothesis.

In addition to HMM and SSM, other mathematical models such as neural networks

³An SSM consists of several micro-segments, usually between 5 and 8, where each observed speech frame is mapped to any of these segments based on the linear time-warping procedure.

(NNs) have also been used. NNs have proved effective when used along with the HMM in a hybrid framework [5] [4]. an NN consists of layers with several nodes in each layer. Each node in the input layer corresponds to a dimension in the acoustic-feature space (e.g., a component of a vector of Mel-frequency cepstra). Each node in the output layer corresponds to a phone being modeled. Nodes from one layer are usually connected to each node in a successive layer. The problem lies in estimating the outputs of each layer and the weights on the connections between nodes in successive layers. In the hybrid framework, NNs are used to estimate the observation probability of an HMM state whereas the HMM captures the time evolution of a phone. NNs offer the advantage of estimating complex observation probabilities and eliminate the need for making assumptions on the shape of these probabilities (e.g., Gaussian or multi-model Gaussian, diagonal covariance matrix or not). These observation probabilities are estimated in the training stage so that the error of guessing the wrong phone based on that observation phone is minimized. Correlation between consecutive frames can be implicitly modeled by including as input to the NNs, speech frames that surround the speech frame whose probability given the HMM state is being estimated. A disadvantage of NNs is that they are expensive in training as they take long time to converge. In addition, an assumption needs to be made regarding the number of nodes at each stage of a NN and on the number of stages (usually, two intermediate stages between the input and output stages are assumed).

Vital to all of the acoustic models is the set of features used in modeling and recognition. In this respect, all statistical systems that we know of utilize Mel-cepstra and cepstrum-related coefficients. Several other signal representations have been tried, keeping everything else fixed, such as the Fourier transform and auditory transforms.

The Mel-cepstral coefficients yielded the best results and therefore became the standard. That is not to say that other representations are not as good since other factors such as the type of observation distributions may affect performance. There is still ongoing work on auditory models and their applications in speech recognition (c.f. [14], [19], [20], [21]). In addition, other methods for computing a signal representation for speech recognition are still explored (c.f. [22]).

In computing cepstra, a 10-30 ms segment of the received speech signal is transformed to the cepstral domain every 5-10 ms (frame rate) resulting in a sequence of cepstral vectors regularly spaced in time. To capture some of the correlation between adjacent frames and the dynamics of the speech process, these vectors are augmented by the time derivatives of the cepstral coefficients and the energy profile. The cepstral coefficients capture the energy concentration in some frequency bands relative to others (e.g., the first coefficient measures the high-frequency energy concentration relative to that of the low frequency). The time difference of the cepstral coefficients captures the change in time of these different energy concentration measures. Once the cepstral vectors are computed, they are used as observation vectors.

In modeling, these observation vectors are used to estimate the different model parameters. Sophisticated training techniques and large sets of training data are relied upon to automatically capture, from these cepstral features, the characteristics that discriminate the different speech units. This is one of the cited advantages of these statistical modeling techniques since they offer the advantage of automatic training and optimization while avoiding the need of acquiring and explicitly modeling acoustic-phonetic knowledge. Knowledge about speech is mainly manifested in the training stage when acoustic models are built. For instance, knowing that the acoustic

manifestations of phones vary depending on the context. a separate model is built for a phone depending on what phone precedes it and what phone proceeds it (triphone model)⁴. In recognition, the models are searched for the most likely one that may have produced the observed cepstral values.

2.2.2 The Acoustic-Phonetic Approach

The acoustic-phonetic approach to speech recognition relies on explicitly modeling our knowledge of the speech signal. This knowledge is reflected in determining the set of features to be extracted, the signal processing tools and methods to extract these features and the decision making that interprets the extracted features in terms of speech events or units. Thus, the acoustic-phonetic approach draws from many areas concerned with human speech such as linguistics and speech perception. Linguistics is concerned with language organization and the determination of the aspects of the speech signal that deliver the intended linguistic message. Particularly, phonology, a branch of linguistics, deals with sound categorization into abstract units, the relationship between these units and with modeling the different phonological processes that result from coarticulation. In addition, phonetics, another branch of linguistics, deals with the motor activity involved in the production of sounds. Speech perception studies are concerned with the psycholinguistic aspects of the signal and with the signal representation in the human auditory system. Motivated by studies on the most powerful speech recognition/production system to date, which is the human

⁴Clustering algorithms are used to reduce the number of triphone models by pooling several context-dependent models of a phone into one. Clustering is automatic but usually takes advantage of linguistic knowledge in the form of hand-specified candidate questions

being. early researchers in speech recognition have advocated the application of the acoustic-phonetic approach. This approach dominated the area of speech recognition from the mid 50's (earliest work in speech recognition) to the late 70's [23]. but it lost ground to the mathematical-based approach (e.g., HMM) mostly pursued today for reasons that will be discussed later in this section. Rather than reviewing the history of speech recognition, the problems that limited the acoustic-phonetic approach will be discussed to motivate this research.

The acoustic-phonetic approach to speech recognition relies on two premises. The first premise is that each sound of the language or phone can be described by a bundle of abstract linguistically-distinct features [24]. The second premise is that each of these features has an acoustic signature in the speech signal. As the linguistic feature theory was evolving, Halle and his colleagues adopted the principle of linearity and researchers in speech recognition adopted the principle of invariance. The linearity principle suggests that words are made up of concatenated phonemes. The invariance principle suggests that a specific set of features describing a sound must always be present (invariant) so that the specific sound is perceived as such. These two principles drove early researchers in speech recognition and speech sciences. Thus, a large effort was dedicated to detecting these **invariant** distinctive features. Moreover, segmentation and labeling of speech intervals, in terms of phonemes, as implied by the linearity principle was the center of any acoustic-phonetic strategy prior to lexical access. A lexicon usually consists of words each of which is described by a sequence of phones. Lexical access is the mechanism by which a word from the lexicon is selected.

In their early investigations, researchers were looking at limited problems such as recognizing some vowels in fixed phonetic contexts (e.g., CVC) or recognizing words

from a small vocabulary. Limited data was available and the task of data analysis was very time consuming and laborious due to many technological limitations. Bound by these restrictions, acoustic-phonetic knowledge was far from being complete. Thus, success achieved with the acoustic-phonetic approach dealing with a small vocabulary and a limited number of speakers, became a disappointment when more ambitious tasks (SUR ARPA project [23]) were undertaken. It became evident that a spoken word is more complex than the mere concatenation of phonemes so the linearity principle does not hold. Coarticulation, sound deletion and other phonological processes appeared to be the major obstacles arguing against the existence of speech invariant units. As a result, a better understanding of speech was gained and lists of phonological processes were compiled by the completion of the SUR project. However, in the SUR project, more emphasis was put on the use of higher speech knowledge (syntax, semantics and pragmatics) and on developing appropriate system architectures that control the use of the different knowledge sources for speech recognition. Consequently, acoustic-phonetic research was not the main focus. Around that time period, HMM and pattern-recognition approaches started emerging as powerful techniques for speech recognition [25] and emphasis was on developing automatic learning algorithms rather than learning more about the acoustic manifestation of speech sounds. So signal processing was kept simple (cepstra), as opposed to the detailed signal analysis required for phonetic-feature extraction.

What are the reasons that caused the acoustic-phonetic approach to lose ground to the mathematical-based approach despite the former's appealing ideas? This question has not been rigorously addressed. However, several reasons are cited that lure people away from taking the acoustic-phonetic path, especially when comparing it

to the mathematical modeling strategy. To build acoustic-phonetic front ends, an extensive knowledge about the speech signal has to be collected a priori. Such knowledge acquisition is time consuming and, to date, it is argued to be at best incomplete and at worst unavailable (e.g., how to recognize devoiced vowels). Moreover, it has been argued that the choice of features and the design of classifiers (decisions deduced from features) are based on ad hoc considerations and methods. That is, there are no optimality criteria in selecting the acoustic parameters or designing the classifiers. Moreover, segmentation and labeling that used to be at the heart of the acoustic-phonetic approach proved to be the center of the problem since they assume that speech is a linear process. In this research, each of the cited problems of the traditional acoustic-phonetic approach is addressed. However, more emphasis is given to the design of acoustic parameters that combine acoustic-phonetic knowledge and statistical methods.

Chapter 3

Phonetic Features and Acoustic Modeling

The approach to acoustic modeling and analysis of speech, pursued in this research, is based on the linguistic theory of phonetic features. Phonetic features are directly related to the way that humans phonate and perceive speech as briefly described in Section 3.1. Furthermore, phonetic features can be described in terms of acoustic properties, as discussed in Section 3.2, that make them suitable speech units for acoustic modeling. The adopted methodology to extract these acoustic properties from the speech signal is addressed in Section 3.3. In Section 3.4, the fuzzy logic and probabilistic approaches are proposed to model the phonetic features in terms of the acoustic properties. The fuzzy logic approach is explored in Chapter 5, whereas the probabilistic approach is proposed here as an alternative although it was not explored. These approaches allow modeling the uncertainty that arises in interpreting the speech signal in terms of the designated phonetic features.

3.1 Phonetic Features: an overview

Phonetic features describe the phonetic capabilities of the human being. A phonetic feature system that sufficiently describes the different sounds of languages was first described by Jakobson *et al.* [24] and later modified by Halle and Chomsky [7]. These phonetic features mainly capture the manner and place of production of the different phonemes and they were suggested as appropriate units for lexical representation based on phonological considerations and across language studies. According to Halle and Chomsky, a phoneme is an abstract symbol that labels a specific feature vector whose different components are assigned binary-values (“+” if a feature is present and “-” if it is not). This role of phonetic features is referred to by Halle and Chomsky as the classificatory role which is used in the representation of lexical items. If lexical items are words, each word will thus be represented by a sequence of phonetic feature vectors, each of which corresponds to a phone in the word. There are different inventories of linguistic features motivated by different purposes (e.g., [26]). However, one can usually draw the correspondence between one feature in an inventory and one or more features in another inventory. This fact suggests that the different feature inventories have the same acoustic consequences which are our main concern. Thus, we select the feature inventory proposed in [7] since it is sufficient for the description of the different English sounds. However, the traditional places of articulation for the stop and nasal consonants (labial, velar, alveolar) supplant the features coronal and anterior in describing these sounds in the feature set used in this research. In addition, we follow the lead of recent advances in phonological theory which suggests that the phonetic features are hierarchically organized [27]. The hierarchical organization of

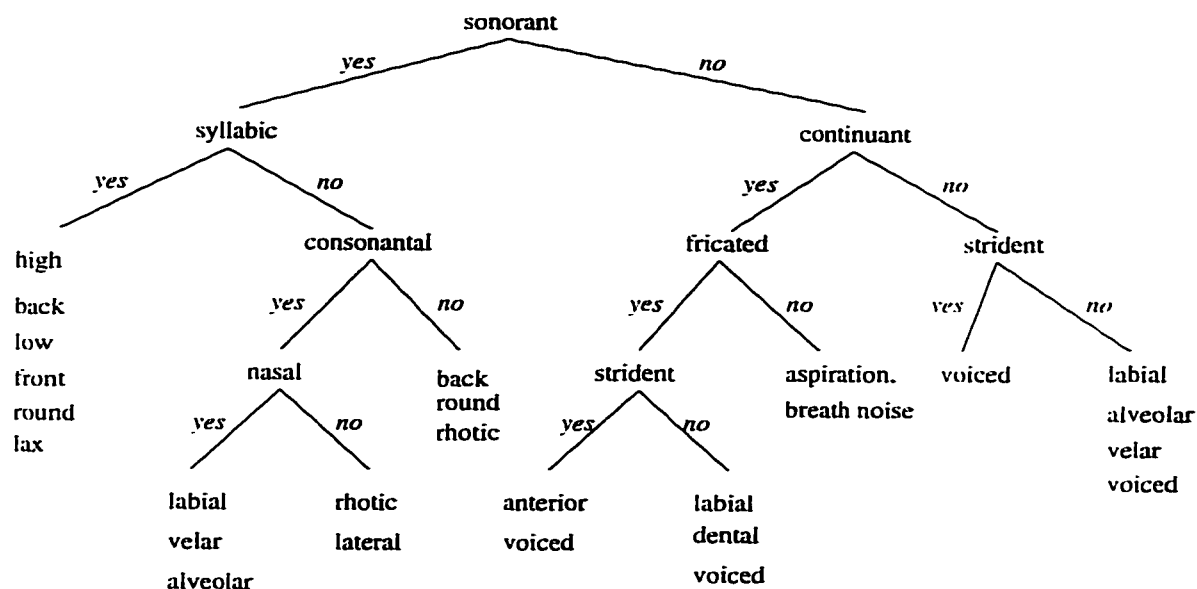


Figure 3.1: The hierarchy of phonetic feature organization.

the different features that we use is depicted in Figure 3.1. This hierarchy suggests that phonemes can be grouped into major classes that dominate the nodes in the tree as depicted in Figure 3.1. Furthermore, the hierarchy also illustrates the dependence among the features (e.g., a syllabic sound must be sonorant) and the discriminatory role that the features play at each node. The articulatory correlates of the different phonetic features are stated in Table 3.1 and Table 3.2.

3.2 Acoustic Properties of Phonetic Features

Phoneticians and researchers in different branches of speech science have been investigating the acoustic properties of the different phonetic features and phonemes. The relevant studies include: (1) perceptual studies that are concerned with the ability to distinguish between sounds [30], (2) spectrographic studies that look at

Table 3.1: Phonetic features and articulatory correlates based on Chomsky and Halle [7], and Ladefoged [28].

Phonetic Feature	Articulatory Correlate
sonorant	No pressure build up in vocal tract, allowing spontaneous vocal fold vibration.
syllabic	A peak in vocal tract opening.
high	Tongue body is raised above the neutral position.
back	Tongue body retracted back toward the pharynx and tongue tip is not in contact with lower teeth or gum ridge.
low	Tongue body is lowered beyond the neutral position so that contact is severed with the teeth.
front	Tongue body is pulled towards the front of the mouth.
round	Lip orifice is narrowed.
lax	Vocal tract is in a neutral or central position.
consonantal	Radical obstruction in the midsagittal region of the vocal tract. The obstruction is at least as narrow as that for the fricatives.

Table 3.2: Phonetic features and articulatory correlates based on Chomsky and Halle [7]. and Ladefoged [28] (cont.).

Phonetic Feature	Articulatory Correlate
nasal	Lowered velum allowing air to escape through the nose.
rhotic	oral constriction formed by either tongue tip or tongue bunching with a sublingual cavity created between the created between the underside of the tongue and the jaw.
noncontinuant	Complete blockage of air flow through the lips.
strident	Air stream is directed towards an obstruction at a great flow rate.
labial	Constriction made at the lips.
dental	Constriction is made at teeth by tongue tip.
alveolar	Constriction made at the alveolar ridge
velar	Constriction made in the velum region
anterior	Constriction in front of the palato-alveolar region.
coronal	Tongue blade is raised above neutral position.
lateral	Air passage is created around the sides of the tongue blade.
voiced	Vocal folds are vibrating.
fricated	Turbulent noise is generated in front of the constriction

the time-frequency characteristics of sounds [31] and (3) acoustic modeling studies which attempt to provide theoretical models for the production of different sounds such as the acoustic-tube modeling of the vocal tract [32] [33]. All of these studies have a bearing on the determination of the acoustic properties of the different sounds and phonetic features. A summary of the acoustic properties for the selected phonetic features is presented in Table 3.3 and Table 3.4 [34]. These acoustic properties may sufficiently describe the acoustic realization of the phonetic features although refinements of some of these properties may be needed. In this research, the acoustic properties listed in Tables 3.3 and 3.4 are used. The difficulty lies in designing acoustic measurements that extract the acoustic properties of the different phonetic features from the speech signal in a speaker-independent manner. This problem is discussed in the following section.

3.3 Philosophy in Designing Acoustic Parameters: Relative Measures

Acoustic parameters (APs) are measures performed on the speech waveform and its time-frequency transformation(s) in order to seek evidence for the acoustic properties of phonetic features. Although the acoustic properties of phonetic features are qualitatively understood, reliable acoustic-measures that provide evidence for these features are still far from being complete. Several attempts have been made to determine the best set of measures needed for the classification of different sounds or sound classes and/or phonetic features. For instance, Glass [35] lists acoustic measures that char-

Table 3.3: Phonetic features and their acoustic correlates.

Feature	Acoustic Correlate
sonorant	Periodicity and high energy at low frequency. No frication noise.
syllabic	A peak in acoustic intensity in the low to mid frequency range.
high	Low first formant.
back	Low second formant.
low	High first formant.
front	High second formant.
round	Lowering of formant frequencies.
lax	Central formant frequencies (e.g. neither low or high) Relatively short duration.
consonantal	A rapid spectrum change at the release of the consonantal configuration.
nasal	Low first formant. Energy is concentrated at low frequency. Energy losses in mid-frequency range due to antiresonances.

Table 3.4: Phonetic features and their acoustic correlates (cont.).

Feature	Acoustic Correlate
rhotic	General lowering of F3 so that it is close in frequency to F2. are close together.
noncontinuant	Rapid onset of energy across all frequencies.
strident	High intensity of turbulent noise.
labial	Lowering of formants. Flat or falling spectrum.
dental	High third and fourth formants.
alveolar	Rising spectrum.
velar	Energy concentrated in the middle part of the spectrum
anterior	Energy is concentrated in the high frequency part of the spectrum.
coronal	Dominant high frequency energy.
lateral	High F3 and low F2. Formants higher than F3 are considerably reduced in intensity.
voiced	Periodicity. Longer duration of vowels preceding voiced consonants.
fricated	High zero crossing rate. Concentration of energy in high frequency bands.

acterize nasal sounds and differentiate them from voice bars and semivowels. Chen [36] lists a set of acoustic measures that distinguish between the different phones that occur in the set of digits. These acoustic measures are related to the acoustic properties of phonetic features that distinguish among the underlying phones. Lahiri *et al.* [37] proposed acoustic measures that distinguish between the labial and alveolar stops. These measures for separating labial and alveolar stops were later modified by Zierten and Espy-Wilson [38] [39]. In addition, phonetically-motivated acoustic-measures that distinguish among the semivowels were developed by Espy-Wilson [40] [41] in the context of a semivowel recognition system. Small databases obtained from a few speakers were usually used to derive these different measures. Furthermore, these measures were not optimized [37] and [38].

A major thrust of this thesis is the development of acoustic measures that accurately extract the different acoustic properties for phonetic features. For speaker-independent speech recognition, the design of these acoustic measures must take into account the intraspeaker and interspeaker variability which can greatly affect the speech signal. Variability is used here to refer to the large number of ways in which a speech utterance can be acoustically realized. Interspeaker variability is due to physiological differences among speakers. For instance, it is well known that the formant frequencies are inversely related to the length of the vocal tract. Consequently, males have lower resonances than females. On the other hand, intraspeaker variability can arise because of changes in the emotional state of the speaker. For instance, the pitch of a person is higher when he/she is in an angry state relative to a normal state. Change in pitch may effect the measurements of periodicity and the spectral balance of the person's speech such as energy at low frequency relative to energy at high

frequency (these acoustic measurements are used to detect the sonorancy feature). A change in a speaker's pitch can be thought of as resulting in another person's voice. Such variability seems to have minimal effects, if any, on the ability of a human listener to decode the acoustic signal and recognize what has been said. Therefore, the acoustic measures must be designed in a way that diminishes the effects of the extralinguistic information on the recognition process. We postulate that this can be achieved by making each acoustic measure relative to a reference value of that measure in time and/or frequency [40] [42]. Furthermore, making relative measurements provides a way to capture the long and/or short term correlation in the speech signal in a natural manner, a weakness in current approaches to speech recognition.

One of the most important characteristics of the undertaken approach lies in explicitly making acoustic measurements that correspond to linguistic units in a manner that follows the phonetic-feature hierarchy. Such explicit measurements can be used to learn about the acoustic signature of a phonetic feature in the speech signal. Furthermore, the feature hierarchy takes into account the dependence among the different phonetic features. For instance, acoustic measurements for syllabicity are carried out in sonorant regions only. Stridency is sought in frication regions and so on. Governing this property extraction strategy is the concept of acoustic events, established by the manner features, that constitute landmarks around which additional acoustic measures can be made. It has been observed [40] [8] [43] that such events are marked by changes in the values of one or more of the acoustic measures from low to high or vice-versa, and often correspond to changes in one or more of the phonetic features. Thus, these acoustic events which occur hierarchically, as implied by the phonetic-feature hierarchy, indicate which features should be sought next as well as

how and where additional measurements should be made to extract these features. Such an approach is demonstrated within the event-based framework to manner-class speech recognition presented in Chapter 5. The conclusion from this work is that the considered phonetic features can be extracted reliably from the speech signal.

3.4 Uncertainty Modeling

The sources of uncertainty to be modeled are discussed in Section 3.4.1. Two different approaches are proposed for modeling uncertainty. The first approach is based on the fuzzy logic framework and is discussed in section 3.4.2. The second approach is based on the probabilistic framework and is discussed in section 3.4.3.

3.4.1 Sources of Uncertainty in the Speech Signal

The sources of uncertainty in the speech signal stem from variability in speech production and they include:

- Physiological differences among speakers
- Changes in the emotional state and/or physical status (sickness) of a speaker.
- Dialect effects.
- Lenition: undershooting in the articulation of sounds such as weakening of constrictions in consonant production.
- coarticulation: temporal overlap of adjacent phones.

- Allophonic variations: systematic context-dependent variation in a phoneme (e.g., an allophone of a word final /t/ is a glottal stop) that does not change the meaning of a word.

Design of relative acoustic measurements was suggested in section 3.3 as a way to diminish the effects of the first two types of variability. These acoustic measurements are used to decide upon the presence or absence of a phonetic feature as implied by the phonetic feature hierarchy. The remaining sources of uncertainty, however, complicate the decision making process. For instance, it is predicted from the lenition process that phonetic features are realized with different degrees of strengths which affect corresponding measurement values. Consequently, the decision about the realization of a feature becomes clouded with ambiguity or uncertainty. Two approaches, the fuzzy logic approach and the more popular probabilistic approach are methods that can be utilized to model such sources of uncertainty.

3.4.2 Fuzzy Logic Framework

Motivation:

The proposition of this approach is motivated by the reasoning of spectrogram reading experts who attempt to decode a spoken utterance represented by its spectrogram into phonemic units. Such an expert looking at a spectrogram will use clauses such as: there is a *strong* dip in energy in this region and there is a *sharp* onset at this time instant. In doing so, the expert uses the labels: *strong*, *sharp* etc. to describe speech events as he/she sees them in the spectrogram. The question that one asks is how strong is strong and how weak is weak? There are no sharp boundaries that define

an interval within which the value of a measurement (e.g. energy dip) is strong and outside of which it is not. There are measurement values that are definitely strong and others that are definitely weak, but there are those cases where no such firm decision can be made. In such borderline cases, one tends to attribute the symbol strong with more confidence as the measurements values move closer to the definitely-strong region. Thus, through this argument, one can see that there is vagueness or “fuzziness” in assigning linguistic values (strong, weak, etc.) to a set of numerical values. This fuzziness is a source of uncertainty. “Fuzziness seems to pervade most human perception and thinking processes” as Parade and Dubois have stated [44]. The fuzzy set theory developed by Zadeh [45] and later expanded by others [44] [46] is a tool designed to deal with such fuzziness that arises in human centered systems.

What is a fuzzy set? A fuzzy set is a linguistic label assigned to a set of objects in a discourse where the boundaries of the set are not well defined. Thus, the elements of the fuzzy set are assigned membership values that describe the compatibility of that element with the linguistic label. Traditionally, the value 1 is assigned to those elements which are definitely compatible with the label, 0 to those which are definitely not compatible with the label and values between 0 and 1 for all other elements. The more an element is compatible with the fuzzy set label, the closer to 1 is its membership value. Thus, a fuzzy set A defined on a universe Ω is completely described by the pairs

$$A = (x, \mu_A(x)), x \in \Omega \quad (3.1)$$

where $\mu_A(x)$ takes values in $[0, 1]$. The membership function thus reflects an ordering of the elements of the universe with respect to the linguistic label. In addition to the

fuzzy set theory. there is a fuzzy logic framework defined on fuzzy sets that allows the representation of the expert's linguistic information. We make use of this fuzzy logic framework in the current research.

Fuzzy modeling of a phonetic feature:

A phonetic feature F_i is characterized by a set of acoustic properties (e.g. strong low-frequency energy) $F_{ij}, j = 1, 2, \dots, l$. Each of the acoustic properties, F_{ij} , has an acoustic measurement procedure m_j and a predetermined membership function that assigns to each measurement value of m_j a degree of compatibility with the property F_{ij} . This membership function will be referred to as $\mu_{F_{ij}}(m_j)$, the compatibility of the j^{th} measurement value with the j^{th} acoustic property of the phonetic feature F_i . The fuzzy model of F_i is given by:

$$\mu_{F_i} = f(\mu_{F_{i1}}(m_1), \mu_{F_{i2}}(m_2), \dots, \mu_{F_{il}}(m_l)) \quad (3.2)$$

where μ_{F_i} is a measure of the degree of compatibility of feature F_i with the considered acoustic pattern. In equation 3.2, F_i is regarded as a fuzzy variable whose membership function is some logical expression, f , that aggregates the membership functions of the acoustic properties in some manner. The function f uses the \wedge and \vee operations defined between two fuzzy variables x and y with respective member functions μ_x and μ_y by:

$$x \wedge y = \min(\mu_x, \mu_y)$$

and

$$x \vee y = \max(\mu_x, \mu_y).$$

Each phonetic feature F_i is thus modeled by a fuzzy rule μ_{F_i} of the form shown in equation 3.2. During phonetic feature recognition, the objective is to assign a phonetic feature to an acoustic object, p , that may consist of one or more speech frames. Thus, the compatibility of p with each of n competing features is computed using the feature models. The feature F_k such that

$$\mu_{F_k} = \max_j \mu_{F_j}, \quad j = 1, 2, \dots, n. \quad (3.3)$$

is assigned to the pattern p .

3.4.3 Probabilistic Approach

The problem at hand is that of sequential decision making represented by the hierarchy of phonetic features. At each node in the hierarchy, a hypothesis is tested regarding the value of the phonetic feature(s) at that node. Accordingly, this problem can be formulated using the traditional Bayesian approach.

Let F_1 and F_2 be the two competing acoustic features at a given node in the hierarchy. Let M be the set of acoustic measurements m_1, m_2, \dots, m_d relevant to the considered phonetic feature. Each measurement can be considered to be a random variable drawn from a probability distribution conditioned on the phonetic feature. These acoustic measurements form a d -dimensional random vector. The probability distribution for each measurement vector given the feature, $P(M/F_i)$, can be estimated a priori from training data. The probability distributions constitute the phonetic feature models. During recognition, a feature can be assigned to the acoustic object p in the following manner:

$$\begin{aligned} \text{if } P(M/F_1) * P(F_1) > P(M/F_2) * P(F_2), & \quad \text{decide } F_1 \\ \text{if } P(M/F_2) * P(F_2) > P(M/F_1) * P(F_1), & \quad \text{decide } F_2 \end{aligned}$$

Chapter 4

Database

The TIMIT database [47] was used throughout this research. TIMIT is a speech corpus developed by the Massachusetts Institute of Technology (MIT) and Texas Instrument (TI) for phonetic studies. This chapter provides a brief description of TIMIT and discusses its advantages and disadvantages in the context of this research.

The development and evaluation of acoustic parameters that target phonetic-features can best be accomplished when the database is labeled in terms of such features. Phonetic-feature labeling identifies the acoustic realization of a sound at the phonetic-feature level capturing (1) the modifications that a phonetic feature, canonically associated with a phoneme, may incur (e.g. going from sonorant to non-sonorant) and (2) coarticulation phenomena that cause phonetic-features associated with one sound to overlap with adjacent sounds. These phenomena are not captured by phone-level labeling where each speech segment is associated with a phone. To the best of our knowledge, a speech database labeled at the phonetic feature level does

not currently exist¹. In existing databases, the phone is the smallest linguistic unit used for labeling. The choices were to either use an existing database labeled in terms of phones, relabel an existing database or to create a new one. The latter two options are time-consuming and costly as a large number of speakers are required to generate the speech material and well trained phoneticians are needed to label speech. Due to lack of resources, the choice was to use an existing database labeled at the phone level and compensate for this disadvantage in the best way possible as explained later in this chapter. Among the available phone-labeled databases, TIMIT was chosen for the following reasons: (1) it has been widely accepted in the speech recognition community as the database of choice for phoneme recognition tasks (c.f. [48], [22]) and (2) it covers a large number of speakers from many geographical regions as opposed to privately collected databases that usually feature very few speakers, all sharing the same or similar dialect.

The TIMIT database consists of 6300 sentences spoken by 630 speakers from eight major dialect regions of the United States. A speaker's dialect region was determined by the geographical area where he/she lived during his/her childhood years. As opposed to speakers from the first seven dialect regions, speakers belonging to the eighth dialect region are people who moved around the United States during their childhood. About 70% of the speakers are male and the rest are female. Each speaker "read" 10 sentences. The sentences are of three types: (1) dialect sentences labeled as "sa" sentences, (2) phonetically compact sentences labeled as "sx" sentences and

¹A database labeled in terms of phonetic features is presently being developed by Professor Kenneth Stevens at the Massachusetts Institute of Technology. However the process is a lengthy one and only 50 sentences have been labeled so far.

(3) phonetically diverse sentences labeled as “si” sentences. Each speaker read the same two “sa” sentences, which were designed to examine inter-dialectal and intra-dialectal differences among speakers across the same phonetic environment. “sx” sentences were designed to provide a good coverage of pairs of phones. There are a total of 450 “sx” sentences. Each speaker read 5 of the “sx” sentences and each “sx” sentence was spoken by 7 speakers. The “si” sentences were selected from existing text to maximize the variety of allophonic contexts. There are a total of 1890 “si” sentences. Each speaker read 3 of these sentences none of which was read by more than one speaker. Thus, TIMIT is a rich database in terms of (1) number of speakers, (2) dialect diversity and (3) gender diversity.

The TIMIT database is divided into two sets: a training set and a test set. The training and test sets are independent such that no utterance or speaker appears in both. The training set consists of 462 speakers from the eight dialect regions whereas the test set consists of 168 speakers from the eight dialect regions. In the development of acoustic parameters and training of classifiers/recognizers, only materials from the training set were used. In evaluating the performance of the acoustic parameters and classifiers/recognizers, only the “si” sentences of the TIMIT test set were used.

Speech in TIMIT was recorded in a quiet environment using a close-talking microphone and was digitized at a 16-kHz sampling rate. Trained phoneticians hand-labeled all sentences. The labels were chosen from a set of 61 symbols. The symbols consist of 52 phones, a pause label, an epenthetic silence label, 6 stop-closure labels and a label used to mark the silences at the beginning and end of a sentence.

For the development of acoustic parameters that target phonetic features and for training and testing phonetic-feature classifiers/recognizers, each occurrence of a

phone in the TIMIT database was mapped to its canonical feature representation [7]. This strategy was adopted as it is impossible to predict a priori how a canonical feature is modified and acoustically realized. Furthermore, it was expected that the analysis of errors declared in classification/recognition results would help us gain insight into the validity of the acoustic parameters and reliability of the phonetic feature classifiers, as well as insight into the effects of context on the phonetic features. For instance, a phone labeled as the fricative "v" is canonically nonsonorant. If the time segment corresponding to this phone is detected by the sonorant phonetic-feature classifier as a sonorant, an error would be declared. However, a "v" can be acoustically realized as a sonorant in certain contexts due to the lenition phenomenon as discussed in Chapter 3. This type of error analysis would also improve our understanding of the coarticulation effects at the phonetic feature level. Error analyses of this type are presented in Chapters 5, 7 and 8. On the other hand, one can argue that the use of contaminated data, i.e. a sonorant "v" considered as "nonsonorant", would negatively affect the development of acoustic parameters and that of the classifiers. We contend that these contaminating cases are very small in comparison to the canonical cases. In addition, a large number of phones that are not affected by such phenomena are lumped in the same set as those phones modified from their canonical form to develop the parameters for relevant phonetic features. For instance, "v"s, as well as other voiceless fricatives that cannot be modified to become sonorant, are lumped into the nonsonorant set. Furthermore, a good portion of the "v"s are still realized as nonsonorant. Thus, the effect of the sonorant "v"s will be minimal on the developed parameters. Results obtained in this thesis support this argument.

Chapter 5

Manner-Class Recognition Based on Phonetic Features

In this chapter, acoustic parameters (APs) that target the manner phonetic features: sonorant, syllabic, fricated and noncontinuant are derived and discussed. In addition, an event-oriented approach to speech recognition is explored. Particularly, an event-based system (EBS) that uses the APs was developed to recognize speech into the manner classes: sonorant consonant, syllabic, noncontinuant and fricated in addition to silence. This event-based system resulted in 72.8% recognition accuracy on the designated task. The confusion errors suggest that a good percentage of them could be explained by contextual variability that alters the acoustic manifestation of a phoneme from its canonical form.

The event-based approach to recognition is based on a signal representation of speech in the phonetic-feature space. This signal representation is obtained by first

making acoustic measurements on the speech signal that capture the acoustic correlates for phonetic features, and then mapping these measurements to evidence in phonetic features. The acoustic measurements themselves are obtained via acoustic parameters as discussed in Section 5.1. The mapping from acoustic measurements to a belief in the implementation of the phonetic features and the event-oriented approach to recognition are discussed in Section 5.2.

In Section 5.3. The performance of EBS is compared to that of a Hidden Markov Model (HMM) system. In this comparison, the undertaken task was also manner-class recognition. In one experiment, the signal representation in the HMM system was composed of Mel-cepstral parameters. In another experiment, the front-end in the HMM system consisted of the APs developed in this chapter. In these experiments, EBS performed better than the HMM system when only cepstral parameters (no derivatives) were used as the front-end to the HMM system. However, the HMM system performed better than EBS when the observation probability distribution given an HMM state was an 8-mixture Gaussian and the cepstral parameters were augmented by their first and second derivatives. In order to compare the APs to Mel-cepstra independent of the recognition strategy, both were separately tested in the HMM framework for the task of manner-class recognition. In these experiments, the APs performed better than the Mel-cepstra when the observation probability distribution given an HMM state was assumed to be unimodal Gaussian. However, Mel-cepstra and APs were similar in performance when they were augmented by their derivatives and 8-mixture Gaussian probabilities were assumed as the observation distributions given an HMM state. Furthermore, gender experiments were conducted whereby speech models were trained on one gender and tested on another. These

experiments showed that the APs are more robust to gender differences than the Mel-cepstra ¹.

5.1 Acoustic Parameters

Acoustic parameters (APs) are exact measures performed on the signal or its time-frequency representation to provide evidence for the acoustic correlates of phonetic features. The computation of the APs is carried out in a relative rather than absolute fashion to focus on the linguistic information in the speech signal and diminish the speaker-dependent effects (see discussion in Section 3.3). For instance, the nonsyllabic measures are intended to measure the energy minimum in a sonorant consonant relative to the energy maximum in the preceding and/or succeeding vowel. This measure accounts for the energy decrease observed in a sonorant consonant, relative to a neighboring vowel, due to the constriction in the vocal tract during sonorant consonant production. As another example, the 100-400 Hz energy measure, being normalized with respect to the maximum in that frequency band across the utterance, accounts for the fact that the sonorant speech segments involve the voicing source of the same speaker.

The benefits of capturing time-correlation are evident by (1) the improvements in recognition accuracy when *MFCC* $\delta 1$ $\delta 2$ coefficients are used as opposed to *MFCC* and (2) the fact that the different speech frames are naturally correlated when they are produced by the same speaker. However, $\delta 1$ and $\delta 2$ coefficients implicitly capture the

¹It should be noted that in all experiments, raw Mel-cepstra were used with no processing, such as cepstral-mean subtraction which may have produced better results.

correlation within only a few frames. In our computations, the acoustic parameters usually capture the correlation within a wider time-window delimited by acoustic events as will be discussed later.

In the research reported here, we determined the acoustic parameters that correspond to the acoustic correlates of the phonetic-features: sonorant, syllabic, fricated and noncontinuant as shown in Table 5.1.

The parameters were determined based on an acoustic analysis of the training set in the TIMIT database and on acoustic studies found in the literature [40] [31]. In the analysis, phones were first described in terms of binary-valued phonetic features (i.e., a feature is present or not) assuming canonical realization. Then, for each feature, all phones that have that feature marked present were clustered in one group while those with that feature marked absent were clustered in another group. The parameters that provided the best separation between the two groups, based on visual analysis of histograms, were selected. Some other parameters were determined based on visual inspection of spectrograms. The fact that this process was done manually greatly limited our ability to investigate a large set of parameters. This problem was later alleviated by the automatic parameter design procedure described in Chapter 6.

The parameter computation was carried out in an event-seeking strategy [49] guided by the feature hierarchy depicted in Figure 5.1. The APs were of two types. Some APs established landmarks that divided the speech waveform into regions. For instance, the parameters related to the feature “sonorant” were computed at every time-frame to segregate regions that are sonorant from those that are not. Other APs established landmarks that pointed up particular instants in time, specifying when the acoustic property of a feature is most evident. For example, the parameters

related to the feature “syllabic” marked energy maxima within sonorant regions. Syllabicity/nonsyllabicity parameters were computed at a 5-ms frame-rate within a 25.6-ms window. The rest of the APs were computed at a 5-ms frame rate within a 10-ms window. The energy parameters were computed with a Hamming window and speech was not preemphasized.

5.2 APs and EBS

In Section 5.1, we discussed how the computation of the APs for the manner phonetic features mark events in the speech signal. Some of the events divide the speech waveform into regions while others mark particular instants in time. These events taken together provide landmarks for the computation of additional APs related to other phonetic features. For instance, since syllabic events generally occur in the middle of a vowel region, they could be used as landmarks around which further analysis is done to capture the acoustic correlates of vowel-related phonetic features such as back, low, high, round and front. This type of event-oriented strategy is derived from spectrogram-reading experiments where experts read spectrograms by first locating landmarks and then asking different questions around these landmarks that pertain to the acoustic properties of the underlying signal. Such a strategy involves event-locating rather than segmentation of the speech signal into phoneme-size units as in the traditional acoustic-phonetic approaches. Thus, no assumption needs to be made regarding speech being composed of juxtaposed phoneme-size segments. As a result, coarticulatory effects may be handled in a straightforward manner.

In Section 5.2.1, combining the APs to provide evidence for the implementation

Table 5.1: The features, their acoustic correlates and the corresponding acoustic parameters.

Feature	Acoustic Correlate	Acoustic Parameter(s)	Parameter Property
Sonorant	strong low-frequency energy [†]	0.1-0.4 kHz	strong
		0-2 kHz energy relative to 2-8 kHz energy	strong
	periodic	Voicing probability [60]	strong
Syllabic	strong mid-frequency energy	peak in 6.4-2.8 kHz and 2-3 kHz energies	strong
Nonsyllabic	weak mid-frequency energy	dip in 0.6-2.8 kHz and/or 2-3 kHz energy	weak
Fricated	turbulent noise in mid-to-high frequency range	zero-crossing rate	high
		energy in 0-2 kHz relative to energy in (2-8 kHz)	weak
		R1: 1 st cross-correlation coefficient [‡] normalized by the zeroth	low
		dip in R1	strong
Noncontinuant	Closure followed by an abrupt spectrum change over some frequency range	<u>Closure:</u> (1) 0.2-3 kHz energy [†] (2) 3-6 kHz energy [†] (3) first normalized cross-correlation coefficient	weak weak low
		<u>Abrupt Onset:</u> sum of positive first-difference values across the STFT channels	large

[†] Each of these parameters is normalized with respect to the maximum value of the parameter across the utterance.

[‡] This parameter was not part of the fricated rule but was part of the silence detection algorithm and was chosen to detect weak fricatives.

of phonetic features is discussed. The EBS for manner-class recognition and experimental recognition results are discussed in Section 5.2.2.

5.2.1 Mapping the APs to the phonetic-feature space

The objective of mapping the speech signal to the phonetic-feature space is to explicitly extract the linguistic-bearing components of the signal while discarding the extralinguistic information. This mapping provides a signal representation for direct lexical access in a feature-based approach to speech recognition² or for modeling higher linguistic units such as phones. This is in contrast to the strategy in today's state-of-the-art speech recognition systems based on HMM where the signal representation is kept simple in terms of cepstra, while sophisticated statistical training algorithms with large bodies of training data are relied upon to filter out the extralinguistic information. Although our general approach is to access the lexicon based on the phonetic-feature representation, we believe that such a representation will also lead to a better use of training data to build context-dependent statistical models of phonemes.

The representation of a speech signal in the phonetic-feature space is obtained by mapping each set of acoustic parameters to the associated phonetic-dimension. Although phonetic features are traditionally considered binary-valued (either present or not) for the purpose of lexical representation, their acoustic correlates can be manifest with different degrees of strength. This variation is due to the speech variability

²In a feature-based approach to speech recognition, words in the lexicon are represented in terms of phonetic features.

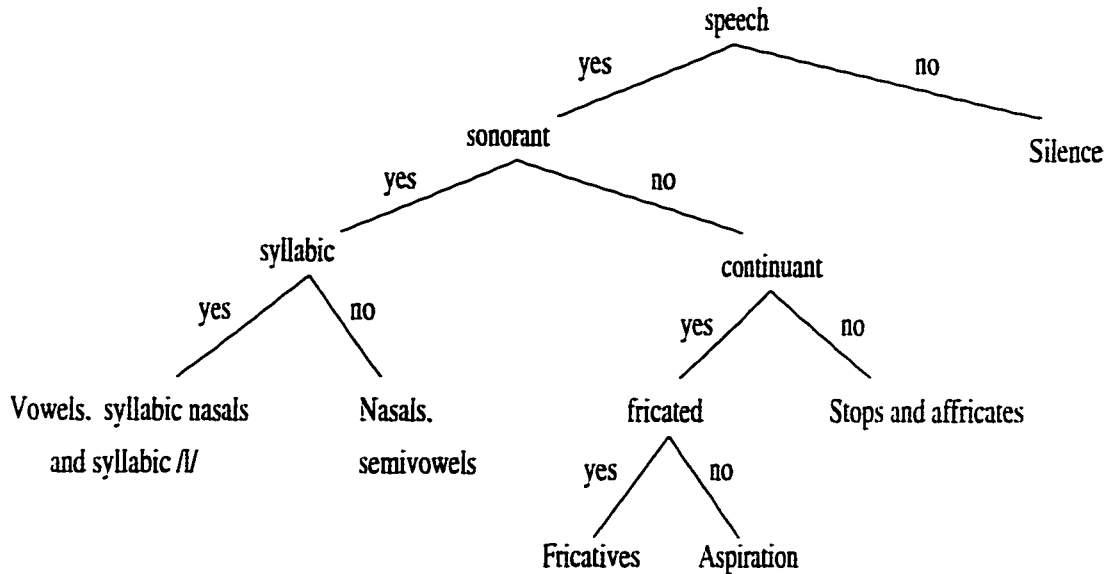


Figure 5.1: The hierarchy of manner feature organization adopted in parameter development.

that arises from the differences among speakers (e.g. dialect, physiological construct, speaking style) and from the different phonological phenomena (e.g. coarticulation) that affect the acoustic manifestation of speech sounds. Such speech variability leads to uncertainty in acoustic-to-feature mapping.

Uncertainty modeling can be handled by a probabilistic framework or a fuzzy logic framework as discussed in Section 3.4. In this work, we chose the latter because (1) it corresponds better to the reasoning of expert spectrogram-readers who use fuzzy terms (e.g. strong, weak) in order to describe acoustic properties and (2) it provides us with a direct control over the mapping which helps us gain a better understanding of the speech process at this stage.

Using the fuzzy logic framework (see Section 3.4), fuzzy rules were developed to

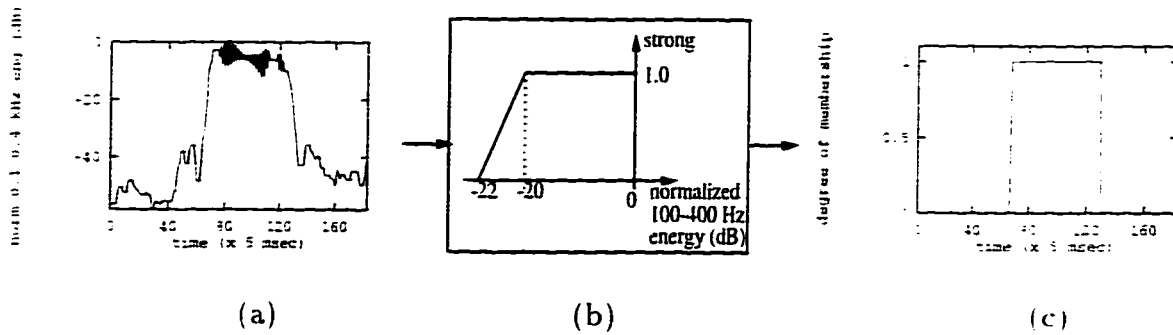


Figure 5.2: (a) parameter: normalized energy (100-400 Hz) (b) membership function for that parameter being strong (c) Membership values assigned to normalized energy.

map the APs to phonetic features. For instance, the “sonorant” feature model is:

$$\begin{aligned} \mu_{\text{sonorant}} = & ((\mu_{\text{strong}}(100 - 400 \text{ Hz energy})) \\ & \vee (\mu_{\text{strong}}(\text{voicing probability})) \\ & \wedge (\mu_{\text{strong}}(\frac{\text{energy}(0 - 2 \text{ kHz})}{\text{energy}(2 - 8 \text{ kHz})})) \\ & \wedge (\mu_{\text{not very weak}}(\text{voicing probability})). \end{aligned}$$

Each membership function has range $[0, 1]$ with a domain that spans the possible values of the considered measurement. There are several shapes of membership functions but we use the S-shape piecewise linear one as illustrated in Figure 5.2.1 (b). In this figure, we also show an example of mapping an acoustic parameter to its compatibility function with an acoustic property. The cutoff points in the membership functions, such as -22 and -20 in Figure 5.2.1, were determined by considering the measurement values obtained from the samples that canonically possess the considered phonetic-feature versus those that do not [50].

Figure 5.3 depicts an example of the representation in the phonetic-feature space for the word “amorist” using the fuzzy rules that we developed. As depicted in the

figure. we start by first segregating speech from silence using our modified version of Rabiner's end-point detection algorithm [51]. This is a binary decision, i.e., either silence or not. Then, the nonsilence regions are divided between those that are sonorant and those that are not. Sonorant regions consist of sonorant speech frames that have a degree of sonorancy of 0.5 or higher. In the sonorant regions, we seek events that indicate syllabicity and nonsyllabicity. These events are peaks and dips, respectively, detected using Mermelstein's convex-hull algorithm [52] and the frequency bands listed in Table 5.1. Note that these events are particular instants in time which indicate the extrema of syllabicity and nonsyllabicity. This is the case since it is often hard to draw boundaries between vowels and adjacent sonorant consonants. In the nonsonorant regions, the degrees of noncontinuancy and frication are measured. A noncontinuant event is detected after a silence and around the boundary of a nonsonorant region if the degree of abruptness is higher than or equal to 0.5. If such a noncontinuant event is detected, the degree of frication is measured during the remaining of the nonsonorant region if that duration is greater than 20 ms and the duration of all the nonsonorant region considered is greater than 50 ms. If no noncontinuant event is detected, the degree of frication during the nonsonorant region is measured. Based on the spectrogram and hand-labeled phonetic transcription, one can see that the signal representation accurately describes the phonetic-content of the underlying utterance.

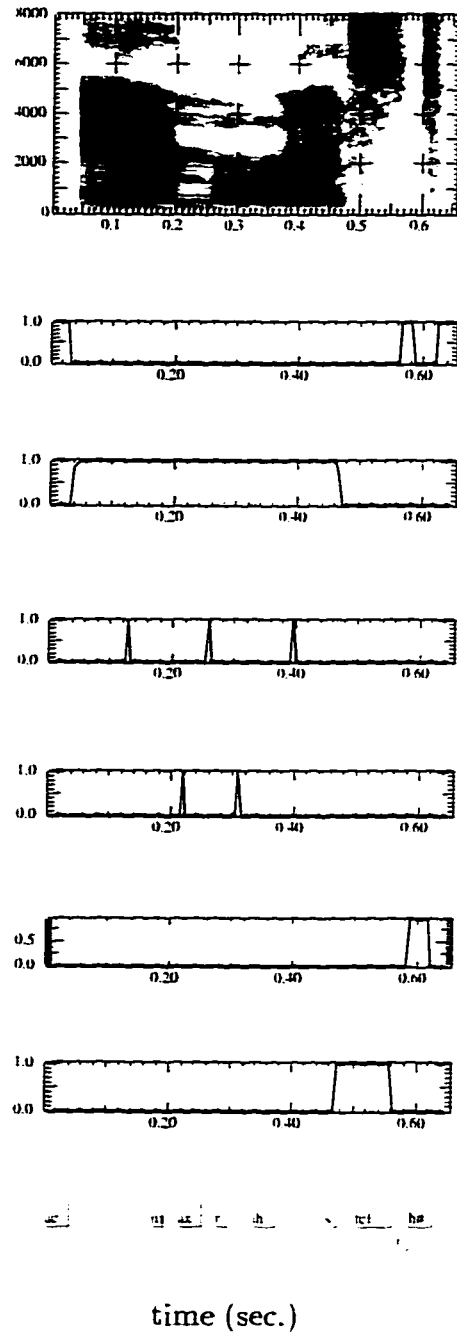


Figure 5.3: An example of a representation in the phonetic-feature space for the word "amorist". The items displayed in this figure from top to bottom are: (1) spectrogram. (2) degree of silence, (3) degree of sonorancy, (4) degree of a syllabic peak. (5) degree of a nonsyllabic dip. (6) degree of noncontinuancy (7) degree of frication and (8) phonetic transcription where a phone label appears at the beginning of the associated time-segment. Note that (4) and (5) mark particular time-instants.

5.2.2 EBS

For comparison with current approaches in speech recognition, an event-based system (EBS) was developed to recognize speech as a sequence of the manner classes: syllabic (vowels, syllabic nasals and /l/), sonorant consonant (nasals and semivowels), fricative, noncontinuant (stops and affricates) and silence (includes stop closures).

The recognition strategy of the manner-class EBS follows the feature-hierarchy of Figure 5.1. Based on this hierarchy, EBS computes a feature-based signal representation as discussed in Section 5.2.1. Then, EBS associates manner-class labels with the events detected in this signal representation. In the context of the hierarchy, the manner-class labels are associated with the hierarchy terminal nodes. At each node in the hierarchy, if the degree of belief in a phonetic feature, as determined from the feature-based signal representation, is 0.5 or higher, EBS decides that the feature is present. Table 5.2 summarizes how the phonetic features are mapped to manner-classes. For the example of “amorist” shown in Figure 5.3, EBS will produce the class sequence “silence syllabic sonorant-consonant syllabic sonorant-consonant syllabic fricative silence noncontinuant silence”.

In testing EBS, the “si” sentences from the TIMIT test set were used. For comparison, the same task was carried out by a commercial HMM recognition system [53]. A set of 3573 “si” and “sx” TIMIT training sentences was used to build manner class context-independent, 3-state, left-to-right HMM models. In doing so, each of the TIMIT phone labels was relabeled in terms of one of the manner classes based on the canonical feature descriptions of the phonemes [7]. Separate HMMs were built for the affricates and stop consonants but were both counted as noncontinuant in recognition.

Table 5.2: Mapping between phonetic features and manner classes.

Phonetic Features	Manner Classes
sonorant + syllabic	syllabic
sonorant + nonsyllabic	sonorant consonant
nonsonorant + fricated	fricative
nonsonorant + noncontinuant	noncontinuant
nonsonorant + nonfricated	aspiration (i.e., no manner class label)

The TIMIT labels /ʔ/, /h/ and /hv/ were deleted prior to training the HMMs and the flaps were mapped to the sonorant consonant class. In the HMM experiments, several HMMs were built. In one experiment, the signal representation consisted of the first 12, after the zeroth, Mel-frequency cepstral coefficients (*MFCC*) and normalized energy, *MFCC_E*, a total of 13 coefficients. The observation probability given an HMM state was unimodal Gaussian (1-mixture) with a diagonal covariance matrix. In the second experiment, *MFCC_E* was also the signal representation or front end but the observation probability given an HMM state consisted of a mixture of 8 Gaussians (8-mixture models), each with a diagonal covariance matrix. In the third and fourth experiments, the signal representation was *MFCC_E_δ1_δ2* (i.e., the *MFCC_E*'s were augmented by their first and second derivatives). However, in the third experiment, the observation probability given an HMM state was a 1-mixture Gaussian and in the fourth experiment it was an 8-mixture Gaussian with diagonal covariance matrices. *MFCC_E* was computed at a 5-ms frame rate with

Table 5.3: Recognition results comparing EBS with APs to the HMM system using *MFCC-E*. In scoring, the splits, merges and synonyms in Table 5.5 were counted as correct.

	EBS	HMM (1 mix)	HMM (8 mix)
% correct	84.6	69.6	74.5
% accuracy	72.8	65.7	69.6

a 10-ms Hamming window while speech was preemphasized with a 0.97 preemphasis coefficient.

The recognition results from these experiments are summarized in Table 5.3 and Table 5.4. In obtaining these results, the splits, merges and synonyms listed in Table 5.5 were allowed. The justification for allowing a syllabic sonorant consonant or a diphthong to split into a syllabic segment followed by a sonorant consonant is discussed in Section 7.4. An affricate, being a composite sound that consists of a stop followed by a fricative, was allowed to be split into a noncontinuant followed by a fricative. In addition, two contiguous sounds that belong to the same manner-class were allowed to be merged³, i.e. recognized as one manner-class. The EBS requires additional mechanisms to detect the occurrence of two contiguous sounds that belong to the same manner class. Such a mechanism may be based on energy change in some frequency bands.

As the recognition results indicate, EBS recognition accuracy was 7% higher than the 1-mixture HMM system and 3% higher than the 8-mixture HMM system with

³The scoring algorithm will also allow one manner class to be split into two as a result.

Table 5.4: Recognition results comparing EBS with APs (13 parameters) to the HMM system using *MFCC_E δ 1 δ 2* (39 parameters) as the front-end. In scoring, the splits, merges and synonyms in Table 5.5 were counted as correct.

	EBS	HMM (1 mix)	HMM (8 mix)
% correct	84.6	75.2	83.8
% accuracy	72.8	68.5	77.6

MFCC_E as the front-end. However, when *MFCC_E δ 1 δ 2* were used as the front-end to the 8-mixture HMM system (number of parameters was increased from 13 to 39), the accuracy of the HMM system was 5% higher than that of EBS. This result may be due to the HMM using more information, 39 parameters, than the EBS which used 13 APs reduced to the 4 phonetic features. However, it is a fact that a statistical approach, such as HMM, makes it easy to modify modeling assumptions (e.g., signal representation) and experiment with them relying on the automatic training procedures to discover relevant information. On the other hand, an approach that does not rely on statistical and automatic learning methods, but requires the designer to discover information and explicitly model it, such as our rule-based EBS, makes such changes complex. Thus, although the experiments reported in this chapter show that an event-based approach is worth pursuing, automatic procedures to help develop various components of a system based on this approach are needed, especially as more complex recognition tasks such as word recognition are considered.

Table 5.5: Splits, merges and synonyms that were scored as correct. Category 1 + category 2 means a sequence of category1 and category2.

TIMIT Labels	allowed to be recognized as
syllabic sonorant consonants (/l/. /r/. /m/. /n/)	syllabic + sonorant consonant
diphthongs (/a ^y /. /a ^w /. /e ^y /. /o ^w /. /ɔ ^y /. /u/)	syllabic + sonorant consonant
/ʒ/. /ʒ/	syllabic + sonorant consonant
affricate	noncontinuant + fricative
/hv/	sonorant consonant or fricative
/h/	fricative
glottal stop (?)	stop or sonorant consonant
?+ syllabic	syllabic
fricative + fricative	fricative
vowel + vowel	syllabic
sonorant consonant + sonorant consonant	sonorant consonant
/dx/	silence + noncontinuant
silence + silence	silence

5.2.3 Error Analysis

An analysis of the event-based results was conducted to determine the error sources. There were three potential error sources identified: contextual variability, scoring algorithm and the decision process. These error sources are:

- Contextual variability: a partial analysis of the declared errors obtained from the EBS recognition results suggests that a good percentage (see Table 5.2.3) of them could be attributed to well known contextual changes that are not traditionally reflected in a phoneme-based hand transcription. For example, many researchers (e.g. [54], [55], [41]) have observed that voiced fricatives and voiced stops may be manifest as sonorants, especially when they occur between two sonorants in a falling stress environment. These acoustic changes, however, have not been traditionally captured in hand transcriptions. On the other hand, other phonological processes such as the flapping of /t/'s and /d/'s are represented. Thus, it is probably the case that most, if not all, of the declared errors listed in Table 5.2.3 are really not errors, but are the result of variability that occurs in speech production and, consequently, in the acoustic manifestation of the articulated sounds. Many of these errors were checked by hand to verify that the acoustic properties had indeed changed. First, voiced stops and voiced fricatives, particularly in intervocalic positions, are often realized as sonorant consonants (c.f. [55], [41]). Second, the fricatives /θ/ and /ð/ are sometimes produced as dental stops as often happens in words like "this" and "the" [31]. Third, semivowels are often coproduced with preceding unvoiced consonants so that they can be partially or completely devoiced (c.f. [55], [56]). Thus, these

semivowels for the most part will be realized as nonsonorant (the devoiced portion). Finally, the reduced vowel /ə/, when occurring between two unvoiced consonants, is often devoiced (c.f. [31]) and may be manifest as a fricative, especially when it is surrounded by two fricatives as the second vowel in the word "thesis".

- Scoring alignment: analysis of the system results revealed problems in the way alignment is done by the scoring algorithm. The scoring algorithm matches the sequence of manner-class labels generated from the TIMIT phonetic transcription with the sequence of recognized labels. This scoring is done without reference to time information. Thus, a recognized "syllabic" label that is correct for a vowel occurring in a particular time region can be matched to a TIMIT "noncontinuant" label for a stop occurring during a different time span, even when there is no time overlap.
- Decision errors: some of the errors were due to wrong decisions made on the basis of the parameters used. For example, the voicing probability measure often had difficulty detecting voicing during /ə/'s that lasted for only one or two pitch periods. Half of the vowels recognized as noncontinuants or fricatives were these short /ə/'s. As another example, the latter portion of front vowels occurring near the end of a sentence were often recognized as nonsonorants. In these cases, the amplitude of F1 at the end of the vowel usually decreased a great deal so that F1 was much weaker than the higher formants. This situation resulted in an error since the sonorant algorithm expects strong energy in the region of F1. As a result, the ends of these vowels were incorrectly recognized

as a noncontinuant or a fricative.

It is our belief along with others [57] that an understanding of contextual effects at the phonetic-feature level will result in better modeling of contextual variability. Such variability modeling can take place in the representation of lexical items by allowing the binary value of a phonetic-feature to be altered depending on the context [8]. In recognition systems utilizing context-dependent phone models to deal with variability, we expect that this knowledge coupled with the feature-based signal representation can improve data sharing among phone models and can lead to a reduction in the model size.

5.3 HMM Recognition System and APs

In this section, the APs are compared directly to the Mel-cepstral coefficients using the HMM framework. The task is again that of manner-class recognition of speech into the manner-classes: syllabic, sonorant-consonant, noncontinuant, fricative and silence. In Section 5.3.1, the modification of the APs developed for an event-based approach to fit into the frame-based HMM framework is discussed. In Section 5.3.2, the conducted experiments are described. The results of the experiments are discussed in Section 5.3.3.

5.3.1 Modification of the APs for the HMM framework

Table 5.1 shows the phonetic features, their corresponding acoustic correlates and acoustic parameters developed in this study. While the APs in Table 5.1 were designed for an event-based system so that they do not necessarily provide information in every

Table 5.6: Error analysis. The errors listed here were deduced from the manner-class recognition results obtained using the EBS. These errors may be explained by speech variability well documented in literature.

TIMIT-labeled Category	Recognized as	% of total errors in each category
Fricative	Sonorant Consonant	83% are weak voiced fricatives <i>/v/. /ð/.</i>
Fricative	Noncontinuant	77% are word-initial <i>/ð/. /θ/.</i>
Fricative	undetected	78% are weak fricatives <i>/f/. /ð/. /θ/. /v/.</i>
Sonorant Consonant	Fricative	62% <i>/l/</i> and <i>/r/</i> occurring after voiceless stop consonants.
Sonorant Consonant	Fricative	9% <i>/y/</i> and <i>/w/</i> occurring after voiceless stop consonants.
Vowel	Fricative	56% <i>/ə/. /ɪ/. /ə/.</i>
Stop	Sonorant Consonant	81% voiced stops <i>/b/. /d/. /g/.</i>
Closure	Sonorant Consonant	69% voiced stop-closures <i>/bcl/. /gcl/. /dcl/.</i>

frame. they can be modified to do so. The only parameters that need modification are the ones that mark particular instants in time which include the peak and dip measures for the phonetic features “syllabic” and “nonsyllabic” and the dip in R1, a measure for the phonetic feature “fricated”. The dip parameters are modified to be dip_to_peak (dtp) measures and the peak parameters are modified to be peak_to_dip (ptd) measures. The modified parameters are listed in Table 5.7.

The dip_to_peak and peak_to_dip APs attempt to estimate the energy levels in a sonorant consonant relative to a neighboring vowel and vice versa. These parameters are based on the assumption that the mid-frequency energy in the vowel region will have a peak relative to an adjacent sonorant consonant since the vowels are realized with a more open vocal tract (i.e., no constriction). On the other hand, sonorant consonants are realized with a mild (e.g., semivowels) to a severe constriction in the vocal tract (e.g., nasals) leading to a decrease in the mid-frequency energy relative to the neighboring vowels. The peak_to_dip and dip_to_peak APs are based on detecting peaks and dips in the appropriate energy profiles (computed at each time frame in the designated band). Significant peaks and dips are detected using the convex-hull algorithm [52] computed recursively across the energy profile. Once peaks and dips are located, the peak_to_dip AP computation proceeds as follows.

1. Assign a zero value to each time instant at which a dip occurred (this zero value in the AP can be thought of as encoding the fact that an energy dip was located at that time instant).
2. For the time stretch between a peak and the immediate dip to its right, compute the difference in dB between the energy profile and the energy value at that dip

location.

3. For the time stretch between a peak and the immediate dip to its left, compute the difference in dB between the energy profile and the energy value at that dip location.
4. The procedure is repeated for each peak resulting in the `peak_to_dip` AP which measures the strength of a peak relative to the immediate dips.

The `dip_to_peak` AP is computed in a similar manner. First, a zero value is assigned to AP at each time instant where a peak was located to encode the fact that a peak was detected at that time instant. Then, for the time stretch between a peak and a dip to its right, the difference in dB is computed between the energy profile and the energy value at the peak time location. For the time stretch between a peak and a left-dip, the difference in dB is also computed between the energy profile and the value and the energy value at the peak time location. At the dip location, the `dip_to_peak` value is computed as the difference in dB between the energy value at that instant and the average energy values of the two immediately surrounding peaks.

Figure 5.4 depicts an example of the APs computed from the utterance “biblical scholars” extracted from the TIMIT database. As this figure shows, the APs capture important characteristics of the speech signal. For instance, the abrupt onset parameter in part (a) shows the highest values at the burst release of the stop consonants /b/ and /k/. The voicing probability, the energy measure in the frequency band 0 – 2 kHz relative to that in the frequency band 2 – 8 kHz and the 100–400 Hz energy measure in parts (k),(l) and (m) of the figure, respectively, have fairly high values during sonorant segments. Furthermore, the peak-to-dip parameters in parts

Table 5.7: The phonetic features, their acoustic correlates and the corresponding acoustic parameters.

Feature	Correlate	Acoustic Parameter
Sonorant	strong low-frequency energy	E0.1-0.4: 100-400 Hz energy † E0-2-2-8: eng(0-2 KHz)-eng(2-8 KHz)
	periodic	Voicing-probability [60]
Syllabic	strong mid frequency energy	ptd0.64-2.8: peak in 0.64-2.8 kHz energy ptd2-3: peak in 2-3 kHz energy.
Nonsyllabic	weak mid-frequency energy	dtp0.64-2.8: dip in 0.64-2.8 kHz energy. dtp2-3: dip in 2-3 kHz energy.
Fricated	turbulence in mid to high frequency range	zcr: zero-crossing rate E0-2-2-8
		R1: first cross-correlation coefficient normalized by the zeroth. dtp_R1: dip-to-peak values of R1
Noncont.	Closure followed by an abrupt spectral change	<u>Closure:</u> E0.2-3: 0.2-3 kHz eng. †. E3-6: 3-6 kHz eng. † R1.
		<u>Abrupt onset :</u> sum of first-difference values across the STFT channels

† normalized with respect to its maximum value across the utterance.

(f) and (g) have peaks during the syllabic segments: /t/, /l/, /a/ and /æ/. Thus, the peak-to-dip parameters coupled with those related to sonorancy will help identify syllabic segments (as indicated by the phonetic-feature hierarchy in Figure 5.1).

5.3.2 Recognition Experiments

In the current work, two sets of experiments were performed to evaluate the acoustic parameters and compare them to cepstral-based parameters. In the first set of

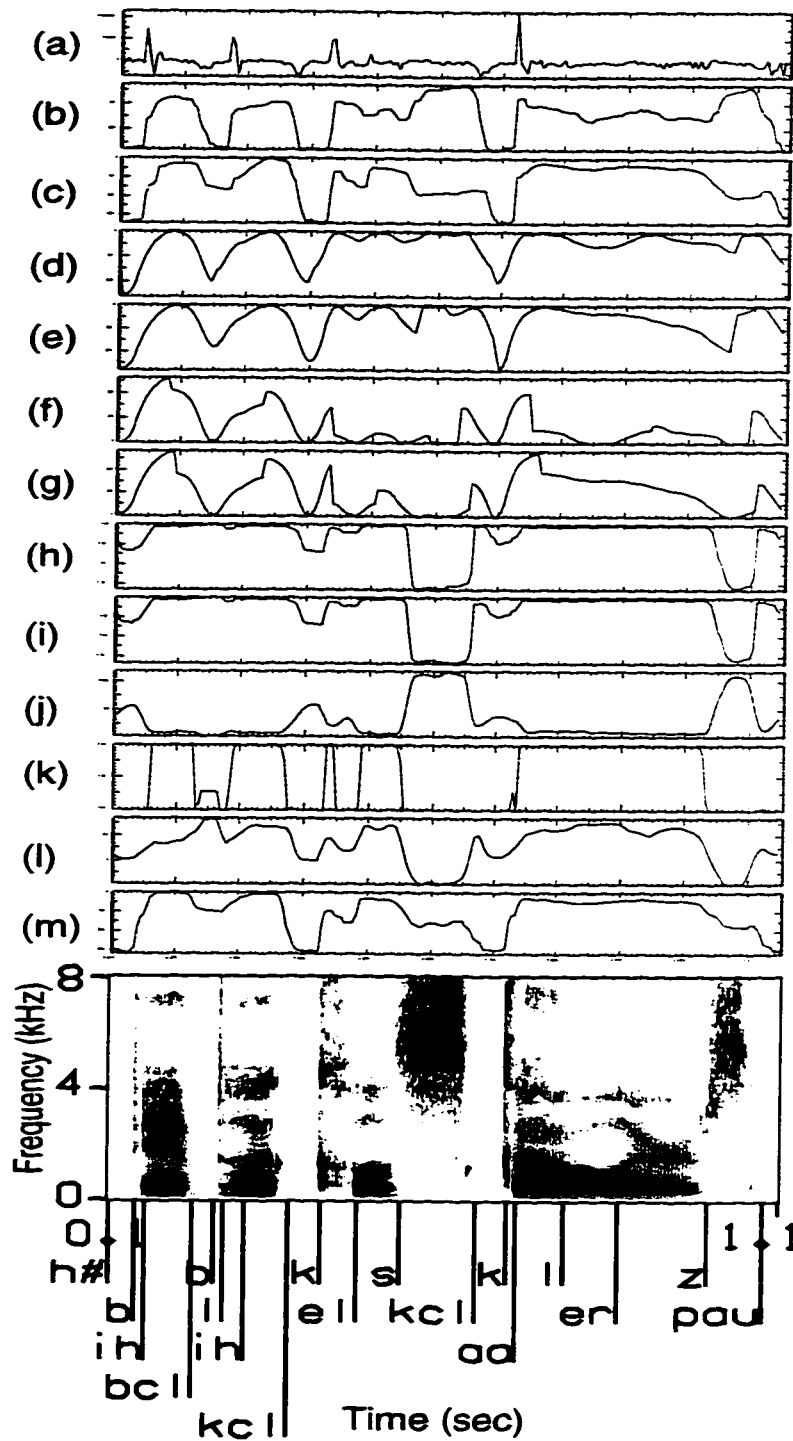


Figure 5.4: This figure illustrates the set of parameters listed in Table 5.7. These parameters are: (a) abrupt onset. (b) E3-6, (c) E0.2-3. (d) dtp_2-3. (e) dtp_0.64-2.8. (f) ptd_2-3. (g) ptd_0.64-2.8, (h) dtp_R1, (i) R1. (j) zcr. (k) voicing-probability. (l) E0-2.2-8 and (m) E0.1-0.4.

experiments, HMMs for the manner-classes were built in the same way described in Section 5.2.2. Recognition tests were carried out using the TIMIT "si" test sentences. In the second set of experiments, to examine robustness to interspeaker variability and more specifically to gender differences, HMM models were built using the TIMIT "si" and "sx" training sentences spoken by females from the New England dialect region (dr1). For testing, recognition was performed on the "si" and "sx" training sentences spoken by males from dr1. All manner-class models were context-independent 3-state HMMs with diagonal-covariance Gaussian mixtures.

5.3.3 Results and Discussion

Table 5.3.3 summarizes the experimental results where the signal representation (front-end) was varied while the modeling and recognition strategy remained the same. The splittings, mergings and synonyms in Table 5.5 were not allowed in obtaining these results. /ʔ/, /h/ and /hv/ were also deleted during recognition. The results for 1 and 8 mixtures show that the acoustic parameters, relative to the cepstral parameters, are better able to reduce speaker variability and target the linguistic information in the speech signal. This is deduced by comparing the small improvement in results going from 1 to 8 mixtures in the case of the APs to the substantial improvement in the case of the cepstral-based parameters. Furthermore, by comparing the *AP* results and the *MFCC_E* results to those obtained using *AP + MFCC_E*, one can argue that the acoustic parameters contain more relevant information than the cepstral parameters.

The results obtained with the APs and their first and second derivatives, in com-

Table 5.8: Recognition results. *MFCC_E* refers to Mel-frequency cepstral coefficients & normalized energy, *MFCC_E δ 1 δ 2* refers to *MFCC_E* & their 1st and 2nd derivatives. *AP* refers to acoustic parameters. *AP δ 1 δ 2* refers to *AP* and their 1st and 2nd derivatives. Each entry contains % correct/% accuracy. No splitting or merging was allowed in scoring.

Signal Representation	1 mix	8 mix
<i>MFCC_E</i>	68.2/61.8	73.3/65.2
<i>MFCC_Eδ1δ2</i>	73.5/63.7	82.8/71.5
<i>AP</i>	75.2/63.9	77.5/66.3
<i>APδ1δ2</i>	78.5/68.1	84.1/71.5
<i>AP + MFCC_E</i>	75.2/63.9	78.1/66.7

parison to APs alone, show that the additional parameters improve results substantially. A preliminary analysis showed that this improved performance was due in large part to better modeling of speech dynamics. For instance, an 11% absolute increase of stop consonant recognition was observed when the probability distribution given the HMM state was assumed to be an 8-mixture Gaussian.

Table 5.9 summarizes the experimental results obtained when the recognizers were trained on speech produced by females and tested on speech produced by males from dialect region dr1. Compared to the results obtained with the cepstral parameters (a 4.7 % absolute degradation from the system trained on both genders), the results obtained with the APs are much closer to the corresponding results listed in Ta-

ble 5.3.3, indicating more robustness to gender variability. Additional experiments were conducted with the APs and their derivatives. In the first experiment, a 1-mixture system, instead of 8 mixtures, was trained using the “si” and “sx” female sentences in the TIMIT training set and tested on the “si” male sentences in the TIMIT test set. An absolute 1% performance degradation was observed compared to the system trained on both males and females. In the second experiment, the same system was trained on all “si” and “sx” male sentences in the TIMIT training set and tested on the “si” female sentences in the test set. No degradation in performance was observed. These experiments indicate the robustness of the APs to gender differences.

In order to compare the HMM framework to the event-based framework, both using APs as the front-end, the HMM results with APs were rescored allowing the splits, merges and synonyms listed in Table 5.5. These results are summarized in Table 5.10. As this table shows, EBS performed better than the HMM system. However, when $AP_{\delta 1 \delta 2}$ was the front-end to the HMM system, HMM outperformed EBS, but note that EBS only uses the APs without their derivatives. The advantage of the HMM framework in this case is that it allowed the “blind” addition of the derivatives to the APs. This cannot be done with the current EBS.

5.4 Concluding Remarks

In this chapter, we discussed a signal representation in the phonetic-feature space. This signal representation was obtained by mapping acoustic parameters that target the manner-of-articulation phonetic features to degrees of belief in these features. We compared an event-based system utilizing this representation to an HMM system

Table 5.9: Recognition results using 8 mixtures. Training done with speech produced by females. Recognition done with speech produced by males.

Signal Representation	%correct/%accurate
<i>MFCC_E_δ1_δ2</i>	81.1/66.8
<i>AP_δ1_δ2</i>	83.3/70.7

utilizing *MFCC_E_δ1_δ2* in a manner-class recognition task. We found that the two systems were comparable. In addition, we compared the acoustic parameters (without mapping them to the phonetic-feature space) to the cepstral parameters in the HMM framework. The task was also manner-class recognition. The results indicate that the signal representation we developed better extracts relevant linguistic information from the speech signal and that relative measures across time or frequency diminish speaker-dependent effects.

Table 5.10: Recognition results obtained when HMM was used as the recognition framework. In scoring, the splittings, mergings and substitutions listed in Table 5.5 were allowed.

Signal Representation	HMM 1 mix	HMM 8 mix	EBS
<i>AP</i>	76.7/69.3	78.9/72.4	84.6/72.8
<i>AP_δ1_δ2</i>	80.2/73.6	85.4/78.0	-

Chapter 6

Parameter Optimization

6.1 Motivation

In Chapter 5, several parameters that relate to the manner-of-articulation features were presented and tested in a manner class recognition task. The recognition results showed that the philosophy of designing parameters based on acoustic-phonetic knowledge and based on relative measures can yield a signal representation that focuses on the phonetic content of the speech signal while reducing all other information. Two shortcomings of the presented parameters are (1) the parameters were designed based on subjective criteria represented by acoustic phonetic knowledge, spectrogram analysis and histogram analysis, and (2) the extent of correlation among the different parameters was not studied to objectively determine whether all used parameters were needed.

It is usually the case that the acoustic measures that distinguish among speech classes are qualitatively known. However, the problem lies in quantifying these mea-

sures. For instance, a sonorant sound has strong low-frequency energy compared to a nonsonorant (obstruent) sound. Thus, low-frequency energy is a qualitative acoustic measure that characterizes sonorancy/nonsonorancy. The problem is to quantify such a measure, i.e., determine the frequency band(s) that should be used in its computation. Therefore, a criterion needs to be used for the selection of a frequency band over another and for eliminating redundant parameters. In this chapter, the problem of acoustic parameter design using objective criteria is addressed. In Section 6.2, the adopted procedure in parameter design is outlined. Two components of this procedure, the Fisher criterion and classification trees are discussed in Section 6.3 and Section 6.4, respectively.

6.2 Procedure

The parameter optimization algorithm has its roots in [58]. The differences between our approach and that in [58] are (1) the APs are based specifically on acoustic cues relevant to phonetic features, (2) the APs are defined in relative terms to reduce the effects of interspeaker variability and (3) classification trees for eliminating redundant parameters are deployed. The objective is to develop parameters that best characterize a phonetic feature and separate it from its antonym(s). To accomplish this task, the Fisher criterion [59] was chosen along with classification trees. Another criterion that was tried in addition to the Fisher criterion is the Fuzzy Evaluation Index (FEI) described briefly in Appendix A. However, preliminary experiments with the FEI suggested that outliers or samples that are realized differently from their canonical form due to contextual variability greatly effect the results of the optimization. Al-

though there are ways to counter the FEI deficiencies as it was defined. the work in this thesis proceeded with the Fisher criterion leaving the possibility of investigating other objective criteria for later work.

The procedure that is used to derive a set of acoustic parameters consists of the following steps:

1. Group the set of all sounds that have a phonetic feature in one group and the sounds that do not have that feature in another group.
2. Based on acoustic phonetics, define a set of generic parameters, with free parameters that are intended to separate the classes from each other. For instance, $E[f_1 : f_2]$ is an energy measure between frequencies f_1 and f_2 (free parameters) with the condition $f_1 < f_2$.
3. The range of frequencies from which f_1 and f_2 are selected is specified from acoustic-phonetic knowledge.
4. For each generic parameter, determine the set of free parameter pairs that result in local maxima in the Fisher criterion (FC) surface.
5. Feed the parameters generated from the previous step as well as any additional parameters obtained through other methods to a classification tree. The objective of the classification tree is to select a subset of these parameters. Classification trees are greedily grown and then pruned back using cross validation.

As an example of acoustic parameter design, consider the place-of-articulation feature that differentiates the anterior stridents (alveolars) /s/ and /z/ from the

nonanterior stridents: /ʃ/, /ʒ/, /ç/ and /j/. The parameter design procedure in this case was based on the following steps:

1. All /s/, /z/ samples from the TIMIT training set were placed in the anterior group while the /ʃ/, /ʒ/, /ç/ and /j/ samples were placed in the nonanterior group.
2. Based on acoustic phonetic knowledge, generic parameters were designed to distinguish between the two classes of sounds. This knowledge specifies that the energy of the nonanterior is concentrated in the third formant ($F3$) region while that of the anterior is concentrated in the fifth formant ($F5$) region. Based on this information, generic parameters were chosen so that energy in a mid frequency band (around $F3$) is measured relative to energy in higher and lower frequency bands and relative to the overall energy. Some of these parameters are computed within the obstruent boundaries relative to the maximum, minimum and average values of the same parameters computed across the utterance.
3. The mid-frequency band, $[f_1 : f_2]$ of the generic parameters was chosen so that f_1 can take any value in the frequency band $[F3-1000, F3+1700](Hz)$ while f_2 can take any value in the frequency-band $[F3-700, F3+2000](Hz)$ with the condition that $(f_2 - f_1) \geq 300(Hz)$. $F3$ was estimated for each TIMIT utterance in the training set separately using the Waves [60] formant tracker. To minimize the effects of errors in formant tracking, the average $F3$ value was computed in two stages. In the first stage, an average $F3$ value was computed in the sonorant regions (as determined from the Waves voicing probability). Then, in the second stage, formant values that deviated by more than a standard

deviation from the average $F3$ value were disregarded and the average $F3$ value was recomputed.

4. For each of the generic parameters, a Fisher surface was computed. As an example, the Fisher surface obtained by computing the energy in the band $[f_{st} : f_{end}]$ relative to the overall energy at the same time frame and then averaged across the utterance is shown in Figure 6.1. From each of the Fisher surfaces, the local maxima were picked. The frequency bands corresponding to these maxima specify a set of candidate parameters that distinguish between the anteriors and nonanteriors. In this case, there were 19 such parameters generated from the 16 generic parameters.
5. The final parameter set is obtained by feeding all 19 parameters, obtained from the previous step, to a classification tree and selecting the ones that significantly contribute to the intended discrimination. As a result, the parameter that measures the energy in the frequency band $[F3 - 187Hz, F3 + 594Hz]$ relative to the overall energy within the obstruent was chosen as the best parameter yielding 91% correct classification. Two additional parameters were chosen by the tree that increase the overall classification rate to 93%.

In order to emphasize the importance of reducing interspeaker variability and specifically gender differences, the density of the best classification parameter obtained with $F3$ normalization and that obtained without $F3$ normalization for the anterior sounds are plotted in Figure 6.2 (a) and (b), respectively. In comparing these two figures, it is clear that the normalized parameter is better able to reduce gender differences.

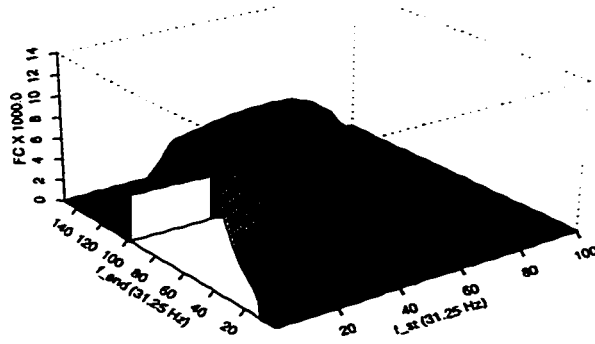


Figure 6.1: Fisher criterion for the parameter which computes the energy between f_{st} and f_{end} relative to the overall energy at a given frame instant. The origin is $F3 - 1000$ (Hz).

6.3 Fisher Criterion: First Stage

The Fisher criterion [59] can be used to evaluate the capability of an N -dimensional feature vector to characterize several competing classes while separating them from each other. The idea of the Fisher criterion is illustrated in Figure 6.3 where samples from three classes denoted by the symbols (x, o and *) are considered. The objective is to have the samples from class C_i clustered around its sample mean m_i while maximizing the distance between the individual m_i 's and the sample mean of the pooled data. The Fisher criterion is defined as the ratio of a measure of the between-class scatter to a measure of the within-class scatter. The within-class scatter is characterized by the matrix:

$$S_W = \frac{1}{n} \sum_{i=1}^C S_i \quad (6.1)$$

where,

$$S_i = \sum_{j=1}^{n_i} (x^{(i)}_j - m_i)(x^{(i)}_j - m_i)^T. \quad (6.2)$$

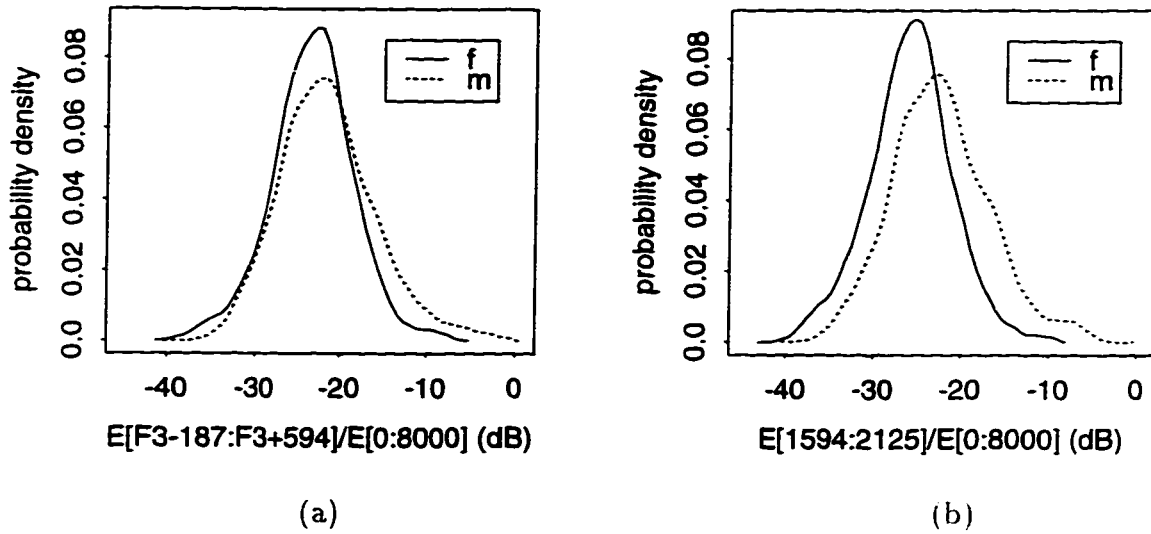


Figure 6.2: Distribution of best parameter computed using the anterior samples for males (m) and females (f). (a) Parameter was defined relative to third formant (F3) location. (b) Parameter was independent of F3.

In equations 6.1 and 6.2, C is the number of classes, m_i is the sample mean for class C_i , $x_j^{(i)}$ is the j^{th} sample from class C_i and n_i is the corresponding number of observed samples. Thus, S_i is proportional to the covariance matrix for class C_i while S_W is a weighted sum of the covariance matrix estimates for the C classes. The variable n denotes the total number of observations and T denotes the transpose operation. The between-class scatter is characterized by the matrix:

$$S_B = \sum_{i=1}^C \frac{n_i}{n} (m - m_i)(m - m_i)^T \quad (6.3)$$

where the sample mean for class C_i is given by

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j^{(i)}$$

and the sample mean for the pooled data is given by

$$m = \sum_{i=1}^C \frac{n_i}{n} m_i.$$

Thus, each diagonal element of S_B is the weighted sum, across classes, of the Euclidean distance between a class sample mean and the global mean along a feature dimension. Thus, maximizing the trace of S_B , $tr(S_B)$, will result in maximizing the sum of these distances across all dimensions. Similarly, each diagonal element of S_W is related to the sum across classes of the class sample variances along a feature dimension. Thus, one form of the Fisher criterion is:

$$(FC)_{tr} = \frac{tr(S_B)}{tr(S_W)} \quad (6.4)$$

According to this criterion, a better feature vector results in a higher $(FC)_{tr}$ value. Another definition of the Fisher criterion is in terms of the determinants of S_B and S_W and is given by:

$$(FC)_{det} = \frac{|S_B|}{|S_W|} \quad (6.5)$$

Maximizing $(FC)_{det}$ results in maximizing/minimizing the variances of S_B/S_W along the principle directions. In this thesis, we adopt the trace definition of FC and we will refer to it simply as FC .

6.4 Classification Trees: Second Stage

Classification trees are used in the parameter optimization procedure as a means to prune down the set of candidate parameters obtained from the first stage based on the Fisher Criterion. In obtaining parameters based on the Fisher criterion, no

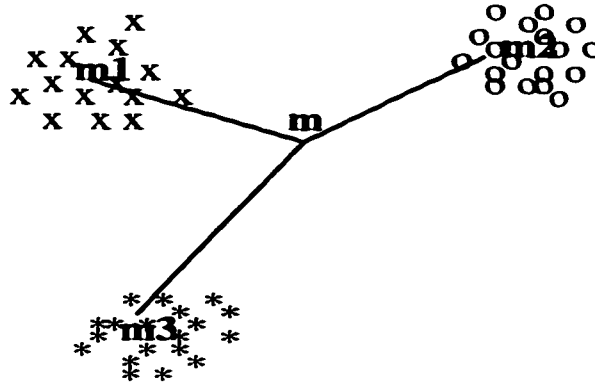


Figure 6.3: Separation between three classes using the Fisher criterion is based on maximizing the distances among the class centroids and the centroid of the pooled data.

attention was given to the redundancy among the different parameters or to the joint contribution of the parameters in making the intended distinction. Classification trees are used to achieve this goal and consequently to eliminate redundant parameters or parameters that become irrelevant when considered jointly with others. Growing a tree is based on a set of data samples computed in the parameter space and an objective criterion that mainly determines the tree structure. A tree growing process first selects from a set of K parameters computed across samples from C classes the parameter that makes the “best” distinction among the C classes, splitting the samples into two groups (see Figure 3.1). Then, the process selects a parameter that distinguishes best among the samples from the group in the left branch of the tree and a parameter that distinguishes best among the samples from the group in the right branch of the tree. The split continues until a split-stoppage criterion is met. As a result, the nodes of the tree will contain the set of parameters, out of the possible K parameters, that distinguish “best” among the C classes.

A classification tree serves as a first-best-entry classifier. Other types of classifiers were considered for this stage. Among those considered were Gaussian classifiers, Gaussian mixture classifiers and linear regression models. The first two classifiers make strong and often invalid assumptions regarding the distribution of the considered parameters. The linear regression models, on the other hand, strongly depend on the way that parameters appear in the model (e.g. square of the parameter, parameter interaction etc.) and may result in unstable models if strongly correlated parameters are considered in the modeling. Thus, to avoid the problems associated with these assumptions, classification trees were chosen. In this section, the tree growing algorithm used and the process of the selection of the final set of parameters out of the possible K parameters are discussed.

6.4.1 Node-Splitting Criterion:

The structure of a classification tree is mainly defined by the objective criterion used in deciding on the “best split” during the tree growing process. Since the concern in this case is to select a set of parameters that best distinguishes among C classes, it is reasonable to use a splitting criterion that is correlated with the classification error rate. The criterion chosen is known as the “Deviance” (D) criterion [61]. This criterion attempts to split the data into contiguous homogeneous regions in the parameter space where homogeneity is synonymous with the considered classes. First, the following definitions are adopted in the tree growing procedure:

- Y is a C -dimensional vector (C is the number of classes) such that every observation of class i is assigned the vector $y_i = (0.0, \dots, 1.0, 0.0)$ where the i^{th}

component of y_i is 1 and every other component takes the value zero.

- The probability of an observation coming from class i and falling in the k^{th} tree node is defined as p_i^k , where.

$$p_i^k = N_i^k / N^k, \quad (6.6)$$

N_i^k is the number of samples from the i^{th} class falling in the k^{th} node and N^k is the total number of samples from all classes falling in that node. That is, p_i^k is the likelihood of a class falling in that node computed from a frequency count.

- The deviance of a sample from class i and falling in the k^{th} node is defined as

$$D^k(y_i) = -2\ln(p_i^k). \quad (6.7)$$

That is the deviance of an observation is based on the log likelihood of an observation being in that node.

- The deviance of a node k is defined as the sum of the deviances of the samples that fall in that node. That is the deviance of the k^{th} node is defined as:

$$D^k = \sum_{j=1}^{N^k} D^k(y_j). \quad (6.8)$$

Thus, the deviance of a node is identically zero if all samples in that node have one color (i.e. belong to one class) and increases as the node becomes more colored. The tree growing procedure proceeds as follows:

1. At each node in the tree, consider each of the parameters as a basis for splitting that node.

2. For each parameter-split, compute the deviance in the right and left nodes resulting from the split as:

$$D_{LR} = D_L + D_R, \quad (6.9)$$

where D_L and D_R are the deviances in the left and right nodes respectively.

3. Let D_P denote the deviance of the node being split. Then, the change in deviance resulting from the split is:

$$\Delta(D) = D_P - D_{LR}. \quad (6.10)$$

The split, i.e. the parameter, that maximizes this deviance change will result in the best split at that node.

6.4.2 Determining Tree Size and Important Parameters

When does the tree growing process stop? Does it stop when each of the terminal nodes has a zero deviance? This certainly can be achieved by having singleton terminal nodes. However, the resulting tree will not be robust since it will be tuned to the training data and will lack generalization to new data sets. As a result, some parameters will appear important when they are not. On the other hand, setting a limit on the size of the tree as a stoppage criterion may lead to the elimination of important parameters. This can be the case due to the greedy nature of the tree growing process. That is, the best parameter is chosen at each node irrespective of what will happen at later stages. In order to avoid these problems, the final tree structure is achieved in two stages. First, a very large tree is grown by allowing a node to be split if its size and deviance exceed some preset liberal value. In all

conducted experiments, a node is allowed to be split if its size (i.e., the number of samples in the node) is greater than 10 and if its deviance is greater than 1% of the root node deviance. The resulting tree will be overly large and may still be overfitting the training data. The second stage of the tree growing process, tree pruning, is used to alleviate this problem. Tree pruning attempts to snip off the tree branches that are tuned to the training data and most likely will not generalize to other data sets.

How should tree pruning be performed? There are two methods to perform tree pruning. The first method uses a totally independent data set from the one used in tree growing while the second method uses the cross validation technique to determine the optimum tree size. In the first method, a sequence of trees is constructed by snipping of branches from the grown tree in the bottom-up direction. The new data is dropped into each tree in the sequence and the corresponding tree deviance is computed. Tree deviance is defined as the sum of all terminal nodes deviances. This process will result in deviance-tree pairs. The tree that gives the minimum deviance can be selected as the tree of choice. The chosen tree can be further inspected for additional manual pruning.

The cross validation method divides the training data randomly into V mutually exclusive subsets. $V - 1$ subsets at a time are used in growing a tree while the held-out subset is used in computing the deviance of the sequence of subtrees derived from this grown tree, therefore the name cross-validation. The sequence of subtrees is obtained based on a set of algorithmically derived complexity factor[62]. For each complexity factor α , the subtree t that minimizes the cost-complexity function

$$D_\alpha = D_t + \alpha \text{size}(t) \tag{6.11}$$

is chosen as the best subtree for that complexity factor. Then, the held-out subset is dropped down the chosen subtree and the corresponding subtree deviance is computed. For a given α , this computation is repeated across all V cross-validation tests and the tree deviance for that complexity factor is compiled across the tests. This process is repeated for all α values. Then, the α that results in the minimum compiled deviance value is chosen as the optimum complexity-factor value. Now, a subtree of the original tree, grown from the full training data set, that minimizes the cost-complexity function in equation 6.11 with the chosen optimum value of α is selected as the most robust tree or right sized tree. The parameters chosen at the tree nodes are considered to be the optimum set of parameters. Since one of the purposes is to have a parsimonious set of parameters, this tree is investigated further to see what effect does snipping off branches (i.e., reducing the set of parameters) have on the classification error rate. The parameters that contribute only about 1% to the classification rate are removed from this set of parameters.

Chapter 7

Optimized Acoustic Parameters

Several acoustic parameters were derived using the parameter optimization procedure outlined in Chapter 6. In Section 7.1, acoustic parameters that target the “sonorant” phonetic feature are derived and tested. Acoustic parameters for the “anterior” place of articulation, distinguishing among strident obstruents, are derived in Section 7.2. Stop place-of-articulation parameters are derived in Section 7.3, and “syllabicity” parameters are derived in Section 7.4. Finally, acoustic parameters that target the “strident” phonetic feature are obtained in Section 7.5.

7.1 Sonorancy

In this section, acoustic parameters that target the sonorant feature in the speech signal are derived. The selection of these parameters is motivated by the manner in which sonorant sounds, as opposed to nonsonorant sounds, are produced.

English sounds can be broadly classified into two categories: the sonorant sounds

and the nonsonorant sounds (or obstruents). Sonorant sounds are produced with a vocal tract configuration that allows spontaneous vocal cord vibration to take place while obstruent sounds are produced with a vocal tract configuration that impedes such vibration [7]. In sonorant sound production, the main source of excitation is a pseudo-periodic source at the glottis while that of obstruents is a noise source at a constriction somewhere forward in the vocal tract. The pseudo-periodic characteristics of the sonorant excitation source is due to the pseudo-periodic movement of the vocal folds opening to let air through from the lungs and closing to prevent such air from passing through.

The vocal tract configuration determines a resonant acoustic structure while the source determines the input to this structure. Thus, the spectral content of the speech signal is influenced by both the resonances of the vocal tract and by the source characteristics. For sonorant sounds, the source is a pseudo-periodic sequence of glottal pulses, called the glottal waveform, that excites all resonances of the vocal tract due to its location at the glottis. The time/frequency characteristics of the glottal waveform has been studied by several researchers (e.g., [32]). The spectrum of the periodic glottal waveform is composed of harmonics at multiples of the pitch frequency, the fundamental frequency at which the vocal folds vibrate. Furthermore, the glottal-waveform spectrum decays at 12dB/octave above approximately 500 Hz [32]. Thus, it is generally the case that, for sonorant sounds (e.g., vowels), the higher resonances of the vocal tract (F_4 , F_5 , etc.) are less excited than the lower ones (F_1 , F_2 and F_3). As a result, the spectral content of the sonorants falls in amplitude as frequency increases and is characterized by strong energy in the region of F_1 and F_2 . Obstruents, on the other hand, have their primary source of excitation above

the glottis, at the constriction. As a result, only the resonances of the vocal cavity in front of the constriction are usually excited. Since this cavity is short, it is the higher formants that have the most energy.

7.1.1 Acoustic Parameters that target the sonorancy feature

Based on the manner of production of the sonorant and obstruent sounds and on several acoustic studies [40] [42], acoustic parameters whose intent is to capture the differences between sonorant and obstruent sounds were proposed.

First, an algorithm that measures the degree of voicing in a given frame of speech was used. This algorithm is implemented by Entropic Research Laboratories [60] as an integrated pitch/formant tracker. The algorithm makes a voicing decision referred to as voicing probability, based on the maximum cross-correlation coefficient [16] computed over a 10 ms window and the rms energy in that window relative to the maximum rms energy across the utterance. The correlation coefficient is intended to capture the pseudo-periodic aspect of sonorant sounds. On the other hand, the rms energy is intended to capture the fact that most sonorant sounds, especially the vowels and semivowels, generally have more overall energy than the obstruent sounds due to the relatively open configuration of the vocal tract.

Second, acoustic parameters that measure the amount of energy concentrated in the lower part of the frequency spectrum relative to the higher part and to the overall energy were designed. These parameters account for the fact that sonorant sounds have strong energy in the lower part of the frequency spectrum as explained above.

Third, acoustic parameters were designed to measure the energy in the region

of the fundamental frequency, F_0 , relative to the maximum, minimum and average energy across the utterance within the same frequency region. These parameters are based on the assumption that all sonorant sounds spoken by the same speaker are produced with more or less the same excitation source. Thus, it is expected that the energy from this source does not change much during an utterance on the order of a few seconds in duration.

7.1.2 Parameter Optimization

As outlined in the parameter optimization procedure described in Chapter 6, the frequency bands that give the higher Fisher criterion value for each generic acoustic parameter were obtained independently from each other. Then, all parameters were fed to a classification tree to further reduce this initial set of parameters. All parameters were computed over the middle two thirds of the sounds in the training data. All sound samples except those labeled as glottal stops ($/ʔ/$), $/h/$, $/hv/$, and $/ə/$ were included in this optimization process. The reasons for excluding the glottal stops are (1) the glottal stop label is not phonemic and (2) the labelers did not distinguish between a glottal stop and glottalization which are spectrally very different. The $/h/$ and $/hv/$ sounds are controversial in terms of sonorancy while the sounds labeled $/ə/$ are generally heavily aspirated $/ə/$'s so they can be manifest as sonorant or nonsonorant. The training data consisted of a total of 12,632 nonsonorant sounds and 24,296 sonorant sounds.

Periodicity-based measure:

For this measure, the voicing probability algorithm described in Section 7.1.1 was

used. Figure 7.1 and Figure 7.2 show the distribution of voicing probability for each phone. As can be seen, most of the canonically sonorant sounds have high voicing probability values while nonsonorant sounds have low voicing probability values. Using a tree-based classifier, this algorithm resulted in 92.9% correct classification rate when tested on the training data. Inspection of the sonorant classification errors revealed that about 48% of the sonorants classified as nonsonorants were the nasals (/n/, /m/, /ŋ/, /ɲ/, /ɳ/, /ɱ/). These errors occurred because of the lower rms energy of the nasal sounds in comparison to the other sonorant sounds. The low rms energy of the nasal sounds is due to their manner of production. Nasal sounds are produced by letting the air flow from the lungs through the “lossy” nasal cavity while completely blocking the air flow out of the mouth.

Energy Concentration in Low Frequency Bands:

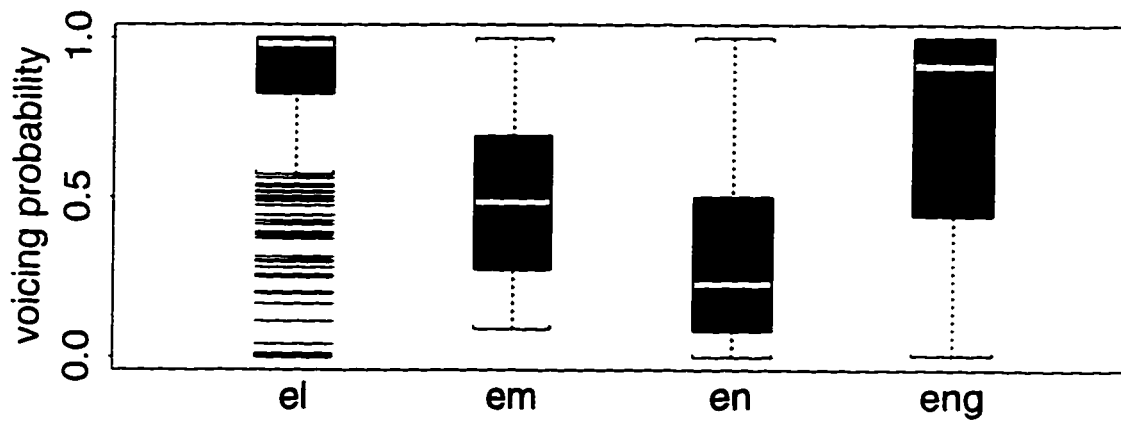
Two generic acoustic parameters were designed to capture the strong energy characteristic of sonorant sounds in the low frequency bands. The first generic acoustic parameter is:

$$E[f_1 : f_2] / E[f_2 : 8000] \quad (7.1)$$

where f_1 and f_2 are free parameters whose values were determined using the Fisher-criterion optimization algorithm described in Chapter 6. The frequency f_1 was chosen from the interval $[0, 4000 Hz]$ while f_2 was chosen from the interval $[500, 4500 Hz]$ such that $(f_2 - f_1) \geq 500 Hz$. Thus, $E[f_1 : f_2]$ was always computed within a minimum bandwidth of $500 Hz$. This restriction limits the overall number of parameters that are spanned by the Fisher-criterion based algorithm and therefore reduces the overall number of computations. In addition, although such a band seems to be arbitrary,



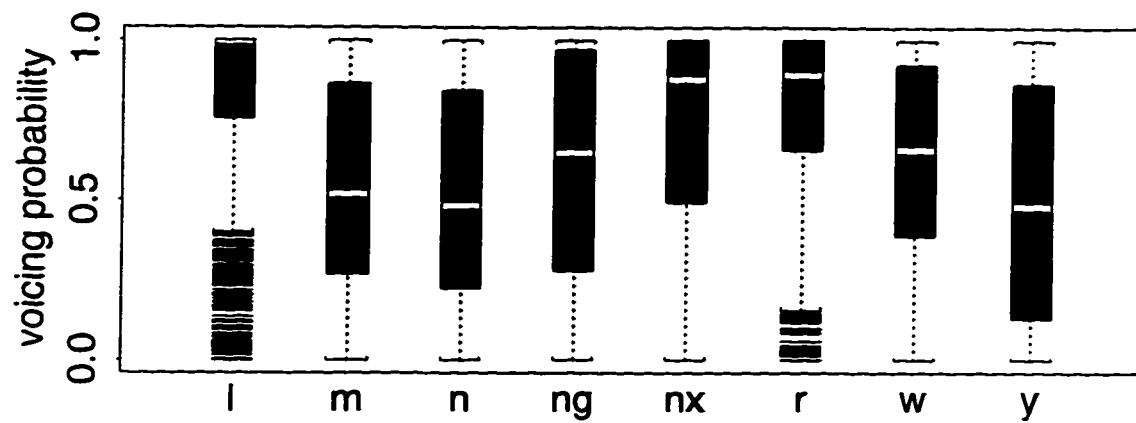
(a)



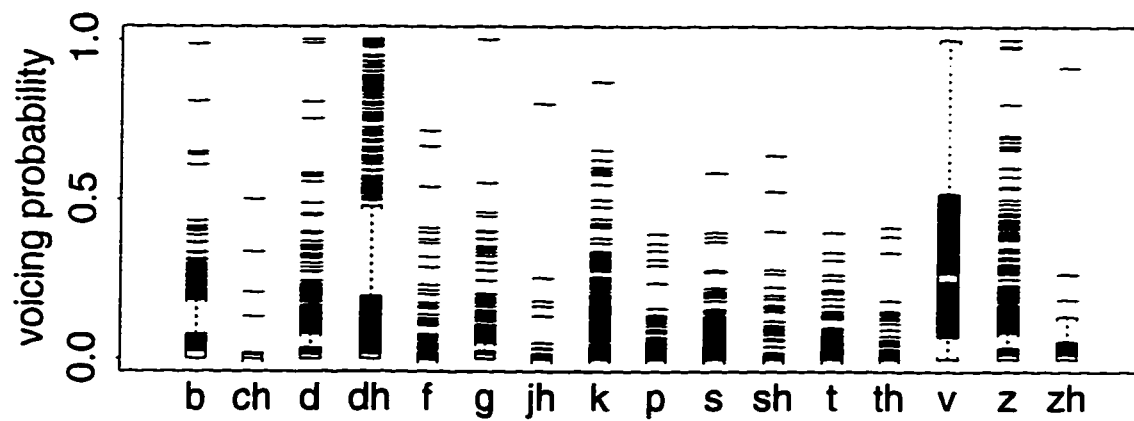
(b)

Figure 7.1: Voicing probability distribution for (a) vowels and (b) syllabic consonants.

The TIMIT symbols for phones are used along the horizontal axis.



(a)



(b)

Figure 7.2: Voicing probability distribution for (a) sonorant consonants and (b) obstruents. The TIMIT symbols for phones are used along the horizontal axis.

Table 7.1: Parameters selected in the Fisher-Criterion stage of the parameter optimization process.

Acoustic Parameter	Fisher-Criterion Selected Parameters
$E[f_1 : f_2]/E[f_2 : 8000]$	E[219:750]/E[750:8000] E[94:2969]/E[2969:8000] E[63:4000]/E[4000:8000]
$E[0 : f_1]/E[f_2 : 8000]$	E[0:688]/E[4000:8000] E[0:719]/E[1219:8000] E[0:656]/E[1156:8000]

the minimum bandwidth chosen is not expected to miss any important parameters because energy concentrations of the sonorant and obstruent sounds are usually in bands wider than 500 Hz . Furthermore, the first parameter computed is the energy in the band $[0 : 500]\text{ Hz}$ relative to that in the band $500 - 8000\text{ Hz}$ and this is intended to capture in one of the parameters the effect of the glottal waveform spectrum decaying beyond 500 Hz at 12 dB/octave . The second acoustic parameter was designed to complement the first acoustic parameter by giving a more detailed description of the spectrum. This second acoustic parameter is given by:

$$E[0 : f_1]/E[f_2 : 8000]$$

where f_1 and f_2 are as specified for the first acoustic parameter. The parameters obtained from each acoustic parameter, based on the Fisher-criterion, are listed in Table 7.1 in order of increasing Fisher-criterion values.

Energy in the F_0 Region:

Two approaches were taken in designing acoustic parameters for measuring energy in the F_0 (pitch) region. The first was based on a general knowledge of the range of F_0 values for males and females without an explicit estimation of the pitch value within a spoken utterance. The second was based on a pitch value estimated from the signal.

It is generally known that the pitch frequency for a male can range between $100Hz$ and $300Hz$ while females have a pitch value in the $150Hz$ to $400Hz$ range. Thus, it seems appropriate to measure the energy in the signal in a 100 to $400Hz$ frequency band. Furthermore, since energy is a function of loudness which may be related to speaker gender, dialect and other factors, we opted to measure the energy in a speech frame in a specified band relative to its maximum, minimum and average value across the utterance. Thus, the generic energy measures were defined as:

$$E[f_1 : f_2] / \text{maximum}(E[f_1 : f_2])$$

$$E[f_1 : f_2] / \text{minimum}(E[f_1 : f_2])$$

$$E[f_1 : f_2] / \text{average}(E[f_1 : f_2])$$

The maximum, minimum and average values are taken across the utterance. For each sonorant and nonsonorant sound considered, $E[f_1 : f_2]$ was computed as the average value of that parameter over the middle two-thirds of the sound. The optimal frequency band defined by f_1 and f_2 was picked using the Fisher-criterion optimization procedure. The values that f_1 and f_2 were allowed to take were liberally constrained by the speech knowledge of pitch range such that $f_1 \in [0Hz, 450Hz]$

Table 7.2: These parameters were selected based on the Fisher-Criterion stage of the parameter optimization process.

Acoustic Parameter	Fisher-Criterion Selected Parameters
$E[f_1 : f_2] / \text{maximum}(E[f_1 : f_2])$	E[156:375]/maximum(E[156:375]) E[156:250]/maximum(E[156:250]) E[0:375]/maximum(E[0:375])
$E[f_1 : f_2] / \text{minimum}(E[f_1 : f_2])$	E[156:375]/minimum(E[156:375]) E[63:375]/maximum(E[63:375])
$E[f_1 : f_2] / \text{average}(E[f_1 : f_2])$	E[0:375]/average(E[0:375]) E[156:375]/average(E[156:375])

and $f_2 \in [50Hz, 500Hz]$ with the constraint that $(f_1 - f_2) > 50Hz$. The bandwidth constraint is due to the fact that a 25 ms Hamming window was used in the analysis, yielding a frequency resolution slightly less than 50Hz. Based on the three measures defined in equation 7.2 and using the Fisher-criterion algorithm, the measures listed in Table 7.2 were obtained.

Another approach was taken to capture the energy in the $F0$ region. In this approach, the frequency band in which the energy was measured was adapted based on the mean $F0$ value estimated from the spoken utterance. In estimating the mean $F0$ value, a two-pass procedure was adopted in order to reduce the effect of errors introduced by false estimation of $F0$ at some time instants. The procedure for estimating the mean $F0$ value consists of the following steps:

- Estimate $F0$ at every speech frame. This was done using the Entropic pitch

tracker algorithm [60] which is based on the cross-correlation method [16].

- Estimate the mean of $F0$, $\bar{F}0$, as:

$$\bar{F}0 = \frac{1}{n} \sum_0^n F0_n$$

where n is the total number of frames at which $F0$ was estimated.

- Estimate the standard deviation of $F0$, σ_{F0} as:

$$\sigma_{F0} = \sqrt{\left(\frac{1}{n} \sum_0^n (F0_n - \bar{F}0)^2\right)}$$

- Reestimate the mean of $F0$, $\hat{F}0$ by filtering out those values of $F0$ that differ from $\bar{F}0$ by more than σ_{F0} and recomputing the average of the remaining $F0$ values.

A generalized acoustic parameter that measures the energy around $\hat{F}0$ was defined as:

$$E[\hat{F}0 + f_1 : \hat{F}0 + f_2] / \text{maximum}(E[\hat{F}0 + f_1 : \hat{F}0])$$

where $f_1 \in [-50Hz, 400Hz]$ and $f_2 \in [0Hz, 450Hz]$ such that $(f_2 < -f_1) \geq 50Hz$.

The optimal parameter determined from this generic measure based on the Fisher-criterion was $E[F0 - 31 : F0 + 156]$.

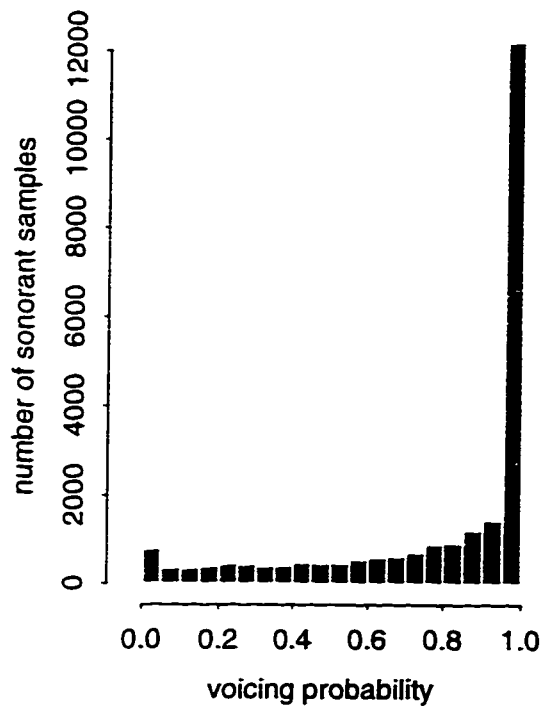
Selected Sonorancy Parameters:

The parameters obtained in Section 7.1.1 were derived independently of each other and therefore may contain a high degree of redundancy. In order to obtain a more parsimonious set of parameters, the tree classification algorithm, the second stage of the parameter optimization process (described in Chapter 6) was applied to this initial parameter set. In addition, since nasals were thought to have lower voicing

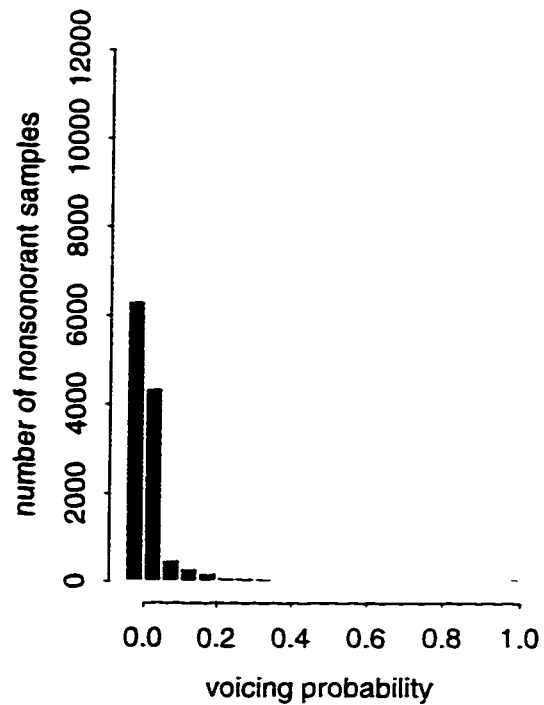
Table 7.3: Parameters selected from the two-stage optimization process to distinguish between sonorant and nonsonorant sounds.

Sonorant Feature Parameter Set
probability of voicing
$E[0 : 688] / E[4000 : 8000]$
$E[0 : 375] / \text{average}(E[0 : 375])$
maximum cross-correlation

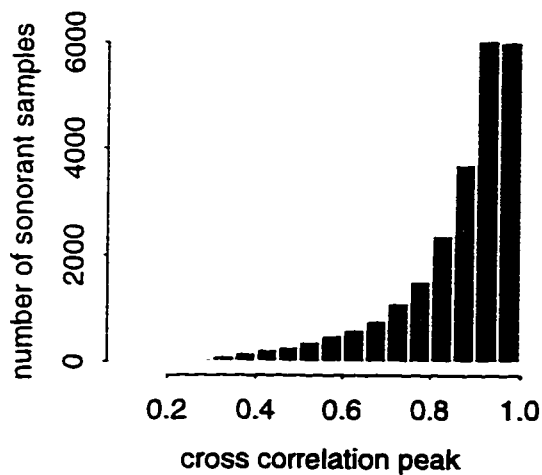
probability compared to the other sonorants due to lower rms values. the peak cross-correlation coefficient was added to this initial parameter set as a separate parameter. Thus, a total of 17 parameters were fed to the tree algorithm. As a result, the parameters listed in Table 7.3 were selected as the optimal set of parameters to distinguish sonorant sounds from obstruent sounds. The resulting classification tree had a correct classification rate of 95.1% on the training data. The distributions of each of the selected parameters for sonorant and nonsonorant sounds are shown in Figure 7.3 and Figure 7.4. It should be pointed out that the energy-based parameters can well be modeled with a Gaussian distribution. However, the voicing probability and the peak correlation coefficient are not Gaussian distributed. In order to examine the sensitivity of each of the selected parameters to gender differences, the gender-dependent distributions of each of these parameters are plotted in Figure 7.5 and Figure 7.6 for sonorant and nonsonorant sounds separately. The figures show that the selected parameters are insensitive to gender differences.



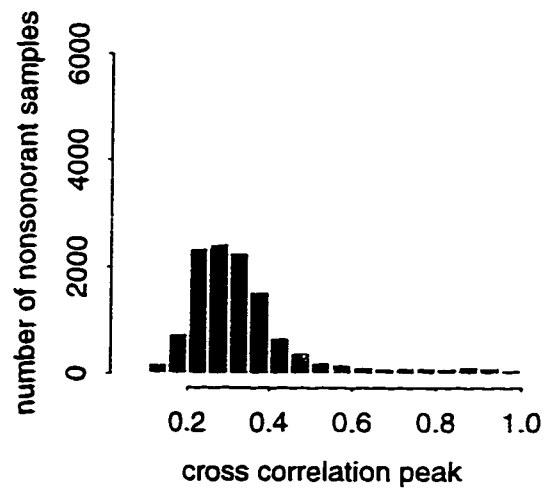
(a) sonorant sounds



(b) nonsonorant sounds

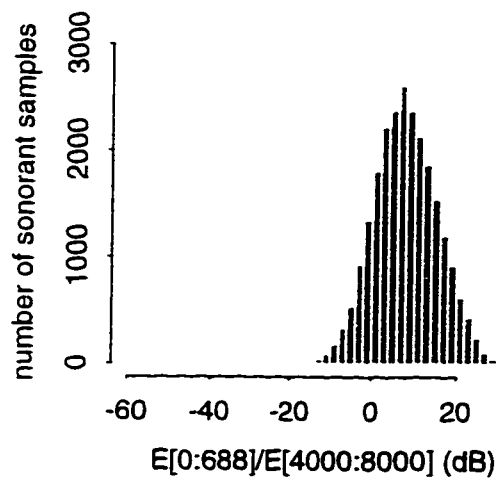


(c) sonorant sounds

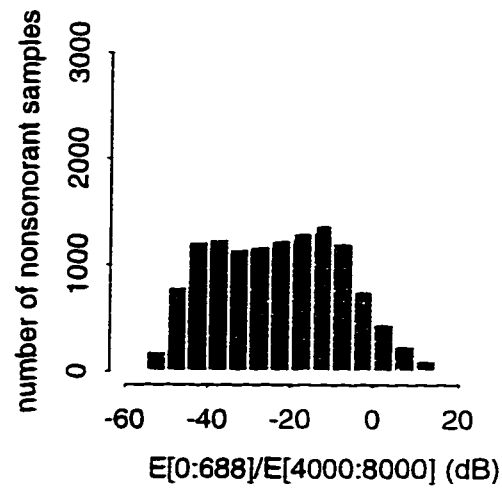


(d) nonsonorant sounds

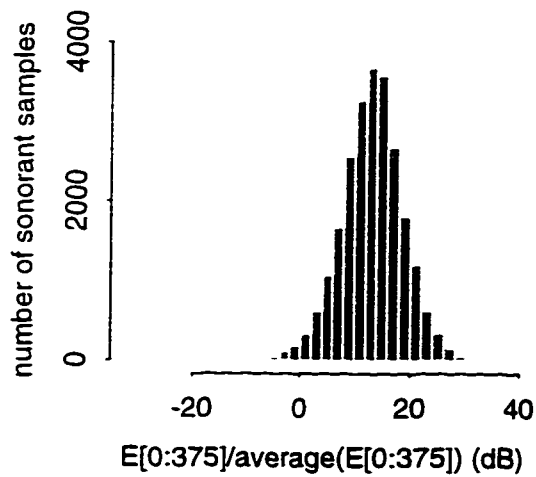
Figure 7.3: Histograms of voicing probability computed over the training data (a) sonorant samples and (b) nonsonorant samples. Histograms of the peak cross correlation coefficient computed over the training data (c) sonorant samples and (d) nonsonorant samples.



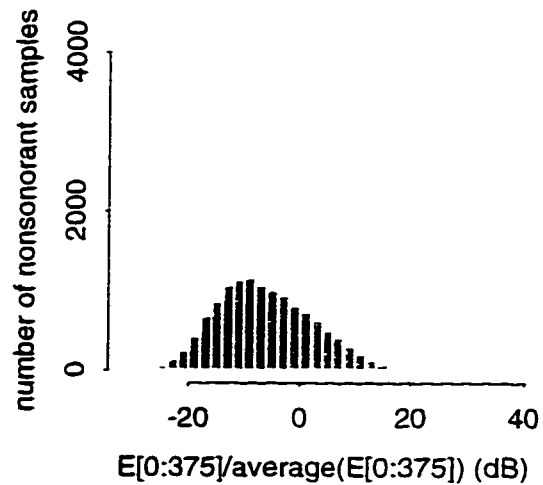
(a) sonorant sounds



(b) nonsonorant sounds

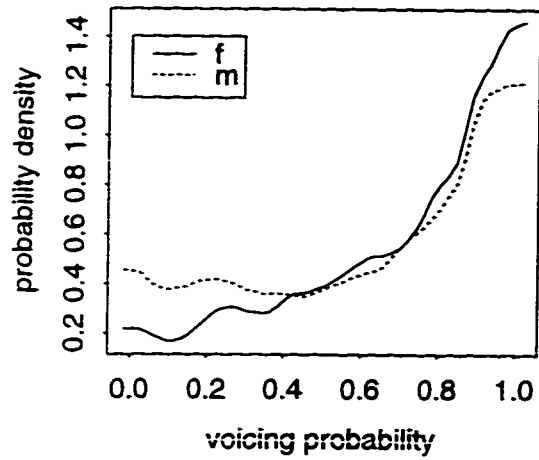


(c) sonorant sounds

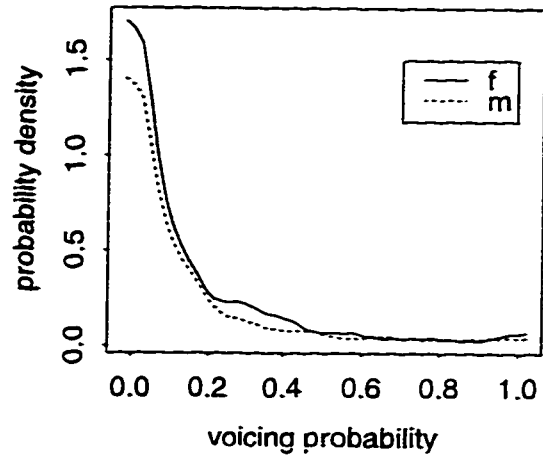


(d) nonsonorant sounds

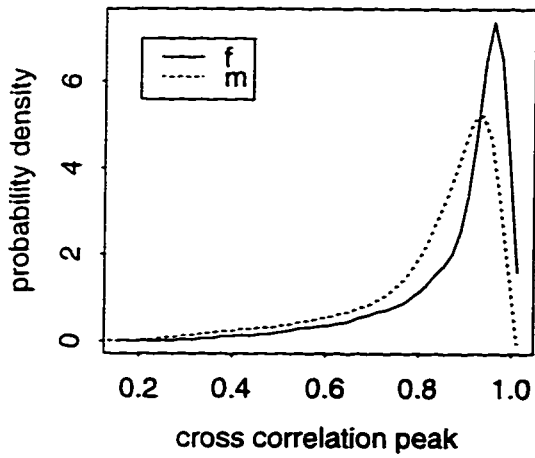
Figure 7.4: Histograms of $E[0:688]/E[4000:8000]$ computed over (a) the sonorant samples and (b) the nonsonorant samples in the training data. Histograms of $E[0:375]/\text{average}(E[0:375])$ computed over (c) the sonorant samples and (d) the nonsonorant samples.



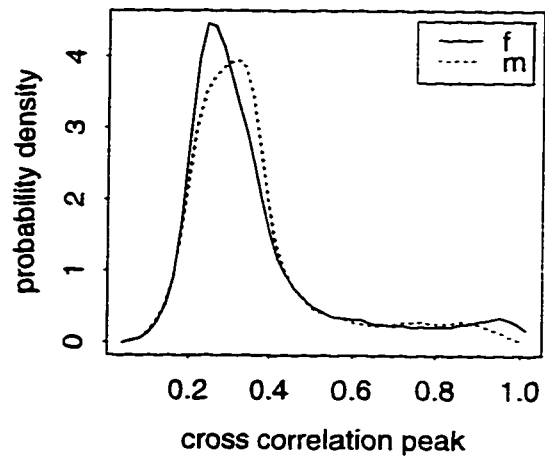
(a) sonorant sounds



(b) nonsonorant sounds

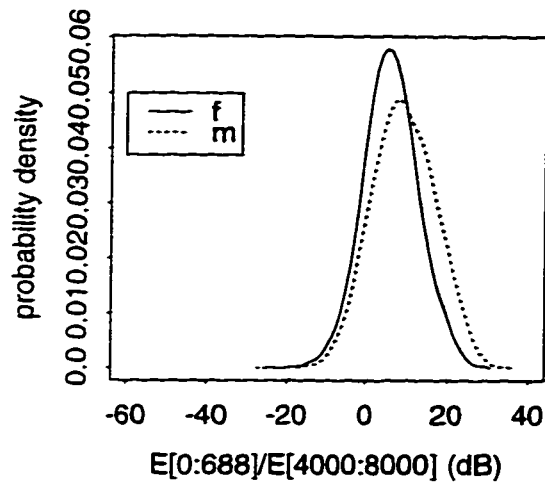


(c) sonorant sounds

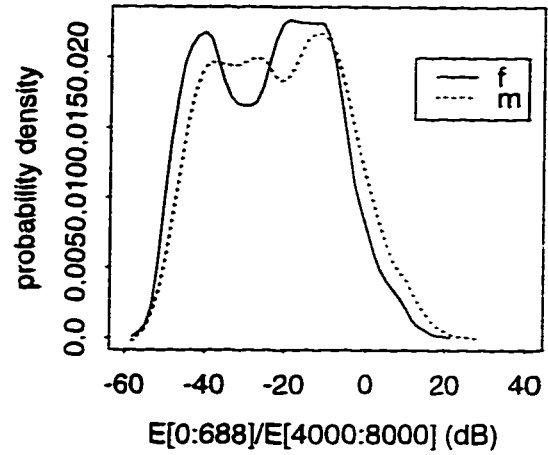


(d) nonsonorant sounds

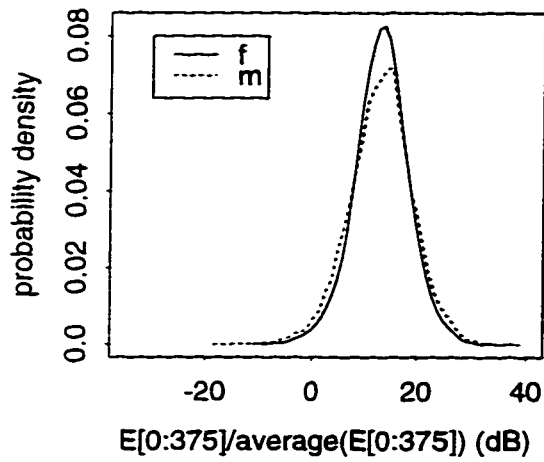
Figure 7.5: The distribution of voicing probability for males and females is shown in (a) for the sonorant samples and in (b) for the nonsonorant samples. The distribution of cross-correlation peak for males and females is shown in (c) for the sonorant samples and in (d) for the nonsonorant samples.



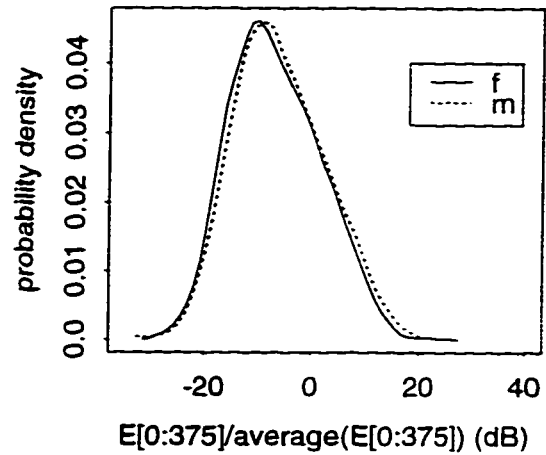
(a) sonorant sounds



(b) nonsonorant sounds



(c) sonorant sounds



(d) nonsonorant sounds

Figure 7.6: The distribution of $E[0 : 688]/E[4000 : 8000]$ for males and females is shown in (a) for the sonorant samples and in (b) for the nonsonorant samples. The distribution of $E[0 : 375]/\text{average}(E[0 : 375])$ for males and females is shown in (c) for the sonorant samples and in (d) for the nonsonorant samples.

7.1.3 Classification Results

In order to evaluate the performance of the selected parameters for the sonorant phonetic feature, the classification tree built in the development stage was used to classify sonorant and nonsonorant samples extracted from the 504 "si" sentences of the TIMIT test set. These sentences are completely independent of the training sentences as explained in Chapter 4. The test set consists of 10730 sonorant samples and 5324 nonsonorant samples. The classification results are summarized in Table 7.4 for both the training set and the test set. Table 7.4. The confusions at the feature level and at the phone level are summarized in Table 7.5. As indicated in Table 7.5, 13.6% of the canonically labeled nonsonorant sounds were classified as sonorant. On the other end, only 1.3% of the canonically labeled sonorant sounds were misclassified as nonsonorant.

Table 7.4: Sonorant-Feature Classification Results

% correct on training data	% correct on test data
95.1	94.6

Table 7.5: Confusion Results among sonorant and nonsonorant sounds

	Sonorant	Nonsonorant
Sonorant	98.7 %	1.3 %
Nonsonorant	13.6 %	86.4 %

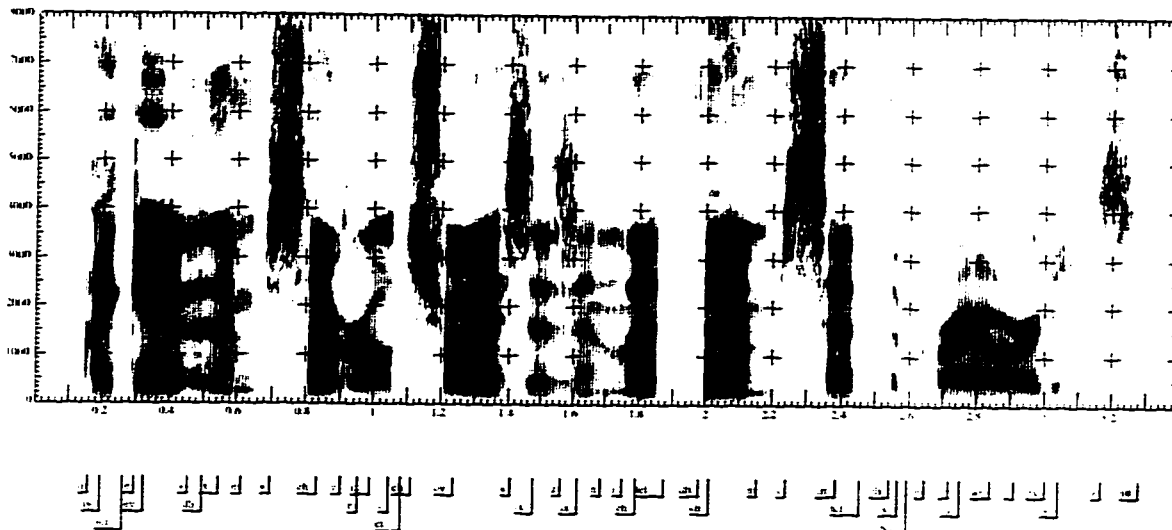


Figure 7.7: This spectrogram shows a sonorant “v” and a sonorant /ð/ between 0.4 and 0.5 seconds. The sentence is “I gave them several choices and let them set the priorities” spoken by a male speaker.

An error analysis revealed that 46% of the nonsonorant sounds classified as sonorant were the two voiced weak fricatives /ð/ and /v/. while the voiced stop consonants /b/, /d/ and /g/ contributed 42.4% to this misclassification. The majority of /v/ and /ð/ that were classified as sonorant occurred between two sonorant sounds and had either a strong voicing probability and strong energy in the F_0 region, as represented by $E[0 : 375]/average(E[0 : 375])$, or a strong voicing probability and relatively weak concentration of energy at high frequency relative to low frequency ($E[0 : 688]/E[4000 : 8000]$). These cases represent 80.5% of /ð/ and /v/ that were misclassified as sonorant. Examples of intervocalic /v/’s and /ð/’s realized as sonorant are depicted in the spectrogram shown in Figure 7.7 between 0.4 and 0.5 seconds. For comparison, Figure 7.8 shows the spectrogram of a canonical fricated /v/ occur-

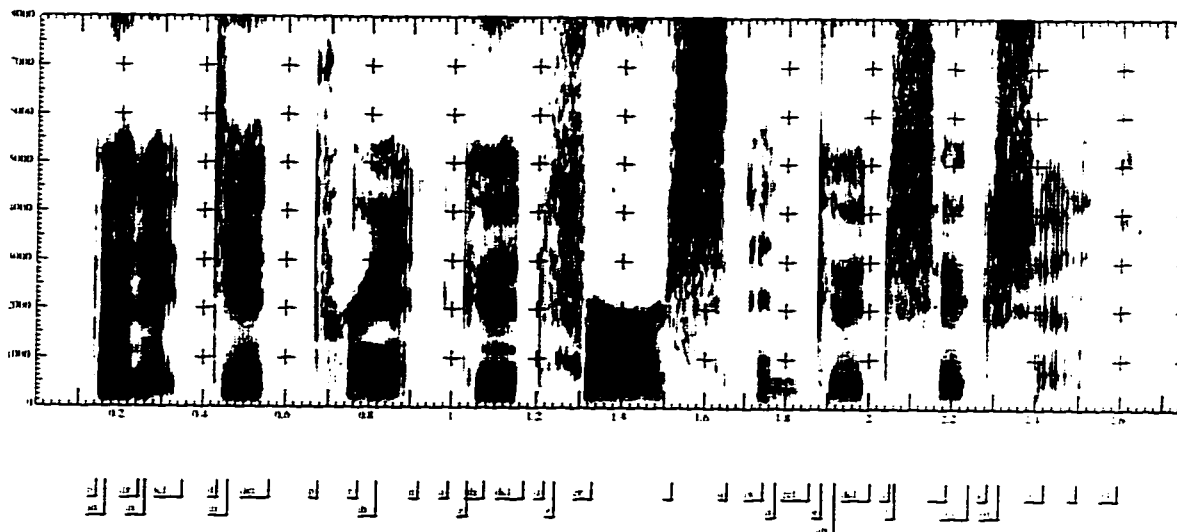


Figure 7.8: This spectrogram shows a canonical fricated “v” at about 1.2 seconds. The sentence is “But it did print good verse and good fiction” spoken by a female speaker.

ing between 1.24 and 1.3 seconds. Most of the misclassified voiced stop consonants have strong cross-correlation peak and relatively weak concentration of energy at high frequency, as represented by $E[0 : 688]/E[4000 : 8000]$. The majority of these stop consonants were followed by a sonorant sound but were preceded by a labeled closure interval. It is possible that this closure interval is also realized as sonorant and that the stop closure and stop release occur between two sonorant sounds. Further examination of these confusions is needed in order to determine if this was the case. In general, a voiced obstruent that occurs next to a sonorant sound, and most often between two sonorant sounds, can be realized with a weakened constriction so that it is manifest as sonorant (as reported in [54], [41]). Such weakened constriction results in a stronger glottal source relative to the pressure source forward in the vocal tract to

be the dominant source in the sound generation. This phenomenon has implications for an event-based approach to recognition where this type of coarticulation must be accounted for in the lexicon or as part of the lexical-access process. In addition, this observation has implications for speech synthesis by rules where the synthesis of voiced obstruents, in some contexts, may sound more natural if stronger voicing energy as compared to frication energy is used.

An analysis of the sonorant misclassifications showed that there is no outstanding trend in the declared errors. This is perhaps due to the very few misclassifications in this case. However, it should be noted that about 21% of the misclassified sonorant sounds with weak voicing probability and strong high-frequency energy were labeled as the front vowels /ɪ/, /iʏ/ and /ɛ/. The strong high-frequency energy of these sounds is due to that fact that they have high second and third formants. However, examining spectrograms of some of these sounds showed that they also have weak energy in the F_0 - F_1 (first formant) especially toward the end of a sentence. In addition, the nasal sounds (/m/, /n/ and /ŋ/) contributed about 34% to the sonorant errors with low voicing probability and strong high-frequency energy. While the reason that some nasal sounds have low voicing probability is justifiable, it is not clear at this time why these sounds can have stronger energy at high frequencies as compared to low frequencies.

7.2 Anterior Place-of-Articulation for Stridents

Acoustic parameters that target the "anterior" phonetic feature were derived to distinguish among the strident obstruents. The acoustic parameters consist of energy

ratios that measure the energy in one frequency band of the spectrum relative to another. Selection of the frequency band was based on the third formant value, F_3 , to reduce the effects of speaker differences on the parameters' values. The selected parameters are listed in Table 7.8.

Strident sounds in American English are the fricatives: /s/, /z/, /ʃ/ and /ʒ/ and the affricates: /tʃ/ and /dʒ/. Strident fricatives and affricates are characterized by a “strong” turbulent noise that distinguishes them from the nonstrident or weak fricatives: /f/, /v/, /θ/ and /ð/. The strident sounds can be classified into two categories: anterior and nonanterior. Anterior refers to the place in the vocal tract in front of the alveolar ridge. Thus, anterior sounds are produced with a constriction in front of the alveolar ridge. This is depicted for the anterior sound /s/ in Figure 7.9. In contrast, nonanterior sounds are produced with a constriction behind the alveolar ridge, as depicted in Figure 7.10 for the /ʒ/ sound.

7.2.1 Acoustic Parameters to Identify the Anterior Place-of-Articulation for Stridents

The constriction location in the vocal tract plays a major role in determining the spectral shape of the strident sounds. In producing these sounds, air flow from the lungs turns into a turbulent noise source at the constriction exciting the vocal cavity in front of it. Thus, the spectral shape of strident sounds is determined by the acoustic structure of this front vocal-cavity. Using acoustic-phonetic theory that models the vocal tract as a concatenation of acoustic tubes [33] [32], it can be shown that the resonances of a vocal cavity are inversely proportional to its length. The longer

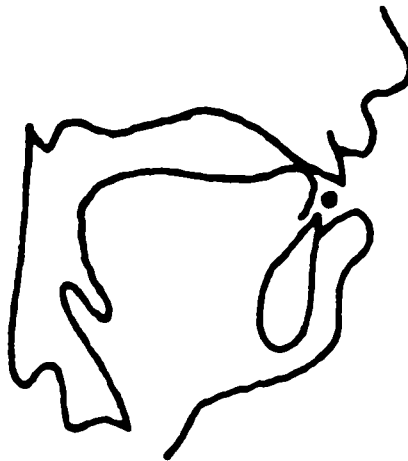


Figure 7.9: This figure shows the constriction formed by the tongue in front of the alveolar ridge during the production of the anterior sound /s/. (Taken from Kent [63])

the cavity, the lower the frequencies of its resonances. Therefore, the anterior sounds, with a shorter front cavity than their nonanterior counterparts (c.f. Figures 7.9 and 7.10), have higher-frequency resonances. For illustration, a spectrogram of the utterance “approach your interview with statuesque composure” is shown in Figure 7.11. Comparing the time segment associated with /s/ to that associated with /ʒ/ in Figure 7.11, it can be seen that /ʒ/ contains more energy than /s/ in the lower part of the spectrum starting around 2300 Hz.

In this study, the focus was on deriving parameters that capture the aforementioned characteristics of the strident sounds. As a result, the anterior sounds, /s/ and /z/, that share the same place of articulation, were considered as one group while the

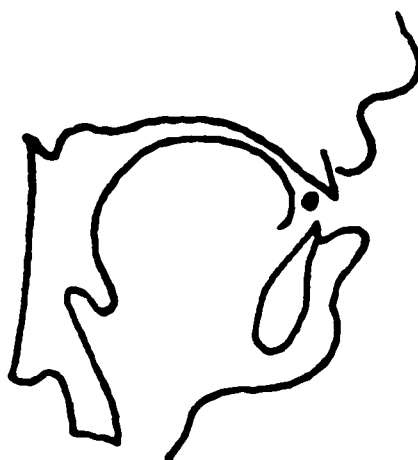


Figure 7.10: This figure shows the constriction formed by raising the tongue dorsum against the palate, behind the alveolar ridge, during the production of the nonanterior sound /ʃ/. (Taken from Kent [63])

nonanterior sounds /ʃ/, /ʒ/, /ç/ and /ʝ/ ¹ were grouped into another.

Several people have studied the spectral characteristics of the fricative sounds in English, including those of the stridents. These studies range from articulatory modeling to acoustic analysis (c.f., [64], [65], [66]). Parameters that have been looked at include the first spectral moment (or center of gravity), and the location of the main spectral peaks. Locating the spectral peak was mainly based on linear prediction analysis. These studies were generally constrained in two ways: (1) the limited number of speakers (usually less than 10 speakers) and utterances, and (2) the constrained contextual environments where only fricatives in word-initial position and in the context of a few vowels were considered.

¹The affricate sounds start with an articulation at their release that is usually more fronted than that of /ʃ/ and /ʒ/ but move fast to a place of articulation that is similar to that of /ʃ/ and /ʒ/. Furthermore, for most of their duration, the affricates look like the nonanterior fricatives on a spectrogram. This is the reason for grouping the affricates with the nonanterior fricatives.



Figure 7.11: Spectrogram of the utterance “Approach your interview with statuesque composure” spoken by a female speaker. The alveolar /s/ fricatives at about 1.5 and 2.1 seconds have strong energy starting at about 4000 Hz, whereas the palatal affricates at 0.55 and 1.82 seconds and the palatal fricative /ʒ/ at 2.77 seconds have strong energy starting at about 2000 Hz and 2300 Hz, respectively.

7.2.2 Optimized Parameters

In this study, the TIMIT database described in Chapter 4 was used to develop parameters that distinguish among the strident obstruents based on place of articulation. All strident sounds were considered independent of context. Parameters suggested in the literature as well as new energy-based parameters were explored. Speech was analyzed using a 25-ms Hamming window, or 400 sample points at 16 kHz sampling rate, and a 5-ms frame rate. The spectrum of the signal within each time window was computed using a 512-point Discrete Fourier Transform (DFT). In this section, the explored parameters are described.

Spectral Center of Gravity:

The spectral center of gravity, also referred to as the first spectral moment, at time

frame i was computed as:

$$M_i = \frac{\sum_{j=32}^{j=256} j A_i[j]}{\sum_{j=32}^{j=256} A_i[j]} * \frac{16000}{512} (Hz) \quad (7.2)$$

where $A_i[j]$ is the spectral amplitude at time i and the DFT bin j . DFT bin 32 corresponds to 1000 Hz while DFT bin 256 corresponds to 8000 Hz, the highest frequency for the TIMIT speech. The mean center of gravity within the strident duration was used as a candidate parameter for identifying the anterior/nonanterior feature. It was expected that this parameter would have lower values for the nonanterior sounds as they contain more energy in the lower part of the frequency spectrum in comparison to their anterior counterparts.

Spectral Peak:

Spectral peak is the frequency location of the highest-amplitude point in the spectrum of a speech frame. This parameter was averaged across the duration of a strident sound and considered as a candidate parameter for the anterior/nonanterior distinction.

Center-of-Gravity and Spectral-Peak in Barks

The Bark scale [67] was invented to mimic the frequency warping that happens in the human auditory system. Various researchers have used this scale in looking at formant values for different vowel sounds. In this study, the mean center of gravity and the mean spectral-peak for each strident obstruent were computed in Barks as well as in Hz. The relationship between frequency F in kHz and frequency B in Barks, for $F > 250 Hz$ is given by:

$$B = 13 * \arctan(0.76F) + 3.5 * \arctan((F/7.5)^2) \quad (7.3)$$

The center-of-gravity and the spectral-peak parameters, in Hz and Barks, were pre-

sented to the tree optimization procedure in two ways: unnormalized and normalized. When normalized, each parameter was measured relative to $F3$. This normalization is intended to reduce the effect of differences in speaker's vocal-tract length on the parameter values. $F3$ estimation was based on a procedure similar to that used for estimating $F0$. This procedure consists of the following steps:

1. Given a speech utterance, estimate the first four formant trajectories using the LPC-based ESPS formant-tracking algorithm [60].
2. Using the ESPS joint formant/pitch tracker, compute the voicing probability (see Section 7.1) of each speech frame.
3. Get the $F3$ value for each speech frame with a voicing probability of 0.8 or larger. This restriction tends to eliminate the nasal sounds and those voiced obstruents with high voicing-probability values. This is desired since the all-pole assumption of the LPC-model does not hold for these sounds.
4. In order to reduce the effect of erroneous $F3$ values and to automatically eliminate /r/'s which have an unusually low $F3$ value. $F3$ is estimated in two passes:
 - (a) Compute the mean and standard deviation of those $F3$ values obtained in step 3.
 - (b) Eliminate the $F3$ values, obtained in step 3, that exceed the mean $F3$ computed in step 4.a by more than a standard deviation. Estimate $F3$ as the mean of the remaining $F3$ values. This estimate will be referred to as $\hat{F}3$ in the rest of this chapter.

Table 7.6: Parameters that were derived from center-of-gravity and spectral peak.

Center-of-Gravity	Spectral Peak
\hat{M} (Hz)	\hat{P} (Hz)
\hat{M} (Barks)	\hat{P} (Barks)
$\hat{M} - \hat{F}3$ (Hz)	$\hat{P} - \hat{F}3$ (Hz)
$\hat{M}/\hat{F}3$	$\hat{P}/\hat{F}3$
$\hat{M} - \hat{F}3$ (Barks)	$\hat{P} - \hat{F}3$ (Barks)

The fourth formant value, $F4$, was also considered as a basis for normalization. However, preliminary analysis with energy-based measures, revealed that $F3$ is a better normalization factor. Measuring the center-of-gravity and the frequency of the spectral peak relative to $\hat{F}3$ was done as a difference and a ratio in the Hertz scale and as a difference in the Bark scale. Thus, there were a total of 10 parameters based on the center-of-gravity and the spectral-peak. These parameters are listed in Table 7.6 where, \hat{M} denotes mean center-of-gravity and \hat{P} denotes mean spectral-peak.

Energy-based parameters

The energy-based parameters are intended to capture the stridents' energy distribution across the frequency spectrum. Acoustic-phonetic knowledge, gathered from acoustic studies on the fricatives [64], spectrogram reading experiments and acoustic-phonetic theory [32], reveals that the energy of nonanterior strident sounds is mainly contained in the region of $F3$ and above. This distribution of strong energy usually starts around 2000 Hz, about 1000 Hz below $F3$ (see Figure 7.10). The energy of anterior strident sounds is contained in the $F4$ - $F5$ region and usually starts around

3500 Hz. Based on these observations, several energy-based parameters were formulated. The explored parameters are listed in Table 7.7 where, \bar{E} , $maximum(E)$ and $minimum(E)$ denote average, maximum and minimum energy across the utterance, respectively. $E[f_1 : f_2]$ denotes energy in the frequency band delimited by f_1 and f_2 . In addition, 0 and 8000 refer to the lowest and highest frequency values in Hz, respectively, since speech was sampled at 16 kHz. The first four parameters attempt to capture the spectral energy distribution within a strident sound. Each of these four parameters was normalized with respect to its average value across the utterance. The normalized parameters correspond to the fifth through the eighth parameter in Table 7.7. Each of the last nine parameters in Table 7.7 measures the energy in different parts of the spectrum relative to its average, minimum and maximum values across the utterance. All but the first four parameters incorporate information about time instants outside the time spans of the stridents. This is an attempt to discover if there is important information outside the strident time span that can help identify the strident place of articulation. Energy computed within a strident sound was always averaged across the strident duration.

To select the best $[f_1, f_2]$ pairs, in the Fisher-criterion sense, the 1268-sentence optimization subset of the TIMIT training set was used. This subset consists of 3379 anterior strident sounds and 1084 nonanterior strident sounds. Two experiments were conducted in order to investigate the effect of $F3$ normalization. In the first experiment, f_1 and f_2 were independent of $F3$ while in the second experiment they were a function of $F3$.

In the first experiment, the range of f_1 was chosen to be [1500, 6700] Hz and that of f_2 was chosen to be [1800, 7000] Hz. The only constraint was that $(f_2 - f_1) \geq 300$

Table 7.7: Energy-Based Parameters for the anterior/nonanterior feature.

$E[f_1 : f_2]/E[f_2 : 8000]$
$E[f_1 : f_2]/E[0 : f_1]$
$E[f_1 : f_2]/E[0 : 8000]$
$E[0 : f_1]/E[f_2 : 8000]$
$(E[f_1 : f_2]/\bar{E}[f_1 : f_2])/(E[f_2 : 8000]/\bar{E}[f_2 : 8000])$
$(E[f_1 : f_2]/\bar{E}[f_1 : f_2])/(E[0 : f_1]/\bar{E}[0 : f_1])$
$(E[f_1 : f_2]/\bar{E}[f_1 : f_2])/(E[0 : 8000]/\bar{E}[0 : 8000])$
$(E[0 : f_1]/\bar{E}[0 : f_1])/(E[f_2 : 8000]/\bar{E}[f_2 : 8000])$
$E[0 : f_1]/\bar{E}[0 : f_1]$
$E[f_1 : f_2]/\bar{E}[f_1 : f_2]$
$E[f_2 : 8000]/\bar{E}[f_2 : 8000]$
$E[0 : f_1]/\text{maximum}(E[0 : f_1])$
$E[0 : f_1]/\text{minimum}(E[0 : f_1])$
$E[f_1 : f_2]/\text{maximum}(E[f_1 : f_2])$
$E[f_1 : f_2]/\text{minimum}(E[f_1 : f_2])$
$E[f_2 : 8000]/\text{maximum}(E[f_2 : 8000])$
$E[f_2 : 8000]/\text{minimum}(E[f_2 : 8000])$

(Hz). The lower edge of the f_1 range was set below 2000 Hz, the empirically-observed frequency at which the nonanterior strident sounds start to have strong energy. The selected frequency ranges ensure the inclusion of spectral bands where nonanterior and anterior sounds have their largest energy concentration. Based on these constraints, the two-stage parameter optimization process described in Chapter 6 was used to select an optimal parameter set. The first-stage, the Fisher-criterion stage, of the optimization process was run on each of the generic parameters described in Table 7.7. As a result, 45 parameters were obtained. These 45 parameters were fed to the second-stage, classification tree algorithm, of the parameter selection process. The classification tree selected 7 parameters yielding an overall correct classification rate of 92.6% on the training data. The first 3 parameters, out of the selected 7, resulted in 91% correct classification of the training data. Since the interest is in a parsimonious representation, only these three parameters were selected as the final set in this experiment. The top parameter contributed about 89% to this classification.

In the second experiment, f_1 and f_2 in Table 7.7 were made dependent on $F3$. Based on visual inspection of spectrograms, the results of the first experiment and the acoustic knowledge by which this study was motivated, the range of f_1 was chosen to be $[F3 - 1000, F3 + 1700]$ (Hz) and that of f_2 was chosen to be $[F3 - 700, F3 + 2000]$ (Hz) with the constraint that $(f_2 - f_1) \geq 300$ (Hz). Thus, the band $[f_1, f_2]$ was moved based on $F3$ values. The Fisher-criterion stage of the parameter optimization process resulted in 19 parameters. These parameters were used in the tree-growing stage of the optimization process to further reduce the parameter set. The classification tree selected 3 parameters that resulted in an overall correct classification rate of 92.6%. Of these 3 parameters, the top parameter contributed 91% to the correct classifi-

cation. Thus, the parameters selected in this experiment gave better classification results in comparison to the non-normalized parameters of the first experiment (first experiment yielded 91 % with three parameters). Furthermore, examination of the top-parameters' distributions for the anterior and nonanterior strident sounds as a function of gender showed that the $F3$ -normalized parameters better reduced speaker-gender effects. The top parameter distributions with and without $F3$ normalization are shown in Figure 6.2, as a function of gender, for the anterior sounds. Thus, the $F3$ -normalized parameters were selected for further studies.

Final Parameter Set:

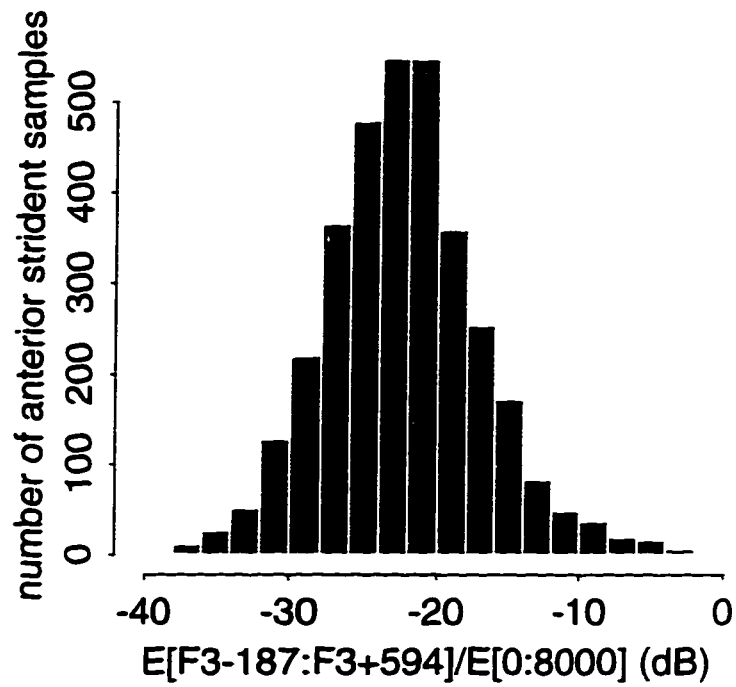
In order to obtain the final parameter set for distinguishing among the anterior and nonanterior strident sounds, the center-of-gravity and spectral-peak parameters listed in Table 7.6 and the 19 $F3$ -normalized energy parameters determined from the Fisher-criterion stage were used to grow a classification tree. Based on this tree, the energy ratios distinguished best among the anterior and nonanterior strident sounds. The selected parameters are shown in Table 7.8. In addition, the contribution of each of the selected parameters to the correct classification is shown in the second column of Table 7.8. That is, the top parameter $E[F3 - 187 : F3 + 594]/E[0 : 8000]$ resulted in 91% correct classification. This parameter distribution is shown in Figure 7.12 for the anterior and nonanterior strident classes and in Figure 7.13 for each of the strident phones. Adding $E[F3 - 125][F3 + 656]/E[F3 + 656 : 8000]$ to the top parameter increased the correct classification rate to 91.6% while the addition of the third parameter $E[F3 - 781 : F3 + 312]/E[0 : F3 - 781]$ increased the overall correct classification to 92.6%.

Error analysis did not reveal any error trend that was dependent on context. How-

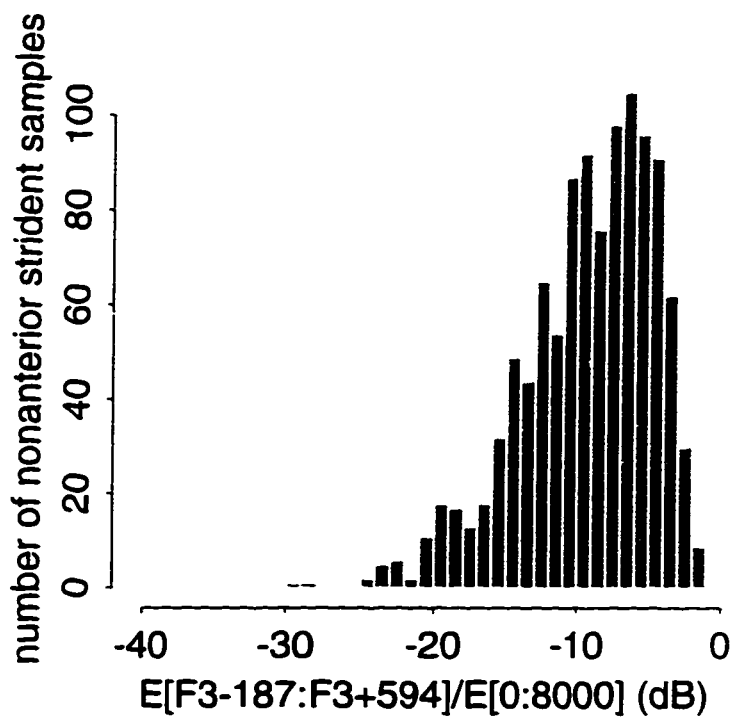
Table 7.8: The parameters selected by the optimization process to distinguish among the anterior and nonanterior strident sounds. The % correct in each row is the correct classification rate obtained by adding the parameter in that row to the parameter(s) in the previous row(s).

Parameter	% correct classification
$E[F3 - 187 : F3 + 594] / E[0 : 8000]$	91.0
$E[F3 - 125][F3 + 656] / E[F3 + 656 : 8000]$	91.6
$E[F3 : 781][F3 + 312] / E[0 : F3 - 781]$	92.6

ever. this error analysis showed that most of the anteriors that were confused with the nonanteriors were spoken by male speakers. In addition, most of the nonanteriors that were confused with the anteriors were spoken by female speakers. The gender-biased errors occurred despite the fact that the $F3$ -based speaker-normalization reduced the differences between females and males in the considered parameter spaces. However, it seems that this normalization was not enough. Examination of spectrograms of sentences where these errors occurred showed that the peak spectral location for an anterior sound was still higher than that of the nonanterior sounds when the two categories occurred in the same sentence. Based on this observation, another parameter set was considered and added to the parameter pool. The added parameters measure the average spectral-peak location and average center-of-gravity across a strident sound relative to their respective maximum values across the utterance. These relative measurements were differences and ratios in the natural-frequency (Hz) scale and differences in the Bark scale.



(a)



(b)

Figure 7.12: Histogram of $E[F3 - 187 : F3 + 594]/E[0 : 8000]$ for (a) the anterior samples and (b) the nonanterior samples

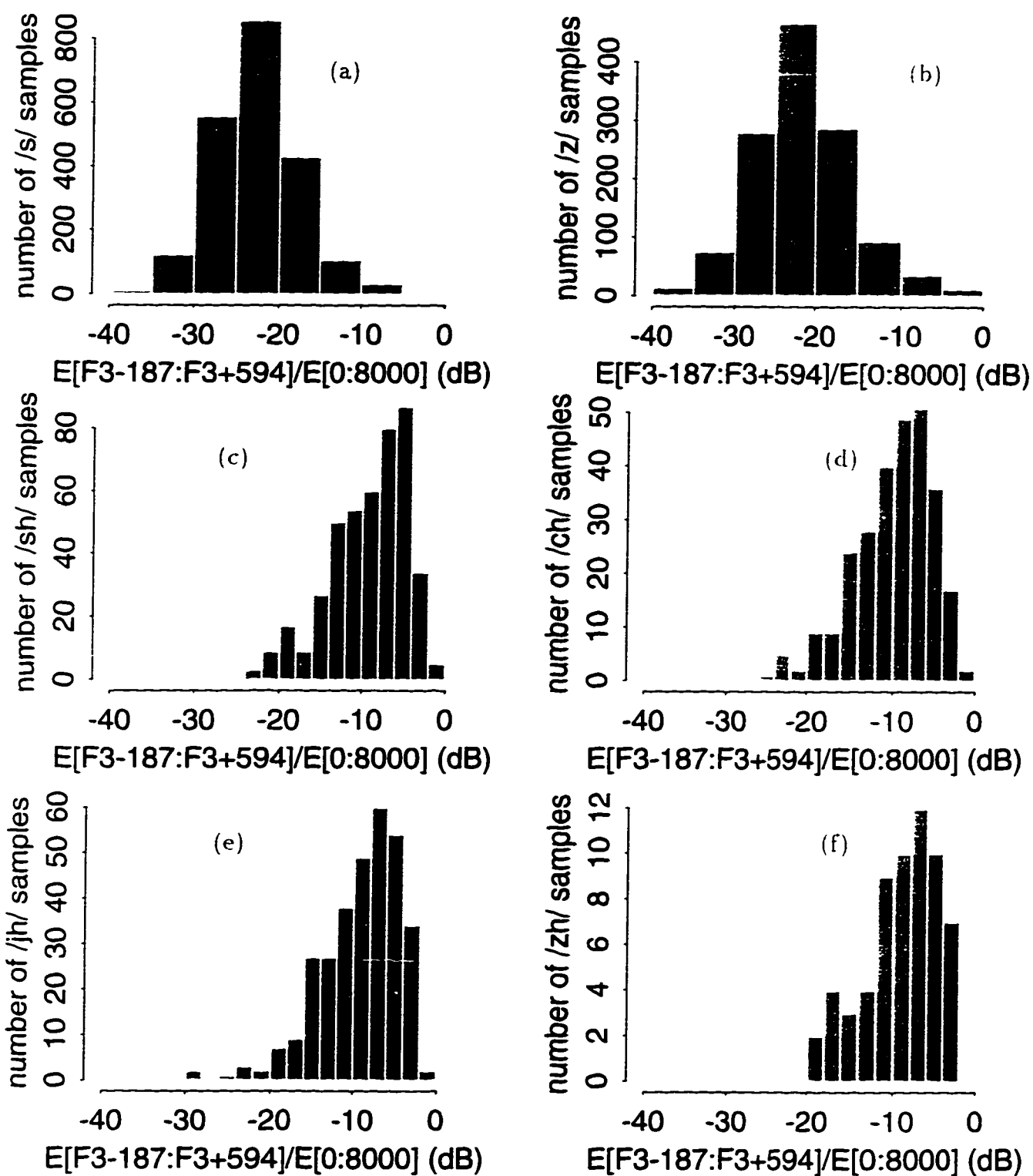


Figure 7.13: The distributions of $E[F3 - 187 : F3 + 594]/E[0 : 8000]$ for the anterior strident phones /s/ and /z/ in (a) and (b), respectively and for the nonanterior strident phones /š/, /č/, /j/ and /ž/ in (c), (d), (e) and (f), respectively.

A classification tree was built with the new parameters in the same pool as the rest of the parameters considered earlier. The resulting tree selected the first two parameters listed in Table 7.8 in addition to $\bar{P}/\max(P)$, the ratio of the spectral-peak location within the strident to the maximum peak location across the utterance. The three selected parameters yielded a correct classification rate of 93.5%, i.e., a correct classification increase of 0.9% over the parameters listed in Table 7.8.

In conclusion, it was clear that more than one cue in the speech utterance can be used to normalize for speaker-dependent effects and to disambiguate confusions that may arise between different sounds. Perceptual experiments to verify this conclusion can be conducted. In these experiments, sounds that were erroneously classified based on the parameters listed in Table 7.8 can be cut and played to human listeners in isolation from the rest of the sentence. Then, the whole sentences containing these sounds would be played to the listeners. In addition, when samples of anterior and nonanterior sounds are available from the same speaker, these sounds can be played in pairs, one anterior and nonanterior sound, and the listener would be asked to identify each sound in the pair. If the second and third experiments result in significantly better identifications of the sounds in question, one can conclude that human listeners rely on information contained in a wide-time window that extends beyond a sound duration to identify that sound.

7.2.3 Classification Results

In order to evaluate the performance of the selected parameters, the tree classifier built during the development stage was used to classify new test data. In this classifier, the

Table 7.9: Classification results on the training and test sets for the anterior and nonanterior strident sounds. The classifier was the classification tree obtained in development. The parameters in Table 7.8 were the only used in the classification tree.

% correct on training data	% correct on test data
92.6	92.0

parameters listed in Table 7.8 were used. The test data consisted of all anterior and nonanterior strident samples extracted from the set of 504 “si” TIMIT test sentences. The results of this classification are shown in Table 7.9. Comparing the results on the test data to those on the training data, it can be seen that little degradation in performance was observed indicating the generality of the parameters across data sets.

7.3 Labial, Alveolar and Velar Place-of-Articulation Parameters for the Stop Consonants

Acoustic parameters that distinguish among the English stop consonants based on place of articulation were derived. The best parameters measure the energy in one frequency band relative to another within the duration of the stop consonant.

The English stop consonants are the labials /b/ and /p/, the alveolars /d/ and /t/ and the velars /k/ and /g/. The stops are canonically realized by first forming

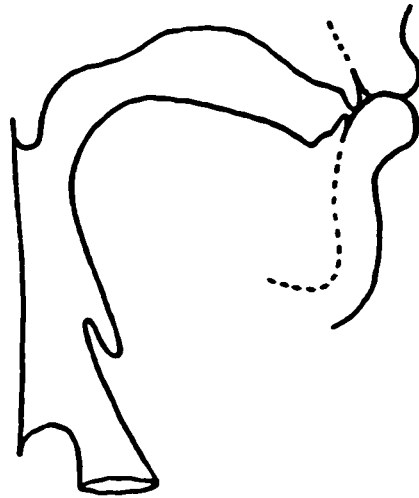


Figure 7.14: The shape of the vocal tract during the production of the labial stop /b/. Labial stop consonants are produced by forming a complete closure at the lips. a complete closure in the vocal tract that blocks the air flow out of the mouth. Air pressure is built-up behind the closure and then abruptly released. The acoustic consequence of this pressure release is observed in the acoustic signal as an abrupt energy change referred to as the stop burst. The stop burst is followed by frication and then aspiration noises that excite the vocal tract section in front of the constriction. Thus, the English stops are characterized by the closure location in the vocal tract. Labial stop consonants are produced with a closure at the lips as depicted in Figure 7.14. The alveolar stops are produced with a closure formed by the tongue tip at the alveolar ridge (see Figure 7.15) while the velar sounds are produced with a closure formed by the tongue dorsum in the velum area (see Figure 7.16)

Acoustic cues that distinguish among the English stop consonants have been studied by many researchers. Most studies were confined to stops occurring in word-initial and prevocalic positions. In addition, the vowel context was most often confined to

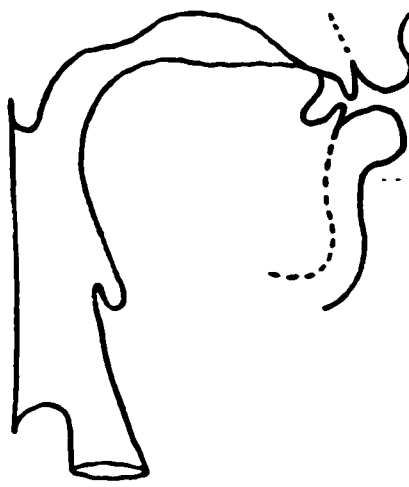


Figure 7.15: The shape of the vocal tract during the production of the alveolar stop /d/. Alveolar stop consonants are produced by forming a complete closure at the alveolar ridge with the tongue tip.

the cardinal set. Among the cues considered were the formant frequencies at the burst release, formant transitions from a stop to a following vowel and the stop-burst spectral shape. In one study, Blumstein and Stevens [68] manipulated the stop burst spectral shape, formant transitions and the duration of the frication-aspiration noise in synthesized stop consonants preceding vowels. They presented the synthesized utterances to human subjects for perceptual tests. Blumstein and Stevens concluded that the overall spectral shape of a stop consonant is the primary cue to its identity while formant transitions can serve as secondary cues to disambiguate confusable cases.

Lamel [69] derived a set of acoustic cues and rules to identify stop consonants. Lamel's rule-based system was able to identify stop consonants with about 71% ac-

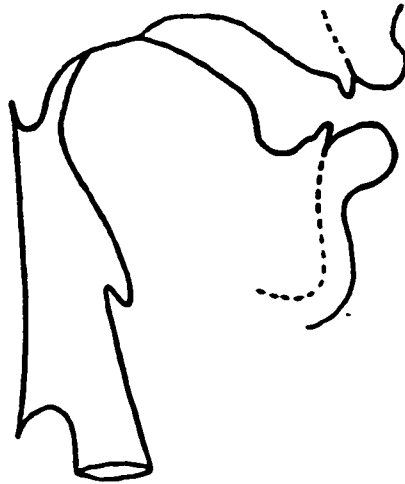


Figure 7.16: The shape of the vocal tract during the production of the velar stop /g/. Velar stop consonants are produced by forming a complete closure in the velum area with the tongue dorsum.

curacy. The acoustic cues were based on formant transitions at stop closures and stop releases in addition to stop spectral shape. These cues are similar to the ones examined by Stevens and Blumstein.

In Kopec and Bush [70], an expert was able to correctly identify a stop consonant in CVC syllables, with the stop being the initial consonant in the syllable, at rates ranging between 76% and 81% depending on the type of information presented to the expert.

Lahiri *et al.* [37] looked at the LPC-smoothed burst spectrum of labial and alveolar stop consonants in word-initial position relative to the spectra at the onset of the following vowel. They were able to identify a stop as alveolar with 94% accuracy and labial with 98% accuracy based on the magnitude of hand-picked peaks at the stop

release relative to the magnitude of corresponding spectral peaks in the following vowel. Lahiri et. al's algorithm was automated and tested by Zierden [39] on a database of words with labials and alveolars in initial-word position. The speech material in this database was recorded by 2 males and 2 females. Zierden found that Lahiri's algorithm, when automated, resulted in a much lower performance.

Sussman [71] derived a set of locus equations to identify the place-of-articulation of a stop consonant. Sussman's utterances were stop-vowel-t sequences where the stop was voiced. The locus equations attempt to model the second formant, F_2 , transition from vowel onset to mid-vowel. In developing and testing the validity of his equations using Linear Discriminant Analysis, Sussman picked formant values by hand. The classification rate achieved by Sussman was 100% given that he had a separate locus equation per stop category for each of his 20 subjects and that his development set and test set were the same. More recently, other researchers have investigated the first four spectral moments [72] as parameters for automatic stop classification but their testing was also confined to prevocalic stops.

The difficulty in comparing the results of the different studies lies in the fact that they all used different speech databases and different evaluation methods. Thus, no conclusion can be made as to what parameters can best characterize the stop consonants.

7.3.1 Acoustic Parameters for Identifying the Stop Place of Articulation

In this research, the focus was on parameters that estimate the spectral shape of the stop consonants. The motivation for such parameters stems from the Blumstein and Stevens perceptual study which found that spectral shape constitutes a primary cue for identifying a stop consonant. Other parameters that rely on formant values were not investigated in this thesis. It is worth pointing out that researchers who had success with formant-based parameters assigned the formant values to speech frames based on their visual inspection of the speech spectra. This was done in order to avoid error-prone formant tracking techniques that may result in incorrect formant estimation leading to incorrect decisions. However, it is our belief, in accordance with Blumstein and Stevens's study, that formant transitions could play a secondary role rather than a primary role in disambiguating stops that are easily confused based on their spectral shape alone. This observation is based on inspection of stop-consonant spectrograms and analysis of classification errors obtained in this research. Thus, formant-based parameters should be further investigated.

Several acoustic studies show that labial stops have a relatively flat spectrum² due to the absence of a cavity in front of the closure place. On the other hand, the alveolar stops are characterized by a strongly rising spectrum due to the short front cavity. Finally, the velars have most of their energy concentrated in the mid-frequency part of the spectrum since the closure-place is almost at the mid-point of the vocal tract.

²The spectrum of a stop consonant could be slowly falling or slowly rising, as opposed to being flat, due to lip perturbations.

In fact, one of the cues to velar-stop identification, used in spectrogram reading, is a distinct energy blob in the mid-frequency region where $F2$ and $F3$ of preceding and proceeding vowels converge. This point of $F2 - F3$ convergence is often referred to as the velar pinch.

A visual spectral analysis of stop consonants occurring in various contexts was also conducted in this research. This spectral analysis revealed that the spectral shape of the stop consonants is most consistent with the theory during the first few frames of frication, when it exists, following the instant of stop release. This frication noise was overlooked in most previous studies of stop consonants. Perhaps, it is the frication spectral shape that led the human subjects in Blumstein and Stevens to better identify stop consonants when frication of sufficient duration was included in the synthesized stops. Thus, parameters that target the spectral shape at stop release and during frication noise were sought. In addition, spectral analysis revealed that the velar stops have a spectral peak in the $F4 - F5$ region. This spectral peak was mostly distinct in velars occurring in back-vowel contexts where the velar pinch, in $F2 - F3$ region, appeared to be lower than in front-vowel contexts.

In order to capture the spectral shape characteristics of the stop consonants, acoustic parameters were designed to measure the energy in parts of the spectrum relative to others. In addition, the first two spectral moments, the skewness and the Kurtosis coefficients utilized in [72] were also considered. A spectral moment is adapted from the probabilistic definition of moments [73]. In this adaptation, a DFT point takes the place of the random variable. The probability of each random variable value is substituted by the energy at that DFT point normalized by the total energy (i.e. sum of magnitude squares of the DFT points). Thus, for a given speech frame,

the first three spectral moments are defined successively as:

$$\begin{aligned}\mu_1 &= \frac{\sum_{i=0}^{N/2} i A^2[i]}{\sum_{i=0}^{N/2} A^2[i]} \\ \mu_2 &= \frac{\sum_{i=0}^{N/2} (i - \mu_1)^2 A^2[i]}{\sum_{i=0}^{N/2} A^2[i]} \\ \mu_3 &= \frac{\sum_{i=0}^{N/2} (i - \mu_1)^3 A^2[i]}{\sum_{i=0}^{N/2} A^2[i]} \\ \mu_4 &= \frac{\sum_{i=0}^{N/2} (i - \mu_1)^4 A^2[i]}{\sum_{i=0}^{N/2} A^2[i]}\end{aligned}$$

where N is the DFT length and $A[i]$ is the magnitude of the i^{th} DFT point. To convert to natural frequency (i.e. Hz), the j^{th} moment is multiplied by $(F_s/N)^j$ where F_s is the sampling rate. The skewness coefficient is defined as:

$$S = \frac{\mu_3}{\sqrt{(\mu_2)^3}} \quad (7.4)$$

whereas the Kurtosis coefficient is defined as:

$$K = \frac{\mu_4}{(\mu_2)^2} - 3 \quad (7.5)$$

Based on knowledge of the stop spectral shape, it is expected that μ_1 would be largest for alveolar stops, intermediate for labial stops (about the mid-point of the spectrum) and lowest for velars (in $F2 - F3$ region). The second moment, μ_2 , is the spectral variance around the center of gravity. It is expected that μ_2 would be largest for the labial stops as their energy is distributed across the spectrum. The skewness and Kurtosis coefficients attempt to capture the spectral tilts.

The generic energy-ratio parameters that attempt to capture the stop consonant spectral shape are listed in Table 7.10. In this table, f_1 and f_2 were subject to the

following constraints:

$$f_1 \in [F3 - 1750, F3 + 3050](Hz),$$

$$f_2 \in [F3 - 1550, F3 + 3250](Hz),$$

$$(f_2 - f_1) \geq 200Hz.$$

The third formant, $F3$ was estimated according to the procedure outlined in Section 7.2. The frequency ranges for f_1 and f_2 were chosen large enough to include a velar resonance sometimes observed around the fifth formant, $F5$. The energy measures estimate the energy distribution across the frequency spectrum at a given speech frame. The frequencies f_1 and f_2 were made dependent on $F3$ as exploratory experimentation showed better classification results with such parameters. Each of the moment-based and energy-based parameters was computed at burst release and during the following frication. The burst release was defined as the instant of largest spectral change detected between the TIMIT-label start-time and two thirds of the

Energy-Ratio Parameters
$E[f_1 : f_2]/E[0 : 8000]$
$E[f_1 : f_2]/E[0 : f_1]$
$E[f_1 : f_2]/E[f_2 : 8000]$
$E[0 : f_1]/E[f_2 : 8000]$

tion 7.2. The frequency ranges for f_1 and f_2 were chosen large enough to include a velar resonance sometimes observed around the fifth formant, $F5$. The energy measures estimate the energy distribution across the frequency spectrum at a given speech frame. The frequencies f_1 and f_2 were made dependent on $F3$ as exploratory experimentation showed better classification results with such parameters. Each of the moment-based and energy-based parameters was computed at burst release and during the following frication. The burst release was defined as the instant of largest spectral change detected between the TIMIT-label start-time and two thirds of the

stop duration (as indicated by the TIMIT labels)³ A parameter computation during the frication noise of a stop consonant was taken as the average value of up to three speech frames following the stop release. If no frication-aspiration noise was present, the parameter value at the stop release was substituted for this computation.

7.3.2 Optimized Parameters

In order to derive parameters that distinguish among the stop consonants, the two-stage parameter optimization process described in Chapter 6 was utilized. The speech material used in development consisted of stop samples in the 1268-sentence optimization subset of the TIMIT training set. This set contains 2213 alveolar stops, 1651 labial stops and 1727 velar stops occurring in unconstrained phonetic contexts.

The first stage of the optimization process, the Fisher-criterion stage, was used to determine optimum values for the free parameters, f_1 and f_2 , in Table 7.10. The frequencies f_1 and f_2 were determined at the instant of stop release and during the frication noise separately. As a result, a total of 38 parameters were obtained. The first spectral-moment, the second spectral-moment, the skewness coefficient and the Kurtosis coefficient were also computed for each stop sample at the instant of burst release and during the frication noise as described in Section 7.3.1. Then, the 38 energy-ratio parameters and the spectral moment-based parameters, a total of 46 parameters, were fed to the second stage of the parameter optimization process, growing a classification tree, to select an optimum parameter set. The classification

³The instant of largest spectral change, rather than the timing of the TIMIT labels, was used to locate stop release due to the observed labelers' inconsistency in locating that event.

Table 7.11: The acoustic parameters selected to distinguish among the stop consonants.

Parameters computed at stop release	Parameters computed during frication
$E[F3 - 1750 : F3]/E[0 : 8000]$	$E[F3 + 31 : F3 + 3250]/E[0 : F3 + 31]$
$E[F3 + 281 : F3 + 1187]/E[0 : F3 + 281]$	$E[F3 + 750 : F3 + 1050]/E[F3 + 1050 : 8000]$

tree selected 20 parameters yielding an overall correct classification rate of 79.2% on the training data. 19 of the 20 parameters were energy-ratios and the remaining one was the second spectral moment computed at stop release. In addition, energy-ratio parameters were at the top of the tree. Since the interest was in obtaining a parsimonious parameter set, the top four parameters out of the 20 were selected. These parameters are listed in Table 7.11. The top parameter was $E[F3 + 31 : F3 + 3250]/E[0 : F3 + 31]$ computed during the frication noise. A new classification tree was grown using the parameters in Table 7.11. The resulting tree had a correct classification rate of 76.4%, a 2.8% absolute degradation from the tree utilizing 20 parameters. The distribution of each of the parameters in Table 7.11 are shown in Figure 7.17-Figure 7.20.

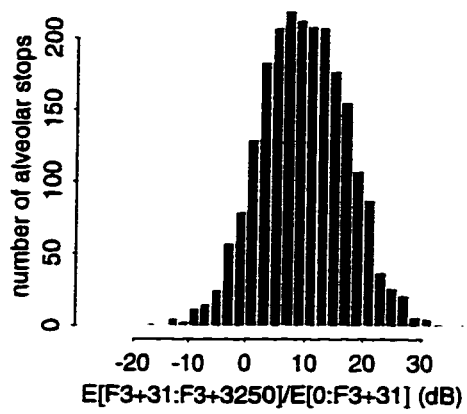
Figure 7.17 shows that $E[F3 + 31 : F3 + 3250]/E[0 : F3 + 31]$ distinguishes the alveolar stops from the labials and velars. Figure 7.18 shows that the velar stop consonants have most of their energy concentrated around $F3$ distinguishing themselves from the alveolars and labials. The two parameters shown in Figure 7.19

and Figure 7.20 contribute to improving the classification of the stop consonants but show larger degree of overlap among the stop consonants, especially the parameter in Figure 7.20, than the first two parameters.

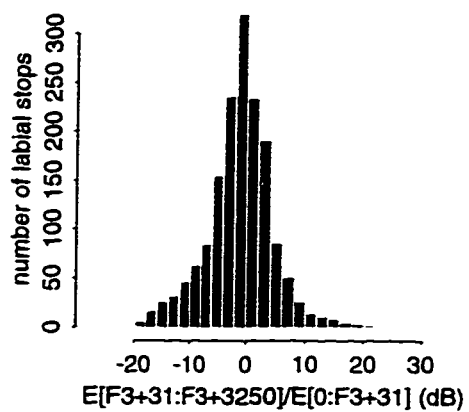
Analysis of classification errors revealed that most of the misclassified stops were voiced. In order to determine the possible cause of these errors, a spectrographic analysis was conducted. This analysis showed that voiced stop consonants manifest high energy concentration at low frequencies when they occur in a sonorant context. This spectral shape was independent of the stop place-of-articulation and could be the result of a weakened closure that allows a relatively strong voicing source to be maintained during the stop as well as the neighboring sonorant sound(s). A Visual inspection of the spectrographic characteristics of these misclassified stops suggested that inclusion of formant transition information between a stop and a following or preceding sonorant sound can ameliorate many of the confusions given that such transition information could be accurately estimated. This observation is consistent with the conclusion of Blumstein and Stevens [68] that formant transitions play a secondary role in stop consonant identification and become important when the stops are hardly identifiable from their gross spectral shapes.

7.3.3 Classification Results

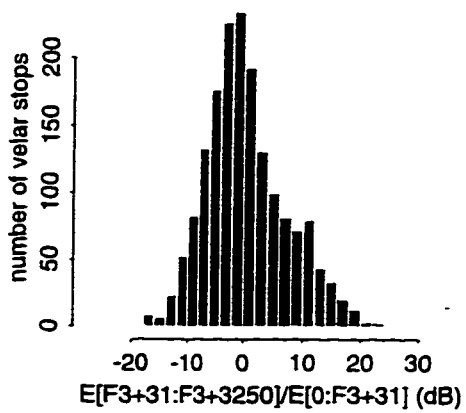
The performance of the developed parameters was evaluated on an independent test set consisting of the 504 TIMIT "si" sentences (c.f. Chapter 4). These sentences contained 1017 alveolar stops, 685 labial stops and 588 velar stops. In this evaluation, the tree classifier built in the parameter development stage was used to classify a stop



(a) alveolar stops



(b) labial stops



(c) velar stops

Figure 7.17: Histograms showing the distribution of $E[F3 + 31 : F3 + 3250]/E[0 : F3 + 31]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.

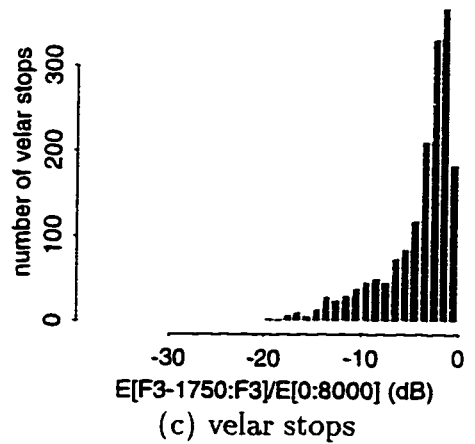
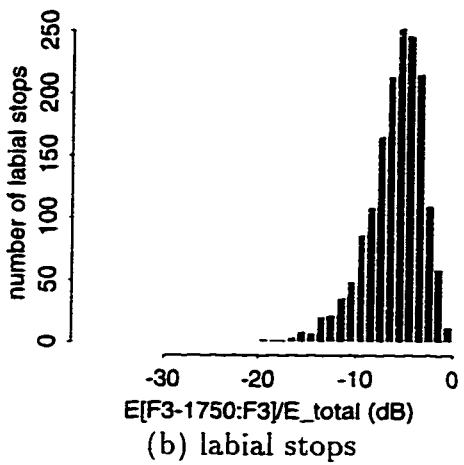
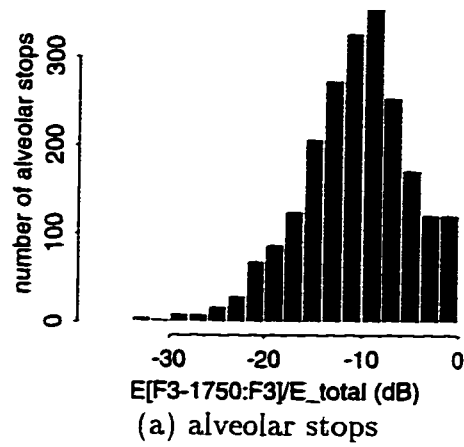
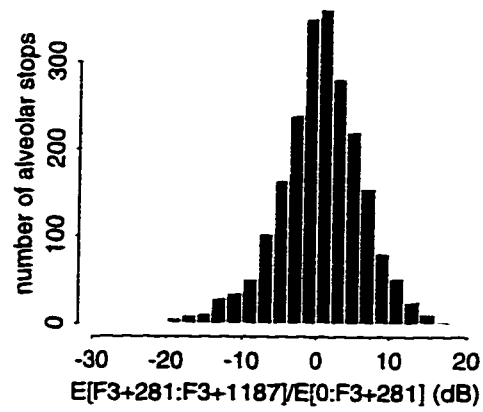
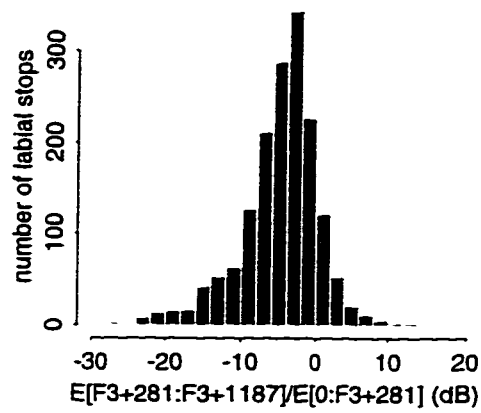


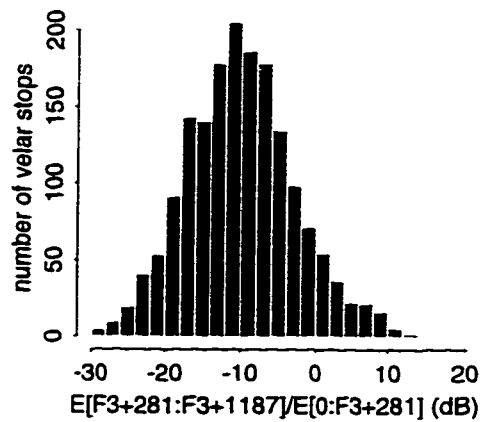
Figure 7.18: Histograms showing the distribution of $E[F3 - 1750 : F3]/E[0 : 8000]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.



(a) alveolar stops

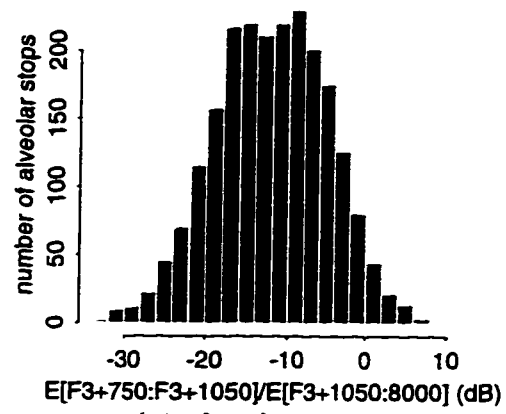


(b) labial stops

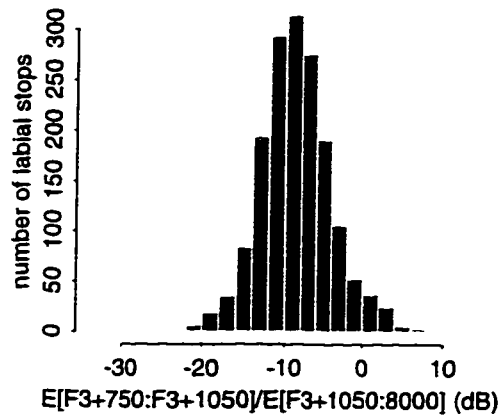


(c) velar stops

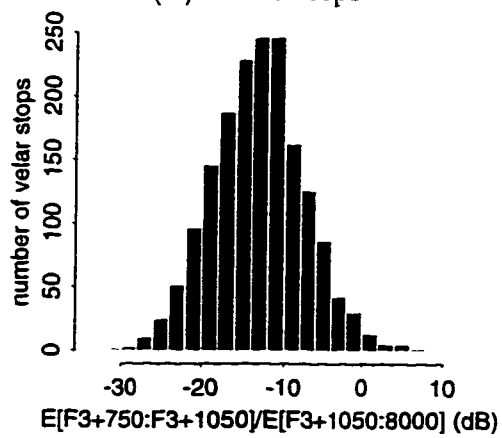
Figure 7.19: Histograms showing the distribution of $E[F3 + 281 : F3 + 1187] / E[0 : F3 + 281]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.



(a) alveolar stops



(b) labial stops



(c) velar stops

Figure 7.20: Histograms showing the distribution of $E[F3 + 750 : F3 + 1050]/E[F3 + 1050 : 8000]$ for: (a) alveolar stops, (b) labial stops and (c) velar stops.

Table 7.12: Classification results on the training and test sets for the stop place of articulation. The classifier was the classification tree obtained in development. The parameters in Table 7.11 were the only ones used in this classification tree.

% correct on training data	% correct on test data
76.4	73.0

consonant occurring in these sentences into one of three categories: labial, alveolar or velar. The results of this classification are summarized in Table 7.12. As these results indicate, there was a 3.4% decrease when the parameters were tested on an independent test set.

7.4 Syllabicity

Phonemes are said to have the “syllabic” phonetic feature when they function as syllable nuclei. The nucleus of a syllable refers to the mandatory part of a syllable that consists of one or more phones. In English, syllable nuclei are mainly vowels. The nasals (/n/, /m/ and /ŋ/) and liquids (/l/ and /r/) are considered to be non-syllabic but can also form syllable nuclei when they occur in certain contexts. All other consonants which include the obstruents and the sonorant glides (/w/, /y/) are nonsyllabic. In this section, parameters that target the “syllabic” phonetic feature in the speech signal are considered.

The “syllabic” phonetic feature is distinctive among the sonorant sounds. Thus, when deriving acoustic parameters that target the “syllabic” phonetic feature, only

phones that are sonorant need to be considered. Accordingly, two groups of phones can be formed. The first group consists of those phones that are sonorant and syllabic while the second consists of phones that are sonorant and nonsyllabic. In the derivation of the "syllabic" acoustic parameters, diphthongs, syllabic nasals and liquids were not considered. This exclusion was motivated by the fact that each of the excluded phones consists of a vowel-sonorant consonant sequence (c.f. [74] p. 20 for discussion on diphthongs and [54] pp. 110-113 and 338-340 for discussion on diphthongs and syllabic consonants). Diphthong vowels, as opposed to monophthong vowels, involve movement of the vocal organs from the position of one vowel to the position of a glide (/y/, /w/). The diphthong vowels [28] are: /a^y/ as in "high", /a^w/ as in "how", /e^y/ as in "bait", /ɔ^y/ as in "boy" and /o^w/ as in "hoe". In addition to these universally recognized diphthongs, the vowel /ʊ/ as in "two" is sometimes considered as a diphthong [75]. The first part of a diphthong that corresponds to a vowel is called the phone kernel and is similar to the corresponding monophthong vowel. The second part of a diphthong is referred to as an off-glide. The degree of similarity between an off-glide and the corresponding independent glide depends on speaker's style and accent. In addition, syllabic nasals (/ŋ/, /ɲ/ and /ŋ/) and syllabic liquids (/l/, /ʒ/ and /ʒ/) are spectrally manifest as one unit in some contexts but often appear to consist of a /ə/ followed by the relevant sonorant consonant. As a result, it would be correct to recognize a diphthong or a syllabic sonorant consonant as a vowel followed by a sonorant consonant. Since the diphthongs and syllabic sonorant consonants may include syllabic and nonsyllabic acoustic events, samples that correspond to these categories were excluded in parameter development. The set of "syllabic" phones considered in development consisted of monophthong vowels while

that of “nonsyllabic” and sonorant phones consisted of the nasals and semivowels.

7.4.1 Algorithm for Detecting a Syllabic/Nonsyllabic Acoustic Event

An algorithm that targets the “syllabic” phonetic feature in the speech signal is outlined in this section. This algorithm is similar to one described in [40] and it is based on the fact that a sonorant consonant always occurs in a sonorant cluster that includes a vowel. The outlined algorithm is motivated by the way vowels and sonorant consonants are produced. Vowels and sonorant consonants are both produced with a pseudo-periodic excitation source at the glottis. However, they differ by the shape of the vocal tract during their production. Vowel sounds are produced with a vocal tract that is not constricted. On the other hand, sonorant consonants are produced with a constriction that is not narrow enough to result in a turbulent noise source (e.g. as in fricatives) but narrow enough to introduce new cavities in the vocal tract that result in antiresonances. In addition, nasal sounds are produced by completely blocking the air flow out of the mouth while letting the air flow out through the lossy nostrils. As a result of these facts, sonorant consonants tend to have less energy in the mid-frequency region compared to vowels. Furthermore, since a sonorant cluster (e.g. arm) that contains sonorant consonants must also include vowels, it seems that any energy measure should be made in the sonorant consonant relative to the first vowel that occurs to the right or left of the sonorant consonant. Such an energy measure will take into account the dynamics of the vocal tract as such sounds are articulated. For instance, in the articulation of a vowel-sonorant consonant sequence,

the vocal tract starts being open during vowel articulation and reaches a point where it is mostly open. Then, the vocal tract becomes gradually constricted as it moves towards the articulation of the sonorant consonant reaching a point of maximum constriction before that constriction is released into the following sound.

In order to capture the stated characteristics of vowels and sonorant consonants, the algorithm is based on the position of a sonorant consonant relative to the vowel [41]. This algorithm attempts to measure the energy difference between a sonorant consonant and two surrounding vowels (intervocalic position), a prevocalic sonorant consonant and the preceding vowel and a postvocalic sonorant consonant and the preceding vowel. If the sonorant-consonant is intervocalic, the algorithm detects the maximum energy in the left vowel (point A in Figure 7.21), the maximum energy in the right vowel (point B in Figure 7.21) and the minimum energy in the sonorant consonant (point C in Figure 7.21). Then, this algorithm measures the minimum energy in the sonorant consonant relative to the smaller of the two detected energy maxima in the surrounding vowels (energy at point C relative to that at point B in Figure 7.21). For sonorant consonants occurring in a prevocalic or postvocalic position, the algorithm detects the maximum energy in the vowel region and the minimum energy in the sonorant consonant region and measures the difference in dB between the two (energy at point A relative to energy at point B in Figure 7.22 for a prevocalic sonorant consonant and Figure 7.23 for a postvocalic sonorant consonant). As can be deduced from the definition of this algorithm, the objective is to detect a nonsyllabic event in a sonorant region. If such a nonsyllabic event cannot be detected, the whole sonorant region is assumed to be syllabic since only a syllabic phone can solely constitute a sonorant region while a sonorant consonant must be always accompanied by

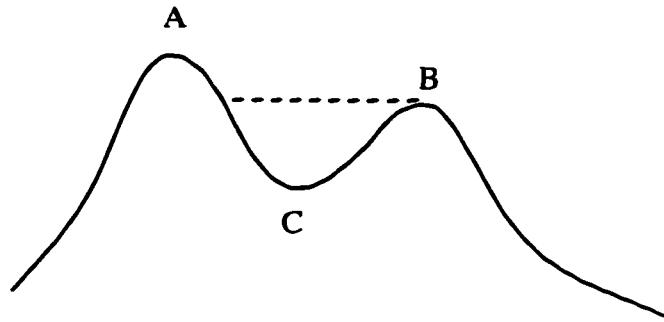


Figure 7.21: Energy profile typical of intervocalic sonorant consonants. The minimum energy value at point C is measured relative to the smaller of the two surrounding maxima at points A and B.

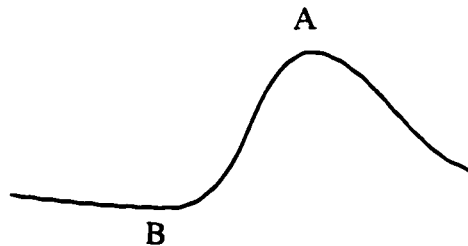


Figure 7.22: Energy profile typical of prevocalic sonorant consonants. The minimum energy value at point B is measured relative to the energy maximum at point A.

a vowel. The question that remains is what frequency band should be selected for computing the energy profile. This question is addressed in the following section.

7.4.2 Optimized Parameters

The algorithm for detecting nonsyllabic events, as defined in Section 7.4.1, requires the computation of an energy profile⁴ in the mid-frequency portion of the spectrum. Thus,

⁴An energy profile refers to energy computed within a frequency band as a function of time.

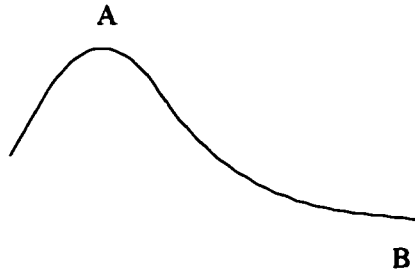


Figure 7.23: Energy profile typical of postvocalic sonorant consonants. The minimum energy value at point B is measured relative to the energy maximum at point A.

the objective was to determine the frequency band(s) within which this energy should be computed. To achieve this objective, samples of monophthong vowels, intervocalic, prevocalic and postvocalic sonorant consonants were extracted from the 1268-sentence development set and used along with the parameter-optimization procedure outlined in Chapter 6. The selected acoustic parameters are listed in Table 7.13 and Table 7.14. In this section, the methodology used in deriving these parameters is discussed.

The generic acoustic parameter for detecting intervocalic nonsyllabic events is depicted in Figure 7.21 and is given by Equation 7.6.

$$E_C[f_1 : f_2] / \text{minimum}(E_A[f_1 : f_2], E_B[f_1 : f_2]) \quad (7.6)$$

where $E[f_1, f_2]$ denotes energy in the frequency band delimited by f_1 and f_2 . The goal is to determine optimum values for the pair (f_1, f_2) in Equation 7.6 so that the false detection of within-vowel nonsyllabic events is minimized while the correct detection of nonsyllabic events is maximized. To do so, The Fisher-criterion stage of the parameter optimization process was used. Since sonorant sounds have most of their energy concentrated below 4 kHz, the following constraints were applied to f_1

and f_2 :

$$\begin{aligned}
 f_1 &\in [100, 3840](Hz) \\
 f_2 &\in [260, 4000](Hz) \\
 f_2 &\geq f_1 + 160(Hz)
 \end{aligned}
 \tag{7.7}$$

All monophthong vowel samples were grouped into one group and the intervocalic sonorant-consonant samples were grouped into another. There were a total of 1358 intervocalic sonorant consonants (vowel, sonorant consonant, vowel sequence) and 11733 vowels in the development set. $E_C[f_1 : f_2] / \text{minimum}(E_A[f_1 : f_2], E_B[f_1 : f_2])$ was computed for each of the vowels and sonorant consonants and for each possible (f_1, f_2) pair. When the parameter in Equation 7.6 was computed within a vowel, $E_C[f_1 : f_2]$ was the minimum value of $E[f_1 : f_2]$ within the vowel whereas $E_A[f_1 : f_2]$ and $E_B[f_1 : f_2]$ were the two surrounding maxima within that vowel. When the parameter in Equation 7.6 was computed for a sonorant consonant, $E_C[f_1 : f_2]$ was the minimum value of $E[f_1 : f_2]$ within the sonorant consonant, whereas $E_A[f_1 : f_2]$ and $E_B[f_1 : f_2]$ were the maximum $E[f_1 : f_2]$ values within the left and right vowel, respectively. The time boundaries of the vowels and sonorant consonants were obtained from the TIMIT label files. The Fisher criterion stage resulted in a total of 11 parameters. These parameters were then fed to the classification-tree stage of the parameter optimization process. As a result, the 4 parameters listed in Table 7.13 were selected. These parameters resulted in a 98.4% correct classification rate on the training data. The top selected parameter, $E[2750 : 3562]$, resulted in a 97% correct classification. The distributions of the two top parameters computed from the energy profiles $E[2750 : 3562]$ and $E[1250 : 2562]$ are shown in Figure 7.24 and

Table 7.13: Energy parameters that are used to detect intervocalic nonsyllabic events.

Energy Parameters
E[2750:3562]
E[1250:2562]
E[1562:2000]
E[1281:3531]

Figure 7.25 for vowels and sonorant consonants. As can be seen in Figure 7.24 and Figure 7.25, $E[2750 : 3562]$ and $E[1250 : 2562]$ show significant energy dips within sonorant consonants relative to the adjacent vowels. However, the energy in the sonorant consonant was sometimes higher than that of an adjacent vowel. These sonorant consonants were often declared as errors in the classification. Inspection of these errors showed that they were mostly semivowels adjacent to the lax vowels /ə/ and /ɪ/.

The generic acoustic parameter for detecting prevocalic and postvocalic nonsyllabic events is depicted in Figure 7.22 and Figure 7.23, respectively. This parameter can be described by:

$$E_A[f_1 : f_2] / E_B[f_1 : f_2] \quad (7.8)$$

In Equation 7.8, $E[f_1 : f_2]$ denotes energy in the frequency-band delimited by the frequencies f_1 and f_2 . The objective, as in the intervocalic case, was to determine values for f_1 and f_2 so that the false detection of nonsyllabic events within vowels is minimized whereas the correct detection of prevocalic and postvocalic nonsyllabic events is maximized. Thus, all 11733 monophthong vowels and all 4098 prevocalic and

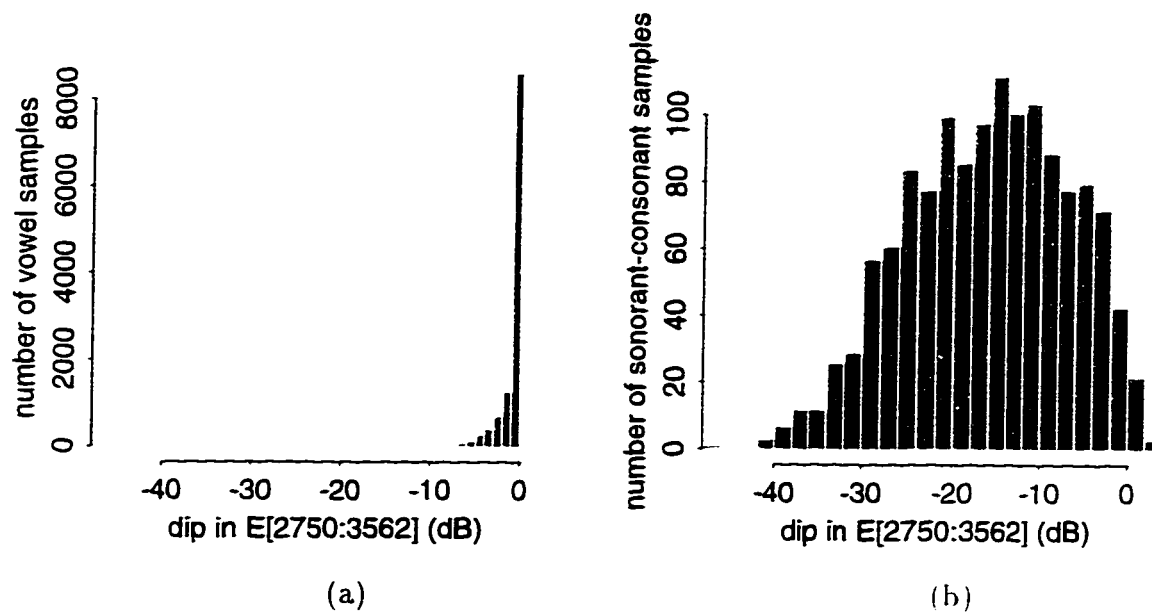


Figure 7.24: (a) Within-vowel energy minimum relative to the smaller of the two surrounding maxima within the same vowel. (b) Sonorant-consonant energy minimum relative to the smaller of the two energy maxima in the left and right-context vowels.

postvocalic sonorant consonants in the 1268-sentence development set were extracted. $E_A[f_1 : f_2]/E_B[f_1 : f_2]$ was computed for each of these samples and for all possible (f_1, f_2) pairs subject to the constraints in equations 7.7. When $E_A[f_1 : f_2]/E_B[f_1 : f_2]$ was computed within a vowel, $E_A[f_1 : f_2]$ was taken as the minimum $E[f_1 : f_2]$ value within the vowel and $E_B[f_1 : f_2]$ was the maximum $E[f_1 : f_2]$ value within that vowel. For prevocalic and postvocalic sonorant consonants, $E_A[f_1 : f_2]$ was taken as the minimum value within the sonorant consonant, whereas $E_B[f_1 : f_2]$ was taken as the maximum value within the vowel. Using the Fisher-criterion stage of the parameter optimization process, (f_1, f_2) pairs were determined so that the parameter described in Equation 7.8 distinguishes best among vowels on one hand and the

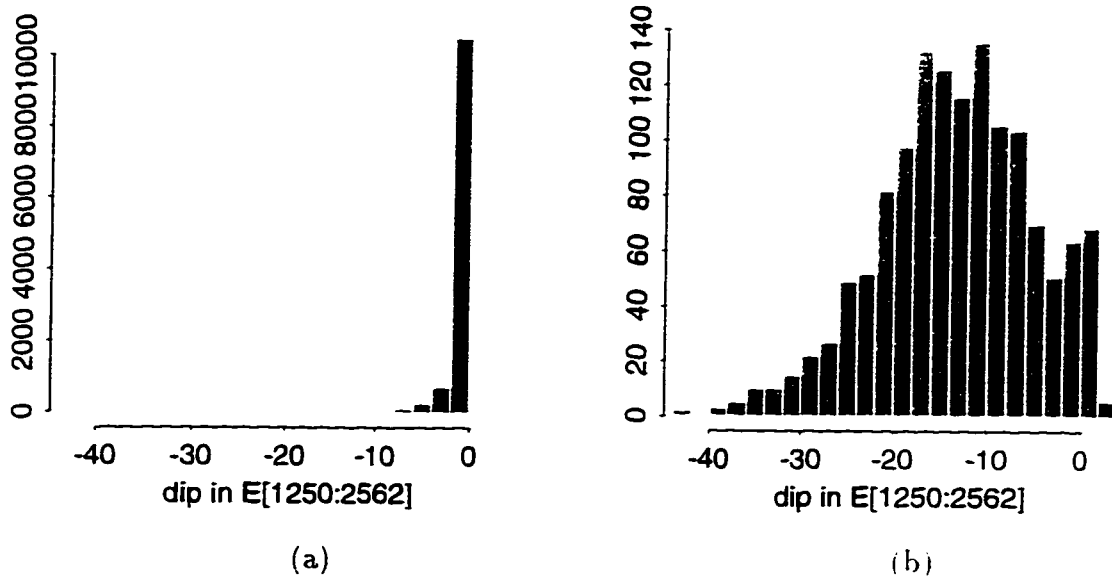


Figure 7.25: (a) Within-vowel energy minimum relative to the smaller of the two surrounding maxima within the same vowel. (b) Sonorant-consonant energy minimum relative to the smaller of the two energy maxima in the left and right-context vowels. group of prevocalic and postvocalic sonorant-consonants on the other hand. The fisher-criterion stage resulted in a total of 8 parameters. These parameters were fed to the classification-tree stage of the parameter optimization process. As a result, the four energy parameters listed in Table 7.14 were selected. The classification tree using the energy dips, as described in Equation 7.8, in these parameters as the predictors had a correct classification rate of 87.4%.

7.4.3 Classification Results

The performance of the developed parameters was evaluated on an independent test set consisting of the 504 TIMIT “si” sentences (c.f. Chapter 4). In this evaluation, the

tree classifiers built in the parameter development stage were used. The classification results are summarized in Table 7.15. These results show an insignificant change between the results obtained on the training set and those obtained on the test set.

Table 7.14: Energy parameters that are used to detect prevocalic and postvocalic nonsyllabic events.

Energy Parameters
E[500:4000]
E[937:3437]
E[2750:4000]
E[2000:4000]

Table 7.15: Classification results on the training and test sets for syllabicity/nonsyllabicity.

Category	% correct on training data	% correct on test data
Intervocalic	98.4	97.8
Nonsyllabic		
Post-/Pre-vocalic	87.4	87.4
Nonsyllabic		

7.5 Stridency

The stridency phonetic feature is characteristic of obstruent sounds with strong turbulent noise. It plays an important role in the correct perception of these sounds. The strident sounds in American English are the fricatives /s/. /z/. /š/. /ž/ and the affricates /č/ and /ǰ/. This feature distinguishes the strong fricatives (i.e., stridents) from the weak fricatives (/f/, /v/, /θ/ and /ð/. It also distinguishes the noncontinuant stridents (affricates) from the other noncontinuant sounds (stop consonants). The distinctive role of the stridency feature was shown through synthesis studies [64] and perceptual studies (e.g. [30], [76]). In addition, an acoustic study that we conducted [77], involving expert spectrogram readers, also showed that this feature plays a distinctive role among the obstruents.

7.5.1 Acoustic Parameters for Stridency: strident obstruents vs. weak fricatives

Strident sounds are characterized by strong energy in the region of F_3 and above (c.f. [64]). Thus, several generic parameters were formulated to capture this property. These parameters attempt to measure the obstruent energy in a frequency band about F_3 and above relative to the maximum, average and minimum energy across the utterance. Acoustic parameters (APs) that target the stridency feature differentiating between the strident obstruents and the weak fricatives were derived separately from APs that differentiate only the affricates from the stop consonants. To determine the APs in the former case, the strident obstruent samples in the 1268-sentence

Table 7.16: Generic acoustic parameters to distinguish between the strident obstruents and the weak fricatives.

Generic Acoustic Parameters
$E[f_2 : 8000]/average(E[f_2 : 8000])$
$E[f_2 : 8000]/maximum(E[f_2 : 8000])$
$E[f_2 : 8000]/minimum(E[f_2 : 8000])$
$E[f_1 : f_2]/average(E[f_1 : f_2])$
$E[f_1 : f_2]/maximum(E[f_1 : f_2])$
$E[f_1 : f_2]/minimum(E[f_1 : f_2])$

development set were grouped in one set and the weak fricatives were grouped in another. The Fisher-criterion stage of the parameter optimization process was used to determine APs from the generic parameters listed in Table 7.16 with the constraints:

$$f_1 \in [F3 - 1000, F3 + 3000]$$

$$f_2 \in [F3 - 700, F3 + 3300]$$

$$f_2 \geq (f_1 + 300)$$

where $F3$ was estimated using the procedure outlined in Section 7.2. In Table 7.16, $E[f_1 : f_2]$ refers to the energy in the frequency band delimited by f_1 and f_2 . From the Fisher-criterion stage, 8 parameters were obtained. These parameters were then fed to the classification-tree stage of the parameter optimization process. As a result, the 5 parameters listed in Table 7.17 were selected. When a parameter is computed within

Table 7.17: Selected acoustic parameters to distinguish between the strident obstruents and the weak fricatives.

Selected Acoustic Parameters
$E[F3 + 94 : 8000]/average(E[F3 + 94 : 8000])$
$E[F3 + 31 : 8000]/minimum(E[F3 + 31 : 8000])$
$E[F3 - 687 : 8000]/maximum(E[F3 - 687 : 8000])$
$E[F3 - 125 : 8000]/maximum(E[F3 - 125 : 8000])$
$E[F3 + 500 : F3 + 3000]/maximum(E[F3 + 500 : F3 + 3000])$

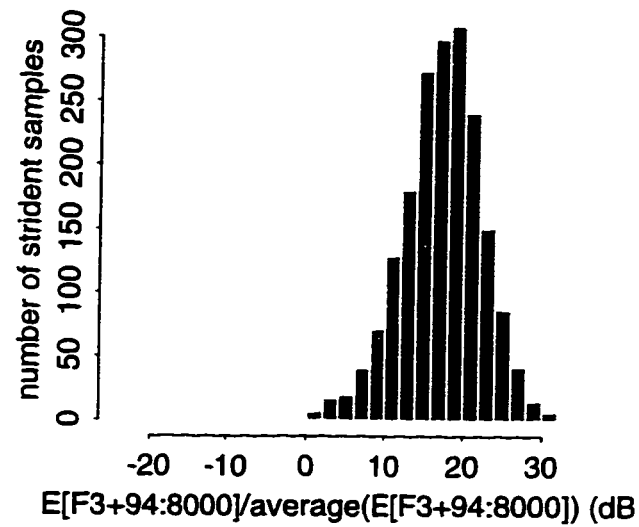
the weak fricative or the strident obstruent, its average value across that obstruent is taken. Classification results using these parameters and the developed tree are summarized in Table 7.18. The classification task was to classify an obstruent, from the set of strident obstruents and weak fricatives, as strident or nonstrident. The test data consisted of all such samples in the 504 "si" TIMIT test sentences. The top parameter alone, $E[F3 + 94 : 8000]/average(E[F3 + 94 : 8000])$, resulted in a 93.4% classification on the training data and 93.3% classification on the test data. The distributions of this parameter for the weak fricatives and strident obstruents are shown in Figure 7.26. In addition, the distributions of this parameter for the weak fricatives and the strident obstruents are shown in Figure 7.27 as a function of gender. Figure 7.27 shows that males and females have very similar distributions for the top parameter.

Table 7.18: Classification results for strident obstruents vs. weak fricatives using the classification tree built in the development stage and the parameters in Table 7.17.

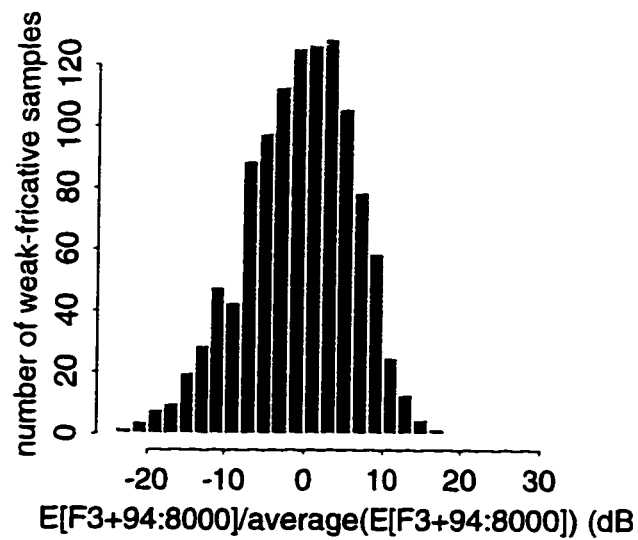
% correct on training data	% correct on test data
94.8	95.0

7.5.2 Acoustic Parameters for Stridency: Affricates vs. Stops

APs that distinguish among the strident noncontinuants (affricates) and nonstrident noncontinuants (stops) were also derived. In this derivation, all affricates in the 1268-sentence development set were considered in one group and all stop consonants were considered in another. Using these samples, the Fisher criterion-stage of the parameter optimization process was deployed to determine APs from the generic parameters in Table 7.16. This Fisher criterion stage resulted in 15 parameters. These 15 parameters along with the duration of each stop and affricate, as an additional parameter, were fed to the classification-tree stage of the parameter optimization process. As a result, the acoustic parameters in Table 7.19 were selected. The classification results for both the training data and the test data, based on these parameters, are summarized in Table 7.20. The classification task was to classify a sample from the set of affricates and stops as strident or nonstrident. The top parameter was $E[F3 + 562 : F3 + 1125]/\text{minimum}(E[F3 + 562 : F3 + 1125])$ followed by duration.

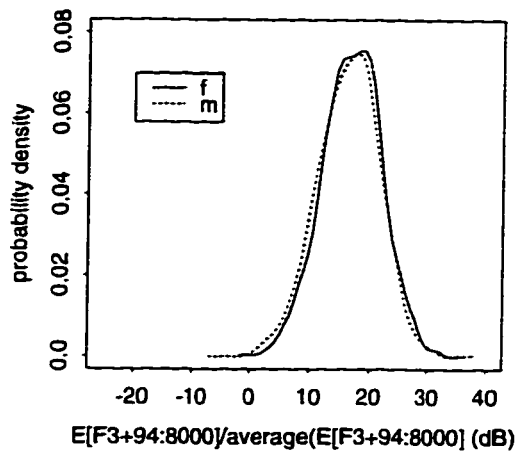


(a) strident obstruents

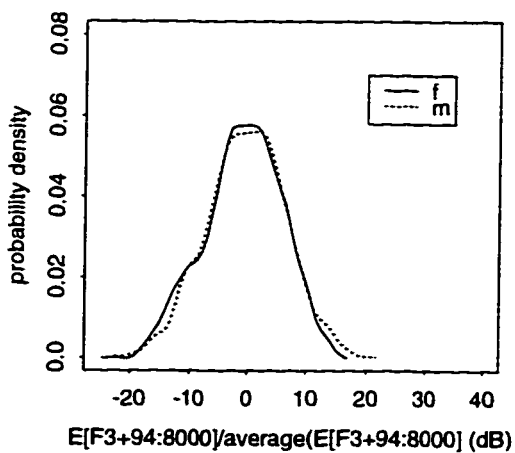


(b) weak fricatives

Figure 7.26: Histograms showing $E[F3 + 94 : 8000]/average(E[F3 + 94 : 8000])$ distributions for: (a) strident obstruents and (b) weak fricatives.



(a) strident obstruents



(b) weak fricatives

Figure 7.27: Probability densities of $E[F3 + 94 : 8000]/average(E[F3 + 94 : 8000])$ for: (a) strident obstruents and (b) weak fricatives as a function of gender (females (f) and males (m))

Table 7.19: Selected acoustic parameters to distinguish between the affricates and the stop consonants.

Selected Acoustic Parameters
$E[F3 + 562 : F3 + 1125] / \text{minimum}(E[F3 + 562 : F3 + 1125])$
duration
$E[F3 - 1000 : F3 - 700] / \text{average}(E[F3 - 1000 : F3 - 700])$
$E[F3 - 250 : 8000] / \text{maximum}(E[F3 - 250 : 8000])$
$E[F3 + 656 : F3 + 1031] / \text{average}(E[F3 + 656 : F3 + 1031])$
$E[F3 + 2000 : 8000] / \text{minimum}(E[F3 + 2000 : 8000])$
$E[F3 + 125 : F3 + 1562] / \text{maximum}([F3 + 125 : F3 + 1562])$

Table 7.20: Classification results for affricates vs. stop consonants using the classification tree built in the development stage and the parameters in Table 7.19.

% correct on training data	% correct on test data
95.5	94.7

Chapter 8

Speech Manner and Obstruent

Place-of-Articulation Recognition

Several experiments are reported in this chapter that examine the performance of a subset of the parameters developed in chapter 7 in a speech manner and obstruent place-of-articulation recognition task. In all reported experiments, the HMM framework, as implemented in HTKv1.5 [53], was used for recognition. These experiments show that the developed parameters are competitive in terms of overall performance with the traditionally used Mel-cepstral parameters. Furthermore, the experiments show that more information pertinent to the stop-consonant place of articulation may be needed in order to further improve the recognition of the three stop-consonant classes: alveolar, labial and velar. This can possibly be achieved by considering dynamic parameters such as characteristics of formant trajectories or moment trajectories. This postulation is based on more recent results obtained in [78]

as well as previous acoustic studies [68] that show that formant movements play a secondary role in keying the place of articulation of stop consonants.

Speech manner and obstruent place-of-articulation recognition, in this thesis, is the task of recognizing a speech utterance as a sequence of the following classes: Silence (SL), Syllabic (SY), Sonorant Consonant (SC), Affricate (A), Labial Stop (LS), Alveolar Stop (AS), Velar Stop (VS), Alveolar Fricative (AF), Palatal Fricative¹ (PF) and Weak Fricative (WF). These are the classes that can be recognized in a given speech utterance based on the APs discussed in chapter 7. The mapping between these classes and the phone labels used in TIMIT is shown in Table 8.1. In chapter 7, classification experiments using these parameters were run on a subset of the TIMIT test set consisting of all phonetically diverse “si” sentences. The same test set is used in the recognition experiments reported in this chapter. As referred to in the speech research literature, classification is the task of identifying the category of a sound with known time boundaries (the time boundaries, in this case, are known from the TIMIT transcription). On the other hand, recognition refers to the task of simultaneous detection and classification of a sound. Thus, in recognition, it is possible to falsely detect the presence of a sound, a phenomenon known as sound insertion, or miss the detection of a sound which is referred to as deletion.

¹Palatal fricatives are the ones produced with a constriction in the palate region. In the context of the parameter derivation in Chapter 6, palatal fricatives are the strident and nonanterior fricatives /ʒ/ and /ʃ/.

8.1 Experimental Objectives

The objectives of the experiments are multifold. First, we want to experiment with the developed parameters in a recognition task in order to discover the weaknesses in the signal representation. Second, we want to compare the developed parameters to the traditionally-used Mel cepstral parameters to see if the two representations are redundant, complementary or one works better than the other. This comparison has led us to use the frame-based HMM framework so that recognition results can be compared and interpreted relative to the front end signal representation. Finally, since the acoustic parameters were developed using a linguistically motivated hierarchical structure, as did the classification, it is worth noting the effect of considering all parameters at once in recognition as is done in the HMM-based framework. That is, in the recognition procedure, every parameter was computed everywhere, whereas in the development and classification, a parameter was looked at given some a priori knowledge of the broader category to which a sound belongs (e.g. only strident obstruents were considered in the development of the anterior/nonanterior parameters).

8.2 Signal Representation

The Mel-cepstrum signal representation of the speech signal consists of the first twelve cepstral coefficients, excluding the zeroth, and the normalized log energy. The Mel-cepstral coefficients are computed every 5 ms within a 10 ms window after passing the speech signal through a high pass FIR filter with a real zero at 0.95. This high pass filter, a preemphasis filter, removes the effect of the radiation loss at the lips.

Table 8.1: This table shows the mapping between the TIMIT labels (represented by IPA symbols) and the speech classes used in this chapter.

TIMIT Labels	Manner-Place Classes
/ɛ/. /ɔ/. /ɑ/. /ʊ/, /ɜ/. /a ^y /. /e ^y /. /a ^w /. /ə/. /ɪ/. /ɪ/. /æ/, /ʌ/, /u/, /ɔ ^y /. /i ^y /. /o ^w /. /ɚ/. /ə/, /l/, /r/, /r/, /r/	syllabic (SY)
/n/. /m/. /ŋ/, /nɪ/. /r/, /hv/ /w/. /l/. /y/. /r/	sonorant consonant (SC)
/tʃ/. /dʒ/	affricate (A)
/b/. /p/	labial stop (LS)
/d/. /t/	alveolar stop (AS)
/g/. /k/	velar stop (VS)
/s/. /z/	alveolar fricative (AF)
/ʃ/. /ʒ/	palatal fricative (PF)
/f/. /v/. /θ/. /ð/, /h/	weak fricative (WF)
h#. kcl. pcl. tcl. dcl, gcl, bcl, pau, epi	silence (SL)

The APs developed in this thesis are computed every 5 ms within a 10 ms window. The energy based parameters are computed after passing the speech signal through a preemphasis filter with a real zero at 0.95. The non-energy based parameters (zero-crossing rate, cross-correlation and voicing probability) are computed from the speech signal prior to preemphasis. The APs used are listed in Table 8.2. These parameters constitute a subset of the ones derived in chapter 7. The choice was made by selecting the most prominent parameters important to the classification of the corresponding phonetic features with the objective of having the number of cepstral-based parameters and their first and second derivatives (39 parameters) close to the number of APs and their first derivatives (40 parameters). In addition, the chosen parameters contribute more than 90% to the correct identification of the corresponding phonetic features.

8.3 Acoustic Models

Acoustic models of all speech categories were context-independent² three-state left-to-right Hidden Markov Models (HMM's). The observation probabilities given an HMM state were continuously distributed Gaussian mixtures with the number of mixtures being one or eight. In all experiments, it was assumed that the components of the observation vector, cepstra or APs, were independent so that the Gaussian

²A context-independent model of a speech category treats the occurrence of that category in any phonetic context the same. In contrast, context-dependent modeling of a speech category results in several models of that category. Each of these models corresponds to a particular phonetic context in which the speech category occurs.

distributions have diagonal covariance matrices. The assumption of independence is unrealistic but was chosen to simplify the computations. Furthermore, the objective was not to develop the best recognition system for the task at hand, but to compare the cepstral-based signal representation to that of the APs in a controlled framework.

8.4 Acoustic Model Training

Training the acoustic models consisted of the three stages implemented in HTKv1.5. All models were bootstrapped from a set of sentences collected from the phonetically diverse “si” and phonetically compact “sx” sentences in the TIMIT training set. Bootstrapping consists of model initialization using Viterbi style training followed by Baum-Welch (or forward backward) training in the second stage. In the third stage, the TIMIT training set was used to reestimate the HMMs built from the first two stages using embedded reestimation. In embedded reestimation, it is assumed that the class label sequence is known but the time segmentation is not. That is, for every label sequence associated with a TIMIT sentence, a large HMM is built by concatenating the HMM’s corresponding to the labels. Then, the HMM’s are updated by mapping the acoustic observations to the HMM states in a manner that maximizes the probability of the large HMM model given the observations. Embedded reestimation is intended to reduce the effect of segmentation errors incurred during the transcription process by human listeners.

8.5 Recognition Experiments

Two sets of experiments were conducted to examine the performance of the derived APs in comparison to cepstra. In the first set of experiments (baseline experiments), speech models were trained on both males and females and tested on an independent test set of males and females. In the second set of experiments, HMMs were tested on speech from one gender and tested on sentences from the other gender. The purpose of the second set of experiments is to compare the APs to cepstra in terms of robustness to gender differences.

8.5.1 Baseline Experiments

In the baseline experiments, the training set consisted of the “si” and “sx” sentences in the TIMIT training set. The test set consisted of all “si” sentences in the TIMIT test set. The insertion/deletion penalty used in recognition was chosen so that the number of insertions and deletions were close to each other as determined by running recognition experiments on a set of 168 “sx” sentences extracted from the TIMIT test set. The recognition results are summarized in Table 8.3. In this table, *MFCC_E* stands for Mel-Frequency Cepstral Coefficients and normalized log Energy. *MFCC_E_δ1_δ2* refers to *MFCC_E* coefficients augmented by their first and second derivatives. *AP* refers to the acoustic parameters derived in this thesis and *AP_δ1* refers to the APs augmented by their first derivatives. In the 1-mixture case, it was assumed that the observation vector given the HMM state is Gaussian distributed, whereas in the 8-mixture case, it was assumed that the probability distribution of the observation vector given the state consists of the combination of 8 Gaussian distributions. The

choice of one mixture and eight mixtures was based on previous experimentation with the manner-class recognition task discussed in chapter 5.

8.5.2 Performance of APs

First, by comparing the recognition results of the AP to those of $AP_{\delta 1}$ in Table 8.3, it is clear that adding the first derivatives increases recognition performance. The improvement in performance is seen in almost every category but is especially prominent in the case of the stop consonants where the correct recognition of the alveolar stops went from 40.3% (c.f. Table 8.6) to 65.2% (c.f. Table 8.7) in the 8 mixture case, an absolute increase by about 25%. Furthermore, increasing the number of Gaussian mixtures resulted in improved recognition results of about 8% for $AP_{\delta 1}$. Comparing the 1 mixture results in Table 8.5 to the 8 mixture results in Table 8.7, it can be seen that the improvement in recognizing the alveolar and velar stop consonants as well as that of the palatal fricatives and weak fricatives is the most prominent. Improved recognition with the higher mixture model can be due to many factors. First, some of the parameters are not Gaussian distributed so the multi-mixture Gaussian distribution provides a better fit. Second, multi-mixture Gaussian distributions may provide better models by capturing the effects of speech variability that arises from (1) interspeaker differences and (2) contextual variability. While we contend that the first type of variability is reduced by appropriate design of the APs, the second type of variability may be captured by the different Gaussian modes of a sound model. An indication of this fact can be deduced from the gender experiments reported in the next section.

Table 8.2: Phonetic features, acoustic correlates, and APs used in the HMM recognition system. A dip_to_peak energy parameter is computed by first locating dips and peaks and then computing, in each frame between the peak and the adjacent dip, the difference in energy between the energy at the peak location and the energy in each frame. A peak_to_dip parameter is computed similarly, but relative to the energy at the dip location instead of the energy at the peak location.

Phonetic feature	Acoustic correlates	APs
sonorant	strong low-frequency energy	voicing-probability (Entropic software) E[0:688]/E[4000:8000] E[0:375]/Eavg[0:375]
nonsyllabic	weak mid-frequency energy	dip_to_peak: E[500:4000], E[2750:3562] E[1250:2562]
syllabic	strong mid-frequency energy	peak_to_dip E[500:4000]
fricated	turbulent noise mainly at high frequencies	zero-crossing R1 = first autocorrelation coefficient dip_to_peak R1
noncontinuant	silence followed by an abrupt onset	<u>silence</u> : E[200:3000]/E _{max} [200:3000] E[3000:6000]/E _{max} [3000:6000] <u>abrupt onset</u> : sum of first difference values across STFT channels
strident	strong turbulent noise around F ₃ and above	E[F3+94:8000]/Eavg[F3+94:8000]
palatal	spectral peak around F ₃	E[F3-187:F3+594]/E[0:8000]
noncontinuant strident	strong turbulent noise in F ₃ region	E[F3+562:F3+1125]/ E _{min} [F3+562:F3+1125]
Stop place	<u>labial</u> : spectrum is fairly flat or falling with increasing frequency <u>alveolar</u> : spectral prominence at high frequencies (above F ₃) <u>velar</u> : spectral prominence in F ₂ -F ₃ region	E[F3+31:F3+3250]/E[0:F3+31] E[F3-1750:F3]/E[0:8000] E[F3+281:F3+1187]/E[0:F3+281] E[F3+750:F3+1050]/E[F3+1050:8000]

Table 8.3: Recognition results. *MFCC_E* refers to 12 Mel-cepstral coefficients normalized & log energy. *MFCC_E δ 1 δ 2* refers to *MFCC_E* & their 1st and 2nd derivatives. *AP* refers to 20 acoustic parameters. *AP δ 1* refers to *AP* and their 1st derivative. Each entry contains % correct/% accuracy.

Signal Representation	1 mixture	8 mixtures
<i>MFCC_E</i>	69.4/52.3	72.6/61.1
<i>MFCC_Eδ1δ2</i>	70.0/54.7	82.0/70.4
<i>AP</i>	71.7/56.1	74.6/61.5
<i>APδ1</i>	75.4/63.4	80.1/69.4

It should be noted from the confusion matrices that the highest confusions arise among the stop consonants, between the weak fricatives and the sonorant consonants and between the palatal fricatives and the affricates. As previously indicated, the confusion among the stop consonants may be reduced by including in the signal representation parameters that relate to formant trajectories. The relatively high confusion between the weak fricatives and the sonorant consonants is due to the fact that many of the voiced weak fricatives that either appear adjacent to a sonorant sound or occur between two sonorant sounds are manifest as more sonorant than fricated. The confusion between the palatal fricatives and the affricates arise because these two sound classes share the same place of articulation which is the palatal region in the mouth. Another source of high confusion is between the weak fricatives and the labial stops. This confusion could be due to the fact that / θ / and / δ / are sometimes realized as noncontinuant (stop-like) and when they do, they appear to have a more

Table 8.4: The confusion matrix when the signal representation consisted of AP and the observation distribution given an HMM state was Gaussian. All numbers are in percentage.

Recognized	SL	SY	SC	A	LS	AS	VS	AF	PF	WF	DEL
TIMIT-Label											
SL	82.2	0.4	3.2	1.0	2.0	0.7	0.6	0.6	0.6	1.8	6.9
SY	1.5	74.5	6.0	0.5	2.3	1.9	0.8	0.3	0.5	0.5	11.2
SC	0.9	1.9	71.4	0.4	4.0	1.9	0.9	0.2	0.3	1.1	17.0
A	0.0	1.9	1.3	69.0	5.7	3.8	2.5	3.8	5.7	0.0	6.3
LS	1.2	0.9	2.2	0.1	70.6	4.1	7.6	0.1	0.3	0.9	12.0
AS	0.79	2.2	1.6	5.8	13.8	50.9	7.1	0.7	0.5	0.9	15.7
VS	2.6	2.2	3.4	3.7	16.8	13.1	43.5	0.2	2.2	1.2	11.0
AF	0.5	0.3	1.3	2.8	0.6	1.8	0.2	86.7	1.4	0.8	3.6
PF	0.4	0.9	0.0	8.1	2.3	7.7	0.9	3.6	71.9	0.4	3.6
WF	3.5	0.8	10.1	0.4	17.4	8.9	4.9	1.3	0.3	36.7	15.7

Table 8.5: The confusion matrix when the signal representation consisted of $AP_{\delta}l$ and the observation distribution given an HMM state was Gaussian. All number are in percentage.

Recognized	SL	SY	SC	A	LS	AS	VS	AF	PF	WF	DEL
TIMIT-Label											
SL	83.4	0.3	3.2	0.6	0.8	0.2	0.4	0.7	0.3	2.7	7.2
SY	0.9	80.8	5.9	0.2	0.6	0.4	1.0	0.2	0.4	0.4	9.1
SC	0.9	0.9	70.5	0.2	2.9	0.5	1.4	0.1	0.4	1.9	20.2
A	0.0	0.0	1.3	77.2	1.3	3.8	0.6	7.0	1.9	0.6	6.3
LS	0.0	0.1	2.3	0.6	78.4	3.6	8.9	0.1	0.0	0.3	5.5
AS	1.1	1.6	2.4	6.4	14.0	50.0	6.3	2.2	0.3	1.7	14.2
VS	1.2	1.9	1.0	2.4	15.5	13.4	47.3	0.3	0.5	1.2	15.3
AF	0.2	0.1	0.8	0.5	0.1	0.3	0.1	93.5	1.0	1.5	2.0
PF	0.9	0.9	2.7	9.0	0.4	0.9	0.9	11.8	67.0	0.9	4.5
WF	2.6	0.6	10.1	0.4	15.0	3.4	4.8	1.7	1.5	47.1	12.8

Table 8.6: The confusion matrix when the signal representation consisted of AP and the observation distribution given an HMM state was a mixture of 8 Gaussians. All numbers are in percentage.

Recognized	SL	SY	SC	A	LS	AS	VS	AF	PF	WF	DEL
TIMIT-Label											
SL	84.5	0.6	2.5	0.8	0.8	0.3	0.5	0.8	0.3	3.2	5.8
SY	1.1	79.4	4.8	0.4	1.6	0.4	0.5	0.4	0.5	0.9	9.9
SC	1.3	3.0	69.6	0.4	2.4	0.8	0.8	0.3	0.3	2.8	18.3
A	1.3	0.6	1.9	69.6	0.6	3.2	1.3	3.2	10.1	1.3	7.0
LS	1.3	0.9	2.5	0.9	77.1	3.2	3.6	0.4	0.0	3.1	7.0
AS	1.3	2.0	3.5	9.2	9.9	40.3	3.6	3.7	1.3	3.3	21.7
VS	1.7	1.2	2.2	5.6	17.8	12.2	46.9	0.5	0.2	2.0	9.5
AF	0.5	0.2	1.5	2.2	0.9	1.6	0.3	86.4	1.4	1.0	3.9
PF	0.4	0.4	4.5	10.9	0	1.3	0.9	4.5	73.3	0.4	3.2
WF	2.9	1.7	9.3	0.8	8.5	3.5	2.3	1.1	0.5	56.4	12.9

Table 8.7: The confusion matrix when the signal representation consisted of $AP_{\delta 1}$ and the observation distribution given an HMM state was a mixture of 8 Gaussians.

All numbers are in percentage.

Recognized	SL	SY	SC	A	LS	AS	VS	AF	PF	WF	DEL
TIMIT-Label											
SL	85.7	0.5	2.3	0.5	0.3	0.4	0.1	0.4	0.2	2.7	6.9
SY	0.6	86.8	3.5	0.0	0.3	0.5	0.2	0.2	0.3	0.5	7.1
SC	0.8	1.0	72.5	0.2	1.7	0.9	0.3	0.2	0.1	3.5	18.7
A	0.0	0.6	1.3	81.0	1.3	5.7	0.0	2.5	1.9	0.6	5.1
LS	0.1	0.9	1.0	1.0	79.0	5.4	4.1	0.1	0.0	1.9	6.4
AS	0.6	1.1	1.5	4.4	8.2	65.2	4.5	1.0	0.2	1.3	12.0
VS	1.2	1.0	1.2	3.1	8.8	12.4	57.6	0.5	0.0	1.5	12.6
AF	0.3	0.1	0.7	1.3	0.1	0.6	0.1	92.2	2.2	1.0	1.4
PF	0.9	0.0	1.8	5.0	0.0	0.0	0.0	6.8	82.3	1.3	1.8
WF	2.3	1.0	7.0	0.1	9.3	5.5	3.9	1.6	0.7	57.9	10.7

or less flat spectrum which is similar to that of labial stops. This problem may be reduced when the dental feature of these weak fricatives is considered.

8.5.3 APs vs. Cepstra: A Performance Comparison

The objective in this section is to compare a cepstral-based signal representation to that of the APs in the speech manner and obstruent place-of-articulation recognition task. Thus, a cepstral based signal representation as described in section 8.2 was generated and 3-state HMMs of all 10 speech classes were trained on the TIMIT “si” and “sx” training sentences and tested on all 504 “si” TIMIT test sentences. The results are summarized in Table 8.3. First, it should be noted that in the 1-mixture case, the APs performed better than their cepstral counterparts. For the 1-mixture case, the recognition accuracy with *AP* was 3.8% higher than with *MFCC_E* and the recognition accuracy with *AP_δ1* was 8.7% higher than with *MFCC_E_δ1_δ2*. It could be the case that the Gaussian distribution provides a worse model for the cepstral observations than for the APs. This can be deduced by comparing the recognition accuracy obtained with 1-mixture models to that obtained with 8-mixture models. In the case of *MFCC_E*, an absolute increase in accuracy of 8.8% was observed (c.f. Table 8.3), while in the case of *MFCC_E_δ1_δ2* an absolute increase of 15.7% was observed in going from unimodal Gaussian to 8-mixture Gaussian models. Moreover, the recognition accuracy with *MFCC_E_δ1_δ2* was 1% higher than that with *AP_δ1* in the 8-mixture case. This is the only case where the cepstral-based representation performed better than the APs.

Better recognition accuracy was achieved with the 8-mixture Gaussian models, in

comparison to the 1-mixture Gaussian models, for both the cepstral signal representation and the AP signal representation. However, the improvement in performance was much more pronounced in the case of the cepstra than in the case of the APs. Multi-mixture Gaussian distributions are used as an observation distribution in order to better model the variability in the acoustic realization of speech sounds. This variability may arise from interspeaker variability (such as gender difference) and contextual variability. Thus, by comparing the 1-mixture results to the 8-mixture results, one can argue that the APs for a specific sound are inherently less variable than the cepstral values for that sound. However, is this reduction in variability due to reduction in the contextual effect on the APs or is it due to reduction in the speaker-dependent effects? In the next section, it is shown through experimental results that the APs are less effected by gender differences than the cepstra and they tend to reduce the effect of the gender differences more than the contextual variability.

In order to hone in on the weaknesses of the APs relative to cepstra, the between-class confusions obtained from the 8-mixture recognition tests (best results) are compared. First, the weak fricative correct recognition with *MFCC_E δ 1 δ 2* was 11.4% higher than with *AP δ 1*. Second, the correct recognition of the three stop consonant classes was higher, for each class, with *MFCC_E δ 1 δ 2* than with *AP δ 1*. In the case of the labial stops, the difference was 3.8%. For the alveolar stops, the difference was 12.1% while for the velar stops the difference was 24.2%. These results suggest that the cepstra are capturing more important acoustic details, relevant to the identification of the stop consonants, than the APs. This is not surprising since the present set of APs provides only a partial representation of the speech signal, whereas the cepstral parameters provide a full representation of the speech signal.

8.5.4 Gender Experiments: APs vs. Cepstra

The set of experiments reported in this section shows that the APs are better able to reduce gender-dependent effects relative to the Mel-cepstral parameters. In this set of experiments, all the TIMIT training sentences from the two dialect regions: dr1, dr2 were used. These sentences were divided into two sets: sentences spoken by females and sentences spoken by males. In each experiment, the acoustic models were trained on one gender and tested on another. Furthermore, $AP_{\delta 1}$ and $MFCC_{E_{\delta 1 \delta 2}}$ were used as the AP-based signal representation and the Mel-cepstral-based signal representation, respectively, since these representations produced the best results. The observation probability given an HMM state was a mixture of 8 Gaussians as this is the distribution, among those tested, that produced the best results in the baseline experiments.

The experimental results are summarized in Table 8.8. It is shown that when training on male speech and testing on female speech, the recognition accuracy with $AP_{\delta 1}$ is about 1% higher than with the Mel-cepstral representation. However, when training on females and testing on males, the recognition accuracy with $AP_{\delta 1}$ is about 3% higher than with the Mel-cepstral representation. In comparison to the baseline results, the recognition accuracy dropped by 1.6% for $AP_{\delta 1}$ and by 3.6% for $MFCC_{E_{\delta 1 \delta 2}}$ when training on males and testing on females. On the other hand, when training on females and testing on males, the recognition accuracy dropped by 4.2% for $AP_{\delta 1}$ and by 8% for $MFCC_{E_{\delta 1 \delta 2}}$. The lower results obtained when training on females could be due to insufficient number of training sentences, 296, in comparison to the 616 male training sentences.

Table 8.8: Recognition results using 8 mixtures. First column. training done with speech produced by males and recognition done with speech produced by females. Second column. training done with speech produced by females and recognition done with speech produced by males.

Signal Representation	%correct/%accurate training:male, testing:female	%correct/%accurate training:female, testing:male
<i>MFCC_Eδ1δ2</i>	78.7/66.9	76.1/62.4
<i>APδ1</i>	79.5/67.8	76/65.2

8.6 Concluding Remarks

It was shown in this chapter that the derived APs can be used in the HMM recognition framework producing comparable results to the Mel-cepstral parameters. Furthermore, it was shown that the derived APs are more robust than the Mel-cepstral parameters to gender differences. The robustness to speaker differences, such as gender, is an important feature of a signal representation used for speaker-independent speech recognition. While similar robustness may be achieved by performing different transformations on the spectral or cepstral representation of the speech signal (e.g., [79], [80], [81], [82]), it should be noted that this is not the sole objective of the derived parameters. The main objective of the derived parameters is to explicitly target the relevant phonetic information in the speech signal and as a byproduct

to reduce speaker-dependent effects. Furthermore, in deriving the APs, knowledge about the acoustic realization of the phonetic features was heavily utilized and in some cases refined. These two properties of the APs differentiate them from other signal representations based on cepstra.

The experiments reported in this chapter were used as a vehicle to compare the APs to Mel-cepstra while keeping the recognition paradigm (HMM) and the model complexity similar. The APs by way of derivation, are not meant to be deployed as the front-end in a frame-based HMM framework, but are tailored for an event-based approach to speech recognition. Using this event-based approach, only APs that are relevant to a sought phonetic feature will be computed at selected instants in time. Such a computation is expected to match the way the APs were developed, i.e. following the phonetic feature-hierarchy. In the HMM recognition framework, all parameters had to be computed at every time frame. Thus, certain APs were computed for sounds that were not considered in the development of the parameters. For instance, the parameters that are relevant to the anterior/nonanterior contrast were developed by considering only the strident fricatives. In the signal representation for HMM-based recognition, these parameters were computed at every time instant without first determining whether that time instant is fricated and strident as the case would be in an event-based approach. However, an event-based paradigm for speech recognition was not the objective of this thesis. Despite this fact, the viability of this paradigm was shown through the manner-class recognition experiments in Chapter 5 and the semivowel recognition experiments in [40].

The computation of all APs at every time, whether they are relevant or not, results in a large dimensional vector for the HMM framework. This problem will become

more acute as more phonetic features relevant to vowels and sonorant consonants are considered. There are a couple of possible ways to deal with the dimensionality problem: (1) use a standard dimensionality reduction techniques such as discriminant analysis or principle decomposition or (2) map each set of acoustic parameters to the phonetic feature they target using the tree classifiers, Neural Nets or fuzzy rules. In the latter case, the input to the HMM system will be degrees of belief (e.g. probability or possibility) in the existence of each phonetic feature.

Chapter 9

Discussion and Conclusions

The lessons that we learned from this research about the speech process are discussed in Section 9.1. The main results obtained in this thesis are summarized and discussed in Section 9.2. Directions for future research are discussed in Section 9.3.

9.1 The Feature-Based Approach to Speech Analysis and Recognition as a learning Tool

The undertaken feature-based approach to speech analysis and recognition was described in Chapter 1 as a learning tool about the speech process. What we learned in the course of this thesis is the following:

- A better understanding of the acoustic correlates of phonetic features.
- An understanding of the signature of the acoustic correlates of phonetic features in the speech signal. This is illustrated by the acoustic parameters we developed

for the manner and obstruent place-of-articulation phonetic features.

- The relative importance of acoustic parameters and therefore acoustic correlates in detecting the acoustic signature of a phonetic feature.
- Acoustic parameters that target phonetic features can be designed to be gender-independent by appropriately normalizing them in time and frequency. The normalization method used depends on the targeted phonetic feature and is based on the articulatory and acoustic correlates of that feature.
- How contextual variability is reflected in the APs and the frequency with which they occur.

9.2 Summary

Three major tasks were undertaken in this thesis: (1) the development of acoustic parameters (APs) that explicitly target the phonetically relevant information in the speech signal. (2) the exploration of the APs in an event-based approach to speech recognition, and (3) the exploration of the APs in the Hidden Markov Model framework for speech recognition.

APs related to the acoustic properties of the manner features: sonorant, syllabic, fricated and noncontinuant were derived. These APs were used in an event-based paradigm for manner-class recognition. The derivation of these APs was motivated by acoustic-phonetic studies and was based on subjective analysis of measurements obtained on the TIMIT database. The mapping from the acoustic parameters to the related phonetic features was done through fuzzy rules motivated by our spectrogram

reading experience.

In recognition, as in development, parameter computation was guided by the phonetic feature hierarchy (see Chapter 5). Thus, not all parameters were computed everywhere. Rather, some parameters were computed only in obstruent regions while others were computed only in sonorant regions. In the manner-class recognition task, the developed system outperformed an HMM system with a Mel-cepstrum front-end and unimodal Gaussians representing the observations' probability distributions. In order to determine whether this performance was due to the front-end signal representation (Mel-cepstra vs. APs) or to the recognition framework (HMM vs. event-based), the APs developed for an event-based system were adapted to fit into the HMM framework and the task of manner-class recognition was carried out. Recognition results showed that the event-based system outperformed the HMM system when both were using the APs as the front-end. These results confirm that an event-oriented paradigm to speech recognition is a viable one.

When the front-end of the HMM system was changed to include the first and second derivatives of the raw signal representation (Mel cepstra or APs) and when the model complexity of the HMM system increased (8-mixture Gaussian observation probability distributions instead of unimodal Gaussians), the HMM system performed better than the event-based system. This means that the dynamic information, derivatives of cepstra or acoustic parameters, contains information relevant to the recognition process. In addition, the multi-mixture observation probabilities were able to capture and better model sources of acoustic variability (e.g., gender and context) leading to improved recognition results.

The HMM system with the Mel cepstra and their derivatives used more informa-

tion (39 parameters) than the event-based system (13 parameters). However, due to the automatic learning capability of the HMM system, this additional information was easily added and the system was relied upon to learn from it. Currently, our hand-designed event-based system does not have that automatic learning capability. Integration of any additional information will require detailed analysis and hand crafting, a laborious and time-consuming process. The original motivation to hand crafting the system was to control the decision process so that we know how parameters contribute to the recognition process. Although such an approach has value, it calls into question our ability to construct a more complex system that will be needed for full speech recognition. On the other hand, the high performance of the event-oriented paradigm is a motivation to further pursue this approach while adding automatic learning capabilities that can make use of our existing acoustic phonetic knowledge and that can enrich that knowledge if possible.

Comparison of the HMM systems using Mel cepstra and Mel-cepstra with their derivatives to those using the APs and the APs with their derivatives, respectively, showed that the APs are able to target the linguistic information in the speech signal and that they are more robust to gender differences. These results motivated the development of APs that target additional phonetic features. However, in order to make the parameter derivation process more objective and to reduce the labor involved, an automatic procedure was developed.

In addition to comparing the APs to Mel-cepstra and the HMM framework to the event-based approach, an error analysis was conducted on the recognition results of the event-based system. This analysis showed that a high percentage of the declared errors, such as a canonically fricative /v/ recognized as a sonorant consonant, were

indicative of systematic contextual variabilities that alter the acoustic realization of a sound from its canonical form. Visual spectrographic inspection of some of these "declared" errors verified that they were not errors, but that they reflected the acoustic manifestation of the corresponding sounds. In addition, this visual inspection suggested that the alteration happened along one or two phonetic-feature dimensions while other features of the phones have not been changed. For instance, inspection of sonorant /v/'s showed that these /v/'s could still be recognized as such based on other phonetic features such as labial, consonantal, nasal etc. Thus, it seems that accounting for this type of variability, in a word or a phone recognition system, could be efficiently done at the lexical level in terms of allowable changes to phonetic features. This is the case as this type of variability usually effects groups of phones rather than one phone. For instance, not only /v/'s could be manifest as sonorant in intervocalic consonant position, but also the canonically voiced fricative /ð/ could be manifest as sonorant in that same or similar context. How such variability could be modeled or accounted for will depend on the recognition paradigm. For instance, in the event-based approach, this type of variability could be accounted for in the lexical description of words if they are defined in terms of phonetic features. In this case, for a word containing an intervocalic /v/, the sonorant feature for that /v/ could be left unspecified so that if a sonorant acoustic event is detected in the speech signal in the time span corresponding to that /v/, a mismatch is not declared. Alternatively, this alteration could be accounted for through rules in the lexical access component. In the HMM framework, if context-dependent models are built, it is feasible that /v/'s in intervocalic position share the same acoustic model independent of the identity of neighboring vowels, thereby reducing the number of acoustic models.

Motivated by the results obtained with the APs related to the manner features and by the need to reduce the subjectivity in parameter derivation as well as the human labor involved, an automatic procedure for the development of APs was devised (see Chapter 6) and tested in this thesis (see Chapter 7). This procedure makes use of statistical methods (Fisher Criterion and automatic classification trees) and acoustic phonetic knowledge to derive, from speech samples, parameters that best distinguish between a phonetic feature and its antonym. Using this procedure, parameters relevant to the phonetic features: sonorant, syllabic, strident, anterior, labial, velar and alveolar were derived. As before, an important attribute of the derived parameters is that they were made relative in time and/or frequency to focus on the phonetic content of the speech signal and reduce speaker-dependent, e.g. gender, effects on the parameter space. The ability of the derived parameters to target the linguistic message in the speech signal was demonstrated by the high correct classification rates obtained at the feature level (see Chapter 7). The ability to reduce speaker-dependent effects was illustrated through the similarity between the female and male distributions in the parameter space.

The acoustic parameters were also used as a front end to an HMM system. In doing so, each parameter was computed as a part of an observation vector at every time frame. Under the assumption of unimodal Gaussian distributions for the observation vector given an HMM state, the HMM system using the acoustic parameters performed better than the HMM system with Mel cepstra as the front end, both with derivatives and without derivatives. However, as the observation distributions were made more complex with 8 Gaussian mixtures, the HMM results were slightly better than the acoustic parameter results (1% difference). This difference in results was

mainly due to insufficient recognition accuracy of the stop consonants. In this case, it was pointed out that additional parameters that target dynamic information such as formant transitions may be needed.

Although the focus in this thesis has been on speech recognition applications, this work has implications in other areas of speech research. The relationships between the abstract phonetic features and their acoustic realizations can be used to develop aids for the hearing impaired and the speech impaired. For instance, in the case of the hearing impaired, the inability of a human subject to distinguish among speech sounds may be due to deficits in his/her ability to perceive certain acoustic properties of the speech signal. Knowing what acoustic properties make a speech sound distinctive from others can guide the development of algorithms that enhance the perception of these properties such as emphasizing certain frequency components. These algorithms can be implemented on a chip that gets implanted in the subject's ear.

The signal representation based on APs can also be used as a visual aid for speech pathologists and their clients to help them understand aspects of the speech signal that are not properly produced. The client in this case may be a person learning English as a second language or a person with speech impediments. The acoustic parameters can also be used in developing devices for the speech impaired. For instance, alaryngeal speakers who use an artificial larynx may not be able to produce important aspects of the speech signal (e.g., alaryngeal speakers usually have problems producing strong turbulent sounds such as /s/). Using the sonorant acoustic parameters, a speech enhancement device may be able to first detect sonorant from nonsonorant segments in the alaryngeal speech signal, a process that can also be done with other speech recognition systems. The parameters can then guide the process-

ing of the speech signal to produce more intelligible speech. For instance, the speech signal can be processed so that the parameter values corresponding to the stridency feature fall in the range of the same parameters measured on normal speech. Finally, this work may have impact on the area of speech synthesis by rule [1]. Knowing what acoustic information needs to be produced in order to relay the linguistic message will help emphasize that information in the synthesized signal. In addition, knowing how different phonetic features relate to each other acoustically and how they vary within one's speech depending on context may lead to synthesizing more natural speech.

9.3 Future Work

The work presented in this thesis can be extended in many ways. The results obtained in this research should be looked at as encouraging, but preliminary since the speech recognition problems addressed were limited in scope. Harder speech recognition tasks such as phoneme recognition and word recognition will necessarily raise many issues that were not addressed in this research.

First, it is clear that phonetic features can be used as a basis for developing acoustic parameters that target the phonetic content in the speech signal. In order to carry phoneme recognition or word recognition, acoustic parameters pertinent to additional phonetic features must be developed. There are 20 or so phonetic features. Based on our experience thus far, there could be on the average of about 2 to 3 parameters per phonetic feature. Thus, about 60 parameters may be needed to target all features. This is about 1.5 times the number of Mel cepstral coefficients used in today's speech recognition systems (39, including derivative features). The effect

of the number of parameters on recognition performance greatly depends on the recognition paradigm. An event based paradigm, such as the one explored in this research, would use only a subset of parameters at different stages of the recognition process. That is, parameters relevant to a phonetic feature will need to be computed if acoustic evidence for that feature is sought. That is, not all parameters need to be computed at once. In such a case, if probabilistic models are to be built for each phonetic feature, only the parameters relevant to that feature need to be considered in this model, and all samples that possess that feature will be used to train these models whereas those that possess the antonym feature will be used to train the competing model(s). On the other hand, if the parameters are to be used in a frame-based recognition framework as the HMM, all parameters will need to be concatenated in one observation vector with 60 dimensions. This large dimensionality may lead to a training problem as large sets of training data will be needed to obtain reliable acoustic models of phones. The problem will become more acute if the derivatives of these parameters are blindly added to the observation space, as was done in this thesis, further increasing the observation-space dimensionality¹.

The dimensionality problem may be alleviated in several ways. One possibility is the use of traditional dimension-reduction techniques, such as linear discriminant analysis, to reduce the dimensionality of the observation space. Alternatively, the acoustic parameters can be mapped to an intermediate signal representation in terms

¹It was empirically shown in this thesis that having APs and their derivatives as a signal representation results in better recognition performance than having the APs alone. In these experiments, The derivative of each AP was blindly added. It is not clear if this is needed as the derivatives of some APs may not be relevant. This issue requires further investigation.

of phonetic features (a vector of 20 components). This representation in the phonetic-feature space may then be used as the front end to the recognition system. The feature signal representation can be in terms of probabilities of phonetic features estimated from the acoustic parameters. The advantage of this approach is that samples from many speech sounds can be pooled together to train feature models leading to an efficient use of training data. This is the case since phonetic features higher up in the feature hierarchy are shared by many phonemes. Efficient use of training data can also be accomplished by sharing data among phones along the acoustic-parameter dimensions related to the phonetic features that they have in common. These alternatives will need to be well formulated and empirically explored. It should be noted that data pooling is done today through context-dependent model clustering for a phone across different contexts without sharing data between phones.

There are additional experiments that need to be done in order to assert how to gain full advantage of the AP signal representation within the HMM framework. In the experiments conducted thus far, diagonal covariance matrices were assumed for the Gaussian distributions. This assumption asserts that the acoustic parameters are independent. Intuitively, this is not the case as there is a high degree of correlation between acoustic parameters, especially those that relate to the same phonetic feature. In addition, the hierarchical organization of features suggest that the phonetic features are dependent on each other and so are there corresponding parameters. Thus, full covariance matrices or block diagonal covariance matrices (capturing the dependence among acoustic parameters related to the same phonetic feature only) need to be experimented with. This could result in better recognition results with the APs in comparison to cepstral-based parameters since cepstral parameters have been shown

to be roughly independent as they approximate principal decomposition. In addition, the role of the APs and their performance in building context-dependent models of phones or speech classes need to be explored. The question in that case is whether knowing about acoustic variability at the phonetic feature level and knowing how phones share phonetic features can lead to a more compact model space.

Several possible experiments in the context of the HMM framework were discussed. However, the ultimate goal is to gain full advantage of a signal representation based on phonetic features in the event-based framework that does not require frame-by-frame decision making, as the HMM does, and does not make an assumption that speech is merely a concatenation of time stretches corresponding to different phones. It is postulated that an event-based approach will allow the detection of speech sounds even if most of the evidence for their existence is overlapping with another sound. For instance, it has been observed from analysis of recognition errors that the strongest evidence of an /r/ in a stop-/r/ sequence may be in the frication of the preceding unvoiced stop consonant. How to detect such an event is still a question that needs to be addressed. Finally, a fully integrated speech recognition system whereby the signal as well as the lexicon are represented in terms of phonetic features still need to be explored. It seems that such an approach may allow better modeling and handling of contextual variability than existing approaches based on context-dependent phone models. However, such an approach still needs to be formulated. Although appealing, the process may be too complicated to be practically implemented for real recognition tasks such as continuous speech recognition. However, it may have a practical application in an N-best rescoring paradigm whereby an initial segmentation of speech is obtained through some other means to produce a list of hypothesized sentences and

this event-base approach is used to rescore or reorder the sentences.

Appendix A

Fuzzy Evaluation Index

The fuzzy evaluation index (FEI) is based on fuzzy set theory and was first proposed in [83] to evaluate the goodness of a **single** feature in discriminating between **two** classes. The term “feature” in this criterion is similar to what we call “acoustic measure” and should not be confused with a “phonetic-feature”. Several FEI’s were defined in [83] based on the index of fuzziness, entropy and π -ness and using *S-shaped* and *π -shaped* membership functions. In all cases, the FEI is defined so that it decreases in value as the feature reliability in characterizing the considered classes increases. In this section, the FEI definition that uses entropy and an *S-type* membership function is considered to illustrate the objective of an FEI. Let C_1 and C_2 be two different classes of interest and let q be the feature being evaluated for its ability to separate between the two classes. Furthermore, let n_1 and n_2 be the number of samples from C_1 and C_2 , respectively. Then, FEI for feature q is defined as:

$$(FEI)_q = \frac{H_{12}^s}{H_1^s + H_2^s} \quad (\text{A.1})$$

where the entropy for class C_i using the S - type membership function $\mu(x)$ is given by.

$$H^s_i = \frac{1}{n_i} \sum_{j=1}^{n_i} K(\mu(x^{(i)}_j)) \quad (\text{A.2})$$

and

$$K(\mu(x^{(i)}_j)) = [-\mu(x^{(i)}_j) \log_2(\mu(x^{(i)}_j)) - (1 - \mu(x^{(i)}_j)) \log_2(1 - \mu(x^{(i)}_j))]. \quad (\text{A.3})$$

H^s_{12} in equation A.1 is computed by pooling the observed samples from C_1 and C_2 together resulting in $n_{12} = n_1 + n_2$ samples. The S - type function involved in the computation of H^s_i is depicted in Figure A.1 where b is the sample mean for class C_i defined as

$$b = \bar{x}^{(i)} = \frac{1}{n_i} \sum_{j=1}^{n_i} x^{(i)}_j$$

The extreme points a and c are defined by:

$$c = b + d$$

$$a = b - d$$

where d is the maximum distance between an observed sample and the sample mean and is given by:

$$d = \max | (\bar{x}^{(i)} - \max_j(x^{(i)}_j)) |, | (\bar{x}^{(i)} - \min_j(x^{(i)}_j)) |.$$

$(FEI)_q$ is minimum when each of H^s_1 and H^s_2 is maximum while H^s_{12} is minimum. The function $K(x)$ involved in the computation of the entropy H^s_i is monotonically increasing in the interval $[0,0.5]$ and monotonically decreasing in the interval $[0.5,1]$ with a range $[0,1]$ as depicted in Figure A.2. Thus, the closer a sample value $x^{(i)}_j$

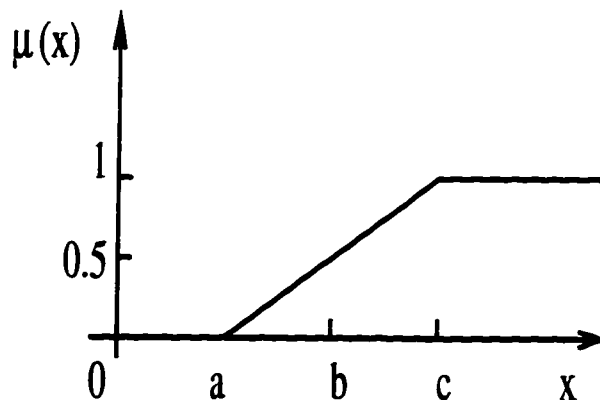


Figure A.1: An S – type membership function.

is to the sample average (note $\mu(\bar{x}^{(i)}) = 0.5$), the higher is the value of $K(\mu(x_j^{(i)}))$. Consequently, H^s_i is higher in value when more sample values are clustered around the sample average. A higher H^s_i value indicates that the intraclass variation for C_i with respect to the feature q is smaller. A characteristic of a good feature is to minimize the intraclass variation. The other characteristic is to maximize interclass separability and this is achieved by minimizing H^s_{12} . The more the samples of both classes are away from the sample average (the sample average in this case is computed over the samples from C_1 and C_2 pooled together), the smaller is the value of H^s_{12} . Consequently, a good feature should have a low FEI value.

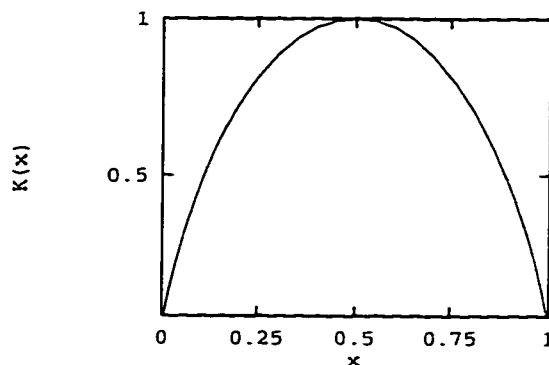


Figure A.2: The function $K(x)$ involved in the computation of the entropy.

Bibliography

- [1] Klatt. Dennis, "Review of Text-to-Speech Conversion for English." *The Journal of the Acoustical Society of America*, 82(3). September 1987.
- [2] Pallet. David *et al.*, "1993 Benchmark Tests for the ARPA Spoken Language Program." *Papers in ARPA Workshop on Human Language Technology*. March 8-11. 1994. Plainsboro. N.J.
- [3] Ostendorf. M. *et al.*, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses." *DARPA Proceedings*. February 1991. pp. 83-87.
- [4] Zavaliagos. G. *et al.*, "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition." *IEEE Transactions on Speech and Audio Processing*, Vol. 2, January 1994. pp. 151-160.
- [5] Bourlard. Herve A. and Morgan, Nelson, *Connectionist Speech Recognition*. Kluwer Academic Publishers, Boston, 1994.
- [6] Cole. R. *al.*. "The Challenge of Spoken Language Systems: Research and Directions for the Nineties." *IEEE Transactions on Speech and Audio Processing*. Vol. 3. No. 1. January 1995, pp. 1-21.

- [7] Chomsky, Noam and Halle, Morris. "The Sound Pattern of English." The MIT Press, Cambridge, Massachusetts, U.S.A.
- [8] Stevens, Kenneth *et al.*, "Implementation of a Model for Lexical Access Based on Features." *Proceedings ICSLP*, Oct. 12-16, 1992, Banff, Alberta, Canada, pp. 499-502.
- [9] Studdert-Kennedy, M., "Speech Perception." *Language and Speech*, v.23, pp. 45-66.
- [10] Cohen, L., "Time-Frequency Distributions-A review." *Proceedings of the IEEE*, Vol. 77, No. 7, pp. 941-81, July 1989.
- [11] Nawab, S. Hamid and Quatieri, Thomas F., "Short-time Fourier Transform." *Advanced Topics in Signal Processing, chapter 6*, edited by Lim and Oppenheim, Prentice Hall, Englewood Cliffs, New Jersey.
- [12] Rioul, Oliver and Vetterli, Martin, "Wavelets and Signal Processing." *IEEE Signal Processing Magazine*, October 1991.
- [13] Claassen, T.A.C.M and Mecklenbrauker, W.F.G., "The Wigner Distribution. A tool for time-frequency Signal Analysis. Part III: Discrete-Time Signals." *Phyllips J. Res.*, Vol. 35, 1980, pp. 276-300.
- [14] Seneff, Stephanie, "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing," *Journal of Phonetics*, 16, 1988, pp. 55-76.
- [15] Davis, Steven B. and Mermelstein, Paul, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences."

- IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-28, No. 4, August 1980.
- [16] Rabiner, L.H and Schafer, R.W, *Digital Processing of Speech Signals*. Prentice Hall, Englewood Cliffs, New Jersey, 1978.
- [17] Rabiner, L.H and Juang, B.H., "An Introduction to Hidden Markov Models." *IEEE Acoustic, Speech and Signal Processing Magazine*. January 1986.
- [18] Ostendorf, Mari and Roukos, Salim. "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition." *IEEE Transactions on Acoustics, Speech and Signal Processing*. Vol. 37, No. 12, December 1989, pp. 1857-1869.
- [19] Ghitza, Oded. "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition." *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 1, Part II, January 1994, pp. 115-132.
- [20] Jankowski, Charles R. Jr., Vo, Hoang-Doan H. and Lippman, Richard P., "A Comparison of Signal Processing Front Ends for Automatic Word Recognition." *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 4, July 1995, pp. 286-292.
- [21] Strobe, Brian and Alwan, Abeer, "A model of Dynamic Auditory Perception and Its Application to Robust Speech Recognition." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA., 1996, pp. 37-40.

- [22] Eide. Ellen et al., "A Linguistic Feature Representation of the Speech Waveform." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1993*. Vol. 2. pp. 483-486.
- [23] Lea, Wayne. "Trends in Speech Recognition." Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.
- [24] Jakobson, R., Fant, G., and Halle, M., "Preliminaries to Speech Analysis." The MIT Press, Cambridge, Massachusetts, U.S.A, 1951.
- [25] Jelinek, F., Bahl, L.R., and Mercer, R.L., "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech." *IEEE Transactions on Information Theory*, Vol. IT-21, 1975, pp. 250-256.
- [26] Sagey, Elizabeth Caroline. "The Representation of Features and Relations in Non-Linear Phonology," PhD Thesis, MIT, 1986.
- [27] Lahiri, Aditi and Marslen-Wilson, William. "The Mental Representation of Lexical Form: A phonological Approach to the recognition Lexicon." *Cognition*, 38, 1991, pp.245-294.
- [28] Ladefoged, Peter, "A course in Phonetics." Second Edition. Harcourt Brace Jovanovich Inc., 1982.
- [29] Delattre, P. and Freeman. "A Dialect Study of American R's by X-ray Motion Picture." *Language*, 44, 1968, pp. 29-68.

- [30] Hedrick. Mark S. and Ohde. Ralph N.. "Effect Of Relative Amplitude Of Frication On Perception Of Place Of Articulation." *The Journal of the Acoustical Society of America*, Vol. 94, No. 4, Oct. 1993.
- [31] Zue. Victor. "Spectrogram Reading Notes," 1985. A course offered at the Massachusetts Institute of Technology.
- [32] Fant. C. G. M.. *Acoustic Theory of Speech Production*. Mouton. The Hague. 1960.
- [33] Stevens. Kenneth. "Acoustic Phonetics". forthcoming. Available in manuscript.
- [34] Stevens. Kenneth. "MIT class notes in speech communication."
- [35] Glass. James Robert, " Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment." Master thesis. MIT. 1984.
- [36] Chen. Francise. "Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary," PhD Thesis. MIT. 1985.
- [37] Lahiri. A.. Gewirth, L. and Blumstein. S.E.. "A Reconsideration of Acoustic Invariance for Place of Articulation in Diffuse Stop Consonants: Evidence from a Cross-Language Study." *The Journal of the Acoustical Society of America*. 76. 1984. pp. 391-404.
- [38] Zierten. Stephanie and Espy-Wilson. C.Y.. "Automatic Classification of Labial and Alveolar Stop Consonants". *The Journal of the Acoustical Society of America*. the 125th meeting of the Acoustical Society of America. Ottawa. May. 1993.

- [39] Zierten Stephanie. "Automatic Classification of Labial and Alveolar Stop Consonants in American English," B.S Thesis. Speech Communication Lab., Boston University. September 1993.
- [40] Espy-Wilson. C. Y.. "An Acoustic-Phonetic Approach to Speech Recognition: Application to the Semivowels." *RLE Technical Report# 531*. MIT. 1987.
- [41] Espy-Wilson. C. Y.. "A Feature-Based Approach to Speech Recognition." *The Journal of the Acoustical Society of America*. Vol. 96. 1994. pp. 65-72.
- [42] Stevens. Kenneth. "Relational Properties as Perceptual Correlates of Phonetic Features." *Proceedings of the Eleventh International Conference of Phonetic Sciences. Tallinn. Estonia. USSR. 1987*. Vol. 4.
- [43] Bitar. Nabil. "A Feature-Based Broad Speech Classifier." *The Journal of the Acoustical Society of America*. the 127th meeting of the Acoustical Society of America. Boston. June. 1994.
- [44] Dubois. Didier and Prade. Henri. "Fuzzy Sets and Systems: Theory and Applications." Academic Press. New York, 1980.
- [45] Zadeh. L. A.. "Fuzzy Sets," *Information and Control*. Vol. 8. 1965. pp. 338-352.
- [46] Zimmerman. Hans J.. "Fuzzy Sets. Decision Making. and Expert Systems." Kluwer Academic Publishers. 1987.
- [47] "TIMIT Acoustic-Phonetic Continuous Speech Corpus." CD-ROM *NIST Speech Disc1-1.1*. National Institute of Standards and Technology. October 1990.

- [48] Diaglakis. V. *et al.*, "Fast Search Algorithms for Phone Classification and Recognition Using Segment-Based Models." *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 40, No. 12, December 1992, pp. 2885-2896.
- [49] Stevens, Kenneth, "Phonetic Evidence for Hierarchies of Features." in *Phonological Structure and Phonetic Form Papers in Laboratory Phonology III*, edited by Patricia A. Keating, Cambridge University Press, 1994.
- [50] De Mori, Renato. *Computer Models of Speech Using Fuzzy Algorithms*. Plenum Press, New York, 1983
- [51] Rabiner, L. R. and Sambur, M. R., "An Algorithm for Determining the End-points of Isolated Utterances," *The Bell System Technical Journal*, Vol. 54, No. 2, February 1975.
- [52] Mermelstein, Paul. "Automatic Segmentation of Speech Into Syllabic Units." *The Journal of the Acoustical Society of America*, Vol. 58, No. 4, Oct. 1975, pp. 880-883.
- [53] *HTK V1.5*. Cambridge University Engineering Department Speech Group and Entropic Research Laboratory Inc., Sept. 1993.
- [54] Olive, Joseph *et. al.*, *Acoustics of American English Speech*. Springer-Verlag, New York, 1993
- [55] Catford, J.C., *A Practical Introduction to Phonetics*. Clarendon Press: Oxford, 1988.

- [56] Boyce, Suzanne and Espy-Wilson, Carol, "Coarticulatory Stability in American English /r/." *The Journal of the Acoustical Society of America*. Vol. 101. No. 6, June 1997.
- [57] Deng, L. & Sun, D., "Phonetic Classification and Recognition Using HMM Representation of Overlapping Articulatory Features for all Classes of English Sounds." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1994*, Vol. 1, pp. 45-48.
- [58] Phillips, Michael and Zue, Victor, "Automatic Discovery of Acoustic Measurements for Phonetic Classification." *Proceedings ICSLP*. Oct. 12-16, 1992. Banff, Alberta, Canada. pp. 795-798.
- [59] Duda, R. and Hart, P., "Pattern Classification and Scene Analysis." John Wiley and Sons, Inc., 1973.
- [60] Waves5.1. Entropic Research Laboratory Inc.
- [61] In *Statistical Models in S*, edited by Chambers, J. M. and Hastie, T. J., Wadsworth & Brooks, 1991.
- [62] Breiman, Leo, Friedman, Jerome H., Olshen, Richard A. and Stone, Charles J., *Classification and Regression Trees*, Wadsworth & Brooks, Monterey, California, USA.
- [63] Kent, Ray D. and Read, Charles, "The Acoustic Analysis of Speech." Singular Publishing Group Inc., San Diego, California, 1992.

- [64] Heinz. John and Stevens, Kenneth. "On the Properties of Voiceless Fricative Consonants." *The Journal of the Acoustical Society of America*, Vol. 33, No. 5, May 1961.
- [65] Shadle, Christine Helen, "The Acoustics of Fricative Consonants." *RLE Technical Report# 506*, The Massachusetts Institute of Technology, 1985.
- [66] Wilde, Lorin. "Analysis and Synthesis of Fricative Consonants." PhD Thesis, MIT, 1995.
- [67] Sydral, A.K. and Gopal, H.S., "A perceptual Model of Vowel Recognition Based on Auditory Representation of American English Vowels." *The Journal of the Acoustical Society of America*, Vol. 79, 1986, pp. 1086-1100.
- [68] Blumstein, Sheila and Stevens, Kenneth. "Perceptual Invariance and Onset Spectra for Stop Consonants in Different Vowel Environments." *The Journal of the Acoustical Society of America*, Vol. 67, February 1980.
- [69] Lamel, Lori. "Stop Identification from Spectrograms." Massachusetts Institute of Technology Project Report, May 1985
- [70] Bush, Marcia, Kopec, Gary and Zue, Victor. "Selecting Acoustic Features for Stop Consonant Identification," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, 1983.
- [71] Sussman, Harvey M. *et al.*. "An Investigation of Locus Equations as a Source of Relational Invariance for Stop Place Categorization." *The Journal of the Acoustical Society of America*, Vol. 90, September 1991, pp. 1309-1325.

- [72] Forrest, Karen *et al.*, "Statistical Analysis of Word-Initial Voiceless Obstruents: Preliminary Data." *The Journal of the Acoustical Society of America*. Vol. 84. No. 1. July 1997.
- [73] Papoulis, Athanasios, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Inc., 1984.
- [74] Trager, George L. and Smith, Henry Lee. *English Structure*.
- [75] Nilsen, Don D.L. and Nilsen, Allen Pace. "Pronunciation Contrasts in English," Regents Publishing Company, Inc.. 1973.
- [76] McCasland, G.P., "Stridency as a Distinctive Feature of American Fricatives." Paper presented at the annual meeting of the Modern Language Association. New York. December 29, 1978.
- [77] Bitar, Nabil *et al.*, "Strident-Feature Extraction in English Fricatives." *The Journal of the Acoustical Society of America*, the 125th meeting of the Acoustical Society of America, Ottawa, May, 1993.
- [78] Hasegawa-Johnson, Mark Allan, "Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification." PhD Thesis. MIT, 1996.
- [79] Eide, Ellen and Gish, Herbert, "A Parametric Approach to Vocal Tract Normalization." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA., 1996.

- [80] Kamm, T., Andreou, A., and Cohen, J.. "Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability." *Proceedings of the Fifteenth Annual Speech Research Symposium*, Baltimore, MD. June, 1995.
- [81] Wegman, Steven *et al.*, "Speaker Normalization on Conversational Telephone Speech." *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA. 1996.
- [82] Lin, Qiguang and Che, Chiwei. "Normalizing the Vocal Tract Length for Speaker Independent Speech Recognition." *IEEE Signal Processing Letters*. Vol. 2. No. 11. November, 1995.
- [83] Pal, Sankar K. and Chakraborty Basabi. "Fuzzy Set Theoretic Measure for Automatic Feature Evaluation." *IEEE transactions on Systems, Man and Cybernetics*. Vol. SMC-16, No. 5, Sept./Oct. 1986, pp. 754-760.

Vita

Nabil Bitar was born in Beirut, Lebanon on October 27, 1966. He completed his high school in 1984 at the Nouvelle Ecole De St. Elie in Beirut earning the Matheleme (Baccalaureate second part). He came to Boston in 1985 to study English and went on to earn his Bachelor degree, his Master degree and his Doctorate, all in Electrical Engineering, at Boston University. Nabil served as a teaching assistant and laboratory assistant for several electrical engineering courses. He worked at GTE Laboratories for about two years in advanced wireless intelligent networks and in wireless data. Then, he moved on to work in the broadband data communication area.