Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition

by

Andrew K. Halberstadt

B.M., Eastman School of Music (1991)
B.S., University of Rochester (1992)
M.S., University of Rochester (1993)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

November 1998
[February, 1999]

Author ....
Department of Electrical Engineering and Computer Science
November 20, 1998

Certified by ....
James R. Glass
Principal Research Scientist
Thesis Supervisor

Accepted by ....
Arthur Smith
Chairman, Departmental Committee on Graduate Students

Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition

by

Andrew K. Halberstadt

Submitted to the Department of Electrical Engineering
and Computer Science
on November 20, 1998, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The acoustic-phonetic modeling component of most current speech recognition systems calculates a small set of homogeneous frame-based measurements at a single, fixed time-frequency resolution. This thesis presents evidence indicating that recognition performance can be significantly improved through a contrasting approach using more detailed and more diverse acoustic measurements, which we refer to as heterogeneous measurements.

This investigation has three principal goals. The first goal is to develop heterogeneous acoustic measurements to increase the amount of acoustic-phonetic information extracted from the speech signal. Diverse measurements are obtained by varying the time-frequency resolution, the spectral representation, the choice of temporal basis vectors, and other aspects of the preprocessing of the speech waveform. The second goal is to develop classifier systems for successfully utilizing high-dimensional heterogeneous acoustic measurement spaces. This is accomplished through hierarchical and committee-based techniques for combining multiple classifiers. The third goal is to increase understanding of the weaknesses of current automatic phonetic classification systems. This is accomplished through perceptual experiments on stop consonants which facilitate comparisons between humans and machines.

Systems using heterogeneous measurements and multiple classifiers were evaluated in phonetic classification, phonetic recognition, and word recognition tasks. On the TIMIT core test set, these systems achieved error rates of 18.3% and 24.4% for context-independent phonetic classification and context-dependent phonetic recognition, respectively. These results are the best that we have seen reported on these tasks. Word recognition experiments using the corpus associated with the JUPITER telephone-based weather information system showed 10-16% word error rate reduction, thus demonstrating that these techniques generalize to word recognition in a telephone-bandwidth acoustic environment.

Thesis Supervisor: James R. Glass
Title: Principal Research Scientist

Acknowledgments

As I look back over the years spent on this thesis, there are many people to thank.

I would like to thank my thesis advisor, Jim Glass, for his insightful suggestions, comments, and feedback throughout this project. I would like to thank Victor Zue for working to bring together the resources and the people to make the Spoken Language Systems group a stimulating and collegial place to conduct graduate research. I would like to thank my entire thesis committee, consisting of Ken Stevens, Victor Zue, and Jim Glass, for teaching me about acoustic phonetics and speech recognition both in the classroom and through their input as this thesis progressed.

All of the members of the SLS group have helped me through engaging discussions, pointers to resources, and help with hardware and software difficulties. Jane Chang, Lee Hetherington, Ed Hurley, Ray Lau, Mike McCandless, Kenney Ng, and Christine Pao deserve special mention for answering many questions about typesetting, debugging, and code development. I would like to thank T.J. Hazen and Stephanie Seneff for their comments on drafts of my thesis.

Finally, I would like to thank my wife, Sandi, and my entire family for their love and encouragement throughout my academic program.

This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center, and by the National Science Foundation under Grant No. IRI-9618731.

Contents

List of Figures

List of Tables

Chapter 1

Introduction, Goals, and Motivation

1.1 Introduction

Automatic speech recognition (ASR) systems are designed to convert a speech waveform into text. Accomplishing this task requires several knowledge sources. Acoustic-phonetic knowledge provides a mapping from the speech waveform to hypothesized phonetic units. Lexical knowledge enables the system to combine the phonetic units into words. A language modeling component stores information about the likelihood of different word sequences. This thesis presents new techniques for the development and implementation of the first of these knowledge sources, namely, the acoustic-phonetic modeling component. Acoustic-phonetic modeling for ASR has been studied extensively, and a number of viable solutions have emerged which are being used in a growing commercial market for speech recognition products. These systems typically calculate a small set of homogeneous frame-based measurements at a fixed time-frequency resolution. In spite of the success of these techniques, this thesis presents evidence indicating that recognition performance can be significantly improved through the use of more detailed and more diverse acoustic measurements which we refer to as heterogeneous measurements. These heterogeneous measurements are combined in such a way as to capture the complementary information which they contain.

From a statistical pattern recognition viewpoint, the entire process of extracting acoustic measurements from the speech waveform can be referred to as preprocessing. In general, preprocessing techniques are designed to reduce the dimensionality of the input data while at the same time retaining as much discriminative information as possible. Of course, these are conflicting conditions, and choosing the proper balance between them is an important aspect of the art of designing pattern recognition systems.

In the speech recognition community, the typical tradeoff between data reduction and information retention has evolved over decades of research. There are a large number of studies in the literature which describe and compare various preprocessing algorithms [5, 11, 34, 50, 51, 60, 61, 69, 76, 84, 89, 90, 96], and these citations are only a small sampling. Some of these studies examine acoustic measurements for particular sound classes, such as the 13 monophthong vowels in [60], or the stop consonants in [69]. These phone-class-specific studies lead naturally to the notion of using different measurements in different phonetic classes. This notion has generally not been pursued by researchers who address the problem of word recognition, in part because data reduction took priority over other factors. To illustrate this point, consider the 1993 survey and tutorial paper "Signal Modeling Techniques in Speech Recognition" by Picone [77]. This paper provides a good overview of the copious literature on acoustic-phonetic modeling, including 116 references and thorough coverage of many systems. The entire paper assumes that *one* set of acoustic measurements will be used with a *single* classifier for the acoustic-phonetic modeling of *all* phonetic classes. Thus, the ASR researchers cited in Picone's paper did not make use of acoustic measurements that were designed for particular phonetic subclasses. In addition, researchers have generally not considered the possibility that competing acoustic measurement candidates may in fact contain complementary information, so that the optimal choice may be to produce a system which can use a variety of measurements simultaneously.

In this thesis, we advocate methods for tipping the scales of the preprocessing tradeoff away from data reduction and toward retention of discriminative information. This change is seen as necessary in order for machines to approach human levels of performance. In addition, this direction for research is being facilitated by the ever increasing computational power of computers. In fact, the current conventional approach, which limits itself to low classifier dimensionality, has been shaped in part by consideration of the feasibility of the computation assuming the use of machines that are now several generations old.

In addition to retaining more acoustic-phonetic information from the speech waveform, the design of experiments in this thesis was guided by two ideas about the evaluation of acoustic measurements. The first idea is that measurements should be evaluated in several different tasks, such as vowel classification and obstruent classification, in addition to the typical overall phonetic classification task. This allows for the possibility that measurements will have different strengths and weaknesses in the various phonetic subclasses. The second idea is that measurements should be considered not only for their own individual merit, but also for how they might be used to complement the information in another set of measurements.

1.2 Thesis Goals

The first goal of this thesis is to develop heterogeneous acoustic measurements to increase the amount of acoustic-phonetic information extracted from the speech signal. Diverse measurements are obtained by varying the time-frequency resolution, the spectral representation, the choice of temporal basis vectors, and other aspects of the preprocessing of the speech waveform. The time-frequency resolution contributes to determining which acoustic attributes of the speech will be blurred and which will be retained in the spectral representation. The temporal basis results in additional temporal blurring of acoustic information. The choice of basis determines which aspects of the temporal trajectory of spectral parameters are actually visible to the classifier. Fundamental frequency and zero-crossing rate are also considered in order to add diversity to the acoustic measurements.

The second goal of this thesis is to develop classifier systems for successfully utilizing high-dimensional heterogeneous acoustic measurement spaces. This is accomplished through the combination of multiple classifiers. In one approach, classifiers are combined in a hierarchical tree with classifiers at each non-leaf node to determine the probabilities on the branches leaving that node. Committee-based approaches using voting, linear combination, or an independence assumption are also presented. Finally, hierarchical and committee-based combination techniques can be combined to form hybrid techniques.

The third goal of this thesis is to increase understanding of the weaknesses of current automatic phonetic classification systems. This is accomplished through perceptual experiments on stop consonants which facilitate comparisons between humans and machines. Error analysis is performed, including examination of particular examples of stop consonants that were classified differently by humans and machines.

1.3 Motivation

Three sources of motivation for this work are presented in this section. Each of these areas contributed to the conception of the ideas for the experiments in this thesis.

1.3.1 Comparisons between Humans and Machines

This section contains comparisons between the performance of humans and machines in various speech recognition and classification tasks. It is known that humans perform better than machines in word recognition tasks, but there are several deeper questions to address: *How much* better at speech recognition are humans when compared with machines? What are the *reasons* for the difference in performance? How much of the difference can be attributed to higher-level knowledge such as semantics and syntax? How much to lexical knowledge? And finally, how much to low-level acoustic-phonetic analysis capability?

The last of those questions is the most important for this thesis. We would like to know how humans and machines compare at the low-level task of acoustic-phonetic

Corpus and Description	Vocabulary Size	Recognition Perplexity	Machine Error (%)	Human Error (%)
TI Digits: Read Digits	10	10	0.72	0.009
Alphabet Letters: Read Alphabetic Letters	26	26	5	1.6
Resource Management: Read Sentences (Word-pair Grammar)	1,000	60	3.6	0.1
Resource Management: Read Sentences (Null Grammar)	1,000	1,000	17	2
Wall Street Journal: Read Sentences	5,000	45	7.2	0.9
North American Business News: Read Sentences	Unlimited	160	6.6	0.4
Switchboard: Spontaneous Telephone Conservations	2,000–Unlimited	80–150	43	4

Table 1.1: Human versus Machine Speech Recognition Performance, after Lippmann [53, 54, 55]

analysis of speech segments, when higher-level knowledge sources are not available. From the literature and from our own work described below, we demonstrate that humans are performing significantly better than machines. This implies that there is low-level phonetic information in the speech signal which machines are not currently extracting, and so this provides motivation for the research performed in this thesis.

Speech Recognition by Humans and Machines

By drawing upon machine recognition and human perception literature, Lippmann [53, 54, 55] has gathered together machine recognition results from several talker-independent speech recognition corpora and paired them with comparable human perception results. Table 1.1 summarizes the characteristics of the corpora and several performance comparisons. The table indicates that machines perform 80 times worse than humans at recognizing digits [12, 48], 5 times worse for alphabetic letters [15, 18], 36 times worse for the 1000-word Resource Management task with a word-pair grammar [37, 72, 100], 9 times worse for Resource Management with a null-grammar [37, 56, 62, 72], 8 times worse for a 5000-word Wall Street Journal task [22, 38, 73], 17 times worse in a 65000-word North American Business News task [20, 94], and 11 times worse for the Switchboard spontaneous speech corpus [52, 57, 58, 74].

In fact, these estimates of error rate differences are conservative, since the machines were allowed to train on closely matched data. While these results clearly indicate that humans are superior to machines, it is difficult to isolate the acoustic-phonetic component from these experiments, since people may be using higher-level knowledge sources. For this reason, the next section examines phonetic classification in order to observe comparisons of low-level acoustic-phonetic analysis capabilities.

Phonetic Classification by Humans and Machines

We expect that humans will perform better than machines in acoustic-phonetic analysis. Hence, human performance often serves to provide an existence proof, showing that the resulting performance level is achievable. This reasoning would indicate, for example, that if humans achieve 95% in a classification task, then obviously 95% accuracy is achievable based on the given information. However, the converse of the above statement is not necessarily true. That is, if 95% is achievable based on the given information, humans may find it difficult or perhaps impossible to achieve 95%. Put another way, this means that human performance in phonetic classification should not be regarded as a strict upper bound on achievable performance. The principal reasons for this conjecture are practical. First, phonetic classification, as opposed to word recognition, is an unnatural task for humans. Second, if the number of phone candidates is large, it stresses a human's ability to sift all the phonetic candidates in the given time. Third, humans may be affected by inattention and fatigue in experimental trials.

Ideally, we would like to compare machine classification and human perception performance on identical speech data over all phonetic sounds. Unfortunately, we did not find any studies in the literature which fit that description, nor was it possible to glean such comparisons by combining the work of several researchers. We did find human and machine results in the task of vowel classification which share many similarities, although the test sets are not identical. We were able to construct human and machine comparisons for stop consonants by comparing perceptual results from the literature with our own machine classification results.[1]

Vowel Classification by Humans and Machines

Cole [16] has performed perceptual studies on vowels excised from TIMIT[40] . For this study, 16 vowel labels from TIMIT were selected, and 168 tokens of each were extracted, for a total of 2688 tokens. The 16 vowels were [iy ih ey eh ae er ah ax aa ao uh uw ow aw ay oy].[2] There were sixteen subjects participating in the experiment, each of whom participated in 10 1-hour sessions over a two-week period. Each subject classified all of the 2688 tokens. The results show that vowels presented in isolation were identified with 54.8% accuracy, while vowels presented with one segment of context were identified with 65.9% accuracy. Cole notes that these perception results are in close agreement with other experiments using excerpts of speech from TIMIT [65, 75, 88, 89].

In [16], the vowel classification accuracy metric is referred to as "listener-labeler" agreement, thus highlighting the fact that it is problematic to claim a single correct phonetic labeling of an utterance (as is done in TIMIT for scoring purposes), since some segments will be ambiguous and experts will not agree on the phonetic label. This is particularly a problem for vowels, due to the nature of the acoustic space from which vowel distinctions are made. This explains part of the reason why listener-labeler agreement is so much lower for vowels than it is for stops, as we describe in the next section.

A number of researchers have reported machine classification results on a 16 vowel classification task from TIMIT [5, 13, 30, 49, 59]. Unfortunately, these machine classification studies and Cole's study did not use the same set of 16 vowels, although the sets differ by only one element; the machine studies included [ux] where Cole used [ax] instead. In addition, the test sets for the human perception and machine results are

[1] Chapter 6 provides additional results on stop consonants, where we performed both the human and the machine components of the experiments.

[2] See Table 2.1 for the International Phonetic Alphabet (IPA) symbols corresponding to these ARPAbet symbols.

not identical. For these reasons, it is difficult to make direct comparisons, yet general comparisons are still valuable. Goldenthal [30] reported performance of 68.9% using gender-specific models. Carlson and Glass [5] reported 68.7% incorporating speaker normalization information. Chun [13] reports 67.6% using speaker-independent models. Meng [59] and Leung [49] reported 65.6% and 64%, respectively, using auditory models.

With human perceptual results at 65.9% accuracy and machine classification results at 64-68.9%, we conclude that listener-labeler error and machine-labeler error for vowel classification under matched conditions produce very similar performance. It is difficult to be conclusive because of differences in the task ([ax] in the perceptual experiments is probably considerably more confusable with other candidates than the [ux] used in the machine studies) and the test sets. Also, with 16 choices, the unnaturalness of the task and the difficulty involved in carefully considering every vowel candidate in the required time on every trial may have reduced the human's performance. This hypothesis is supported by the fact that the humans in Cole's study [16] appeared to show some improvement over time, thus indicating that they were learning how to perform the task more accurately. In contrast to these similar human and machine performance results for vowels, we will see in the next section that the differences in performance for stop consonants are much larger.

Stop Classification by Humans and Machines

Nossair and Zahorian [69] have compared human and machine performance in the task of classification of syllable-initial stops in *isolated* CVC syllables. They collected a 30-speaker database containing isolated CVC syllables. Each of the 6 stops appears with each of eleven vowels for a total of 66 relevant tokens per speaker. With these highly constrained conditions, they obtained machine error rates that were approximately 2 times worse than human performance, although this result is dependent upon which conditions in the perceptual experiments are considered comparable to the machine experiments.

In [41], Lamel reported results from perceptual experiments on stop consonants

Phonetic Context	Human Error (%)	Machine Error (%)	Error Rate Increase Factor
V–S–V	3.4	11.3	3.3
V–F–S–V	12.2	32.9	2.7
V–N–S–V	7.6	18.7	2.5

Table 1.2: Human vs machine stop classification, after [31]: Comparing the baseline system from Chapter 3 with Lamel's perceptual results[41]

extracted from TIMIT. Machine comparisons with Lamel's results were reported by this author in [31]. The comparisons are summarized in this section. Lamel's perceptual results are broken down according to phonetic context. We will consider three phonetic contexts, namely syllable-initial stops in a vowel-stop-vowel sequence, vowel-fricative-stop-vowel sequences, and non-syllable-initial stops in homorganic nasal clusters. Only short speech segments of three phones (V–S–V) or four phones (V–F–S–V and V–N–S–V) were presented to the listener. These sequences were sometimes across word boundaries, so listeners could not use lexical, syntactic, or semantic information. We obtained a list of the testing tokens so that we could use the same set. We trained our classifier on speakers that were not in the test set under consideration to ensure that the system remains speaker-independent. The machine classifier is the baseline system which will be presented in Chapter 3. The system was trained on stops from all contexts to provide a context-independent result.

Table 1.2 summarizes the results. For syllable-initial singleton stops followed by a vowel, Lamel reports that human listeners achieved 3.4% error. The machine classifier performed more than three times worse, obtaining 11.3% error. For vowel-fricative-stop-vowel sequences, human listeners obtained 12.2% error, while the machine classification performed more than two and a half times worse, obtaining 32.9% error. For non-syllable-initial stops in homorganic nasal clusters, human listeners obtained 7.6% error on TIMIT tokens, while the machine obtained 18.7% error, which is two and a half times worse than the humans.

These results indicated that, for stop consonants, there is a significant amount of low-level acoustic phonetic information which the automatic classifier is not effectively extracting. This experimental outcome is consistent with results in the literature comparing human and machine performance in a variety of speech recognition tasks [54]. These results motivated our attempts to extract more low-level acoustic-phonetic information from the speech signal through the use of heterogeneous measurements.

1.3.2 Time-Frequency Resolution Considerations

Analysis of the performance of classification systems shows that within-phone-class performance is dependent upon the time-frequency resolution of the Fourier analysis. Furthermore, the optimal settings for individual phone classes differ substantially [31]. Figure 1-1 shows stop and nasal classification error versus the Hamming window duration of the short-time Fourier analysis. Stop classification is optimized with increased time resolution, while the nasal classification performance favors increased frequency resolution. This result indicates that using acoustic measurements with a single fixed time-frequency resolution is suboptimal, and suggests that the use of heterogeneous measurements may lead to classification improvements.

1.3.3 Typical Classifier Confusions

Figure 1-2 is a bubble plot of a confusion matrix from a TIMIT phonetic classification experiment. Nearly 80% of the confusions occur by choosing an alternative within the same manner class.[3] Another 7% occur due to confusions involving the closure/silence class. This illustrates that phonetic classification can be broken down into the subproblems of vowel classification, nasal classification, etc. Performance improvements within these subtasks should result in overall performance improvements, since there is only a small amount of "leakage" error which occurs between different subclasses. In general, the lower the error between classes, the greater confidence we can have that subtask improvement will lead to overall improvement. Thus, one might consider an easier breakdown into classes, such as sonorants, obstruents, and

[3]See Chapter 2 for a listing of the phonetic labels with their manner class membership.

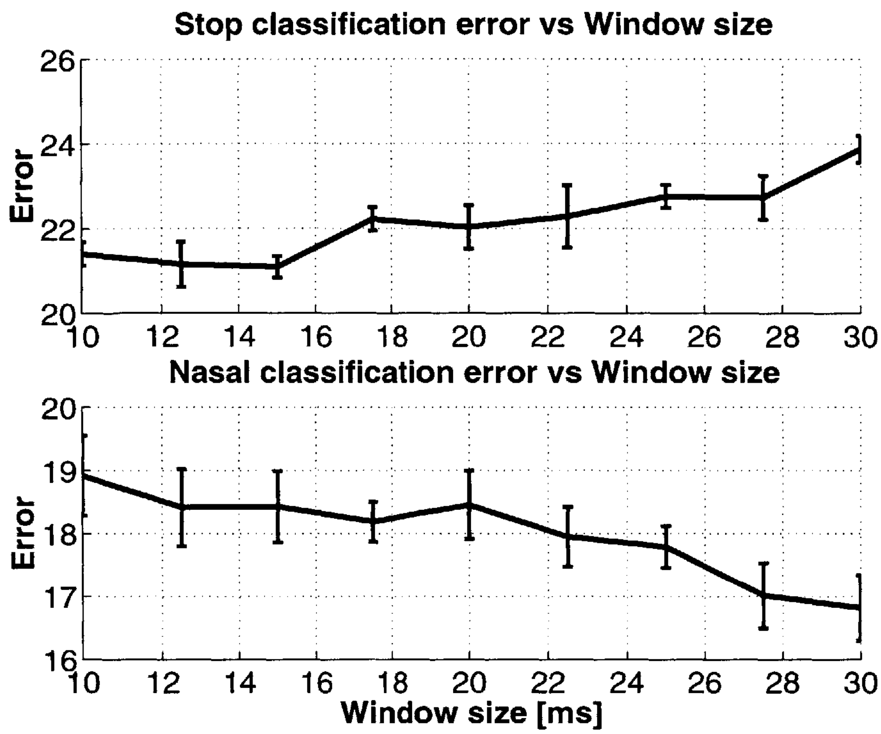

Figure 1-1: Comparison of within-class stop and nasal classification as a function of Hamming window duration. The vertical bars show one standard deviation in the performance calculated over five trials of mixture model training.

silence.

The hierarchical methods presented in this thesis directly pursue the approach of improving classification within phonetic manner classes. The committee-based methods do not do this directly.

1.4 Thesis Structure

This thesis examines new techniques in acoustic-phonetic modeling of speech. Chapter 2 provides background information for the experimental work. Chapter 3 presents hierarchical approaches. Chapter 4 presents acoustic measurements designed for use in committee-based systems, and presents several committee-based and hybrid techniques for combining multiple classifiers. Chapter 5 presents experiments evaluating the various classifier combination techniques in the task of TIMIT phonetic classi-

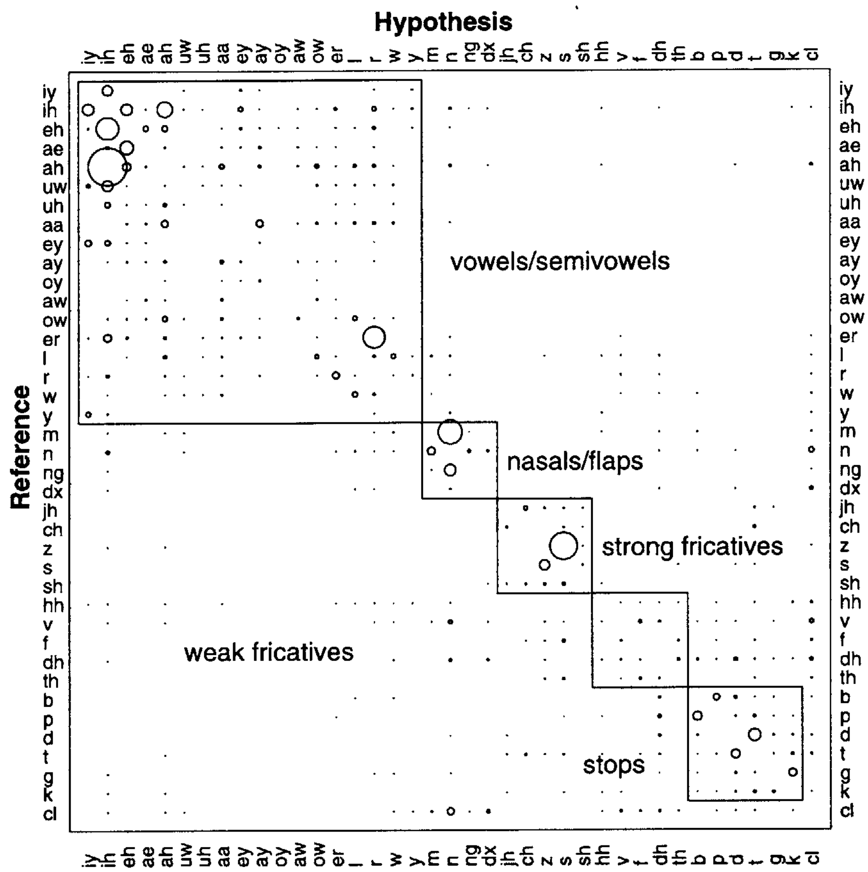

Figure 1-2: TIMIT phonetic classification confusions, with radii linearly proportional to the error. The largest bubble is 5.2% of the total error.

fication. In addition, this chapter contains TIMIT phonetic recognition results and word recognition results on utterances spoken to MIT's JUPITER weather information system [101]. Chapter 6 contains the design and results from perceptual experiments with stop consonants, comparisons with machine classifiers, and an analysis errors made by machines. Chapter 7 summarizes the thesis and discusses contributions and ideas for future work.

Chapter 2

Experimental Background

This chapter contains background information for the experimental work presented in this thesis. This includes information about the TIMIT and JUPITER databases, the SUMMIT speech recognition system, and the training of acoustic models.

2.1 The TIMIT Database

The TIMIT acoustic-phonetic continuous speech corpus [40] was recorded at Texas Instruments (TI), transcribed at the Massachusetts Institute of Technology (MIT), and verified and prepared for CD-ROM production by the National Institute of Standards and Technology (NIST). It contains speech from 630 speakers representing 8 major dialect divisions of American English, each speaking 10 phonetically-rich sentences. There are 438 male speakers and 192 female speakers. The corpus includes the speech waveform files with corresponding time-aligned orthographic and phonetic transcriptions.

2.1.1 TIMIT Phones and Phone Classes

Table 2.1 shows the IPA and ARPAbet symbols for the 61 phones in the TIMIT corpus. The ARPAbet symbols will be used throughout the thesis. In accordance with common practice [43], we collapsed the 61 TIMIT labels into 39 labels before

IPA	ARPAbet	Example	IPA	ARPAbet	Example
[ɑ]	aa	b*o*b	[ɨ]	ix	deb*i*t
[æ]	ae	b*a*t	[i^{y}]	iy	b*ee*t
[ʌ]	ah	b*u*t	[ǰ]	jh	*j*oke
[ɔ]	ao	b*ough*t	[k]	k	*k*ey
[ɑ^{w}]	aw	b*ou*t	[k$^{\square}$]	kcl	k closure
[ə]	ax	*a*bout	[l]	l	*l*ay
[ə^{h}]	ax-h	p*o*tato	[m]	m	*m*o*m*
[ɚ]	axr	butt*er*	[n]	n	*n*oo*n*
[ɑ^{y}]	ay	b*i*te	[ŋ]	ng	si*ng*
[b]	b	*b*ee	[ɾ̃]	nx	wi*nn*er
[b$^{\square}$]	bcl	b closure	[o^{w}]	ow	b*oa*t
[č]	ch	*ch*oke	[o^{y}]	oy	b*oy*
[d]	d	*d*ay	[p]	p	*p*ea
[d$^{\square}$]	dcl	d closure	[□]	pau	pause
[ð]	dh	*th*en	[p$^{\square}$]	pcl	p closure
[ɾ]	dx	mu*dd*y	[ʔ]	q	glottal stop
[ɛ]	eh	b*e*t	[r]	r	*r*ay
[l̩]	el	bott*le*	[s]	s	*s*ea
[m̩]	em	bott*om*	[š]	sh	*sh*e
[n̩]	en	butt*on*	[t]	t	*t*ea
[ŋ̍]	eng	Wash*ing*ton	[t$^{\square}$]	tcl	t closure
[▯]	epi	epenthetic silence	[θ]	th	*th*in
[ɝ]	er	b*ir*d	[ʊ]	uh	b*oo*k
[e^{y}]	ey	b*ai*t	[u^{w}]	uw	b*oo*t
[f]	f	*f*in	[ü]	ux	t*oo*t
[g]	g	*g*ay	[v]	v	*v*an
[g$^{\square}$]	gcl	g closure	[w]	w	*w*ay
[h]	hh	*h*ay	[y]	y	*y*acht
[ɦ]	hv	a*h*ead	[z]	z	*z*one
[ɪ]	ih	b*i*t	[ž]	zh	a*z*ure
-	h#	utterance initial and final silence			

Table 2.1: IPA and ARPAbet symbols for phones in the TIMIT corpus with example occurrences

scoring. The mapping is shown in Table 2.2. In addition, glottal stops were ignored for classification experiments, but included for recognition experiments. We have decided to divide the TIMIT phonetic labels into 6 broad, manner classes: vowels and semivowels (VS), nasals and flaps (NF), strong fricatives (SF), weak fricatives and aspirants (WF), stops (ST), and closures. Alternatively, we have partitioned the phones into three broad classes: sonorants (SON), obstruents (OBS), and silences (SIL). Table 2.3 shows the membership of each of these phonetic classes.

1	iy	20	n en nx
2	ih ix	21	ng eng
3	eh	22	v
4	ae	23	f
5	ax ah ax-h	24	dh
6	uw ux	25	th
7	uh	26	z
8	ao aa	27	s
9	ey	28	zh sh
10	ay	29	jh
11	oy	30	ch
12	aw	31	b
13	ow	32	p
14	er axr	33	d
15	l el	34	dx
16	r	35	t
17	w	36	g
18	y	37	k
19	m em	38	hh hv
39	bcl pcl dcl tcl gcl kcl q epi pau h# not		

Table 2.2: Mapping from 61 classes to 39 classes for scoring of results, after [43].

2.1.2 TIMIT Data Sets

The sentences in the TIMIT corpus are divided into three types: dialect (SA), phonetically-compact (SX), and phonetically-diverse (SI). The dialect sentences were designed to reveal the dialectical variation of the speakers, and were read by all 630

Phone Class	# TIMIT labels	TIMIT labels
Vowel/Semivowel (VS)	25	aa ae ah ao aw ax axh axr ay eh er ey ih ix iy ow oy uh uw ux el l r w y
Nasal/Flap (NF)	8	em en eng m n ng nx dx
Strong Fricative (SF)	6	s z sh zh ch jh
Weak Fricative (WF)	6	v f dh th hh hv
Stop (ST)	6	b d g p t k
Closure (CL)	9	bcl dcl gcl pcl tcl kcl epi pau h#
Sonorant (SON)	33	Vowel/Semivowel + Nasal/Flap
Obstruent (OBS)	18	Strong Fric + Weak Fric + Stop
Silence (SIL)	9	Same as Closure

Table 2.3: Phonetic subsets which will be used in subsequent esperiments.

speakers. The two dialect sentences were "She had your dark suit in greasy wash water all year." and "Don't ask me to carry an oily rag like that." The phonetically-compact (SX) sentences were hand-designed to be phonetically comprehensive as well as compact, in the sense of brevity. The phonetically diverse (SI) sentences were selected from existing text sources. Table 2.4 indicates the number of unique sentence orthographies of each type, the number of speakers per unique sentence orthography, and the number of sentences of each type spoken by each speaker.

Sentence Type	# Sentences	# Speakers/ Sentence	Total	# Sentences/ Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342	-	6300	10

Table 2.4: TIMIT speech material according to sentence type.

The core test set was selected to include 2 males and 1 female from each of the 8 dialect regions. Table 2.5 shows the 24 speakers in the core test set, along with their dialect region. There are 8 texts for each speaker (dialect sentences were excluded), for a total of 192 utterances in the core test set.

Dialect	Speakers
New England	mdab0 mwbt0 felc0
Northern	mtas1 mwew0 fpas0
North Midland	mjmp0 mlnt0 fpkt0
South Midland	mlll0 mtls0 fjlm0
Southern	mbpm0 mklt0 fnlp0
New York City	mcmj0 mjdh0 fmgd0
Western	mgrt0 mnjm0 fdhc0
Army Brat (moved around)	mjln0 mpam0 fmld0

Table 2.5: 24 speakers in the TIMIT core test set, with their dialect region.

1	faks0	11	fdac1	21	fjem0	31	mgwt0	41	mjar0
2	mmdb1	12	mmdm2	22	mpdf0	32	fcmh0	42	fkms0
3	mbdg0	13	mbwm0	23	mcsh0	33	fadg0	43	fdms0
4	fedw0	14	mgjf0	24	mglb0	34	mrtk0	44	mtaa0
5	mtdt0	15	mthc0	25	mwjg0	35	fnmr0	45	frew0
6	fsem0	16	mbns0	26	mmjr0	36	mdls0	46	mdlf0
7	mdvc0	17	mers0	27	fmah0	37	fdrw0	47	mrcs0
8	mrjm4	18	fcal1	28	mmwh0	38	fjsj0	48	majc0
9	mjsw0	19	mreb0	29	fgjd0	39	fjmg0	49	mroa0
10	mteb0	20	mjfc0	30	mrjr0	40	fmml0	50	mrws1

Table 2.6: 50 speakers in the TIMIT development set.

The NIST "complete" test set was formed by including all 7 repetitions of the SX texts in the core test set. This procedure resulted in adding another 144 speakers to the core set, for a total of 168 speakers in the complete test set. This set was not used in this thesis. The reason for this is that we made use of a development set which overlaps with this definition of the "complete" test set.

The NIST training set consists of the 462 speakers which are not included in either the "core" or "complete" test sets. With the exception of the dialect (SA) sentences, which are excluded from classification and recognition experiments, there is no overlap between the texts read by the training and testing speakers.

We made extensive use of a 50-speaker development set. The core set was reserved

Set	# Speakers	# Utterances	# Hours
Train	462	3,696	3.14
Development	50	400	0.34
Core Test	24	192	0.16
"Full" Test	118	944	0.81

Table 2.7: Number of speakers, utterances, and hours of speech in the TIMIT training, development, core test, and "full" test sets.

Phones	462-speaker Train	50-speaker Development	24-speaker Core	118-speaker "full" Test
VS	58,840	6,522	3,096	15,387
NF	14,176	1,502	731	3,566
SF	13,157	1,326	661	3,169
WF	8,990	1,014	467	2,323
ST	16,134	1,685	799	4,022
CL	28,928	3,008	1,461	7,230
glottal(q)	2,685	277	118	650
Total	142,910	15,334	7,333	36,347

Table 2.8: Token counts in phonetic subclasses for the TIMIT training, development, core test, and "full" test sets.

for final testing in order to avoid biasing results toward the core set. Thus, experiments for system design and modification were performed using the development set. The speakers in this set are disjoint from both the training set and the core test set. Table 2.6 lists the 50 speakers in the development set. In Chapter 3, we make use of a 118-speaker test set which consists of the "complete" test set, minus our development set. We refer to this 118-speaker set as the "full" test set.

Table 2.7 summarizes the number of speakers, the number of utterances, and the number of hours of speech in each of the sets used in the experiments in this thesis. Table 2.8 indicates the number of tokens in each of the data sets. These totals are helpful for the detailed results in Section 5.1.3, where the exact number of errors is reported along with the percent error.

2.2 The JUPITER Database

Word recognition experiments were performed on the database associated with the JUPITER telephone-based weather information system [101, 29]. JUPITER allows a user to access and receive on-line weather information over the phone. The system has weather information for more than 500 cities, mostly within the United States, but also selected major cities around the world. It can give a weather report for a particular day or several days, and answer specific questions about weather phenomena such as temperature, wind speed, precipitation, pressure, humidity, sunrise times, etc. Table 2.9 shows an example dialogue between a user and JUPITER. Data has been collected from calls made to the system via a toll-free number. Apart from a small amount of read speech collected for initial system development, these utterances are spontaneous speech from self-trained or untrained users.

Table 2.10 indicates the number of utterances and the number of hours of speech in the training and test sets. From the standpoint of the number of hours of speech, the JUPITER corpus is considerably larger than the TIMIT corpus. The recognition vocabulary had 1893 words. A class-bigram language model was used, which has a perplexity of 15.2 on the test set. The test set included only within-vocabulary, spontaneous utterances collected from the operational system.

Experiments using the JUPITER database are included in order to demonstrate that the acoustic-phonetic modeling techniques developed in this thesis generalize to word recognition. In addition, at the same time, these experiments show that these techniques are effective in a telephone-bandwidth acoustic environment.

2.3 The SUMMIT Speech Recognition System

The experiments in this thesis were carried out using the SUMMIT segment-based speech recognizer [27]. One of the potential advantages of a segment-based approach over conventional frame-based methods is that it provides more flexibility in choosing *what* acoustic attributes to extract, and *where* to extract them from the speech signal.

User: Hello, what is the weather in Philadelphia today? JUPITER: "In Philadelphia today: partly sunny and hot with 30 percent chance of thunderstorms afternoon. high mid 90s center city. tonight: partly cloudy with 30 percent chance of thunderstorms evening. lows mid or upper 70s center city. Is there something else?"
User: What about the hurricane? JUPITER: "Hurricane advisories have been issued for the following places today: South Carolina, and Georgia. Please choose one if you'd like to know more."
User: How about South Carolina? JUPITER: "Hurricane advisories have been issued for the following places in South Carolina today: Charleston, and Myrtle Beach. Please choose one if you'd like to know more."
User: What what is the wind speed in Atlantic City New Jersey? JUPITER: "In Atlantic City today: wind southwest 15 miles per hour. tonight: wind southwest 10 miles per hour. What else?"
User: What is it like in San Francisco? JUPITER: "In San Francisco tonight: coastal low clouds and fog spreading over the city through the night. lows 50s. today: low clouds and patchy fog most areas. high 60s the coast and low 70s in the warmest areas. tonight. low clouds spreading in the evening over the city. lows 50s. Can I help you with something else?"
User: What about Chile? JUPITER: "I know of only one city in Chile: Santiago. today: partly cloudy. high 59 and low 40. Is there something else I can do for you?"

Table 2.9: Example of user interaction with the JUPITER weather information system.

It is believed that this flexibility enhances the system's ability to take full advantage of the acoustic-phonetic information encoded in the speech signal.

The essential element of the segment-based approach is the use of explicit segmental start and end times in the extraction of acoustic measurements from the speech signal. In order to implement this measurement extraction strategy, segmentation hypotheses are needed. The SUMMIT system has made use of a variety of segmentation algorithms [28, 7, 46] for producing segmentation hypotheses. In one algorithm [28, 46], major segment boundaries are hypothesized when a measure of spectral change exceeds a global threshold. Minor segment boundaries are hypothesized between the major boundaries based again upon spectral change, but this time

Data Set	# Utterances	# Hours
Train	17,807	16.2
Test	1,806	1.6

Table 2.10: Number of utterances and number of hours of speech in the training and test sets used for the JUPITER word recognition experiments.

utilizing a local threshold which depends on the analysis of the signal between the major boundaries. All segment boundaries between major boundaries are fully interconnected to form a network of possible segmentations. The size, or depth, of the network is determined by the thresholds. This segmentation algorithm may be referred to as "acoustic segmentation," since it depends on acoustic distance metrics. The JUPITER word recognition experiments presented in this thesis make use of the acoustic segmentation. In more recent work [7, 46] a technique called "probabilistic segmentation" has been developed which produces a segment graph as a result of frame-based recognition processing. This recognition processing does not necessarily need to be full word recognition, but might instead be phonetic recognition, or even further backed-off to phonetic broad-class recognition. In general, it has been found that probabilistic segmentation techniques produce much more accurate segment networks than acoustic segmentation techniques. The TIMIT phonetic recognition experiments in this thesis all make use of probabilistic segmentation.

Frame-based measurements from the speech signal give rise to a *sequence* of observations. Since there is no overlap in the observations, every path through the network accounts for all observations. In contrast, segment-based measurements from a hypothesized segment network lead to a *network* of observations. For every path through the network, some segments are on the path and others are off the path. To maintain probabilistic integrity when comparing different paths it is necessary for the scoring computation to account for all observations by including both on-path and off-path segments in the calculation. In [29], all off-path segments are accounted for by the use of a single antiphone model. In [7], off-path segments are accounted for by the use of near-miss models.[1] The TIMIT phonetic recognition experiments presented in this thesis make use of both antiphone and near-miss modeling techniques.

2.4 Training of Acoustic Models

Throughout this thesis, normalization and principal components analysis were performed on the acoustic measurement vectors in order to whiten the space prior to modeling. The whitened measurement vectors were modeled using mixture distributions composed of multivariate Gaussian probability density functions. The general form of the n-dimensional Gaussian probability density function (pdf) is:

$$\mathcal{N}(\vec{f};\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{f}-\vec{\mu})^T\Sigma^{-1}(\vec{f}-\vec{\mu})\right\} \tag{2.1}$$

where $\vec{\mu}$ is the n-dimensional mean of the distribution, and Σ is the $n \times n$ covariance matrix. A mixture Gaussian pdf containing M components is given by

$$p(\vec{f}) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\vec{f};\mu_m,\Sigma_m) \tag{2.2}$$

where the mixture weights, ω_m, satisfy

$$\sum_{m=1}^{M} \omega_m = 1 \tag{2.3}$$

$$0 \leq \omega_m \leq 1. \tag{2.4}$$

The use of M Gaussian kernels in a mixture Gaussian pdf may be referred to as "M components," "M mixture components," or even "M mixtures," although the final form is not preferable since it can be ambiguous in some contexts.

In most of the experiments reported in this thesis, the covariance matrix was restricted to be diagonal. In comparison with full covariance Gaussian kernels, the use of diagonal covariance Gaussian kernels enables the use of more mixture components

[1]Appendix B explains antiphone modeling and near-miss modeling in greater detail.

because there are many fewer parameters to train per component. In addition, the computations required, per component, for training and testing become simpler and require fewer operations.

Gaussian mixture models were trained by a two-step process. In the first step, the K-means algorithm [21] was used to produce an initial clustering of the data. In the second step, the results of the K-means algorithm were used to initialize the Estimate-Maximize (EM) algorithm [19, 21] which iteratively maximizes the likelihood of the training data and estimates the parameters of the mixture distribution. The EM algorithm converges to a local maximum. There is no guarantee of achieving the global optimum. The outcome of the EM algorithm is highly dependent on the initial conditions obtained from the K-means algorithm. In order to improve the performance and robustness of the mixture models, we used a technique called aggregation, which is described in Appendix C.

2.4.1 Acoustic Modeling Parameters

The acoustic modeling parameters used for the various experiments in this thesis are summarized in Table 2.11. The modeling parameters were standardized for the committee-based and hybrid experiments in Chapters 4 and 5. The aggregation of 4 models at a time was a design decision based on the tradeoff between the increased performance versus greater computation that comes with aggregation. In general, aggregation produces diminishing returns as more and more models are aggregated. Thus, the aggregation of a very large number of acoustic models as in Chapters 3 and 6 is not usually advisable. Those large models were a result of experiments testing the limits of how aggregation affects performance. The particular number of models aggregated varies because it was chosen to optimize performance on the development set.

Experiment/ Measurements	Full or Diagonal Covariance	Maximum # Mixture Components per Phonetic Unit	Minimum # Data Vectors per Mixture Component	# of Models Aggre-gated
Chapter 3:				
SB	Full	12	500	20
SV	Full	12	500	27
SN	Full	12	500	23
SS	Full	12	300	20
SF	Full	12	500	20
VTLN:SB	Diagonal	96	61	8
Chapters 4 & 5:				
TIMIT:				
S1-S8,SVa,SN,SS,SF	Diagonal	96	61	4
B1-B5	Diagonal	100	10	4
JUPITER:				
B1,B3a,B4	Diagonal	50	50	1
Chapter 6:				
System A	Full	12	500	20
System B	Full	12	300	20
System C	Full	12	500	10
Heterogeneous	As in Chapters 4 & 5			

Table 2.11: Table of the acoustic modeling parameters used for the experiments in this thesis, except where indicated otherwise in Section 5.1.4.

Chapter 3

Hierarchical Methods: Measurements, Classifier Structures, and Experiments

Hierarchical approaches to the problems of measurement selection and classifier combination are presented in this chapter. Hierarchical approaches emphasize the notion that the overall phonetic classification task can be subdivided into smaller tasks. Figure 1-2, as discussed previously, shows that most automatic classification errors occur within the correct phonetic manner class. If the division into subtasks can be performed reliably, then it follows that a hierarchical scheme can be used to take advantage of classifiers which are designed for individual subtasks. This is the strategy which is explored in this chapter. Related work and the design of class-specific acoustic measurements are discussed first, followed by a presentation of hierarchical approaches for the combination of classifiers. Class-specific vocal tract length normalization is presented as a variation on the idea of using different measurements for different phonetic classes. All of these techniques are evaluated on the TIMIT phonetic classification task.

3.1 Related Work

There have been a number of attempts to solve the phonetic classification problem by a hierarchical approach. Some early work in this direction addressed the problem of isolated-word, alphanumeric classification. This task has only a 36-word vocabulary (the English letters A through Z and the digits zero through nine), yet it is difficult because all but three of the items are monosyllabic, and it contains several highly confusable subsets [42]. The following six confusable subsets have been identified [79]: (a) {B C D E G P T V Z 3}, (b) { A J K 8 H}, (c) {L M N}, (d) {F S X 6}, (e) {I Y 5}, and (f) {Q U 2}. Nearly all errors occur within these sets. The first of these confusable sets is known as the "E set." This is analogous to the existence of confusable sets pointed out in Figure 1-2 for the overall phonetic classification problem. In both cases, the observation suggests a hierarchical approach.

In a paper by Cole, Stern, and Lasry (1986) [17], a hierarchical system called FEATURE is described which addresses the problem of classifying letters and digits. They make reference to earlier work by Rabiner and Wilpon (1981) [80] which used a two-pass approach for the same problem. The first pass assigned the utterance to one of a group of pre-defined confusable sets. The second pass attempted to provide optimal separation among members in each class. Recognition improvements of 3% to 7% were obtained. Bradshaw, Cole, and Li [4] (1982) used a similar procedure on the "E set," and obtained a decrease in recognition error from 37% to 16% on that set. Returning again to the FEATURE system, it made use of Bayesian decisions at each node of a hierarchical decision tree. The decisions at each node were based upon a subset of 50 acoustic measurements. Each node could use a different subset based upon which were considered most beneficial for discrimination among the letters at that node. The overall set of 50 acoustic measurements were developed through the incorporation of knowledge from spectrogram reading. Their final results on an 8-speaker database resulted in a speaker-independent error rate of 11% on the task of classifying alphabetic letters (without the numbers). The system achieved 16.7% error on the confusable "E set." Regardless of whether one deemed the FEATURE system to be a success or not, there are several observations that are relevant for the work in this thesis. First, the use of a Bayesian hierarchical decision tree classifier is computational equivalent to the MAP hierarchical approach presented in this thesis, although the theoretical justification is not identical. Second, the FEATURE system did not address the full phonetic classification problem. Third, in the years following 1986, hierarchical approaches were not widely used in the speech recognition community.

More recently, Chun [13] (1996) explored hierarchical approaches for the full phonetic classification problem. Although Chun's systems produced small gains of less than half a percent in absolute performance, none of the changes were reported to be statistically significant. For vowels, the baseline MFCC-based acoustic measurements were first replaced by a measurement vector consisting of frequencies, amplitudes, and derivatives of formants, log duration, and fundamental frequency. This measurement vector performed worse than the baseline. This is not surprising, since formant measurements are known to be subject to greater risk of severe measurement errors, such as mistaking the third formant for the second formant. Chun's system did not have a mechanism for handling these types of formant measurement errors. As an alternative, the MFCC-based feature vector was augmented with fundamental frequency. Measurement vectors for improving the classification of the [s z],[sh zh], and [ch jh] pairs were also explored, again without producing any statistically significant improvements.

In [96], Zahorian (1997) reports TIMIT phonetic classification results using a single large neural network (LNN) versus using binary-pair partitioned neural networks (BPP). The phones were scored using 39 phone classes, so there were "39 choose 2" pairs of phones, which is 741. One small network was trained for each of these 741 phone pairs. The LNN took 8 times longer to train than the BPP system. The BPP structure made it possible to use different acoustic measurements for each of the pairwise phonetic distinctions. Two experiments were performed. In the first experiment, the LNN and BPP systems used the same set of 60 measurements. The LNN obtained 23.0% error on the core set, which was slightly better than 23.5% error for the BPP. In the second experiment, 35 measurements were selected for each phone pair, out of a pool of 96 measurements. The measurements were selected in an attempt to optimally discriminate between each pair of phones. The performance with the "best 35" measurements was 24.3% error, which is only a small drop in performance considering the drop in dimensionality from 60 to 35. Some other experiments were also reported, but the use of phone-pair specific measurements never surpassed the 23.0% error obtained using a single set of 60 measurements.

In summary, interest in hierarchical approaches for phonetic classification was strong in the late 1970's and early 1980's. However, papers from that time generally did not address the complete phonetic classification task. In addition, hierarchical approaches have been relatively uncommon since that time. The recent work by Chun and Zahorian did not produce statistically significant improvements in performance. Thus, at the time of the work performed in this thesis, there was a definite lack of recent experimental studies using hierarchical approaches which demonstrate improvements in performance in the full phonetic classification task.

3.2 Heterogeneous Measurements for Hierarchical Systems

There are at least two potential advantages to the use of hierarchical methods with class-specific measurements. First, the use of hierarchical measurements provides the opportunity to develop diverse measurements which focus on the phonetically relevant information for discriminating among sounds in a particular phone class. For example, stop measurements could focus on the burst. Vowel measurements could either encode the gradual temporal evolution of the formants or attempt to capture vowel landmarks [86, 85], i.e., target formant locations. Second, class-specific measurement design permits the removal of dimensions that are unnecessary in a particular sound class, thus making better use of the training data and reducing the computational requirements of the system. For example, in the experiments in this chapter we find 50-dimensional stop consonant measurements which perform better than the baseline 61-dimensional measurements.

The measurements used in this section are calculated using a two step process. The first stage is referred to as frame-based processing. In this stage, a spectral representation of the speech signal is calculated at regular intervals (usually 5 ms in this work). The second stage is segment-level processing. In segment-level processing, a frame-based spectral representation and knowledge of hypothesized segment boundaries are combined to produce segmental measurements. We will examine these two stages separately.

3.2.1 Frame-based Spectral Representations

Short-time spectral representations of the speech signal have been widely used for speech recognition. In this thesis we make use of two well-known representations, Mel-frequency cepstral coefficients (MFCCs) [61] and perceptual linear prediction cepstral coefficients (PLPCCs) [34], as well as a third MFCC-like representation described in [96, 90, 68] which we refer to as "Discrete Cosine Transform Coefficients" (DCTCs). In this chapter, we will restrict ourselves to MFCCs and DCTCs.

The steps used to calculate the MFCCs can be summarized as follows:

1. Remove the DC component and normalize the magnitude of the waveform.

2. Preemphasize the waveform with the finite impulse response filter

$$y[n] = x[n] - 0.97x[n-1].$$

3. Calculate the magnitude squared of the STFT with a frame interval of 5 ms and a Hamming window of length 10–30 ms.

4. Compute the Mel-Frequency Spectral Coefficients (MFSCs) using inner products with triangular basis functions, illustrated in Figure 3-1. These MFSCs are designed to incorporate a Mel-warping of the frequency axis [77]

$$f' = 2595 \log_{10}\left(1 + \frac{f}{700}\right).$$

5. Compute the logarithm of the MFSCs, $10\log_{10}(\cdot)$.

6. Take the cosine transform of the log MFSCs to obtain the Mel-frequency cepstral coefficients (MFCCs). Let $s[k]$ be the log MFSCs for $k = 0, 1, \ldots, K-1$, then the MFCCs $c[k]$ are given by

$$c[m] = \sum_{k=0}^{K-1} \cos\left(\frac{\pi m(k - \frac{1}{2})}{K}\right) s[k].$$

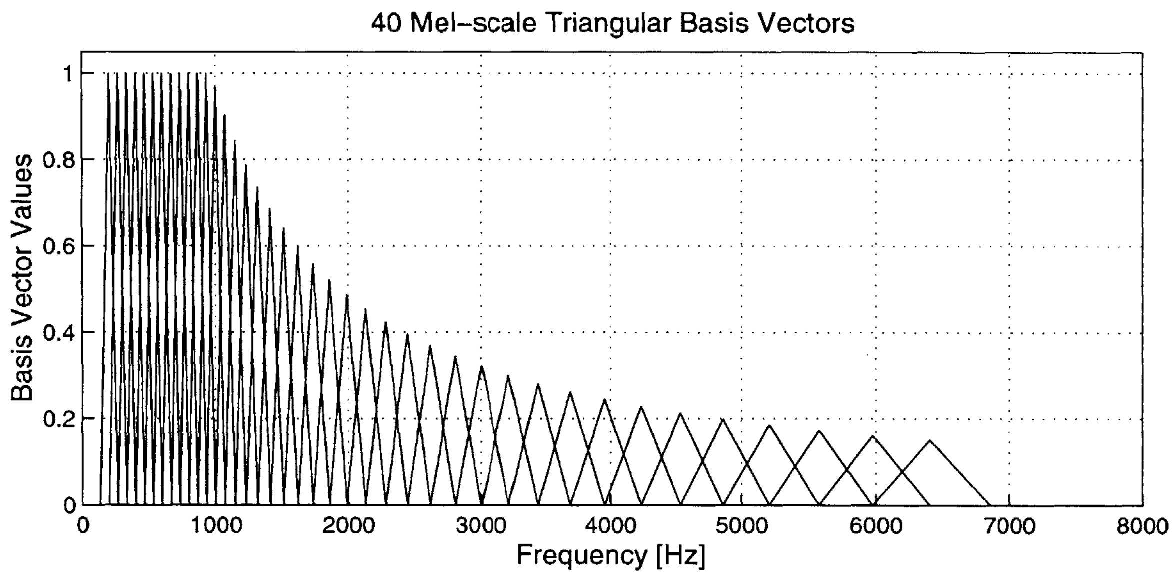

Figure 3-1: Illustration of 40 "Triangular" Mel-scale basis vectors which were used for the calculation of MFSCs from the magnitude of the STFT.

The DCTCs are similar to the MFCCs, in that each step of processing has an analogous step in the MFCC calculation, except for the morphological dilation in the fourth step. The steps used to calculate the DCTCs [96] can be summarized as follows:

1. Remove DC component and normalize the magnitude of the waveform.

2. Preemphasize the waveform with the infinite impulse response filter

$$y[n] = x[n] - 0.95x[n-1] + 0.494y[n-1] - 0.64y[n-2].$$

3. Take the magnitude squared of the STFT at a 3ms frame rate and a 10 ms Hamming window.

4. Apply a morphological dilation operation [68] to the output of the magnitude of the STFT. This has the effect of emphasizing local spectral peaks and eliminating local minima. To define this operation, let f be a continuous independent variable, and let B represent the width over which the dilation is performed. Given a function $s[f]$, the dilated function $s_d[f]$ is defined for each f_o as

$$s_d[f_o] = \max_{|f-f_o|<B} s[f].$$

For the discrete case, we perform a discrete maximization analogously. The dilation width is 80 Hz.

5. Compute the logarithm of the dilated STFT values, $10\log_{10}(\cdot)$.

6. Apply a bilinear-warped cosine transform in the frequency dimension. This is analogous to combining the MFSC and MFCC transforms into a single step. Figure 3-2 shows the shape of the first five of these basis vectors, using a bilinear warp factor of 0.45.

The calculation of the PLPCCs is described in detail in [34]. This technique incorporates three perceptual effects: the critical-band spectral resolution, the equal-loudness curve, and the intensity-loudness curve. PLPCCs use a Bark-scale warping of the frequency axis [83]

$$f' = 6 \ \ln\left\{\frac{f}{600} + \left[\left(\frac{f}{600}\right)^2 + 1\right]^{0.5}\right\}.$$

The purpose of this warping is analogous to that of the Mel-scale warping used for MFCCs. PLPCC analysis makes use of linear prediction analysis to produce an all-pole model of the spectrum, and is therefore model-based. Finally, the linear prediction coefficients are converted to cepstral coefficients corresponding to the cepstrum of the impulse response of the all-pole model.

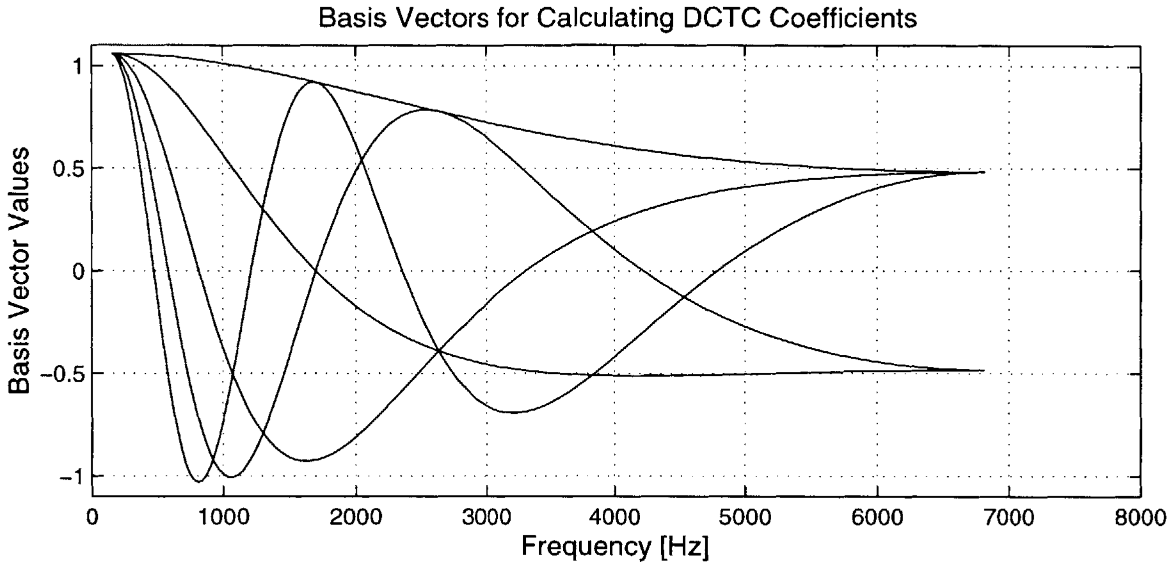

Figure 3-2: Illustration of basis vectors used to calculate the DCTCs, after [96].

3.2.2 Vocal Tract Length Normalization (VTLN)

Vocal tract length normalization (VTLN) has proved to be a useful method to account for across-speaker variations in the speech signal [23, 44, 45, 91, 98, 99]. VTLN is an efficient form of speaker adaptation, in that it only requires one parameter per speaker. In this work, we seek to extend the VTLN concept within a hierarchical framework by allowing a small number of phone-class-specific parameters per speaker. At first, it may seem counterintuitive to allow two or more VTLN warping parameters, since clearly each speaker has only one vocal tract. However, this strategy is reasonable because the relevant portion of the vocal tract can be different for different sounds. For example, when producing an [s], the length of the cavity in front of the articulatory constriction is crucial, but the length of the cavity behind the constriction is less relevant. Furthermore, empirical results show that iterative optimization of the VTLN warp parameters produces different results for different sound classes.

There are a variety of ways to implement VTLN. In each case, the fundamental idea is to linearly warp the frequency axis of the speech spectral data. In these experiments, the warping is implemented through interpolation of the short-time

Fourier transform (STFT) representation for the signal. Let f represent the original frequency axis, normalized such that $f = 1$ corresponds to π in the discrete-time STFT. Let f' correspond to the warped frequency axis. Let k be the warp factor, with $k = 1.0$ corresponding to no warping. Warp factors less than one correspond to frequency compression, and are frequently chosen for female speakers. Male speakers usually have warp factors greater than one, which corresponds to frequency expansion. The mapping $f \rightarrow f'$ is piecewise linear, given by

$$f' = \begin{cases} kf & \text{if } 0 < f < \ell \\ hf + (1-h) & \text{if } \ell < f < 1 \end{cases}$$

where $\ell = 0.875$ and the slope h is given by $h = (k(1-\ell))/(k-\ell)$. The use of two linear segments was suggested in [91]. Figure 3-3 illustrates this frequency warping for several values of k.

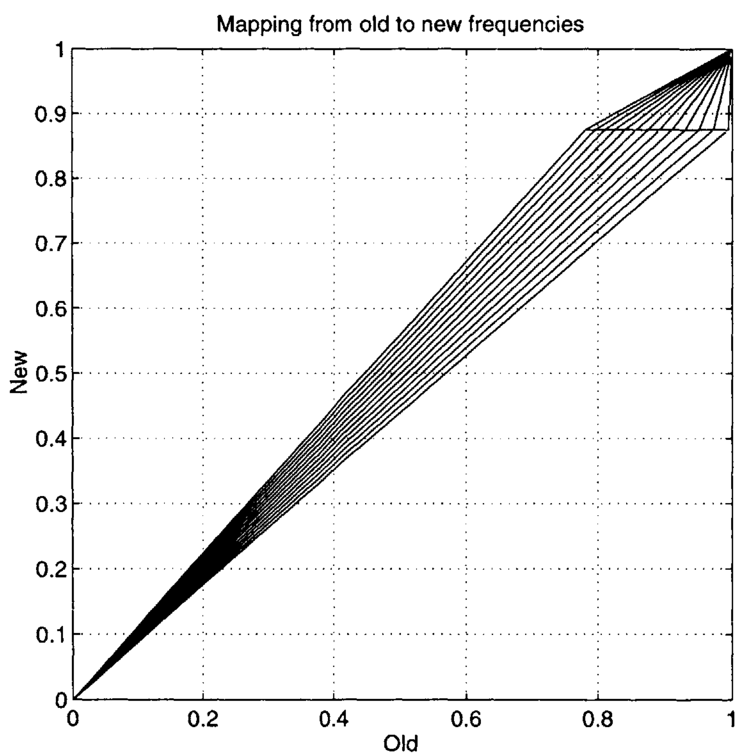

Figure 3-3: Illustration of frequency warping with warp factor ranging from 0.88 to 1.12, and l=0.875 determining boundary between the two linear segments.

In the implementation, we want to produce the values of the STFT at

$$f' = \frac{2n}{N}, \quad n = 0, 1, \cdots, \frac{N}{2}.$$

For each such f', we use the inverse mapping $f' \rightarrow f$ to obtain f. Generally, the DFT will not be calculated at exactly the desired frequencies. To obtain the desired value, linear interpolation is performed based on the two neighboring values in the calculated DFT. The bin resolution of the DFT should be large enough to achieve good interpolation results. The bin resolution can be increased by zero-padding the short-time section and taking a longer DFT. Note that increased bin resolution does not increase the frequency resolution of the STFT. The frequency resolution is given by the length and the shape of the data window. In these experiments there were 1024 frequency bins for normalized frequency in the range $0 \leq f \leq 1$ to use in the interpolation.

3.2.3 Segmental Measurements for Baseline System

A baseline classifier was first established using homogeneous measurements. This measurement set is based on previously reported phonetic classification work [50, 13]. A 61-dimensional homogeneous measurement vector was calculated for each phonetically labeled segment in the TIMIT transcriptions. These segmental measurements were calculated from the frame-based MFCC measurements calculated with a 5 ms frame rate and a 20.5 ms Hamming window. The measurement vector consisted of three MFCC averages computed approximately over segment thirds (actually in a 3–4–3 proportion), two MFCC derivatives computed over a time window of 40 ms centered at the segment beginning and end, and log duration. The derivatives of the MFCC tracks were calculated using linear least-squared error regression. Let $c[n, k]$ represent the MFCCs, with k indexing the cepstral number in "quefrency" [2], and n indexing the temporal sequence of frames. Let n_i and n_f represent the frame indices for the start and end times of the derivative calculation, and let $N = n_f - n_i + 1$.

Reference Phone Partition	% of Hypotheses in the Incorrect Phone Class
{SON,OBS+SIL}	1.7
{SON,OBS,SIL}	2.3
{VS,NF,ST,SF+WF+CL}	3.4
{VS,NF,SF,WF,ST,CL}	4.1

Table 3.1: Shows the percentage of errors made by the baseline classifier when dividing the development set into each of several phonetic partitions.

The expression for the derivative [10, 78, 66] is:

$$\Delta(k) = \frac{NS_{xy} - S_x S_y}{NS_{xx} - S_x^2}$$

where $S_x = \sum_{n=n_i}^{n=n_f} n, \quad S_y = \sum_{n=n_i}^{n=n_f} c[n,k], \quad S_{xx} = \sum_{n=n_i}^{n=n_f} n^2, \quad S_{xy} = \sum_{n=n_i}^{n=n_f} n\ c[n,k].$

This baseline configuration obtained classification errors of 21.1% and 21.6% on the development and core test sets, respectively. These results will be used for comparison with phone-class-specific measurements described in the next section.

3.2.4 Choosing Phonetic Subsets

Measurement design was carried out in two stages. First, a partitioning of the phones into subsets was chosen. Second, measurements were developed for each of the chosen subsets.

Consider the first stage, where the division of the phones into subsets was chosen. One would like to have many different subsets in order to make the acoustic measurements as specific as possible to particular phonetic confusions. On the other hand, one would like a baseline classifier to be able to reliably sift unknown phonetic tokens into the proper phone class. Placing tokens into the correct class becomes more and more difficult as the number of classes increases. Thus, these conditions are conflicting, and therefore the design will try to achieve a compromise solution. In the experiments, the phonetic classes that were considered were: sonorants (SON),

Reference Phone Class	% of Hypotheses which were Out-of-Class	% of Hypotheses which were Phonetic Errors	% of Phonetic Errors which were Out-of-Class
SON	1.4	26.6	5.2
OBS	3.6	23.7	15.3
SIL	3.2	3.2	100.0
OBS+SIL	2.1	14.9	13.9
VOW	1.4	27.7	5.2
NAS	6.8	21.5	31.6
SFR	2.3	20.7	11.3
WFR	19.9	27.0	74.2
STP	5.8	24.2	23.8
SFR+WFR+SIL	4.0	12.0	33.0
SFR+WFR+ SIL+NAS	2.4	14.1	17.1

Table 3.2: For each reference phone class on the left, the three columns show what percentage of the hypotheses were outside of the given reference phone class, what percentage of the hypotheses were phonetic errors, and what percentage of the phonetic errors were out-of-class.

obstruents (OBS), silents (SIL), vowels/semivowels (VS), nasals/flaps (NF), weak fricatives (WF), strong fricatives (SF), stops (ST), and closures (CL). Table 2.3 shows the membership of these classes. As usual in phonetic classification experiments, the glottal stop is ignored. Given these classes, partitions of the phone space can be notated as {SON,OBS,SIL}, {SON,OBS+SIL}, {VS,NF,SF,WF,ST,CL}, or {VS,NF,ST,SF+WF+CL}. Table 3.1 shows the errors that occur when the baseline classifier divides the development set into each of these partitions. The division into only 2 or 3 classes given by the {SON,OBS+SIL} or {SON,OBS,SIL} partitions achieve only 1.7 or 2.3% error, but these classes are so broad that a further division into smaller subsets was desirable. The {VS,NF,SF,WF,ST,CL} partition was judged to have a sufficiently small number of classes for class-specific measurement development, but the number of errors was too high when the phones were partitioned into these classes. Looking at Table 3.1, 4.1% error seems, on the surface, to be reliable enough partitioning. This classifier has an overall phonetic error rate of 21.1%, shown later in Table 3.4. Since $(21.1 - 4.1)/21.1 = 0.81$, this implies that 81% of the errors occur within the correct manner class. However, Table 3.2 shows the within-class error performance in more detail, where it can be seen that 74% of the weak fricative errors were in the wrong class. This was judged to be far too high a percentage. To alleviate this problem, the weak fricatives, strong fricatives, and closures were combined into a single class. This subset has 33% of the errors out-of-class, which is much improved from 74%. More reduction in the percentage of out-of-class errors could have been achieved by also merging the nasals with the fricatives and closures. Phonetic knowledge tells us that there are very significant differences between nasals and fricatives, and thus it was desirable to keep them separate for the measurement development.

3.2.5 Phone-class-specific Segmental Measurements

The results in Tables 3.1 and 3.2 led to the selection of the {VS,NF,ST,SF+WF+CL} partition of the phones into four subsets for the purposes of measurement development. These symbols represent phonetic manner classes, as defined in Table 2.3. This partition is derived from the six manner classes except that the weak fricatives, strong fricatives, and closures have been combined into a single class. In the following paragraphs we describe phone-class-specific measurements and report within-class classification error on the development set. We compare the performance of these measurements to the baseline and also report the McNemar [26] significance level of the difference.

For vowel/semivowel measurements, we used 62 dimensions. The first 60 dimensions were calculated as in [96]. These involve calculation of MFCC-like frame-based measurements, followed by a cosine transform in the time dimension to encode the trajectories of the frame-based features. The use of a tapered, fixed length (300ms), centered temporal window for the cosine transform results in capturing some contextual information which can be modeled in an unsupervised manner through the mixtures in the Gaussian models. In addition to these 60 measurements, duration and average pitch were also included for a total of 62 measurements. The pitch measurement was calculated using a cepstral-based method. These measurements resulted in

Measurement Set	# Dims	STFT [ms]	Spectral Representation	Temporal Basis
Baseline (SB)	61	20.5	12 MFCCs	3 averages (3-4-3), 2 derivatives
Vowel/Semivowel (SV)	62	10	12 DCTCs	5 tapered cosines 300ms wide
Nasal/Flap (SF)	62	28.5	12 MFCCs	3 averages (3-4-3) 2 derivatives
Stop (SS)	50	10	12 MFCCs	2 averages (halves) 2 derivatives
Fricative/Closure/ Silence (SF)	62	26.5	12 MFCCS for averages 11 MFCCs for derivatives	3 averages + 2 derivatives

Table 3.3: Summary of baseline (SB) and hierarchical segmental measurements. See text for full description, including additional dimensions such as energy, log duration, zero-crossing rate, and low-frequency energy.

a vowel/semi-vowel error of 25.7% on the development set, which improves upon the 26.9% (0.02 significance level) obtained by the baseline system, and is competitive with previously reported results [82].

For nasals, baseline measurements were altered by changing the Hamming window duration to 28.5 ms and adding a measure of average pitch, giving a total of 62 measurements per segment. These nasal-optimized measurements achieved 14.8% on the development set, compared to 16.6% obtained by the baseline system (0.001 significance level).

In our stop classification experiments, we increased the time resolution by using a 10 ms Hamming window, and used a 50 dimensional feature vector, composed of MFCC averages over halves of the segment (24 dimensions), time-derivatives of the MFCC tracks at the segment boundaries and at the start of the previous segment (24 dimensions), a measure of low-frequency energy in the previous segment (1 dimension), and log duration. We found that averaging over halves of the segment instead of thirds did not cause a drop in performance for the stops. Due to smaller dimensionality (50 dimensions), we adjusted the classifier by lowering the minimum number of tokens per Gaussian kernel from 500 to 300. In a six-way stop classification task, these measurements obtained an error of 16.6% on the development set, compared to 20.4% for the baseline (10^{-4} significance level), and compare favorably to previously reported results [90].

For fricatives and closures, a 26.5 ms Hamming window was used for frame-based calculations. Time derivatives of only 11 MFCCs (instead of 12) were extracted at the segment boundaries. Three new measurements were added: the zero-crossing rate, the total energy of the entire segment (which is similar but not the same as the information in the first MFCC coefficient), and a time derivative of the low frequency energy at the beginning of the segment. This 62-dimensional measurement set obtained 8.8% on the development set, compared to 9.1% for the baseline (0.1 significance level).

Table 3.3 summarizes the key points about these measurement sets. In the table, the measurement sets are given labels (SB, SV, etc.) for ease of reference.

3.3 Hierarchical Classifier Combination

3.3.1 MAP Hierarchical Scoring

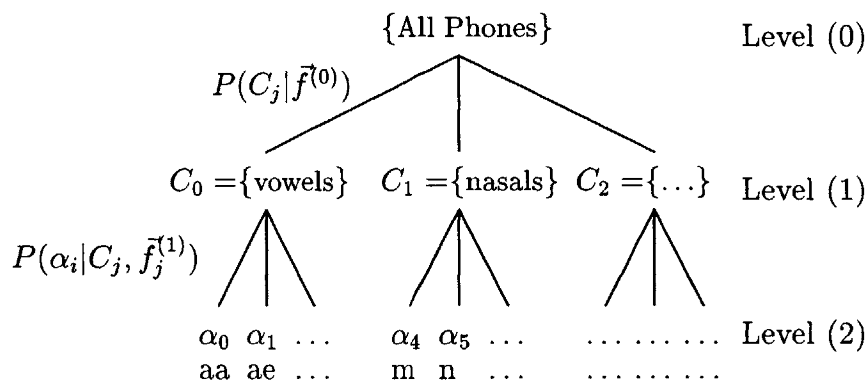

Figure 3-4: Portion of a hierarchical tree for classification.

The second major challenge which must be addressed in order to use heterogeneous measurements is to define a framework for overall classification which makes use of these diverse measurements. The goal of phonetic classification is to determine the most probable phone, α^*, given the acoustic feature vector $\vec{f}$. We can expand the decoding procedure over a set of phone classes C_i according to the expression

$$\alpha^* = \arg\max_{\alpha} P(\alpha|\vec{f}) = \arg\max_{\alpha} \sum_i P(\alpha|C_i, \vec{f}) P(C_i|\vec{f}).$$

If each phone belongs to only one class, then the summation over i becomes trivial since for each phone there is only one i such that $P(\alpha|C_i, \vec{f})$ is nonzero.

In these expressions, $\vec{f}$ represents all of the measurements that might be used by the system. Thus, each set of heterogeneous measurements is a subset of $\vec{f}$. In fact, we can cast the above decoding as a hierarchical process [13]. Thus, at level zero, a single measurement set $\vec{f}^{(0)} \subset \vec{f}$ is used to determine the probability of membership in class j at level one, that is

$$P(C_j^{(1)}|\vec{f}) \approx P(C_j^{(1)}|\vec{f}^{(0)}).$$

In this expression we have decided to approximate $P(C_j^{(1)}|\vec{f})$ by $P(C_j^{(1)}|\vec{f}^{(0)})$ based on practical considerations, such as problems with high classifier dimensionality and superfluous measurement dimensions. These considerations led us to the assumption that each class probability can be more accurately estimated in practice using a subset of the features contained in $\vec{f}$. This assumption does not necessarily hold from a purely theoretical standpoint, where issues stemming from finite training data can be ignored. Continuing at level one, the feature set used to determine the conditional probability of level two class membership can depend upon the conditioning level one class index, j. We indicate the feature set dependence on j using the notation $\vec{f}_j^{(1)}$. Thus the conditional probability of level two class membership is obtained using the approximation

$$P(C_i^{(2)}|C_j^{(1)}, \vec{f}) \approx P(C_i^{(2)}|C_j^{(1)}, \vec{f}_j^{(1)}).$$

Using this notation and the above approximations, our previous decoding equation becomes

$$C_{i^*}^{(2)} = \arg\max_i \sum_j P(C_i^{(2)}|C_j^{(1)}, \vec{f}_j^{(1)})P(C_j^{(1)}|\vec{f}^{(0)}).$$

Figure 3-4 illustrates this decoding computation. This process can be iterated to as many stages as desired. In the present implementation, the level two classes $C_i^{(2)}$ are the individual phones, so no further iteration is required.

3.3.2 Using Binary Decisions

The MAP hierarchical scoring procedure above makes a "soft-decision" at each node of the tree. In contrast, we also implemented a procedure which finds the maximum score at each non-terminal level of the tree and only explores candidates further down the tree from that point. This can be regarded as a variation of the MAP procedure. Referring to Figure 3-4, it implies that

$$P(C_j|\vec{f}^{(0)}) = \begin{cases} 1 & \text{for } j = j_o \\ 0 & \text{for } j \neq j_o \end{cases}$$

where

$$j_o = \arg\max_j P(C_j|\vec{f}^{(0)}).$$

3.4 Experimental Results

In this section, we examine the performance of the measurements and classifier structures presented in this chapter in the task of TIMIT phonetic classification.

3.4.1 TIMIT Phonetic Classification Hierarchical Scoring

Tables 3.4 and 3.5 summarize the performance of the MAP hierarchical scoring technique. This MAP framework for combining heterogeneous measurements achieved 20.0% error on the development set compared to 21.1% for the baseline (10^{-5} significance level), and was also used for final testing on the NIST core set to obtain 21.0%. When compared to the NIST core baseline of 21.6%, the significance level

Task	% Error on dev set		
	Baseline	Heterogeneous	Significance
Overall	21.1	20.0	0.00001
Vowel/Semivowel	26.9	25.7	0.02
Nasal/Flap	16.6	14.8	0.001
Stop	20.4	16.6	0.0001
Fric/Clos/Sil	9.1	8.8	0.1

Table 3.4: Classification error on the 50 speaker development set

Task	% Error on full test set		
	Baseline	Heterogeneous	Significance
Overall	21.6	21.0	0.001
Vowel/Semivowel	27.8	27.3	0.18
Nasal/Flap	16.5	15.4	0.004
Stop	19.6	17.9	0.002
Fric/Clos/Sil	9.1	8.8	0.06

Table 3.5: Classification error on the 118 speaker test set.

was 0.16. However, we suspected that the small size of the core set made significance testing somewhat coarse. Therefore, we also compared the baseline and heterogeneous framework results on the 118 speaker test set, which includes all data not in the training or development sets, with results summarized in Table 3.5. The overall results of 21.6% and 21.0% were the same as for the core set, but with better significance levels. These results confirm that heterogeneous measurements are producing significant improvements on independent testing data.

The above MAP framework allows for some interaction between the scores at different levels. Alternatively, we have implemented a strict framework in which the first classifier makes a hard decision about the level one class membership. This strict framework also achieved 20.0% error on the development set, and fewer than 1% of the testing tokens were classified differently from the MAP framework. The strict framework requires the computation of only one level one feature set $\vec{f}_j^{(1)}$ for each segment, which provides an opportunity for computational savings compared to the

MAP framework. This strict framework can be thought of as a strategy for pruning the full MAP framework, and other pruning strategies could also be devised which save computation with minimal effect on performance [13].

3.4.2 TIMIT Phonetic Classification with Class-specific VTLN

Vocal Tract Length Normalization (VTLN) is an efficient speaker normalization technique. We experimented with using phone-class-specific VTLN in order to determine if there could be any benefit from using a normalization factor in different phonetic classes. The baseline acoustic measurements described in this chapter were used in all of these experiments.

VTLN warp factors for the training data were developed based on three different phonetic classes: sonorants (SON), obstruents (OBS), and sonorants and obstruents together (S+O). Table 2.3 indicates the division of the TIMIT phone set into sonorant, obstruent, and silence classes. Figure 3-5 indicates the trajectory of the average warp factors over the training speakers under each training condition. There are differences between the average male and female warp factors under each condition.

Figure 3-6 summarizes the error rate reduction that was achieved under the three different training conditions, for the case of using the reference transcription together with 8 utterances for each speaker at testing time. Using the reference transcription removes any possible deleterious effects from determining the testing warp factors on incorrectly labeled data. We wanted to be sure to use correctly labeled data in order to see if class-specific VTLN warpings provide any advantages over using a single VTLN warp factor. Also, it can be seen from Table 3.6 that significant error rate reduction can also be achieved using hypothesized transcriptions and instantaneous adaptation. For the overall and sonorant tasks, training the warp factors on sonorants only is best. For the obstruent classification task, VTLN warping produced only small changes in performance under all three training conditions.

Before presenting more detailed results, the training and testing procedures will be described. The warping of the training data was obtained using the following procedure:

1. **Initialization**: Assign one warping factor, k_n to each training speaker. Let $k_n(i)$ denote the warping factor for the The boundary models

   n-th speaker on the i-th iteration. Initialize warp factors to 1.0, that is, $k_n(0) = 1.0 \ \forall \ n$.

2. **Train models with current warp factors**: Train models using $k_n(i)$.

3. **ML choice of new warping factors**: For each speaker, and for each warp factor, obtain the acoustic likelihood of that speaker's training data. For each speaker, find the warp factor which produces the maximum acoustic likelihood. That specifies $k_n(i + 1)$.

4. **Increment** i, and **repeat** steps 2 and 3 until the change in the warp factors is below a threshold.

To make the training iterations specific to a particular phone class, such as sonorants, only sonorants are used in step 3.

Once the warp factors for the training data have been determined by the procedure above, models can be trained using the warped data. The testing procedure includes three steps:

1. **Transcribe using neutral warping:** Obtain a transcription of the test data using a warping factor of 1.0 on the test data, which is equivalent to no warping

2. **ML search for warp factor.** Using the transcription from the previous step, select the warp factor which produces the highest acoustic likelihood.

3. **Re-transcribe using the warp factor from ML search.** The result of this transcription is the final answer.

There are a number of variations on this procedure. In step 1, models trained on warped data or unwarped data could be used. It is better to use models trained on unwarped data, but it requires having two sets of models, since the models trained on warped data will be used in step 2. In step 2, for development purposes, the reference transcription could be used instead of a hypothesized transcription. Testing can also be done "instantaneously," where the warp factor for each testing utterance is determined independently, or in "batch" mode, where a group of utterances which are known to be from the same speaker are used to choose the warp factor. The results will be given for both instantaneous adaptation and batch adaptation using eight utterances at a time.

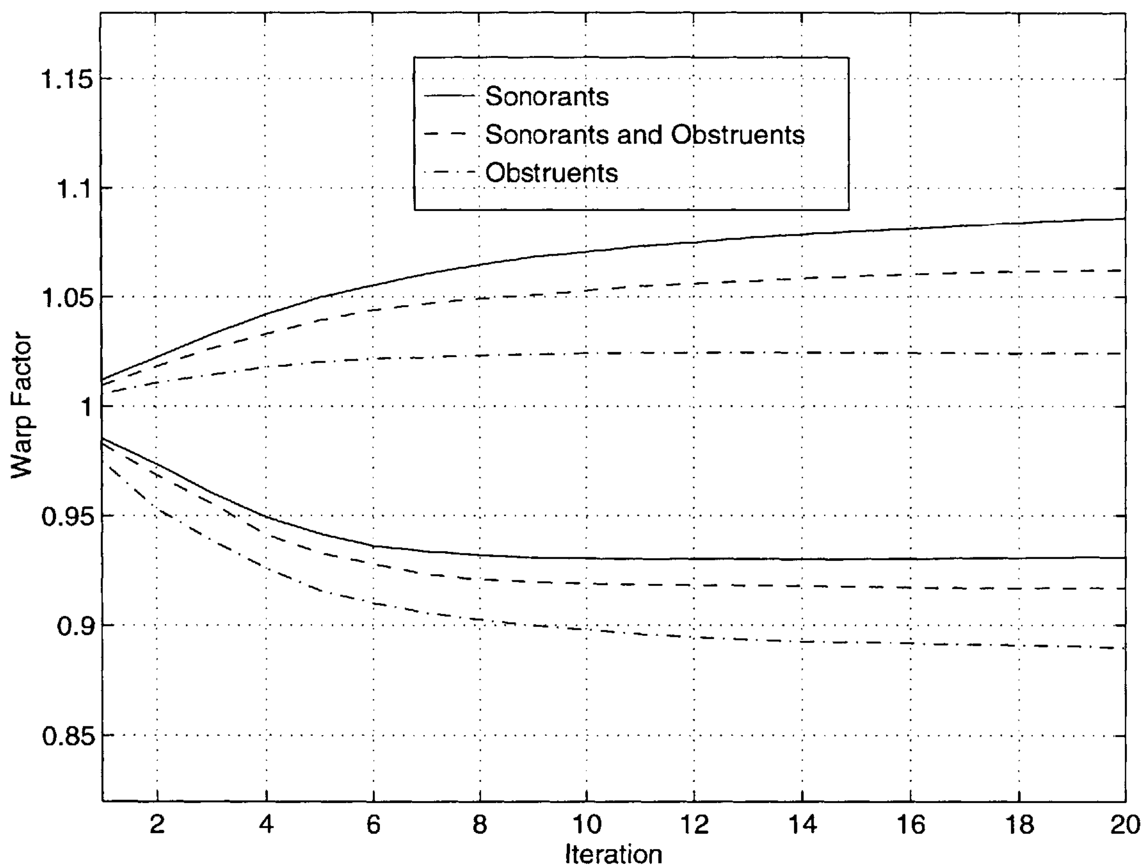

Figure 3-5: Trajectory of the average warp factors of the training speakers as a function of the training iteration. The legend indicates which phonetic labels were used in the warp factor training procedure. Males and females are graphed separately. Average male warp factors are greater than one, average female warp factors are less than one.

Table 3.6 shows detailed absolute error rate performance on the development set for a variety of training and testing conditions. The baseline classification level of 21.7% is different from the 21.1% in Table 3.4 because the classifiers are different.

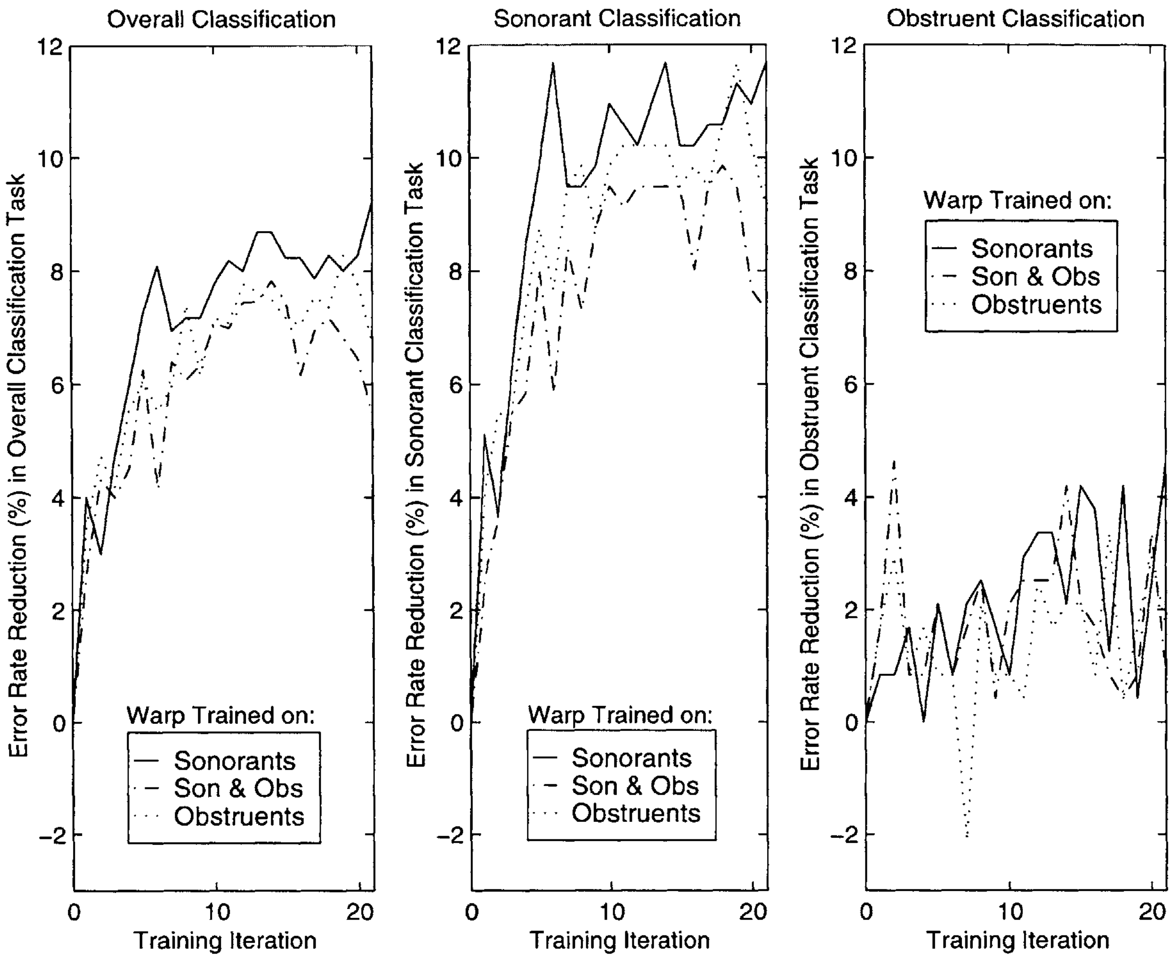

Figure 3-6: Results of phone-class-specific vocal tract length normalization experiments

In this section, the classifier uses diagonal-covariance Gaussian models with 8-fold aggregation, whereas the 21.1% result used full-covariance Gaussian models with 20-fold aggregation.[1] These classifier parameters are summarized in Section 2.4.1. Table 3.6 is particularly helpful for showing the performance differences stemming from different training and testing conditions. The results confirm that it is better to use models trained on unwarped data for the first-pass classification. When the warp factor was trained on sonorants and the reference transcription of the test utterance was used to find the warp factor of the test utterance, the system obtained 19.7%

[1] Aggregation is described in Appendix C.

overall classification error. When the transcription hypothesis of the testing utterance was given by unwarped models and a bigram phonotactic model, the system obtained 19.9%. These two results are quite similar and correspond to 9.2% and 8.4% relative error rate reduction, respectively. The similarity indicates that warp factors can be found reliably even when the true transcription is not known.

In conclusion, these results do not indicate that using different VTLN warpings for different phone classes will be helpful. On the positive side, they show that training the warp factor on sonorants is the best of the three training conditions. They also show that the standard linear warping produces very little performance improvement in the obstruents.

3.5 Summary

These experiments demonstrate the viability of using heterogeneous, phone-class-specific measurements to improve the performance of acoustic-phonetic modeling techniques. In order to be able to compare with other results in the literature, we do not include vocal tract length normalization. The hierarchical system achieved 21.0% error on the core test set in the task of context-independent phonetic classification. This performance compares favorably with the best results reported in the literature. Zahorian [96] reports 23.0% on the core test set, while Leung et al. [50] report 22.0% on a different test set.

This chapter explored hierarchical approaches to the two challenges of measurement development and classifier combination. Heterogeneous measurements were used across different phone classes. In subsequent chapters, we will apply heterogeneous measurements both within and across phonetic classes in committee-based and hybrid approaches.

Warp Trained on:	Testing Conditions	Warped models for first pass?	Overall Classification		Sonorant Classification		Obstruent Classification	
			Error	ERR	Error	ERR	Error	ERR
SON	baseline	–	21.74	–	27.4	–	23.8	–
SON	8/ref	–	19.73	9.2	24.2	11.8	22.7	4.6
SON	1/ref	–	19.89	8.4	24.3	11.5	23.1	3.2
SON	8/hyp(bi)	Yes	20.18	7.2	24.9	9.2	22.9	3.8
SON	1/hyp(bi)	Yes	20.34	6.4	25.0	8.8	23.3	2.3
SON	8/hyp(uni)	Yes	20.27	6.8	25.1	8.7	23.0	3.5
SON	1/hyp(uni)	Yes	20.46	5.9	25.2	8.0	23.3	2.3
SON	8/hyp(bi)	No	19.92	8.4	24.5	10.6	22.7	4.6
SON	1/hyp(bi)	No	20.08	7.6	24.6	10.2	23.1	3.2
SON	8/hyp(uni)	No	19.92	8.4	24.5	10.7	22.8	4.4
SON	1/hyp(uni)	No	20.14	7.4	24.6	10.2	23.2	2.7
OBS	baseline	–	21.74	–	27.4	–	23.8	–
OBS	8/ref	–	20.56	5.4	25.4	7.5	23.6	0.8
OBS	1/ref	–	20.69	4.8	25.8	6.0	23.3	2.2
S+O	baseline	–	21.74	–	27.4	–	23.8	–
S+O	8/ref	–	20.28	6.7	24.9	9.3	23.4	1.9
S+O	1/ref	–	20.14	7.3	24.7	9.9	23.3	2.3

Table 3.6: Error rate performance and relative error rate reduction (ERR) of VTLN under a variety of training and testing conditions. The second column indicates the number of utterances from the current speaker used at testing time to determine the warp factor for that utterance and the method of obtaining the "first-pass" transcription of the testing utterance so that the warp factor could be chosen. "ref" refers to using the reference transcription, "hyp(uni)" refers to using a transcription hypothesized with a unigram phonotactic model, and "hyp(bi)" refers to using a transcription hypothesized with a bigram phonotactic model. The third column also pertains to testing and indicates whether warped or unwarped models were used for the "first-pass" transcription.

Chapter 4

Committee-based and Hybrid Methods: Measurements and Classifier Structures

Measurements and classifier structures for committee-based and hybrid combinations of multiple classifiers are presented in this chapter. First, some related work from the literature is discussed. Second, a variety of acoustic measurement algorithms are described. Third, methods for combining classifiers are presented.

4.1 Related Work

There has recently been an increase in interest in committee-based methods in speech recognition. These efforts have taken a variety of forms, such as combining phone-based and syllable-based recognizers [95], combining recognizers operating in different frequency bands [3], or combining the outputs of recognizers developed at different research sites [25]. In [95], the phone-based and syllable-based recognizers were combined at the whole-utterance level. On the OGI numbers corpus, the phone-based system had 6.8% error, the syllable-based system had 9.8% error, and the combined system had 5.5% error, so the combination was successful for reducing the error rate. In [3], recognizers operating in different frequency bands could be combined at a variety of levels. The results do not show improvements in performance, but they do show increased robustness to narrow-band noise.

Committee-based approaches gained greater attention following the February 1997 Large Vocabulary Continuous Speech Recognition (LVCSR) Hub5-E Benchmark Test Evaluation conducted by the National Institute of Standards and Technology (NIST). Each site participating in the evaluation submitted their speech recognition output to NIST, including a word transcription with a confidence score for each word. NIST developed a post-processing system that takes multiple recognition outputs and combines them to yield a new transcription [25]. The system is referred to as "ROVER," standing for "Recognizer Output Voting Error Reduction." The ROVER system had several variations, depending on how the confidence scores were used. In system NIST1, confidence information was ignored so that ties were resolved by frequency of occurrence. In system NIST2, the system resolved ties by a weighted average between frequency of occurrence and average confidence for each word. In NIST3, the system used a weighted average between frequency of occurrence and the maximum confidence scores. The ROVER system was used to post-process 5 submissions to the 1997 Hub5-E Benchmark Test Evaluation. The submissions were from BBN, Carnegie Mellon University (CMU), Cambridge University (CU-HTK), Dragon Systems, and SRI, with word error rates of 44.9%, 45.1%, 48.7%, 48.9%, and 50.2%, respectively. The ROVER systems achieved error rates of 39.7%, 39.5%, and 39.4% for NIST1, NIST2, and NIST3, respectively. These results are relative error rate reductions of 11.8%, 12.0%, and 12.5%, respectively, compared to the best individual system word error rate of 44.9%. Recall that the NIST1 system does not use any confidence information, but rather is based only on voting with frequency of occurrence.

The favorable results from the NIST ROVER system resulted in others adopting some form of ROVER-like aspect to their individual systems in more recent evaluations. System descriptions from the September 1998 evaluation on Switchboard and CallHome corpora [67] indicate that at least three sites used ROVER-like approaches in their systems. BBN rescored N-best lists using different frame rates at the acoustic modeling level (80, 100, or 125 frames per second), then combined the results using a modified ROVER system which has weights on the three scoring results. HTK used ROVER to combine the outputs from a triphone and a quinphone system. AT&T used ROVER to combine two systems based on different lexicons.

In summary, committee-based methods for speech recognition have been explored along a variety of dimensions, and favorable results have been obtained. However, most of these approaches have been limited to combining final word recognition outputs. The techniques proposed in this thesis use multiple classifiers at the acoustic modeling level, so that the change in acoustic modeling is invisible to the remainder of the system. In addition, performing the combination at the acoustic modeling level allows the system to provide more accurate acoustic-phonetic information to the first-stage search.

4.2 Measurements

This section presents a variety of measurement sets which were designed to contain complementary phonetic information. This design was accomplished through phonetic and signal processing knowledge combined with empirical verification. The measurement sets presented here vary in several ways, namely, using different time-frequency resolutions, different temporal basis vectors, and/or different spectral representations. This section describes the algorithms for calculating these measurements and presents the performance of individual measurement sets in phonetic classification tasks.

The measurements in this section make use of Mel-frequency cepstral coefficient (MFCCs) and perceptual linear prediction cepstral coefficients (PLPCCs) spectral representations. The calculation of these representations is described in Chapter 3. Recall that PLPCCs make use of an all-pole model of the spectrum and a Bark frequency scale. In contrast, MFCCs are not model-based and make use of a Mel-frequency spectrum. In addition, PLPCCs attempt to account for addition perceptual effects through the use of the intensity-loudness power law. In summary, MFCCs and PLPCCs are similar, but it is hypothesized that the variation between them will cause them to contain some complementary information.

Both MFCCs and PLPCCs use a Hamming window

$$w[n] = 0.54 - 0.46\cos(2\pi n/M), \quad 0 \leq n \leq M$$

in the initial short-time Fourier analysis of the speech signal. The functional form and the length of $w[n]$ determines the frequency resolution of the spectral representation. Let $W(e^{j\omega})$ be the discrete-time Fourier transform (DTFT) of $w[n]$. When the short-time section of speech is multiplied by $w[n]$, the corresponding operation in the frequency domain is a convolution of the DTFT of the speech signal with $W(e^{j\omega})$. The width of the mainlobe of $W(e^{j\omega})$ is $\frac{8\pi}{M}$ in digital radian frequency. Thus, if the Hamming window has a length of 10, 20, or 30 milliseconds, the corresponding mainlobe widths in Hertz (given a sampling rate of 16 kHz) are 200, 100, and 66 Hz, respectively. This frequency resolution will affect all of the subsequent processing. Any details in the speech signal which are blurred because of poor frequency resolution cannot be recovered in subsequent processing. For this reason, we calculate multiple measurement sets using different Hamming window lengths in order to achieve a variety of time-frequency resolutions.

We divide the acoustic measurements obtained from the spectral representations into two types: segmental measurements, which are calculated based on a start and end time, and boundary, or landmark, measurements which are calculated using a single reference time specification. Table 4.1 summarizes the characteristics of eight segmental (S1–S8) and five boundary (B1–B5) measurements used in subsequent experiments. In all measurements, a frame rate of 200 frames per second (5 ms per frame) was used for short-time Fourier transform (STFT) analysis. The first column is a label for ease of reference. The second column indicates the number of dimensions in the measurement set. For B1 and B2, the notation $104 \Rightarrow 60$ indicates that principal components analysis was used to reduce the dimensionality of the measurements from 104 to 60. The third column indicates the duration in milliseconds of the Hamming window for short-time Fourier transform analysis. The fourth column includes

	# Dims	STFT [ms]	Spectral Representation	Temporal Basis
S1	61	10	12 MFCC	5 avg
S2	61	30	12 MFCC	5 avg
S3	61	10	12 MFCC	5 cos $\pm$ 30ms
S4	61	30	12 MFCC	5 cos $\pm$ 30ms
S5	64	10	9 MFCC	7 cos $\pm$ 30ms
S6	61	30	15 MFCC	4 cos $\pm$ 30ms
S7	61	20	12 PLPCC	5 avg
S8	61	20	12 PLPCC	5 cos $\pm$ 30 ms
B1	104 $\Rightarrow$ 60	30	12 MFCC+ energy	8 avg 5 10 20 40
B2	104 $\Rightarrow$ 60	20	12 PLPCC+ energy	8 avg 5 10 20 40
B3	60	30	12 MFCC	5 cos $\pm$ 75ms
B4	60	30	12 MFCC+ZC+ energy+LFE	4 cos $\pm$ 50ms
B5	60	10	10 MFCC	6 avg 20 20 20

Table 4.1: Segmental and boundary measurement set summary.

the spectral representation, which may include MFCCs or PLPCCs, energy, low frequency energy (LFE), and/or zero-crossing (ZC) rate. The fifth column indicates the temporal basis that was applied. In each case, the temporal basis was applied as an inner product with the frame-based spectral representation. For the segmental measurements, the cosine temporal basis extends 30 ms beyond the start and end of the segment on both sides, indicated by ± 30. For the boundary measurements, the cosine basis extended 50 or 75 ms to either side of the boundary. For segmental measurements, the "5 avg" basis consists of averages over the segment in a 3-4-3 proportion, and also includes a 30 ms average on either side of the segment. For the boundary measurements, the "8 avg" basis consists of symmetric, non-overlapping averages over 5, 10, 20, and 40 milliseconds (indicated by 5 10 20 40) [27], for a total extension of 75 ms to either side of the boundary. The width of the average is increasing as the distance from the boundary increases. Similarly, the "6 avg" basis consists of symmetric, non-overlapping averages over 20, 20, and 20 milliseconds, for a total extension of 60 ms to either side of the boundary.

Table 4.2 shows classification results for each of the eight segmental acoustic measurements. For each measurement set, the error rates shown are all from a single classification experiment, and thus the results for phonetic subclasses are for the case of unrestricted hypotheses. This implies, for example, that some of the errors in reference stop consonants may be due to hypothesizing a phone which is not a stop. This is in contrast to the case where the hypotheses are restricted to be within-class, thus creating a smaller classification problem. In the case of stops, the "restricted hypothesis" case is a 6-way classification task. Results using restricted hypotheses are reported in Section 5.1.3 and Chapter 6.

Examination of the within-class performance of these measurement sets, as shown in Table 4.2, reveals differences among them. In the sonorants, the use of a cosine basis in the time dimension in sets S3 and S4 is superior to the use of a piecewise-constant basis, as in S1 and S2. In the weak fricatives and stops, the piecewise constant temporal basis of measurement sets S1 and S2 is superior to the cosine basis of S3 and S4. These results indicate that these measurements contain complementary information. When examining only the overall classification results, these differences are not apparent, since all four of S1-S4 have very similar overall classification error rates. This is an example of the importance of examining detailed performance results when looking for differences among classifiers. Measurements S5 and S6 both perform slightly worse than S1-S4, but they were retained because they are designed to capture different acoustic information. S5 has increased time resolution and decreased frequency resolution through the use of a 10ms Hamming window, a smaller number of MFCCs, and a greater number of temporal cosine basis vectors. S6 has decreased time resolution and increased frequency resolution through the use of a 30 ms Hamming window, more MFCCs, and a smaller number of temporal basis vectors. S7 and S8 are comparable to S5 and S6 in overall performance. S7 and S8 are based on the PLPCC spectral representation instead of the MFCCs. As in S1-S2 versus S3-S4, the cosine basis in S8 produces better performance in the sonorants in comparison to the piece-wise constant basis in S7.

Acoustic Measurements	% Error: Unrestricted Hypotheses								
	ALL	VOW	NAS	SFR	WFR	STP	SIL	SON	OBS
S1	21.52	28.4	22.4	19.8	26.9	23.7	3.8	27.3	23.2
S2	21.60	28.9	22.2	18.2	28.4	23.7	3.4	27.7	23.1
S3	21.46	27.9	20.4	18.9	29.0	26.5	3.8	26.5	24.6
S4	21.47	27.9	20.7	19.5	30.1	24.8	4.0	26.6	24.4
S5	22.10	28.8	23.1	20.2	30.0	24.9	3.6	27.8	24.6
S6	22.64	28.4	24.6	20.8	32.5	26.6	4.4	27.7	26.2
S7	22.68	29.6	25.7	20.4	30.3	25.2	3.3	28.9	24.9
S8	22.08	28.3	25.8	19.4	30.3	24.6	3.9	27.8	24.3

Table 4.2: Context-independent classification performance on the development set for each of the 8 segmental measurement sets shown in Table 4.1.

4.3 Committee-based Classifier Structures

In this section we present several committee-based methods for combining multiple classifiers. The methods covered are voting, weighted linear combination of likelihood ratios (WLCLR), and finally the use of an independence assumption. In each case, the methods are structured so that the acoustic scores are combined before the search. This allows the changes in acoustic modeling to be invisible to the remainder of the system.

We will now define the notation that will be used in this section to describe the algorithms for combining classifiers. Let $A = \{\alpha_1, \alpha_2, \ldots\}$ be an ordered set of linguistic labels. Let $|A|$ denote the number of elements in A. Consider N classifiers which have been trained to discriminate among the elements of A. These classifiers may, in general, be defined over different measurement input spaces. Thus, for each input token and each $n \in \{1, 2, \ldots N\}$ there is a corresponding vector of measurements we denote by $\vec{f}_n$. For each token, let $\vec{f}$ be the column vector containing all of the measurements, that is,

$$\vec{f} = [\vec{f}_1^T \ \vec{f}_2^T \ \ldots \vec{f}_n^T]^T, \tag{4.1}$$

where $\cdot^T$ denotes the transpose operator. For each $\alpha \in A$ and each classifier n, let $p_n(\vec{f}_n|\alpha)$ be the scalar value of the conditional probability density function (pdf) of

$\vec{f}_n$. For notational convenience, we form these conditional pdf values into a matrix. Let $\mathbf{P}$ denote the $|A| \times N$ matrix of conditional pdf values

$$\mathbf{P} = \begin{bmatrix} p_1(\vec{f}_1|\alpha_1) & p_2(\vec{f}_2|\alpha_1) & \cdots & p_N(\vec{f}_N|\alpha_1) \\ p_1(\vec{f}_1|\alpha_2) & p_2(\vec{f}_2|\alpha_2) & & \\ \vdots & & \ddots & \\ & & & \end{bmatrix} \tag{4.2}$$

For each input token, the output of the acoustic modeling system is a vector of scores with one entry for each linguistic unit, that is,

$$\vec{s} = \begin{bmatrix} s(\alpha_1) \\ s(\alpha_2) \\ \vdots \end{bmatrix}. \tag{4.3}$$

The acoustic scores $\vec{s}$ may be an arbitrary function of $\mathbf{P}$ and other parameters affecting the combination of pdf values, which we denote by $\mathbf{\Gamma}$. Thus, we can write

$$\vec{s}(\vec{f}) = \mathcal{F}(\mathbf{P}, \mathbf{\Gamma}). \tag{4.4}$$

The dependence of $\vec{s}$ on all of the input measurements $\vec{f}$ is shown explicitly for emphasis. Each of the following sections will describe a different algorithm for the function $\mathcal{F}$ which produces the scores from the classifier outputs and other free parameters.

4.3.1 Voting

Voting is the simplest of the procedures presented here. In spite of its simplicity, there are two issues which need to be resolved. First, the algorithm needs to output a vector of scores for each linguistic unit, as opposed to the naive voting concept which would imply returning only the single most "popular" linguistic unit. Secondly, a method for resolving tied votes is needed. These two issues are addressed in the following algorithm.

Let $\vec{b} = [b_1 b_2 \dots b_N]^T$ be defined as

$$b_n = \arg \max_{k \in \{1,2,\dots|A|\}} p_n(\vec{f}_n|\alpha_k), \tag{4.5}$$

which is the index of the most likely phonetic label for each classifier. Let $\vec{v}$ be a $|A|$-dimensional integer-valued vector which is used to count votes. The individual elements of $\vec{v}$ are given by

$$v_k = \sum_{n=1}^{N} \delta[b_n - k], \quad \forall\, k \in \{1, 2, \dots |A|\} \tag{4.6}$$

where the discrete delta function δ is given by

$$\delta[n-m] = \begin{cases} 1, & \text{for } m = n \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

Let $v^* = \max_{k \in \{1,2,\dots|A|\}} v_k$. Let ℓ be the smallest integer such that $v_\ell = v^*$. The output scores $\vec{s}(\vec{f})$ are the ℓth column of $\mathbf{P}$, that is,

$$\vec{s}(\vec{f}) = \begin{bmatrix} p_\ell(\vec{f}_\ell|\alpha_1) \\ p_\ell(\vec{f}_\ell|\alpha_2) \\ \vdots \end{bmatrix}. \tag{4.8}$$

To state it in words, the tie is resolved by finding the classifier with the lowest index among all those classifiers whose top scoring linguistic unit received at least as many votes as any other linguistic unit.

4.3.2 Weighted Linear Combination of Likelihood Ratios

The weighted linear combination of likelihood ratios (WLCLR) method makes use of additional parameters, $\mathbf{\Gamma}$, in defining the function $\mathcal{F}$, as introduced in Equation 4.4.

The equation for the WLCLR scores is

$$s(\alpha_k) = \sum_{n=1}^{N} g_{k,n} \left(\frac{p_n(\vec{f}_n|\alpha_k)}{\sum\limits_{a \in A} p_n(\vec{f}_n|a)} \right), \tag{4.9}$$

where the weights $g_{k,n}$ have the property

$$\sum_{n=1}^{N} g_{k,n} = 1 \quad \forall k, \tag{4.10}$$

so that each row of the $|A| \times N$ parameter matrix

$$\mathbf{\Gamma} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,N} \\ g_{2,1} & g_{2,2} & & \\ \vdots & & \ddots & \\ & & & \end{bmatrix} \tag{4.11}$$

will sum to one. In the general case, the weights can be classifier-specific and/or linguistic-unit specific. If equal weights are used, then

$$g_{k,n} = \frac{1}{N} \quad \forall k, n. \tag{4.12}$$

Alternatively, the weights could be trained on a development set using a Maximum Likelihood (ML) or Linear Least-Square Error (LLSE) criterion. The likelihood ratio serves to normalize the absolute magnitude of the pdf values across classifiers.

4.3.3 Assuming Statistical Independence

Let $\vec{x}$ and $\vec{y}$ be random vectors. Let $\vec{z} = [\vec{x}^T \vec{y}^T]^T$. If $\vec{x}$ and $\vec{y}$ are statistically independent, then the joint density $p(\vec{z})$ can be factored

$$p(\vec{z}) = p(\vec{x})p(\vec{y}). \tag{4.13}$$

In a similar fashion, one method of combining classifiers is to assume statistical independence among the n random vectors $\vec{f_1}, \vec{f_2}, \ldots \vec{f_n}$ which leads to the expression

$$s(\alpha_k) = \prod_{n=1}^{N} p_n(\vec{f_n}|\alpha_k) \quad \forall k. \tag{4.14}$$

The feature vectors $\vec{f_1}, \vec{f_2}, \ldots \vec{f_n}$ in our experiments will seriously violate the independence assumption. Empirical results demonstrate that in spite of the faulty assumption in the derivation, this algorithm is still an effective method for combining the outputs of multiple classifiers.

In practice, the combination method in Equation (4.14) can be implemented by summing the log probabilities. Furthermore, in this thesis, the implementation was altered to average the log probabilities. In the case of phonetic classification with unigram priors, this alteration produces no change in performance. However, in systems with a language model, the dynamic range of the acoustic scores will interact with the language model. Averaging the log scores keeps the same dynamic range in the sense that if all input classifiers had the same score vector, the output score vector would be identical to each of the inputs.

We would like to gain some insight into how to interpret the use of the independence assumption in this context, where we know that the random vectors are not truly independent. In particular, we would like to understand the differences between training a single classifier using the entire vector $\vec{f}$ versus training multiple classifiers using $\vec{f_n}$ and combining the likelihood estimates using Equation (4.14). First, if all of the vectors were modeled with a Gaussian distribution, then assuming independence among the $\vec{f_n}$ would be identical to assuming a block-diagonal structure for the covariance matrix of the overall feature vector $\vec{f}$. The actual situation is considerably more complicated, however, because the experiments use *mixture* Gaussian models to parameterize the probability density function of the random vectors.

Given the use of mixture Gaussian models, let us consider what conditions lead to equivalent models when training a single classifier using the entire vector $\vec{f}$ versus training multiple classifiers using $\vec{f_n}$ and combining with Equation (4.14). Assume that each $p_n(\vec{f}_n|\alpha)$ is modeled as a mixture of Gaussian probability density functions with L_n components,

$$p_n(\vec{f}_n|\alpha) = \sum_{\ell=1}^{L_n} \omega_{n,\ell}\, \mathcal{N}(\vec{f}_n; \mu_{n,\ell}, \Sigma_{n,\ell}), \tag{4.15}$$

where $\mathcal{N}(\vec{f}; \mu, \Sigma)$ denotes a Gaussian (Normal) distribution with mean vector μ and covariance matrix Σ. The statistical independence assumption implies

$$p(\vec{f}|\alpha) = \prod_{n=1}^{N} \left[\sum_{\ell=1}^{L_n} \omega_{n,\ell}\, \mathcal{N}(\vec{f}_n; \mu_{n,\ell}, \Sigma_{n,\ell}) \right]. \tag{4.16}$$

We now show by construction that there exists an equivalent mixture Gaussian probability density function for $p(\vec{f}|\alpha)$ of the form

$$p(\vec{f}|\alpha) = \sum_{s \in S} \omega_s\, \mathcal{N}(\vec{f}; \mu_s, \Sigma_s) \tag{4.17}$$

Consider the integer-valued n-tuples $(1, 1, \ldots 1)$ through $(L_1, L_2, \ldots L_N)$. There are a total of $(\prod_{n=1}^{N} L_n)$ of these n-tuples. Denote the set of all of these n-tuples by the letter S, and let $s \in S$ with $s = (s[1], s[2], \ldots s[N])$. To complete the construction, let

$$\omega_s = \omega_{1,s[1]}\, \omega_{2,s[2]}\, \cdots\, \omega_{N,s[N]} \tag{4.18}$$

$$\mu_s = [\mu_{1,s[1]}^T \quad \mu_{2,s[2]}^T \quad \cdots \quad \mu_{n,s[N]}^T]^T \tag{4.19}$$

and

$$\Sigma_s = \begin{bmatrix} \Sigma_{1,s[1]} & \mathbf{0} & \mathbf{0} & \cdots & \\ \mathbf{0} & \Sigma_{2,s[2]} & \mathbf{0} & & \\ \mathbf{0} & \mathbf{0} & \Sigma_{3,s[3]} & & \\ \vdots & & & \ddots & \\ & & & & \Sigma_{N,s[N]} \end{bmatrix} \tag{4.20}$$

where $\mathbf{0}$ refers to zero-matrices of the appropriate dimensions to maintain the block-diagonal structure of Σ_s.

Consider the following example. Assume we have trained 8 classifiers ($N = 8$)

where each has 100 mixtures ($L = 100$), yielding a model in the form of Equation (4.15). The above construction says that it would take 100^8 mixtures to construct precisely the same model in the form of Equation (4.17). Of course, due to physical limitations, it is not possible to construct such a large model, which in this case would require *100 million billion* mixtures. In contrast, the original set of $N = 8$ independent models in Equation (4.15), using a total of 800 mixtures, is very practical. Consider further the case in which all covariance matrices $\Sigma_{n,\ell}$ are diagonal. In this case, correlations among dimensions within a particular $\vec{f}_n$ are modeled completely by the placement of the mean vectors, with no contribution from off-diagonal elements in the covariance matrix.

In conclusion, the decision regarding how to partition the full acoustic measurement space $\vec{f}$ into subspaces $\vec{f}_n$ is equivalent to deciding which correlations to model. Put another way, it is equivalent to deciding which correlations are important for discriminating among classes.

4.4 Hybrid Classifier Structures

Two types of hybrid classifier structures were implemented which combine hierarchical and committee-based techniques. The first method is to build a hierarchy-of-committees, where a committee of classifiers is used at each node of a hierarchical tree. The second method is to use a hierarchical classifier as one member of a committee of classifiers.

4.5 Summary

This chapter presented measurements and classifier structures for committee-based and hybrid approaches for using heterogeneous measurements. Voting, linear combination, and the use of an independence assumption are the three committee-based methods that were presented. The next chapter presents empirical evaluation of classifier structures for combining multiple measurement sets.

Chapter 5

Committee and Hybrid Methods: Experiments

This chapter begins with an evaluation of committee-based and hybrid methods for incorporating heterogeneous measurements in the task of TIMIT phonetic classification. The most promising methods are then further evaluated in TIMIT phonetic recognition and JUPITER word recognition tasks.

5.1 TIMIT Phonetic Classification

Comparisons among the algorithms for combining classifiers were performed on the TIMIT phonetic classification task. Unigram, bigram, and trigram phonotactic models were used in order to observe the interaction of these techniques with higher-level knowledge sources.

5.1.1 Comparing Voting, WLCLR, and Independence

We compare the results of using voting, linear combination with equal weights, or an independence assumption for combining multiple classifiers in the task of TIMIT phonetic classification. Figures 5-1, 5-2, and 5-3 show the performance of all possible subsets of the eight segmental measurements sets S1–S8 listed in Table 4.1. Error

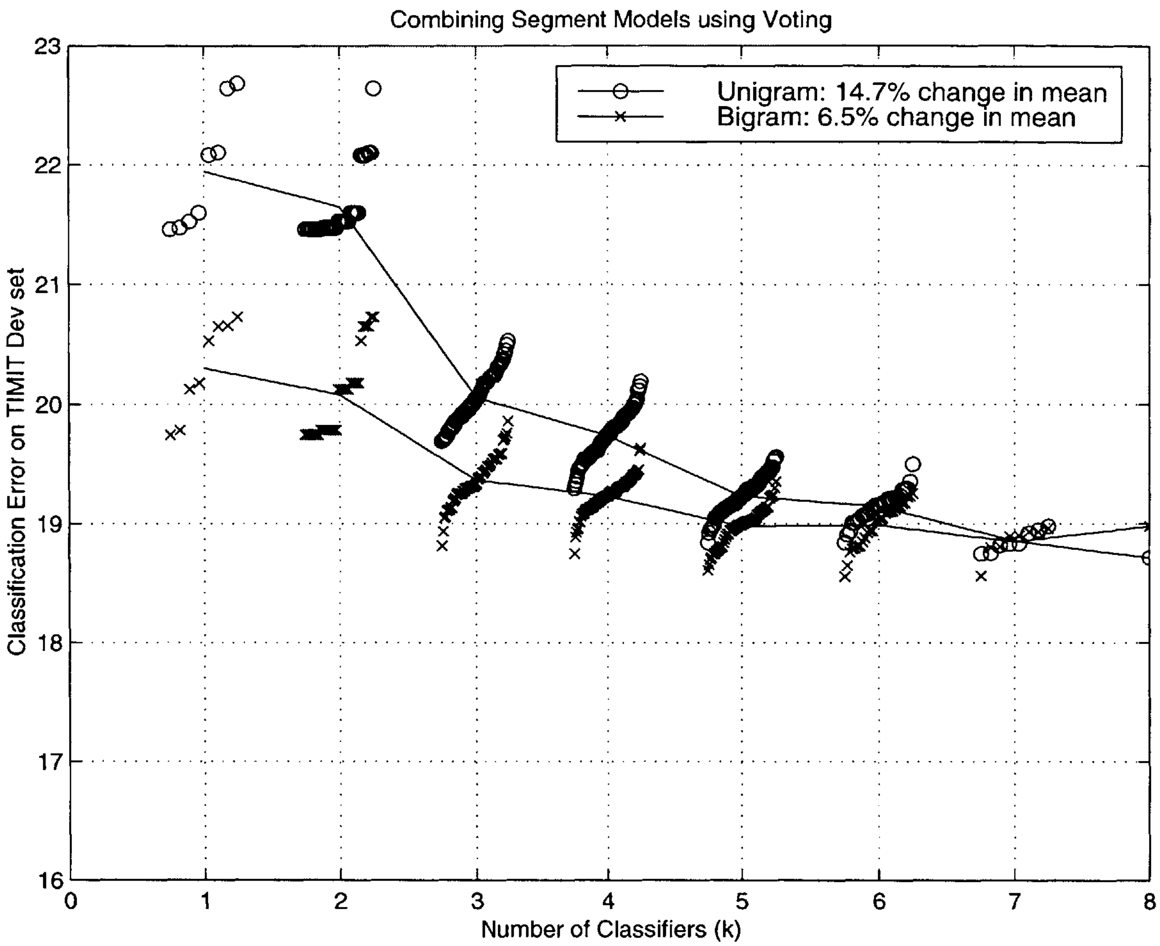

Figure 5-1: Performance of classifier combination using voting for all possible subsets of 8 classifiers. Error rate is shown as a function of the number of classifiers in the set. Results are shown for both unigram and bigram phonotactic model conditions.

rates are shown as a function of the number of classifiers in the subset. In each figure, individual data points are evenly spaced along the x-axis in the vicinity of the appropriate value of k, where k represents the number of classifiers being combined. Lines connect the mean values. The total number of experiments for each phonotactic model is

$$\sum_{k=1}^{8} \binom{8}{k} = \sum_{k=1}^{8} \frac{8!}{k!(8-k)!} = 255.$$

As k varies from 1 to 8, the number of individual classification experiments is 8, 28, 56, 70, 28, 8, 1, respectively. In addition to Figures 5-2 and 5-3, Table 5.1 summarizes

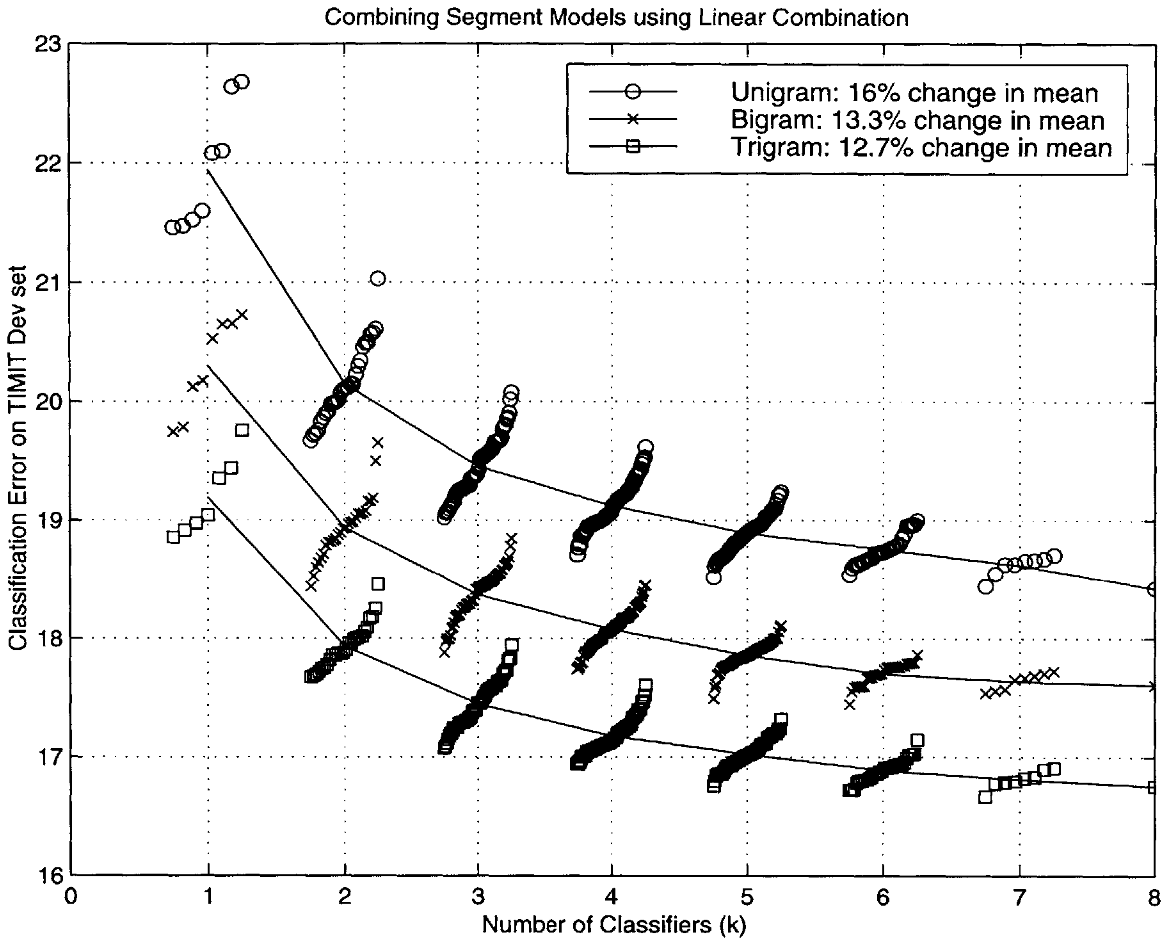

Figure 5-2: Performance of classifier combination using linear combination for all possible subsets of 8 classifiers. Error rate is shown as a function of the number of classifiers in the set. Results are shown for unigram, bigram, and trigram phonotactic model conditions.

some of the best CI classification results.

Voting can produce different results depending upon the ordering of the classifiers which is used to resolve tied votes. Figure 5-1 shows the performance where the classifiers were ordered from best to worst based on their individual performance. This explains why the mean of the voting performance with two classifiers (which is actually equivalent to replacement with the first classifier) is better than the mean with one classifier. For $k = 3, 4, \ldots, 8$, the average absolute difference in error rate between the voting results using a "best-to-worst" versus a "worst-to-best" ordering

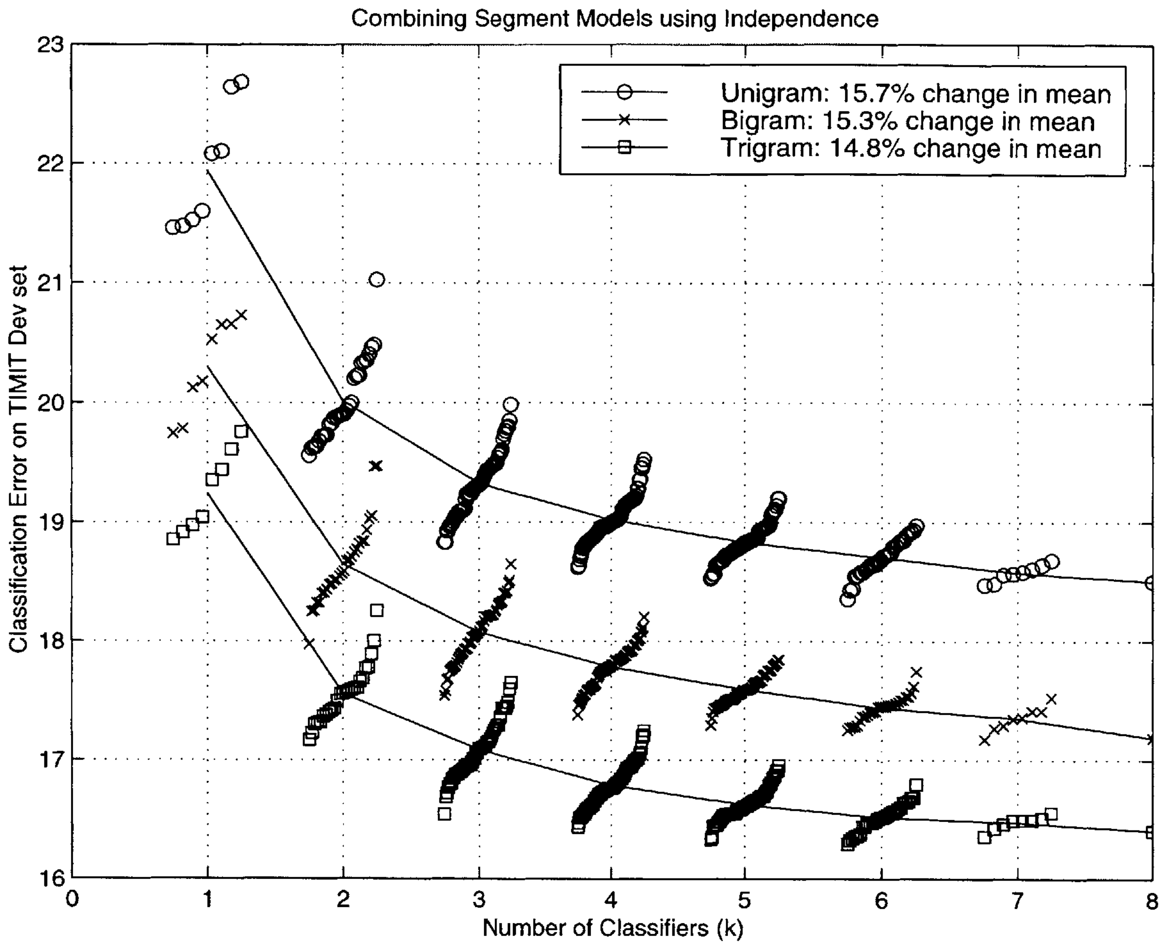

Figure 5-3: Performance of classifier combination using an independence assumption for all possible subsets of 8 classifiers. Error rate is shown as a function of the number of classifiers in the set. Results are shown for unigram, bigram, and trigram phonotactic model conditions.

is only 0.04%. Therefore, the ordering is not crucial to the comparison of voting with other methods.

In the unigram case the three methods of voting, WLCLR, and independence produce similar performance. In contrast, in the bigram case, voting with 8 classifiers obtained 18.9%, which is actually worse than the 18.6% that was obtained with voting in the unigram case. This is probably because voting lacks soft-decision capability, and thus it does not improve the quality of the entire vector of scores, but rather focuses only on the quality of the top choice. The WLCLR and independence methods produce favorable trends with all three phonotactic models, although the in-

Methods	% Error Dev	core
Voting (S1–S8)	18.6	–
Linear Combination (S1–S8)	18.4	–
Independence (S1-S8)	18.5	–
Hybrid: Committees at nodes of tree	18.3	–
Hybrid: S1–S8 + Hierarchy	18.1	18.3

Table 5.1: Summary of TIMIT CI classification results.

dependence assumption performs slightly better on average. The results indicate that indirect learning of phonotactic information has very little effect, since using multiple classifiers improves phonetic discrimination regardless of which phonotactic model is used. In addition, the independence method is less expensive to implement, since the log scores can simply be added together, and it does not require calculation of a likelihood ratio. For these reasons, the remaining experiments with hybrid techniques, phonetic recognition, and word recognition all use the independence assumption to combine committees of classifiers.

5.1.2 Hybrid Methods

In this section, two ways to combine hierarchical and committee-based approaches are tested. The first hybrid approach uses a committee of classifiers at each node of a hierarchical tree. This approach might also be called a "hierarchy-of-committees." We implemented phone-class specific classifiers which use different measurements for different phonetic classes, as listed in Table 3.3, except that MFCCs were used as the spectral representation for the vowel measurements instead of DCTCs. The performance of each is similar. Let us refer to these phone-class-specific hierarchical measurement sets as SVa (the lower-case "a" indicating the alteration to use MFCCs), SN, SF, and SS, representing segmental vowel, nasal, fricative, and stop measurements, respectively. A hierarchy-of-committees classifier was formed using: S1, S2, S4, and S5 at the root node; SVa, S1, S4, S6, and S8 at the vowel node; SN, S2, S3, and S4, at the nasal node; SF, S1, S2, and S3 at the fricative/closure node;

and SS, S1, S2, S5, and S8 at the stop node. Each of the committees was combined using an independence assumption. This resulted in a performance of 18.3% on the development set, as shown in Table 5.1. This hierarchical configuration suggests that computation can be reduced with minimal degradation in performance by targeting the measurements toward particular phone classes.

The second hybrid approach is to use a hierarchical classifier as one member of a committee. A hierarchical classifier was constructed using one set of measurements at each node. The measurements were S2, SVa, SN, SS, and SF for the root, vowel/semivowel, nasal/flap, stop, and fricative/closure nodes, respectively. This hierarchical classifier was added as a ninth member to the previously 8-member segmental measurements committee. The 9 classifiers were combined using independence to obtain 18.1% on the dev set, and 18.3% on the core set. Table 5.1 summarizes these results. This result is a 12.9% improvement over our previous best reported result of 21.0% [31]. The next best result that we have seen in the literature reporting TIMIT CI classification on the core test set is 23.0% [97].

5.1.3 Detailed Phonetic Classification Performance Metrics

Context-independent phonetic classification on the TIMIT database has been used extensively for comparing and contrasting various classification techniques. This is made possible over time through detailed reporting by various researchers. In order to facilitate a variety of performance comparisons by future researchers, this section includes more detail about the performance of our best context-independent classification result on both the core test set and the development set. These results are from the hybrid system which is a 9-member committee consisting of classifiers using measurements S1-S8 plus a hierarchical classifier using phone-class-specific measurements. The overall error rates of 18.1% and 18.3% on the development and core test sets, respectively, are shown in Table 5.1. Complete confusion matrices are given in Appendix D.

Tables 5.2 and 5.3 show detailed results of classifier performance within various phonetic classes. All of these results use the 39-classes in Table 2.2 for scoring. The performance depends upon whether the hypotheses are restricted to be within the correct phonetic class or not. The column of results labelled "unrestricted hypotheses" all came from a single classification experiment over all phones. The "restricted hypotheses" columns came from experiments where the hypotheses were restricted to be within the same phonetic class as the reference. For example, the restricted hypothesis experiment for stops is a 6-way classification experiment. Restricting the hypotheses makes the task easier, leading to lower error rates. The "restricted hypotheses" results are good for comparing with studies that look only at one sound class.

Another way to look at the results is to see how well the system can sort the phones into one of various phonetic partitions. Table 5.4 shows these results, which can be compared with the results from the baseline system in Chapter 3, shown in Table 3.1. Thus, there is 3.3% error when dividing the development set into phonetic manner classes. Interestingly, summing up the "restricted hypotheses" errors listed in Table 5.2, it can be seen that given perfect phonetic manner class partitioning, the phone-class specific classifiers presented here could achieve classification error of 15.6% on the development set.

Reference Phonetic Class	Unrestricted Hypotheses		Restricted Hypotheses	
	# Errors	% Error	# Errors	% Error
All	2729	18.1	2729	18.1
Sonorant	1835	22.9	1795	22.4
Obstruent	791	19.6	716	17.8
Vowel/Semivowel	1584	24.3	1544	23.7
Nasal	251	16.7	197	13.1
Strong Fricative	223	16.8	207	15.6
Weak Fricative	242	23.9	112	11.1
Stop	326	19.4	287	17.0
Silent	103	3.4	0	0

Table 5.2: Detailed context-independent classification results on the development set for the hybrid system composed of a nine-member committee. "Restricted hypotheses" refers to constraining the classifier to choose a phone in the same phonetic class as the reference. Scores are calculated using the 39-classes in Table 2.2.

Reference Phonetic Class	Unrestricted Hypotheses		Restricted Hypotheses	
	# Errors	% Error	# Errors	% Error
All	1320	18.3	1320	18.3
Sonorant	900	23.5	879	23.0
Obstruent	381	19.8	345	17.9
Vowel/Semivowel	768	24.8	741	23.9
Nasal	132	18.1	110	15.1
Strong Fricative	108	16.3	100	15.1
Weak Fricative	118	25.3	48	10.3
Stop	155	19.4	131	16.4
Silent	39	2.7	0	0

Table 5.3: Detailed context-independent classification results on the core test set for the hybrid system composed of a nine-member committee. "Restricted hypotheses" refers to constraining the classifier to choose a phone in the same phonetic class as the reference. Scores are calculated using the 39-classes in Table 2.2.

5.1.4 Training One High-dimensional Classifier

This section addresses an issue that often arises when considering the results in the previous sections. We have shown significant improvements by combining multiple classifiers. However, the number of acoustic measurements and the total number of model parameters vary greatly between the single classifier and the multiple classifier systems. Acoustic modeling experiments often try to keep the number of measure-

Reference Phone Partition	% of Hypotheses in the Incorrect Phone Class	
	Development	Core
{SON,OBS+SIL}	1.3	1.3
{SON,OBS,SIL}	1.9	1.7
{VS,NF,ST,SF+WF+CL}	2.7	2.9
{VS,NF,SF,WF,ST,CL}	3.3	3.4

Table 5.4: Shows the percentage of errors made by the 9-member committee hybrid classifier when dividing the development and core test sets into each of several phonetic partitions.

ments and the size of the acoustic models constant in order to make comparisons as equitable as possible. It is generally expected that using more measurements and/or increasing the number of parameters in the models provides potential for improvement, as long as the resulting measurement/modeling combination is still sufficiently trainable. This section shows that typical techniques for utilizing all the measurements within a single classifier do not lead to similar performance improvements.

Three segmental measurement sets were selected for this comparison. They are S2, S4, and S5 from Table 4.1. Table 5.5 shows that the error rate performance of these measurements when tested separately is 21.5–22.7%. When these measurements were combined using independence, the error rate is 18.9%. The model parameters shown in Table 5.5 are: the minimum number of data points per mixture component, the maximum number of mixtures per phone, the number of training trials aggregated.[1] The labels A1–A4 have been used for ease of reference to the four "all-in-one" measurement and classifier configurations. The "$\Rightarrow$" notation refers to dimensionality reduction through the use of principal components analysis. First, in conditions A1 and A2, which use all 185 measurements, the performance is poor. This is probably due to the poor trainability of the models due to high dimensionality. That is, the models tend to overfit the training data and thus they do not produce good generalization performance on the test data. The minimum number of data points per mixture component was adjusted between A1 and A2 in order to attempt to circumvent the overfitting problem and to adjust the model size (in megabytes, MB) in order to have a model which is comparable in size to the total models of the S2+S4+S5 system. Conditions A3 and A4 try an alternate approach, where dimensionality reduction is used to attempt to capture the information from the 185 measurements in a reduced set of 61 measurements. This approach was better than A1 and A2, but did not come close to performing as well as the S2+S4+S5 system. These results indicate that conventional techniques for training a single classifier from a large number of measurements are not successful. Thus, the multiple classifier techniques presented in this thesis are producing results that single classifiers have not been able to achieve.

[1] Aggregation is described in Appendix C.

Section 4.3.3 provides theoretical insight into why the "all-in-one" modeling of 185 measurements leads to lower performance than the separate modeling of the three measurement sets. From one vantage point, it is simply the exponential growth of the space which comes with high-dimensionality. From another viewpoint, it is shown in Section 4.3.3 that with mixture Gaussian models and typical numbers of mixture components, it would require an impractically large number of mixtures for the "all-in-one" model to produce an acoustic model which is equivalent to the acoustic model produced by modeling S2, S4, and S5 separately, then combining them using independence.

Acoustic Measurements	# Dims	Model Parameters	Model Size [MB]	% Error
S2:	61	61/96/4	4.8	21.5
S4:	61	61/96/4	4.8	21.5
S5:	64 ⇒ 61	61/96/4	4.8	22.7
S2+S4+S5: Independence			14.4	18.9
A1: "all-in-one"	185	185/96/4	4.6	24.6
A2: "all-in-one"	185	61/96/4	12.8	23.9
A3: "all-in-one"	185 ⇒ 61	61/96/4	4.8	21.4
A4: "all-in-one"	185 ⇒ 61	61/96/12	14.4	21.0

Table 5.5: Comparing training of a single classifier with 185 dimensions versus using three separate classifiers. The model parameters are: minimum # of data points per mixture, maximum # mixtures per phone, # of training trials aggregated. The arrow "⇒" indicates dimensionality reduction through principal component analysis.

5.1.5 Deciding which Measurements to Combine

Figures 5-2 and 5-3 indicate that combining measurements nearly always produces improvements in performance, regardless of which of the S1–S8 are combined. Looking more closely, it can be seen that some combinations of measurements perform better than others, which suggests that it would be advantageous to be able to predict which measurement combinations will perform the best. In this section, we test two methods of predicting classifier combination performance. The first is based on the average performance of the constituent classifiers, the second is based on a generalized distance metric which measures the degree of agreement among the constituent classifiers.

	Value of correlation coefficient						
	Number of Classifiers						Avg
	2	3	4	5	6	7	$\lvert \cdot \rvert$
(a) Avg Individual Error	0.15	-0.07	-0.10	-0.13	-0.20	-0.21	0.14
(b) Distance Metric	-0.61	-0.63	-0.58	-0.55	-0.54	-0.45	0.56

Table 5.6: Correlation coefficients between the error rate of systems using the combination of multiple classifiers and (a) the average error rate of the individual constituent classifiers, and (b) a generalized distance metric measuring the diversity of the hypotheses of the individual constituent classifiers. The final column shows the average magnitude of the correlation coefficients. These results are for combination using independence with a unigram phonotactic model.

We define a generalized distance metric in the following way. Consider a pairwise distance metric between classifiers as the number of tokens which they classify differently on the development set. Now generalize this metric to N classifiers by adding the pairwise distance between all classifier pairs in the set. This generalized distance metric measures how much diversity there is among the hypotheses made by the classifiers.

Consider first what correlations we expect to observe. First, naively one might expect to see that the performance of the combined classifiers is positively correlated with the performance of the constituent individual classifiers. However, we do not expect to see this since we hypothesize that the diversity among classifiers is the more important factor. Thus, we expect to see only weak correlation between the error rates. Second, we expect that a greater distance metric implies more diverse information and thus the potential for greater improvement when those classifiers are combined. This corresponds to expecting a significant negative correlation between the distance metric and the error rate.

Table 5.6 shows the correlation coefficients calculated for the case of combining classifiers with independence using only unigram phonotactic information. The corre-

Acoustic Measurements	Core Test set % Error	% Sub
avg of 1 seg + antiphone	30.1	19.0
avg of 1 seg + near-miss	28.7	18.0
5 segs + antiphone	27.7	16.7
avg of 1 bound	27.1	16.5
5 segs + near-miss	26.4	16.1
5 bounds	24.9	14.9
5 segs + 5 bounds + near-miss	24.8	15.0
5 segs + 5 bounds + antiphone	24.4	14.7

Table 5.7: Summary of TIMIT phonetic recognition results obtained in this work.

lation coefficients must be in the range -1 to 1, with 1 representing perfect correlation. As expected, the results in the second row, labelled (b), show significant negative correlation between the error rate of the classifier combination and the distance metric, with the average magnitude of the correlation being 0.56. This implies that the distance metric is useful as a predictor of which classifier combinations are likely to perform well. Now consider the results in the first row, labelled (a). When the number of classifiers was in the range 3 through 7, it was actually the case that a better average performance of the constituent classifiers was correlated with worse performance of the combination. This is counter to the naive expectation, but consistent with our hypothesis that diversity of the constituent classifiers is more important than looking for the best individual classifiers. It is also consistent with our hypothesis because the average magnitude of the correlations was relatively small, only 0.14.

In conclusion, given a set of classifiers, the generalized distance metric can be used to predict which classifier combinations are likely to perform well. This information makes it likely that combinations that perform well can be found without an exhaustive experimental search.

5.2 TIMIT Phonetic Recognition

Our TIMIT phonetic recognition experiments make use of a segment network produced by a first-pass recognition system. We refer to this step as probabilistic segmentation [6, 7, 47]. Either antiphone modeling [27] or 1-state near-miss modeling [6, 7] was used with segment models in order to account for both on-path and off-path segments in the segment network. Antiphone modeling and near-miss modeling are described in more detail in Section 2.3 and Appendix B. The phonetic recognition results make use of a phone bigram with a perplexity of 15.8 on the core set.

Table 5.7 summarizes a series of phonetic recognition experiments. The acoustic features for these experiments were S1, S2, S4, S5, S7 and B1–B5 from Table 4.1. The "avg of 1 seg" and "avg of 1 bound" rows refer to the average performance over 5 experiments where each measurement set was used by itself. For the segmental performance, we report cases of using both near-miss modeling and antiphone modeling. When using the antiphone, combining 5 segmental measurements reduced the error rate from 30.1% to 27.7%, which is a 7.9% reduction. This change in performance is smaller than what was observed in classification. However, the substitution error fell from 19.0% to 16.7%, which is a 12.1% reduction. Thus, combining multiple classifiers has a significant effect in reducing substitution errors. Combining 5 boundary measurements reduced the error rate from 27.1% to 24.9%, and substitution errors fell from 16.5% to 14.9%. Adding segment models to the boundary models did not produce much additional gain, probably because the segment models were context independent (CI), while the boundary models were context dependent (CD). Near-miss models were better than the antiphone when using only segment models, but were worse when using segment and boundary models together. The final phonetic recognition result of 24.4% compares favorably with results in the literature. Table 5.8 compares this result with the best results reported in the literature.

5.3 JUPITER Word Recognition

Finally, we performed experiments using the database associated with the telephone-based JUPITER weather information system [101, 29]. This evaluation is important because it demonstrates two types of generalization of the previous work. First, changing the task from phonetic classification or recognition to word recognition

Method	% Error core
Triphone CDHMM [39]	27.1
Recurrent NN [81]	26.1
Bayesian Triphone HMM [64]	25.6
Near-miss [7]	25.5
Heterogeneous Measurements	24.4

Table 5.8: Comparison of phonetic recognition results on TIMIT core set with the best results from the literature.

Acoustic Measurements	# Dimensions	% Error	% Sub
B1	104⇒50	11.3	6.4
B4	60⇒50	12.0	6.7
B3 (altered)	60⇒50	12.1	6.9
3 bounds: B1 + B4 + B3(alt)	-	10.1	5.5

Table 5.9: Summary of JUPITER word recognition results.

corresponds to linguistic or lexical generalization. Secondly, there are significant changes in the nature of the speech data, the acoustic environment, and the bandwidth of the system. Specifically, TIMIT contains clean, read speech recorded with a close-talking microphone at a 16kHz sampling rate. In contrast, the JUPITER database contains noisy, spontaneous speech recorded over-the-telephone at an 8kHz sampling rate.

For the experiments in this thesis, the recognizer was configured to use an 1893-word vocabulary and a class bigram language model with a perplexity of 15.2 on the 1806 utterances in the test set. Additional information about the training and test sets is available in Chapter 2. In [29], a word error rate performance of 11.8% was reported based on the same test set as in this work, but with a larger training set. Their training set consisted of 20,064 utterances, which is about 13% larger than the 17,807 utterances that were used for training for the experiments in this thesis.

We trained three sets of boundary acoustic models (see Table 4.1), corresponding to B1, B4, and a variation of B3 with the STFT analysis window changed to 10 ms. Principal components analysis was used to reduce the dimensionality of each of these individual measurement sets to 50 dimensions. Table 5.9 summarizes the results.

Individually, the three systems achieved 11.3%, 12.0%, and 12.1% word error rate. These results are consistent with the 11.8% reported in [29]. Combining these three boundary models led to a word error rate of 10.1%. This corresponds to word error rate reductions of 10–16%, and substitution error rate reductions of 14–20%. It is encouraging to see that the reduction in substitution errors is similar to the error rate reductions that were observed in the TIMIT phonetic classification experiments.

These results confirm that the techniques proposed in this thesis generalize well to word recognition in a telephone bandwidth acoustic environment. One of the strengths of these methods is that they are extremely general. The experiments in this thesis report only a small sampling of the practically limitless possibilities of system designs which incorporate the same principles: multiple complementary acoustic measurements harnessed through multiple classifiers for improved recognition. This particular set of three measurement sets for use with the JUPITER system was chosen for its diversity, with good results. It is likely that further research will lead to better results as more acoustic measurement combinations are explored.

5.4 Summary

We have shown that heterogeneous measurements can be incorporated using committee-based and hybrid methods to improve the performance of phonetic classification, phonetic recognition, and word recognition systems. Diverse measurements were obtained by using different time-frequency resolutions, different temporal basis vectors, and/or different spectral representations. Empirical results confirmed that these measurements contain complementary acoustic-phonetic information.

Chapter 6

Stop Consonant Classification by Humans and Machines

The stop perception experiments described in this chapter were designed to achieve two goals. The first goal was to provide a benchmark for comparison with machine performance. We did not expect humans or machines to be able to achieve perfectly correct classification because of inherent phonetic ambiguity in the speech signal. Thus, given any particular classification result, it is often difficult to know how much room there is for improvement. Human perceptual experiments alleviate this uncertainty by providing a high benchmark which is known to be achievable based on the acoustic evidence. The second goal of these experiments was to facilitate analysis of the classification errors made by machines, with a view toward postulating how these errors might be avoided in future systems.

The stop consonants for these experiments were extracted from the TIMIT database with no special conditions on the phonetic contexts. This is important in order to consider the full phonetic diversity that occurs in continuous speech.

6.1 Related Work

There are a number of studies in the literature which examine stop perception performance by human listeners, but none has all of the characteristics which are needed to fit the design goals of these experiments. Some studies severely restrict the phonetic context of the stop consonants. In other cases it is not possible to obtain an automatic classification result on the same database for the purposes of comparison. In the "Rhyme Test" and "Modified Rhyme Test," Fairbanks (1958) [24] and House(1965) [36], respectively, worked with limited contexts. Work by Nusbaum (1984) [70], Miller and Nicely (1955) [63], and Clark (1983) [14] all considered CV syllables. Winitz (1972) [92] worked with voiceless stops isolated from conversational speech, but obtaining the database for calculating a corresponding machine result is difficult.

The most relevant perceptual studies are those by Lamel (1988) [41], and Nossair and Zahorian (1991) [69]. Nossair and Zahorian's paper [69] is particularly relevant because it includes comparisons between humans and machines. Unfortunately, however, their experiments were limited to syllable-initial singleton stops. In [41], Lamel reported detailed perceptual results from a variety of contexts. Not all contexts were represented, but some of the stops were from the TIMIT database, which made it convenient to obtain corresponding machine results. In fact, machine comparisons with Lamel's results were reported by this author in [31], and summarized in this thesis in Section 1.3.1. Note that Lamel performed the perceptual experiments, then we performed corresponding machine experiments about ten years later. The experiment was helpful for illuminating the gap between human and machine stop classification performance. However, there were some elements that were awkward because the comparisons with machines were added as an afterthought, instead of being integrated into the experimental design. Out of this experience, we decided to design a joint experiment which would include both perceptual experiments and automatic classification by machine. The jointly designed experiments are the main subject of this chapter.

6.2 Perceptual Experiment Design

6.2.1 Data Set Selection

The selection of specific stop consonant tokens for inclusion in these perceptual experiments was guided by the desire to learn about phonetic classification performance differences between humans and machines. There was also a desire to keep the total number of stops to a small enough number so that the volunteer human subjects would not experience too much inattention or fatigue. The most interesting examples are tokens where humans perform well but machines perform poorly. In order to increase the likelihood of finding these situations, tokens were selected on the condition that they were difficult for machines to classify. Stop consonant candidates were classified by three different automatic stop consonant classification systems. Tokens were selected if they were misclassified by at least one of the systems.

For ease of reference, the three automatic systems used in the token selection process were designated as systems A, B, and C. All three systems used the 462-speaker NIST training set for training. The stop consonant tokens for this perceptual experiment were drawn from the 50-speaker development set in order to avoid any overlap with the standard training set, which would disturb the ease of making comparisons with machines. Table 6.1 indicates the number of tokens of each stop in the training and development sets.

Table 6.2 summarizes the characteristics of the measurements used in these three systems. Systems A and B use measurements which are the baseline and stop-specific measurements, respectively, described in Chapter 3 on hierarchical techniques. The measurements for system C are similar but not identical to measurements used elsewhere in this thesis. System C uses 5 tapered, centered, cosine basis vectors which are 300 ms long. These basis vectors are identical to those used in measurement set "SV" in Table 3.3. This measurement vector was supplemented with average pitch, log duration, log energy, and zero-crossing rate for a total of 64 dimensions.

The acoustic modeling for systems A, B, and C made use of a full-covariance mixture Gaussian classifier with phone priors from the training data (e.g., a phone

Stop	Training	Development
b	2,181	249
d	2,432	239
g	1,191	127
p	2,588	281
t	3,948	413
k	3,794	376
Total	16,124	1,685

Table 6.1: Token counts for training and development sets

unigram). Normalization and principal component analysis were performed to whiten the feature space. For each trial of model training, a maximum of 12 full-covariance Gaussian kernels were allowed per phone. The mixture kernels were seeded via randomly initialized K-means clustering and trained using the EM algorithm. The number of mixtures was selected to achieve a minimum of approximately 500 tokens per kernel for systems A and C, 300 tokens for system B. Multiple trials of model training were aggregated to produce more robust models, as summarized in Section 2.4.1.[1]

	# Dims	STFT [ms]	Spectral Representation	Temporal Basis
A: "SB" in Table 3.3	61	20.5	12 MFCCs	3 averages (3-4-3), 2 derivatives
B: "SS" in Table 3.3	50	10	12 MFCCs	2 averages (halves), 2 derivatives
C:	64	20.5	12 MFCCs	5 tapered cosines 300 ms wide

Table 6.2: Segmental measurements for systems A, B, and C used in design of perceptual experiments. See text for full details.

Table 6.3 summarizes the results for the automatic systems, and shows several subsets of the development data that can be defined according to these results. For these perceptual experiments, we chose to use the 490 stops for which at least one automatic classifier made an error. These tokens are listed on the second-to-last line

[1]Aggregation of acoustic models is described in Appendix C.

of the table, referred to as the "hard+harder+hardest" tokens. There are no special conditions on the phonetic context of these tokens. Appendix A provides a complete list of these tokens along with the human and machine hypotheses collected in these experiments.

Description of token sets		Error Rate (%)			
Number of Classifiers (out of systems A, B, and C) that correctly identified these tokens	Number of stop tokens	System A	System B	System C	Voting using A, B, and C
exactly 3 correct (easy)	1,195	0.0	0.0	0.0	0.0
exactly 2 correct (hard)	223	48.0	23.8	28.3	0.0
exactly 1 correct (harder)	131	76.3	68.7	55.0	93.1
exactly 0 correct (hardest)	136	100.0	100.0	100.0	100.0
hard+harder+hardest	490	70.0	56.9	55.3	52.7
All of the development set	1,685	20.4	16.6	16.1	15.3

Table 6.3: Definition of token subsets according to performance of automatic systems.

6.2.2 Preparation of Acoustic Data

The choice of how much context to provide to the listener in the perceptual experiment is important. The first objective was to make it difficult for humans to use phonotactic or lexical knowledge. To reach this goal, the contexts had to be short. The second objective was to provide approximately as much context as the automatic classification systems use. To reach this objective, the contexts had to be sufficiently long so that they would be comparable to that of the automatic systems which make use of information from the segment boundaries in the measurement extraction process. Obviously, a balance was needed between these two conflicting conditions.

The final design resulted in the following waveform preparation. First, for each stop token, three or four phonetic segments from the TIMIT phonetic transcription files were selected. These three or four segments included the stop itself, one segment after the stop, and either one or two segments before the stop. If the segment immediately preceding the stop was a closure, then two segments before the stop were included. In all other cases, only one segment before the stop was included. Next, 15% of the first and last segments were trimmed away. The purpose of this is to avoid situations where the next sound on either side can be inferred from hearing the phonetic transition. If these sounds could be identified, then as many as five phonemes might be inferred, which would potentially foil the objective of not allowing subjects to make use of lexical knowledge. Next, the beginning and ending 15 ms of each speech segment were tapered in time using a raised cosine window. This was done to avoid the perception of "clicking noises" at the beginning and end of the example. The waveforms were normalized so that the loudness of the examples would be approximately consistent across tokens.

6.2.3 Presentation of Acoustic Data

The 490 tokens were randomly divided into 5 lists of 98 tokens. The first ten tokens were repeated at the end, so that there were a total of 108 examples to respond to in each list. For scoring, the first five and the last five responses were disregarded. The test was administered using headphones at a computer workstation. A response was required for every token before the computer would proceed to the next example. Each example was played twice, with a 1 second interval between first and second playback. Subjects could request to have the example replayed as many times as desired before giving their response. Each subject was asked to take the test at their convenience at their workstation, spreading the lists over several days.

The instructions were:

```
INSTRUCTIONS: Over the headphones you will hear short segments
of speech.  Each example will be played twice.  Your task is to
identify the stop consonant which occurs during this brief speech
segment.  You must enter one of the six choices (p,t,k,b,d, or
g) before the computer will proceed to the next speech segment.
Each list of stops has 108 stops to identify, which normally
takes about 15-20 minutes to complete.

These stop consonants were extracted from continuous speech, and
can occur in any phonetic context.  For example, they may be word
internal, across word boundaries, or at the start or end of a
```

```
sentence.  This variety of context may occasionally make iden-
tifying the stop difficult.  Usually, but not always, you will
hear some brief sounds on either side of the stop.  Generally,
you will be on your own, except for one rule and one hint:

RULE: If you perceive any mismatch between the stop closure and
the stop release (or burst), you should make your choice based
upon your best assessment of the identity of the BURST.

HINT: If the stop is at the very beginning of the speech seg-
ment, then you will be alerted of this fact with the message:
Listen for the stop at the beginning of the segment.  Of the
108 stops in each list, only between 5 and 9 of the stops will
be in this category.

Finally, listen carefully.  In this experiment we would like to
measure the best that humans can do at identifying these stops.

Press any key to start when you are ready to begin.
```

6.2.4 Characteristics of Human Listeners

The test was taken by 7 students and staff in the Spoken Language Systems group at MIT during the summer of 1997. Two of the subjects had no background in speech science. The other subjects had some knowledge of acoustic-phonetics. Any advantage gained from knowledge of speech science is acceptable, since the experiment is designed to measure achievable perceptual performance, not "typical" performance.

6.3 Perceptual Experiment Results

Figure 6-1 indicates the results for each of the seven listeners in the tasks of 6-way stop identification, 3-way place identification, and 2-way voicing identification. The average performance and the performance obtained by voting are shown in Table 6.4 over the same three tasks. The average performance is generated from all $7 \times 490 = 3430$ listener responses. Table 6.5 shows the total listener responses in a confusion matrix. Table 6.6 shows the confusion matrix resulting from the majority vote. In the process of generating the voting results, there were seven cases where a tie needed to be broken. In each case, the reference was favored. This could affect the score by as much as 1.4% (7/490) in stop identification. Only five of these ties involved place

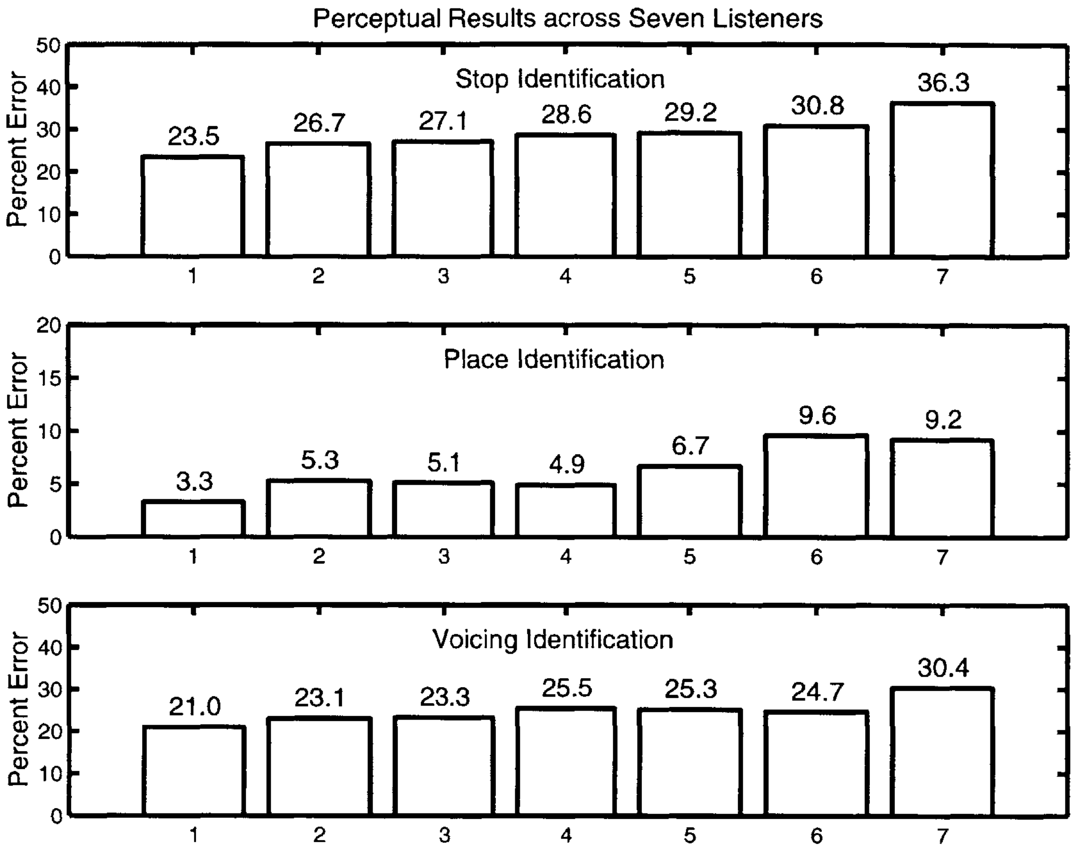

Figure 6-1: Perceptual results across seven listeners.

identification, so the bias toward the reference in tie-breaking could have changed the place-identification score by 1.0% (5/490).

	Task		
	Stop	Place	Voicing
Average Listener	28.9	6.3	24.7
Listener Voting	23.1	2.2	21.2

Table 6.4: Error rate results for the average of the 7 listeners and for a majority vote result from the 7 listeners.

Tables 6.7, 6.8, and 6.9 indicate the McNemar significance of the performance difference between each pair of listeners for each of the three tasks. The "Y," for "Yes," indicates that the difference is statistically significant at the 0.05 level, which implies that there is less than a five percent chance that the two error rates are equivalent.

Answer	Number of tokens	Percent Error	Listener's response					
			b	p	d	t	g	k
b	70	35.5	316	122	18	11	19	4
p	100	19.3	114	593	7	9	3	9
d	68	36.9	7	4	442	219	19	9
t	105	29.8	4	11	137	447	20	18
g	91	25.2	4	1	5	2	356	108
k	56	27.3	4	6	5	17	75	285
Total	490	28.9						

Table 6.5: Human listener confusion matrix for all 490 tokens and all seven subjects, showing the average performance.

Reference	Number of tokens	Percent Error	Hypothesis					
			b	p	d	t	g	k
b	70	31.4	48	19	2		1	
p	105	7.6	8	97				
d	100	36.0	1		64	34		1
t	91	22.0		2	16	71	1	1
g	68	20.6					54	14
k	56	23.2		1		1	11	43
Total	490	23.1						

Table 6.6: Majority vote among human listeners for all 490 tokens.

Conversely, the "N," for "No," indicates that the difference between the two performance levels is *not* statistically significant. An "N" implies that the listener's results are equivalent. Looking at Table 6.7, we see that there is no significant difference between listeners 2 through 6. Listeners 1 and 7 are outliers, performing the best and worst, respectively. The significance results are similar, though not identical, for the place of articulation and voicing tasks. Note that the actual significance levels are given in small print. Choosing a cutoff level of 0.05 for the binary "Y" or "N" decision is arbitrary, but it is helpful in providing a high-level view of the differences between pairs of listeners.

Pairwise McNemar Significance: 6-way Stop Consonant Identification						
	2	3	4	5	6	7
1	N 9.7×10^{-2}	Y 4.1×10^{-2}	Y 4.6×10^{-3}	Y 1.8×10^{-3}	Y 1.9×10^{-4}	Y 5.3×10^{-7}
2		N 9.1×10^{-1}	N 4.0×10^{-1}	N 2.6×10^{-1}	N 5.7×10^{-2}	Y 1.5×10^{-4}
3			N 5.2×10^{-1}	N 3.6×10^{-1}	N 7.9×10^{-2}	Y 5.7×10^{-4}
4				N 8.4×10^{-1}	N 3.2×10^{-1}	Y 3.5×10^{-3}
5					N 4.9×10^{-1}	Y 7.6×10^{-3}
6						Y 4.1×10^{-2}

Table 6.7: McNemar significance of the difference between listeners in stop identification. Listeners are referenced by numerical indices corresponding to Figure 6-1. "Y" or "N" indicates whether or not the difference is significant at the 0.05 level. The exact significance level is given in small print.

Pairwise McNemar Significance: 3-way Stop Place Identification						
	2	3	4	5	6	7
1	N 6.4×10^{-2}	N 1.1×10^{-1}	N 1.7×10^{-1}	Y 3.3×10^{-3}	Y 1.6×10^{-6}	Y 9.0×10^{-6}
2		N 1.0	N 8.7×10^{-1}	N 3.2×10^{-1}	Y 3.1×10^{-3}	Y 9.4×10^{-3}
3			N 1.0	N 2.7×10^{-1}	Y 2.6×10^{-3}	Y 1.0×10^{-2}
4				N 1.8×10^{-1}	Y 1.4×10^{-3}	Y 5.5×10^{-3}
5					N 8.1×10^{-2}	N 1.3×10^{-1}
6						N 9.0×10^{-1}

Table 6.8: McNemar significance of the difference between listeners in identifying the place of articulation of stop consonants. Listeners are referenced by numerical indices corresponding to Figure 6-1. "Y" or "N" indicates whether or not the difference is significant at the 0.05 level. The exact significance level is given in small print.

6.4 Stop Classification using Heterogeneous Measurements and Multiple Classifiers

We now describe a system using heterogeneous measurements which will be compared with the perceptual results. We chose to use segmental measurements S1-S8 and boundary measurements B1-B5, as listed in Table 4.1. The multiple classifiers were combined using an independence assumption. A bigram phonotactic model was used. The segmentation was taken from the TIMIT transcriptions in order to produce a context-dependent classification result of 15.0%. Note that the bigram makes use of

Pairwise McNemar Significance: 2-way Stop Voicing Identification						
	2	3	4	5	6	7
1	N 2.8×10^{-1}	N 2.1×10^{-1}	Y 1.0×10^{-2}	Y 1.4×10^{-2}	Y 4.7×10^{-2}	Y 8.7×10^{-5}
2		N 1.0	N 2.1×10^{-1}	N 2.6×10^{-1}	N 4.3×10^{-1}	Y 1.4×10^{-3}
3			N 2.5×10^{-1}	N 3.2×10^{-1}	N 4.9×10^{-1}	Y 3.8×10^{-3}
4				N 1.0	N 7.4×10^{-1}	N 5.7×10^{-2}
5					N 8.2×10^{-1}	Y 4.3×10^{-2}
6						Y 2.2×10^{-2}

Table 6.9: McNemar significance of the difference between listeners in identifying the voicing of stop consonants. Listeners are referenced by numerical indices corresponding to Figure 6-1. "Y" or "N" indicates whether or not the difference is significant at the 0.05 level. The exact significance level is given in small print.

hypothesized phonetic context. This use of hypothesized context is in contrast to some context-dependent classification work in the literature which may have used *reference* context in the application of the phonotactic model [8, 9]. A context-dependent classification error rate of 17.7% is reported in [8, 9], using a non-standard test set consisting of a 20-speaker subset of the NIST TIMIT training set.

Context-dependent classification performance is a good indicator of the substitution error that will occur in TIMIT phonetic recognition. The 15.0% error result obtained above is the same as or close to the 15.0% and 14.7% substitution error results reported in the last two lines of Table 5.7. This correspondence was also found with other acoustic measurements. This correspondence is important because it validates the use of phonetic classification results for the purpose of improving phonetic recognition.

In order to make direct comparisons with the perceptual results, the Viterbi search was altered. This is necessary since humans were performing six-way classification, whereas normally the recognizer is considering all phone candidates for every segment. Thus, for this purpose, the reference transcription was consulted. If the reference phone was a stop consonant, then the hypothesis was restricted to be one of the six stop consonants. This allowed the phones surrounding the stop bursts (including the stop closures) to be hypothesized over the full phone set, while at the same time restricting the hypotheses for the actual stop bursts. Using this restriction on the stop consonants, the system achieved 14.8% classification error on the development set. The performance on the stop consonants was 11.1%, as shown in Table 6.10.

6.5 Progress in Machine Classification of Stops

Literature/ Experiment	Task		
	stop	voicing	place
Johnson,1996[32]	-	-	11.0
Chun,1996[13]	21.6	-	-
System A	20.4	15.1	7.2
Zahorian,1996[90]	18.8	-	-
System C	16.1	10.9	6.5
Voting: A, B, and C	15.3	-	-
Heterogeneous	11.1	7.8	3.8

Table 6.10: Stop consonant identification, voicing, and place identification error rates on the development set.

Table 6.10 summarizes progress in the automatic classification of stop consonants. Conveniently, there are three results from the literature which report on the same training and development sets as used in this thesis. The system reported by Johnson [32] with a place identification error rate of 11% makes use of only 10 manually extracted formant and burst spectral measurements. The classification was performed using linear discriminant analysis. Johnson found that the place identification error using this small set of measurements degraded significantly to 24% when these measurements were automatically extracted. The system reported by Chun [13] uses measurements identical to those used in system A, shown in Table 6.2, except that Chun may have averaged the MFCCs literally over thirds of the segment rather than a 3-4-3 proportion. Chun's results are considerably worse than system A because system A makes use of the aggregation of multiple trials in the classifier, as mentioned in Section 6.2.1, resulting in a model which is larger and considerably more accurate. Zahorian's results reported in [90] made use of a frequency-dependent time-warping in the acoustic measurements and a binary-pair partitioned neural network classifier. The "Voting: A, B, and C" line in Table 6.10 indicates the performance of obtained by voting with systems A, B, and C.

The heterogeneous result is not directly comparable to the other machine system results since it includes the addition of a bigram and context-dependent models. However, the purpose in this section is not to separate the contributions of heterogeneous measurements, the bigram, and the context-dependent models. The purpose, rather, is to obtain a system which is suitable for comparison with the human perceptual results. Ideally, the human and machine systems would have access to identical information. It seems reasonable that humans would make use of phonotactic knowledge. Therefore, a bigram is included. The precise conditions which would give the humans and machines access to the same information are, in fact, difficult to determine and subject to debate, especially when one considers the possible intrusion of higher-level phonotactic and lexical knowledge. In summary, the heterogeneous system is a reasonable candidate for comparison with humans.

6.6 Narrowing the Gap: Human versus Machine Performance

Figure 6-2 shows the comparison of 4 machine systems versus the average listener performance and the majority vote listener performance. Machine systems A, B, and C were used in the data set selection. The fourth machine is the system using heterogeneous measurements discussed in Sections 6.4 and 6.5. Consider first the comparison of machine systems A, B, and C with the human listeners. It can be seen that automatic stop place identification is 3.5 to 10 times worse than that of humans. At the beginning of this investigation, this result provided motivation, since it indicates that there is significant room for improvement in automatic identification of place of articulation. In contrast, voicing identification by machine is only 1.5 to 2.4 times worse than listeners. The net result of these two kinds of errors (place and

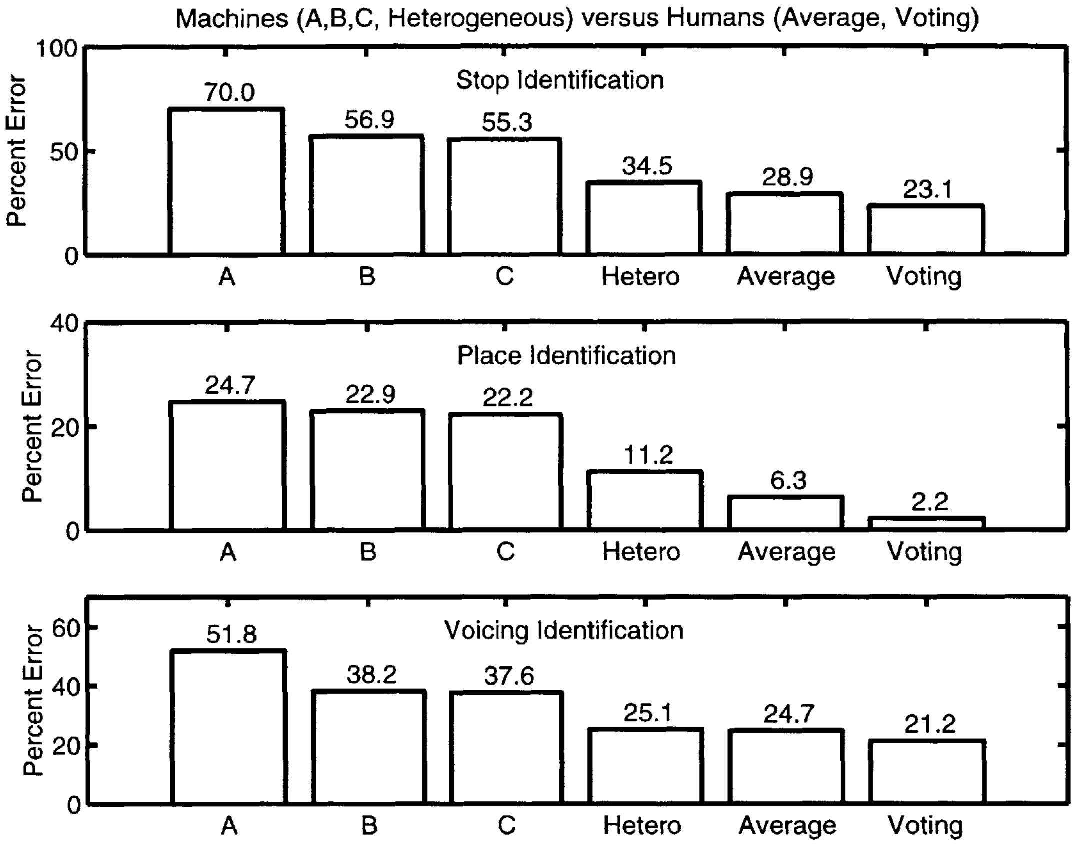

Figure 6-2: Comparing 4 machine systems (A, B, C, and heterogeneous measurements) with human listeners (average and voting).

voicing) leads to 6-way stop identification by machine that is 1.9 to 3.0 times worse than listeners. All of these results indicate that much improvement of machines is needed in order to reach human levels of performance.

The heterogeneous measurements system produces error rate reductions of 38%, 50%, and 33% relative to system C in the tasks of stop, place, and voicing identification, respectively. Due to those very significant improvements, the heterogeneous measurements system is now much closer to the human listener results. Place identification is 1.8 to 5.1 times worse, voicing identification is only 1.0 to 1.2 times worse, and 6-way stop identification is 1.2 to 1.5 times worse.

Figure 6-3 compares the heterogeneous measurements system with each individual listener. Table 6.11 shows the McNemar statistical significance of the difference between the heterogeneous measurements system and the individual listeners for each

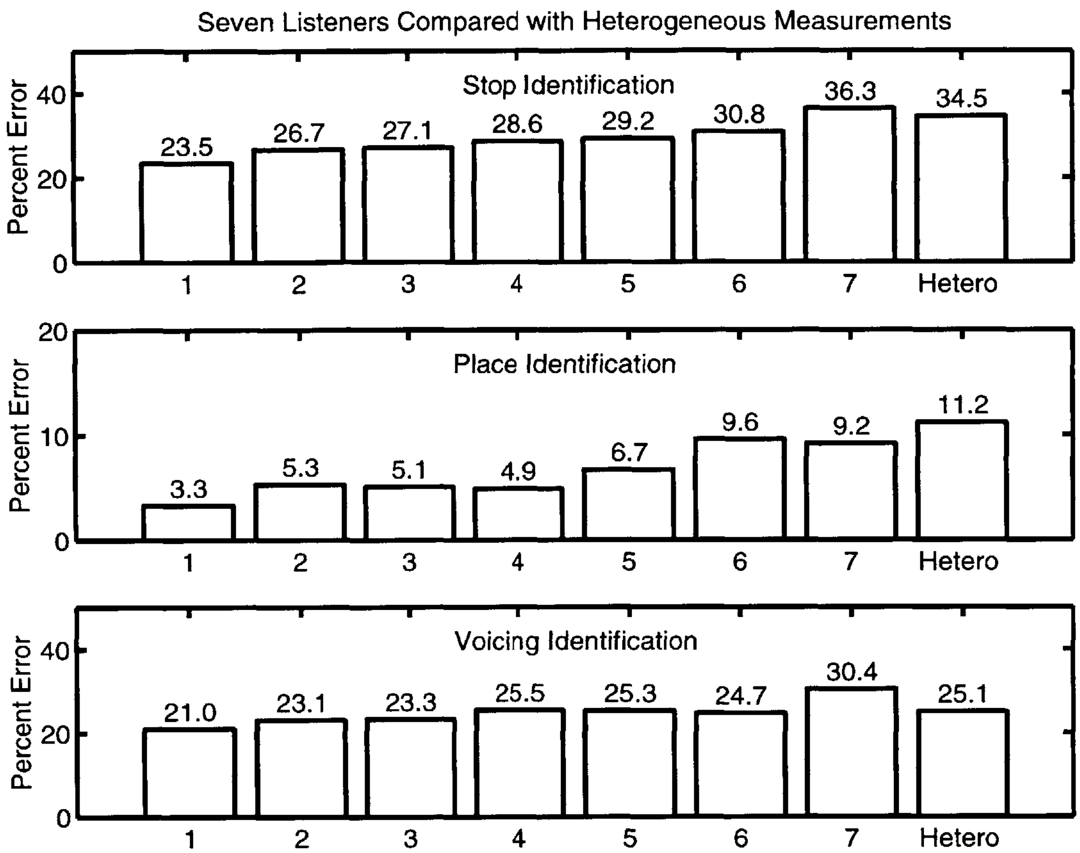

Figure 6-3: Comparing seven individual listeners with heterogeneous measurements machine result. Note that the machine is sometimes better than the listeners.

of the three tasks. In the voicing identification task, statistical significance tests show that the machine is equivalent to 5 of the listeners, better than the worst listener, and worse than the best listener. Thus, all in all, the machine is the same or better than 6 out of 7 listeners in voicing identification. The voicing identification performance of listener number one is 21.0%. This is the best result. Even the majority vote perceptual result was no better, at 21.2%. An error rate reduction of only 15% in the machine system would result in performance equal to the perceptual voting system. Thus, the gap between human and machine voicing identification performance under these conditions is very small, but not entirely eliminated.

There is still a significant gap between human and machine place identification performance. The machine is never better than any of the listeners, but significance tests indicate that there is no significant difference between two of the listeners and

Speaker #	McNemar Significance					
	Stop		Voicing		Place	
1	Y	6.8×10^{-6}	Y	5.9×10^{-2}	Y	1.2×10^{-6}
2	Y	1.8×10^{-3}	N	3.8×10^{-1}	Y	6.4×10^{-4}
3	Y	3.6×10^{-3}	N	4.5×10^{-1}	Y	2.3×10^{-4}
4	Y	1.6×10^{-2}	N	9.2×10^{-1}	Y	1.9×10^{-4}
5	Y	4.2×10^{-2}	N	1.0	Y	1.4×10^{-2}
6	N	1.6×10^{-1}	N	9.2×10^{-1}	N	4.4×10^{-1}
7	N	5.7×10^{-1}	Y	5.1×10^{-2}	N	3.3×10^{-1}

Table 6.11: McNemar Significance of difference between automatic system using heterogeneous measurements and each of seven human listeners for each of three tasks. "Y" and "N" indicate presence or absence of a significant difference at the 0.05 level.

the machine system. This achievement is less important due to the fact that there is a very wide range in place identification performance among the listeners. Listener number 6 is 2.9 times worse than listener number 1. The perceptual majority vote result of 2.2% error is the best. An error rate reduction of 80% is still required to make the machine performance equivalent to the voting system, and this is in spite of the fact that the heterogeneous measurements system has already benefited from a voting-like procedure through the use of multiple classifiers.

The 6-way stop identification task is, or course, a blending of the voicing and place results discussed above. The heterogeneous measurements system achieves performance which is statistically equivalent to listeners 6 and 7. The listener voting result of 23.1% is only slightly better than the best individual listener, 23.5%. An error rate reduction of 33% is required to make the machine system achieve the same performance as the perceptual voting result. Although this sounds promising, the analysis above shows that the actual situation is more difficult, since voicing identification is comparable, but place identification has a lot of ground to cover in order to fill in the gap in performance.

The interpretation of results comparing human and machine performance is subject to some debate due to varying viewpoints regarding the equity of the comparison. Phonetic classification of excerpts of speech is an unnatural task for humans, which is a disadvantage for humans. On the other hand, the machines in these experiments have access to only about 16,000 examples of stop consonants for training, whereas the human subjects have listened to millions of examples over the course of their 20 or more years of life. Future work might consider the effects of giving humans task-specific training, or training a machine classifier with millions of examples. The results obtained here are conclusive within the scope of the parameters of the experiment.

In summary, the heterogeneous measurements system achieves voicing identification equivalent to 6 out of 7 listeners. Place and stop identification were equivalent to 2 out of 7 listeners. Future work in stop consonant identification should continue to focus on the place identification problem in order to achieve human-like performance.

6.7 Examining Place of Articulation Errors

The results from the previous section show that place of articulation errors are the biggest source of difference between human and machine performance. In this section, particular examples of errors are examined. For the purpose of this analysis, the performance of the heterogeneous measurements system will be contrasted with the perceptual majority vote result. The results will be generically referred to as from "humans" or "machines," for ease of reference.

The errors can be divided into three categories based on whether they were incorrectly classified by humans only, machines only, or both humans and machines. Out of the 490 tokens in the data set, both humans and machines were correct for 427 tokens. There were 52 tokens where only the machine was wrong, 8 where only the humans were wrong, and 3 where both were wrong. These errors are examined in more detail in the sections which follow.

6.7.1 Humans correct, Machine incorrect

Table 6.12 lists the 52 tokens where the perceptual majority vote had the correct place of articulation, but the heterogeneous measurements system was incorrect. A large number of these are situations where the phonetic context may have contributed to the difficulty. More specifically, 22 have adjacent liquids or glides, 14 have adjacent fricatives or aspirants, 6 have adjacent nasals, and 6 are phonemic double-stops, leading to a mismatch between the closure and burst. Some of these contexts occur together, such that there are 10 additional stops which don't fall into any of the above-mentioned categories. Of these, 4 occur at the beginning of the sentence, one occurs at the end, and that leaves just 6 that are simply intervocalic singleton stops.

One way to categorize the machine errors is to see if any humans made the same error. Recall that the tokens in the list being considered here are those for which the heterogeneous measurements system is always wrong, and at least one of the machine systems A, B, or C must also be incorrect since that was a condition for inclusion in this perceptual experiment. The remaining 2 machine hypotheses may be correct or incorrect. In order to characterize the performance, each of the four machine hypotheses can be placed into one of three categories. First, the hypothesis may be correct. Second, the hypothesis may be incorrect, but at least one human may have made the same error. Third, the hypothesis may be incorrect, and no human made the same error. Using this division, the hypotheses were 29% correct, 21% with errors in common with humans, and 50% errors not in common with humans. The errors not in common with humans are of the greatest concern. In fact, there are 11 cases where all four machine hypotheses were incorrect and no humans made the same errors. These are numbers 7, 11, 13, 16, 26, 31, 32, 34, 40, 42, and 50. If we include one case with context of [hv], then all of these tokens have a context of a fricative, a liquid, or a glide. This result indicates that these contextual situations create a great deal of difficulty for machines while at the same time humans do not have the same problem. Consider the spectrograms of several of these cases:

- Figure 6-4 shows number 7 from Table 6.12. The machines substituted [t] or [k] for the [p] at 1.78 seconds in "syrup for." The machines probably had difficulty because the time resolution of most of the automatic measurement algorithms was such that the frication from the [f] could interfere with the acoustic evidence for the [p] burst.

- Figure 6-5 shows number 11 from Table 6.12. This realization of the /k/ at 0.45 seconds in "include" lacks a true velar closure because of the [n] on the left. Thus, the acoustic evidence for the [k] occurs in and after the burst. The extreme formant locations of the [l] result in a [k] burst which is atypical with respect to all [k] bursts, and thus difficult for the machine.

- Figure 6-6 shows number 16 from Table 6.12. The high second formant in the [iy] and the low third formant in the [r] produce an unusual [t] at 1.1 seconds in "retract." There is a larger proportion of energy in the region of the second and third formant than is typical for all /t/ bursts. This change in the energy spectrum made the [t] look somewhat like a [k]. The machines made this error, but the humans were able to account for the context and know this was a [t].

- Figure 6-7 shows number 32 from Table 6.12. The machines mistook the [k] at 1.45 seconds in "panic they" for a [d]. This situation is similar to number 7 in Table 6.12 in that acoustic evidence from a right context fricative causes machines to give the place of articulation as alveolar. Humans did not have any problem with this. It seems that the machine is not able to perform the burst analysis in a sufficiently short-time manner.

After looking at these examples, two hypotheses emerge which may help to explain the remaining gap between human and machine place of articulation identification performance. The first hypothesis is that extreme formant locations in neighboring phones produce stop tokens which are statistical outliers. The classifiers in this study typically do not have enough training data to robustly model the "edges" of the acoustic space. This problem is made more acute when, for example, an outlier of a [t] is similar to a typical [k], as in number 16 in Table 6.12. If this is simply an issue of training data, then these problems with outliers will be alleviated through the addition of more training data. The second hypothesis is that the machines are not able to accurately measure and model the extremely short-time phenomena at the stop burst which are sometimes critical to correct identification of a stop consonant. The phonemic double stops in Table 6.12 support this hypothesis. The humans

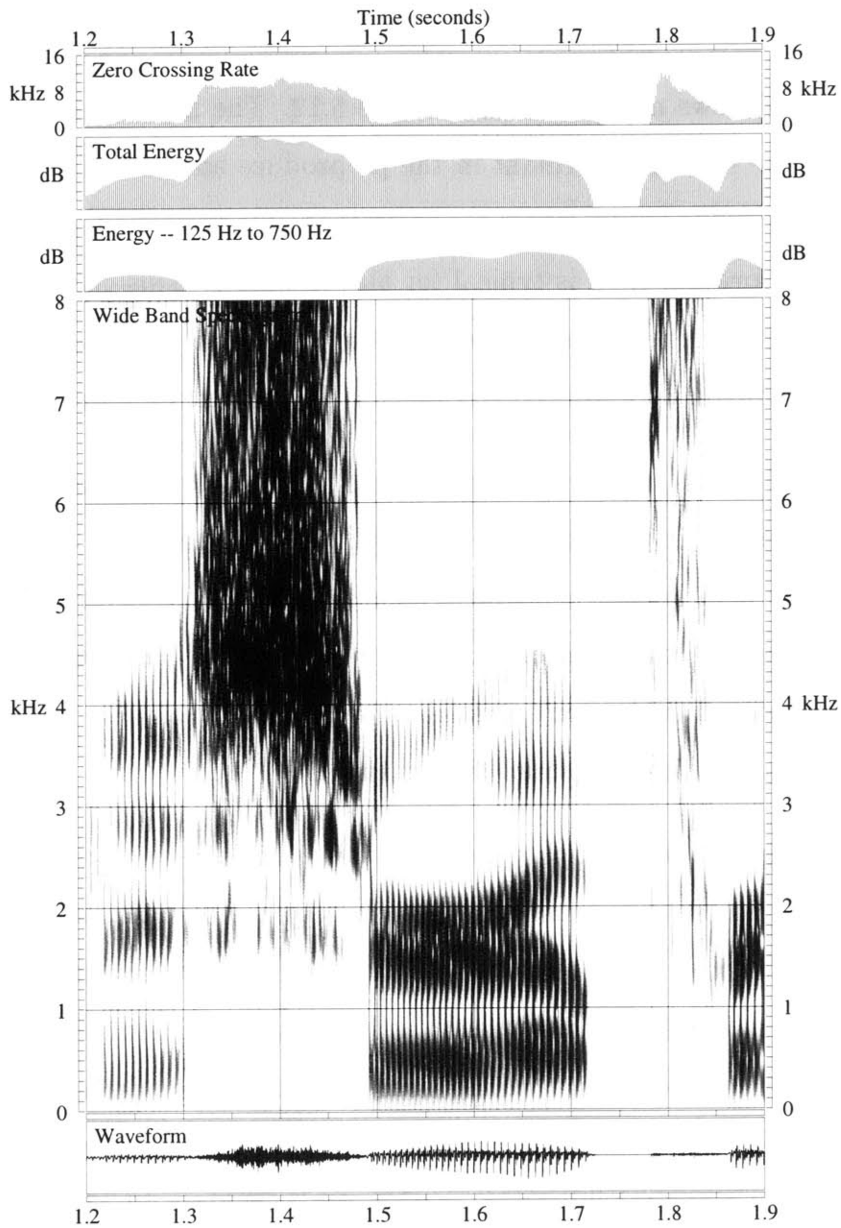

Figure 6-4: This is a spectrogram of "syrup for," corresponding to number 7 from Table 6.12. The burst of the [p] is located at 1.78 seconds. The machines substituted [t] or [k] for the [p]. This was probably due to the effects of the neighboring [f].

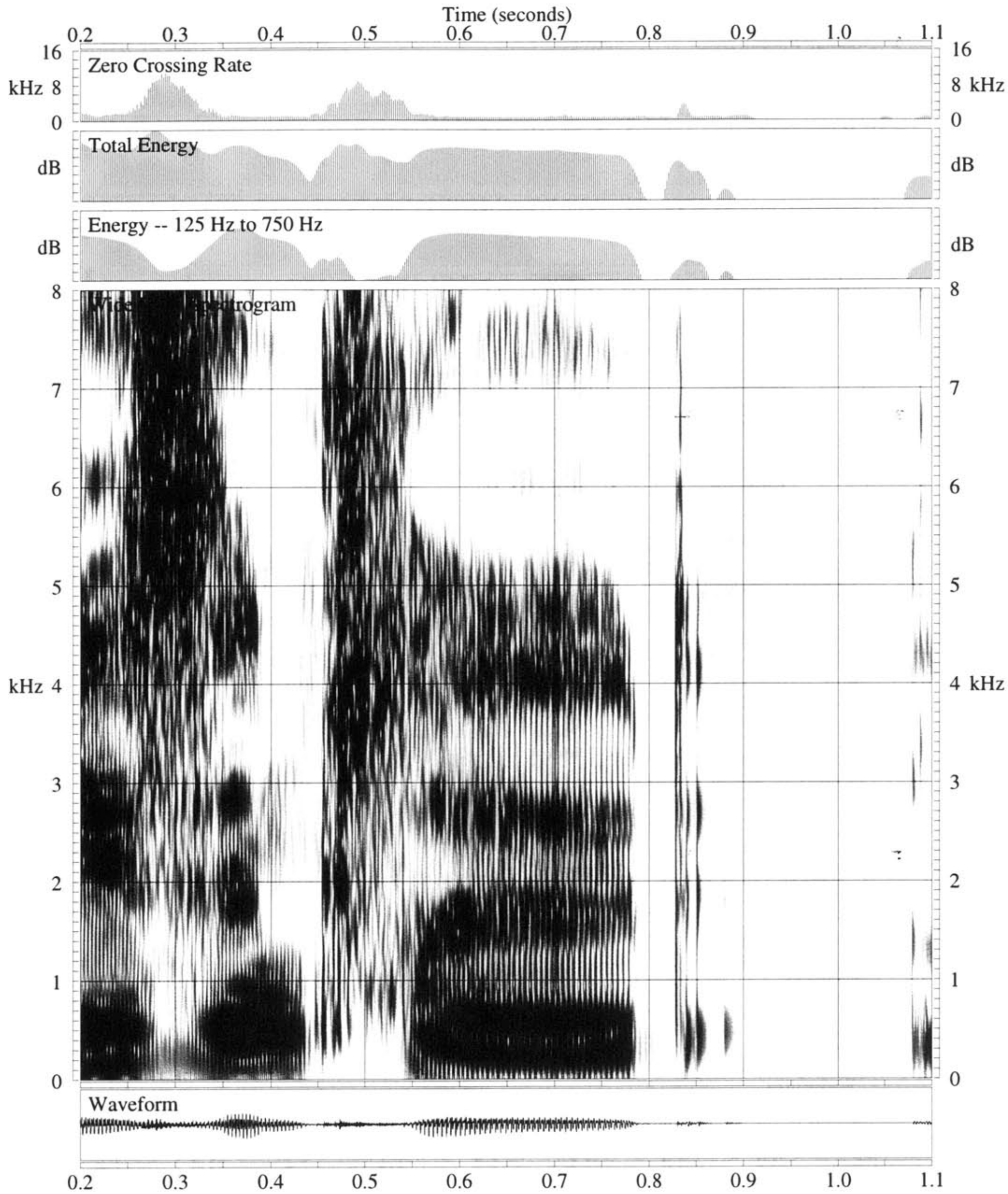

Figure 6-5: This spectrogram of "include," corresponding to number 11 from Table 6.12, shows the lack of a true velar closure because of the [n] on the left of the [k] burst at 0.45 seconds.

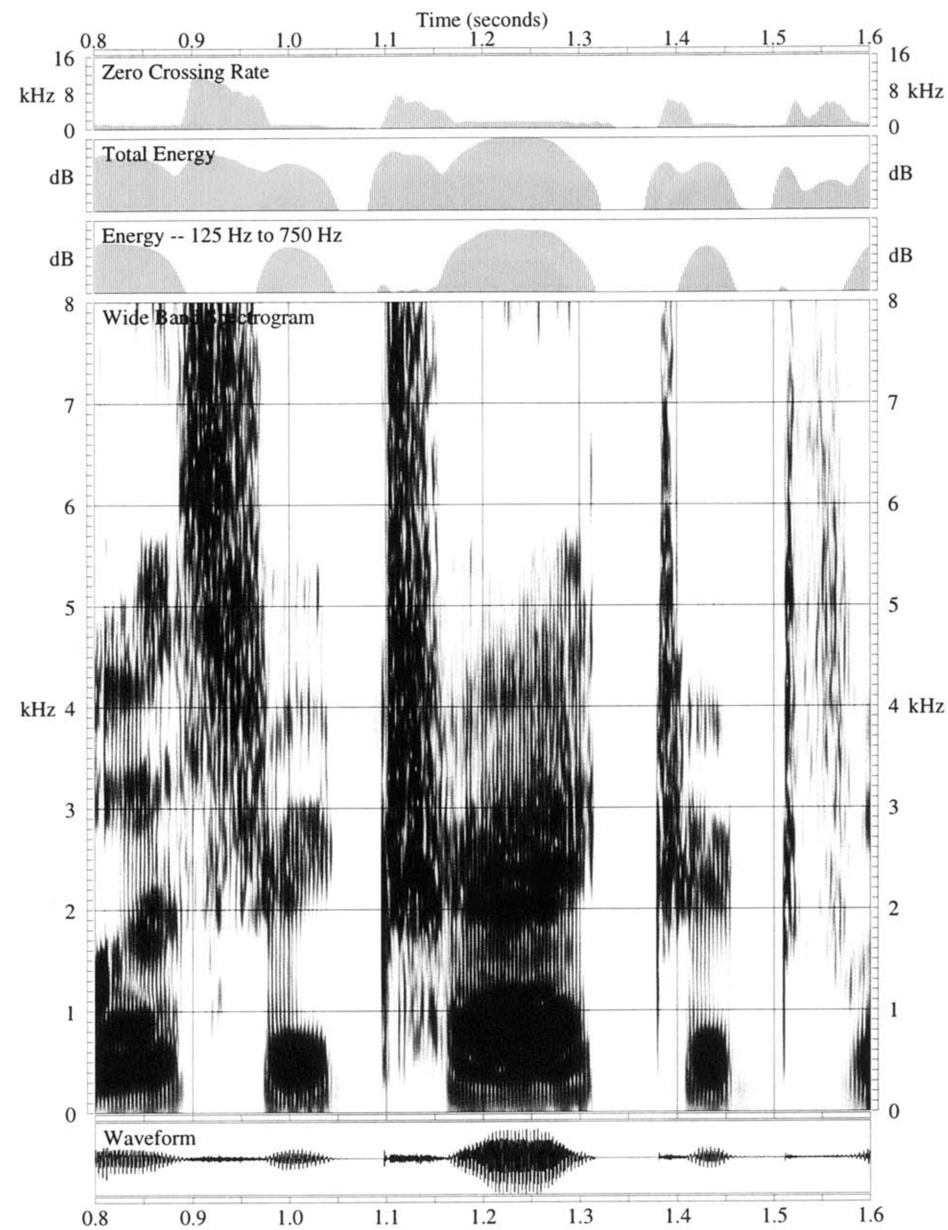

Figure 6-6: Spectrogram of "retract," corresponding to number 16 in Table 6.12. The /t/ within the word "retract" is the sound of interest. The high second formant in the [iy] and the low third formant in the [r] produce an unusual [t], at about the 1.1 second mark, that is difficult for the machines to classify.

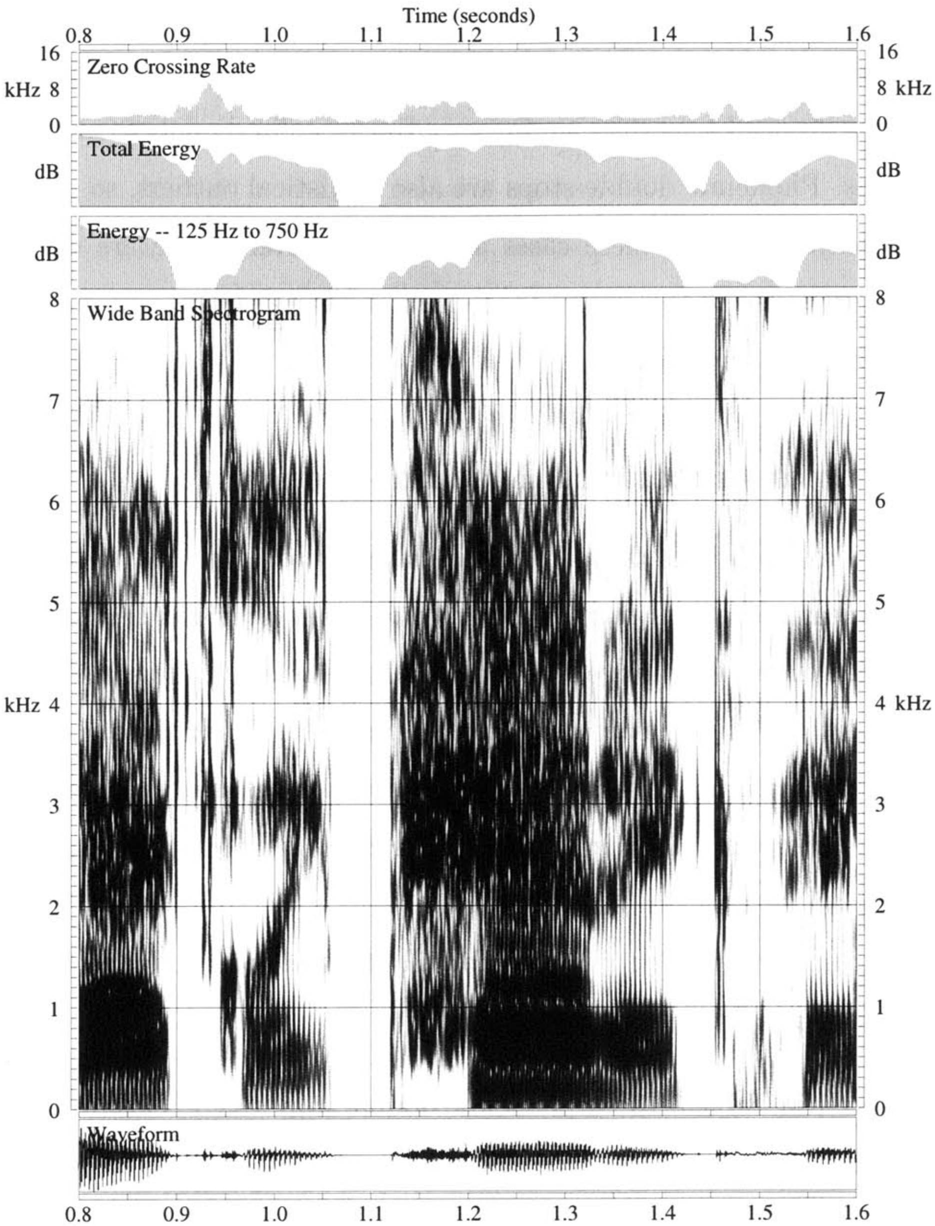

Figure 6-7: Spectrogram of "panic they," corresponding to number 32 in Table 6.12. The machines mistook the [k] at 1.45 seconds for a [d]. Acoustic evidence from the right-context fricative may have been the cause of the migration to an alveolar place of articulation. It seems that the machine was not able to perform the burst analysis in a sufficiently short-time manner.

were able to recognize the mismatch between stop closure and stop burst information and choose the place of the burst. The machine systems were generally not able to separate the acoustic evidence from the burst and the closure. For example, the acoustic measurements for the closure include information from both before and after the closure. Thus, the burst acoustics are included with the closure acoustics. Machine systems need a mechanism for very time-localized analysis at burst and closure locations. Phonemic double-stops are also statistical outliers, so clearly the first hypothesis plays a role in these cases as well. However, there are other examples where the chief problem for machines seems to be the short-time nature of the most important acoustic evidence.

6.7.2 Humans incorrect, Machine correct

There were 8 tokens where the perceptual majority vote was incorrect, but the heterogeneous measurements system was correct. In the previous section, the severity of the machine errors was gauged by whether humans made the same errors. In this case, the severity of the human errors can by gauged by looking at how many of the humans were incorrect. It turns out that there is only one example where all the humans were incorrect. Thus, in 7 out of these 8 cases, at least one human did select the correct place of articulation.

Spectrograms will not be provided for all of these tokens, but several examples will be discussed:

- Number 1 in Table 6.13: A phonemic double stop, the [k] closure confused some listeners. The non-speech scientists who took the test may not have been clear about the distinction between listening for the burst identity as opposed to the closure. In fact, 3 out of 4 listeners were correct, and there are 6 examples of phonemic double stops in Table 6.12 where the majority of humans were correct. Thus, it seems likely that the incorrect listener hypotheses here could have been from misunderstanding or fatigue.

- Number 2 in Table 6.13: This instance of /b/ in the word "ability" sounds and

	# in Appendix A	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
1	31	si497-b-mers0	ah pcl p ax-h s	the puppets delight	bppt	ppppbp
2	35	sx44-b-mbwm0	axr tcl t ey n	an entertaining sport	ttkk	ttttttt
3	39	sx403-b-mjfc0	r dcl d r ow	her wardrobe consists	gggg	dgdddgd
4	42	sx413-b-mdlf0	# h# d ae f	# daphne's swedish	bddb	ddddddd
5	47	si1454-b-fcmh0	ay pcl p sh y	main types used	pkkt	bbbbbbb
6	56	si1555-b-mmdm2	iy pcl p iy v	a sleepy voice	ppkk	ppbpppp
7	63	si728-b-mjar0	ah pcl p f axr	this syrup for	ttkt	ppppppp
8	74	si1542-b-mpdf0	# h# d ih f	# the feet	dbpb	ddddtdd
9	75	si1542-b-mpdf0	iy tcl t w ao	the feet wore	dddg	tgttttt
10	96	si1628-b-mdls0	r bcl d w ey	an absorbed way	kbbb	dgdgddd
11	101	si860-b-fkms0	n kcl k l ux	these include leaves	tttt	kkkkkkk
12	114	sx298-b-fsem0	n bcl b iy ow	than be overcharged	dbdd	bdbbbbb
13	121	sx389-b-fmah0	ih gcl g hv aw	a big house	dddd	gkkkgkg
14	131	si1084-b-fedw0	ey pcl p s iy	other shapes evolve	dpdt	bpppbbp
15	146	sx113-b-mbdg0	s kcl k ix l	a muscular abdomen	dgdt	ktkkkkk
16	165	sx119-b-fmah0	iy tcl t r ae	was retracted with	kkkk	ttttttt
17	166	sx119-b-fmah0	ae kcl t axr dcl	was retracted with	tktk	tdttdtt
18	176	si859-b-mrjm4	n kcl k en tcl	non contributory plan	gkkd	kkkkkkk
19	179	sx396-b-mdvc0	iy pcl p f r	to leap frantically	ttdt	tpbpppp
20	183	sx279-b-mgwt0	ow tcl t w axr	is noteworthy #	pppk	pdtdkdt
21	194	sx313-b-faks0	# h# d r aa	# drop five	tddb	ddddddd
22	203	sx383-b-fdrw0	ah gcl g h# #	oriental rug #	kkkd	ggggggg
23	207	sx298-b-mmjr0	n bcl b iy ow	than be overcharged	ggbd	ppdbdbp
24	233	sx113-b-fdrw0	uh dcl d f axr	is good for	kdkg	tgddddd
25	238	si1181-b-fjmg0	uh kcl k ih z	present book is	kttt	ggggggg
26	241	si2255-b-mmdb1	f tcl t iy n	some fifteen feet	kkkk	ttttttt
27	245	si494-b-mwjg0	s tcl t w ah	at least one	pptk	ttttttt
28	295	sx101-b-fjmg0	# h# k ih n	# kindergarten children	kpkt	kkkkktk
29	299	si1670-b-fmml0	# h# g aa q	# got no	kbpp	gggggkg
30	306	sx228-b-fdms0	uh kcl k dh iy	viewpoint overlooked the	kptp	bgttgkg
31	333	sx379-b-mcsh0	z dcl d r ay	gives driving lessons	gggp	dttdtdd
32	335	si1502-b-fdms0	ih kcl k dh ey	exactly panic they	dddd	kkkkkkk
33	338	sx110-b-frew0	aa bcl b l ax	extra problems #	gbbd	bbbbbbb
34	343	sx208-b-fsem0	aa bcl b s eh	# rob sat	tddd	bbbbbbb
35	346	si818-b-fgjd0	ix tcl p ao r	transmit poorly and	tppt	ppppppp
36	352	si1133-b-mteb0	eh kcl t l ao	can project long	kkkk	tttktkt
37	382	sx143-b-mdlf0	ux gcl g el s	jennifer's bugle scared	gbgb	ggggggg
38	410	sx28-b-fsem0	ae tcl g aa r	that guard for	ggdd	ggggggg
39	412	sx28-b-fsem0	v gcl g ae s	of gas #	kttt	dgggggg
40	417	sx102-b-mmdm2	ae pcl p axr s	from kidnappers #	gggg	bbbppdt
41	419	sx47-b-mroa0	s tcl t r ao	mom strongly dislikes	kktk	ttttttt
42	428	si677-b-mroa0	el tcl t r eh	comparable trends can	kkkk	ttttttt
43	442	si1910-b-frew0	ix n b r iy	and breathed for	bgbg	bbbbbbb
44	449	sx233-b-mrcs0	ng gcl g h# #	same thing #	dddd	gkdggdg
45	455	sx413-b-mteb0	iy dcl d ix sh	daphne's swedish needlepoint	ggdg	ddddddd
46	461	sx365-b-mmdb1	ae tcl g ah n	at gunpoint for	tgdt	gggggtk
47	465	sx29-b-mers0	ax-h pcl p ax-h tcl	of potatoes #	pttt	pptpttp
48	466	sx29-b-mers0	ax-h tcl t ey dx	of potatoes #	ttpp	ttttttt
49	470	si2275-b-mthc0	ih dcl d l uh	all did look	bddb	ddgdddd
50	471	si2275-b-mthc0	ih pcl p sh ey	and shipshape #	kggk	bpppppp
51	484	si459-b-mmwh0	n tcl t el hh	for experimental hospitals	tddp	ttttdtt
52	486	sx282-b-mrjr0	axr gcl g aa tcl	fairy forgot to	ggdd	ggggggg

Table 6.12: 52 tokens where perceptual majority vote had the correct place of articulation but heterogeneous measurements system was incorrect.

looks confusable with a [d]. Perhaps the speaker made an alveolar closure before closing the lips. Both machine and human labellings are mixed.

- Number 3 in Table 6.13: This sentence-final [t] in "employment" is very weak in energy. This may explain in part why some listeners confused it with [p] in the absence of prosodic context. All automatic systems chose alveolar place of articulation.

- Number 4 in Table 6.13: Some humans thought that this [d] in "reads" was a [b]. This is apparently related to the high-front vowel left context obscuring formant motion toward alveolar place, combined with a very short voice onset time moving into the [z]. The machine systems all had the correct place of articulation.

- Number 5 in Table 6.13: The alveolar place of the left context [z], and the right context reduced vowel [ix] result in faint or missing formant transition cues for labial place. Faint cues still favor [b], but [d] is a very close second choice. Both human and machine hypotheses are mixed.

- Number 6 in Table 6.13: The acoustic evidence for [k] in the word "strength" in this example is almost non-existent. This could perhaps be better labeled as a [kcl], without a stop burst, followed immediately by a [th]. None of the humans chose velar place. Some chose alveolar, and others chose labial. Three out of the four machines were also incorrect, which is understandable based on the lack of evidence.

- Number 7 in Table 6.13: The [t] at the end of "viewpoint" was mistakenly labeled [g] by 5 out of the 7 listeners. Perhaps the combination of formant motion in the /oy/ and the presence of a glottal stop after the /t/ led listeners to choose velar place. The machines all chose alveolar place.

- Number 8 in Table 6.13: Three out of 7 of the listeners were correct, but the other four mistook the second [t] in "exultantly" for a [p]. The combination

	# in Appendix A	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
1	60	si649-b-fadg0	eh kcl t f er	and respect for	dkkt	kkkttkt
2	77	sx279-b-fgjd0	ax bcl b ih l	alice's ability to	bbdb	tdbbddd
3	94	sx229-b-mrjm4	n tcl t h# #	for employment #	tddd	dpppdpp
4	135	sx278-b-mdls0	iy dcl d z dh	occasionally reads the	ttdd	bddbdbb
5	161	si2119-b-mrjm4	z bcl b ix n	has been lost	ddbb	dbbdddt
6	273	sx409-b-mrjm4	ng kcl k th m	increases strength miraculously	pddk	tptppdp
7	305	sx228-b-fdms0	oy n t q ow	the viewpoint overlooked	dddd	ggdgdgg
8	368	si1496-b-mrws1	ix n t l ix	was exultantly easy	kdtd	tpppddp

Table 6.13: 8 stop consonant examples where the majority vote of the humans gave the incorrect place of articulation, but the heterogeneous measurements system had the correct place of articulation.

of nasal and liquid context along with a very short voice onset time may have mislead listeners. One machine system chose velar place.

In summary, the human errors are usually due to acoustic-phonetic ambiguity, or a combination of misunderstanding and fatigue in identifying a phonemic double-stop. There was only one case where all the humans were wrong, and in that case the acoustic evidence was simply absent (number 6). Thus, humans are making very efficient use of the available information.

6.7.3 Humans incorrect, Machine incorrect

There were 3 tokens in which both the perceptual majority vote and the heterogeneous measurements system were incorrect. Consider each of these cases:

- Number 1 in Table 6.14: The [y] in the right context led some humans and machines to mistake the [b] in "vocabulary" for a [g]. The hypotheses of both humans and machines were mixed.

- Number 2 in Table 6.14: The [d] in "seemed foolish" has ambiguous acoustic cues. There is a concentration of energy at the burst that makes it look possibly velar. Both human and machine hypotheses were mixed.

- Number 3 in Table 6.14: The initial stop in "combine" was unanimously recognized by humans as [t]. This is probably due to misarticulation by the speaker.

	# in Appendix A	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
1	55	sx53-b-mrcs0	ae bcl b y ax	simple vocabulary contains	gddg	gbdgggg
2	256	si2203-b-faks0	m dcl d f uw	dive seemed foolish	dkdk	gpgkkkt
3	340	sx118-b-mmjr0	# h# k ax bcl	# combine all	tttp	ttttttt

Table 6.14: 3 stop consonant example where both the majority vote of the humans and the heterogeneous measurement system had the wrong place of articulation.

This example will be examined again in the section concerning the role of phonetic ambiguity.

In summary, these situations were caused by difficult contexts, ambiguous phonetic cues, and/or misarticulation.

6.8 Role of Phonetic Ambiguity

Phonemes are linguistic categories which are the building blocks of the lexicon. In contrast, phones are acoustic categories which are used by linguistics to describe the actual sounds which occur in speech. For example, consider the question "Did you?" which can be transcribed phonemically as /d ih d y uw/. This phrase might be realized phonetically as [d ih dcl d y uw] or [d ih dcl jh ax].

There are times when the phonetic evidence conflicts with the phonemic representation in the lexicon. In this study, these instances were found when all human listeners agreed about the identity of the phone, but their unanimous conclusion disagreed with the reference transcription. Table 6.15 lists the 27 tokens that had this experimental outcome. In general, in cases with incomplete or conflicting acoustic evidence, the TIMIT reference transcriptions are biased toward the phonemic category. This is in spite of the fact that the TIMIT reference transcriptions were created with the goal of obtaining a true phonetic transcription. The perceptual results are more likely to produce the true phonetic category, since they are based only on acoustic evidence. Thus, these examples show conflicts between phonetic and phonemic categories.

	# in Appendix A	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
1	3	sx290-b-mjsw0	s bcl b r ow	this brochure is	bppp	ppppppp
2	32	si497-b-mers0	s dcl d ax l	puppets delight the	tttt	ttttttt
3	47	si1454-b-fcmh0	ay pcl p sh y	main types used	pkkt	bbbbbbb
4	65	sx191-b-fjmg0	th dcl d r ao	# withdraw all	ttgt	ttttttt
5	84	si2120-b-fkms0	z dcl d pau dh	he paused then	tttt	ttttttt
6	125	si1994-b-mtdt0	z tcl t ix gcl	was to get	dddd	ddddddd
7	128	si773-b-fcal1	s dcl d r ay	projects dry slowly	tttt	ttttttt
8	149	si909-b-mgwt0	n tcl t eh tcl	compulsory retirement at	dddd	ddddddd
9	180	sx319-b-mrjm4	ow tcl t q ay	big goat idly	dddt	ddddddd
10	186	sx364-b-fjem0	iy tcl t ix ng	a heating system	dgdd	ddddddd
11	206	sx298-b-mmjr0	aa tcl b ay dh	not buy these	pppp	ppppppp
12	208	sx298-b-mmjr0	jh dcl d h# #	be overcharged #	ttdt	ttttttt
13	209	si2064-b-mteb0	z gcl g ih f	was gifted with	ggkg	kkkkkkk
14	238	si1181-b-fjmg0	uh kcl k ih z	present book is	kttt	ggggggg
15	246	si1653-b-fedw0	ow kcl k s uw	he spoke soothingly	dkkk	ggggggg
16	261	si1946-b-majc0	v dcl d h# #	means solved #	tttt	ttttttt
17	270	sx214-b-fdac1	n dcl d ax-h pau	web glistened in	tttt	ttttttt
18	313	sx379-b-fadg0	s dcl d r ay	gives driving lessons	tttt	ttttttt
19	340	sx118-b-mmjr0	# h# k ax bcl	# combine all	tttp	ttttttt
20	348	sx102-b-mpdf0	ae pcl p axr z	from kidnappers #	gggp	bbbbbbb
21	350	sx109-b-fadg0	ng kcl k el q	her ankle on	kgkk	ggggggg
22	401	si2293-b-mjfc0	eh tcl t q ao	ones that all	dddd	ddddddd
23	403	si2293-b-mjfc0	eh tcl t l aw	laughed at loudest	dbdd	ddddddd
24	421	sx14-b-fcmh0	z dcl d ih z	before thursday's exam	tttt	ttttttt
25	431	sx189-b-fgjd0	# h# d ih s	# destroy every	ttdt	ttttttt
26	453	sx317-b-mroa0	ih tcl t iy ih	very pretty in	dddd	ddddddd
27	463	si904-b-mglb0	n tcl t ix v	came into view	dddd	ddddddd

Table 6.15: Tokens with questionable TIMIT reference transcriptions.

When Table 6.15 is examined more closely, it can be seen that most of these instances are caused by voicing ambiguity, but not place ambiguity. In fact, 26 out of 27 involve only voicing errors. There is only one that involves a place of articulation error (number 19 in Table 6.15). In many cases, the machine results agreed with the same phonetic category as the perceptual results.

Now consider a few particular examples.

- Number 5 in Table 6.15: This is a good example of a conflict between the phonemic and phonetic evidence. The orthography is "he paused then," and the relevant phoneme is the /d/ at 0.93 seconds in "paused." Figure 6-8 shows a spectrogram of these words. Examination of the spectrogram shows that there is no pre-voicing before the stop burst, and the voice onset time is not available because the speaker actually paused after saying the word "paused." Thus, acoustic-phonetic evidence indicates that this phone is [t], and both humans and machines confirmed this. The people labeling the TIMIT transcriptions, however, were biased toward the phonemic /d/ indicated by the word "paused."

- Number 10 in Table 6.15: Figure 6-9 shows the spectrogram of "heating." The /t/ at 4.30 seconds was transcribed as [tcl t] in spite of the presence of voicing and a short voice onset time. Perceptual results indicate that it is closer to a [dcl d]. From a speech production point of view, this example was almost realized as a flap, [dx], which is common for intervocalic /t/ or /d/ sounds.

- Number 18 in Table 6.15: This is an example of contextual effects leading to an apparent phonetic/phonemic conflict for two phones in a row. Figure 6-10 shows the spectrogram of "gives driving." The portion of interest the /d/ at 0.79 seconds in /z dcl d r ay/. The TIMIT transcribers decided that voicing was absent from the /z/, so they labeled it as [s]. However, they still labeled the stop as [dcl d]. Perceptual results show that this stop was perceived as [tcl t]. Humans may have been biased toward perception of [t] because /s t r/ is allowable at the start of a syllable, whereas /s d r/ is not. The human listeners were told that the stops could have been extracted across syllable boundaries, but without training it may have been difficult for humans to suppress the intuitive perception of short segments as being syllable-initial. On the other hand, there are other cases, such as number 109 in Appendix A, where some listeners did overcome a similar bias involving /s b/ versus /s p/.

- Number 19 in Table 6.15: This is interesting because it is the one case where there was a unanimous place of articulation error. Figure 6-11 shows a spectrogram of the beginning of the word "combine." The stop of interest is the sentence initial /k/ at 0.14 seconds. This example appears to be a case of misarticulation. The change in place of articulation cannot be attributed to contextual effects since it occurs at the beginning of the sentence. In spite of the phonemic /k/, what the speaker actually produced is apparently much closer to a phonetic [t].

In practice, word recognition systems have several different ways of handling the variability in the way that people pronounce words. Some systems absorb the variability into context-dependent acoustic models. Other systems attempt to model the lexicon more precisely by explicitly considering alternative pronunciations. The results here illuminate the difficulty involved in creating a single "true" phonetic transcription in light of phonetic ambiguity, and highlight the need for appropriate ways to deal with these effects in word recognition.

6.9 Summary

Perceptual experiments with stop consonants were designed in order to explore differences between human and machine performance. The experiments focused on tokens which were difficult for machines to classify.

Comparisons of listener results with a machine system using heterogeneous measurements showed that voicing identification performance by machine was equivalent or better than that of six out of seven listeners. Place identification performance was equivalent to that of two out of seven listeners. Heterogeneous measurements were helpful in narrowing the gap between human and machine performance. More work is needed in order to improve place of articulation performance to the point where it is comparable to human performance.

Particular errors in place of articulation identification by humans and machines were examined. Two hypotheses were developed to explain why there is still a significant gap between human and machine place of articulation identification. The first hypothesis has to do with the lack of training data for rare phonetic contexts. The second hypothesis asserts that current systems do not adequately measure and model extremely short-time phenomena such as is found at the stop burst. This is not a problem for most stops, since there are typically several redundant acoustic cues for the place of articulation. There are times, however, when these short-time cues are non-redundant and critical to the identity of the place. In these situations, the shortcomings of the machines are apparent. These two hypotheses should be helpful in the design of future systems which aim to eliminate the gap between human and machine performance in place of articulation identification. In addition, the investigation also uncovered examples of phonetic ambiguity, misarticulation, and questionable TIMIT reference transcriptions.

This investigation and the specific examples herein provide perspective regarding the acoustic difficulties and phonetic ambiguities which arise in continuous speech, and hypotheses regarding how to potentially improve future systems.

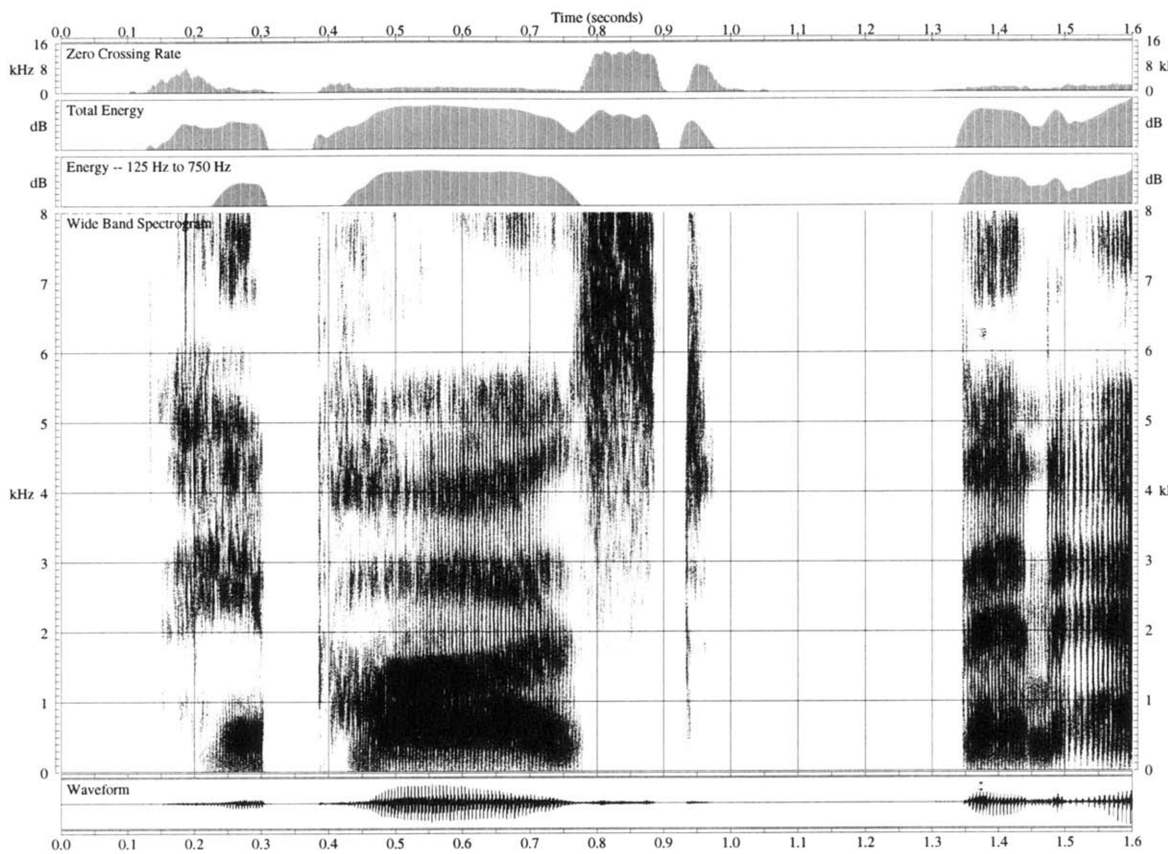

Figure 6-8: Spectrogram of the words "he paused then," corresponding to number 5 in Table 6.15. The stop consonant of interest is the /d/ at 0.93 seconds in "paused." There is no pre-voicing before the stop burst, and the voice onset time is not available because the speaker actually paused after saying the word "paused." Thus, acoustic-phonetic evidence indicates that this phone is [t], and both humans and machines confirmed this. This is an example of conflict between phonemic and phonetic evidence.

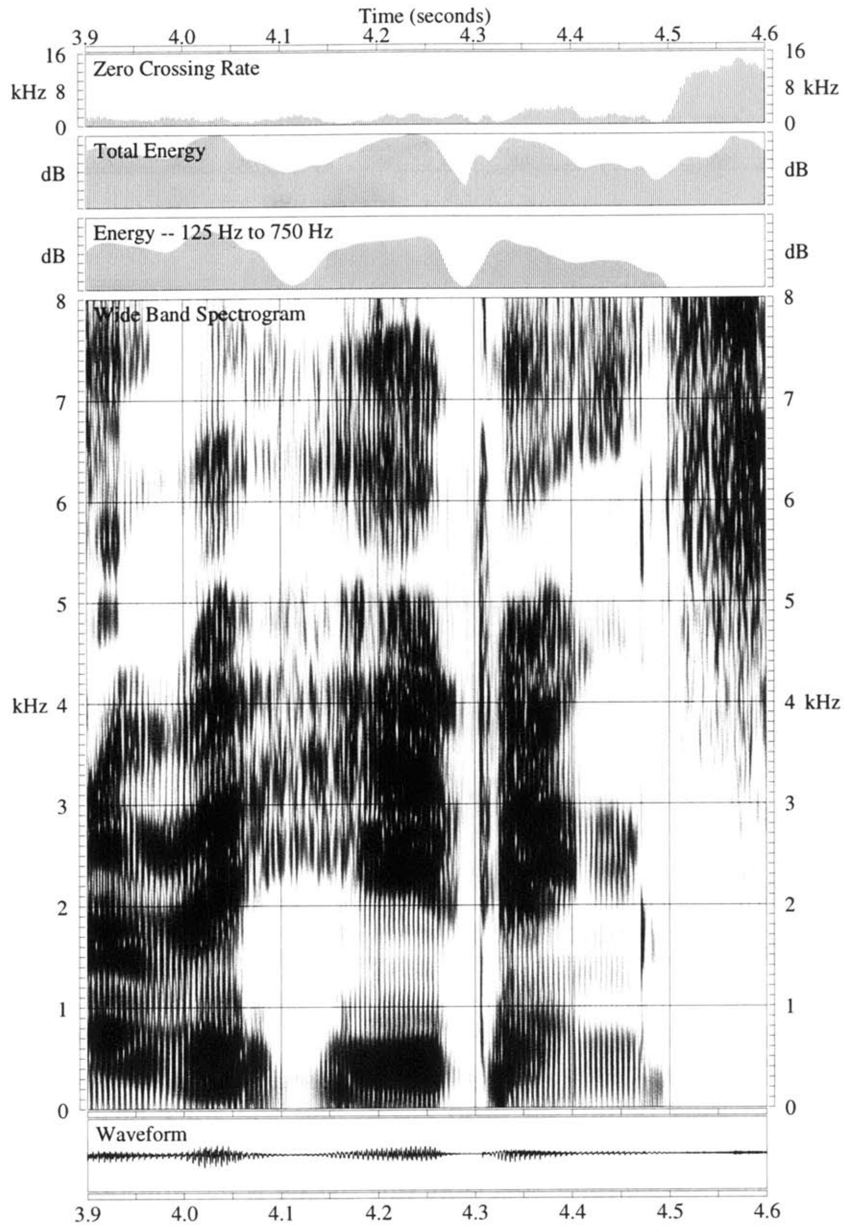

Figure 6-9: Spectrogram of "heating," corresponding to number 10 in Table 6.15. The /t/ 4.30 seconds was transcribed as [tcl t] in spite of the presence of voicing and a short voice onset time. Perceptual results indicate that it is closer to a [dcl d]. From a speech production point of view, this example was almost realized as a flap, [dx], which is common for intervocalic /t/ or /d/ sounds.

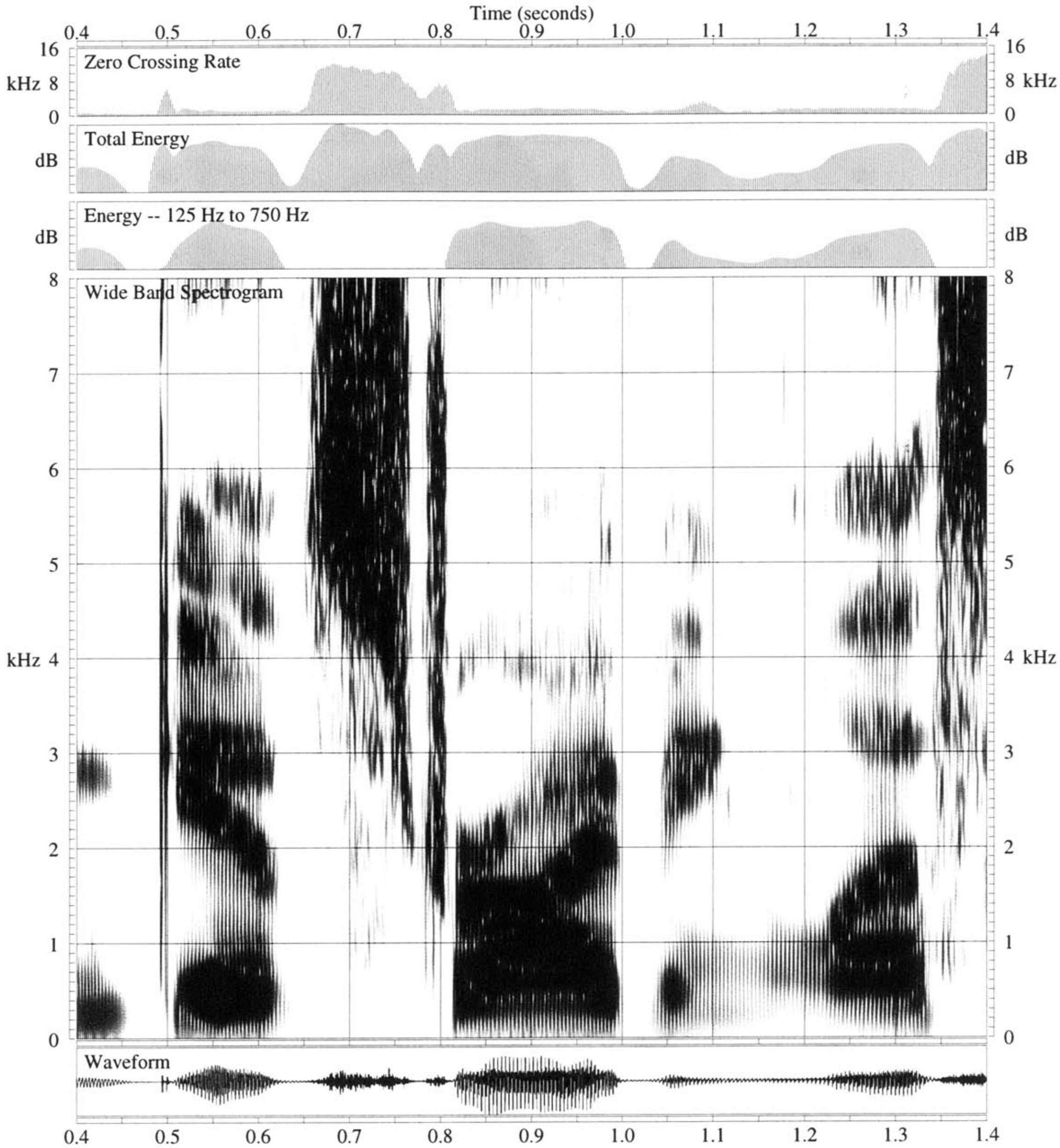

Figure 6-10: Spectrogram of "gives driving," corresponding to Number 18 in Table 6.15. The stop of interest is the /d/ at 0.79 seconds. This is an example of contextual effects leading to an apparent phonetic/phonemic conflict for two phones in a row. The TIMIT reference transcribers decided that voicing was absent from the /z/, so they label it as [s]. However, they still labeled the stop as [dcl d]. Perceptual results show that this stop was perceived as [tcl t].

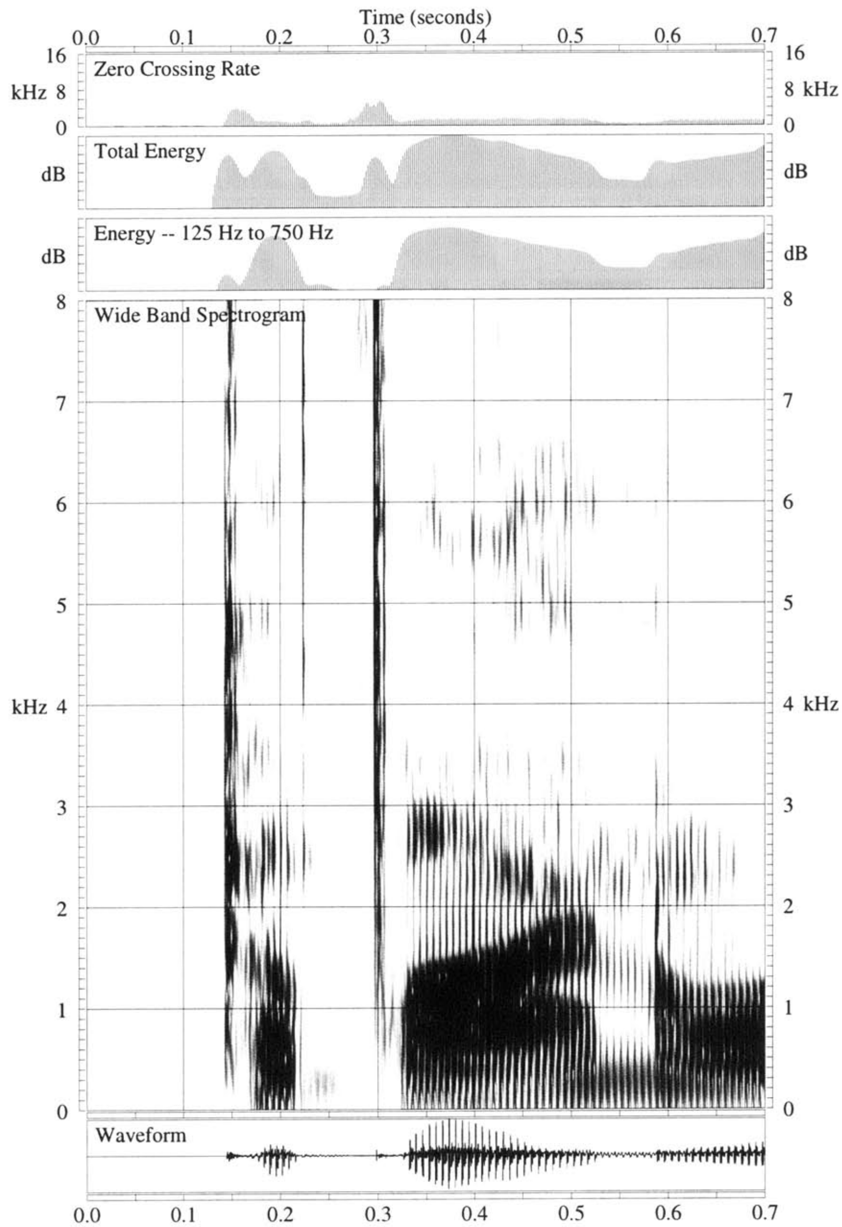

Figure 6-11: Spectrogram of "combine," corresponding to Number 19 in Table 6.15. The stop of interest is the sentence initial /k/ at 0.14 seconds. This is the one case in Table 6.15 where there was a unanimous place of articulation error by humans. It appears to be a case of misarticulation. In spite of the phonemic /k/, what the speaker actually produced is apparently much closer to a phonetic [t].

Chapter 7

Conclusion

7.1 Summary and Contributions

The work in this thesis was motivated by three observations. First, perceptual experiments showed that humans are still much better than machines in low-level acoustic phonetic tasks. Second, experiments using a variety of time-frequency resolutions showed that the typical, fixed time-frequency resolution used in most systems is a compromise among the settings that are optimal in different phonetic classes. Third, most phonetic confusions occur within the correct manner class, so the phonetic classification task can be broken up into smaller subproblems. These motivational observations may also serve to mobilize and guide the efforts of other researchers.

Two ideas regarding the evaluation of classifiers set this work apart from other acoustic modeling studies. The first idea flows from the third motivation. Simply stated, it is the notion that one should not simply look at the overall phonetic classification results of a particular classifier, but should also examine the performance within phonetic subclasses in order to more completely evaluate the strengths and weaknesses of each particular classifier. This analysis resulted in additional insights into the design of good acoustic measurements. For example, in Section 4.2, examining the performance of classifiers within phonetic manner classes resulted in the observation that cosine temporal basis vectors were better than linear or constant basis functions for the encoding of MFCC trajectories of sonorants. Examining only the overall classification result does not reveal this. The second change in the way that classifiers were evaluated is even more important than the first. In the past, acoustic measurement evaluation experiments were carried out with the objective of finding and using the one-best acoustic measurement set. In contrast to this, the committee-based experiments in this thesis explicitly aim to develop a suite of acoustic measurement vectors that can be used simultaneously in order to take advantage of the complementary information among them. In fact, acoustic measurements were designed with the objective of *maximizing* the complementary phonetic information. This design objective produces a new way of thinking in which different measurement sets are no longer viewed as independent and in competition, but rather as interrelated and complementary. Different measurement sets should now be designed jointly, not independently. Instead of having to choose one single time-frequency resolution, we select several which will extract different information from the speech signal. This change in the evaluation of acoustic measurements is implicit in the design of all of the committee-based experiments.

The notion of using multiple acoustic information sources has led to encouraging results in a wide variety of experiments. These include experiments using different lexical representations such as incorporating syllabic knowledge [95], or using different phonetic inventories (AT&T), different frame rates (BBN), multiple levels of acoustic context (triphone and quinphone, HTK), and the ROVER system as originally proposed by NIST [25].[1]

Hierarchical approaches to measurement development and classifier combination were explored in Chapter 3. Classifiers were combined by expanding the usual MAP decoding over a set of phonetic classes. Phone-class-specific acoustic measurements were developed through phonetic knowledge combined with experimental verification for four phone subsets: vowels/semivowels, nasals/flaps, stops, and fricatives/closures. Increasing the time resolution and lowering the dimensionality of the segmental stop consonant measurements was particularly effective, leading to 8.7% error rate reduc-

[1]The systems from AT&T, HTK, and BBN were part of the September 1998 Benchmark Tests [67] mentioned in Chapter 4.

tion relative to the baseline on the full test set. Overall, these techniques achieved 21% error on the TIMIT core test set, which was the best reported result at that time [31].

Several committee-based approaches for combining multiple classifiers were described in Chapter 4, including voting, linear combination, and the use of an independence assumption. This chapter also presented 8 segmental measurement sets and 5 boundary measurement sets which were designed to contain complementary acoustic-phonetic information. A theoretical discussion looked at the differences between attempting to train one classifier using the union of the measurements in the individual feature vectors, versus training separate classifiers followed by combining the outputs of the classifiers. In addition to the committee-based approaches, two hybrid approaches which combine elements of the committee and hierarchy approaches were also presented.

Techniques for committee-based classifier combination were evaluated in Chapter 5. Both the linear combination and independence assumption techniques performed well in comparative experiments in the task of TIMIT phonetic classification. A hybrid system using a a nine-member classifier committee produced the best phonetic classification result of 18.3% error on the TIMIT core test set, which is a 12.9% improvement over our previous best reported result of 21.0% [31]. The next best result that we have seen in the literature reporting TIMIT CI classification on the core test set is 23.0% [97]. Subsequent experiments in TIMIT phonetic recognition and JUPITER word recognition used an independence assumption for combining multiple classifiers. The final TIMIT phonetic recognition result of 24.4% on the core test set is the best that has been reported in the literature. In the JUPITER telephone-based word recognition task, word error rate reductions of 10-16% were observed using a committee of three classifiers. The experiments with JUPITER are important because they show the generalization of these techniques to word recognition in a telephone-bandwidth acoustic environment.

Several sections in Chapter 5 make additional contributions. Section 5.1.3 provides detailed performance results from our best context-independent, classification system without vocal tract length normalization. Section 5.1.4 shows experimentally that training a single high-dimensional classifier is typically not a successful way to utilize high-dimensional measurement spaces due to the exponentially growing size of high-dimensional spaces. Section 5.1.5 shows that individual classifiers which show mutual diversity are good candidates for classifier combination. Given a set of individual classifiers and their performance on a development set, one would like to be able to predict which subset of those classifiers might perform well when combined into a single system. We found that the diversity among a subset of classifiers was a good predictor of the success of the combination of those classifiers. In fact, diversity was a better predictor than the individual classifier performance on the development set. This is reasonable, since two classifiers which have very good performance but are nearly identical will not lead to much benefit when they are combined because they do not contain much complementary information.

Perceptual experiments on stop consonants and comparable machine classification results are presented in Chapter 6. The final results comparing our best machine classification results with humans shows that machines perform almost as well as humans in identification of voicing of stops, but still lag by a factor of 1.8-5.1 in identifying the place of articulation. Error analysis of the differences between human and machine performance resulted in two hypotheses about why there remains a gap between human and machine stop classification performance. The first hypothesis is that machines do not adequately handle cases where there are extreme formant locations in neighboring phones. These cases produce statistical outliers which are generally not adequately modeled based on the limited amount of training data. The second hypothesis is that the machines are not able to accurately measure and model the extremely short-time phenomena at the stop burst which are sometimes critical to correct identification of a stop consonant. The error analysis investigation also uncovered examples of phonetic ambiguity, misarticulation, and questionable TIMIT reference transcriptions. Taken as a whole, the perceptual studies provide perspective concerning the acoustic difficulties and phonetic ambiguities which arise in continuous speech. These studies will help to guide the direction of further research.

7.2 Future Extensions

There are a large number of areas for extension of this work. The concept of designing and jointly using heterogeneous measurements and multiple classifiers is relatively new and unexplored. Acoustic measurements from the past need to be re-evaluated in the context of a search for complementary phonetic information. New measurements need to be developed which may excel at particular phonetic distinctions, without being concerned that they are poor for other distinctions. New techniques are needed for utilizing high-dimensional acoustic spaces. This thesis used the paradigm of combining multiple classifiers which were trained separately. Intuitively, it should be possible to improve this classifier structure through a mechanism for joint training of all classifiers in a unified framework. This unified framework may be able to eliminate the notion of multiple classifiers and replace it with a single high-performance classifier structure which is able to overcome the difficulties normally associated with training a classifier of such high-dimensionality. Alternatively, other classifier structures could be explored, such as the use of stacked generalization [93].

The ROVER system [25], which was developed by NIST and used by a number of research sites, combines the final word recognition output from multiple recognizers. In contrast, the methods in this thesis combine multiple classifiers at the acoustic modeling level, so that the changes in acoustic modeling are invisible to the rest of the system. Future research might compare and contrast these two approaches in a large vocabulary word recognition task. Early combination of acoustic models should provide better information for the search. On the other hand, late combination allows one to potentially benefit from the *independent* interaction of each acoustic measurement set with the remainder of the system.

Multiple acoustic measurements could be sought for the purpose of improving the noise robustness of a system. This requires the design of measurements which are robust to a variety of types of noise. There is work in the literature which combines acoustic measurements based on multiple independent frequency bands [87, 3, 71]. Those studies usually do not show performance improvements, but they do show increased noise robustness. This method of achieving noise robustness could potentially be extended by adding other types of measurements.

The techniques presented in this thesis should be useful for speech scientists who develop specialized acoustic measurements. These researchers could *augment* existing measurement sets with the alternative measurements they develop. This allows them to seek complementary information, rather than competing directly with conventional acoustic measurements. There are studies which have moved in this direction, such as combining formant measurements and MFCCs [32, 35, 82], but they have not used powerful multiple classifier structures such as those presented here.

In this thesis, we have not been concerned about the computational complexity of using multiple classifiers. If all computation is performed, then using N classifiers will increase the acoustic modeling cost of a recognizer by a factor of N. It should be possible to realize much of the performance improvement from multiple classifiers without full computation of every classifier for every acoustic observation, using ideas analogous to "phonetic fast-match." Conventional fast-match results in evaluating the acoustic models for only a subset of all the possible phones. This new fast-match would retain that quality, but add another dimension, in which a variable number of classifiers may be consulted for each acoustic observation. The variable number could be driven by a measure of confidence or some other criterion.

Exploration of the use of heterogeneous measurements and multiple classifiers is inherently more complex and time consuming than the acoustic measurement studies in the past which only attempted to find a one-best measurement set. This is, in part, due to the fact that the space of possible sets of measurements and ways of combining them is subject to combinatorial explosion. Consider, for example, the exhaustive search experiments in this thesis, where every possible subset of 8 segmental measurements was combined in a variety of ways. The number of experiments required to try all classifier combinations was

$$\sum_{k=1}^{8} \binom{8}{k} = \sum_{k=1}^{8} \frac{8!}{k!(8-k)!} = 255.$$

If one wanted to do a similar exhaustive search over all combinations of 24 sets of measurements, it would require over *16 million* experiments, rather that 255. Thus, exhaustive search quickly becomes impractical. In spite of these difficulties, we expect that future research will continue to uncover sets of measurements which are particularly suited for joint use with multiple classifiers.

7.3 Final Comments

This thesis provides motivation for new directions in acoustic modeling research. It has been demonstrated empirically that typical homogeneous acoustic measurements do not capture some of the low-level acoustic phonetic information in the speech signal. The solution to this problem is to retain more information, at the expense of having more measurements. In this thesis, signal processing knowledge and acoustic-phonetic knowledge were combined to motivate a new agenda for acoustic modeling research. That new agenda is to design heterogeneous measurements to maximize the coverage of the acoustic input space and to design classifier structures in order to use those measurements simultaneously. In the future, acoustic-phonetic knowledge should continue to guide the process of determining where more acoustic detail is needed. Signal processing knowledge should be used to determine how to obtain that level of detail. Data-driven techniques, such as the evaluation on the TIMIT and JUPITER databases in this thesis, should continue to be used to evaluate the effectiveness of various measurement sets. Acoustic modeling according to this broad strategy is a fertile area for further research.

There is still a large gap between human and machine speech recognition ability. In automated systems, the blurring or loss of low-level acoustic-phonetic information cannot be regained by subsequent processing, although the loss of acoustic-phonetic information may be masked by the application of higher-level lexical or linguistic constraints. Consideration of the results in this thesis in light of the fundamental limits on time-frequency resolution and the non-invertibility of most preprocessing algorithms suggests that speech recognition systems of the future will be designed to incorporate heterogeneous acoustic measurements. This will result in the retention of more acoustic-phonetic information from the speech signal, and ultimately in the narrowing of the gap between human and machine speech recognition performance.

Appendix A

Results of Perceptual Experiments

This appendix provides the raw data from the perceptual experiments and the comparison with machine performance. The table on the following pages lists each token individually with the utterance tag, TIMIT phonetic labels, the orthography, and the machine and human hypotheses from these experiments. Chapter 6 presents these results in summary form along with error analysis and discussion.

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
1	sx53-b-fcal1	m pcl p el v	a simple vocabulary	ggbb	bbbbbpp
2	sx139-b-fnmr0	ng gcl g ax l	the bungalow was	dggg	ggggggg
3	sx290-b-mjsw0	s bcl b r ow	this brochure is	bppp	ppppppp
4	sx290-b-mjsw0	ih kcl k ix l	is particularly informative	gkkk	kkkgkkk
5	sx290-b-mjsw0	s pcl p eh kcl	a prospective buyer	bbpp	pbppppp
6	sx290-b-mjsw0	eh kcl t ix v	a prospective buyer	tdtt	tttttdt
7	sx320-b-mrws1	aa tcl b iy w	not be within	pbpb	bbbbbbp
8	si2270-b-mjsw0	ao gcl g z dcl	# dogs did	kgkg	ggggggg
9	si2270-b-mjsw0	z dcl d ih dcl	dogs did something	ddtd	tdttttt
10	si2270-b-mjsw0	ng tcl t ax w	something to one's	ttdt	tttddtt
11	sx410-b-mbns0	iy kcl d ey m	each weekday morning	tddt	dddddkd
12	si1301-b-mmwh0	# h# d iy n	# deans can	tddd	ddddddd
13	si1301-b-mmwh0	en tcl b r ih	important bridge between	pbpb	bbbbbbb
14	si1301-b-mmwh0	ae kcl k el tcl	the faculty #	pgkk	kgkkkkk
15	sx410-b-fkms0	# h# g r eh	# greg buys	kggg	gkkgkgk
16	sx410-b-fkms0	l kcl k q iy	fresh milk each	ggkk	kkkkkkk
17	sx410-b-fkms0	iy kcl d ey m	each weekday morning	ddgd	ddddkkd
18	sx140-b-fmml0	ux dcl d ix dcl	unevenly distributed #	tttd	tddddtd
19	sx94-b-fedw0	ah ng g s tcl	# youngsters love	kkkg	ggkkgkk
20	sx94-b-fedw0	n dcl d iy q	common candy as	dbdd	ddddbd
21	si1218-b-fdms0	axr dcl d w eh	moreover dwell on	tttt	ddddddtd
22	sx143-b-fcal1	ux gcl g el s	jennifer's bugle scared	dggg	ggggggg
23	si1539-b-mgwt0	el dcl d hv ih	fists pummeled him	tdbd	ddddddd
24	si1539-b-mgwt0	s tcl t ae gcl	he staggered forward	dttt	ttttttt
25	si1539-b-mgwt0	axr dcl d f ao	he staggered forward	tddd	dgddddd
26	si2166-b-mmjr0	w iy t hh aa	look sweetheart some	dddd	dtddddd
27	si2166-b-mmjr0	r tcl t pau s	look sweetheart some	dddt	tgtttdt
28	si497-b-mers0	# h# d ix s	# displayed as	tddd	ddddtdd
29	si497-b-mers0	s pcl p l ey	# displayed as	bppp	pbppppp
30	si497-b-mers0	m pcl p s pau	as lamps the	dpgp	bpbpppb
31	si497-b-mers0	ah pcl p ax-h s	the puppets delight	bppt	pppppbp
32	si497-b-mers0	s dcl d ax l	puppets delight the	tttt	ttttttt
33	sx44-b-mbwm0	# h# b ae s	# basketball can	pbbb	bbbtbbb
34	sx44-b-mbwm0	ix tcl b ao l	# basketball can	pppp	ppbpppp
35	sx44-b-mbwm0	axr tcl t ey n	an entertaining sport	ttkk	ttttttt
36	sx44-b-mbwm0	s pcl p ao r	entertaining sport #	bppp	ppppppp
37	sx28-b-mmjr0	eh tcl g aa r	that guard for	kkkg	kgggkgg
38	sx28-b-mmjr0	v gcl g ae s	of gas #	kkkk	gkggggg
39	sx403-b-mjfc0	r dcl d r ow	her wardrobe consists	gggg	dgdddgd
40	sx408-b-fdms0	ey pau b ah dx	subway but i	dbbb	bbbbbbb
41	sx408-b-fdms0	ax n t ix n	i haven't enough	dddd	ttttttt
42	sx413-b-mdlf0	# h# d ae f	# daphne's swedish	bddb	ddddddd
43	sx413-b-mdlf0	s kcl k aa r	needlepoint scarf matched	kkgg	kkkkkkk
44	si674-b-mbwm0	s pcl p eh kcl	no spectator type	ttpp	pbppppp
45	si674-b-mbwm0	eh kcl t ey dx	no spectator type	ttpt	ttttttt
46	si674-b-mbwm0	ay pcl p ix kcl	spectator type experience	ppbp	ppppppp
47	si1454-b-fcmh0	ay pcl p sh y	main types used	pkkt	bbbbbbb
48	si1454-b-fcmh0	m pcl p ax z	various compositions of	tpbp	ppppppp
49	si1454-b-fcmh0	v pcl p aa l	of polyesters #	bbbp	ppppppp
50	sx19-b-fadg0	n bcl b iy f	often be flimsy	bddb	bbbbbbb

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
51	sx274-b-mtdt0	ah gcl g zh y	the luxurious massage	gdgg	ggggggg
52	sx134-b-mbwm0	# h# d ix s	# december and	tdtd	ddddddd
53	sx134-b-mbwm0	ah n t s tcl	nice months to	dddd	ttttttt
54	sx134-b-mbwm0	s pcl p eh n	to spend in	tttp	pbppppp
55	sx53-b-mrcs0	ae bcl b y ax	simple vocabulary contains	gddg	gbdgggg
56	si1555-b-mmdm2	iy pcl p iy v	a sleepy voice	ppkk	ppbpppp
57	si649-b-fadg0	ae kcl k ix v	a lack of	bkkk	gkgkggk
58	si649-b-fadg0	er pcl p ix s	of purpose and	btpp	ppbbppp
59	si649-b-fadg0	s pcl p eh kcl	and respect for	bbtb	ppppppp
60	si649-b-fadg0	eh kcl t f er	and respect for	dkkt	kkkttkt
61	si649-b-fadg0	ix kcl k l iy	for heroic leadership	gkkk	kgkkkkk
62	si649-b-fadg0	ih pcl p h# #	heroic leadership #	pptp	bpbppbb
63	si728-b-mjar0	ah pcl p f axr	this syrup for	ttkt	ppppppp
64	si728-b-mjar0	s pau d r ey	minutes drain them	tddt	tdttttd
65	sx191-b-fjmg0	th dcl d r ao	# withdraw all	ttgt	ttttttt
66	sx191-b-fjmg0	ae kcl k y ix	phony accusations at	kpkk	kkkkkkk
67	sx50-b-fkms0	ix kcl k pau eh	# catastrophic economic	tkkk	kkkgkkk
68	sx50-b-fkms0	eh kcl k ax n	catastrophic economic cutbacks	ddtk	kkkkkkk
69	sx50-b-fkms0	ih gcl g l eh	cutbacks neglect the	tggg	ggggggg
70	sx143-b-mrcs0	z bcl b y ux	jennifer's bugle scared	bbpb	bpbbbbp
71	sx143-b-mrcs0	s kcl k eh r	bugle scared the	gkkg	kkkkkkk
72	sx143-b-mrcs0	ae n t el ow	the antelope #	kdkd	tktkttt
73	sx143-b-mrcs0	ow pcl p h# #	the antelope #	tptp	bpppppp
74	si1542-b-mpdf0	# h# d ih f	# the feet	dbpb	ddddtdd
75	si1542-b-mpdf0	iy tcl t w ao	the feet wore	dddg	tgttttt
76	si1542-b-mpdf0	s dcl d ih s	obvious disrepair #	tttt	tdttttt
77	sx279-b-fgjd0	ax bcl b ih l	alice's ability to	bbdb	tdbbddd
78	sx279-b-fgjd0	uw pcl p er v	without supervision is	ppkp	ppdpppp
79	sx364-b-fedw0	ax bcl b r ey	the breakdown of	ppbb	bbbbbbb
80	sx364-b-fedw0	ey kcl d aw n	the breakdown of	tddd	ddddddd
81	sx364-b-fedw0	r tcl t h# #	system part #	ttdt	tgttttt
82	sx43-b-mjfc0	l dcl d uh l	# elderly people	bddd	ddddddd
83	si1015-b-mthc0	iy dcl d h# #	livestock feed #	tdtd	dtddddd
84	si2120-b-fkms0	z dcl d pau dh	he paused then	tttt	ttttttt
85	si2120-b-fkms0	ih dcl d pau eh	then added everything	tddd	ddddddd
86	si2120-b-fkms0	ih pcl p pau ih	a ship is	ddtp	bbpbpbp
87	si2120-b-fkms0	eh pcl p ax n	a weapon #	bppp	pbbpppp
88	si844-b-fdac1	s pcl p eh n	the expense of	bbbp	pbppppp
89	si844-b-fdac1	s pcl p ih r	actual experience #	bbbp	pbppppp
90	sx53-b-mdlf0	m pcl p el v	a simple vocabulary	bbbb	ppbbbpp
91	sx314-b-mbwm0	iy pcl p el w	if people were	ppbp	ppppppp
92	sx205-b-mthc0	s pcl p eh kcl	are expected to	bppp	ppppppp
93	sx229-b-mrjm4	m pcl p l oy	for employment #	pggp	ppbpppp
94	sx229-b-mrjm4	n tcl t h# #	for employment #	tddd	dpppdpp
95	si1628-b-mdls0	m dcl d hv ix	he informed him	dtdd	tptddgk
96	si1628-b-mdls0	r bcl d w ey	an absorbed way	kbbb	dgdgddd
97	sx200-b-mjsw0	s tcl t aa pcl	the stopwatch from	ddtt	tdttttt
98	sx293-b-fdrw0	# h# p l iy	# please take	kppp	pkppkpk
99	sx293-b-fdrw0	s dcl d er dx	this dirty tablecloth	tttt	tdttttt
100	sx293-b-fdrw0	ey bcl b ax kcl	dirty tablecloth to	dbdb	bbbbbgb

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
101	si860-b-fkms0	n kcl k l ux	these include leaves	tttt	kkkkkkk
102	sx407-b-mroa0	f pau d ae n	laugh dance and	bbbd	ddddddd
103	si1280-b-frew0	l tcl t axr sh	the shelter shown	tdtt	ttttttt
104	si1280-b-frew0	axr tcl t eh kcl	fallout protection #	ttkt	tttkttt
105	sx320-b-mbns0	aa pcl b iy w	not be within	pbpb	bbbbbbb
106	sx383-b-mbdg0	r pcl p ix tcl	the carpet cleaners	bppp	ppppppp
107	sx383-b-mbdg0	m pcl p uw dx	cleaners shampooed our	ptpp	ppppppp
108	sx313-b-mjfc0	# h# d r aa	# drop five	tttd	gdddgdd
109	sx313-b-mjfc0	s bcl b ax-h f	box before you	bbpp	bppbpdp
110	sx313-b-mjfc0	iy gcl g ow aw	you go out	dggg	ggggggg
111	sx369-b-mgwt0	ey pcl p axr dcl	bob papered over	pbpp	pbppppp
112	sx281-b-mgjf0	ae gcl g w ix	played tag with	kkkk	ggggggg
113	sx281-b-mgjf0	ch bcl b ao l	beach balls for	pppp	bgbbpbb
114	sx298-b-fsem0	n bcl b iy ow	than be overcharged	dbdd	bdbbbbb
115	sx372-b-mrjr0	s tcl t ah dx	much study #	pttt	ttttttt
116	si1750-b-mrtk0	m bcl b ae s	the investor himself	bbdb	tgbtbbt
117	si1750-b-mrtk0	f bcl b iy dh	himself be the	pkpp	pbppptp
118	si1750-b-mrtk0	ax pcl p r ay	the prime minister	pkpp	ppppppp
119	sx278-b-mjar0	s tcl t iy dcl	his prestige he	dttt	tdttttt
120	sx389-b-fmah0	ah bcl b ih gcl	a big house	dddb	bbbbbbb
121	sx389-b-fmah0	ih gcl g hv aw	a big house	dddd	gkkkgkg
122	sx229-b-fnmr0	em pcl p l oy	for employment #	ppkp	ppppppp
123	si1988-b-mjar0	s tcl t r ay	how's it strike	tktt	ttttttt
124	sx99-b-fgjd0	z bcl b ey s	was based on	tbtp	pbppppp
125	si1994-b-mtdt0	z tcl t ix gcl	was to get	dddd	ddddddd
126	si1994-b-mtdt0	s pcl p eh n	you just spent	bbbp	pbppppp
127	sx230-b-fmml0	axr pau b ix tcl	here but rationalize	pbpb	pbtbptp
128	si773-b-fcal1	s dcl d r ay	projects dry slowly	tttt	ttttttt
129	si593-b-mrcs0	n tcl t hh ae	the gallant half	tddd	ttttdtt
130	si593-b-mrcs0	z dcl d ay ih	is dying on	tttd	dtddddd
131	si1084-b-fedw0	ey pcl p s iy	other shapes evolve	dpdt	bpppbbp
132	sx394-b-fdac1	s tcl t r ao	teeth strong #	dttt	ttttttt
133	sx278-b-mdls0	# h# t ux f	# to further	kttt	ttttttt
134	sx278-b-mdls0	z pcl p r eh	his prestige he	ptpp	pppppbp
135	sx278-b-mdls0	iy dcl d z dh	occasionally reads the	ttdd	bddbdbb
136	si500-b-mrws1	ae kcl k s aa	obvious nymphomaniacs on	dkkk	kkkgggk
137	sx95-b-mmdb1	iy gcl g w aa	# iguanas and	kggg	gkggggk
138	sx95-b-mmdb1	ix gcl g ey dx	and alligators are	kkkk	ggkgkbk
139	sx364-b-mglb0	ey kcl d aw nx	the breakdown of	ttdt	dddddgd
140	sx364-b-mglb0	em pcl p aa r	system part #	bbbp	ppppppp
141	sx49-b-mrjm4	th dcl d ey w	twelfth day we'll	ttdt	tdttttt
142	sx94-b-fjem0	n dcl d iy eh	common candy as	gdgd	ddddddd
143	si1490-b-fkms0	ae pcl p s ax	# perhaps it	pkpp	ppppppb
144	si1490-b-fkms0	ae pcl p s ix	right perhaps it	tdkp	bbbpbbb
145	sx409-b-fnmr0	iy tcl t iy ng	# eating spinach	ttkt	ttttttt
146	sx113-b-mbdg0	s kcl k ix l	a muscular abdomen	dgdt	ktkkkkk
147	sx113-b-mbdg0	z gcl g uh dcl	is good for	kkkk	kgkkkgg
148	sx113-b-mbdg0	uh dcl d f axr	is good for	kdgd	ddddddd
149	si909-b-mgwt0	n tcl t eh tcl	compulsory retirement at	dddd	ddddddd
150	si909-b-mgwt0	eh tcl t s ih	retirement at sixty	ttkt	tgttttt

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
151	si909-b-mgwt0	ih kcl k s tcl	at sixty five	gggk	tkkkgkk
152	si1648-b-mmjr0	el dcl d ix s	electrical display in	ddtd	ddddddd
153	si1648-b-mmjr0	s pcl p l ey	electrical display in	pbpp	ppppppp
154	sx410-b-fmml0	# h# g r eh	# greg buys	kggg	ggggggg
155	sx410-b-fmml0	ix kcl k dcl d	each weekday morning	tttk	kbbbkkk
156	si2196-b-mdvc0	z gcl g uh dx	as good an	kkkk	kgkkkkk
157	si2196-b-mdvc0	z dcl d ae tcl	be sneezed at	ttdt	tdttttt
158	si2119-b-mrjm4	ng kcl t ix dcl	instinct to discipline	gktt	ttttttt
159	si2119-b-mrjm4	ix dcl d ih s	to discipline has	bddd	ddddddd
160	si2119-b-mrjm4	ax pcl p l ix	to discipline has	ppkp	ppppppp
161	si2119-b-mrjm4	z bcl b ix n	has been lost	ddbb	dbbdddt
162	sx102-b-mrjr0	s pcl p eh sh	# special task	bpbb	bbpbppp
163	sx102-b-mrjr0	s kcl k f ao	special task forces	kgkk	kpkkkkk
164	sx413-b-fcall	s kcl k aa r	needlepoint scarf matched	tkkk	kgkkkkk
165	sx119-b-fmah0	iy tcl t r ae	was retracted with	kkkk	ttttttt
166	sx119-b-fmah0	ae kcl t axr dcl	was retracted with	tktk	tdttdtt
167	sx385-b-mtaa0	axr bcl b r ow	were broken #	ppbb	pbppbpb
168	si2247-b-mjar0	iy kcl k n ih	# weakness in	kktk	kkkkkkk
169	si824-b-fcmh0	m pcl p r eh	enzyme preparation through	ppbp	ppppppp
170	sx143-b-mteb0	z bcl b y ux	jennifer's bugle scared	dbbb	bbgbbbb
171	sx143-b-mteb0	ux gcl g el s	jennifer's bugle scared	gddg	ggdgggg
172	si1399-b-fnmr0	ih gcl g l aa	the agglomeration of	gkgg	gggggkg
173	sx385-b-mreb0	l tcl t eh z	shoulder felt as	tdpd	dddbbdd
174	si2313-b-mrjr0	er pau d ih dcl	junior didn't he	gdtd	ddddddd
175	si2313-b-mrjr0	n dcl d ux dh	even do that	tttd	ddddddd
176	si859-b-mrjm4	n kcl k en tcl	non contributory plan	gkkd	kkkkkkk
177	sx49-b-fnmr0	th dcl d ey pau	twelfth day we'll	dttt	tdttdtt
178	sx396-b-mdvc0	sh bcl b iy gcl	fish began to	bppp	pbppppp
179	sx396-b-mdvc0	iy pcl p f r	to leap frantically	ttdt	tpbpppp
180	sx319-b-mrjm4	ow tcl t q ay	big goat idly	dddt	ddddddd
181	sx319-b-mrjm4	el dcl d th r	idly ambled through	tddd	ttdtdkt
182	sx279-b-mgwt0	ih tcl t iy tcl	alice's ability to	kttt	dttttdt
183	sx279-b-mgwt0	ow tcl t w axr	is noteworthy #	pppk	pdtdkdt
184	sx364-b-fjem0	ix bcl b r ey	the breakdown of	pppp	ppbbbbb
185	sx364-b-fjem0	ey kcl d aw nx	the breakdown of	tddd	ddddddd
186	sx364-b-fjem0	iy tcl t ix ng	a heating system	dgdd	ddddddd
187	sx50-b-fmml0	eh kcl k ix n	catastrophic economic cutbacks	kpkk	kkkkkkk
188	sx50-b-fmml0	ah tcl b ae kcl	economic cutbacks neglect	pppp	bpbppbb
189	sx50-b-fmml0	ix gcl g l eh	cutbacks neglect the	kkgg	ggkgggk
190	si1663-b-mjfc0	ao tcl t hh iy	he thought he	dddd	tgtttdt
191	si1307-b-mroa0	n tcl t ax-h pcl	conformational entropy #	dtdt	ttttttt
192	si1307-b-mroa0	ax-h pcl p iy h#	conformational entropy #	pkpp	pptpppp
193	si1220-b-mbns0	s tcl t iy zh	money prestige and	dttt	ttttttt
194	sx313-b-faks0	# h# d r aa	# drop five	tddb	ddddddd
195	sx313-b-faks0	s bcl b ax f	box before you	bbpp	pbppppp
196	sx313-b-faks0	ix gcl g ow aw	you go out	kkgg	ggggggg
197	sx413-b-mrcs0	# h# d ae f	# daphne's swedish	tttt	ddddddd
198	sx413-b-mrcs0	er tcl t h# #	her skirt #	ttdt	ttttdtt
199	sx115-b-majc0	m bcl b l ah	the emblem depicts	gdgb	gbdbgbg
200	sx115-b-majc0	ih pcl p ih kcl	emblem depicts the	pkpp	ppppppp

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
201	sx98-b-mdls0	# h# g eh s	# guess the	kkkk	tgggggg
202	sx383-b-fdrw0	aa pcl p ix tcl	the carpet cleaners	tppp	gbpbppp
203	sx383-b-fdrw0	ah gcl g h# #	oriental rug #	kkkd	ggggggg
204	si1198-b-fsem0	s pcl p ay axr	awe inspiring #	bbbb	pbppppp
205	sx298-b-mmjr0	ay dcl d r ae	# i'd rather	ddgd	gdddddd
206	sx298-b-mmjr0	aa tcl b ay dh	not buy these	pppp	ppppppp
207	sx298-b-mmjr0	n bcl b iy ow	than be overcharged	ggbd	ppdbdbp
208	sx298-b-mmjr0	jh dcl d h# #	be overcharged #	ttdt	ttttttt
209	si2064-b-mteb0	z gcl g ih f	was gifted with	ggkg	kkkkkkk
210	si2064-b-mteb0	f tcl t ih dcl	was gifted with	ttdt	tdddttt
211	sx371-b-mgjf0	aa tcl b iy dh	not be the	bbpb	pbppppp
212	sx388-b-fsem0	ax bcl b eh er	the barometric pressure	bbpb	bbbbbbb
213	sx36-b-mdvc0	m pcl p l ix	most accomplished artists	pbbp	ppppppp
214	sx36-b-mdvc0	sh tcl t q aa	most accomplished artists	pptt	ttttttt
215	sx224-b-mwjg0	aw gcl g uh dcl	how good is	kgkk	kgggkgg
216	sx224-b-mwjg0	uh dcl d ix q	how good is	dbdd	ddddddd
217	sx224-b-mwjg0	n dcl d uh r	your endurance #	ttdd	ddddddd
218	sx289-b-fadg0	f gcl g ax l	weatherproof galoshes are	kgkk	kkkgkkk
219	sx368-b-mjar0	s tcl t ax m	lori's costume needed	dttt	ttttttt
220	sx368-b-mjar0	ae kcl g l ah	black gloves to	kkkk	kgkkkkk
221	sx368-b-mjar0	iy kcl k ax m	be completely elegant	dkkk	kkkkkkk
222	sx368-b-mjar0	m pcl p l iy	be completely elegant	pttp	ppppppp
223	sx115-b-mreb0	ax gcl g l ow	all aglow #	gkkk	gkggggg
224	sx319-b-fnmr0	el dcl d th r	idly ambled through	tddd	tgdddbd
225	sx320-b-fmml0	ih gcl g aa gcl	nearest synagogue may	ggdg	ggggggg
226	si1463-b-mbdg0	s pcl p eh kcl	its expected value	bbpp	pbppppp
227	sx134-b-fjsj0	s tcl t ix s	months to spend	tddt	ttttttt
228	sx368-b-mdls0	ae kcl g l ah	black gloves to	kkgk	kkkgkkk
229	sx368-b-mdls0	ax-h bcl b iy kcl	to be completely	dbbb	bbbbbbb
230	sx368-b-mdls0	iy kcl k em pcl	be completely elegant	tkkk	kkkkkkk
231	sx368-b-mdls0	ix gcl g ix n	completely elegant #	gdgg	ggggggg
232	sx113-b-fdrw0	z gcl g uh dcl	is good for	gkgg	gkggkkg
233	sx113-b-fdrw0	uh dcl d f axr	is good for	kdkg	tgddddd
234	sx99-b-mgwt0	s tcl t ao n	was based on	tptt	ttttttt
235	si1181-b-fjmg0	# h# p r aa	# properly used	pbpp	ppppppp
236	si1181-b-fjmg0	aa pcl p axr l	# properly used	bppp	pbppppp
237	si1181-b-fjmg0	z dcl d ix pcl	properly used the	tttt	ddtdttt
238	si1181-b-fjmg0	uh kcl k ih z	present book is	kttt	ggggggg
239	si2255-b-mmdb1	z tcl t w ao	eyes toward the	ppkt	ttttttt
240	si2255-b-mmdb1	eh dcl d r uw	the bedroom some	ggdd	ddddddd
241	si2255-b-mmdb1	f tcl t iy n	some fifteen feet	kkkk	ttttttt
242	sx101-b-mgjf0	l dcl d r ix	kindergarten children decorate	tdtd	ttddddd
243	sx101-b-mgjf0	n dcl d eh kcl	children decorate their	tttd	ddddddd
244	si494-b-mwjg0	# h# b ey kcl	# bake slowly	pbbb	bbbbbbb
245	si494-b-mwjg0	s tcl t w ah	at least one	pptk	ttttttt
246	si1653-b-fedw0	ow kcl k s uw	he spoke soothingly	dkkk	ggggggg
247	si2169-b-mgwt0	ah pcl p s tcl	# upstairs busy	pkpp	bppbppp
248	si2169-b-mgwt0	iy tcl t sh aw	busy feet showering	kttt	bbttttt
249	si2169-b-mgwt0	ay kcl k r ey	showering like raindrops	kkgg	kggggggk
250	si1645-b-mthc0	em pcl p l iy	was completely unjustified	ptpp	ppppppp

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
251	si1645-b-mthc0	ay dcl d h# #	completely unjustified #	ttdd	dgddddd
252	si1010-b-mjsw0	s pcl p er m	hard endosperm of	ppbp	pbppppp
253	si1010-b-mjsw0	eh gcl g s ay	the egg sized	kggg	ggggggg
254	sx209-b-fmah0	ay kcl k el kcl	# michael colored	kgkk	kgkgkkk
255	sx209-b-fmah0	ix bcl b eh dcl	the bedroom wall	dbdb	bbbbbbb
256	si2203-b-faks0	m dcl d f uw	dive seemed foolish	dkdk	gpgkkkt
257	si1828-b-fsem0	axr gcl g aa tcl	he forgot #	gdgg	ggggggg
258	si1304-b-mbwm0	ix kcl k s pcl	was exposed to	kkgk	gkkkkkk
259	si1304-b-mbwm0	s pcl p ow z	was exposed to	bbbp	pbppppp
260	si1304-b-mbwm0	iy gcl g ae s	velocity gas jet	dggg	ggggggg
261	si1946-b-majc0	v dcl d h# #	means solved #	tttt	ttttttt
262	sx4-b-fedw0	r dcl d h# #	working hard #	dtdd	ddtdddd
263	sx44-b-mwjg0	# h# b ae s	# basketball can	pbbb	bbbbbbb
264	sx44-b-mwjg0	s kcl k ix tcl	# basketball can	ttkk	kkkkkkk
265	sx44-b-mwjg0	axr tcl t ey n	an entertaining sport	tkkt	ttttttt
266	si995-b-mmdb1	ae tcl d ih f	at different times	ttdd	dtddttt
267	si995-b-mmdb1	z gcl g l iy	as glee club	kkkk	gkkkkkk
268	si995-b-mmdb1	r gcl g ix nx	as organist #	gkkg	kgkgggk
269	sx214-b-fdac1	s pcl p ay dx	the spider web	bpbb	ppppppp
270	sx214-b-fdac1	n dcl d ax-h pau	web glistened in	tttt	ttttttt
271	sx389-b-mers0	n tcl t r ix	remote countryside #	tptt	ttttktt
272	sx409-b-mrjm4	s pcl p ih nx	eating spinach nightly	pbpp	pbppppp
273	sx409-b-mrjm4	ng kcl k th m	increases strength miraculously	pddk	tptppdp
274	si1030-b-frew0	# h# b r ih	# bring me	pbbb	bbbbbbb
275	si1283-b-fdrw0	l tcl t uh dh	a shelter the	dtdt	tdttttt
276	si1283-b-fdrw0	aw tcl d ix kcl	fallout decays rapidly	ddtt	ttttdtt
277	si1283-b-fdrw0	ae pcl p ix dcl	decays rapidly at	bppp	ppppppp
278	si2172-b-mpdf0	iy gcl g ae n	strength began to	gkgg	ggggggg
279	si2172-b-mpdf0	ih pcl p q ah	to zip up	dppp	pbbbppp
280	si2172-b-mpdf0	ah pcl p ix n	zip up and	bbpp	pppppbp
281	sx134-b-mwjg0	s pcl p eh n	to spend in	bktp	pbppppp
282	sx199-b-fadg0	s pcl p ow zh	avoid exposure to	bbbb	ppppppp
283	sx199-b-fadg0	ix kcl k ix n	to contagious diseases	tkkk	gggkkkk
284	sx199-b-fadg0	s dcl d ix z	contagious diseases #	tttt	tdttttt
285	sx403-b-faks0	r dcl d r ow	her wardrobe consists	ddgd	ddddddd
286	si2005-b-mreb0	n tcl t el hh	would gentle her	ttdt	ttttdtt
287	si769-b-fnmr0	ux dcl d r ay	to dry with	ttdd	ddddddd
288	sx205-b-majc0	s pcl p eh kcl	are expected to	bbbp	pbppppbp
289	sx205-b-majc0	eh kcl t ix dcl	are expected to	dttt	tddkdgt
290	sx140-b-mrws1	ix dcl d h# #	unevenly distributed #	dtdd	bdddddd
291	si776-b-mgjf0	# h# d ux n	# do not	ddtd	dtttttt
292	si776-b-mgjf0	ao tcl d r ao	not draw yarn	tdtt	tdttddt
293	sx283-b-mrtk0	n dcl p er ix	planned parenthood organizations	pktp	pkpppkp
294	sx283-b-mrtk0	z pcl p axr m	organizations promote birth	pptp	pppppdp
295	sx101-b-fjmg0	# h# k ih n	# kindergarten children	kpkt	kkkkktk
296	sx101-b-fjmg0	l dcl d r ix	kindergarten children decorate	dgdd	ddddddd
297	sx101-b-fjmg0	n dcl d eh kcl	children decorate their	tddd	ddddddd
298	sx101-b-fjmg0	ax dcl d ey s	all holidays #	bbdd	ddddddd
299	si1670-b-fmml0	# h# g aa q	# got no	kbpp	ggggggkg
300	sx380-b-frew0	ch gcl g aa r	such garbage #	gkgg	kgggkgg

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
301	sx205-b-mtaa0	s pcl p eh kcl	are expected to	bbbp	ppppppp
302	sx205-b-mtaa0	eh kcl t ix dcl	are expected to	dttt	ttttttt
303	sx205-b-mtaa0	ey gcl g ah v	obey government orders	kkgg	kggggk
304	sx388-b-mmjr0	ix kcl k pcl p	the barometric pressure	kktk	kpkkkkk
305	sx228-b-fdms0	oy n t q ow	the viewpoint overlooked	dddd	ggdgdgg
306	sx228-b-fdms0	uh kcl k dh iy	viewpoint overlooked the	kptp	bgttgkg
307	sx199-b-mcsh0	s pcl p ow zh	avoid exposure to	bbbb	pbppppp
308	si1915-b-mtaa0	eh gcl g f axr	renunciation beg forgiveness	ggkg	bgggggg
309	sx372-b-mpdf0	ae tcl d ay eh	that diagram makes	bbdd	dtddddd
310	sx372-b-mpdf0	s tcl t ah dx	much study #	dttt	ttttttt
311	sx314-b-mwjg0	uh dcl b iy n	would be no	gbpp	bbbbbbb
312	sx314-b-mwjg0	iy dcl d f axr	no need for	kkkt	dgddddd
313	sx379-b-fadg0	s dcl d r ay	gives driving lessons	tttt	ttttttt
314	sx379-b-fadg0	z dcl d ey z	on thursdays #	tttt	tdttttt
315	sx205-b-mreb0	s pcl p eh kcl	are expected to	bppp	ppppppp
316	sx43-b-faks0	iy pcl p el er	elderly people are	bppp	pbbbppp
317	sx43-b-faks0	ix dcl d h# #	often excluded #	ktdd	ddddtd
318	sx29-b-fmah0	eh gcl t ix pcl	greg to pick	ddtd	gdddgdd
319	sx29-b-fmah0	ih kcl k ix pcl	to pick a	gdkk	kggggk
320	si1423-b-fdrw0	iy gcl g aa dx	he got up	kgdg	ggggggg
321	si568-b-fsem0	sh tcl t q ah	a crushed unlined	dttt	dtttttt
322	sx53-b-mteb0	n tcl t ey n	vocabulary contains symbols	dttd	tttddtt
323	si659-b-fmah0	s tcl t r iy	his history is	kttt	tdttttt
324	si659-b-fmah0	eh kcl k ix n	must recognize his	gdkk	kggggkk
325	si998-b-mdls0	v dcl d ao nx	they thrived on	tttt	tddttdt
326	sx224-b-fjsj0	ah gcl g uh dcl	how good is	gkgg	ggggggg
327	sx34-b-fdac1	n q d ux tcl	don't do charlie's	tddd	ddddddd
328	sx34-b-fdac1	z dcl d er dx	charlie's dirty dishes	tttt	ddttttt
329	si1573-b-faks0	ae gcl g er dcl	and haggard and	ggdg	ggggggg
330	sx203-b-fdrw0	ix bcl b r ah	ended abruptly at	ppbb	bbbbbbb
331	sx203-b-fdrw0	ah pcl p l iy	ended abruptly at	ptpp	pbppbpp
332	sx184-b-fedw0	ix kcl k ae tcl	the cat's meow	ktkk	kkkkkkk
333	sx379-b-mcsh0	z dcl d r ay	gives driving lessons	gggp	dttdtdd
334	si1502-b-fdms0	ae kcl k l ix	not exactly panic	kdkk	ggkggkg
335	si1502-b-fdms0	ih kcl k dh ey	exactly panic they	dddd	kkkkkkk
336	si1502-b-fdms0	ey gcl g ey v	they gave way	kggg	ggggggg
337	sx110-b-frew0	eh kcl k s tcl	solve extra problems	gkkk	gkkkkkg
338	sx110-b-frew0	aa bcl b l ax	extra problems #	gbbd	bbbbbbb
339	si1723-b-mrtk0	ix dcl d r ih	idea drifted in	ddgd	ddddddd
340	sx118-b-mmjr0	# h# k ax bcl	# combine all	tttp	ttttttt
341	sx118-b-mmjr0	n tcl t s pau	the ingredients in	gddt	ggtttdt
342	si2104-b-fdac1	sh dcl d ix v	lush divorcee at	dtdt	tdttttt
343	sx208-b-fsem0	aa bcl b s eh	# rob sat	tddd	bbbbbbb
344	sx208-b-fsem0	s kcl k eh tcl	and sketched the	kkdg	kkkkdkk
345	sx208-b-fsem0	ey gcl g iy s	stray geese #	kkkg	ggggggg
346	si818-b-fgjd0	ix tcl p ao r	transmit poorly and	tppt	ppppppp
347	sx102-b-mpdf0	s pcl p eh sh	# special task	bbbp	pbpppbp
348	sx102-b-mpdf0	ae pcl p axr z	from kidnappers #	gggp	bbbbbbb
349	sx109-b-fadg0	ih pcl t eh n	she slipped and	tddd	tdtdttt
350	sx109-b-fadg0	ng kcl k el q	her ankle on	kgkk	ggggggg

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
351	sx109-b-fadg0	s tcl t iy pcl	the steep slope	dttt	tdttttt
352	si1133-b-mteb0	eh kcl t l ao	can project long	kkkk	tttktkt
353	si1133-b-mteb0	jh gcl g ow l	range goals for	dkkg	gpgkkgg
354	sx404-b-fjsj0	ey pcl p er en	burned paper and	ppgp	ppppppp
355	sx404-b-fjsj0	ax bcl b ih gcl	a big bonfire	ddbb	bbkbbbb
356	si1375-b-mreb0	n pau d aa m	often dominant stress	ttdt	dddddt
357	si1375-b-mreb0	m pcl p l iy	stress simply indicates	pkkp	pppkpkp
358	si1853-b-mrcs0	eh n t s ix	thus events occurred	dttt	tkdtdtt
359	si1853-b-mrcs0	v dcl d h# #	ones arrived #	tttt	ttdttdt
360	sx227-b-mroa0	axr tcl t ax-h kcl	healthier to cook	dddd	dgddddd
361	sx227-b-mroa0	uh gcl g axr h#	without sugar #	bbgg	ggggggg
362	sx140-b-mbns0	eh gcl g axr kcl	# agricultural products	dggg	ggggggg
363	si551-b-fjmg0	s tcl t ix kcl	an autistic child	gttt	kgttttt
364	sx48-b-fdms0	aw tcl g r ux	grandmother outgrew her	kkkk	kkgkgkg
365	sx48-b-fdms0	ah pcl b r iy	her upbringing in	pppp	pppppbp
366	si1850-b-mbns0	uh dcl d m ao	a good morrow	dgdd	ddpdddg
367	si1496-b-mrws1	ix gcl g z ah	was exultantly easy	kkkk	ggkkkgg
368	si1496-b-mrws1	ix n t l ix	was exultantly easy	kdtd	tpppddp
369	sx290-b-frew0	s bcl b r ow	this brochure is	pppp	ppppptp
370	sx290-b-frew0	ih kcl k y axr	is particularly informative	dkkk	dkdkgdk
371	sx290-b-frew0	ax tcl t ih v	particularly informative for	ttdd	ttttttt
372	sx290-b-frew0	s pcl p eh kcl	a prospective buyer	pptp	pbppppp
373	sx290-b-frew0	eh kcl t ix v	a prospective buyer	dttt	tdtdgdd
374	sx115-b-mtaa0	m dcl d ix pcl	emblem depicts the	tttt	tdtdttt
375	sx115-b-mtaa0	ah gcl g l ow	all aglow #	kggg	gkggggg
376	sx138-b-fdms0	s pcl p ih l	customer spilled some	bbpp	ppppppp
377	sx138-b-fdms0	s pcl p eh n	some expensive perfume	bbtp	pbppppp
378	sx109-b-mcsh0	ih pcl t ih n	she slipped and	ddtt	tddttpt
379	sx109-b-mcsh0	s pcl p r ey	and sprained her	bbbp	pbppppp
380	sx109-b-mcsh0	s tcl t iy pcl	the steep slope	bdpt	ttttttt
381	sx109-b-mcsh0	iy pcl p s l	the steep slope	pkpp	ppbpppp
382	sx143-b-mdlf0	ux gcl g el s	jennifer's bugle scared	gbgb	ggggggg
383	sx143-b-mdlf0	s kcl k eh r	bugle scared the	kkgg	kkkkkkk
384	sx44-b-fjsj0	# h# b ae s	# basketball can	pbbb	bbbbbbb
385	sx44-b-fjsj0	s kcl k ax bcl	# basketball can	tkkk	kkkgkkk
386	sx44-b-fjsj0	l kcl k em bcl	basketball can be	tkkk	kkkkkkk
387	sx44-b-fjsj0	iy ng g s pcl	an entertaining sport	kkkg	gkkkkgk
388	si634-b-fjem0	ay tcl p axr s	quite persuasively the	tppp	ppppppp
389	sx320-b-fkms0	aa gcl g q m	nearest synagogue may	ddgg	ggggggg
390	sx320-b-fkms0	aa tcl b iy w	not be within	bbpb	bbbbbbb
391	sx373-b-mrtk0	iy pcl p ix ng	superb sleeping pill	bbkp	ppppppp
392	sx373-b-mrtk0	ng pcl p ih l	sleeping pill #	ppkp	ppppppp
393	si2300-b-fmml0	z bcl b ix s	was beside herself	pppp	ppppptp
394	si2114-b-fjsj0	s tcl t ey kcl	# stake me	dttt	ttttttt
395	si2114-b-fjsj0	z dcl d ay s	those dice #	ttdd	ddtdddd
396	si1279-b-fadg0	# h# b r ih	# bricks are	pbbb	pbppppb
397	sx385-b-mthc0	axr bcl b r ow	were broken #	bpbb	bbbbbdb
398	sx318-b-fdms0	z dcl d ih gcl	please dig my	dttt	tdttttt
399	sx318-b-fdms0	ay pcl p ax tcl	my potatoes up	dbtp	gdbbpdp
400	sx318-b-fdms0	ah pcl b iy f	up before frost	bbpb	bpbbbbb

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
401	si2293-b-mjfc0	eh tcl t q ao	ones that all	dddd	ddddddd
402	si2293-b-mjfc0	f tcl t eh tcl	grownups laughed at	dttt	ddtddtd
403	si2293-b-mjfc0	eh tcl t l aw	laughed at loudest	dbdd	ddddddd
404	si1019-b-mers0	z pcl p ih l	as pilgrims #	ptkp	ppppppp
405	sx404-b-mwjg0	r bcl b bcl b	# barb burned	dppp	ppbbbpb
406	sx404-b-mwjg0	b bcl b er n	barb burned paper	pbbp	bbbbbbb
407	sx404-b-mwjg0	ix bcl b ih gcl	a big bonfire	bbpp	bbbgbbb
408	sx404-b-mwjg0	ih gcl b aa n	big bonfire #	bppp	bpbbbgb
409	sx28-b-fsem0	# h# b ey gcl	# beg that	pbpp	gdbbddb
410	sx28-b-fsem0	ae tcl g aa r	that guard for	ggdd	ggggggg
411	sx28-b-fsem0	n gcl g ae l	one gallon of	gddg	ggggggg
412	sx28-b-fsem0	v gcl g ae s	of gas #	kttt	dgggggg
413	sx380-b-mjsw0	ch gcl g aa r	such garbage #	bggg	ggggggg
414	sx410-b-mrws1	iy kcl d ey m	each weekday morning	tttt	ddddkdd
415	sx102-b-mmdm2	s pcl p eh sh	# special task	bbpb	pbppppp
416	sx102-b-mmdm2	s kcl k f ao	special task forces	kgkk	kkkkkkk
417	sx102-b-mmdm2	ae pcl p axr s	from kidnappers #	gggg	bbbppdt
418	si1285-b-mtaa0	l bcl b iy n	will be necessary	pbbb	pppbpbp
419	sx47-b-mroa0	s tcl t r ao	mom strongly dislikes	kktk	ttttttt
420	sx47-b-mroa0	ae pcl p tcl t	dislikes appetizers #	dtpp	ppppppp
421	sx14-b-fcmh0	z dcl d ih z	before thursday's exam	tttt	ttttttt
422	si854-b-fjsj0	n kcl k er ix	# encourage them	tkkk	kkkkkkk
423	si854-b-fjsj0	r bcl b eh n	their benefits #	dbbb	bbbbbbb
424	sx99-b-mmwh0	z bcl b ey s	was based on	pbbb	bbbbpbp
425	sx314-b-fjsj0	f pcl p iy pcl	if people were	ptpp	bbppptp
426	sx314-b-fjsj0	iy pcl p el q	if people were	gpdp	ppppkpp
427	sx314-b-fjsj0	iy dcl d f ao	no need for	tkkd	dgddddd
428	si677-b-mroa0	el tcl t r eh	comparable trends can	kkkk	ttttttt
429	si677-b-mroa0	z kcl k en b	trends can be	kktk	kkkkkkk
430	si590-b-mbns0	# h# b r ih	# bridges tunnels	pbpp	pbbppbp
431	sx189-b-fgjd0	# h# d ih s	# destroy every	ttdt	ttttttt
432	sx274-b-fedw0	ax gcl g zh ux	the luxurious massage	kkkk	gggggkg
433	sx364-b-mtdt0	s tcl t ix m	heating system part	dttt	ttttttt
434	sx200-b-frew0	ix n t r ow	ralph controlled the	tptt	ttttttt
435	sx224-b-mbwm0	n tcl t s h#	your endurance #	dtdt	gktkttt
436	si1264-b-fjem0	# h# d ae n	# dances alternated	tddd	ddddddd
437	sx208-b-mmjr0	aa bcl b s ae	# rob sat	pbbb	bgbbbbb
438	sx208-b-mmjr0	n dcl d pau en	the pond and	dttd	ptttdtt
439	sx185-b-mmdb1	# h# d ix dcl	# did you	tdtd	ddddddd
440	sx115-b-mthc0	iy pcl p ih kcl	emblem depicts the	ttpp	pppppkp
441	si2093-b-mbdg0	s tcl t ix kcl	was unenthusiastic #	dttt	ttttttt
442	si1910-b-frew0	ix n b r iy	and breathed for	bgbg	bbbbbbb
443	si1910-b-frew0	ng kcl k ah m	mining company #	kkdk	kkkkkkk
444	si1910-b-frew0	m pcl p ix n	mining company #	ptpp	ppppbdp
445	sx139-b-mrjm4	ng gcl g ax l	the bungalow was	bbgg	ggbdgbg
446	sx50-b-mrws1	tcl bcl b ae kcl	economic cutbacks neglect	pbpb	bbbkbbb
447	sx50-b-mrws1	ix gcl g l eh	cutbacks neglect the	kgkk	kgggggg
448	si943-b-faks0	s pcl p ix kcl	below expectations #	bbpp	pbppppp
449	sx233-b-mrcs0	ng gcl g h# #	same thing #	dddd	gkdggdg
450	si1909-b-fadg0	z bcl b ix n	rolls beneath the	ppbp	pbppppp

	Utterance Tag	TIMIT Phonetic labels	Orthography	Machine Hypotheses A,B,C,Hetero	Perceptual Hypotheses 7 subjects
451	sx110-b-mjsw0	ax-h bcl b eh s	the best way	dddb	bbbbbbb
452	sx110-b-mjsw0	s tcl t ax-h s	is to solve	ttdt	ddttttt
453	sx317-b-mroa0	ih tcl t iy ih	very pretty in	dddd	ddddddd
454	sx230-b-mbns0	pau bcl b ax tcl	here but rationalize	bbpb	bbbbpbb
455	sx413-b-mteb0	iy dcl d ix sh	daphne's swedish needlepoint	ggdg	ddddddd
456	sx369-b-mmwh0	# h# b aa bcl	# bob papered	ppbp	bgbbkbb
457	si1124-b-mwjg0	ow tcl t hh ix	me quote him	dtdt	ttttddd
458	sx365-b-mmdb1	ix gcl g ah n	the gunman kept	tkkk	kkgkgkk
459	sx365-b-mmdb1	ih kcl t em kcl	his victim cornered	dttt	tdtgkkt
460	sx365-b-mmdb1	em kcl k ao r	victim cornered at	kkgk	kkkkkkk
461	sx365-b-mmdb1	ae tcl g ah n	at gunpoint for	tgdt	gggggtk
462	si904-b-mglb0	ow dcl t r ae	railroad tracks came	kttt	ttttttt
463	si904-b-mglb0	n tcl t ix v	came into view	dddd	ddddddd
464	sx29-b-mers0	ih kcl k ix pcl	to pick a	kktk	kkkkkkk
465	sx29-b-mers0	ax-h pcl p ax-h tcl	of potatoes #	pttt	pptpttp
466	sx29-b-mers0	ax-h tcl t ey dx	of potatoes #	ttpp	ttttttt
467	sx293-b-mbdg0	ey kcl k dh ih	please take this	kkgk	ggtgggt
468	sx293-b-mbdg0	s dcl d er dx	this dirty tablecloth	dttt	tdttttt
469	sx223-b-mjfc0	axr bcl b l aa	butcher block table	pbbb	pbpbbpb
470	si2275-b-mthc0	ih dcl d l uh	all did look	bddb	ddgdddd
471	si2275-b-mthc0	ih pcl p sh ey	and shipshape #	kggk	bpppppp
472	si2275-b-mthc0	ey pcl p h# #	and shipshape #	tptp	dpttpbp
473	si1484-b-fjsj0	iy kcl k axr z	honest seekers after	gkkk	kgkkggg
474	sx191-b-mgjf0	th dcl d r ao	# withdraw all	tttt	dttttd
475	sx191-b-mgjf0	ae kcl k ux z	phony accusations at	ktkk	tkkkkkk
476	si1970-b-mroa0	ax bcl b w q	clearly identifiable enemy	ggkb	bgbbkgb
477	si1970-b-mroa0	iy kcl k en tcl	enemy continued on	tttk	kkkkkkk
478	si1970-b-mroa0	ux dcl d aa n	enemy continued on	tdgt	ddddddd
479	si2179-b-mcsh0	aa pcl p s epi	never wops me	tkpp	pgppppp
480	si459-b-mmwh0	axr pcl p axr pcl	further proposed grants	bppp	ppppppp
481	si459-b-mmwh0	z dcl g r ae	proposed grants of	kkgg	kgkkkkk
482	si459-b-mmwh0	s pcl p eh s	an unspecified sum	bbbp	bbppppp
483	si459-b-mmwh0	s pcl p er m	for experimental hospitals	bbbp	pbppppp
484	si459-b-mmwh0	n tcl t el hh	for experimental hospitals	tddp	ttttdtt
485	si459-b-mmwh0	s pcl p ix dx	experimental hospitals #	bbpp	ppppppp
486	sx282-b-mrjr0	axr gcl g aa tcl	fairy forgot to	ggdd	ggggggg
487	si745-b-mreb0	# h# b r oy	# broil or	pppp	bbbbbpb
488	sx188-b-mjar0	s pcl p eh n	unlimited expense account	bbpp	pbpbppp
489	si1714-b-fedw0	s tcl t eh pcl	normally step in	dttt	tdttttt
490	si1714-b-fedw0	eh pcl p ih ix	normally step in	pbbp	bbpbpbp

Appendix B

Antiphone Modeling and Near-miss Modeling

Antiphone modeling and near-miss modeling both refer to probabilistic frameworks for segment-based speech recognition. This discussion is drawn from [6, 7]. The material is reproduced here to aid the reader in following the phonetic recognition experiments in Chapter 5.

Speech recognition systems are designed to find the sequence of words, W^*, that maximizes the *a posteriori* probability of the acoustic observations, A. This is written mathematically as

$$W^* = \arg\max_W P(W|A) = \arg\max_W P(A|W)P(W). \tag{B.1}$$

where $P(A|W)$ is the acoustic model and $P(W)$ is the language model.

In frame-based systems, the acoustic observations A are a non-overlapping time-series. Thus, every segmentation of the speech signal accounts for all observations. In segment-based systems, A is a temporal network. For each possible path through the network, some segments are on the path and others are off the path. Thus,

$$P(A|W) = P(A_{on}A_{off}|W) \tag{B.2}$$

where A_{on} refers to the on-path segments, and A_{off} refers to the off-path segments for a particular path through the network. We now describe two different ways to compute $P(A_{on}A_{off}|W)$.

B.1 Antiphone Modeling

This algorithm uses a single non-lexical acoustic model, referred to as the "not" model, or the "antiphone" to account for all off-path segments. Furthermore, this is done in an efficient manner by assuming independence between the on-path and off-path acoustic observations. Let $\widehat{w}$ represent the nonlexical unit, then

$$P(A_{on}A_{off}|W) = P(A_{on}|W)P(A_{off}|\widehat{w})\frac{P(A_{on}|\widehat{w})}{P(A_{on}|\widehat{w})} = K\frac{P(A_{on}|W)}{P(A_{on}|\widehat{w})} \tag{B.3}$$

where K is a constant for all paths through the network. Instead of scoring A_{on} against the lexical models and A_{off} against the antiphone model, A_{on} is scored against both lexical and nonlexical models.

B.2 Near-miss Modeling

Although antiphone modeling is efficient, it uses only a single non-lexical acoustic model to account for all off-path segments. The idea of antiphone modeling has been generalized to near-miss modeling, which uses multiple nonlexical units to model off-path segments as "near-misses" of lexical units that are on the path. In order to calculate $P(A_{on}A_{off}|W)$, every off-path segment is associated with one and only one on-path segment. Let the on-path lexical units be $W = \{w_1w_2...w_N\}$. Let $A_{\widehat{w}_n}$ be the acoustic observations associated with segments in the near-miss subset of w_n, so that

$$A_{\widehat{w}_n} \bigcap A_{\widehat{w}_m} = \emptyset \text{ for } n \neq m \tag{B.4}$$

$$A_{off} = \bigcup_{n=1}^{N} A_{\widehat{w}_n} \tag{B.5}$$

The scoring them proceeds as

$$P(A_{on}A_{off}|W) = P(A_{on}|W) \prod_{n=1}^{N} P(A_{\widehat{w}_n}|\widehat{w}_n). \tag{B.6}$$

In this scenario, off-path segments are explicitly scored against the appropriate near-miss models. For further discussion and comparison of near-miss and antiphone modeling techniques, please consult the references.

Appendix C

Aggregation of Probabilistic Models

C.1 Introduction

Model aggregation is a means of improving the performance and robustness of mixture Gaussian models.[1] This technique produces models that are more accurate and more robust to different test sets than traditional cross-validation using a development set. In this appendix, a theoretical justification for this technique is presented with some discussion of its effectiveness.

Mixture Gaussian models are typically trained using a procedure that combines supervised and unsupervised training. Supervision is provided through the class labels of the training data which are known during training. In contrast, the weights which determine how much each data point contributes to the training of the mean and covariance of each individual mixture component are not known when training begins, but rather are determined during the training process in an unsupervised manner. The algorithms used to determine these weights, such as the K-means clustering algorithm and the Expectation-Maximization (EM) algorithm, do not guarantee a globally optimal solution. These algorithms often converge to a locally optimal solu-

[1]This appendix draws from joint work with T. J. Hazen, reported in [33].

tion, where the exact local optimum that will be reached is highly dependent on the initialization of the unknown weights at the beginning of the training process. As a specific example of this phenomenon, word recognition accuracies on the Resource Management task were obtained from 24 trials of training mixture Gaussian models by randomly-initialized K–means clustering followed by EM iterations. The average word error rate was 4.55%, with a maximum, minimum and standard deviation of 4.83%, 4.33%, and 0.127, respectively. The error rate reduction achieved in traversing from the worst trial to the best trial is 10.4%. In light of this variation, one may ask: What is a good strategy for using the results on development data to choose which models to use on an independent test set?

In the past, many have simply chosen the model training trial that performs the best on the development data, i.e. cross-validation. One problem with this strategy is that noise on the development data contributes a random component to the performance. As a result, better performance on the development set may not indicate models which are better matched to the true underlying distribution of the data. Instead, it may only indicate that the models are superficially better matched to the idiosyncrasies of the development set. As an example of this, TIMIT phonetic recognition accuracies were calculated on development and test sets for 24 independently trained models, using different random initializations in the K–means clustering. The correlation between development set and test set accuracy was indeed weak (correlation coefficient 0.16), and in this case the simple "pick-the-best-on-development-data" cross-validation strategy was particularly disappointing since the trial performing best on the development set performed worst on the test set. A second disadvantage of simple cross-validation is that computation is wasted, since the results of only one training trial are kept, while the models from the other trials are thrown away [1].

To counter the problems discussed above, an algorithm is needed which produces a mixture density function which can be proven to yield better classification accuracy, on average, than any randomly initialized density function trained using standard techniques. At the very least, it is desirable to show the new algorithm can reduce the value of some error function which is strongly correlated with classification accuracy. Aggregation is a technique which meets this criterion.

C.2 Theoretical Basis

Aggregation of probabilistic classifiers is performed by averaging the outputs of a set of independently trained classifiers. The proof that follows will demonstrate that an aggregate classifier is guaranteed to exhibit an error metric which is equal to or better than the average error metric of the individual classifiers used during aggregation. Though the empirical evidence presented in this paper uses only mixture Gaussian classifiers, this proof is valid for any type of probabilistic classifier. This proof is also completely independent of the test data being presented to the classifier. Thus, the method is robust because it improves performance regardless of the test set being used.

To begin, assume a set of N different classifiers have been trained. In most of the experiments in this paper, multiple classifiers are generated from the same data set by using different random initializations in the K–means clustering prior to EM training of the mixtures. However, the proof does not depend in any way on how the classifiers are generated. This set will be called Φ and can be represented as:

$$\Phi = \{\vec{\varphi}_1(\vec{x}), \vec{\varphi}_2(\vec{x}), \ldots, \vec{\varphi}_N(\vec{x})\} \tag{C.1}$$

Within Φ, each classifier $\vec{\varphi}_n(\vec{x})$ takes an observation vector $\vec{x}$ as its input. The underlying class that $\vec{x}$ belongs to will be defined as $c(\vec{x})$. To classify $\vec{x}$, each $\vec{\varphi}_n(\vec{x})$ contains a probability density function for each of the D different classes from which $\vec{x}$ might be drawn. Each classifier $\vec{\varphi}_n(\vec{x})$ outputs a D dimensional vector containing the *a posteriori* probabilities of $\vec{x}$ belonging to each of the D classes. This output vector for $\vec{\varphi}_n(\vec{x})$ can be represented as

$$\vec{\varphi}_n(\vec{x}) = \begin{bmatrix} \varphi_{n,1}(\vec{x}) \\ \varphi_{n,2}(\vec{x}) \\ \vdots \\ \varphi_{n,D}(\vec{x}) \end{bmatrix} = \begin{bmatrix} \mathrm{P}_n(c(\vec{x}) = 1 \mid \vec{x}) \\ \mathrm{P}_n(c(\vec{x}) = 2 \mid \vec{x}) \\ \vdots \\ \mathrm{P}_n(c(\vec{x}) = D \mid \vec{x}) \end{bmatrix} \tag{C.2}$$

In order to evaluate the performance of a classifier, an appropriate metric must be defined. Let $\vec{y}(\vec{x})$ be a D dimensional vector which indicates the reference class, or "correct answer", through a binary representation. That is,

$$\vec{y}(\vec{x}) = [y_1(\vec{x}), y_2(\vec{x}), \ldots, y_D(\vec{x})]^T \tag{C.3}$$

where

$$y_d(\vec{x}) = \begin{cases} 1 & \text{if } \mathrm{c}(\vec{x}) = d \\ 0 & \text{otherwise} \end{cases} \tag{C.4}$$

Ideally, a classifier's *a posteriori* probability for the correct class should be as close to 1 as possible while all of the incorrect classes should have *a posteriori* probabilities as close to 0 as possible. The error metric used in this proof is the squared distance between the classifier's output and the "correct answer". Thus, the *error* when input $\vec{x}$ is presented to the nth classifier is defined as

$$e_n(\vec{x}) = \|\vec{y}(\vec{x}) - \vec{\varphi}_n(\vec{x})\|^2 \tag{C.5}$$

This error can be expressed as

$$e_n(\vec{x}) = \sum_{d=1}^{D} (y_d(\vec{x}) - \varphi_{n,d}(\vec{x}))^2 \tag{C.6}$$

Using the error metric defined above, the mean error over all N classifiers in Φ

for an input vector $\vec{x}$ is

$$e(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}\left(y_d(\vec{x}) - \varphi_{n,d}(\vec{x})\right)^2 \tag{C.7}$$

For notational ease, the remainder of this development will drop the explicit dependence on the input vector $\vec{x}$. It should be understood that the analysis proceeds identically given any input vector. Continuing, the average error of the N classifiers expands to

$$e = \sum_{d=1}^{D}\left[y_d^2 - \left(\frac{2y_d}{N}\sum_{n=1}^{N}\varphi_{n,d}\right) + \left(\frac{1}{N}\sum_{n=1}^{N}\varphi_{n,d}^2\right)\right] \tag{C.8}$$

The aggregate classifier, $\vec{\varphi}_A$, simply averages the outputs of all N classifiers in Φ. This is expressed as

$$\vec{\varphi}_A = \frac{1}{N}\sum_{n=1}^{N}\vec{\varphi}_n \tag{C.9}$$

The error for the aggregate classifier model is

$$e_A = \|\vec{y} - \vec{\varphi}_A\|^2 = \sum_{d=1}^{D}(y_d^2 - 2y_d\varphi_{A,d} + \varphi_{A,d}^2) \tag{C.10}$$

By substituting in the definition of $\vec{\varphi}_A(\vec{x})$ from Equation (C.9) the error of the aggregate classifier can be rewritten as

$$e_A = \sum_{d=1}^{D}\left[y_d^2 - \left(\frac{2y_d}{N}\sum_{n=1}^{N}\varphi_{n,d}\right) + \left(\frac{1}{N}\sum_{n=1}^{N}\varphi_{n,d}\right)^2\right] \tag{C.11}$$

By comparing the expressions in Equations (C.8) and (C.11), it can be seen that e_A will be less than or equal to e if

$$\left(\frac{1}{N}\sum_{n=1}^{N}\varphi_{n,d}\right)^2 \leq \frac{1}{N}\sum_{n=1}^{N}\varphi_{n,d}^2 \tag{C.12}$$

In fact, this condition is always true for any arbitrary vector because it is a special case of the Cauchy–Schwarz inequality. Given any two vectors $\vec{a} = [a_1, a_2, \ldots a_N]^T$

and $\vec{b} = [b_1, b_2, \ldots b_N]^T$, then by the Cauchy–Schwartz inequality

$$\left|\sum_{n=1}^{N} a_n b_n\right|^2 \leq \left(\sum_{n=1}^{N} a_n^2\right)\left(\sum_{n=1}^{N} b_n^2\right) \tag{C.13}$$

Now let $b_n = 1$ for all n so that $\sum_{n=1}^{N} b_n^2 = N$ to obtain

$$\left(\sum_{n=1}^{N} a_n\right)^2 \leq N \sum_{n=1}^{N} a_n^2 \tag{C.14}$$

which is the desired result. Note that equality holds in Equation (C.12) only if the $\varphi_{n,d}$ matrix is constant along each row n, i.e. every classifier is giving exactly the same *a posteriori* probabilities. Thus, in practical situations with classifiers that produce different probabilities, the inequality becomes strict.

This derivation proves that, for *any* input token $\vec{x}$, the error e_A of the aggregate classifier created from the classifier set Φ is always smaller than the average error e of the N individual classifiers in Φ, provided that the N classifiers do not all produce the same *a posteriori* probabilities.

C.3 Experimental Effectiveness

The previous theoretical discussion leads us to expect aggregation to be helpful for performance, but it does not predict the magnitude of the improvements that will be observed. Thus, empirical studies are needed in order to obtain a sampling of the improvements that can be obtained in typical speech recognition tasks. These empirical studies are found in [33], and they report error rate reductions of up to 12% in the tasks of phonetic classification, phonetic recognition, and word recognition. These performance improvements come at the price of increased computation. If N trials of model training are aggregated, then the final aggregated model is N times larger than the original, thus requiring N times more computation. Pruning and fast-match techniques may be used help reduce the burden of the increase in computational requirements.

Appendix D

Confusion Matrices

This appendix contains detailed performance results for the committee of nine classifiers discussed in Section 5.1.3. Confusion matrices calculated on the development and core test sets are given in Tables D.1 and D.2. The reference phonetic class is given along the rows, so that the percent error for each phonetic label (after mapping to the 39 classes given in Table 2.2) can be calculated from the numbers along each row. The hypotheses are given along the columns.

	iy	ih	eh	ae	ax	uw	uh	ao	ey	ay	oy	aw	ow	er	l	r	w	y	m	n	ng	v	f	dh	th	z	s	zh	jh	ch	b	p	d	dx	t	g	k	hh	cl		% err
iy	502	45				2			9						1			5																						iy	11.0
ih	45	978	48	2	74	5	1		17				2	11	1	5		5		4	2			1																ih	18.6
eh	5	68	236	26	29			2	9	1	1	2	1	2		9																								eh	39.6
ae		8	39	161	8			3	3	8		5		2																										ae	32.1
ax		144	31	6	382		2	24		10		5	13	2	8	10			1	3													1	1					8	ax	41.3
uw	11	34			6	88		3	1				6	4	7	1	3																							uw	46.3
uh		22	4	1	11	2	7	1					1	3	1		1																							uh	87.0
ao			3	5	16			383		17		5	8	3	12	3	3																					1		ao	16.6
ey	17	17	8			1			185	2																														ey	19.6
ay		2	5	1	12			11	3	175		2		1		1																						1		ay	18.2
oy					1			1		4	24																													oy	20.0
aw			1	7	4			10		1		33	5																											aw	45.9
ow		3	7	1	18	2		12	1		1	12	97	2	12																									ow	42.3
er	1	30	7		13	1	2	2	2				2	383	2	67		1		1																				er	25.5
l					16	2	1	8			1	1	15		545	10	28	1	4	2		3		6										1				1	2	l	15.8
r	2	13	1		6	1		2		4				32	3	468	1	2	1	1		3		1										1					1	r	13.8
w		1			1	3		7							38	6	217		1			1		2							1								1	w	22.2
y	24	8														1		74		2																	1	3	2	y	35.7
m		1			1										1	5	1		329	86	3	1		2							4								1	m	24.4
n		16		1	4	1									6	1	1		23	713	15	2		1							1		1	11					10	n	11.6
ng		1																	2	32	63																			ng	35.7
v		1			1										1	1	1		2	12		156	10	4	1	2					3	1	2	6		1		2	16	v	30.0
f																						5	253	1	7	1	5				1							1	6	f	9.6
dh		1			1										1					8		7	2	187	5						7	3	19	5	6	2	2	2	3	dh	28.4
th																						2	14	8	34	5	3								5				5	th	55.3
z					1																			1		262	97	1												z	27.6
s																							4		1	48	600	2	2						4					s	9.2
zh																										5	8	135	2	3										zh	11.8
jh																										3	3	2	58	14			1		5					jh	32.6
ch																										1	3	1	7	48					4					ch	25.0
b															1		1							8							190	34	12	1		2				b	23.7
p														1		1								7							26	229	1		10		5	1		p	18.5
d					1																			7					3	1	5	2	168		41	7	2		2	d	29.7
dx					2											1				7				1										146				1	4	dx	9.9
t		1																					2	2		2	3		2	6	1	6	32		340		11		5	t	17.7
g		1															1			1				1							1	1	8		2	80	30		1	g	37.0
k																							1							1		3	3		7	7	352		2	k	6.4
hh	2	1						1	1							1	2	1	1	2		1	3	2								3		1	1		4	142	5	hh	18.4
cl		2			4												3	2	9	36	1	13	2	8	1						1	1	1	12	1			6	2905	cl	3.4
	iy	ih	eh	ae	ax	uw	uh	ao	ey	ay	oy	aw	ow	er	l	r	w	y	m	n	ng	v	f	dh	th	z	s	zh	jh	ch	b	p	d	dx	t	g	k	hh	cl		18.12

Table D.1: Confusion matrix on the development set for the 9-member committee of classifiers described in Section 5.1.3. The reference labels are on the left, the hypothesis labels are along the top.

	iy	ih	eh	ae	ax	uw	uh	ao	ey	ay	oy	aw	ow	er	l	r	w	y	m	n	ng	v	f	dh	th	z	s	zh	jh	ch	b	p	d	dx	t	g	k	hh	cl		% err
iy	213	18			1	3			5									3																						iy	12.3
ih	20	460	21	2	38	5	1		9				1	6	2	5	1	2		7																				ih	20.7
eh	2	31	113	15	18		1	2	3					2		2																								eh	40.2
ae	1	4	20	61	4			1	6	5		3																												ae	41.9
ax		48	15	3	207			11		1	1		7	5	9	6	1			3																		1	3	ax	35.5
uw	3	17		1	3	43								2	1	1	1	1		1																				uw	41.9
uh		9	2		10		4	1		1				1		1																								uh	86.2
ao			1	6	12			183		9	1	3	5	3	2	3																								ao	19.7
ey	9	7	4	1		1			90	2																														ey	21.1
ay			1	2	4			8	4	69			1																											ay	22.5
oy			1							2	12			1																										oy	25.0
aw				4	1			1				22	2																											aw	26.7
ow		1	2	1	10	1		8				2	55	1	7		1																							ow	38.2
er		18	3	2	2	2		2		1			2	172		26	1	1		1																			1	er	26.5
l		2			5		1	4		1		1	6	1	233	10	13		4	4		5		1																l	19.9
r		1												11	1	253	1				1	2																		r	6.3
w					2	1		6							26	2	104					2									1									w	27.8
y	12	2													1			34																1						y	32.0
m					1											3			149	42	1	2		1							3								2	m	27.0
n		4			1	1									1				14	336	13			1							1	2		8					3	n	12.7
ng																			3	15	32													1					1	ng	38.5
v																			2	11		64		3	2	3							2	1					5	v	31.2
f																						3	116	3	2		1					3			1		2			f	11.5
dh															1					2		5	2	87		2					1	1	15	6	5			1		dh	32.0
th																						1	3	3	20	1	2								5				3	th	47.4
z																									1	134	44	2												z	26.0
s																							1		1	17	297	1		1					3					s	7.5
zh																							1			2	6	64		3					1					zh	16.9
jh																											1	2	33	5					1					jh	21.4
ch																											3	2	4	25					5		1			ch	37.5
b																	2							2							106	17	3				1	1		b	19.7
p																							1	5							13	105	2		5		4		1	p	22.8
d		1																						8					2		3		75		19	1	3		1	d	33.6
dx																		1		5														82					2	dx	8.9
t																								2			4		1			3	10		166				1	t	11.2
g																		1						1								1	6			42	15			g	36.4
k																														1		3	2		3	5	150		1	k	9.1
hh		2													1	1		1	1	3				1											1		3	62	1	hh	19.5
cl						1								1	1		1		5	7	3	6	1	3	1	1								6				2	1422	cl	2.7
	iy	ih	eh	ae	ax	uw	uh	ao	ey	ay	oy	aw	ow	er	l	r	w	y	m	n	ng	v	f	dh	th	z	s	zh	jh	ch	b	p	d	dx	t	g	k	hh	cl		18.30

Table D.2: Confusion matrix on the core test set for the 9-member committee of classifiers described in Section 5.1.3. The reference labels are on the left, the hypothesis labels are along the top.

Bibliography

[1] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[2] B. P. Bogert, M. J. R. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudoautocovvariance, cross-cepstrum, and saphe cracking. In M. Rosenblatt, editor, *Proc. Symposium Time Series Analysis*, pages 209–243, New York, 1963. John Wiley and Sons.

[3] H. Bourland and S. Dupont. Subband-based speech recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1251–1254, 1997.

[4] G. L. Bradshaw, R. A. Cole, and Z. Li. Comparison of learning techniques in speech recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 554–557, 1982.

[5] R. Carlson and J. Glass. Vowel classification based on analysis-by-synthesis. In *Proc. Int. Conf. Spoken Language Processing*, pages 575–578, Banff, Canada, October 1992.

[6] J. Chang and J. Glass. Segmentation and modeling in segment-based recognition. In *Proc. European Conf. Speech Communication and Technology*, pages 1199–1202, Rhodes, Greece, September 1997.

[7] J. W. Chang. *Near-Miss Modeling: A Segment-Based Approach to Speech Recognition.* Ph. D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, June 1998.

[8] R. Chengalvarayan and L. Deng. HMM-based speech recognition using state-dependent, discriminatively derived transforms on Mel-warped DFT features. *IEEE Trans. Speech and Audio Processing*, 5(3):243–256, May 1997.

[9] R. Chengalvarayan and L. Deng. Use of generalized dynamic feature parameters for speech recognition. *IEEE Trans. Speech and Audio Processing*, 5(3):232–242, May 1997.

[10] B. Chigier. Phonetic classification on wide-band and telephone quality speech. In *Proc. 5th DARPA Speech and Nat. Lang. Workshop*, February 1992.

[11] B. Chigier and R. Brennan. The effects of signal representations, phonetic classification techniques, and the telephone network. In *Proc. Int. Conf. Spoken Language Processing*, pages 97–100, Banff, Canada, October 1992.

[12] W. Chou, C. H. Lee, and B. H. Juang. Minimum error rate training of inter-word context dependent acoustic model units in speech recognition. In *Proc. Int. Conf. Spoken Language Processing*, pages S09:3.1–3.4, Yokohama, Japan, September 1994.

[13] R. T. Chun. *A Hierarchical Feature Representation for Phonetic Classification.* M.Eng. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, March 1996.

[14] J. E. Clark. Intelligibility comparisons for two synthetic and one natural speech source. *Journal of Phonetics*, 11:37–49, 1983.

[15] R. Cole, M. Fanty, Y. Muthusamy, and M. Gopalakrishnan. Speaker-independent recognition of spoken English letters. In *Proc. Int. Joint Conf. Neural Networks*, volume II, pages 45–51, San Diego, CA, June 1990.

[16] R. A. Cole and Y. K. Methusamy. Perceptual studies on vowels excised from continuous speech. In *Proc. Int. Conf. Spoken Language Processing*, pages 1091–1094, Banff, Canada, October 1992.

[17] R. A. Cole, R. M. Stern, and M. J. Lasry. Performing fine phonetic distinctions: Templates versus features. In J. S. Perkell and D. H. Klatt, editors, *Invariance and variability in speech processes*, pages 325–345. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[18] N. Daly. *Recognition of Words from their Spellings: Integration of Multiple Knowledge Sources.* M.S. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, May 1987.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39:1–38, 1977.

[20] N. Deshmukh, A. Ganapathiraju, R. J. Duncan, and J. Picone. Human speech recognition performance on the 1995 CSR Hub-3 corpus. In *Proc. DARPA Speech Recognition Workshop*, pages 129–134, Harriman, NY, February 1996. Morgan Kaufmann.

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

[22] W. J. Ebel and J. Picone. Human speech recognition performance on the 1994 CSR spoke 10 corpus. In *Proc. ARPA Spoken Lang. Sys. Tech. Workshop*, pages 53–59, Austin, TX, January 1995. Morgan Kaufmann.

[23] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 346–349, Atlanta, GA, May 1996.

[24] G. Fairbanks. Test of phonemic differentiation: The rhyme test. *J. Acoust. Soc. Amer.*, 30(7):596–600, 1958.

[25] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding 1997*, 1997. http://www.nist.gov/speech/papers.htm.

[26] L. Gillick and S. J. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 532–535, Glasgow, Scotland, May 1989.

[27] J. Glass, J. Chang, and M. McCandless. A probabilistic framework for feature-based speech recognition. In *Proc. Int. Conf. Spoken Language Processing*, pages 2277–2280, Philadelphia, October 1996.

[28] J. R. Glass. *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition.* Ph. D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, 1988.

[29] J. R. Glass and T. J. Hazen. Telephone-based conversational speech recognition in the jupiter domain. In *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998. To appear in 1998.

[30] W. D. Goldenthal. *Statistical Trajectory models for Phonetic Recognition.* Ph.D. thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, September 1994.

[31] A. K. Halberstadt and J. R. Glass. Heterogeneous measurements for phonetic classification. In *Proc. European Conf. Speech Communication and Technology*, pages 401–404, Rhodes, Greece, September 1997.

[32] M. A. Hasegawa-Johnson. *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification.* Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, September 1996.

[33] T. J. Hazen and A. K. Halberstadt. Using aggregation to improve the performance of mixture Gaussian acoustic models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages II–653–656, Seattle, WA, May 1998.

[34] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Amer.*, 87(4):1738–1752, April 1990.

[35] J. N. Holmes, W. J. Holmes, and P. N. Garner. Using formant frequencies in speech recognition. In *Proc. European Conf. Speech Communication and Technology*, pages 2083–2086, September 1997.

[36] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter. Articulation testing mthods: Consonantal differentiation with a closed-response set. *J. Acoust. Soc. Amer.*, 37(1):158–166, 1965.

[37] X. D. Huang, K.F. Lee, H.W. Hon, and M. Y. Hwang. Improved acoustic modeling with the SPHINX speech recognition system. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 345–348, Toronto, Canada, May 1991.

[38] F. Kubala. Design of the 1994 CSR benchmark tests. In *Proc. ARPA Spoken Lang. Sys. Tech. Workshop*, pages 41–46, Austin, Texas, January 1995. Morgan Kaufman.

[39] L. Lamel and J. Gauvian. High performance speaker-independent phone recognition using CDHMM. In *Proc. European Conf. Speech Communication and Technology*, pages 121–124, 1993.

[40] L. Lamel, R. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proc. of the DARPA Speech Recognition Workshop*, Palo Alto, February 1986. Report No. SAIC-86/1546.

[41] L. F. Lamel. *Formalizing Knowledge used in Spectrogram Reading: Acoustic and perceptual evidence from stops.* Ph.D. thesis, Department of Electrical and Computer Engineering, Massachusetts Institute of Technology, Cambridge, May 1988.

[42] L. F. Lamel and V. W. Zue. Performance improvement in a dynamic-programming-based isolated word recognition system for the alpha-digit task. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 558–561, 1982.

[43] K. F. Lee and H. W. Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(11):1641–1648, November 1989.

[44] L. Lee and R. Rose. Speaker normalization using frequency warping procedures. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 353–356, 1996.

[45] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Trans. Speech and Audio Processing*, 6(1), January 1998.

[46] S. C. Lee. *Probabilistic Segmentation for Segment-Based Speech Recognition.* M. S. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, May 1998.

[47] S. C. Lee and J. R. Glass. Probabilistic segmentation for segment-based speech recognition. In *Proc. Int. Conf. Spoken Language Processing*, Sydney, Australia, 1998. To appear in 1998.

[48] R. G. Leonard. A database for speaker-independent digit recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 42.11.1–41.11.4, San Diego, CA, March 1984.

[49] H. Leung. *The Use of Artificial Neural Networks for Phonetic Recognition.* Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, May 1989.

[50] H. Leung, B. Chigier, and J. Glass. A comparative study of signal representations and classification techniques for speech recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 680–683, Minneapolis, April 1993.

[51] H. Leung, J. Glass, M. Phillips, and V. Zue. Detection and classification of phonemes using context-independent error back-propagation. In *Proc. Int. Conf. Spoken Language Processing*, pages 1061–1064, Kobe, Japan, November 1990.

[52] Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA. *SWITCHBOARD: A User's Manual*, 1995. Catalog Number LDC94S7.

[53] R. Lippmann. Recognition by humans and machines: miles to go before we sleep. *Speech Communication*, 18:247–248, 1996.

[54] R. Lippmann. Speech perception by humans and machines. In *Proc. of the ESCA Workshop on the "Auditory Basis of Speech Perception"*, pages 309–316, Keele University, United Kingdom, July 1996.

[55] R. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, July 1997.

[56] R. Lippmann, L. Braida, and N. Durlach. A study of multichannel amplitude compression and linear amplification for persons with sensorineural hearing loss. *J. Acoust. Soc. Amer.*, 69:524–534, 1981.

[57] F. Liu, M. Picheny, P. Srinivasa, M. Monkowski, and J. Chen. Speech recognition on Mandarin call home: A large-vocabulary, conversational, and telephone speech corpus. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 157–160, Atlanta, May 1996.

[58] A. Martin. Personal communication, 1996. Cited by Lippmann in July 1996 in *Proc. of the ESCA Workshop on the "Auditory Basis of Speech Perception"*, Keele University, United Kingdom.

[59] H. Meng. *The Use of Distinctive Features for Automatic Speech Recognition.* M.S. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, September 1991.

[60] H. Meng and V. Zue. Signal representation comparison for phonetic classification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 285–288, Toronto, Canada, May 1991.

[61] P. Mermelstein and S. Davis. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(4):357–366, August 1980.

[62] G. A. Miller. Decision units in the perception of speech. *Institute of Radio Engineers Transactions on Information Theory*, 8:81–83, 1962.

[63] G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some english consonants. *J. Acoust. Soc. Amer.*, 27(2):338–352, 1955.

[64] J. Ming and F. J. Smith. Improved phone recognition using bayesian triphone models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages I–409–412, Seattle, WA, May 1998.

[65] Y. Muthusamy, R. Cole, and M. Slaney. Speaker independent vowel recognition: Spectrograms versus cochleagrams. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 533–536, Albuquerque, New Mexico, April 1990.

[66] M. Muzumdar. *Automatic Acoustic Measurement Optimization for Segmental Speech Recognition.* M.Eng. thesis, Department of Electrical and Computer Engineering, Massachusetts Institute of Technology, Cambridge, May 1996.

[67] NIST. NIST September 1998 Hub5-E Benchmark test results for recognition of conversational speech over the telephone, in English. Available via ftp from jaguar.ncsl.nist.gov, in /evaluations/hub5/sept98.

[68] Z. B. Nossair, P. L. Silsbee, and S. A. Zahorian. Signal modeling enhancements for automatic speech recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 824–827, Detroit, MI, May 1995.

[69] Z. B. Nossair and S. A. Zahorian. Dynamic spectral shape features as acoustic correlates for initial stop consonants. *J. Acoust. Soc. Amer.*, 89(6):2978–2991, 1991.

[70] H. C. Nusbaum, M. J. Dedina, and D. B. Pisoni. Perceptual confusions of consonants in natural and synthetic cv syllables. Research on speech perception progress report no. 10, Indiana University, 1984.

[71] S. Okawa, E. L. Bocchieri, and A. Potamianos. Multi-band speech recognition in noisy environments. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 641–644, 1998.

[72] D. S. Pallett. DARPA resource management and ATIS benchmark test poster session. In *Proc. DARPA Speech and Nat. Lang. Workshop*, pages 49–58, Pacific Grove, CA, February 1991. Morgan Kaufmann.

[73] D. S. Pallett et al. 1994 benchmark tests for the ARPA spoken language program. In *Proc. ARPA Spoken Lang. Sys. Tech. Workshop*, pages 5–36, Austin, TX, January 1995. Morgan Kaufmann.

[74] B. Peskin, S. Connolly, L. Gillick, S. Lowe, D. McAllaster, V. Nagesha, P. van-Mulbregt, and S. Wegmann. Improvements in switchboard recognition and topic identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 867–870, Atlanta, May 1996.

[75] M. Phillips. Speaker-independent classification of vowels and dipthongs in continuous speech. In *Proc. of the 11th Int. Congress of Phonetic Sciences*, Tallinn, Estonia, U.S.S.R., August 1987.

[76] M. Phillips and V. Zue. Automatic discovery of acoustic measurements for phonetic classification. In *Proc. Int. Conf. Spoken Language Processing*, pages 795–798, Banff, Canada, October 1992.

[77] J. W. Picone. Signal modeling techniques in speech recognition. *Proc. of the IEEE*, 81(9):1215–1247, September 1993.

[78] L. Rabiner, J. Wilpon, and F. Soong. High performance connected digit recognition using hidden Markov models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1988.

[79] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon. Speaker-independent recogniiton of isolated words using clustering techniques. *IEEE Trans. Acoust., Speech, Signal Processing*, 27:336–349, 1979.

[80] L. R. Rabiner and J. G. Wilpon. Isolated word recognition using a two-pass pattern recognition approach. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 724–727, 1981.

[81] A. Robinson. An application of recurrent neural nets to phone probability estimation. *Transactions on Neural Networks*, 5:298–305, March 1994.

[82] P. Schmid. *Explicit N-best Formant Features for Segment-Based Speech Recognition.* Ph.D. thesis, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Portland, October 1996.

[83] M. R. Schroeder. *Recognition of Complex Acoustic Signals, Life Sciences Research Report 5.* Abakon Verlag, Berlin, 1977.

[84] C. L. Searle, J. Z. Jacobson, and S. G. Rayment. Stop consonant discrimination based on human audition. *J. Acoust. Soc. Amer.*, 65(3):799–809, 1979.

[85] K. N. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17:3–45, 1989.

[86] K. N. Stevens, A. M. Liberman, M. Studdert-Kennedy, and S. E. G. Öhman. Cross-language study of vowel perception. *Language and Speech*, 12:1–23, 1969.

[87] S. Taibrewala and H. Hermansky. Sub-band based recognition of noisy speech. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1255–1258, 1997.

[88] G. N. Tajchman and M. A. Bush. Contextual effects in the perception of naturally produced vowels. *J. Acoust. Soc. Amer.*, 88(Supplement 1):S55, Fall 1990. Program of the 120th meeting of the Acoust. Soc. Amer., San Diego, November 1990.

[89] G. N. Tajchman and M. A. Bush. Effects of context and redundancy in the perception of naturally produced english vowels. In *Proc. Int. Conf. Spoken Language Processing*, pages 839–842, Banff, Canada, October 1992.

[90] X. Wang, S. A. Zahorian, and S. Auberg. Analysis of speech segments using variable spectral/temporal resolution. In *Proc. Int. Conf. Spoken Language Processing*, pages 1221–1224, Philadelphia, October 1996.

[91] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 339–342, Atlanta, GA, May 1996.

[92] H. Winitz, M. E. Scheib, and J. A. Reeds. Identification of stops and vowels for the burst portion of /p, t, k/ isolated from conversational speech. *J. Acoust. Soc. Amer.*, 51:1309–1317, 1972.

[93] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[94] P. C. Woodland, M. J. F. Gales, D. Pye, and V. Valtchev. The HTK large vocabulary recognition system for the 1995 ARPA H3 task. In *Proc. DARPA Speech Recognition Workshop*, pages 99–104, Harriman, NY, February 1996. Morgan Kaufmann.

[95] S. Wu, B. E. D. Kingsbury, N. Morgan, and S. Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 721–724, 1998.

[96] S. A. Zahorian, P. Silsbee, and X. Wang. Phone classification with segmental features and a binary-pair partitioned neural network classifier. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1011–1014, Munich, Germany, April 1997.

[97] S. A. Zahorian, P. L. Silsbee, and X. Wang. Phone classification with segmental features and a binary-pair partitioned neural network classifier. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1011–1014, 1997.

[98] P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1039–1042, Munich, Germany, April 1997.

[99] P. Zhan, M. Westphal, M. Finke, and A. Waibel. Speaker normalization and speaker adaptation – a combination for conversational speech recognition. In *Proc. European Conf. Speech Communication and Technology*, pages 2087–2090, Rhodes, Greece, September 1997.

[100] V. Zue, J. Glass, M. Phillips, and S. Seneff. The MIT SUMMIT speech recognition system: A progress report. In *Proc. DARPA Speech and Nat. Lang. Workshop*, pages 179–189, Philadelphia, February 1989. Morgan Kaufmann.

[101] V. Zue, S. Seneff, L. Hetherington J. Glass, E. Hurley, H. Meng, C. Pao, J. Polifroni, R. Schloming, and P. Schmid. From interface to content: Translingual access and delivery of on-line information. In *Proc. European Conf. Speech Communication and Technology*, pages 2227–2230, Rhodes, Greece, September 1997.