ABSTRACT

Title of dissertation: SPEECH RECOGNITION BASED ON PHONETIC FEATURES AND ACOUSTIC LANDMARKS

Amit Juneja, Doctor of Philosophy, 2004

Dissertation directed by: Carol Espy-Wilson
Department of Electrical and Computer Engineering

A probabilistic and statistical framework is presented for automatic speech recognition based on a phonetic feature representation of speech sounds. In this acoustic-phonetic approach, the speech recognition problem is hypothesized as a maximization of the joint posterior probability of a set of phonetic features and the corresponding acoustic landmarks. Binary classifiers of the manner phonetic features - syllabic, sonorant and continuant - are applied for the probabilistic detection of speech landmarks. The landmarks include stop bursts, vowel onsets, syllabic peaks, syllabic dips, fricative onsets and offsets, and sonorant consonant onsets and offsets. The classifiers use automatically extracted knowledge based acoustic parameters (APs) that are acoustic correlates of those phonetic features. For isolated word recognition with known and limited vocabulary, the landmark sequences are constrained using a manner class pronunciation graph. Probabilistic decisions on place and voicing phonetic features are then made using a separate set of APs extracted using the landmarks.

The framework exploits two properties of the knowledge-based acoustic cues

of phonetic features: (1) sufficiency of the acoustic cues of a phonetic feature for a decision on that feature and (2) invariance of the acoustic cues with respect to context. The probabilistic framework makes the acoustic-phonetic approach to speech recognition suitable for practical recognition tasks as well as compatible with probabilistic pronunciation and language models. Support vector machines (SVMs) are applied for the binary classification tasks because of their two favorable properties - good generalization and the ability to learn from a relatively small amount of high dimensional data. Performance comparable to Hidden Markov Model (HMM) based systems is obtained on landmark detection as well as isolated word recognition. Applications to rescoring of lattices from a large vocabulary continuous speech recognizer are also presented.

SPEECH RECOGNITION BASED ON PHONETIC FEATURES AND ACOUSTIC LANDMARKS

by

Amit Juneja

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Commmittee:

Carol Espy-Wilson, Chair/Advisor
Min Wu
Shihab Shamma
Amy Weinberg
Lise Getoor

To the proponents of the method of science

ACKNOWLEDGMENTS

I would like to thank a number of people without whose contribution this work would have been far from feasible.

Many thanks to Carol Espy-Wilson for the support she gave me through the project with resources, knowledge and advice. I really appreciate the independence she gave me in designing the theoretical framework developed in this work. I would also like to thank her for being very flexible and understanding during some tough personal situations that I faced.

Thanks to Om Deshmukh and Tarun Pruthi for help with numerous technical things that helped in speedy completion of this work and for being excellent roommates. Thanks to Tarun for making me watch a lot of movies!

Thanks to Mark Hasegawa-Johnson for involving me at the JHU summer workshop. Thanks to Mark Hasegawa-Johnson, Karen Livescu, Katrin Kirchoff, Steven Greenberg, James Baker, Kemal Sonmez and Ken Chen for excellent discussions we had during the course of the workshop.

Thanks to the members of my PhD proposal committee - Rama Chellappa and Min Wu - for their valuable comments and suggestions. Thanks to the system administrators, especially Shantanu Ray and Peggy Jayant, for their relieable and timely support.

My teachers during the middle school, especially Mr. V. K. Saini, deserve a

special thanks for raising my interest in science and research.

Thanks to my wife for being very cooperative during the last two years of the doctoral work when we had to live apart. Thanks to my parents and my brother and his family for constant support during the course of my PhD.

Thanks to National Science Foundation and Honda Initiation Grant for the funding of the project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, motivation is built up for the probabilistic and statistical framework of the acoustic-phonetic approach to automatic speech recognition (ASR) presented in this work. The approach, named as the event-based system (EBS), is based on the concept of representation of speech sounds by bundles of phonetic features (*Chomsky and Halle*, 1968) and acoustic landmarks (*Stevens*, 2002). EBS uses knowledge-based acoustic parameters (APs) that target the acoustic correlates of the binary manner features - *sonorant*, *syllabic* and *continuant* - to obtain multiple probabilistic landmark sequences for a speech signal. The landmarks are then used to extract APs for other manner features such as nasal and strident, and for place and voicing features, and the probabilities of these features are obtained using another set of binary classifiers. Posterior probabilities of words are then found by a combination of these probabilities. The most salient feature of the framework is its utilization of the context invariance property of the knowledge-based APs which is explained and mathematically formalized in Chapter 3.

Phonetic features (discussed in detailed in Section 1.1) are more fundamental units of speech than phones, phonemes or triphones that have been used conventionally in automatic speech recognition (*Rabiner and Juang*, 1993). Unlike phonemes, phonetic features have clear articulatory and acoustic correlates, and many of the the acoustic correlates can be automatically extracted. Also, phonetic features can describe all languages in the world while phonemes differ highly from language to language. There is evidence of the use of phonetic features in human speech perception (*Delgutte and Kiang*, 1984). There is also evidence from human perceptual studies that splitting speech recognition problem into the recognition of manner, place and voicing features can be advantageous in noisy environments (*Miller and Nicely*, 1955).

The landmark and knowledge based approach offers a number of advantages. First, by carrying out the analysis only at significant locations, the landmark based approach to speech recognition utilizes strong correlation among the speech frames. Second, analysis at different landmarks may be done with different APs that are computed at different resolutions. For example, analysis at stop bursts to determine the place of articulation requires a higher resolution than that required at syllabic peaks to determine the tongue tip and blade features. Third, the approach provides very straightforward analysis of errors. Given the physical significance of the APs and a recognition framework that uses only the relevant APs, error analysis can determine whether the APs need to be refined or the decision process didn't take into account a certain type of variability that occurs in the speech signal. In fact, this landmark and knowledge-based approach to recognition is a tool itself for

understanding speech variability. There is evidence from studies of human speech perception that analysis of speech is carried out at certain events like stop closures, stop releases and vowel onsets (*Ohde and Stevens*, 1983; *Tartter et al.*, 1983).

A good amount of work has gone into automatic extraction of knowledge based acoustic parameters (*Espy-Wilson*, 1987; *Bitar*, 1997; *Ali*, 1999; *Carbonell et al.*, 1987; *Glass*, 1984; *Chen*, 2000; *Hasegawa-Johnson*, 1996) as well as detection of acoustic landmarks (*Espy-Wilson*, 1987; *Liu*, 1996; *Bitar*, 1997; *Salomon et al.*, 2004; *Ali*, 1999; *Mermelstein*, 1975; *Niyogi*, 1998). However, the use of these ideas in practical automatic speech recognition (ASR) systems is far from realized. An attempt is made in this work to build a recognition system that explicitly uses knowledge based APs as well as carries out word level recognition. The framework for EBS has been designed to allow the use of prior language and pronunciation models with a knowledge based approach and scalability to large vocabulary recognition.

The production of speech by the human vocal tract and the concept of phonetic features are introduced in Section 1.1, and the concepts of acoustic landmarks and the acoustic correlates of phonetic features are discussed in Section 1.2. In Section 1.3 the basic ideas of acoustic phonetic knowledge based ASR are presented. The various drawbacks of the acoustic phonetic approach that have led the ASR community to abandon the approach and some ideas of solving those problems are briefly discussed in Section 1.4. The basics and the terminology of the state-of-the-art ASR, that is based largely on Hidden Markov Models (HMMs) are presented in Section 1.5 and the performance of the state-of-the-art systems is compared with human speech recognition in Section 1.6. An introduction to support vector machines (SVMs) is

presented in Section 1.7. A literature survey of the previous ASR systems that utilize acoustic phonetic knowledge is presented in Chapter 2. Chapter 3 presents the probabilistic acoustic-phonetic knowledge-based framework for speech recognition. Chapter 4 discusses the implementation and experiments for the landmark-detection system. Classification of place and voicing phonetic features is discussed in Chapter 5. Finally, word recognition results are presented in Chapter 6, and the conclusions and suggestions for future work appear in Chapter 7.

1.1 Speech Production and Phonetic Features

Speech is produced when air from the lungs is modulated by the larynx and the supra-laryngeal structures. Figure 1.1 shows the various articulators of the vocal tract that act as modulators for the production of speech. The characteristics of the excitation signal and the shape of the vocal tract filter determine the quality of the speech pattern one hears. In the analysis of a sound segment, there are three general descriptors that are used - source characteristics, manner of articulation and place of articulation. Corresponding to the three types of descriptors, three types of articulatory phonetic features can be defined - manner of articulation phonetic features, source features, and place of articulation features. The phonetic features, as defined by *Chomsky and Halle* (1968) are minimal binary valued units that are sufficient to describe all the speech sounds in any language. In the description of phonetic features, examples are given using American English phonemes. A list of American English phonemes appears in Appendix A with examples of words where

the phonemes occur.

1. **Source**

   The source or excitation of speech can be periodic when air is pushed from the lungs at a high pressure that causes the vocal folds to vibrate, or aperiodic when either the vocal folds are spread apart or the source is produced at a constriction in the vocal tract. The sounds that have the periodic source or vocal fold vibration present are said to possess the value '+' for the *voiced* feature and the sounds with no periodic excitation have the value '-' for the feature *voiced*. Both periodic and aperiodic sources may be present in a particular speech sound, for example, the sounds /v/ and /z/ are produced with vocal fold vibration but a constriction in the vocal tract adds an aperiodic turbulent noise source. The main (dominant) excitation is usually the turbulent noise source generated at the constriction. The sounds with both the sources are still *+voiced* by definition because of the presence of the periodic source.

2. **Manner of articulation**

   Manner of articulation refers to how open or close is the vocal tract, how strong or weak is the constriction and whether the air flow is through the mouth or the nasal cavity. Manner phonetic features are also called articulator-free features (*Stevens*, 2002) which means that these features are independent of the main articulator and are related to the manner in which the articulators are used. The sounds in which there is no sufficiently strong constriction so as to produce turbulent noise or stoppage of air flow are called sonorants which include

Phonetic feature	Articulatory correlate	Vowels	Sonorant consonants (nasals and semi-vowels)	Fricatives	Stops
sonorant	No constriction or constriction not narrow enough to produce turbulent noise	+	+	-	-
syllabic	Open vocal tract	+	-		
continuant	Incomplete constriction			+	-

Table 1.1: Broad manner of articulation classes and the manner phonetic features

vowels and the sonorant consonants (nasals and semi-vowels). Sonorants are characterized by the phonetic feature $+$*sonorant* and the non-sonorant sounds (stop consonants and fricatives) are characterized by the feature $-$*sonorant*. Sonorants and non-sonorants can be further classified as shown in Table 1.1 that summarizes the broad manner classes (vowels, sonorant consonants, stops and fricatives), the broad manner phonetic features - *sonorant*, *syllabic* and *continuant* and the articulatory correlates of the broad manner phonetic features.

Table 1.2 shows finer classification of phonemes on the basis of the manner

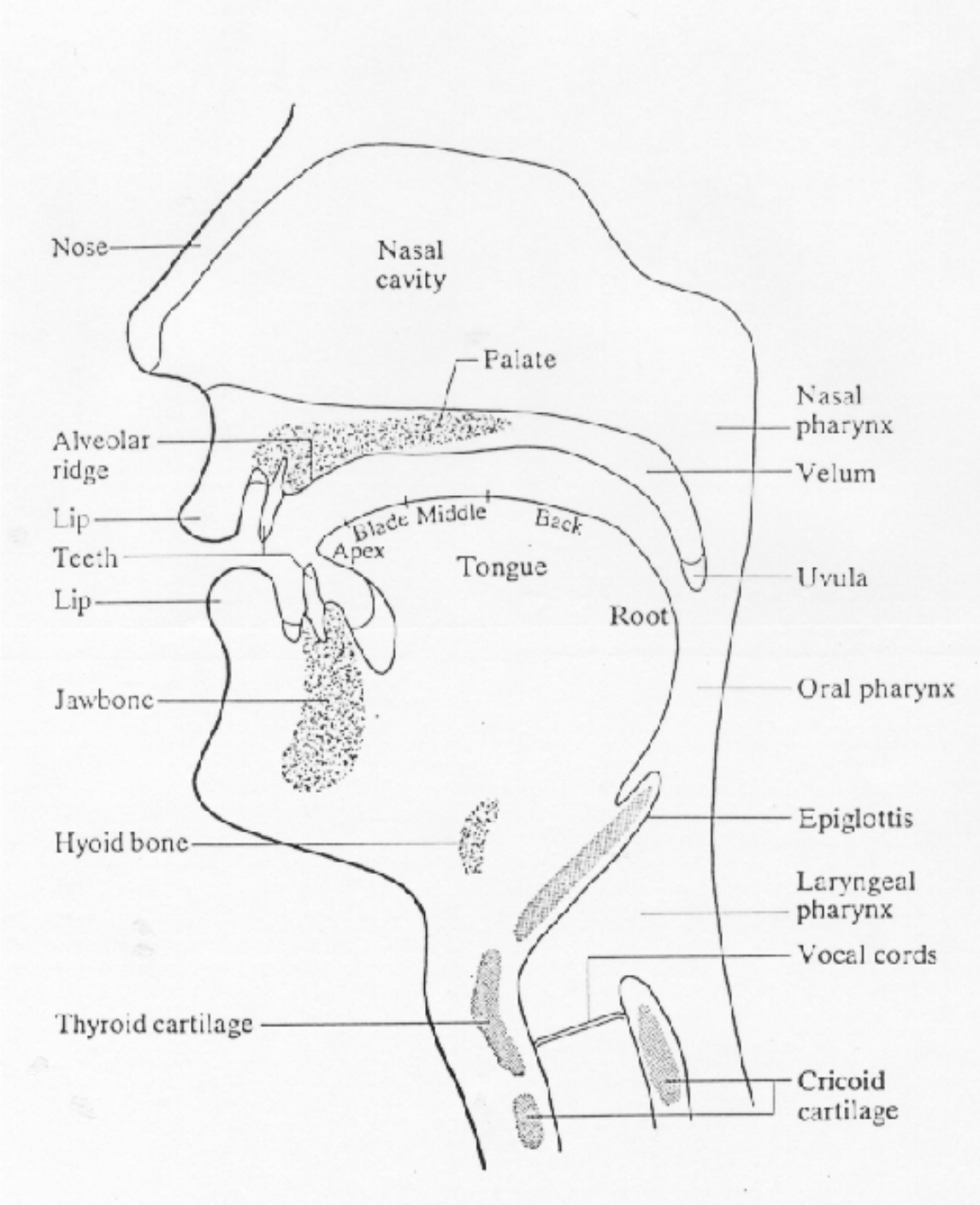


Figure 1.1: The vocal tract

Phonetic feature	s, sh	z, zh	v, dh	th, f	p, t, k	b, d, g	vowels	w r l y	n ng m
voiced	-	+	+	-	-	+	+	+	+
sonorant	-	-	-	-	-	-	+	+	+
syllabic							+	-	-
continuant	+	+	+	+	-	-			
strident	+	+	-	-	-	-			
nasal								-	+

Table 1.2: Classification of phonemes on the basis on manner and voicing phonetic features

phonetic features and the voicing feature. As shown in Table 1.2, fricatives can further be classified by the manner feature *strident*. The $+$*strident* feature signifies greater degree of frication or greater turbulent noise, that occurs in the sounds /s/, /sh/, /z/, /zh/. The other fricatives /v/, /f/, /th/ and /dh/ are $-$*strident*. Sonorant consonants can be further classified by using the phonetic feature $+$*nasal* or $-$*nasal*. Nasals, with $+$*nasal* feature - /m/, /n/, and /ng/ - are produced with a complete stop of air flow through the mouth. Instead the air flows out through the nasal cavities.

3. **Place of articulation**

   The third classification required to produce or characterize a speech sound is the place of articulation, that refers to the location of the most significant constriction (for stops, fricatives and sonorant consonants) or the shape and position of the tongue (for vowels). For example, using place phonetic features

, stop consonants may be classified (see Table 1.3) as (1) alveolar (/d/ and /t/) when the constriction is formed by the tongue tip and the alveolar ridge (2) labial (/b/ and /p/) when the constriction is formed by the lips, and (3) velar (/k/ and /g/) when the constriction is formed by the tongue dorsum and the palate. The stops with identical place, for example the alveolars /d/ and /t/ are distinguished by the voicing feature, that is, /d/ is $+voiced$ and /t/is $-voiced$. The place features for other classes of sounds - vowels, sonorants consonants and fricatives - are tabulated in Appendix B.

All the sounds can, therefore, be represented by a collection or bundle of phonetic features. For example, the phoneme /z/ can be represented as a collection of the features

$$\{-sonorant, +continuant, +voiced, +strident, +anterior\}.$$

Moreover, words may be represented by a sequence of bundles of phonetic features. Table 1.4 shows the representation of the digit 'zero', pronounced as /z I r ow/, in terms of the phonetic features. Phonetic features may be arranged in a hierarchy such as the one shown in Figure 1.2. The hierarchy enables us to describe the phonemes with a minimal set of phonetic features, for example, the feature *strident* is not relevant for sonorant sounds.

Phonetic feature	Articulatory correlate	b p	d t	g k
velar	Constriction between tongue body and soft palate	-	-	+
alveolar	Constriction between tongue tip and alveolar ridge	-	+	-
labial	Constriction between the lips	+	-	-

Table 1.3: Classification of stop consonants on the basis of place phonetic features

/z/	/I/	/r/	/o/	/w/
$-sonorant$	$+sonorant$	$+sonorant$	$+sonorant$	$+sonorant$
$+continuant$	$+syllabic$	$-syllabic$	$+syllabic$	$-syllabic$
$+voiced$	$-back$	$-nasal$	$+back$	$-nasal$
$+strident$	$+high$	$+rhotic$	$-high$	$+labial$
$+anterior$	$+lax$		$+low$	

Table 1.4: Phonetic feature representation of phonemes and words. The word 'zero' may be represented as the sequence of phones /z I r ow/ as shown in the top row or the sequence of corresponding phonetic feature bundles as shown in the bottom row.

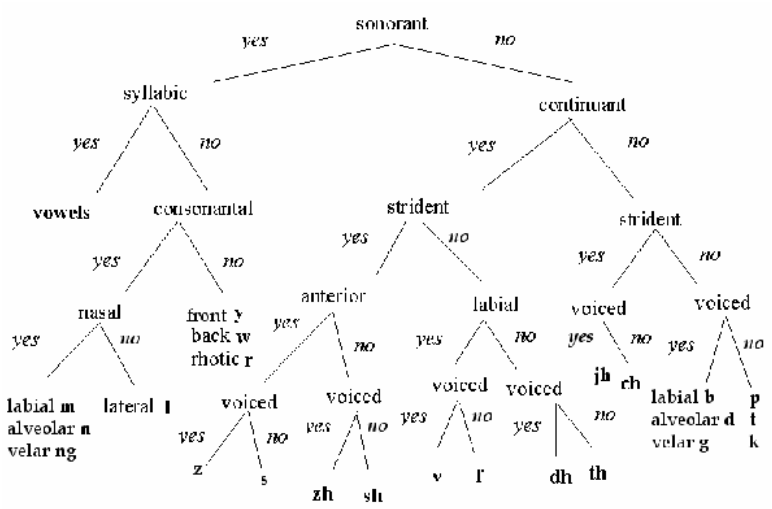


Figure 1.2: Phonetic feature hierarchy

1.2 Acoustic correlates of phonetic features

The binary phonetic features manifest in the acoustic signal in varying degrees of strength. There has been considerable research in the understanding of the acoustic correlates of phonetic features, for example, Stevens (*Stevens et al.*, 1999; *Stevens*, 1980; *Espy-Wilson*, 1987; *Glass*, 1984). In this work, the term Acoustic Parameters or APs is used for the acoustic correlates that can be extracted automatically from the speech signal and there has been some success in finding these automatically extracted acoustic correlates, for example, (*Ali*, 1999; *Bitar*, 1997; *Hasegawa-Johnson*, 1996; *Liu*, 1996; *Deshmukh et al.*, to appear). In EBS, the APs related to the broad manner phonetic features - *sonorant*, *syllabic* and *continuant* - are extracted from every frame of speech. Table 4.1 provides examples of APs for manner phonetics features (*Bitar*, 1997; *Deshmukh et al.*, to appear), and later used in Support Vector Machine (SVM) based segmentation of speech (*Juneja and Espy-Wilson*, 2003,

2004).

The APs for broad manner features and the decision for the positive or negative value for each feature is used to find a set of landmarks in the speech signal. Figure 1.3 illustrates the landmarks obtained from the acoustic correlates of the manner phonetic features. There are two kinds of manner landmarks (1) landmarks defined by an abrupt change, for example, burst landmark for stop consonants (shown by ellipse 1 in the figure), and vowel onset point (VOP) for vowels, and (2) landmarks defined by the most prominent manifestation of a manner phonetic feature, for example, a point of maximum low frequency energy in a vowel (shown by ellipse 3) and a point of lowest energy in in a certain frequency band (*Bitar*, 1997) for an intervocalic sonorant consonant (a sonorant consonant that lies between two vowels).

The acoustic correlates of place and voicing phonetic features are extracted using the locations provided by the manner landmarks. For example, the stop consonants /p/, /t/ and /k/ are all unvoiced stop consonants and they differ in their place phonetic features. /p/ is $+labial$, /t/ is $+alveolar$ and /k/ is $+velar$. The acoustic correlates of these three kinds of place phonetic features can be extracted using the burst landmark (*Stevens et al.*, 1999) and the VOP. The acoustic cues for place and voicing phonetic features are most prominent at the locations provided by the manner landmarks, and they are least affected by contextual or coarticulatory effects at these locations. For example, the formant structure typical to a vowel is expected to be most prominent at the location in time where the vowel is being spoken with the maximum loudness.

In a broad sense, the landmark based recognition procedure involves three

steps (1) location of manner landmarks, (2) analysis of the landmarks for place and voicing phonetic features and (3) matching the phonetic features obtained by this procedure to phonetic feature based representation of words or sentences. This is the approach to speech recognition that is followed in this work. The landmark based approach is similar to human spectrogram reading (*Zue and Cole*, 1995) where an expert locates certain events in the speech spectrogram, and analyze those events for significant cues required for phonetic distinction. By carrying out the analysis only at significant locations, the approach utilizes strong correlation among the speech frames. The approach has been advocated by Stevens (*Stevens et al.*, 1992; *Stevens*, 2002) and further pursued by Liu (*Liu*, 1996) and Bitar and Espy-Wilson (*Bitar*, 1997; *Espy-Wilson*, 1994).

1.3 Definition of acoustic-phonetic knowledge based ASR

All the approaches to ASR can be classified as either 'static' or 'dynamic'. In the static approach, explicit events are located in the speech signal and the recognition of units - phonemes or phonetic features - is carried out using a fixed number of acoustic measurements extracted using those events. In the static method, no dynamic models like HMMs are used to model the time varying characteristics of speech. In this thesis, the acoustic phonetic approach to ASR is defined as a static approach where analysis is carried out at explicit locations in the speech signal and EBS

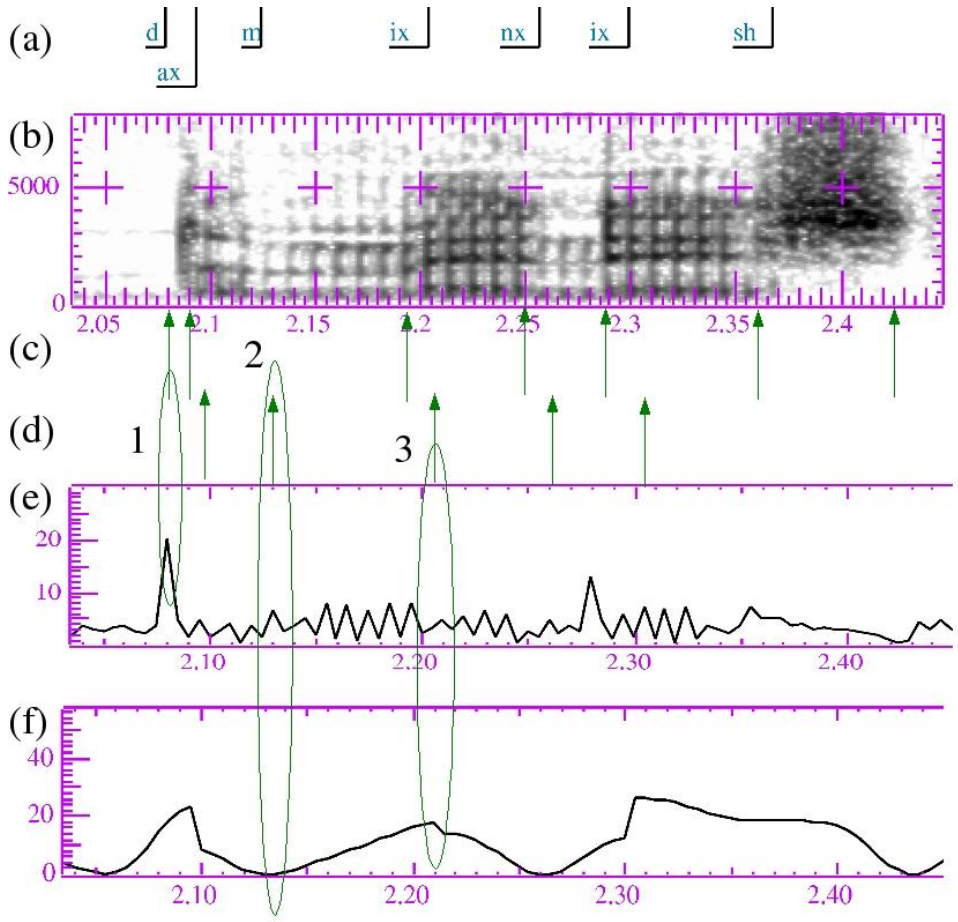


Figure 1.3: Illustration of manner landmarks for the utterance "diminish" from the TIMIT database (*NIST*, 1990). (a) Phoneme Labels, (b) Spectrogram, (c) Landmarks characterized by sudden change, (d) Landmarks characterized by maxima or minima of a correlate of a manner phonetic feature, (e) Onset waveform (an acoustic correlate of phonetic feature $-$*continuant*), (f) E[640,2800] (an acoustic correlate of *syllabic* feature). Ellipse 1 shows the location of stop burst landmark for the consonant /d/ using the maximum value of the onset energy signifying a sudden change. Ellipse 2 shows how minimum of E[640,2800] is used to locate the syllabic dip for the nasal /m/. Similarly, ellipse 3 shows that the maximum of the E[640,2800] is used to locate a syllabic peak landmark of the vowel /ix/.

belongs to this category. In the dynamic approach, speech is modeled by statistical dynamic models like HMMs and this approach is discussed further in Section 1.5.

A detailed discussion of the past acoustic phonetic ASR methods and other methods that utilize acoustic phonetic knowledge (for example, HMM systems that use acoustic phonetic knowledge) is presented in Section 2. A typical acoustic-phonetic approach to ASR has the following steps (this is similar to the overview of the acoustic-phonetic approach presented by Rabiner (*Rabiner and Juang*, 1993) but it is defined here more broadly):

1. Speech is analyzed using any of the spectral analysis methods - Short Time Fourier Transform (STFT), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), etc. - using overlapping frames with a typical size of 10-25ms and typical overlap of 5ms.

2. Acoustic correlates of phonetic features are extracted from the spectral representation. For example, low frequency energy may be calculated as an acoustic correlate of sonorancy, zero crossing rate may be calculated as a correlate of frication, and so on.

3. Speech is segmented by either finding transient locations using the spectral change across two consecutive frames, or using the acoustic correlates of source or manner classes to find the segments with stable manner classes. The earlier approach , that is, finding acoustic stable regions using the locations of spectral change has been followed by Glass et al. (*Glass and Zue*, 1988). The latter method of using broad manner class scores to segment the signal has been

used by a number of researchers (*Bitar*, 1997; *Liu*, 1996; *Fohr et al.*; *Carbonell et al.*, 1987). Multiple segmentations may be generated instead of a single representation, for example, the dendograms in the speech recognition method proposed by Glass (*Glass and Zue*, 1988). (The system built by Glass et al. is included here as an acoustic phonetic system because it fits the broad definition of the acoustic-phonetic approach, but this system uses very little knowledge of acoustic phonetics.)

4. Further analysis of the individual segmentations is carried out next to either recognize each segment as a phoneme directly or find the presence or absence of individual phonetic features and using the intermediate decisions to find the phonemes. When multiple segmentations are generated instead of a single segmentation, a number of different phoneme sequences may be generated. The phoneme sequences that match the vocabulary and grammar constraints are used to decide upon the spoken utterance by combining the acoustic and language scores.

1.4 Hurdles in the acoustic-phonetic approach

A number of problems have been associated with the acoustic-phonetic approach in the literature. Rabiner (*Rabiner and Juang*, 1993) lists at least five such problems or hurdles that have made the use of the approach minimal in the ASR community. The problems with the acoustic phonetic approach and some ideas for solving them provide much of the motivation for the present work. These documented problems of

the acoustic-phonetic approach are now listed and it is argued that either insufficient effort has gone into solving these problems or that the problems are not unique to the acoustic-phonetic approach.

- It has been argued that the difficulty in proper decoding of phonetic units into words and sentences grows dramatically with an increase in the rate of phoneme insertion, deletion and substitution. This argument makes the assumption that phoneme units are recognized in the first pass with no knowledge of language and vocabulary constraints. This has been true for many of the acoustic phonetic methods, but this is not necessary since vocabulary and grammar constraints may be used to constrain the speech segmentation paths (*Glass et al.*, 1996).

- Extensive knowledge of the acoustic manifestations of phonetic units is required and the lack of completeness of this knowledge has been pointed out as a drawback of the knowledge based approach. While it is true that the knowledge is incomplete, there is no reason to believe that the standard signal representations, for example, Mel-Frequency Cepstral Coefficients (MFCCs), used in the state-of-the-art ASR methods (discussion in Section 1.5) are sufficient to capture all the acoustic manifestations of the speech sounds. Although the knowledge is not complete, a number of efforts to find acoustic correlates of phonetic features have obtained excellent results. Most recently, there has been significant development in the research on the acoustic correlates of place of stop consonants and fricatives (*Stevens et al.*, 1999; *Ali*, 1999; *Bitar*, 1997),

nasal detection (*Pruthi and Espy-Wilson*, 2003), and semivowel classification (*Espy-Wilson*, 1994). The knowledge from these sources may be adequate to start building an acoustic-phonetic speech recognizer to carry out word recognition tasks, and that was the focus of this work. It should be noted that because of the physical significance of the knowledge based acoustic measurements, it is easy to pinpoint the source of recognition errors in the recognition system. Such an error analysis is close to impossible in MFCC like front-ends.

- The third argument against the acoustic-phonetic approach is that the choice of phonetic features and their acoustic correlates is not optimal. It is true that linguists may not agree with each other on the optimal set of phonetic features, but finding the best set of features is a task that can be carried out instead of turning to other ASR methods. The phonetic feature set used in this work will be based on the distinctive feature theory and it will be optimal in that sense.

- Another drawback of the acoustic-phonetic approach as pointed out in (*Rabiner and Juang*, 1993) is that the design of the sound classifiers is not optimal. This argument probably assumes that binary decision trees with hard knowledge-based thresholds are used to carry out the decisions in the acoustic-phonetic approach. Statistical pattern recognition methods that are no less optimal than the HMMs have been applied to acoustic-phonetic approaches as discussed further in Section 2. Statistical pattern recognition methods have been applied in some acoustic phonetics knowledge based methods, for exam-

ple, (*Niyogi*, 1998; *Fohr et al.*) although scalability of these methods to bigger recognition tasks has not been accomplished.

- The last shortcoming of the acoustic-phonetic approach is that no well defined automatic procedure exists for tuning the method. The acoustic-phonetic methods can be tuned if they use standard data driven pattern recognition methods, and this can be possible in the presented approach. But the goal of this work was to design an ASR system that does not require tuning except under extreme circumstances, for example, accents that are extremely different from standard American English (assuming the original system was trained on native American speakers).

1.5 State-of-the-art ASR

ASR using the acoustic modeling by HMMs has dominated the field since the mid 1970s when very high performance on certain continuous speech recognition tasks was reported by Jelinek (*Jelinek*, 1976) and Baker (*Baker*, 1975). A very brief review of HMM based ASR, starting with how isolated word recognition is carried out using HMMs is presented here. Given a sequence of observation vectors $O = \{o_1, o_2, ..., o_T\}$, the task of the isolated word recognizer is to find from a set of words $\{w_i\}_{i=1}^{V}$, a word w_v^* such that

$$w_{v^*} = \arg \max_{w_i} P(O/w_i)P(w_i). \tag{1.1}$$

One of the ways to carry out isolated word recognition using HMMs is to build a 'word model' for each word in the set $\{w_i\}_{i=1}^{V}$. That is, an HMM model $\lambda_v = (A_v, B_v, \pi_v)$ is built for every word w_v. An HMM model λ is defined as a set of three entities (A, B, π) where $A = \{a_{ij}\}$ is the transition matrix of the HMM, $B = \{b_j(o)\}$ is the set of observation densities for each state, and $\pi = \{\pi_i\}$ is the set of initial state probabilities. Let N be the number of states in the model λ, and the state at instant t be denoted by q_t, a_{ij}, $b_j(o)$ and π_i are defined as

$$a_{ij} = P(q_{t+1} = j | q_t = i), \qquad 1 \leq i, j \leq N \tag{1.2}$$

$$b_j(o) = P(o_t = o | q_t = j) \tag{1.3}$$

$$\pi_i = P(q_1 = i), \qquad 1 \leq i \leq N \tag{1.4}$$

The problem of isolated word recognition is then to find the word w_{v^*} such that

$$v^* = \arg \max_i P(O|\lambda_i) P(w_i). \tag{1.5}$$

Given the models λ_v for each of the words in $\{w_i\}_{i=1}^{V}$, the problem of finding v^* is called the decoding problem. The Viterbi algorithm (*Viterbi*, 1967; *Forney*, 1973) is used to find the estimate of the probabilities $P(O|\lambda_i)$, and the prior probabilities $P(w_i)$ are known. The training of HMMs is defined as a task of finding the best model λ_i, given an observation sequence O or a set of observation sequences for each word w_i and it is usually carried out using the Baum-Welch algorithm (derived from Expectation Maximization algorithm). Multiple observation sequences, that is, multiple instances of the same word are used for training the models by sequentially carrying out the iterations of the Baum-Welch over each instance. Figure 1.4 shows

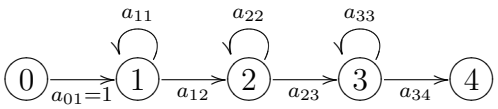


Figure 1.4: A typical topology of an HMM used in ASR with non-emitting start and end states 0 and 4

a typical topology of an HMM used in ASR. There are two non-emitting states - 0 and 4 - that are the start and the end states, respectively, and the model is left-to-right, that is, no transition is allowed from any state to a state with lower index.

For continuous or connected word speech recognition with small vocabularies, the best path through a lattice of HMMs of different words is found to get the most probable sequence of words given a sequence of acoustic observation vectors. A language or grammar model may be used to constrain the search paths through the lattice and improve recognition performance. Mathematically the problem in continuous speech recognition is to find a sequence of words $\hat{W}$ such that

$$\hat{W} = \arg\max_{W} P(O|W)P(W). \tag{1.6}$$

The probability $P(W)$ is calculated using a language model appropriate for the recognition task, and the probability $P(O|W)$ is calculated by concatenating the HMMs of the words in the sequence W and using the Viterbi algorithm for decoding. A silence or a 'short pause' model is usually inserted between the HMMs to be concatenated. Figure 1.5 illustrates the concatenation of HMMs. Language models are usually composed of bigrams, trigrams or probabilistic context free grammars (*Jurafsky and Martin*, 2000).

When the size of the vocabulary is large, for example, 100,000 or more words, it is impractical to build word models because a large amount of storage space is required for the parameters of the large number of HMMs, and a large number of instances of all the words is required for training the HMMs. But words highly differ in their frequency of occurrence in speech corpora, and the number of available training samples is usually insufficient to build acoustic models. HMMs have to be built for subword units like monophones, diphones (centers of sequences of phone pairs), triphones (phones in context of two adjoining phones) or syllables. A dictionary of pronunciations of words in terms of the subword units is constructed and the acoustic model of each word is then the concatenation of the subword units in the pronunciation of the word, as shown in Figure 1.6. Monophone models have shown little success in ASR with large vocabularies and the state-of-the-art in HMM based ASR is the use of triphone models. There are about 40 phonemes in American English. Therefore, approximately 40^3 triphone models are required.

An enormous number of modifications and improvements over the basic HMM method for ASR have been suggested in the past two decades, but these methods are not discussed here. The goal of this work is an acoustic-phonetic knowledge based system that will operate very differently from the HMM approach. It is now briefly discussed why the performance of the HMM based systems is far from that of human speech recognition (HSR), and what is the difference in the performance of ASR and HSR.

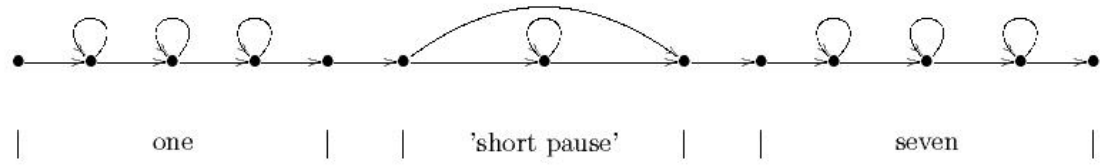


Figure 1.5: Concatenation of word level HMMs for the words - 'one' and 'seven' - through a 'short pause' model. To find the likelihood of an utterance given the sequence of these two words, the HMMs for the words are concatenated with an intermediate 'short pause' model and the best path through the state transition graph is found. Similarly the three HMMs are concatenated for the purpose of training and the Baum-Welch algorithm is run through the composite HMM

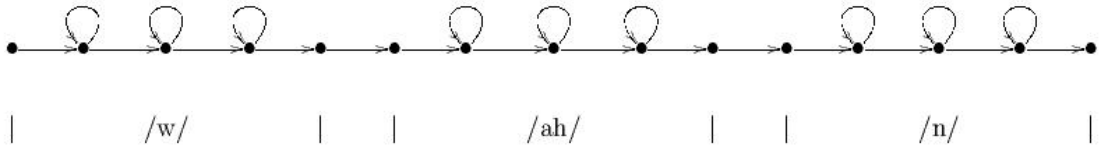


Figure 1.6: Concatenation of phone level HMMs for the phonemes - /w/, /ah/ and /n/ - to get the model of the word 'one'. To find the likelihood of an utterance given the word 'one', the HMMs for the these phonemes are concatenated and the best path through the state transition graph is found. Similarly the three HMMs are concatenated for the purpose of training and the Baum-Welch algorithm is run through the composite HMM

1.6 ASR versus HSR

ASR has been an area of research over the past 40 years. While significant advances have been made, especially since the advent of the HMM based ASR systems, the ultimate goal of performance equivalent to humans is nowhere near. In 1997, Lippmann (*Lippmann*, 1997) compared the performance of ASR with HSR. The comparison is still valid today given only incremental improvements to HMM based ASR have been made since that time. Lippmann showed that humans perform approximately 3 to 80 times better than machines using word error rate (WER) as the performance measure. The conclusion made by Lippmann that is most relevant to this work is that the gap between HSR and ASR can be reduced by improving low level acoustic-phonetic modeling. It was noted that ASR performance on a continuous speech corpus - Resource Management - drops from 3.6% WER to 17% WER when the grammar information is not used (i.e., when all the words in the corpus have equal probability). The corresponding drop in the HSR performance was from 0.1% to 2%, indicating that ASR is much more dependent on high level language information than HSR. On a connected alphabet task, the recognition performance of HSR was reported to be 1.6% WER while the best reported machine error rate on islolated letters is about 4% WER. The 1.6% error rate of HSR on connected alphabet can be considered to be an upper bound of human performance on isloated alphabet. On telephone quality speech, Ganapathiraju (*Ganapathiraju*, 2002) reported an error rate of 12.1% on connected alphabet which represents the state-of-the-art. Lippmann also points out that human spectrogram reading per-

formance is close to ASR performance although, it is not as good as HSR. This indicates that the acoustic-phonetic approach, inspired partially from spectrogram reading, is a valid option for ASR.

Further evidence that humans carry out highly accurate phoneme level recognition comes from perceptual experiments carried out by Fletcher (*Fletcher and Steinberg*, 1929). On clean speech, a recognition error of 1.5% over the phones in nonsense consonant-vowel-consonant (CVC) syllables was reported. (Machine performance on nonsense CVC syllables is not known.) Further, it was reported that the probability of correct recognition for a syllable is the product of the probability of correct recognition of the constituent phones. Allen (*Allen*, 1994, 2002) inferred from this observation in his review of Fletcher's work that individual phones must be correctly recognized for a syllable to be recognized correctly. Allen further concluded that it is unlikely that context is used in the early stages of human speech recognition and that the focus in ASR research must be on phone recognition. Fletcher's work also suggests that recognition is carried out separately in different frequency bands and the phone recognition error rate by humans is the minimum of error rate across all the frequency bands. That is, recognition of intermediate units that Allen calls phone features (not the same as phonetic features) is done across different channels and combined in such a way that the error is minimized. In HMM based systems the recognition is done using all the frequency information at the same time and in this way HMM based systems work in a very different manner from HSR. Moreover, the state-of-the-art of the technology is more concentrated on recognizing triphones because of the poor performance of HMMs at phoneme recognition.

The focus of EBS is on the recognition of phonetic features and the correct recognition of phonetic features will lead to correct recognition of phonemes. The recognition system presented in this work is not based on processing different frequency bands independently, but all the available information is not used at the same time for recognizing all the phones. That is, different information (acoustic correlates of phonetic features) is used for recognition of different features to get partial recognition results (in terms of phonetic features) and at times this information may belong to different frequency bands. The goal in building a phonetic feature and landmark based system is to capture the low level information with a satisfactory accuracy.

1.7 Support Vector Machines

SVMs are maximum margin classifiers. These have been applied in this work as binary classifiers of phonetic features for both obtaining the acoustic landmarks and detecting the place of articulation. Figure 1.7 illustrates the difference between large margin classifiers and small margin classifiers. For linearly separable data lying in space $\mathbb{R}^n$, the goal of SVM training for two class pattern recognition is to find a hyperplane defined by a weight vector $\mathbf{w}$ and a scalar b

$$\mathbf{w}.\mathbf{x} + b = 0, \;\; x \in \mathbb{R}^n \tag{1.7}$$

such that the margin $2/||\mathbf{w}||$ between the closest training samples with opposite labels is maximized. Figure 1.7 shows two types of classifiers for linearly separable data (1) a linear classifier without maximum margin and (2) a linear classifier with

maximum margin. It is easy to see in Figure 1.7 that the classifier in (b) is more robust to noise because a larger amount of noise is required to let a sample point cross the decision boundary. It has been argued (*Vapnik*, 1995) that the maximization of the margin leads to the minimization of a bound on the test error by the principle of Structural Risk Minimization (discussed in Section 1.7.1).

In general, SVMs select a set of N_{SV} support vectors $\{\mathbf{x}_i^{SV}\}_{i=i}^{N_{SV}}$ that is a subset of l vectors in the training set $\{\mathbf{x}_i\}_{i=1}^{l}$ with class labels $\{y_i\}_{i=1}^{l}$, and find an optimal separating hyperplane $f(\mathbf{x})$ (in the sense of maximization of margin) in a high dimensional space $\mathcal{H}$,

$$f(\mathbf{x}) = \sum_{i=1}^{N_{SV}} y_i \alpha_i K(\mathbf{x}_i^{SV}, \mathbf{x}) - b. \tag{1.8}$$

The space $\mathcal{H}$ is defined by a linear or non-linear kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ that satisfies the Mercer conditions (*Burges*, 1998). The weights α_i, the set of support vectors $\{\mathbf{x}_i^{SV}\}_{i=1}^{N_{SV}}$ and the bias term b are found from the training data using quadratic optimization methods.

The mapping $\Phi : \mathbb{R} \mapsto \mathcal{H}$ can be explicitly defined for certain kernels but it is usually difficult. The space $\mathcal{H}$ may be infinite dimensional but that is handled elegantly because K is a scalar, and the training is straightforward because of the linearity of the separating function $f(\mathbf{x})$ in K in Equation 1.8. Two commonly used kernels are radial basis function (RBF) kernel and linear kernel. For RBF kernel,

$$K(\mathbf{x}_i, \mathbf{x}) = exp(-\gamma |\mathbf{x}_i - \mathbf{x}|^2) \tag{1.9}$$

where the parameter γ is usually chosen empirically by cross-validation from the

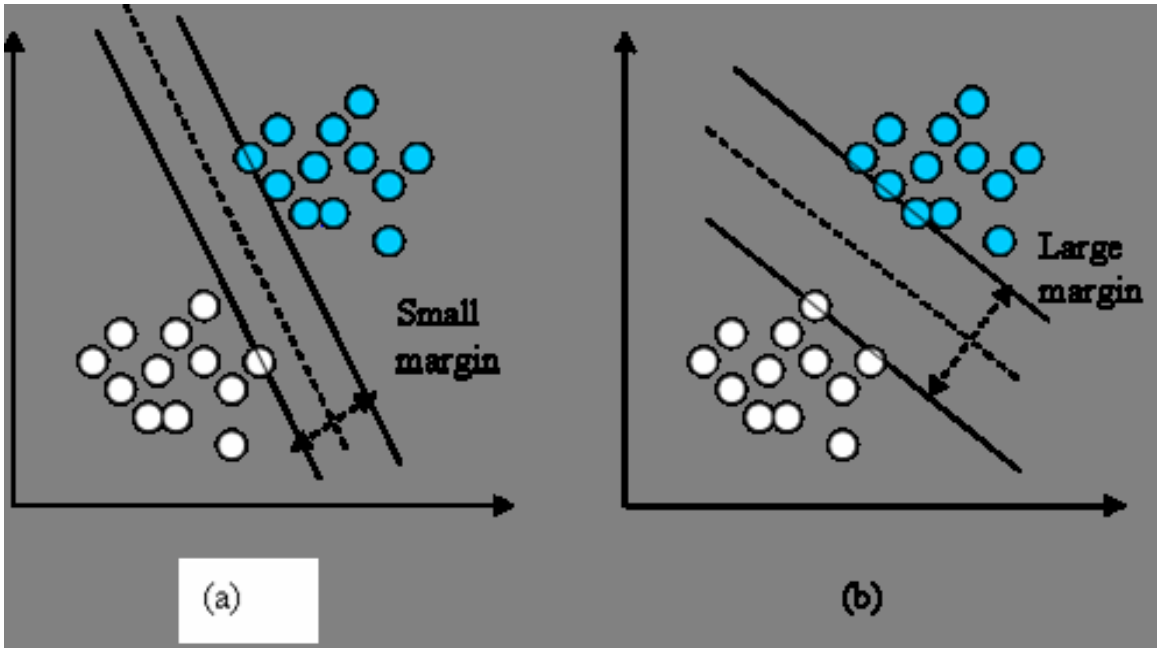


Figure 1.7: (a) small margin classifiers, (b) maximum margin classifiers

training data. For the linear kernel,

$$K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i.\mathbf{x} + 1 \tag{1.10}$$

1.7.1 Structural Risk Minimization (SRM)

Given a set of training vectors $\{\mathbf{x}_i\}_{i=1}^{l}$, and the corresponding class labels $\{y_i\}_{i=1}^{l}$ such that

$$y_i \in \{-1, +1\} \text{ and } \mathbf{x}_i \in \mathbb{R}^n,$$

assume that the samples $\{\mathbf{x}_i\}_{i=1}^{l}$ and the class labels $\{y_i\}_{i=1}^{l}$ are produced by a joint probability distribution $P(\mathbf{x}, y)$ (note that $dP(\mathbf{x}, y) = p(\mathbf{x}, y)d\mathbf{x}dy$ where $p(\mathbf{x}, y)$ is

the probability density). For a possible function $f(\mathbf{x}, \alpha)$ that attempts to find the class labels for given vector a $\mathbf{x}$, the expected risk of the function or the expected error on unseen data is defined as

$$R(\alpha) = \int \frac{1}{2}|y - f(\mathbf{x}, \alpha)| dP(\mathbf{x}, y). \tag{1.11}$$

With a probability η $(0 \leq \eta \leq 1)$, the following bound on the expected risk exists (*Vapnik*, 1995),

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(log(2l/h) + 1) - log(\eta/4)}{l}} \tag{1.12}$$

where h is called the Vapnik Chervonenkis (VC) dimension and the second term on the right side is called the VC confidence. $R_{emp}(\alpha)$ is the empirical risk

$$R_{emp}(\alpha) = \frac{1}{2l} \sum_{i=1}^{l} |y_i - f(\mathbf{x}_i, \alpha)|. \tag{1.13}$$

The VC dimension h depends on the class of functions $f(\mathbf{x}, \alpha)$ and the empirical risk is defined for a particular α under consideration. h is defined as the maximum number of samples that can be separated by a function from the class of functions $f(\mathbf{x}, \alpha)$ with any arbitrary labeling of those samples. The principle of structural risk minimization consists of finding the class of functions and a particular function belonging to that class (defined by a particular value of α), such that the sum of VC confidence and the empirical risk is minimized. SVM training finds a separating hyperplane by maximizing margin across the two classes and this process of finding a maximum margin classifier has been linked to the SRM principle. There is no concrete proof however that SVMs actually minimize the expected bound on test data error (*Burges*, 1998).

Chapter 2

Previous acoustic-phonetic methods

A number of ASR procedures have appeared in the literature that make use of acoustic phonetics knowledge. These procedures can be classified into three broad categories that will make it easy for the reader to contrast these methods with this work - (1) the acoustic phonetic approach to recognition, (2) the use of acoustic correlates of phonetic features in the front-ends of dynamic statistical ASR methods like HMMs, and (3) the use of phonetic features in place of phones as recognition units in the dynamic statistical approaches to ASR that use standard front-ends like MFCCs.

2.1 Acoustic-phonetic approach

The acoustic phonetic approach is the recognition strategy that was outlined in Section 1.3. It is characterized by the use of spectral coefficients or the knowledge based acoustic correlates of phonetic features to first carry out the segmentation of speech and then analyze the individual segments or linguistically relevant landmarks for phonemes or phonetic features. This method may or may not involve the use of statistical pattern recognition methods to carry out the recognition task. That is, these methods include pure knowledge based approaches with no statistical modeling. The acoustic phonetic approach has been followed and implemented for recognition in varying degrees of completeness or capacity of application to real world recognition problems. Figure 2.1 shows the block diagram of the acoustic phonetic approach. As shown in Table 2.1, most of the acoustic phonetic methods have been limited to the second and third modules (i.e., landmark detection and phone classification). Only the SUMMIT system (discussed below) is able to carry out recognition on continuous speech with a substantial vocabulary. But the SUMMIT system uses a traditional front end with little or no knowledge based APs. Also most systems that have used or developed knowledge based APs do not have a complete set of APs for all phonetic features.

2.1.1 Landmark detection or segmentation systems

Bitar (*Bitar*, 1997) used knowledge based acoustic parameters in a fuzzy logic framework to segment the speech signal into the broad classes - vowel, sonorant conso-

nant, fricative and stop - in addition to silence. Performance comparable to an HMM based system (using either MFCCs or APs) was obtained on the segmentation task. Bitar also optimized the APs for the discriminative capacity on the phonetic features the APs were designed to analyze. APs were also developed and optimized for the phonetic features *strident* for fricatives, and *labial* and *alveolar* for stop consonants. Many of the APs developed by *Bitar* (1997) are used in this work. However, some of them have been refined. A recognition system for word recognition was not developed in this work.

Liu (*Liu*, 1996) proposed a system for detection of landmarks in continuous speech. Three different kinds of landmarks were detected - glottal, burst and sonorant. Glottal landmarks marked the beginning and end of voiced regions in speech, the burst landmark located the stop bursts, and the sonorant landmarks located the beginning and end of sonorant consonants. The three kinds of landmarks were recognized with error rates of 5%, 14% and 57% respectively, when compared to hand-transcribed landmarks and counting insertions, deletions and substitions as errors. It is difficult to understand these results in the context of ASR since it is not clear how the errors will affect word or sentence recognition. A system using phonetic features and acoustic landmarks for lexical access was proposed by Stevens et al, (*Stevens et al.*, 1992; *Stevens*, 2002) as discussed in Section 1.2. However, a practical framework for speech recognition was not presented in either of these works.

Salomon (*Salomon*, 2000) used temporal measurements derived from the average magnitude difference function (AMDF) computed in each frequency channel

to obtain measures of periodicity, aperiodicity, energy onsets and energy offsets. This work was motivated by the perceptual studies that humans are able to detect manner and voicing events in spectrally degraded speech with considerable accuracy, indicating that humans use temporal information to extract such information. An overall detection rate of 70.8% was obtained and a detection rate of 87.1% was obtained for perceptually salient events. The temporal based processing proposed in this work, and developed further by Deshmukh et at (*Deshmukh et al.*, to appear) have been used in the proposed project.

Ali (*Ali*, 1999) carried out segmentation of continuous speech into broad classes - sonorants, stops, fricatives and silence - with an auditory-based front end. The front end was comprised of mean rate and synchrony outputs obtained using a Hair Cell Synapse model (*Seneff*, 1988). Rule based decisions with statistically determined thresholds were made for the segmentation task and an accuracy of 85% was obtained that is not directly comparable to (*Liu*, 1996) where landmarks, instead of segments are found. Using the auditory based front end, Ali further obtained very high classification accuracies on stop consonants (86%) and fricatives (90%). The sounds /f/ and /th/ were put into the same class, and so were /v/ and /dh/ for the classification of fricatives. Glottal stops were not considered in the stop classification task. One of the goals of this work was to show noise robustness of the auditory-based front end and it was successfully shown that the auditory based features perform better than the traditional ASR front ends. An acoustic phonetic speech recognizer to carry out recognition of words or sentences was not designed as a part of this work.

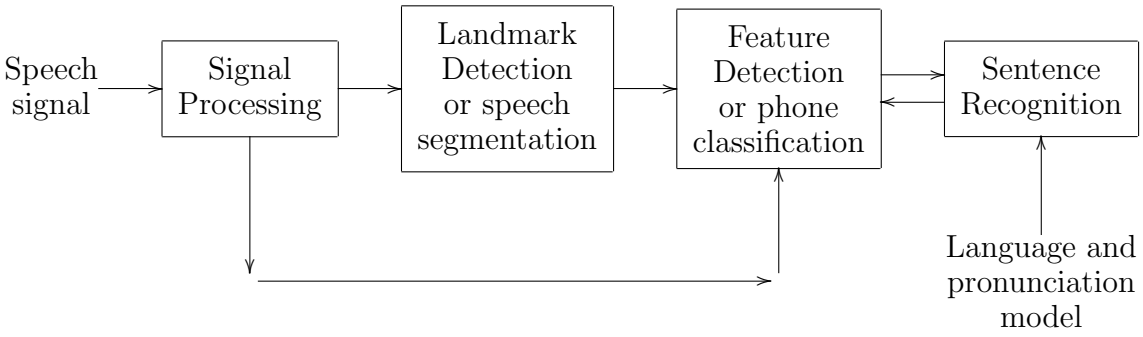


Figure 2.1: Block diagram of acoustic phonetic approach

Mermelstein (*Mermelstein*, 1975) proposed a convex hull algorithm to segment the speech signal into syllabic units using maxima and minima in a loudness measure extracted from the speech signal. The basic idea of the method was to find the prominent peaks and dips. The prominent peaks were marked as syllabic peaks and the points near the syllabic peaks with maximal difference in the loudness measure were marked as syllable boundaries. Although this work was limited to segmenting the speech signal into syllabic units rather than recognizing the speech signal, the idea of using the convex hull was utilized later by Espy-Wilson (*Espy-Wilson*, 1994), Bitar (*Bitar*, 1997) and Howitt (*Howitt*, 2000) in locating sonorant consonants and vowels in the speech signal.

2.1.2 Word or sentence recognition systems

The SUMMIT system

The SUMMIT system (*Zue et al.*, 1989; *Glass et al.*, 1996; *Halberstadt*, 1998; *Chang*, 1998) developed by Zue et al. uses a traditional front-end like MFCCs or auditory-based models to obtain multilevel segmentations of the speech signal. The segments

are found using either - (1) acoustic segmentation (*Glass and Zue*, 1988) method finds time instances when the change in the spectrum is beyond a certain threshold and (2) boundary detection methods that use statistical context dependent broad class models (*Chang and Glass*, 1997; *Lee*, 1998). The segments and landmarks (defined by boundary locations) are then analyzed for phonemes using Gaussian Mixture Models (GMMs) or multi-layer perceptrons. Results comparable to the best state-of-the-art results in phoneme recognition were obtained using this method (*Glass et al.*, 1996) and, with the improvements made by Halderstadt (*Halberstadt*, 1998), the best phoneme recognition results to date were reported. A probabilistic framework was proposed to extend the segment based approach to word and sentence level recognition. SUMMIT system has produced good results on continuous speech recognition as well (*Halberstadt*, 1998; *Chang*, 1998). This probabilistic framework is discussed below in some detail because the probabilistic framework used in the present work is similar to it in some ways, although there are significant differences that are discussed in brief towards the end of this section.

Recall that the problem in continuous speech recognition is to find a word sequence $\hat{W}$ such that

$$\hat{W} = \arg \max_{W} P(W|O) \tag{2.1}$$

Chang (*Chang*, 1998) used a more descriptive framework to introduce the probabilistic framework of the SUMMIT system. In this framework, the problem of ASR is written more specifically as

$$\hat{W}\hat{U}\hat{S} = \arg \max_{WUS} P(WUS/O), \tag{2.2}$$

where U is a sequence of subword units like phones, diphones and triphones. S denotes the segmentation, that is, the start and end of each unit in the sequence. The observation sequence O has a very different meaning from that used in the context of HMM based systems. Given a multilevel segment-graph, and the observations extracted from the individual segments, the symbol O is used to denote the complete set of observations from all segments in the segment graph. This is a very different situation from HMM based systems where the observation sequence is the sequence of MFCCs and other parameters extracted at each frame of speech, identically for every frame. In the SUMMIT system, on the other hand, the acoustic measurements may be extracted in different ways in each segment.

Using successive applications of Bayes rule and because $P(O)$ is constant relative to the maximization, Equation 2.2 can be written as

$$\hat{W}\hat{U}\hat{S} = \arg\max_{WUS} P(O/WUS)P(S/WU)P(U/W)P(W) \tag{2.3}$$

$P(O|WUS)$ is obtained from the acoustic model, $P(S|UW)$ is the duration constraint, $P(U|W)$ is the pronunciation constraint, and P(W) is the language constraint. The acoustic measurements used for a segment are termed as 'features' for that segment and acoustic models are built for each segment or landmark hypothesized by a segment. This definition of 'features' is vastly different from the phonetic features used in this thesis. A particular segmentation (sequence of segments) may not use all the features available in the observation sequence O. Therefore, a difficulty is met in comparing the term $P(O/WUS)$ for different segmentations. Two different procedures have been proposed to solve this problem - Near-Miss Modeling

Module	Bitar	Liu	Ali	Salomon	Mermelstein	APHODEX	Fanty et al	SUMMIT	Chang
Knowledge based APs	Partial	Partial	Partial	Partial	No	Partial	Partial	No	Partial
Landmark detection	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Feature detection or phone classification	Partial	No	Partial	No	No	Partial	Yes	Yes	Yes
Sentence recognition	No	No	No	No	No	No	Partial	Yes	Yes

Table 2.1: The previous acoustic-phonetic methods and the scope of those methods

(*Chang*, 1998) and anti-phone modeling (*Glass et al.*, 1996).

A two-level probabilistic hierarchy, consisting of broad classes - vowels, nasals, stops, etc. - at the first level and phones at the second level was used in the SUMMIT system by Halberstadt (*Halberstadt*, 1998) to improve the performance of the recognition systems. Different acoustic measurements for phonemes belonging to different broad classes were used to carry out the phonetic discrimination. This is similar to a typical acoustic-phonetic approach to speech recognition where only relevant acoustic measurements are used to analyze a phonetic feature. But the acoustic measurements used in this system were the standard signal representation like MFCCs or PLPs, augmented in some cases by a few knowledge based measurements.

EBS is similar to SUMMIT in the sense that both the systems generate multiple segmentations and then use the information extracted from the segments or landmarks to carry out further analysis in a probabilistic manner. There are five significant factors that set the systems apart. First, SUMMIT is a phone based recognition system while EBS is a phonetic feature based system. That is, phonetic feature models are built in EBS instead of phone models. Secondly, although EBS uses a similar idea of obtaining multiple segmentations and then carrying further analysis based on the information obtained from those segments, it concentrates on linguistically motivated landmarks instead of analyzing all the front-end parameters extracted from segments and segment boundaries. Third, EBS utilizes the sufficiency and invariance properties of acoustic parameters in such a way that it does not need to account for all acoustic observations for each segmentation. Fourth, in

EBS, binary phonetic feature classification provides a uniform framework for speech segmentation, phonetic classification and lexical access. This is very different from the SUMMIT system where segmentation and analysis of segmentations are carried out using different procedure. Fifth, the SUMMIT system uses standard front-ends for recognition with a few augmented knowledge based measurements, and the proposed system uses only the relevant knowledge based APs for each decision.

Other Methods

A neural network based recognizer *Fanty et al.* (1992) that can be classified as an acoustic-phonetic approach was reported for word recognition. Speech is analyzed frame by frame for broad categories of phonemes using neural network classifiers. These categories are decided on the basis of perceptual and acoustic similarity rather than articulatory phonetic features. Speech is segmented on the basis of the frame level analysis, and the segments are then analyzed for the constituent phonemes using another set of neural networks. Different neural networks are used for each category of phonemes. Signal parameterization is composed of PLP coefficients augmented by certain knowledge based measurements. For certain acoustic measurements, landmarks like location of maximum zero crossing rate for fricatives are also used. On the studio quality ISOLET spoken letter corpus (ref) 96% accuracy was achieved. Performance on the telephone quality speech of the CSLU Whitepages corpus was reported at 89.1%, the best result at that time (1992) on the spoken alphabet task.

The system in (*Fanty et al.*, 1992) was the more advanced version of the FEA-

TURE system (*Cole et al.*, 1983) developed in the early 1980s for isolated letter recognition. The FEATURE system used some knowledge based measurements like energies in different frequency bands, zero crossing rate, etc. Four points were located in the utterance containing the isolated digit - the beginning of the utterance, the onset of the vowel, the vowel offset and the end of the utterance. A probabilistic classification tree based on grouping similar letters together was constructed. At each node of the tree, likelihoods were computed for the utterance to belong to the node using multivariate Gaussian probability distributions. Only relevant features were extracted at each node of the tree, that is a typical characteristic of a hierarchical acoustic-phonetic approach. Probabilities at each node leading to a terminal node were multiplied to come up with the probability of the terminal node representing a spoken letter. Although this is classified here as an acoustic phonetic approach, it should be noted here that this was not an articulatory feature based system.

A rule-based acoustic phonetic speech recognition system (APHODEX) in which speech is segmented into coarse classes - voiced plosives, unvoiced plosives, vowels, unvoiced fricative, voiced fricatives and sonorant consonants - was reported *Fanty et al.* (1992). The segments are then analyzed using two kinds of acoustic cues - strong cues and weak cues. If strong cues provide sufficient information about the phoneme in a broad class segment, a decision is made irrespective of the weak cues. If the strong cues do not provide sufficient information, weak cues are used for decoding. The acoustic cues used in decoding are knowledge based measurements like formant transitions and spectral peaks. The system outputs a phoneme lattice

that can be used for hypothesizing words and sentences. Recognition results at the word level were not presented for this system.

Log critical-band energies were used in a syllable-based speech recognition system (*Chang*, 2002) to obtain the manner level segmentation, classification of place of segments and identification of syllables. For manner segmentation, a frame classification accuracy of 85% was obtained and for place classification, accuracies ranging from 44% to 96% were obtained. A syllable-matching algorithm was used to get scores of different words. It was shown in this work that word errors in current large vocabulary recognizers depend directly on phone errors providing further evidence of the need for conducting fine acoustic-phonetic analysis in speech. Further, it was shown that the tolernance of the recognition systems to errors was dependent on the part of the syllable - onset or coda - where the articulatory feature is present, which shows the need to find accurate landmarks including vowel onsets and offsets. The approach in this work differs significantly from the work presented here because EBS is significantly more knowledge intensive and it utilizes the properties of knowledge based acoustic parameters appropriately.

2.2 Knowledge based front-ends

Some researchers have utilized acoustic cues that are correlates of phonetic features to form the front-end in HMM based ASR methods and other statistical methods. These methods traditionally use standard front-ends like MFCCs and LPC coefficients. The use of acoustic phonetic knowledge in the fronts-ends in these systems

led to improvement in performance using certain performance criteria.

Bitar and Espy-Wilson (*Bitar*, 1997) showed that acoustic-phonetic knowledge based acoustic parameters perform better than the standard MFCC based signal representation on the task of broad class segmentation using an HMM based back end. In particular, it was shown that the decrease in performance was much less dramatic for the knowledge based front-end than for MFCCs when cross-gender testing was carried out, that is, when training was done on males and testing was done on females, and vice versa. These experiments were extended to isolated word recognition (*Deshmukh et al.*, 2002) and a similar pattern was observed not only for cross gender testing but also for testing across adults and children whose speech can be from different databases.

Hosom (*Hosom*, 2000) augmented a PLP based front-end with five knowledge based acoustic measurements - intensity discrimination, voicing, fundamental frequency, glottalization and burst-related impulses - in a hybrid framework of HMMs and Artificial Neural Networks (ANNs). Three different ANNs were built, one for each of the multivalued distinctive features - Manner, Place and Height - and the outputs of these networks were combined to produce phoneme probabilities using fuzzy logic rules (a model called Fuzzy-Logic Model of Perception (*Massaro*, 1993) was used for combination). The observation probabilities of HMM states were estimated from these phoneme probabilities. Three more networks were used for the same distinctive features to estimate the phoneme transition probabilities that were further used to estimate the state transition probabilities in the HMM framework. A relative reduction in error rate of 26% was obtained on the task of automatic

alignment of phonemes in the TIMIT database over a baseline HMM/ANN system. When the time-alignment system was used to train the hybrid HMM/ANN for the OGI alphadigit task, a relative reduction in error rate of 10% was obtained.

2.3 Phonetic features as recognition units in statistical methods

In this category of ASR methods, the usual statistical frameworks use phonetic features as an intermediate units of recognition, and then use the outputs of the intermediate classifiers to recognize phonemes, words or sentences. These methods use no explicit knowledge of the acoustic correlates of phonetic features.

Deng (*Deng and Sun*, 1994) used five multi-valued articulatory features and their overlapping patterns to guide the topology of HMMs in an MFCC and HMM based speech recognizer. An HMM state is constructed for each bundle of phonetic features and those bundles are determined by a canonical representation of phonemes in terms of phonetic features as well as linguistic rules for change in the feature values for overlapping phonemes. For each phoneme sequence (a sentence), a graph of hidden states is constructed using the mapping of phonemes to feature bundles. The composite HMM is then trained using the Baum-Welch algorithm. An improvement in phoneme classification accuracy in the range 15%-27% was obtained over a baseline context-independent recognition system.

Eide et al. (*Eide et al.*, 1993) proposed a method of phoneme classification us-

ing a phonetic feature bundle representation of phonemes. Probabilities of phonetic features at each frame in a phoneme segment were estimated using Gaussian mixture models. Probabilities of different phonemes for given hand-segmented phoneme regions were estimated from the phonetic feature probabilities at each frame within the segments under analysis. The latter estimate was obtained using the frequency of the phonetic features occurring in the phoneme segment in the training data. A phoneme classification result of 70% was obtained. This is not a direct acoustic-phonetic approach because it lacks the use of landmarks and knowledge based signal representation.

Kirchoff (*Kirchhoff*, 1999) used five multivalued articulatory features as intermediate classification units in a hybrid HMM/ANN approach. The observation densities of HMM states in this system were modeled using ANNs instead of Gaussian mixtures. The posterior probabilities of each feature value at each HMM state were obtained from the output of the ANNs. These posterior probabilities were then combined to extract the posterior phone probabilities, that were converted to likelihoods. An improvement over a baseline HMM/ANN system was observed, especially when the signal was corrupted with noise.

2.4 Conclusions from the literature survey

While there have been many attempts at an acoustic-phonetic approach to ASR, only one of them - the SUMMIT system - has been able to match the performance of HMM based methods on practical recognition tasks. The other acoustic-phonetic

methods were stopped at the level of finding distinctive acoustic correlates of phonetic features, detection of landmarks or broad class recognition. Although the SUMMIT system carries out segment based speech recognition with some knowledge based measurements, it is not a landmark based system in the strict sense, nor a phonetic feature based system. Like HMM based systems, it uses all available acoustic information (for example, all the MFCCs) for all decisions. But the success of SUMMIT has been motivating because it appears to be the only 'static' approach that actually works on practical tasks. Acoustic phonetics knowledge and the concept of phonetic features has been used with HMM based systems with some success, but that only marginally adds to these systems an enhanced ability to recognize at the level of phonemes. In conclusion, there is no acoustic-phonetic approach to ASR that explicitly targets linguistic information in the speech signal as well as carries out practical recognition tasks.

Chapter 3

A Probabilistic Framework

The problem of recognition of bundles of features can be expressed as maximizing the posterior probability of landmarks and the corresponding feature bundles, given the observation sequence O. That is,

$$\hat{U}\hat{L} = \arg \max_{UL} P(UL|O) = \arg \max_{UL} P(L|O)P(U|OL), \tag{3.1}$$

where $L = \{l_i\}_{i=1}^{M}$ is a sequence of landmarks and $U = \{u_i\}_{i=1}^{N}$ is the sequence of phonemes or bundles of features corresponding to the phoneme sequence. The meanings of these symbols is illustrated in Table 3.1 for the digit "zero". There are several points to note with regard to the notation in Table 3.1.

1. l_i denotes a set of related landmarks that occur together. For example, the syllabic peak (syllable nucleus) and the VOP occur together. Also certain landmarks may be repeated in the sequence. For example, when a vowel follows a sonorant consonant, the sonorant consonant offset and the vowel onset are identical.

2. Each set of landmarks l_i is related to a broad class B_i of speech selected from the set {vowel (V), fricative (Fr), sonorant consonant (SC), stop burst (ST), silence (SIL)} as shown in Table 3.2. For example, the syllabic peak and the VOP are related to the broad class V. Let $B = \{B_i\}_{i=1}^{M}$ denote the sequence of broad classes corresponding to the sequence of sets of landmarks L. Note that ST denotes the burst region of a stop consonant, and the closure region is assigned the broad class SIL.

3. The number of the set of landmarks M and the number of bundles of phonetic features N may not be the same in general. This difference may occur because a sequence of sets of landmarks and the corresponding broad class sequence, for example, SIL-ST, may correspond to one set of phonetic features (the closure and the release constitute one stop consonant) or two bundles (closure corresponds to one stop consonant and release corresponds to another stop consonant, e.g, the cluster /kt/ in the word "vector"). Also, one set of landmarks or the corresponding broad class may correspond to two sets of place features, for example, in the word "omni" with the broad class sequence V-SC-V, the SC will have the features of the sound /m/ (calculated using the SC onset) as well the sound /n/ (calculated using SC offset).

The landmarks and the sequence of broad classes can be obtained deterministically from each other, for example, the sequence B = {SIL,Fr,V,SC,V,SC,SIL} for "zero" in Table 3.1 will correspond to the sequence of sets of landmarks L shown. Therefore

$$P(L|O) = P(B_L|O) \tag{3.2}$$

Table 3.1: An illustrative example of the symbols B, L and U

	/z/	/I/	/r/	/o/	/w/
$U \Rightarrow$	u_1	u_2	u_3	u_4	u_5
	$-sonorant$	$+sonorant$	$+sonorant$	$+sonorant$	$+sonorant$
	$+continuant$	$+syllabic$	$-syllabic$	$+syllabic$	$-syllabic$
	$+strident$	$-back$	$-nasal$	$+back$	$-nasal$
	$+voiced$	$+high$	$+rhotic$	$-high$	$+labial$
	$+anterior$	$+lax$		$+low$	
$B \Rightarrow$	Fr	V	SC	V	SC
$L \Rightarrow$	l_1	l_2	l_3	l_4	l_5
	Fon	VOP	Son	VOP	Son
	Foff	P	D	P	D
			Soff		Soff

[ht]

Table 3.2: Landmarks and corresponding broad classes.

Broad Class Segment	Landmark Type
Vowel	Syllabic peak (P)
	Vowel onset point (P)
Stop	Burst
SC	Syllabic dip (D)
	SC onset (Son)
	SC offset (Soff)
Fricative	Fricative onset (Fon)
	Fricative offset (Foff)

where B_L is a sequence of broad classes for which the landmark sequence L is obtained. Note that there is no temporal information contained in B, U and L except for the order in which the symbols occur. This equivalence of broad classes and landmarks is not intended as a general statement and it holds only for the landmarks and broad classes shown in Table 3.2.

3.1 Segmentation using manner phonetic features

Given a sequence of T frames $O = \{o_1, o_2, ..., o_T\}$, where o_t is the vector of APs at time t, the most probable sequence of broad classes $B = \{B_i\}_{i=1}^{M}$ and their durations $D = \{D_i\}_{i=1}^{M}$ have to be found. The frame o_t is considered as the set of all APs computed at frame t, although EBS does not use all the APs in each frame. EBS

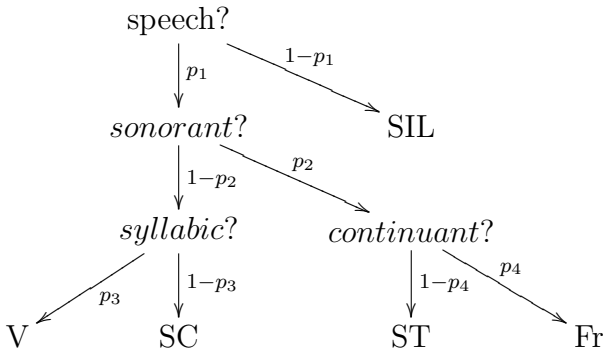


Figure 3.1: Probabilistic Phonetic Feature Hierarchy

uses the probabilistic phonetic feature hierarchy shown in Figure 3.1 to segment speech into the five manner classes. The broad class segmentation problem can be stated mathematically as,

$$\hat{B}\hat{D} = \arg \max_{BD} P(BD/O) \tag{3.3}$$

Provided that the frame at time t lies in the region of one of the manner classes, the posterior probability of the frame being part of a vowel at time t can be written as

$$P_t(V|O) = P_t(\text{speech}, \textit{sonorant}, \textit{syllabic}|O) \tag{3.4}$$

$$= P_t(\text{speech}|O) P_t(\textit{sonorant}|\text{speech}, O) P_t(\textit{syllabic}|\textit{sonorant}, O) \tag{3.5}$$

where P_t is used to denote the posterior probability of a feature or a set of features at time t. Similar expression can be written for each of the other manner classes.

Calculation of the posterior probability for each feature requires only the acoustic correlates of that feature. Furthermore, to calculate the posterior probability of a manner phonetic feature at time t, only the acoustic correlates of the feature in a set

of frames $\{t-s, t-s+1, ..., t+e\}$, using s previous frames and e following frames along with the current frame t, are required to be used. Let this set of acoustic correlates extracted from the analysis frame and the adjoining frames for a feature f be denoted by x_t^f. Then equation 3.5 can be rewritten as

$$\begin{aligned} P_t(V|O) =& P_t(\text{speech}|x_t^{speech}) P_t(sonorant|\text{speech}, x_t^{sonorant}) \\ & P_t(syllabic|sonorant, x_t^{syllabic}) \end{aligned} \tag{3.6}$$

The probability $P(BD|O)$ can now be expanded in terms of the underlying manner phonetic features of each broad class. Denote the features for class B_i as the set $\{f_1^i, f_2^i, ..., f_{N_{B_i}}^i\}$, the broad class at time t as b_t, and the sequence $\{b_1, b_2, ..., b_{t-1}\}$ as b^{t-1}. Note that B is the broad class sequence with no duration information. On the other hand, b_t denotes a broad class at time t. Therefore, the sequence b^t includes duration information. Making a stronger use of the definition of acoustic correlates by assuming that the acoustic correlates of a manner feature at time t are sufficient even if b^{t-1} is given,

$$P(BD|O) = \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} P_t(B_i|O, b^{t-1}) \tag{3.7}$$

$$= \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i|x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i, b^{t-1}) \tag{3.8}$$

In the above equation, $\sum_{j=1}^{i-1} D_j$ is the sum of the durations of the $i-1$ broad classes before the broad class i, and $\sum_{j=1}^{i} D_j$ is the sum of durations of the first i broad classes. Therefore, $\sum_{j=1}^{i-1} D_j - \sum_{j=1}^{i} D_j$ is the duration of the i^{th} broad class and hence the numbers $\{1+\sum_{j=1}^{i-1} D_j, ..., D_i+\sum_{j=1}^{i-1} D_j\}$ are the frame numbers of the

frames that occupy the i^{th} broad class. Now expanding the conditional probability,

$$= \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i, x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i, b^{t-1})}{P_t(x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i, b^{t-1})}. \tag{3.9}$$

Splitting the priors,

$$P(BD|O) = \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i|f_1^i, ..., f_{k-1}^i, b^{t-1}) \frac{P_t(x_t^{f_k^i}|f_1^i, ..., f_k^i, b^{t-1})}{P_t(x_t^{f_k^i}|f_1^i, ..., f_{k-1}^i, b^{t-1})}. \tag{3.10}$$

Clearly

$$\prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i|f_1^i, ..., f_{k-1}^i, b^{t-1}) = P(BD) = P(B)P(D|B) \tag{3.11}$$

Now given the set $\{f_1^i, ..., f_{k-1}\}$ or the set $\{f_1^i, ..., f_k\}$, $x_t^{f_k^i}$ is assumed to be independent of b^{t-1}. The independence of the APs given the set $\{f_1^i, ..., f_k\}$ is hard to establish, but it can be shown to hold better for the knowledge-based APs than mel-frequency cepstral coefficients (MFCCs) under certain conditions as discussed in Section 3.3. In words, this independence means that the APs for a phonetic feature are assumed to be invariant with the variation of the broad class of neighboring frames, for example, the APs for the feature sonorant are assumed to be invariant of whether the sonorant frame lies after vowel, nasal or fricative frames. This is further discussed in Section 3.3. Making this independence or invariance assumption,

$$P(BD|O) = P(B)P(D|B) \prod_{i=1}^{M} \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i|x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i)}{P_t(f_k^i|f_1^i, ..., f_{k-1}^i)}. \tag{3.12}$$

The posteriors $P_t(f_k^i|x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i)$ are directly obtained in this work from the SVM based classifiers using binning (*Drish*, 1998). The discriminant space of the

SVMs is split into bins and the posterior of a particular class is estimated as the ratio of the number of samples of that class in the bin to the total number of samples in that bin.

The term $P_t(f_k^i|f_1^i, ..., f_{k-1}^i)$ normalizes the imbalance of the number of positive and negative samples in the training data. For example, if equal number of training samples were used to find the posterior in the binning method, the estimate of the posterior probability is not biased toward a particular class. But, for example, if the number of training samples of class $+1$ is twice that of the number of samples of the class -1, the estimate of the posterior of the $+1$ class is $4/3$ times that of the case where equal number of samples were used. Similarly, the estimate of the posterior of the -1 class is $2/3$ times that of the case where equal number of samples were used. The denominator in this case will divide the posterior of class $+1$ by $2/3$ and the posterior of the class -1 by $1/3$. Assume a particular bin where the correct posterior is $1/2$ for both the classes, then the scores of $2/3 \div 2/3$ for the class $+1$ and $1/3 \div 1/3$ for the class -1 are obtained using this normalization but these are not posteriors because these can be greater than one. These can be considered as likelihoods because this normalization is equivalent to conversion of a posterior probability to a likelihood by division from a prior. But note that here the likelihood is that of only the relevant observations and not all observations.

The computation of $P(BD|O)$ for a particular B and all D is a very computationally intensive task in terms of storage and computation time. Therefore, an approximation is made that is similar to the approximation made by Viterbi decoding in the HMM based recognition systems and the SUMMIT system (*Glass et al.*,

1996),

$$P(B|O) \approx \max_{D} P(BD|O) \tag{3.13}$$

Because the probabilities $P(B|O)$ calculated this way for different B will not add up to one, the more correct approximation is

$$P(B|O) \approx \frac{\max_D P(BD|O)}{\sum_B \max_D P(BD|O)}, \tag{3.14}$$

although the term in the denominator is not relevant to the maximization in Equation (3.1).

A Viterbi-like probabilistic segmentation algorithm presented in the next section takes as input the probabilities of the broad manner phonetic features - *sonorant*, *syllabic* and *continuant* - and outputs the probabilities $P(B|O)$ under the assumption of Equation 3.13.

3.2 Probabilistic segmentation algorithm

The probabilistic segmentation algorithm is similar to (*Lee*, 1998) and the primary difference is that it operates only on binary posterior probabilities of phonetic features in each frame instead of calculating a 'segment score' which is a likelihood of observations in a segment. The algorithm has the four steps listed below. Please note that the algorithm below computes the probability $P(B|O)$ a bigram model for the prior $P(B)$. The prior can also be obtained from a sophisticated language model in constrained vocabulary recognition.

Denote by n the number of unique broad classes (five in this case) and call them β_i with i varying from 1 to n. A segmentation path will be denoted by a tuple

(B, D, Π) with the sequence of broad classes B, a sequence of durations D and the posterior probability of the segmentation Π. Let N^{best} denote the number of most probable paths required from the algorithm. It is assumed a bigram language model for the priors $P(B)$ is available but that is not necessary and the algorithm can be modified to consider other language models. Denote by D_{last} the last element in the sequence D and by B_{last} the last element in B.

1. Location of transition points

   Form a sequence of times when the probability ranking of the broad classes changes. Call the set of these times $\Gamma = \{\tau_i\}_{i=1}^{K}$ where K is the number of such locations. The change of a broad class along a segmentation path will only be allowed at these locations. This does not imply however that a class must change at a transition point.

2. Initialization

   Form a sequence of segmentations $\mathbb{S} = \{S_i\}_{i=1}^{N}$ where S_i is the segmentation (B^i, D^i, Π^i) such that $B^i = \{\beta_i\}$ and $D^i = \{\tau_1 - 1\}$. That is for each broad class, a path is defined with that single broad class in the class sequence and a duration given by the length of time before the first transition point. Set Π^i as

   $$\Pi^i = \prod_{t=1}^{\tau_1 - 1} P_t(\beta_i|O) P(\tau_1 - 1|\beta_i) P(\beta_i) \tag{3.15}$$

   and use Equation 3.12 to evaluate $P_t(\beta_i|O)$.

3. Forward computation

   for k from 1 to K, (begin loop 1)

(a) Initialize an empty set of segmentation paths $\mathbb{S}'$

(b) for i from 1 to n, (begin loop 2)

for each segmentation $S_j = (B^j, D^j, \Pi^j)$ in $\mathbb{S}$, (begin loop 3)

i. Create a new path $S' = (B', D', \Pi') = (B^j, D^j, \Pi^j)$,

ii. if B^j_{last} is same as β_i

- Assign $p_{dur} = P(D'_{last} + \tau_{k+1} - \tau_k | B^j_{last}) / P(D'_{last} | B^j_{last})$
- Assign $D'_{last} = D'_{last} + \tau_{k+1} - \tau_k$
- Assign $p_{trans} = 1$

else

- Append $\tau_{k+1} - \tau_k$ to D'
- Append β_i to B'
- Assign $p_{trans} = P(\beta_i | B'_{last})$
- Assign $p_{dur} = P(D'_{last} | B'_{last})$

iii. Update Π' as

$$\Pi' = \Pi^j \prod_{t=\tau_k}^{\tau_{k+1}-1} P_t(\beta_i|O) p_{dur} p_{trans} \tag{3.16}$$

and again using Equation 3.12 to evaluate $P_t(\beta_i|O)$.

iv. Append the path S' to the sequence of paths $\mathbb{S}'$

end loop 3

end loop 2

(c) For each path S' in $\mathbb{S}'$, if another path exists with same broad class sequence and greater probability, delete the path S' from $\mathbb{S}'$. This step implements the approximation in Equation 3.13

(d) Select the N^{best} paths in $\mathbb{S}'$ and delete the rest of the paths in $\mathbb{S}'$.

(e) Assign $\mathbb{S} = \mathbb{S}'$

end loop 1

4. The sequence $\mathbb{S}$ gives the N^{best} most probable segmentations.

3.3 Sufficiency and Invariance

Although it is not clear how sufficiency and invariance can be rigorously established for certain parameters, some idea can be obtained from classification and scatter plot experiments. For example, sufficiency of the four APs used for sonorant feature detection - periodicity, aperiodicity, energy in (100Hz,400Hz) and ratio of the energy in (0,F3) to the energy in (F3, half of sampling rate) - can be viewed in relation to 13 mel-frequency cepstral coefficients (MFCCs) in terms of classification accuracy of the sonorant feature. Using SVMs, a frame classification accuracy of 94.39% was obtained on TIMIT 'sx' sentences which compares well to 94.68% accuracy obtained using MFCCs, when all other test conditions were kept identical. In both the cases, a set of 10,000 randomely selected samples of each of the $+sonorant$ and $-sonorant$ frames were used for training and the same number of samples were extracted from the test set for testing.

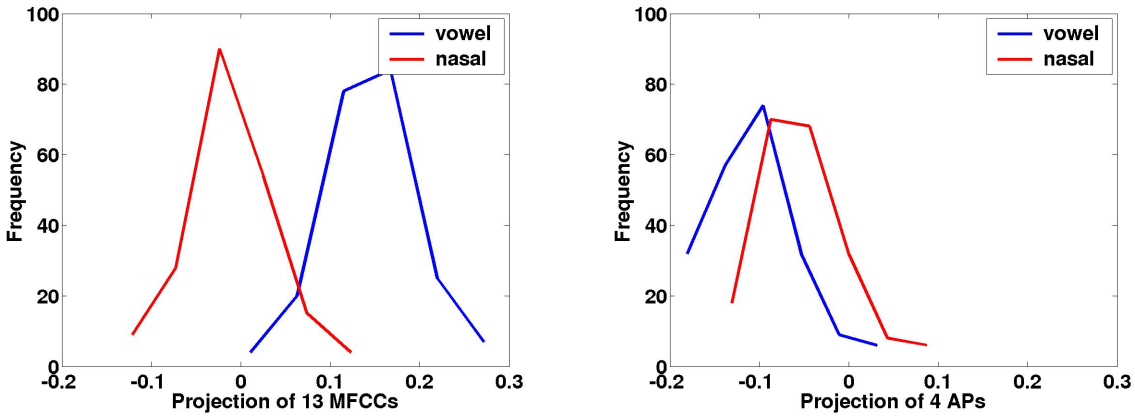
Invariance was assumed with variation in previous broad class frames in Equation 3.12 where the APs $x_t^{f_k^i}$ for a manner feature were assumed to be independent of the manner class labels of preceding frames b^{t-1} when $\{f_1^i, ..., f_k\}$ or $\{f_1^i, ..., f_{k-1}\}$ was given. First consider the case where $\{f_1^i, ..., f_k\}$ is given, that is, the value of the feature whose APs are being investigated is known. A typical case where the assumption may be hard to satisfy is when the APs for the sonorant feature are assumed to be invariant of whether the analysis frame lies in the middle of a vowel region or the middle of a nasal region, that is, b^{t-1} is composed of nasal frames in one case and vowel frames in the other case.

Such independence can roughly be measured by the similarity in the distribution of vowels and nasals based on the APs for the feature sonorant. To test this independence, 200 sets of *sonorant* APs for each nasals and vowels were extracted randomly from the TIMIT train set. Each set of APs was extracted from a single frame located at the center of the vowel or the nasal. The APs were then used to discriminate vowels and nasals using Fischer Linear Discriminant Analysis (LDA). Figure 3.2(a) shows the distribution of the projection of the 13 MFCCs into a one-dimensional space using LDA. A similar projection is shown for the four *sonorant* APs in Figure 3.2(b). It can be seen from these figures that there is considerably more overlap in the distribution of the vowels and the nasals for the APs of the *sonorant* feature than for the MFCCs. Thus, the APs for the sonorant feature are more independent of the manner context than are the MFCCs.

But there are certainly cases where neither APs nor MFCCs may satisfy the invariance assumption. For example, when using multiframe observations, the APs for

the *sonorant* feature may have different distributions at the fricative-vowel boundary and at the middle of a vowel. At the fricative-vowel boundary, the frames before the boundary frame are $-sonorant$ frames and at the middle of a vowel, the frames before the middle frame are $+sonorant$ frames. The multiframe acoustic observations are clearly different in the two cases if some frames previous to the current analysis frames are included. Boundaries are a small portion of the speech signal and it is hoped that the breakdown of this assumption should have little effect on recognition performance. Also note that this assumption is similar to the assumption in the HMM based approach where the likelihood of an observation is assumed to be dependent only on the current state.

The invariance of the APs $x_t^{f_k^i}$ for a manner feature f_k^i with the manner class labels of preceding frames b^{t-1} when only the features $\{f_1^i, ..., f_{k-1}\}$ is given is now considered. If only single frame observations are used, the observations may depend strongly on the broad class of the current frame b_t. But multiframe observations, especially if frames preceding the current analysis frame are used, are clearly dependent on the broad class sequence b^{t-1}. In most cases, multiframe observations are used in this work and this particular assumption will not be satisfied. As shown in Chapter 4, reasonable results are still obtained on broad class segmentation.



(a) (b)

Figure 3.2: (a) Projection of 13 MFCCs into a one-dimensional space with vowels and nasals as discriminating classes, (b) Similar projection for four APs used to distinguish +sonorant sounds from -sonorant sounds. Because APs for the sonorant feature discriminate vowels and nasals worse than MFCCs, they are more invariant

3.4 Constrained Landmark Detection for Word Recognition

For isolated word or connected word recognition, manner class segmentation paths can be constrained by a pronunciation model such as a Finite State Automata (FSA) (*Jurafsky and Martin*, 2000). The remaining phonetic features can then be estimated from the landmarks obtained in the segmentation process. Figure 3.3 shows an FSA based pronunciation model for the digit 'zero' and the canonical pronunciation /z I r ow/. The broad manner class representation corresponding to the canonical representation is Fr-V-SC-V-SC where it is assumed that the the offglide of the final vowel /ow/ may be recognized as a sonorant consonant. One

transition is made for each frame of speech, starting from the initial state $S0$, and the transition probability is equal to the posterior probability of the manner class that labels the transition. Starting with the start state S0, the best path through the FSA for 'zero' can be calculated using (1) the posterior probability of a manner class for each frame as a transition probability, and (2) the posterior probabilities of the features listed below each state once the search algorithm has exited out of that state and the next state (that is, when sufficient information is available for obtaining landmarks for those features).

Figure 3.3 is a simple case where only one set of features is associated with each broad class. Often two sonorant consonants may occur consecutively so that two sets of features have to be associated with the broad class SC. In such a case, the first set of features (for example the features $+labial$ and $+nasal$ for the sonorant consonant /m/ in the word "omni") are computed using the landmark associated with the onset of SC the second set of features associated with /n/ are computed using the consonant release. For connected word recognition, the FSAs of all the words can be connected through a SILENCE state and the best path can be found using the composite FSA. The probabilistic segmentation algorithm has been modified such that only those transitions allowed by the automata are made at each transition point.

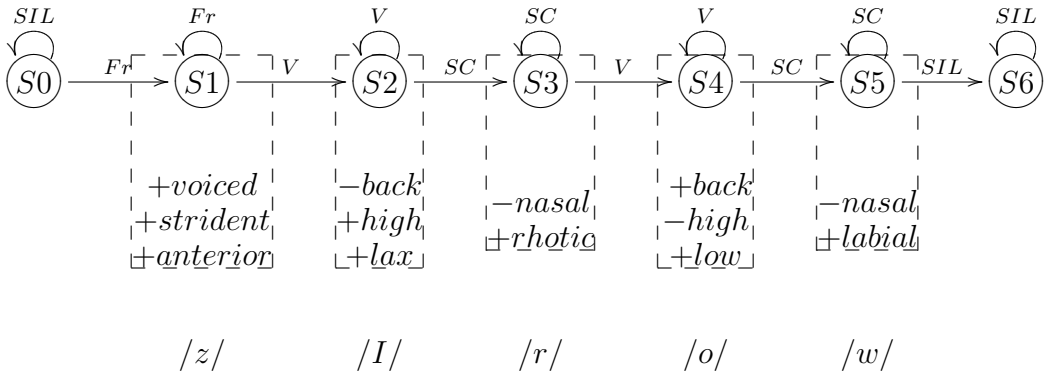


Figure 3.3: A phonetic feature based pronunciation model for the word 'zero'.

3.5 Probabilistic place and voicing detection

Using the acoustic landmarks obtained in the broad class recognition system, the probabilities of other manner phonetic features, and place and voicing features can be obtained. For example, given a manner class segmentation $B = \{\text{V,SC,V}\}$ or more explicitly, the corresponding sequence of landmarks $L = \{l_1, l_2, l_3\}$, and the observation vector O, to find the probability that the intervocalic SC is a nasal, the following acoustic observations need to be made (*Pruthi and Espy-Wilson*, 2003). (1) the energy offset at the SC onset, (2) the density of formants (resonances) at the SC syllabic dip, (3) an energy ratio at the SC syllabic dip, (4) the energy onset at the SC offset (vowel onset) and (5) the stability of the spectrum in the SC region. Let the set of APs extracted from the set of landmarks l_2 for a feature f be denoted by $x_{l_2}^f$ and the probability that the SC in the sequence V-SC-V is the phoneme /n/ be denoted by $P_2(/n/)$ (we use the index 2 because SC is the second broad class in the segmentation V-SC-V), we can write

$$P_2(/n/|O, L) = P(nasal|l_2, x_{l_2}^{nasal})P(alveolar|nasal, l_2, x_{l_2}^{alveolar}) \tag{3.17}$$

The assumption has been made that the SC landmarks and the acoustic correlates of the *nasal* and *alveolar* are sufficient to find the posterior probability of those features. In general, only the landmarks from adjoining broad class segments may be needed. For example, to find the probability that the SC in a V-SC-V sequence is an /r/ the measurement of the third formant (F3) in the adjoining vowels may be needed because /r/ is characterized by a sharp decline in F3 relative to the adjoining vowel. Therefore,

$$P_2(/r/|O,L) = P(-nasal|l_2, x_{l_2}^{nasal}) P(rhotic|-nasal, l_1, l_2, l_3, x_{l_1,l_2,l_3}^{alveolar}) \tag{3.18}$$

In general, if the bundle of features below the level of broad manner phonetic features for a phoneme u_i is represented by $\{f^i_{N_{B_i}+1}, f^i_{N_{B_i}+2}, ..., f^i_{N_i}\}$, then, given a sequence of landmarks $L = \{l_i\}_{i=1}^M$ and the observation sequence O, the conditional probability of the sequence of phonemes can be written as

$$P(U/OL) = \prod_{i=1}^{M} \prod_{k=N_{B_i}+1}^{N_i} P_i(f^i_k | f^i_{N_{B_i}+1}, ..., f^i_{k-1}, L, x^{f^i_k}_{l_{i-1},l_i,l_{i+1}}, u^{i-1}) \tag{3.19}$$

where the sufficiency of the acoustic correlates $x^{f^i_k}_{l_{i-1},l_i,l_{i+1}}$ has been assumed. This can be rewritten as

$$P(U/OL) = \prod_{i=1}^{M} \prod_{k=N_{B_i}+1}^{N_i} P_i(f^i_k | f^i_{N_i+1}, ..., f^i_{k-1}, L, u^{i-1}) \frac{P(x^{f^i_k}_{l_{i-1},l_i,l_{i+1}} | f^i_k, f^i_{N_{B_i}+1}, ..., f^i_{k-1}, L, u^{i-1})}{P(x^{f^i_k}_{l_{i-1},l_i,l_{i+1}} | f_{N_{B_i}+1}, ..., f^i_{k-1}, L, u^{i-1})} \tag{3.20}$$

It is straightforward to see that

$$\prod_{i=1}^{M} \prod_{k=N_{B_i}+1}^{N_i} P_i(f^i_k | f^i_{N_{B_i}+1}, ..., f^i_{k-1}, L, u^{i-1}) = P(U|L) \tag{3.21}$$

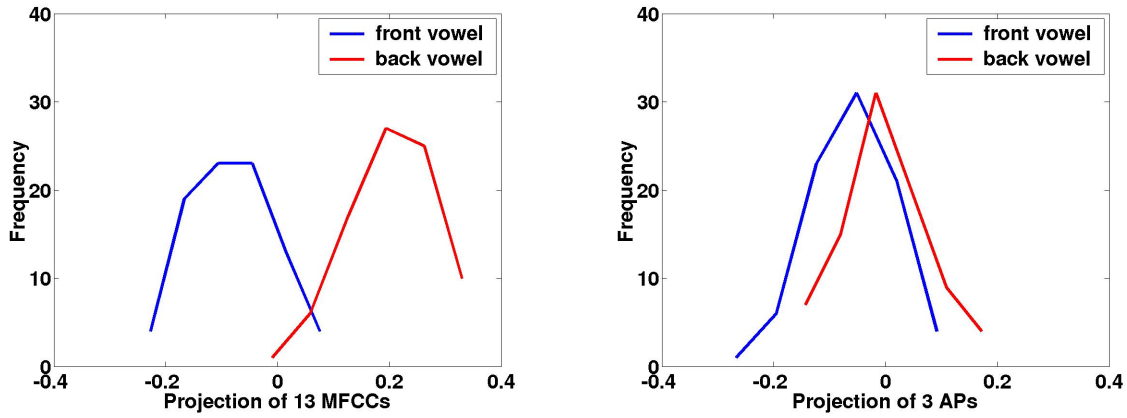
If the APs of the place features are assumed to be invariant of the place features of the place context, the term u^{i-1} can be ignored. Furthermore, the acoustic correlates

may depend on the manner of the current sound and the adjoining sounds, therefore, instead of keeping the complete landmark sequence, only the landmarks l_{i-1}, l_i, l_{i+1} may be kept in the above equation. For example, the acoustic correlates of the feature *alveolar* at a stop release may be dependent only on the presence of the closure, the release and whether the following sound is a vowel or a fricative, and not on the sound that is present before the stop closure. Making these assumptions,

$$P(U/OL) = P(U|L) \prod_{i=1}^{M} \prod_{k=N_{B_i}+1}^{N_i} \frac{P(f_k^i | x^{f_k^i}_{l_{i-1},l_i,l_{i+1}}, f^i_{N_{B_i}+1}, ..., f^i_{k-1}, l_{i-1}, l_i, l_{i+1})}{P(f_k^i | f_{N_{B_i}+1}, ..., f^i_{k-1}, l_{i-1}, l_i, l_{i+1})} \tag{3.22}$$

Again the numerator is obtained from the outputs of an SVM and the denominator is obtained from the fraction of positive or negative samples used in SVM training.

The invariance of APs of the place phonetic features extracted using the manner landmarks can also be assessed by the scatter of the APs with the change in context. First, the invariance of the acoustic correlates $x^{f_k^i}$ with the place features of the neighboring sounds is investigated when the feature f_k^i is given. To show the invariance, 200 samples of the stop consonant /t/ in prevocalic contexts were extracted in each of the two vowel contexts - front and back. LDA was then used to discriminate the two vowel contexts using three APs - Av, Ahi and Ahi-A23 - extracted from four frames each at the stop onset and at the vowel onset. These are the APs relevant for the distinction of the stop features *labial* and *alveolar*, and the feature *+alveolar* is assumed to be given (that is why the consonant /t/ is used). The same experiment was repeated by replacing APs by 12 MFCCs along with energy. As shown in in Figure 3.4, APs overlap considerably more than MFCCs across

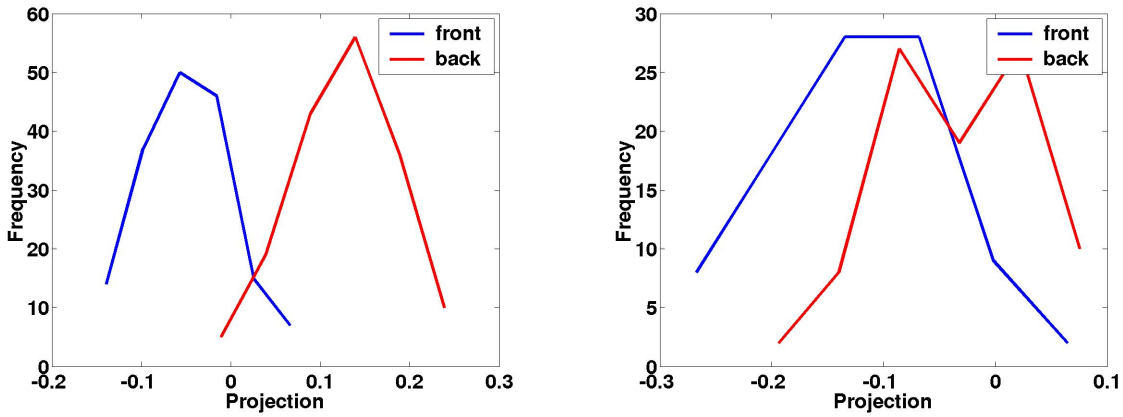


(a) (b)

Figure 3.4: (a) Projection of 13 MFCCs using Fisher LDA into a one-dimensional space with front and back vowel contexts as discriminating classes, (b) Similar projection for the three APs used to distinguish *+labial* stops from *+alveolar* stops. Because APs for stop place considerably overlap in different vowel contexts, they are more invariant of the vowel context. Samples of only the sound /t/ were used to obtain these plots.

the two vowel contexts showing that they are more invariant than the MFCCs.

Second, the case where the invariance of the acoustic correlates $x^{f_k^i}$ with the place features of the neighboring sounds is investigated when the feature f_k^i is not known. The experiment of discriminating the the vowel contexts was repeated but instead of using only the sound /t/, both the stop consonants /p/ and /t/ were used, that is, the value of the features *labial* and *alveolar* were not known. That is, only the features $f_1^i, ..., f_{k-1}^i$ were given. It can be seen from Figure 3.5 that the APs for distinguishing the place *labial* and *alveolar* of stop consonants is still more invariant than MFCCs even when the stop place is not known.



(a) (b)

Figure 3.5: (a) Projection of 13 MFCCs using Fisher LDA into a one-dimensional space with front and back vowel contexts as discriminating classes, (b) Similar projection for the three APs used to distinguish *+labial* stops from *+alveolar* stops. Because APs for stop place considerably overlap in different vowel contexts, they are more invariant of the vowel context. Samples of both the sounds /p/ and /t/ were used to obtain these plots.

APs are compared with MFCCs for the performance on the classification of the features *labial* and *alveolar* in Table 5.1. The three APs mentioned above perform reasonably well (78.24%) compared to the 13 MFCCs (84.53%) but the gap in performance is significant. Therefore, APs may not be truely sufficient for recognition but with certain improvements sufficiency can be reached.

Chapter 4

Landmark Detection Experiments

4.1 Database

The phonetically rich 'si' sentences from the training section of the TIMIT database was used for training and development. The 'sx' sentences from the test section of the TIMIT database were used for testing. The 2230 isolated digit utterances from the TIDIGITS training corpus were used for cross-database limited vocabulary testing. For the purpose of training, TIMIT phoneme labels were mapped to broad class labels.

4.2 Experiments and results

For binary classification experiments, one SVM was trained for each of the phonetic features and the corresponding positive and negative samples shown in Figure 3.1. Syllabic sonorant consonants (/em/, /el/, /en/, /er/ and /eng/) and diphthongs

(/iy/, /ey/, /ow/, /ay/, /aw/, and /uw/) were not used in the training of the feature *syllabic*, and affricates (/jh/ and /ch/) and glottal stops were not used in training of the feature *continuant*, but these sounds were used for frame-based testing. The APs in Table 4.1 were used for classification and both linear and Radial Basis Function (RBF) SVMs (*Vapnik*, 1995) were used for all the nodes in the feature hierarchy (see Figure 3.1) - *speech*, *sonorant*, *syllabic*, and *continuant*. For the feature *continuant*, the stop burst frame identified as the first frame of a stop consonant using TIMIT labeling was trained against all fricative frames. For the other features, all frames for each of the classes were extracted as training samples. Training was performed on randomly picked samples from the 'si' sentences of the TIMIT training set, and testing was performed on randomly picked samples from the 'sx' sentences of the TIMIT test set. The number of adjoining frames used for classification of each feature were optimized by minimizing the error on a separate set of randomly picked frames from the training 'si' sentences.

4.2.1 Frame-based results

Figure 4.1 shows how the classification results vary as the number of previous frames s is varied for each of the four manner classifiers. Similar plots were obtained for the number of following frames e. The optimal values were chosen as the ones where tthe first dip in the plots appeared. The values of the two variables were then used to get binary classification results on the complete 'sx' portion of the TIMIT database (instead of using randomly picked samples). The binary classification results at the

Table 4.1: APs used in broad class segmentation. f_s : sampling rate, F3 : third formant average, [a,b]: frequency band [aHz,bHz], E[a,b]: energy in the frequency band [aHz,bHz]

Phonetic Feature	APs
Silence	(1) E[0,F3-1000], (2) E[F3,f_s/2], (3) ratio of spectral peak in [0,400Hz] to the spectral peak in [400,f_s/2], (4) Energy onset (5) Energy offset
sonorant	(1) Temporal measure of periodicity, (2) Temporal measure of aperiodicity (3) Ratio of E[0,F3-1000] to E[F3-1000,f_s/2], (4) E[100,400]
syllabic	(1) E[640,2800] (2) E[2000,3000] (3) Temporal measure of periodicity (4) Temporal measure of aperiodicity (5) Total energy
continuant	(1) Temporal onset measure, (2) Temporal offset measure, (3) E[0,F3-1000], (4) E[F3-1000,f_s/2]

Table 4.2: Binary classification results for manner features in %

Feature	s	e	Accuracy on middle frames	Accuracy on all frames
sonorant	4	1	96.55	94.39
syllabic	16	24	86.44	81.69
Speech/silence	3	2	94.74	93.47
continuant	4	4	-	95.58

optimal values of s and e are shown in Table 4.2 in two cases - (1) when all the frames were used for testing and (2) when only the middle one-third portion of each broad class was used for testing. The difference in the results indicates the percentage of errors that are made due to boundary or coarticulation effects. Note that in the presented landmark-based system, it is not important to classify each frame correctly. The results on the middle one-third segment are more representative of the performance of the system because if the frames in a stable region are correctly recognized for a particular manner feature, this would mean that the corresponding landmarks may still be correctly obtained. For example, if the middle frames of an intervocalic sonorant consonant are correctly recognized as $-syllabic$, then the correct recognition of frames near the boundary is not significant because landmarks for the sonorant consonant will be obtained accurately. For the feature *continuant*, the classification error on middle frames is not relevant because the SVM is trained to extract the stop burst as opposed to a certain stable region of speech.

Figures 4.2 and 4.3 show the most significant sources of error for each of the phonetic features. The errors include misclassifications of the $+feature$ sounds

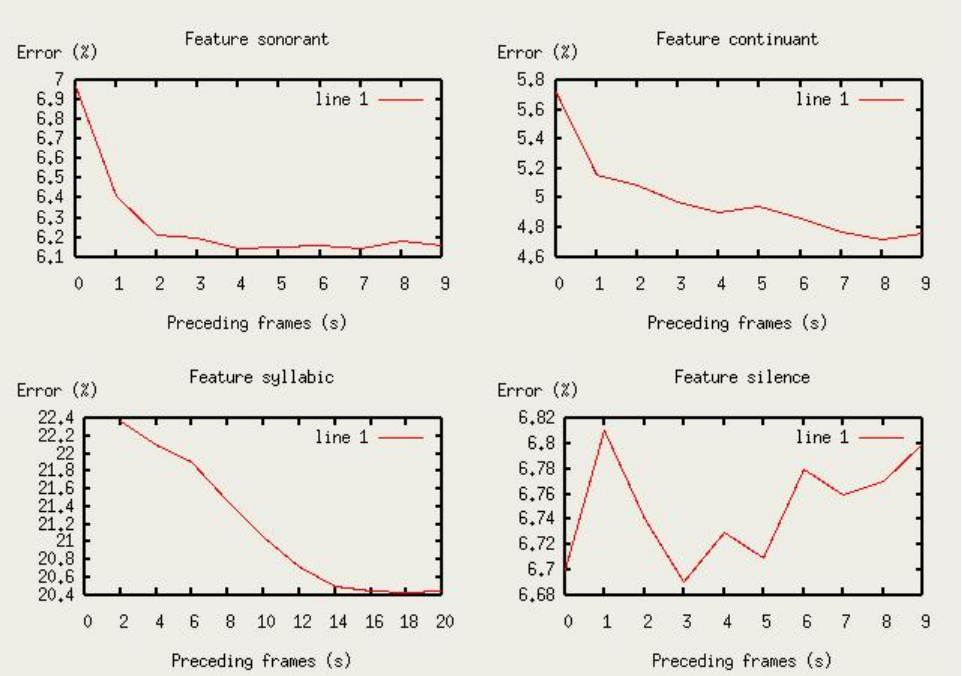


Figure 4.1: Variation in error with the number of preceding frames

as $-feature$, and vice versa. For the feature *sonorant*, it can be seen that the sounds /v/ and the glottal stop /q/ are often detected as $+sonorant$. A separate detector is required either at the broad class recognition level or further down the hierarchy to recognize glottalization because it can be significant for lexical access, especially in the detection of the consonant /t/. The sound /v/ is many times manifested as a sonorant consonant so that the assignment of $+sonorant$ for /v/ is expected. For the feature *syllabic*, classification accuracy for nasals as -syllabic is above 90%. But the semivowels - /y/, /r/, /l/ and /w/ have lower accuracies which is expected because of the vowel-like behaviour of these sounds. About 15% of the frames of reduced vowels are also misrecognized as sonorant consonants. This typically happens when there is a sonorant consonant in the intervocalic context of a stressed vowel and a reduced vowel such that the reduced vowel is confused as a

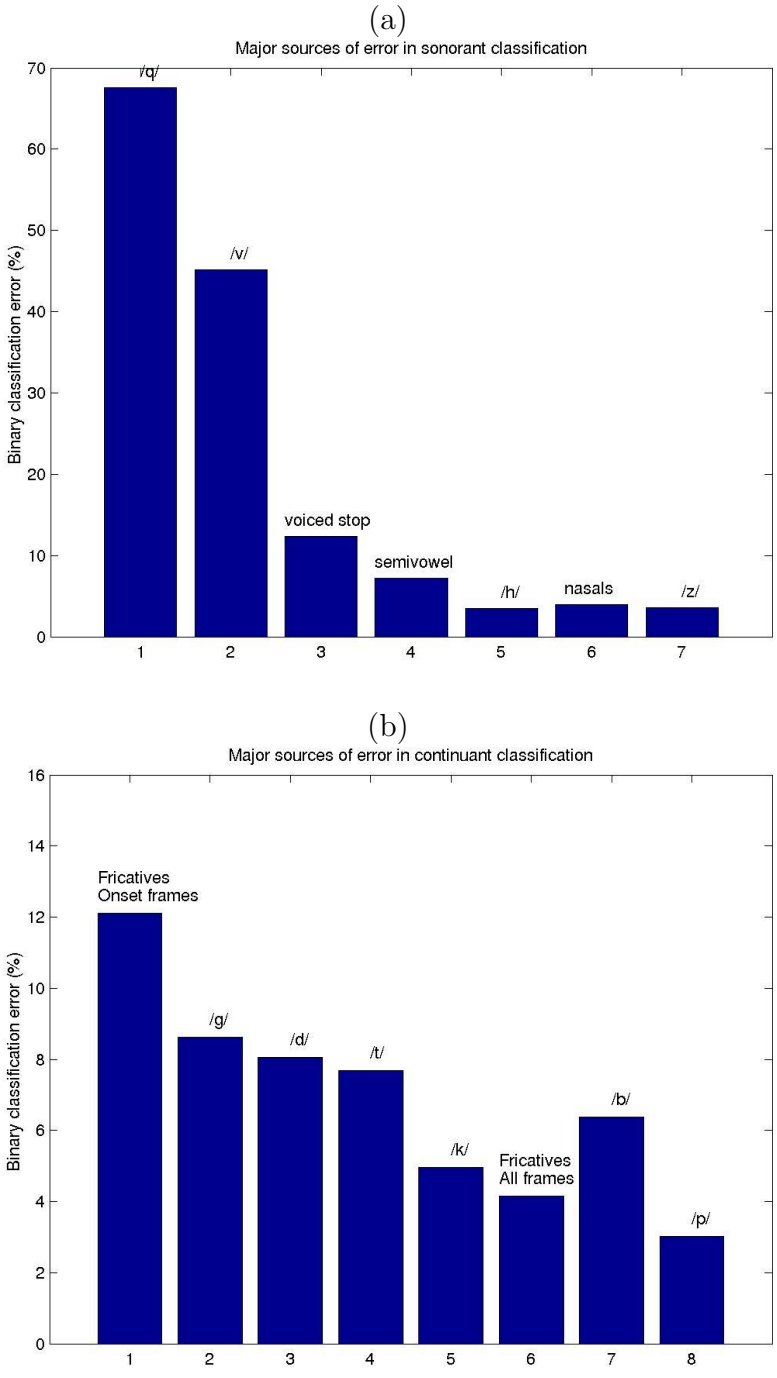
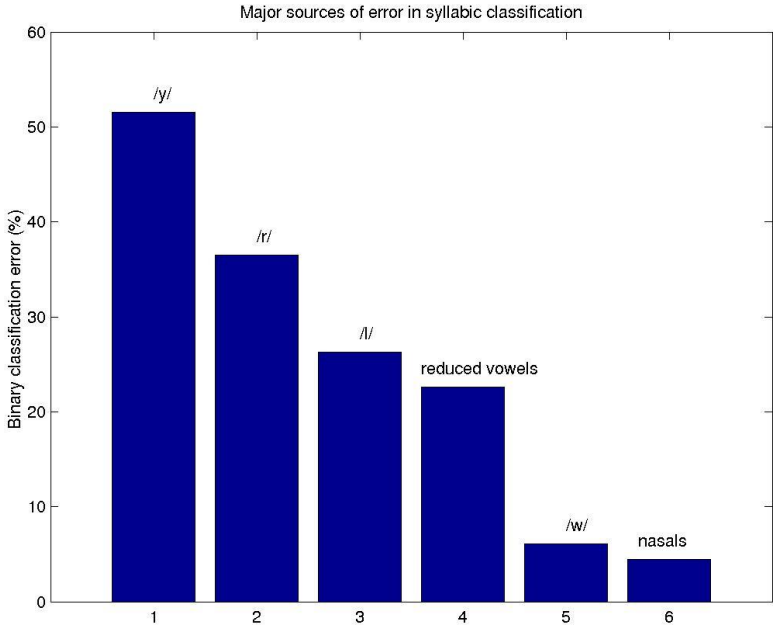


Figure 4.2: Sounds with high error percentages for the features (a) *sonorant* and (b) *continuant*.

(a)



(b)

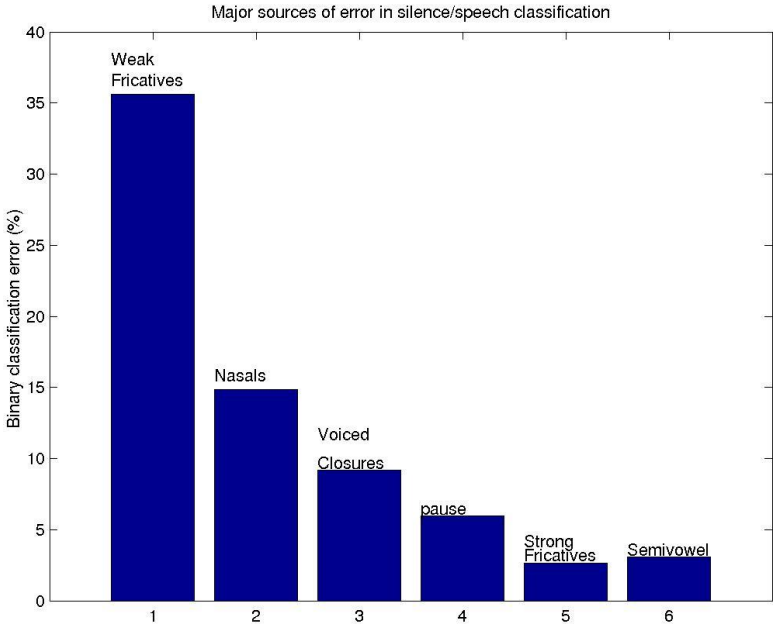


Figure 4.3: Sounds with high error percentages for the features (a) *syllabic* and (b) silence.

continuation of the sonorant consonant. A similar result was shown earlier (*Howitt*, 2000) where the reduced vowels showed maximum error in the deletion of vowel landmarks. The performance of the feature *continuant* is 95.58% which indicates the accuracy on classification of onset frames of all non-sonorant sounds. That is, an error was counted if a stop burst was wrongly classified as $-$*continuant* or a fricative onset was wrongly classified as a stop burst. The major source of error is the misclassification of 13.74% of fricative onsets as stop bursts. This is usually expected in word-initial fricatives.

4.2.2 Sequence-based results

The SVM models obtained in the frame-based analysis procedure were used to obtain broad class segmentation as well as the corresponding landmark sequences for the 840 'sx' sentences of the TIMIT test set using the probabilistic segmentation algorithm. Not all broad class sequences were allowed as the segmentation paths were constrained using a pronunciation graph such that (1) SCs only occur adjacent to vowels, (2) ST is always preceded by SIL and (3) each segmentation path starts and ends with silence. The duration probability for each broad class was modeled by a mixture of Rayleighs using a single Rayleigh density for the classes SC, V, Fr and ST, and a mixture of two Rayleigh densities for SIL (one density targets short silence regions like pauses and closures and the other density targets beginning and ending silence). The parameter for each Rayleigh density was found using the empirical means of the durations of each of the classes from the the TIMIT training data.

Table 4.3: Allowed splits, merges and substitutions

Reference	Allowed hypothesis	Reference	Allowed hypothesis
V+V	V	SC + SC	SC
Fr + Fr	Fr	SIL + SIL	SIL
/q/ + V, V + /q/	V	/q/	ST, SC
/t/, /p/, /k/, /g/, /d/	ST+Fr	/v/	SC, Fr
/em/, /en/, /er/, /el/	V+SC	/ch/, /jh/	ST+Fr
/hv/	SC, Fr	/dx/	SC
/dx/	SILEN + ST	/iy/, /ow/, /ey/, /oy/, /aw/, /uw/, /ow/	V+SC

All allowable broad class sequences were considered to be equiprobable, that is, priors were not used in the landmark detection procedure. The 'score' of a particular sequence of broad classes B and its durations was thus computed as

$$\bar{P}(B|O) = \prod_{i=1}^{M} P(D_i|B_i) \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i|x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i)}{P_t(f_k^i|f_1^i, ..., f_{k-1}^i)}. \tag{4.1}$$

For the purpose of scoring, the reference phoneme labels from the TIMIT database were mapped to manner class labels. Some substitutions, splits and merges as shown in Table 4.3 were allowed in the scoring process. Specifically, note that two identical consecutive broad classes were allowed to be merged into one since the distinction between such sounds is left to the place classifiers. Also note that affricates were allowed to be recognized as ST+Fr as well as Fr, and similarly diphthongs -

/iy/, /ey/, /ow/, /ay/, /aw/, and /uw/ - were allowed to be recognized as V+SC as well as V because the off-glides may or may not be present. Scoring was done purely on the sequences of hypothesized symbols without using time information.

The same knowledge based APs were used to construct a 14-parameter front-end for an HMM based broad class segmentation system. The comparison with the HMM-based system is not for the purpose of establishing that the presented system performs better than the HMM-based systems, but to show an acceptable level of performance. All the HMMs were context-independent 3-state (excluding entry and exit states) left-to-right HMMs with diagonal covariance matrices and 8-mixture observation densities for each state. All the 'si' utterances from the TIMIT training set were used for training the HMM broad classifier. The segmentation was identically constrained for both the HMM system and EBS while testing on TIDIGITS as well as TIMIT. The results are shown in Table 4.4. The results are also shown for EBS for two different front-ends - AP and MFCC. The performance of all the systems, except when EBS is used with MFCCs, is comparable although the HMM-MFCC system gives the maximum accuracy. However as shown in the next section, the MFCC based systems show worse generalization in cross-database testing. The inferior performance of MFCCs with EBS is perhaps because of the better agreement of APs with the invariance assumptions of the probabilistic framework (*Juneja and Espy-Wilson*, 2004). Similarly, better performance of MFCCs in the HMM framework may be because of better agreement with the diagonal covariance assumption of the HMM system applied here. That is, APs are not processed by a diagonalization step prior to application to the HMM systems while MFCCs go

Table 4.4: Broad class segmentation results

	EBS (RBF)	EBS (linear)	HMM
	Corr/Acc	Corr/Acc	
AP	86.7/79.5	84.3/77.8	83.4/78.1
MFCC	76.4/68.0	-	87.7/80.3

Table 4.5: Confusion matrix for segmentation with exclusion of affricates, syllabic sonorant consonants, /v/, glottal stop /q/, diphthongs and flap /dx/

	Total	Fr	SILEN	V	SC	ST	Deletions	Correct (%)
Fr	3627	3179	6	0	80	115	247	88.20
SILEN	6102	7	5614	11	56	0	414	92.00
V	6565	34	35	5724	10	23	739	87.19
SC	5504	81	30	0	4565	32	796	82.94
ST	3417	195	0	10	75	2755	382	80.63
Insertions		394	520	167	616	520		

through such a process. These are possible explanation of these results and they are open to further investigation.

An example of landmarks generated by EBS on a test sentence of TIMIT is shown in Figure 4.4 which also shows how errors in the system can be easily analyzed. Two kinds of errors are shown in this picture. First, based on the dip in the measure E[2000,3000], the pattern recognizer detects an intervocalic SC, even though the SC is postvocalic. Second, based on the AP E[2000,3000] which is meant

Table 4.6: Confusion matrix for affricates, syllabic sonorant consonants (SSCs), /v/, glottal stop /q/, diphthongs and flap /dx/. Empty cells indicate that those confusions were scored as correct but the exact number of those confusions were not available from the scoring program.

	Total	Fr/ ST+Fr	SILEN	V/ V+SC	SC	ST	Deletions	Correct (%)
/q/	534	2		0	5			99.63
Diph	2557	23	17	2310	9	2	196	90.34
SSCs	789	7	14	647	12	1	107	82.00
/v/	392		15	0		0	39	86.22
/dx/	336	11	2	0			51	80.95
/ch/, /jh/	396	393	0	1	0	0	2	99.24

Table 4.7: Broad class results on TIDIGITS

	EBS (linear)	EBS (RBF)	HMM-AP	HMM-MFCC
Constrained	91.8/85.1	91.5/85.1	91.3/85.8	92.3/84.2
Unconstrained	93.3/74.6	92.3/78.2	88.7/79.5	88.3/74.8

to find /r/-colored regions, the pattern recognizer proposes a SC at the beginning of the sonorant region of "Charlie" (ellipse 2). Inspection of the spectrogram shows that the vowel and /r/ are completely merged and further analysis is required to unravel the merged sounds.

The confusion matrix for EBS using the AP front-end is shown in Table 4.5 without including the sounds - diphthongs, syllabic sonorant consonants, flaps, /v/, affricates and the glottal stop /q/. For these latter set of sounds the confusion matrix is shown in Table 4.6.

4.2.3 Word-level results

Figure 4.5 shows an example of the output of the unconstrained probabilistic segmentation algorithm for the utterance 'two' with canonical pronunciation /t uw/. The two most probable landmark sequences obtained from the algorithm are shown in this figure. The landmark sequence obtained with the second highest probability for this case is the correct sequence. It is hoped that once probabilistic place and voicing decisions are made, the second most probable sequence of landmarks will yield an overall high posterior word probability for the word "two".

To get the results on constrained segmentation, the segmentation paths were

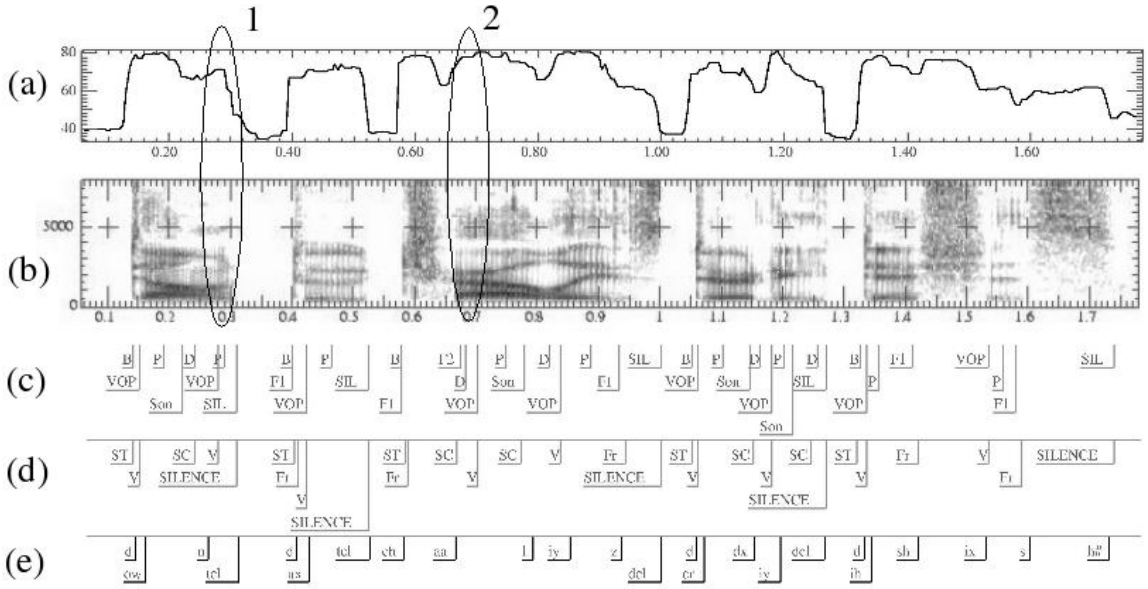


Figure 4.4: (a) E[2000,3000], (b) Spectrogram of the utterance, "don't do Charlie's dirty dishes", (c) Landmark labels, (d) broad class labels, and (e) phoneme labels. Note that the broad class and phoneme labels are marked at the beginning of each sound, and the landmark labels show the time instant of each landmark. The ellipses 1 and 2 show the two errors made by the system on this utterance. In 1, E[2000,3000] dips in the nasal region and then rises sharply indicating the presence of a vowel although no vowel is present. In 2, E[2000,3000] does not dip in the region of vowel /aa/ (although the vowel is /r/-colored as shown by low F3) but the pattern recognizer gets a syllabic dip.

constrained using the broad class label pronunciation models for the digits - 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Using the models trained on TIMIT, 2230 isolated digit utterances were tested using the constrained segmentation algorithm. The results are shown in Table 4.7 for EBS (with linear as well as RBF kernels) and for the HMM systems trained on TIMIT and tested on TIDIGITS. On moving from unconstrained to constrained segmentation, a similar improvement in performance of the EBS (RBF) and HMM-AP systems can be seen in this table. This result shows that EBS can be constrained in a successful manner like the HMM systems. The overall performance of EBS using RBFs is also very close to the HMM-AP system, and considerably better that the HMM system that uses the MFCC front-end. The improvement over the latter may solely be due to the better speaker independence of the APs as compared to the MFCCs (*Deshmukh et al.*, 2002). Note the relative performance of the HMM-AP and the HMM-MFCC systems.

Finally, word level accuracies were obtained for all the systems. A segmentation for a digit was scored as correct if it was an acceptable segmentation for that digit. A word-level correctness of 70% was obtained using the EBS-AP system and about 85% of the digits had a correct segmentation among the top 2 choices. The result is substantial since no information from the TIDIGITS database was used in training. A correctness of 72% was obtained by the HMM-AP system and a correctness of 63% was obtained by the HMM-MFCC system. These results further confirm the equivalent performance of EBS and HMM-AP system, and better performance of EBS over HMM-MFCC system.

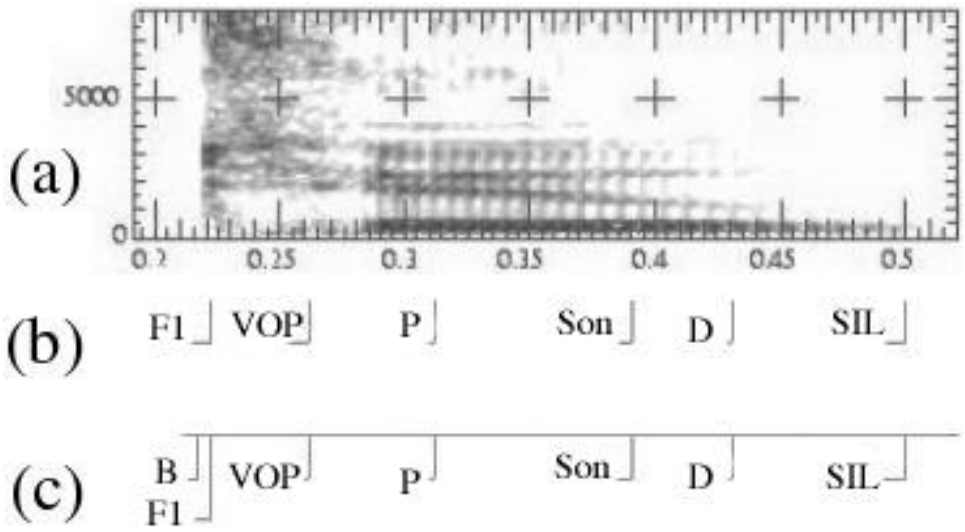


Figure 4.5: A sample output of the probabilistic landmark detection for the digit 'two'. Two most probable landmark sequences (a) and (b) are obtained by the probabilistic segmentation algorithm. The first most probable sequence (a) has a missed stop consonant but the second most probable sequence gets it.

4.3 Discussion

A system has been described for generating multiple landmark sequences of a speech utterance along with the posterior probability of each utterance. The landmark sequences can be constrained using broad class pronunciation models. For unconstrained segmentation on TIMIT, an accuracy of 79.5% is obtained assuming certain allowable splits, merges and substitutions that may not affect the final lexical access. The results assume a correct labeling of the phonemes although the TIMIT labeling has some incorrect labels. Higher performance indices and better trained models may be obtained if databases correctly labeled for landmarks are available. EBS performs significantly better with APs than with MFCCs because APs satisfy the assumptions of the probabilistic framework more closely. Moreover, the EBS-AP system shows a performance very similar to the HMM-AP system even though it uses the parameters selectively for each decision. On cross database constrained detection of landmarks, a correct segmentation was obtained for about 70% of the words. An incorrect most probable segmentation of a word does not show that the final word recognition will be wrong since the place probabilities may significantly affect the overall posterior word probabilities. But the overall performance can only be verified after complete implementation of the system.

The comparison with previous work on feature detection is very difficult because of the different test conditions and definitions of features used by different researchers. The 94.39% classification accuracy on the *sonorant* feature compares well with Bitar (*Bitar*, 1997) who obtained an accuracy of 94.6% for sonorancy de-

tection on all the 'si' sentences from the TIMIT database. The *continuant* result of 95.58% is not directly comparable with previously obtained stop detection results (*Bitar*, 1997; *Liu*, 1996; *Niyogi*, 1998) because this only shows the frame accuracy on binary classification with only stops and fricatives as the two competing classes. A 81.69% accuracy on the *syllabic* feature may seem low, but note that there is usually no sharp boundary between vowels and semivowels. Therefore, a very high accuracy at the frame level for this feature is not only very difficult to achieve, but also it is not very important as long as sonorant consonants are correctly spotted.

Chapter 5

Classification of features at landmarks

This chapter focuses on classification of place and voicing phonetic features and the manner phonetic features - *nasal* and *strident* at the acoustic landmarks. Knowledge based acoustic parameters are compared with MFCCs for the performance at the classification of the distinctions (1) *labial/alveolar* for stop consonants and (2) *anterior* for strident fricatives. Experiments were also carried out on conversational telephone speech in preparation for and the Johns Hopkins University CLSP summer workshop of 2004 and these are presented here as well. While experiments were conducted for classification of a large number of phonetic features, special attention was given to stop place and fricative place classifications and these are discusses in extra detail. A major reason for focusing on stop place and fricative place classification is that knowledge-based APs were available for these features. The APs are still under development for nasal place, fricative voicing and some other phonetic

features.

In general parameters were extracted from multiple frames centered at the landmarks to get reasonable accuracies. When parameters like MFCCs are extracted from multiple frames the dimension of the acoustic feature space becomes very high - sometimes comparable to or greater than the number of training samples - and SVMs are shown to not have an adverse effect of the increase in dimension, as expected from the theory. Experiments were conducted on three different databases

- TIMIT was used for experiments on 16kHz read speech
- NTIMIT was used for telephone bandwidth read speech experiments
- ICSI transcribed part of Switchboard database was used for conversational telephone speech (*Greenberg et al.*, 1996)

All the three databases had the phoneme labels although the ICSI labels had the stop consonants marked as one single segment instead of separate closures and releases. Unmarked stop releases made the experimentation difficult on switchboard data, therefore, using the phone labels along with the output of a stop burst detector, stop release labels were automatically generated. The original ICSI labels marked the stop consonants as one big segment starting at the closure and ending at the vowel onset in case there was a following vowel. The stop burst was hypothesized at the location of the maximum value of the probability of the stop burst obtained using the phonetic feature hierarchy and manner SVMs. Figure 5.1 shows that very accurate alignments were obtained for stop release labels. When there were two

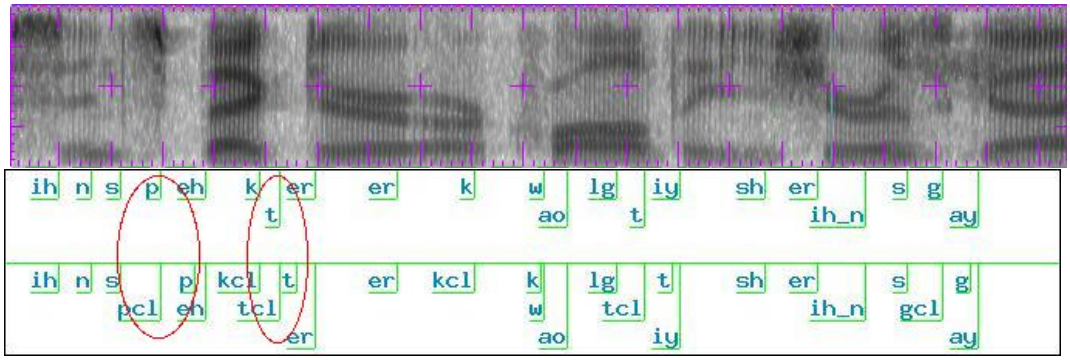


Figure 5.1: Top: spectrogram, Middle: phone labels from ICSI transcriptions, Bottom: realigned labels with stop releases marked. In the ellipse to the left, the segment /p/ is split into the closure /pcl/ and /p/ . In the ellipse to the right a sequence of /k/ and /t/ is split into the sequence /kcl/, /tcl/ and /t/ such that the release of /k/ is not marked. The figure shows that the stop release labels generated using the phone labels along with the outputs of the manner SVMs are very accurate.

consecutive stop consonants, the release of the first consonant was not marked, as shown by consecutive stops /k/ and /t/ in the figure.

5.1 Stop place classification

The problem of stop place of articulation classification has been addressed a number of times in the past by various researchers on data sets of different sizes. The goal in this section is not to invent new acoustic observations for stop place of articulation but it is to test various acoustic features for performance with SVMs on different data sets that are in general larger than the ones previous researchers have used.

For each SVM classification, an equal number of positive and negative samples are used so that results are not unnecessarily biased. Experiments were first conducted on TIMIT with SVMs trained on the 'si' sentences of the train set and tested on the 'si' sentences of the test set.

APs were first compared with MFCCs for performance on stop place of articulation classification. Two APs have been suggested for the distinction of labial and alveolar stop consonants - Ahi-A23 and Av-Ahi. Ahi captures the amplitude of the high frequency peak at the burst spectrum, Av is the low frequency peak of the vowel spectrum and A23 is the amplitude of the burst spectrum in the range of F3. Table 5.1 shows the results with these APs when the Av was computed at the vowel onset and Ahi and A23 were computed at the stop burst. In a different experimental setup, Ahi and A23 were computed across 5 frames starting at the stop burst and going toward the vowel. Similarly Av was computed at multiple frames starting at the vowel onset. In the third experiment, the energy ration parameter E[0,F3]/E[F3,SF/2] was added. Finally, formant measurements are added in the fourth experiment in each of the frames. Table 5.1 also shows the results when 13 MFCCs with and without their delta and acceleration coefficients replace the APs but the frames where the parameters are picked up are kept identical. The classification results show that APs perform considerably well and close to MFCC based classifier even though they are very small in number. The slight drop in results when formant estimates were added to the APs can be explained by the fact that the formant tracker used in the classification tasks is far from perfect. MFCCs, on the other hand, are implicitly modeling formant measurements by the distribution of

Acoustic parameters	Accuracy	Number of parameters per frame	Number of context frames
Ahi-A23, Ahi, Av	70.66	3	1
Ahi-A23, Ahi, Av	78.24	3	5
Ahi-A23, Ahi, Av, E[0,F3]/E[F3,SF/2]	81.34	4	5
Ahi-A23, Ahi, Av, E[0,F3]/E[F3,SF/2], F1, F2, F3	81.24	7	5
13 MFCCs	84.53	13	5
13 MFCCs + delta + acceleration	87.62	39	5

Table 5.1: Classification of *labial/alveolar* place of articulation on the TIMIT database. The number of context frames indicate the number of frames at both the stop burst and the vowel onset from where the APs mentioned in the first column. The total number of APs used in SVM classification is two (vowel onset and stop burst) times the number of parameters times the number of context frames.

energies in various frequency bands. Although MFCCs cannot model formant movements exactly, these parameters are measured consistently in all cases and there is no scope of "error" in their measurement.

In another experiment, it was tested whether computing Ahi and A23 at a resolution higher than the usual frame step of 5ms helps in stop place of articulation classification. In a separate classification experiment, values of Ahi and A23 computed at all 1ms frames that were being skipped when these were computed at the rate of 5ms. A drop in performance to 81.04% was observed indicating that the higher resolution of these acoustic observations may not be as necessary as it has been hypothesized (*Stevens et al.*, 1999).

There have been experiments (*Hasegawa-Johnson et al.*, 2005) where a large number of context frames starting at the stop burst frame were used instead of separately selecting frames from the stop burst and the vowel onset. To test if this helps, MFCCs were extracted from ten consecutive frames starting at the stop burst and the classification results were compared with the earlier case where these parameters were extracted from 5 context frames each at the stop burst and the vowel onset. A classification accuracy of 87.33% was obtained which is slightly and insignificantly lower than the accuracy obtained with with separate parameter extraction from the stop release and the vowel onset.

Acoustic parameters	Accuracy	Number of parameters per frame
Ahi-AF3, E[0,F3]/E[F3,SF/2], E[F3-187,F3+584], E[F3+1500,f_s/2]	83.91	4
Ahi-AF3, E[0,F3]/E[F3,SF/2], E[F3-187,F3+584], E[F3+1500,f_s/2]	84.78	4
13 MFCCs	91.96	13
13 MFCCs + delta + acceleration	92.17	39

Table 5.2: Classification of *anterior* place of articulation for strident fricatives. Four context frames were used in each classification. Two frames were picked from each of the fricative and the adjoining vowel. The two frames were picked at the distances of 5ms and 15 ms from the boundary in each of the vowel and the fricative.

5.2 Fricative place of articulation classification

Similar experiments comparing APs with MFCCs were conducted for the *anterior* place of articulation of strident fricatives. Table 5.2 shows the results on the TIMIT database with APs as well as MFCCs. Same pattern as with stop place of articulation was observed, that is, MFCCs perform somewhat better than the APs. But even though the number of APs used is very small, the performance is comparable. For this feature too, there is further scope of improvement in the design of the APs.

5.3 Classification of various features: results from JHU CLSP workshop 2004

In this section, the drop in performance of phonetic feature classification is studied when speech is filtered by the telephone channel. Performance is compared on TIMIT and NTIMIT on a number of features in tables 5.3 and 5.4 for pre-vocalic and post-vocalic contexts respectively. All classifications were conducted using 13 MFCCs and their delta and acceleration coefficients extracted from the landmark and the nearby frames listed in the last columns of these tables. Equal number of samples for each of the $+1$ and -1 classes were used in training as well testing. The classification accuracies in prevocalic contexts vary from about 79% to about 95% on the TIMIT database, and from 73% to 93% on the NTIMIT database. Even on the TIMIT databse, certain features require significant improvement in classification performance, for example, the feature *velar* for stop consonants and the feature *labial* for nasals. The drop in performance from TIMIT to NTIMIT is particularly significant when information above 4000Hz is important for classiifcation of a feature, for example, fricative *strident* classification. For classification of nasal place, the drop is insignificant since much of the information is contained in the movement of the formants and the spectrum of the nasal murmur. Numbers in postvocalic contexts are generally lower than those in prevocalic contexts perhaps because syllable codas are usually less stressed than syllable onsets.

Experiments were also conducted at WS04 to compare the relative effectiveness of using MFCCs with that of using the rate-scale representation motivated by the

Feature	NTIMIT	TIMIT	Landmark and context frames
Stop Voicing	81.09	85.93	Stop burst: [-5,-3,+1,+3,+5,+7], Vowel onset: [+1,+2,+3,+4,+5,+6]
Stop Velar	73.21	79.82	Stop burst: [0,2,4,6,8,10]
Stop Labial/Alveolar	76.30	87.11	Stop burst: [0,2,4,6,8,10
Fricative voicing	76.35	81.01	Release: [-3,-2,-1,0,1,2,3]
Fricative strident	82.30	88.31	Release: [-3,-2,-1,0,1,2,3]
Fricative anterior	84.48	83.37	Release: [-3,-2,-1,0,1,2,3]
Nasal	92.74	94.81	Release: [-3,0,3]
Nasal Labial	78.60	79.88	Release: [-3,-1,1,3,5,7,9]

Table 5.3: Results on NTIMIT and NTIMIT for various classifications at prevocalic landmarks

Feature	NTIMIT	TIMIT	Landmark and context frames
Stop Velar	67.53	72.12	Closure: [-7,-5,-3,-1]
Stop Labial/Alveolar	64.64	76.02	Closure: [-7,-5,-3,-1]
Fricative voicing	77.84	83.08	Closure: [-3,-2,-1,0,1,2,3]
Fricative strident	72.52	92.26	Closure: [-3,-2,-1,0,1,2,3]
Fricative anterior	83.19	86.94	Closure: [-3,-2,-1,0,1,2,3]
Nasal	95.74	97.78	[-3,0,3]
Nasal Labial	67.30	71.95	Closure: [-7,-5,-3,-1,1,3]
Nasal Alveo-lar/Velar	82.44	86.99	Closure: [-7,-5,-3,-1,1,3]

Table 5.4: Results on NTIMIT and NTIMIT for various classifications at postvocalic landmarks

auditory cortex (*Mesgarani et al.*, 2004). Identical number of context frames were used for the classification of each phonetic feature at the landmarks and the APs relevant for each task were appended to the parameters. APs have not been explicitly designed for telephone bandwidth speech and some of them use information above 4000Hz. This made the direct use of APs impossible in the form they were available before the workshop. A simple ad hoc change was carried out in the computation of APs to make them more suitable for the telephone bandwidth speech. A number of APs involve computation of energy in a frequency band starting at a certain frequency and ending at half of the sampling rate or above 4kHz. The computation of energies in these bands was forced to end at the frequency of 4kHz. This change represents a significant change in the APs and for certain classifications, for example, the feature *strident* for fricatives, it might have had an adverse effect in classification performance. The change is not optimal since there may be a frequency band available that may provide better classification performance.

Table 5.5 shows the accuracies obtained using SVMs on the test part of the NTIMIT database using either combination in both pre-vocalic and post-vocalic contexts. It can be seen that the performance of the two kinds of parameters - MFCC and rate-scale - is similar and no significant pattern can be noticed. APs perform better for some features while MFCCs perform better for other features. Rate-scale representation has been shown to be more robust to noise (*Mesgarani et al.*, 2004) and these results may provide a starting point for comparing noise robustness of the two kinds of parameters.

Table 5.6 shows the performance of some of the classifiers on the Switchboard

database and compares it with the NTIMIT database. There is drop in performance from read speech to conversational speech but it should be noted that a combination of NTIMIT and Switchboard databases was used for training. The reason to use a combination of data was that only a small part of Switchboard has been carefully transcribed at the phonetic level. It can be expected that when a large amount of phonetically transcribed Switchboard data is available, significant improvements in classification of features may be obtained.

5.4 Summary

Classification results for place and voicing features have been obtained on different databases using APs, MFCCs, rate-scale representation and combinations of these parameters. Binary classification accuracies range from 70% to 95%. An average absolute drop of about 5% is observed when switching from 16kHz studio speech to telephone speech. Further drop is observed when testing on conversational telephone speech. In spite of a lot of research that has been reported on detection of place of articulation of stop consonants, there is still a tremendous scope of improvement, especially on telephone bandwidth speech. The amount of phonetically labeled data for conversational telephone speech is small compared to the amount of data available for read speech. Further improvements in classifications will also require phonetic annotation of large amounts of conversational telephone speech.

Pre-vocalic contexts			Post-vocalic contexts		
Feature	MFCC+ APs	Rate-scale+ APs	Feature	MFCC+ APs	Rate-scale+ APs
Stop Voicing	83.15	85.26	Stop Velar	66.37	67.50
Stop Velar	72.55	82.20	StopAlveolar	62.95	63.30
Stop Alveolar	73.90	73.13	Stop Labial	65.00	73.05
Stop Labial	71.48	69.85	Fricative voicing	77.25	77.95
Fricative voicing	79.72	75.75	Fricative strident	78.50	73.20
Fricative strident	83.05	82.15	Fricative anterior	83.04	82.67
Fricative anterior	85.92	78.10	Fricative Labial	70.15	74.96
Fricative Labial	73.50	84.74	Nasal	88.83	87.45
Nasal	88.89	75.45	Nasal Labial	67.05	66.03
Nasal Labial	74.14	86.50	Nasal Alveolar	74.02	73.85
Nasal Alveolar	75.86	74.29	Nasal Velar	80.22	80.76
Nasal Velar	83.33	77.03	Lateral	76.65	75.55
Lateral	73.20	78.70	Rhotic	83.48	79.09
Rhotic	82.39	70.73	Round	80.48	83.97
Round	78.06	73.25	Palatal	91.04	90.30
Palatal	91.20	76.00			

Table 5.5: A comparison of MFCCs with rate-scale representation for classification of features at landmarks

Feature	NTIMIT	Switchboard	Context
Fricative anterior	82.67	75.71	Prevocalic
Nasal Labial	74.29	77.00	Prevocalic
Nasal Alveolar	77.03	72.00	Prevocalic
Nasal Velar	78.70	72.22	Prevocalic
Fricative Labial	-	77.50	Postvocalic
Fricative Labial	-	69.50	Prevocalic

Table 5.6: Comparison of results on read speech and conversational speech

Chapter 6

Word Recognition

The design of the isolated word recognizer that combines the landmark detection module with the place and voicing detectors is described in this chapter. The isolated word recognition data set used in this chapter has equal probabilities for all of the words, therefore, the priors were neglected. The landmark detection module described in Chapter 4 provides the probability of landmarks without the prior probabilities included. That is, it provides the following probability (denote it by $\bar{P}(B|O)$)

$$\bar{P}(B|O) = \prod_{i=1}^{M} P(D_i|B_i) \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i|x_t^{f_k^i}, f_1^i, ..., f_{k-1}^i)}{P_t(f_k^i|f_1^i, ..., f_{k-1}^i)}. \tag{6.1}$$

where B is the broad class sequence corresponding to the landmark sequence L and other variables have their usual meaning as in Chapter 4. To obtain a score of each word, the above probability of the landmark sequence was multiplied by the probability of the sequence of phonetic features given the landmarks. The probability of the phonetic feature sequence was computed without the priors as well (denote this

Table 6.1: Classification of place and voicing features on E-set utterances

Feature	Accuracy (APs)	Accuracy (MFCCs)
Stop voicing	98.08%	98.73
labial/alveolar for unvoiced stops	100%	96.16%
Fricative voicing	84.21%	88.16%
strident	83.75%	88.75%
anterior	87.50%	88.75%
labial/alveolar for voiced stops	88.46%	89.74%
aspiration/frication distinction	86.54%	94.23%

probability by $\bar{P}(U/OL)$

$$\bar{P}(U/OL) = \prod_{i=1}^{M} \prod_{k=N_{B_i}+1}^{N_i} \frac{P(x^{f_k^i}_{l_{i-1},l_i,l_{i+1}} | f_k^i, f^i_{N_{B_i}+1}, ..., f^i_{k-1}, L, u^{i-1})}{P(x^{f_k^i}_{l_{i-1},l_i,l_{i+1}} | f_{N_{B_i}+1}, ..., f^i_{k-1}, L, u^{i-1})} \tag{6.2}$$

The score obtained by multiplying the two probabilities above has also been applied to rescoring of lattices from an HMM based large vocabulary continuous speech recognizer as described in Section 6.2. In lattice rescoring experiments, the stream weight was provided by the language model and hence the score obtained by multiplying the above two expressions was used as an acoustic score.

6.1 E-set experiments

The probabilistic framework was first applied to recognition of eight E-set utterances - B, C, D, G, P, T, V, Z. This is a small but challenging task because a small transient region at the beginning of the utterance is where all of the word confusions lie. For

initial experiments, the SVMs were trained on the speakers F1-F4 and M1-M4 of the TI46 database and the models were tuned for good binary classification with the development set consisting of the speakers F5, F6, M5 and M6. Table 6.1 shows the binary classification results on these features on the development set using the APs described in Chapter 5. This table also shows the classification accuracies obtained using 39 MFCC coefficients including the delta and acceleration coefficients. The accuracies are considerably better than the corresponding accuracies obtained on the TIMIT database in Chapter 5 because all of the classifications in Table 6.1 were in the context of the vowel /iy/.

The APs for the feature *aspiration* were not available, therefore, for distinguishing the aspiration following the stop consonants from the frication noise in the sounds /z/, /s/ and /jh/, the APs for the feature *strident* were used. Exact APs have also not been developed for the feature *voiced* for fricatives, and the following measures were used for this feature (1) E[100,400], (2) E[0,F3]/E[F3,SF/2], (3) Pitch, (4) zero crossing rate and (5) zero crossing rate of high pass filtered signal. All of these acoustic measurements target the presence of periodic or sonorant energy in the signal.

Once manner and place classifiers are available, a number of system parameters can be varied in the implementation of the probabilistic framework. Some of these system parameters and their effect on the system performance is studied here:

- **Fixed or flexible manner class representation**

  Two sets of experiments were conducted with E-set recognition. In the first

experiment, the fixed manner class representation of each of the phonemes - /b/, /c/, /d/, /p/, /t/, /jh/, /v/, /z/ - was assumed except that /jh/ was allowed to have a stop burst with an unspecified place and appearance of aspiration after stop releases was kept optional. These settings gave a word recognition accuracy of 77.22%. A large number of the sounds /z/ and /s/ were recognized as /jh/ because many of the speakers pronuounce these sounds with a sharp onset. Therefore, in the second set of experiments, all of the fricatives - /z/ and /s/ - were allowed to have a stop burst with an unspecified place. This increased the word recognition accuracy to 78.48%. This is consistent with the results in the previous chapter where manner of the digits of TIDIGITS was shown to be highly variable. Both of these experiments were conducted with linear SVMs, no optimization of regularization parameter C, and using the histogram method of conversion of SVM discriminant to probabilities. It should be noted that for this small vocabulary and the small lengths of broad class sequences involved, such knowledge-based changes in broad class representations of sounds is easy but it becomes difficult in large vocabulary systems. To be more specific, when the vocabulary is large it is difficult to store a large number of different broad class representations, and a way must be found to predict flexible pronunciations in a generative manner similar to (*Livescu and Glass*, May 2004). An attempt to integrate EBS with the generative model in (*Livescu and Glass*, May 2004) is discussed in Section 6.3.1.

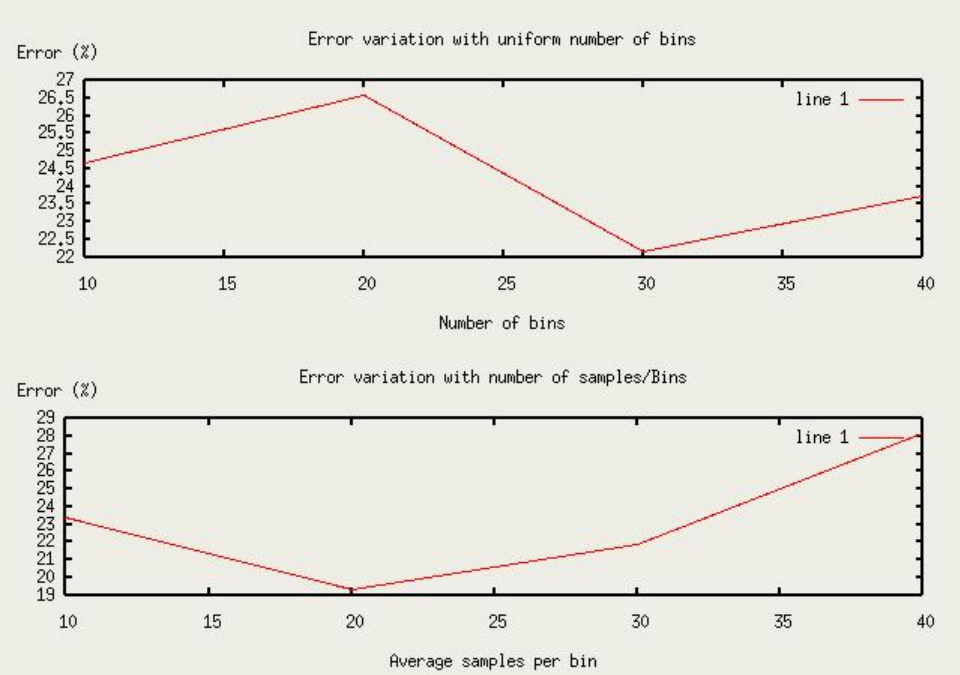


Figure 6.1: Variation of error with number of bins

- **The number of bins in histogram method**

  Four values of the number of bins in histogram method of conversion of SVM discriminant to probability were tested and the effect was studies on the word error rate. It was expected that the error would first drop with the increase in the number of bins and then it would rise. This is because for low number of bins the resolution and hence the accuracy of the mapping of discriminant to probability is low. For very high number of bins, the probabilities become erroneous again because the number of samples in each bin is not sufficiently high. But as observed in Figure 6.1(a) the behavior is more erratic possibly because the accuracy of bins vary largely with the number of training samples, and the number of training samples were very different for the manner classifiers as compared to the place classifiers. Especially, there were many more number

of frames available for *sonorant* classification than the number of landmarks available for stop place classification. In the second experiment, the number of bins for each classifier were varied as a factor of the number of training samples. That is, the number of bins were calculated as $N_{samples}/N_{samples/bins}$ where $N_{samples}$ is the number of training samples and $N_{samples/bins}$ is the expected number of samples in each bin. The variation of word error rate with $N_{samples/bins}$ is shown in Figure 6.1(b). The variation in word error rate is as per expectation in this case, that is, it first drops and then rises.

- **The choice of probability conversion method**

  The histogram method was compared to the mapping of SVM discriminant to probability using a sigmoid function. The function tried in this experiment was $f(g(x)) = 1/(1 + exp(-g(x)))$ where $g(x)$ is the SVM discriminant. A word recognition accuracy of 69.93% was obtained which is sign ificantly inferior to the accuracy of 80.69% obtained using the best histogram settings on the development data.

- **Optimization of the regularization parameter C**

  Now using the optimal parameter settings of the histogram method along with the linear SVMs, the effect of the optimal choice of C was studied on the word error rate. The development data was split into two parts - one with speakers M5 and F5 and other with speakers M6 and F6. The optimal value of C for each SVM was chosen such that the binary classification accuracy was minimum on the first set. The optimal values were then used for word

Manner	Place	Word accuracy
Linear	Linear	80.69
RBF	RBF	63.29
Linear	RBF	75.63
RBF	Linear	77.85

Table 6.2: Effect of SVM kernel on word accuracy

recognition on the speakers M6 and F6. Accuracies of 77.98% and 83.02% were obtained with and without C optimization respectively. The reverse was expected but this behaviour may be because of the hold-out cross validation method used here. That is, C was optimized to minimize error on a data set separately held out from the training data. C optimized using leave-one-out cross validation may provide a better value of C. But in the rest of the experiments, C was not optimized and the default value provided by the SVM Light toolkit was used.

- **Choice of SVM kernel**

  Four sets of experiments were conducted to study the effect of the SVM kernel on word recognition - by combination of linear and RBF kernels for manner detection and landmark classification. The results are shown in Table 6.2 which compares the four combinations. The combination that has both kinds of classifiers as linear gives the lowest error rate on the development set, therefore, all linear classifiers are used in the rest of the experiments.

Table 6.3: Word recognition performance on E-set development set using TIMIT trained models

	EBS	HMM
APs	80.69	80.12
MFCCs	75.64%	88.23%

6.1.1 HMM-based system

An HMM system was built to recognize the E-set and the performance of EBS with the system. Context independent monophone models similar to (*Deshmukh et al.*, 2002) were built for all of the consonants - /b/, /d/, /g/, /t/, /p/, /v/, /z/, /c/ - as well as the vowel /iy/ and the closures - /bcl/ and /dcl/ - and the closures for unvoiced stop consonants - /pcl/ and /tcl/. These models are significantly different from the standard monophone models where there is a single HMM for a stop consonant. Separate models were built here for the closures and releases of the stop consonants to model the detailed dynamic acoustic manifestation of stop consonants. All models were three-state 8-mixture Gaussian density models with one skip transition from the first state to the third state. Models were first trained using the segment boundaries specified by the phonetic labels and then embedded re-estimation was conducted without using the hand-transcribed time boundaries. Figure 6.2 shows the variation of error with the number of re-estimation steps with '0' steps referring to the case where the models were trained only using the manual phonetic transcriptions. It can be seen that embedded re-estimation increases the recognition error substantially. It's hard to say whether this result shows that

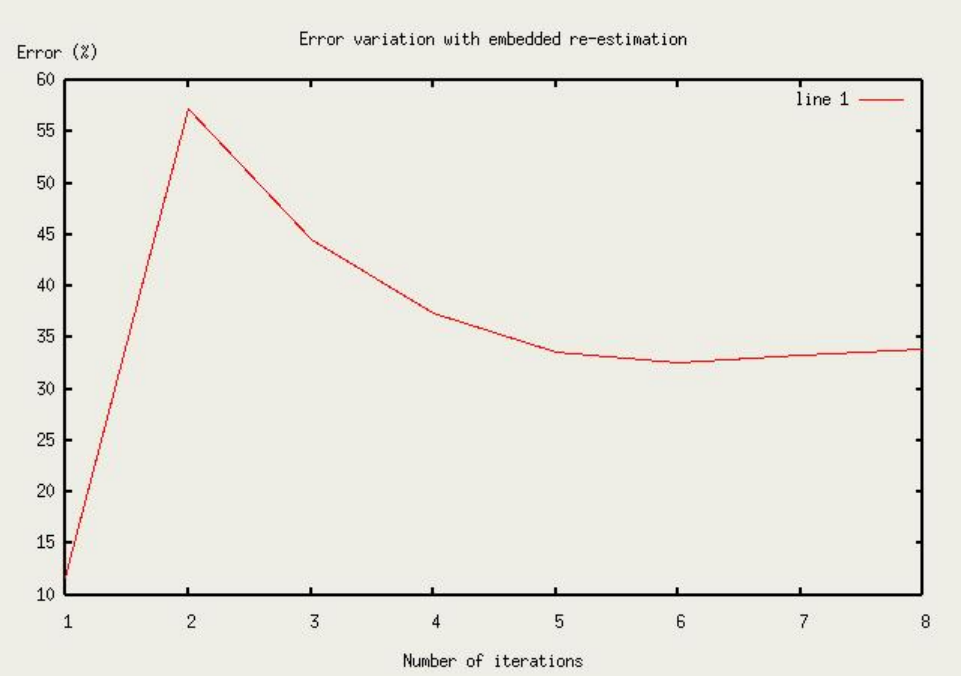


Figure 6.2: Variation of error with re-estimation iterations

the lack of time aligned phonetic labels in HMM based speech recognition hurts the performance seriously in general. This result is in agreement with (*Hosom*, 2000) where it was shown that for connected alpha-digit task the knowledge of accurate phonetic labels can significantly improve performance. For the final test data evaluation the models trained with no re-estimation were used.

The word recognition results on the development data are shown in Table 6.3. Similar pattern as in landmark detection is observed, that is, the HMM system with MFCCs performs better than the HMM system with APs. EBS performs better with the APs than the MFCCs. Overall HMM-MFCC system gives the best performance that is considerably better than all other systems.

Table 6.4: Word recognition performance on E-set test set

	EBS	HMM
APs	84.64	77.14
MFCCs	80.29	90.93%

6.1.2 Test data results

The trained models were finally applied to the test data composed of the E-set utterances from the TI46 speakers that were not used in either training or development, that is, F7-F8 and M7-M8. The word error rate of isolated word recognition is shown in Table 6.4. The HMM system using 39 MFCCs including the delta and acceleration coefficients gives the best performance on word recognition, followed by EBS using the knowledge-based APs. The EBS-AP system again performs significantly better than the EBS-MFCC system showing that EBS is able to utilize the interesting properties that APs possess and MFCCs do not. This is more apparent from the fact that the EBS-MFCC system does not give better performance than the EBS-AP system even though MFCCs give better classifications accuracies on phonetic features (see Table 6.1). Table 6.5 shows the confusion matrix of the E-set utterances from the results generated by EBS. It is easy to observe from these confusions that most of the errors - confusions between B and V, and G and T - are due to unreliable classification of the feature *aspiration*. Because the APs for the feature *aspiration* have not been developed, there is no reliable way of distinguishing the frication in the sounds /v/ and /jh/ from the aspiration that may follow /b/ and /t/. Development of APs for this feature may lead to significant gains in

Table 6.5: Confusion matrix of the E-set test data

	B	S	D	G	P	T	V	Z
B	37	0	3	0	0	0	0	0
S	0	38	0	1	0	0	1	0
D	1	0	38	0	0	1	0	0
G	0	0	0	29	0	11	0	0
P	0	0	1	0	33	5	0	0
T	0	0	0	2	0	38	0	0
V	11	3	2	1	0	0	21	2
Z	0	5	0	0	0	0	0	35

performance.

6.2 Rescoring of switchboard lattices

The experiments in the previous sections were limited to read speech with very small vocabulary. At the CLSP workshop of 2004, EBS was applied to rescore the lattice output of the SRI speech recognizer (*Stolcke et al.*, 2003). Lattices were available for the RT03 development and evaluation data (*NIST*). Each branch in the lattice consisted of a word, its phone-level representation, the acoustic score and the language scores of the word from the SRI recognizer. The task of the rescoring process was to provide a score from EBS to supplement the scores already in the lattice. An algorithm was then used to find optimal stream weights for each of the scores in the lattice including the EBS scores (*Hasegawa-Johnson et al.*, 2005) such that the

WER was minimized. A fixed phonetic feature bundle representation of each phone was used with the mapping of the manner features given in Chapter 4, and the place features as listed in Appendix B were used. The lattice had different branches for different pronunciations of words and these were scored separately using EBS so that some variation in pronunciation was taken into account while rescoring. The SVMs trained on NTIMIT (listed in Chapter 5) were used for the phonetic feature classifications. The duration statistics were recomputed using the ICSI transcribed part of the switchboard database before application to the probabilistic segmentation algorithm.

Figure 6.3 shows the broad class output of EBS forcedalignment on the multi-word sequence "i_think_it". The labels are very well aligned and there is only one problem in the broad class outputs. Since a fixed mapping of phones to manner and place features was assumed, the /t/ in the word final position was forced to have a separate burst and a closure. A lack of pronunciation variability rules made EBS produce forced alignments that can significantly affect the likelihoods generated by EBS for many of the words. A stream weight of 10^{-5} was assigned to EBS that was negligible compared to the weights of approximately 1 and 8 of the acoustic model and the language model respectively. This stream weight did not lead to any drop in the word error rate.

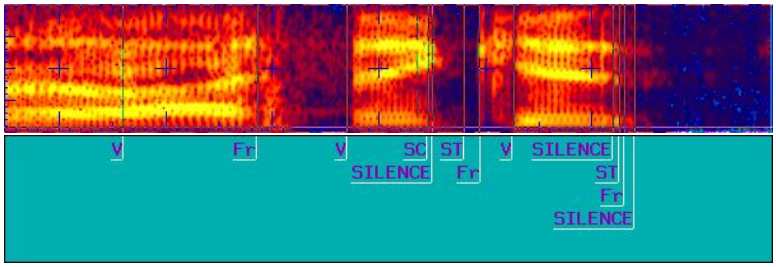


Figure 6.3: A example of a landmark forced alignment by EBS on RT03 development data on the utterance "i_think_it"

6.3 Application to discriminative lattice rescoring

A system was designed and implemented by Kirchoff (*Hasegawa-Johnson et al.*, 2005) for reducing substitution errors in the lattices using phonetic feature classifiers for selecting among confusable words. In this method, the most common word confusions were identified and each confusion was converted to a binary relation either between two broad classes of sounds or between a broad class and a place feature. The task of EBS was then to carry out constrained detection to give out probabilities of each of the features. Figure 6.4 shows how EBS was constrained to compute the probabilities of the feature pair {vowel, +low}. The beginning and the final states were allowed to take any broad class and their probabilities were ignored, and the probabilities were picked only from the relevant states in the middle. The probabilities thus generated were used in a maximum entropy classifier by Kirchoff to rescore the lattices. On the RT03 development data, a statistically insignificant reduction in word error rate of less than 0.05% or about 14 words was acheived.

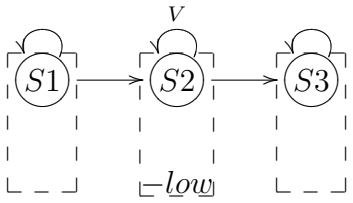


Figure 6.4: A FSA for computation of probabilities of a pair of features

6.3.1 Combination with a generative pronunciation model

The lack of availability of a phonetic feature based pronunciation model and the assumption of canonical pronunciations was largely responsible for poor performance of EBS in lattice rescoring. A Dynamic Bayesian Network (DBN) based generative pronunciation model (*Livescu and Glass*, May 2004) that models pronunciation variability by allowing overlapping of articulatory features was applied by Livescu (*Hasegawa-Johnson et al.*, 2005). EBS outputs were used in two different ways by Livescu at the workshop:

1. Manner segmentations were generated by EBS using the probabilistic segmentation algorithm. The probabilities of the manner phonetic features *sonorant*, *continuant*, *syllabic* and silence were provided to the DBN in each frame of speech and the probabilities of the place and voicing features were provided at the landmarks.

2. In this system, the probabilities of all phonetic features, whether place, manner or voicing - were provided to the DBN in each speech frame and the task of locating the landmarks and using the place probabilities at the appropriate landmarks was left to the DBN.

Excellent alignment of the articulators was obtained by in this work in both the cases and a drop in WER from 27.9% to 27.2% was reported on a subset of the RT03 development set consisting of three speakers.

6.4 Summary

Experiments on word recognition using the probabilistic framework developed in Chapter 3 have been presented. It has been shown that EBS performs better than the HMM system using APs but worse using MFCCs. EBS is better able to utilize the properties of the APs than the HMM system, and that was the motivation in the design of the probabilistic framework. The system has been applied to lattice rescoring over the ouptut of an HMM based large vocabulary recognizer. No reduction in word error rate was observed when EBS was directly applied to rescoring of lattices. Some positive trend in recognition performance was observed by other researchers when they applied the output of EBS with their pronunciation models (*Hasegawa-Johnson et al.*, 2005).

Chapter 7

Conclusions

An acoustic-phonetic speech recognition system has been developed with various exciting properties. To the best of our knowledge, this is the first statistical recognizer which uses only relevant knowledge-based acoustic observations at relevant locations in time. Moreover, the system provides a mathematical framework for understanding context-invariance of acoustic parameters. For place phonetic features, invariance is assumed with variation of the place of neighboring sounds. For example, the acoustic cues of stop place are assumed to be independent of the place features of the following vowel. For manner phonetic features, invariance of acoustic cues for a feature in a particular frame is considered with respect to the variation of manner features below that feature in the phonetic feature hierarchy. The probabilistic framework formalizes the need for the search of high accuracy context invariant acoustic parameters that acoustic phonetic researchers have tried to find over many years. Some of the knowledge-based APs have been shown to approximately satisfy the invariance property required by the probabilistic framework. Especially it has

been shown that APs satisfy the invariance property significantly better the the mel-frequency cepstral coefficients.

Performance very close to HMM based systems has been achieved on segmentation of continuous speech and detection of acoustic landmarks. A number of errors in the landmark detection system are due to the reductions and the coarticulations that usually occur in continuous speech. For example, the sound /r/ may merge completely with the adjacent vowel to cause an /r/-colored vowel that may not show a dip in energy in the sonorant region of the syllable under investigation. Similarly word initial fricatives are often released with a sudden burst that is classified as a stop burst by the landmark detection system, but that is counted as an error by the scoring program. This variation in manner with context and speaking rate or style, that leads to reductions of coarticulation, poses a significant challenge for landmark-based speech recognition. For example, if a stop burst is detected along with frication noise following it, a separate module is required to check whether the stop release is lexically distinctive or whether it was produced due to a sudden release of a fricative. High level information can be very useful in disambiguation of such cases. For example, if it is known that the burst is at the beginning of the word, then it is likely that the stop burst is not lexically distinctive. But if the stop burst-frication noise pair occurs in the middle of a word, the burst and the noise are parts of separate sounds and the stop burst is distinctive. This indicates that a significant amount of work is required in integrating high level information in landmark-based speech recognition.

Classifiers have been developed for a number of phonetic features and their ac-

curacies have been tested on read speech as well as conversational telephone speech. Reasonable accuracies have been obtained on classifications of most of the phonetic features, but the scope of improvement is tremendous in both the knowledge-based design of acoustic parameters as well as the performance of the statistical classifiers. The probabilistic framework for word recognition system has not performed as well as the HMM based system. The results on word recognition are consistent with how well the acoustic features satisfy the invariance assumptions of the probabilistic framework. APs perform better than MFCCs with the EBS probabilistic framework because they satisfy the invariance assumption better. On the other hand, MFCCs perform better than the APs in the HMM framework because they satisfy the property of lack of correlation across feature dimensions.

The system has been applied to telephone speech for lattice rescoring and good alignment of landmarks has been obtained. Because of the assumption of canonical pronunciations, an insignificant stream weight was assigned to EBS in lattice rescoring on the RT03 development set. Significant improvement may be expected with an appropriate pronunciation model.

7.1 Suggestions for future work

There is a huge scope of improvement that provides a lot of opportunity for further research in all aspects of landmark-based speech recognition. Many of these ideas are listed below.

1. **Better acoustic observations**

   Most of the phonetic feature classifiers need a lot of improvement before landmark-based speech recognizers can be applied to practical speech recognition tasks. A lot of this improvement can arise from design of better knowledge based acoustic parameters. Significant progress was made in phonetic feature classification at JHU 2004 summer workshop (*Hasegawa-Johnson et al.*, 2005) but most of this improvement was achieved by combination of a large number of diverse acoustic observations. While improvements have been reported using that method, such a large number of acoustic observations are not likely to be invariant of context, especially since the same measurements were used for classifications of most of the place phonetic features.

2. **Manner independent cues for place recognition**

   As it has been observed, manner can be highly variable from speaker to speaker for the same speech sound. For example, a stop release may be weak enough to look only like an aspiration segment. What may really distinguish it from the sound /h/, for example, is the strong movement of formants. Therefore, whether or not a sudden energy burst is observed in the speech signal, a way should be determined to find stop consonants, for example, by directly using the formant movements. Those formant movements may then be directly used to find the place of the stop consonants.

3. **Pronunciation modeling**

   Significant advances are required in pronunciation modeling for landmark

based speech recognition. Conventional approach to handle pronunciation variation is to store many possible phone-based pronunciations for each word. This approach lacks systematic knowledge of how pronunciations can be varied and therefore, cannot predict unseen pronunciations. A generative model has been developed (*Livescu and Glass*, May 2004) that has the capability of predicting many different pronunciations on the basis of an overlapping articulatory feature model. While this model was designed to handle frame-based observations, it has also been fused with the landmark-based approach so that it used only relevant observations at each landmark. Promising results were reported, but further research is required in combining the landmark-based approach with such generative pronunciation models to build a stand-alone speech recognition system. Other possibility in pronunciation modeling is to incorporate all the feature-based pronunciation rules (*Zhang*, 1998), but this approach may encounter the same obstacle as the one that stores many different pronunciations. This approach may still be worthy of further investigation because it fits really well with the landmark-based approach.

4. **Better probabilistic modeling**

   Support vector machines were not designed to be Bayesian classifiers even though we have used them in a Bayesian framework by converting the SVM discriminant to a posterior probability. To the best of our understanding, this computation of posterior using a histogram method is not well studied. Better methods of converting SVM outputs to probabilities are available (*Kwok*,

2000) that may be more compatible with the Bayesian probabilistic framework presented in this work. Also, the Bayesian framework may be avoided entirely and a new statistical framework may be designed entirely on the idea of the VC dimension as shown by the label sequence training method for force-aligning labels (*Altun and Hofmann*, 2003).

5. **Relative significance of each phonetic feature**

   The probabilistic framework developed in this work gives equal weight or importance to each phonetic feature. Different phonetic features may have different weights in their contribution to speech understanding. It was demonstrated by Kirchoff (*Hasegawa-Johnson et al.*, 2005) in a lattice rescoring framework that certain phonetic distinctions may be more significant for removing confusions in the output of an HMM based speech recognizer as compared to other phonetic features. While this was tested in a rescoring framework instead of a direct decoding framework, methods may be devised to put different weights on different phonetic features in a direct landmark-based recognition method. The significance of different features may change considerably with the environment, for example, in noise, and methods may be found to adapt the weights to changing environment.

Appendix A

Tables of place and voicing features

Feature	Articulatory correlate	v	f	dh	th	z	zh	s	sh
voiced	Vocal vold vibration	+	-	+	-	+	+	-	-
strident	Airstream from the constriction hits an obstacle	-	-	-	-	+	+	+	+
alveolar	Tongue tip against alveolar ridge	-	-	+	+	+	-	+	-
labial	Constriction at lips	+	+	-	-	-	-	-	-

Table A.1: The features *strident*, *voiced* and the place features for fricative consonants

Feature	Articulatory correlate	w	r	l	y	n	m	ng
nasal	Closed oral cavity, flow through nasal cavity	-	-	-	-	+	+	+
labial	Constriction at lips					-	+	-
alveolar	Tongue tip against alveolar ridge					+	-	-
rhotic	Curled up tongue	-	+	-	-			
lateral	Lateral airflow around one or both sides of tongue	-	-	+	-			
round	Lip rounding	+	-	-	-			

Table A.2: The place and manner features for sonorant consonants

Feature	Articulatory correlate	iy	ih	ey	eh	ae	aa	ao	ow	ah	uw	uh
back	Tongue positioned towards back of mouth	-	-	-	-	-	+	+	+	+	+	+
low	Low tongue position	-	-	-	-	+	+	+	-	-	+	+
high	High tongue position	+	+	-	-	-	-	-	-	-	-	-
tense	Tense articulators	+	-	+	-	-			+	-	+	-
round	Lip rounding	-	-	-	-	-	-	+	+	-	+	+

Table A.3: The place features for vowels

Appendix B

User manual of the toolkit of landmark-based speech recognition

The following is a manual of a part of the landmark based speech recognition toolkit written to implement and test the ideas presented in this thesis. The complete manual can be found at http://www.ece.umd.edu/~juneja/apfactmanual.pdf . This manual does not include help on the part of the code for word recognition but the online version will eventually contain that help. The following manual is for the part of the code that can be used for binary classification experiments.

B.1 Synopsis

System Requirements:

A. SVM Light must be installed on the system B. Phoneme label files in TIMIT format must be availabe C. Frame-by-frame computed acoustic features in binary format (explained below) or HTK format D. Python 2.2 E. *nix (Unix, Linux, etc.)

. It may run on Windows but I never tested it.

1. train_config.py

   Usage: train_config.py <Config File>

   This is the main executable for phonetic feature classification. It can (a) create files for use with MATLAB, SVM Light and LIBSVM by picking up acoustic parameters either by frame-by-frame basis or on the basis of landmarks, (b) train SVM classifiers (available only for SVM Light, and LIBSVM has to be run separately) while optimizing the kernel parameter and the penalty (bound on alphas) with different methods - minimum XiAlpha estimate of error, minimum number of support vectors, minimum cross-validation error, (c) do SVM classification on test files created by the code in a separate pass, (d) create histograms. SVMs for multiple phonetic features can be trained and tested at the same time. Please read the help in README.config for formatting the config file because this is the most crucial step.

2. print_landmarks.py

   Usage: print_landmarks.py <Config File>

   This will use the same config file as needed by train_config.py . It will create a landmark label file for each utterance in a list of utterances provided in the

config file. The landmarks can be generated in one of the two ways: (a) using knowledge based acoustic measurments (b) using only the phoneme labels.

3. collate_aps.py

   Usage: collate_aps.py

   Combines two streams of acoustic parameters, for example, one stream of MFCCs and one stream of knowledge based acoustic measurements, by choosing only specified set of measrements from both the streams. It can also compute and append delta and acceleration coefficients for the selected measurements from both the streams. Binary and HTK format for both input and output are accepted. To create output files in HTK format, ESPS must be installed on the system, especially, the 'btosps' and 'featohtk' commands must be available. To customize the command opent the file collate_aps.py and follow the instructions.

4. phn2lab.py

   Usage: phn2lab.py <phn file> <lab file>

   Converts phn labels to ESPS format labels that can be displayed in xwaves.

5. batch_phn2lab.py

Usage: batch_phn2lab.py <phn file list> Converts label files in .phn format to ESPS .lab format given an input list of .phn files. It assumes that the input files have 3 character extension.

6. findScalingParameters.py

   findScalingParameters.py <Config File>

   Uses the same config file as in train_config.py to compute the scaling parameters for all of the acoustic measurements. This script must be run before running the train_config.py if scaled parameters are to be used.

7. File formats

   Binary: This is plain binary format. Acoustic parameters are written frame-by-frame with each parameter in 'float'. For example, if there are 500 frames and 39 parameter per frame, then 39 parameters for the first frame are written first, followed by the 39 parameters of the second frame, and so on. Note (1) each parameter is written in float (2) as far as this toolkit is concerned, linux and unix generated acoustic parameter files in binary format are not cross-compatible on these systems because the two systems use a different byte order.

B.2 Configuration files parameters

A number of values can be set in a config file that goes as input to the executables train_config.py . These are discusses here. Three examples of a config file are config_broadclass_hie.py, config_mfc_hie.py and context_config.py provided along with the scripts. The config variables are set in python format which has a very easy and obvious syntax. The code can be used for frame-based and landmark-based training and testing. Many experiments can be carried out by both frame-based and landmark based methods. Landmarks are computed by the sytsem automatically for each phoneme by first converting a phoneme into a broad class label and then finding a set of landmarks for each broad class. The following landmarks are computed : Vowel (V) : [Vowel onset point (VOP), Peak] Sonorant consonant (SC - nasal or semivowel) : For postvocalic case, [Syllabic peak of previous vowel, SC onset, syllabic dip which is the mid point of the SC segment in this case], For prevocalic case, [syllabic dip which is the mid point of the SC segment in this case, SC offset (vowel onset), syllabic peak of the following vowel]. Intervocalic case: [Syllabic peak of previous vowel, SC onset, syllabic dip which is the mid point of the SC segment in this case, SC offset (vowel onset), syllabic peak of the following vowel] Stop (ST) : [Burst, Release] Fricative: [start frame, 1/4 frame, middle frame, 3/4 frame, end frame] Silence: [Silence start, silence end] The silence landmarks are useful for classification of the stop place features in postvocalic contexts.

The landmarks shown above for each broad class must be noted because this knowledge is essential for doing landmark-based experiments. In landmark based

experiments, you need to specify where acoustic parameters are to be picked at. For example, if acoustic parameters 1,23,27 (this numbering is for the order in which the parameters are stored in parameter files starting with 1) are to be picked at Peak of the vowel, then the value of the Parameters variable below for such a class has to be set as [[], [1, 23, 27]] such that nothing is picked at the vowel onset point. In addition if a number of adjoining frames is to be used at Peak landmark then the value of Adjoins is set as [[], [-4, -2, 0, 2, 4]] and then the parameters [1, 23, 27] will be picked from (Peak - 4)th frame, (Peak - 2)nd frame and so on. For a particular classification, the current version of the code has a constraint that if the number of parameters at a landmark for a broad class are non-zero: then the number of parameters and the number of adjoins for that landmark must be the same as other non-zero ones. For example, if some parameters have to be picked from the VOP, then it should also have three parameters (considering above example) computed using the adjoins of size five, for example [-4, -1, 0, 1, 4]. Of course, the parameters and the adjoins may be different.

A single config file can be used for a number of SVM classification experiments. In the config file you specify a list of SVM Light formatted data files, a list of model files names, indices of parameters to be extracted for each classification, etc. The i'th element of each of these lists determine how the i'th experiment is done.

1. Flags and values related to kinds of tasks and various inputs (labels and acoustic parameters)

   outputDir

The full path of the directory containing the acoustic parameter files. A misnomer because this directory is more of an input.

labelsDir

The full path of the directory containing the label files in TIMIT format.

modelDir

The outout directory where model files and SVM Light formatted data files will be written.

filelist

Full path of a list of acoustic parameter files.

shuffleFilesFlag

If this is set to 1, the list of files will be shuffled before use

apFileExtLen

This an integer telling the length of extension of each acoustic parameter file. The code takes off this many number of characters and appends the label extension (refLabelExtension) to find the label file in the directory labelsDir .

refLabelExtension

The extension of the label file, for example, 'phn'

SkipDataCreationFlag

If this flag is test to 1, then no SVM formatted data files are created. This is used to only run the SVM Light, for example, to optimize the value of gamma or C .

SkipModelTrainingFlag

Setting this to 1 will skip model training. This can be used to (1) only create

the SVM Light formatted data files so as to test with other toolkits such as LIBSVM of MATLAB externally, (2) create SVM Light formatted data files that can be used as validation files for SVM training in a separate pass.

SkipBinningFlag

Setting this to 1 will skip creation of bins for probabilistic modeling of SVM outputs. This not relevant for this version of teh code.

binaryClassificationFlag

If this flag is set to 1, SVMs will be run on the files in the array SvmInputFilesDevel

classificationType = 2

1: Non-Hierarchical 2: Hierarchical . Please ignore this flag in this version of the toolkit. It is only relevant in the full-version

nBroadClasses

Please ignore this value in this version of the toolkit. It is only relevant in the full-version. Give it any value but do include it in the config file.

nBroadClassifiers = 4 # Not relevant for classification

Please ignore this value in this version of the toolkit. It is only relevant in the full-version. Give it any value but do include it in the config file.

nClasses

The number of SVMs . Not required but it can ease writing of certain variables in the config file that are same across all the SVMs to be trained. For example in python, a=['z']*5 will assign ['z', 'z', 'z', 'z', 'z'] to a .

selectiveTraining

The code allows for carrying out the designated tasks on a specified set of features instead of all the features. Even if config file is written for 20 SVMs (features), you can specify which features to analyze. For example, selectiveTraining = [0,3,5,6]

apDataFormat

0: binary, 1: HTK .

2. Values related to the names of SVM Light format files and model files to be created

SvmInputFiles

The names of SVM Light formatted files to be created. For example, SvmInputFiles = ['LightSonor', 'LightStops', 'LightSC', 'LightSilence']

SvmInputFilesDevel

The names of files used for validation. When optimizing a kernel related parameter, these files will be used to minimize the error on. For example, SvmInputFilesDevel = ['LightSonorDevel', 'LightStopsDevel', 'LightSCDevel', 'LightSilenceDevel']

modelFiles

The names of models. For example, modelFiles = ['rbf_model_sonor', 'rbf_model_stop', 'rbf_model_sc', 'rbf_model_sil']

3. Values and flags related to the parameters used in each classification

Parameters

The list of parameters to be used for each classification. For example, [[1, 2, 15, 16, 19], [4, 5, 17, 18], [8, 13, 14, 15, 16], [9, 4, 5, 6, 7]] where each list is a list of parameter for the corresponding index of model file, SVM data file, etc. These examples are good only for frame-based training. For landmark based testing, parameters are specified for each landmark as exemplified in the synopsis above. More examples can be found in the config_mfc_hie.py (example file) file provided with the toolkit.

Doublets = [[]]*nClasses

Not tested in a while and better not to use. Assign Doublets = [[]]*nClasses to have the code ignore it.

Adjoins

The number of adjoining frames along with the current frame to be used for classification. For example, [[-4, -3, -2, -1, 0, 1], [-4, -3, -2, -1, 0, 1, 2, 3, 4], [-16, -12, -8, -4, 0, 4, 8, 12, 16, 20, 24], [-3, -2, -1, 0, 1, 2]]. For landmark-based training, adjoins have to be specified for each landmark as stated in the synopsis above.

numberOfParameters

The number of parameters per frame in each acoustic parameter file.

stepSize

The step size of the frames in milliseconds. Required for reading the labels.

classes_1

The +1 class members (phonemes/broad classes) from which the parameters are to be extracted. For example, classes_1 = [['V', 'SC', 'N'],['ST', 'VST'], ['n', 'm', 'y', 'w', 'r', 'l', 'ng'], ['start-end', 'VB', 'epi', 'CL']]. See the file labels.py for the mapping used for phonemes to broad classes.

classes_2

The -1 class members (either phonemes or broad classes but not both in any classification) from which the parameters are to be extracted. For example, classes_2 = [['V', 'SC', 'N'], ['ST', 'VST'], ['n', 'm', 'y', 'w', 'r', 'l', 'ng'], ['start-end', 'VB', 'epi', 'CL']] . See the file labels.py for the mapping used for phonemes to broad classes.

useDurationFlag

A flag for each classification, for example, [0, 0, 0, 0]. A flag can take a value 1 only when the corresponding parameterExtractionStyles flag is set to 7 (landmark based training) .

specificDataFlags

If broad classes are used in classes_1 and classes_2 for any of the classification, set it to 0 otherwise set it to 1, for that classification.

parameterExtractionStyles

0: Frame based training, 1: IGNORE, not tested in a while, 7: landmark-

based testing .

useDataBound

Setting this flag to 1 will use an upper bound on the number of samples extracted for each classification . The number is set by the values maxclass1 and maxclass2 explained below .

placeVoicingSpecifications

This selects the kind of landmark training for each classifier for which landmark training is chosen. For vowels the options are 'generic' (all vowels will be used), 'preSConly' (vowels with no following sonorant consonant will be used and postSConly (vowels with no preceding vowels will be used). For fricatives, the options are 'generic' (all fricatives), 'genericPreVocalic' (fricatives before vowels and sonorant consonants), 'genericPostVocalic' (fricatived after vowels or sonorant consonants), 'genericIsolated' (fricatives with no adjoining sonorants). For sonorant consonants, the options are 'genericInterVocalicSC' (as the name suggests - note that there are five landmarks in this case), 'genericPreVocalicSC' (three landmarks) , 'genericPostVocalicSC' (three landmarks). For stops, the only valid option is 'genericPreVocalic'. The variable placeVoicingSpecifications will be removed in the forthcoming versions of the code and the framework will allow the user to specify any context.

init1

For frame-based training this is the list of numbers of initial frames to be extracted for each classifier. If for any classifier this value is set to non-zero, then only that number of initial frames will be used from classes_1 . The mid-

dleFlag1 will be ignored. For example, init1 = [0, 1, 0, 0] # Only relevant for frame-based training

init2

For frame-based training this is the list of numbers of initial frames to be extracted for each classifier. If for any classifier this value is set to non-zero, then only that number of initial frames will be used from classes_2 . The middleFlag2 will be ignored. For example, init2 = [0, 1, 0, 0] # Only relevant for frame-based training

delstart1

Delete an initial number of frames when picking frames for frame-based training from a label in classes_1. For example, delstart1 = [0, 0, 0, 0] . Only relevant for frame-based training. Ignored if a corresponding init1 value is set to non-zero .

delstart2

Delete an initial number of frames when picking frames for frame-based training from a label in classes_2. For example, delstart2 = [0, 0, 0, 0] . Only relevant for frame-based training. Ignored if a corresponding init2 value is set to non-zero.

delend1

Similar to delstart1 but for end frames.

delend2

Similar to delstart2 but for end frames.

contextFlag1

Specify the left and right context of eaach of the labels in classes_1. Only the phonemes/broad classes with the specified context will be used. If the ith element of the list contains 'left' or 'right' or both, then only those phonemes will be used that have the phonemes or broad classes specified in the context1 dictionary in the designated context. Currently this is only implemented for frame-based training. For landmark based training, use placeVoicingSpecification . The example file context_config.py shows an example of how to use context. If phonemes are specified in classes_1 and classes_2, then the context must also be phonemes, and the same for broad classes.

contextFlag2

Specify the left and right context of eaach of the labels in classes_2. Only the phonemes/broad classes with the specified context will be used. If the ith element of the list contains 'left' or 'right' or both, then only those phonemes will be used that have the phonemes or broad classes specified in the context2 dictionary in the designated context. Currently this is only implemented for frame-based training. For landmark based training, use placeVoicingSpecification . The example file context_config.py shows an example of how to use context. If phonemes are specified in classes_1 and classes_2, then the context must also be phonemes, and the same for broad classes.

context1

Specify the context . Relevant only if contextFlag1 is not empty. The element corresponding to to the ith classifier is a dictionary in python format. For example, an element may be 'left': ['iy', 'ow'], 'right': ['k', 'g']. Many examples

of using context are in the file context_config.py.

context2

Specify the context . Relevant only if contextFlag2 is not empty. The element corresponding to to the ith classifier is a dictionary in python format. For example, an element may be 'left': ['iy', 'ow'], 'right': ['k', 'g']. Many examples of using context are in the file context_config.py.

randomSelectionParameter1

Instead of picking all frames pick frames randomly. For example, randomSelectionParameter1 = [0, 0, 0, 0]. This feature has not been tested in a while, so please prefer not to use it. # Only relevant for frame-based training

randomSelectionParameter2

Instead of picking all frames pick frames randomly. For example, randomSelectionParameter2 = [0, 0, 0, 0] . This feature has not been tested in a while, so please prefer not to use it. Only relevant for frame-based training

middleFlag1

Specify if only the frames from a middle portion of each label is to be used for training. 1: middle 1/3 segment, 2: middle 2/3 segment, 3: only the center frame. Example, middleFlag1 = [0, 0, 0, 0] # Only relevant for frame-based training

middleFlag2

Specify if only the frames from a middle portion of each label is to be used for training. 1: middle 1/3 segment, 2: middle 2/3 segment, 3: only the center frame. Example, middleFlag1 = [0, 0, 0, 0] # Only relevant for frame-based

training

maxclass1

Maximum number of samples to be extracted for class +1. Example, maxclass1 = [20000, 5000, 20000, 20000] # Only relevant for frame-based training

maxclass2

Maximum number of samples to be extracted for class -1. Example, maxclass2 = [20000, 5000, 20000, 20000] # Only relevant for frame-based training

4. SVM parameter settings

trainingFileStyle = 'Light'

Choice between 'Light' and MATLAB . If MATLAB is chosen then a binary file is written .

kernelType = [2, 2, 2, 2]

Same usage as SVM Light. 10 : Use known optimal gammas. Set the optimumGammaValues below For example, kernelType = [2, 2, 2, 2]

gammaValues

The set of values from which optimal is to be found. For example, gammaValues = [0.05, 0.01, 0.005, 0.001, 0.0005, 0.00001]

optimumGammaValues

If optimal gamma value is known for each or some of the classifications, set it

here. For example, [0.01, 0.001, 0.001, 0.01] will set 0.01 as the optimal value for classification 0, 0.001 as optimal value fot the classification of index 1 and so on.

cValuesArray = [0.05, 0.5, 1.0, 10]

Values of C from which best C is to be chosen. For example, cValuesArray = [0.05, 0.5, 1.0, 10]

flagCheckForDifferentC

If set to 0, default C found by SVM Light will be used .

svmMinCriterion

If set to 'numSV' the minimum number of support vectors will be used to get the optimum value of C as well as gamma . 'crossValidation' will cause the code to use validation across the files in SvmInputFilesDevel . The files in SvmInputFilesDevel need to be created in a separate run of the code by specifying the same names in the SvmInputFiles

BinsFilenames

The names of files that will contain the histogram binning information. For example, BinsFilenames= ['BinsSonor30RBF', 'BinsStops30RBF', 'BinsSC30RBF', 'BinsSilence30RBF'] . Binning is not relevant for this version of the code.

probabilityConversionMethod

Choice of 'bins' or 'trivial' . Trivial will use linear mapping from [-1,1] to [0,1]

binningBound

Bins will be constructed between -binningBound and +binningBound

5. Parameters for scaling

   parameterScalingFlag

   If this is set to 1, the parameters will be scaled by their empirical mean and variance. If set to 1, findScalingParameters.py must run before train_config .

   scaleParameterFile

   The full path of file to be created by findScalingParameters.py and to be read by train_config.py . For example, modelDir+'/'+'scalesFile'

   scalingFactor

   The value at which standard deviation of the scaled parameters is set.

   scalingToBeSkippedFor

   A list of indices of features for scaling is not to be used. For example, [0,4,5]

6. Parameter Addition Specifications : Deprecated: should be ignored but not deleted

   addParametersFlag = 0

addDirectory = '/dept/isr/labs/nsl/scl/vol05/TIMIT_op/train'

temporalStepSize = 2.5

fileExts = ['aper.bin', 'per.bin', 'pitch.bin', 'soff.bin', 'son.bin']

channels = [1,1,1,1,1]

7. Ap specifications for landmark detection

useLandmarkApsFlags

Before landmark-based analysis is done, the code finds out the landmarks using the phoneme labels and optionally using knowledge based acoustic measurments. Landmarks are defined corresponding to broad classes vowel, fricative sonorant consonant (nasal or semivowel), silence and stop burst. If you want to use knowledge based measurements along with the phoneme labels for finding landmarks for any of the broad classes, set the corresponding flags as 1. For example, useLandmarkApsFlags = 'V':0, 'Fr':0, 'ST':1, 'SILENCE':0, 'SC':1 will cause the code to use measements for the landmarks for ST and SC, and only the phoneme labels will be used to find the other landmarks. The parameters defined by the landmarkAps will be used.

landmarkAps

The index of the parameter for each of the measurements - onset, offset, totalEnergy, syllabicEnergy, sylEnergyFirstDiff - has to be set below. For example,

landmarkAps = 'onset': 17, 'offset': 18, 'totalEnergy': 18, 'syllabicEnergy': 13, 'sylEnergyFirstDiff': 32 . Note that the first parameter is 1 and not zero. The maximum value of 'onset' parameter will be used to find stop burst. The maximum value of totalEnergy will be used to find the vowel landmark its minimum value will be used to find the dip of an intervocalic sonorant consonant. The maximum value of the sylEnergyFirstDiff will be used to find the SC offset (while moving from SC to vowel) and its minimum value will be used to find the SC onset (while moving from vowel to SC).

References

Isolet, release version 1.3, http://cslu.cse.ogi.edu/corpora/isolet/version.html.

Ali, A. M. A., Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition, Ph.D. thesis, University of Pennsylvania, 1999.

Allen, J. B., How do humans process and recognize speech?, *IEEE Trans. on Speech and Audio Proc.*, *2(4)*, 567–577, 1994.

Allen, J. B., From lord rayleigh to shannon: How do humans decode speech?, *International Conference on Acoustics, Speech and Signal Processing*, 2002, presentation, http://auditorymodels.org/jba/PAPERS/International Conference on Acoustics, Speech and Signal Processing.

Altun, Y., and T. Hofmann, Large margin methods for label sequence learning, *Eurospeech*, 2003.

Baker, J. K., The dragon system - an overview, *IEEE Trans. Acoustics, Speech, Signal Proc.*, *23(1)*, 24–29, 1975.

Bitar, N., Acoustic analysis and modelling of speech based on phonetic features, Ph.D. thesis, Boston University, 1997.

Burges, C., A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, pp. 121–167, 1998.

Carbonell, N., D. Fohr, and J. P. Haton, Aphodex, an acoustic-phonetic decoding expert system, *International Journal of Pattern Recognition and Artificial Intelligence*, *1(2)*, 1987.

Chang, J., Near-miss modeling: A segment-based approach to speech recognition, Ph.D. thesis, Massachusetts Institute of Technology Department of Electrical Engineering and Computer Science, 1998.

Chang, J., and J. Glass, Segmentation and modeling in segment based recognition, *Eurospeech*, pp. 1199–1202, 1997.

Chang, S., A syllable, articulatory-feature, stress-accent model of speech recognition, Ph.D. thesis, University of California, Berkeley, 2002.

Chen, M. Y., Nasal detection module for a knowledge-based speech recognition system, *International Conference on Spoken Language Processing*, *4*, 636–639, 2000.

Chomsky, N., and N. Halle, *The Sound Pattern of English*, Harper and Row, 1968.

Cole, R., R. Stern, M. Phillips, S. Brill, A. Pilant, and P. Specker, Feature-based speaker-independent recognition of isolated english letters, *International Conference on Acoustics, Speech and Signal Processing*, pp. 731–734, 1983.

Delgutte, B., and N. Y. S. Kiang, Speech coding in the auditory nerve: Iv. sounds

with consonant-like dynamic characteristics, *J. Acoustical Soc. of Amer.*, pp. 897–907, 1984.

Deng, L., and D. X. Sun, A statistical framework for automatic speech recognition using the atomic units constructed from overlapping articulatory features, *J. Acoust. Soc. Am.*, *100(4)*, 2500–2513, 1994.

Deshmukh, O., C. Espy-Wilson, and A. Juneja, Acoustic-phonetic speech parameters for speaker independent speech recognition, *International Conference on Acoustics, Speech and Signal Processing*, pp. 593–596, 2002.

Deshmukh, O., C. Espy-Wilson, and A. Salomon, Use of temporal information: Detection of the periodicity and aperiodicity profile of speech, *IEEE Trans. on Speech and Audio Processing*, to appear.

Drish, J., Obtaining calibrated probability estimates from support vector machines, 1998, web document, http://citeseer.nj.nec.com/drish01obtaining.html.

Eide, E., J. Rohlicek, H. Gish, and S. Mitter, A linguistic feature representation of the speech waveform, *International Conference on Acoustics, Speech and Signal Processing*, *93*, 483–486, 1993.

Espy-Wilson, C., An acoustic phonetic approach to speech recognition: Application to the semivowels, Ph.D. thesis, Massachusetts Institute of Technology, 1987.

Espy-Wilson, C., A feature-based semivowel recognition system, *J. Acoust. Soc. Am.*, *96*, 65–72, 1994.

Fanty, M., R. A. Cole, and K. Roginski, English alphabet recognition with telephone speech, *Advances in Neural Information Processing Systems*, 1992.

Fletcher, H., and J. C. Steinberg, Articulation testing methods, *Bell Syst. Tech. J.*, *88*, 806–854, 1929.

Fohr, D., J. Haton, and Y. Laprie, Knowledge -based techniques in acoustic-phonetic decoding of speech: interests and limitations, *International Journal of Pattern Recognition and Artificial Intelligence*, *8*, 133–153.

Forney, G. D., The viterbi algorithm, *Proc. IEEE*, *61*, 268–278, 1973.

Ganapathiraju, A., Support vector machines for speech recognition, Ph.D. thesis, Mississippi State University, 2002.

Glass, J., Nasal consonants and nasalized vowels: an acoustic study and recognition experiment, Master's thesis, Massachusetts Institute of Technology, 1984.

Glass, J., and V. Zue, Multi-level acoustic segmentation of continuous speech, *International Conference on Acoustics, Speech and Signal Processing, New York, NY*, pp. 429–432, 1988.

Glass, J., J. Chang, and M. McCandless, A probabilistic framework for feature-based speech recognition, *International Conference on Spoken Language Processing*, pp. 2277–2280, 1996.

Greenberg, S., J. Hollenback, and D. Ellis, Insights into spoken language gleaned

from phonetic transcription of the switchboard corpus, *International Conference on Spoken Language Processing*, 1996.

Halberstadt, A. K., Heterogenous acoustic measurements and multiple classifiers for speech recognition, Ph.D. thesis, Massachusetts Institute of Technology, 1998.

Hasegawa-Johnson, M., Formant and burst spectral measurements with quantitative error models for speech sound classification, Ph.D. thesis, Massachusetts Institute of Technology, 1996.

Hasegawa-Johnson, M., et al., Landmark-based speech recognition: Report of the 2004 johns hopkins summer workshop", *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2005, submitted.

Hosom, J. P., Automatic time alignment of phonemes using acoustic-phonetic information, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, 2000.

Howitt, A. W., Automatic syllable detection for vowel landmarks, Ph.D. thesis, Massachusetts Institute of Technology, 2000.

Jelinek, F., Continuous speech recognition by statistical methods, *Proc. IEEE. 64*, *4*, 1976.

Juneja, A., and C. Espy-Wilson, Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines, *International Joint Conference on Neural Networks*, 2003.

Juneja, A., and C. Espy-Wilson, Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition, *From sound to sense: 50+ years of discoveries in speech communication, MIT, Cambridge MA*, pp. C–151 to C–156, 2004.

Jurafsky, D., and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2000.

Kirchhoff, K., Robust speech recognition using articulatory information, Ph.D. thesis, U. of Bielefeld, Germany, 1999.

Kwok, J. T., The evidence framework applied to support vector machines, *IEEE Transactions on Neural Networks*, *11(5)*, 1162–1173, 2000.

Lee, S., Probabilistic segmentation for segment-based speech recognition, Master's thesis, Massachusetts Institute of Technology, 1998.

Lippmann, R. P., Speech recognition by machines and humans, *Speech Communication*, *22*, 1–15, 1997.

Liu, S. A., Landmark detection for distinctive feature based speech recognition, *J. Acoust. Soc. Am.*, *100(5)*, 3417–, 1996.

Livescu, K., and J. Glass, Feature-based pronunciation modeling for speech recognition, *HLT/NAACL*, May 2004.

Massaro, D. W., The paradigm and the fuzzy logical model of perception are alive and well, *Journal of Experimental Psychology*, pp. 115–125, 1993.

Mermelstein, P., Automatic segmentation of speech into syllabic units, *J. Acoust. Soc. Am., 58(4)*, 860–883, 1975.

Mesgarani, N., M. Slaney, and S. A. Shamma, Speech discrimination based on multiscale spectrotemporal features, *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2004.

Miller, G. A., and P. E. Nicely, An analysis of perceptual confusions among some english consonants, *J. Acoustical Soc. of Amer., 27*, 338–352, 1955.

NIST, nIST Spoken Language Technology Evaluations, http://www.nist.gov/speech/tests/index.htm.

NIST, Timit acoustic -phonetic continuous speech corpus, *NTIS Order No. PB91 -5050651996*, 1990.

Niyogi, P., Distinctive feature detection using support vector machines, *International Conference on Acoustics, Speech and Signal Processing*, pp. 425–428, 1998.

Ohde, R. N., and K. N. Stevens, Effect of burst amplitude on the perception of stop consonant place of articulation, *J. Acoustical Soc. of Amer., 74*, 706–714, 1983.

Pruthi, T., and C. Espy-Wilson, Automatic classification of nasals and semivowels, *International Conference on Phonetic Sciences, Barcelona, Spain*, 2003.

Rabiner, L., and B. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

Salomon, A., Speech event detection using strictly temporal information, Master's thesis, Boston University, 2000.

Salomon, A., C. Espy-Wilson, and O. Deshmukh, Detection of speech landmarks from temporal information, *J. Acoust. Soc. Am.*, *115*, 1296–1305, 2004.

Seneff, S., A joint synchrony/mean-rate model of auditory speech processing, *J. of Phonetics*, *16*, 55–76, 1988.

Stevens, K. N., Acoustic correlates of some phonetic categories, *J. Acoust. Soc. Am.*, *68*, 836–842, 1980.

Stevens, K. N., Toward a model for lexical access based on acoustic landmarks and distinctive features, *J. Acoust. Soc. Am.*, *111(4)*, 1872–1891, 2002.

Stevens, K. N., S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, Implementation of a model for lexical access based on features, *International Conference on Spoken Language Processing*, 1992.

Stevens, K. N., S. Manuel, and M. Matthies, Revisiting place of articulation measures for stop consonants: Implications for models of consonant production, *International Congress on Phonetic Sciences*, *2*, 1117–1120, 1999.

Stolcke, A., H. Franco, and R. Gadde, 2003, speech-to-text research at SRI-ICSI-UW, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/sri+-rt03-stt.pdf.

Tartter, V. C., D. Kat, A. G. Samuel, and B. H. Repp, Perception of intervocalic

stop consonants: the contributions of closure durations and formant transitions, *J. Acoustical Soc. of Amer.*, *74*, 715–725, 1983.

Vapnik, V., *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.

Viterbi, A. J., Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Trans. Information Theory*, *IT-13*, 260–269, 1967.

Zhang, Y., Towards implementation of a feature-based lexical-access system, Master's thesis, Massachusetts Institute of Technology, 1998.

Zue, V., and R. Cole, Experiments on spectrogram reading, *International Conference on Acoustics, Speech and Signal Processing*, 1995.

Zue, V., J. Glass, M. Philips, and S. Seneff, The mit summit speech recognition system: A progress report, *DARPA Speech and Natural Language Workshop*, pp. 179–189, 1989.