

Speech Recognition using Acoustic Landmarks and Binary Phonetic Feature Classifiers

October 31, 2003

Amit Juneja
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742, USA
juneja@glue.umd.edu

Ph.D. Thesis Proposal

Abstract

In spite of decades of research, Automatic Speech Recognition (ASR) is far from reaching the goal of performance close to Human Speech Recognition (HSR). One of the reasons for unsatisfactory performance of the state-of-the-art ASR systems, that are based largely on Hidden Markov Models (HMMs), is the inferior acoustic modeling of low level or phonetic level linguistic information in the speech signal. An acoustic-phonetic approach to ASR, on the other hand, explicitly targets linguistic information in the speech signal. But an acoustic phonetic system that carries out large ASR speech recognition tasks, for example, connected word or continuous speech recognition, does not exist. We propose a probabilistic and statistical framework for ASR based on the knowledge of acoustic phonetics for connected word ASR. The proposed system is based on the idea of representation of speech sounds by bundles of binary valued articulatory phonetic features. The probabilistic framework requires only binary classifiers of phonetic features and the knowledge based acoustic correlates of the features for the purpose of connected word speech recognition. We explore the use of Support Vector Machines (SVMs) for binary phonetic feature classification because of the favorable properties well suited to our recognition task that SVMs offer. In the proposed method, probabilistic segmentation of speech is obtained using SVM based classifiers of manner phonetic features. The linguistically motivated landmarks obtained in each segmentation is used for classification of source and place phonetic features. Probabilistic segmentation paths are constrained using Finite State Automata (FSA) for isolated or connected word recognition. The proposed method could overcome the disadvantages encountered by the early acoustic-phonetic knowledge based systems, that led the ASR community to switch to ASR systems highly dependent on statistical pattern analysis methods.

Contents

1	Introduction	4
1.1	Speech Production and Phonetic Features	4
1.2	Acoustic correlates of phonetic features	7
1.3	Definition of acoustic-phonetic knowledge based ASR	8
1.4	Hurdles in the acoustic-phonetic approach	10
1.5	State-of-the-art ASR	12
1.6	ASR versus HSR	13
1.7	Overview of the proposed approach	16
2	Literature Survey	17
2.1	Acoustic-phonetic approach	17
2.1.1	Landmark detection or segmentation systems	17
2.1.2	Word or sentence recognition systems	19
	The SUMMIT system	19
	Other methods	20
2.2	Knowledge based front-ends	21
2.3	Phonetic features as recognition units in statistical methods	22
2.4	Conclusions from the literature survey	23
3	Method	24
3.1	Segmentation using manner phonetic features	24
3.1.1	The use of Support Vector Machines (SVMs)	27
3.1.2	Duration approximation	28
3.1.3	Priors and probabilistic duation	29
3.1.4	Initial experiments and results	30
3.1.5	Probabilistic segmentation algorithm	31
3.2	Detection of features from landmarks	33
3.2.1	Initial experiments with place and voicing feature detection	34
3.3	Framework for isolated and connected word recognition	34
3.3.1	Evolving ideas on the use of probabilistic language model	36
3.4	Project Plan	37
	References	39
A	American English Phonemes	44
B	Tables of place and voicing features	46
C	Support Vector Machines	47
C.1	Structural Risk Minimization (SRM)	47
C.2	SVMs	47

1 Introduction

In this section, we will build up the motivation of the proposed probabilistic and statistical framework for our acoustic-phonetic approach to Automatic Speech Recognition (ASR). The proposed approach to ASR is based on the concept of bundles of articulatory phonetics features and acoustic landmarks. 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 we present the basic ideas of acoustic phonetic knowledge based ASR. The various drawbacks of the acoustic phonetic approach that have led the ASR community to abandon the approach and our ideas of solving those problems are briefly discussed in Section 1.4. We present the basics and the terminology of the state-of-the-art ASR, that is based largely on Hidden Markov Models (HMMs) in Section 1.5 and compare the performance of the state-of-the-art systems with human speech recognition in Section 1.6. Finally we give an overview of the proposed approach in Section 1.7. A literature survey of the previous ASR systems that utilize acoustic phonetic knowledge is presented in Section 2. Section 3 presents the proposed acoustic phonetic knowledge based framework for phoneme and connected word speech recognition.

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 we hear. 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 [1] are minimal binary valued units that are sufficient to describe all the speech sounds in any language. In the description of phonetic features, we give examples 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 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

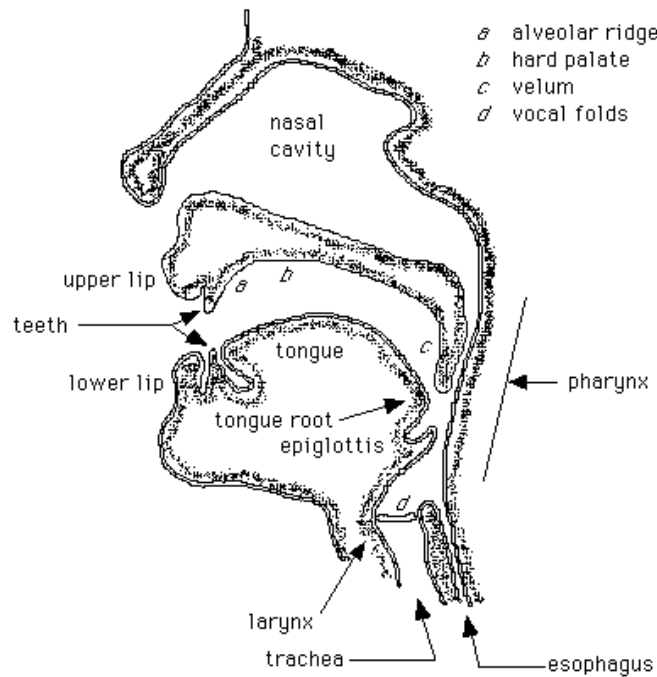


Figure 1.1: The vocal tract

the constriction and whether the air flow is through the mouth or the nasal cavity. Manner phonetic features are also called articulator-free features [4] 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 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 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.

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

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

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

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

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

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

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
<i>velar</i>	Constriction between tongue body and soft palate	-	-	+
<i>alveolar</i>	Constriction between tongue tip and alveolar ridge	-	+	-
<i>labial</i>	Constriction between the lips	+	-	-

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

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

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.

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, Bitar [50], Stevens [59], Espy-Wilson [2], Ali [34]. In this work, we will use the term Acoustic Parameters (APs) for the acoustic correlates that can be extracted automatically from the speech signal. In our recognition framework, the APs related to the broad manner phonetic features - *sonorant*, *syllabic* and *continuant* - are extracted from every frame of speech. Table 1.5 provides examples of APs for manner phonetics features that were developed by Bitar and Espy-Wilson [50], and later used by us in Support Vector Machine (SVM) based segmentation of speech [5].

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 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 [50] 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 [59] 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

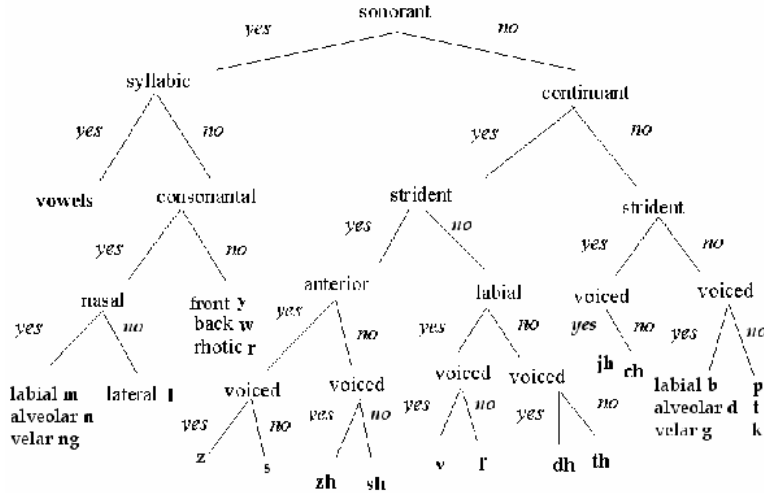


Figure 1.2: Phonetic feature hierarchy

Phonetic Feature	APs
<i>sonorant</i>	(1) Probability of voicing [51], (2) First order autocorrelation (3) Ratio of $E[0, F3-1000]$ to $E[F3-1000, f_s/2]$, (4) $E[100, 400]$
<i>syllabic</i>	(1) $E[640, 2800]$ and (2) $E[2000, 3000]$ normalized by nearest syllabic peaks and dips
<i>continuant</i>	(1) Energy onset, (2) Energy offset, (3) $E[0, F3-1000]$, (4) $E[F3-1000, f_s/2]$

Table 1.5: APs for the features *sonorant*, *syllabic* and *continuant*. ZCR : zero crossing rate, f_s : sampling rate, F3 : third formant average. $E[a, b]$ denotes energy in the frequency band [aHz, bHz]

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 we will follow in the proposed project. The landmark based approach to speech recognition is similar to human spectrogram reading [7] 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 landmark based approach to speech recognition utilizes strong correlation among the speech frames. The landmark based approach to speech recognition has been advocated by Stevens [3, 4] and further pursued by Liu [6] and Bitar and Espy-Wilson [50, 2].

1.3 Definition of acoustic-phonetic knowledge based ASR

We can broadly classify all the approaches to ASR as either 'static' or 'dynamic'. In the static approach, explicit events are located in the speech signal and the recognition of units - phonemes

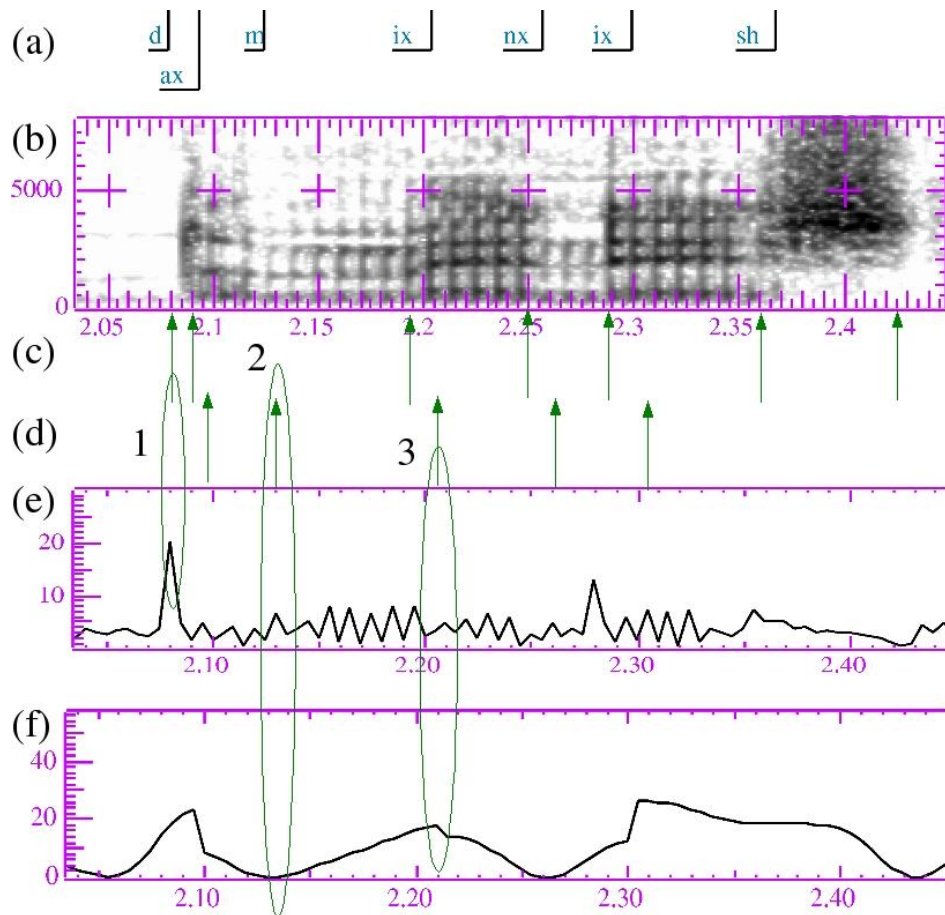


Figure 1.3: Illustration of manner landmarks for the utterance "diminish" from the TIMIT database [35]. (a) Phoneme Labels, (b) Spectrogram, (c) Landmarks characterized by sudden change, (d) Landmarks characterized by maxima or minima of a correlate of a manner phonetic feature, (e) Onset waveform (an 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/.

or phonetic features - is carried out using a fixed number of acoustic measurements extracted using those events. In the static method, no statistical dynamic models like HMMs are used to model the time varying characteristics of speech. In this proposal, we define the acoustic phonetic approach to ASR as a static approach where analysis is carried out at explicit locations in the speech signal. Our landmark based approach to ASR belongs to this category. In the dynamic approach, speech is modeled by statistical dynamic models like HMMs and we discuss this approach further in Section 1.5. Acoustic-phonetic knowledge has been used in dynamic systems but we refrain from calling such methods as acoustic-phonetic approaches because there is no explicit use of acoustic events and acoustic correlates of articulatory features in these systems.

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 [31] but we define it 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. [8]. The latter method of using broad manner class scores to segment the signal has been used by a number of researchers [50, 6, 9, 10]. Multiple segmentations may be generated instead of a single representation, for example, the dendograms in the speech recognition method proposed by Glass [8]. (We include the system proposed by Glass et al. 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 and is largely statistical.)
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 to ASR in the literature. Rabiner [31] 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 our ideas for solving them provide much of the motivation for the proposed work. We now list these

documented problems of the acoustic-phonetic approach and argue that either not sufficient 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 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 we will show that this is not necessary. Vocabulary and grammar constraints may be used to constrain the speech segmentation paths, as will be shown by the recognition framework we propose.
- 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 [59, 34, 50], nasal detection [11], and semivowel classification [2]. We believe the knowledge from these sources is adequate to start building an acoustic-phonetic speech recognizer to carry out big recognition tasks, and that will be a focus of the proposed project. The knowledge based acoustic correlates of phonemes or phonetic features offer a significant advantage that the standard front ends are not able to offer. 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 we will use in our work will be based on the distinctive articulatory feature theory and it will be optimal in that sense. But the proposed system will be flexible to take as a design parameter a different set of features. Such flexibility will make the system usable as a test bed to find an optimal set of features although that is not the focus of the proposed work.
- Another drawback of the acoustic-phonetic approach as pointed out in [31] is that the design of the sound classifiers is not optimal. This argument assumes that binary decision trees 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 we shall discuss in Section 2. Statistical pattern recognition methods have been applied in some acoustic phonetics knowledge based methods, for example, [23, 9] 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 will be possible in the proposed approach. But the goal of our work is 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 [12] and Baker [13]. We will present a very brief review of HMM based ASR, starting with how isolated word recognition is carried out using HMMs. Given a sequence of observation vectors $O = \{o_1, o_2, \dots, o_T\}$, 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

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

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 , we can define a_{ij} , $b_j(o)$ and π_i as

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

$$b_j(o) = P(o_t = o | q_t = j) \quad (1.3)$$

$$\pi_i = P(q_1 = i), \quad 1 \leq i \leq N \quad (1.4)$$

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

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

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 [14, 15] 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 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

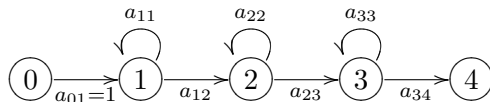


Figure 1.4: A typical topology of an HMM used in ASR with non-emitting start and end states 0 and 4

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

$$\hat{W} = \arg \max_W P(O|W)P(W). \quad (1.6)$$

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 [67].

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 (a set of two phones), triphones (a set of three 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.

We have presented the basic ideas of HMM based approach to ASR. An enormous number of modifications and improvements over the basic HMM method for ASR have been suggested in the past two decades, but we refrain from discussing these methods here. The goal of the proposed work is an acoustic phonetic knowledge based system that will operate very differently from the HMM approach. We now discuss briefly 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.

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 [16] compared the performance of ASR with HSR. The comparison is still valid today given only incremental improvements to HMM based

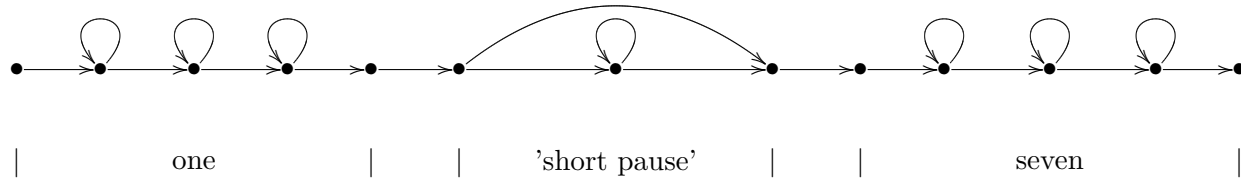


Figure 1.5: Concatenation of word level HMMs for the words - 'one' and 'seven' - through a 'short pause' model. To find the likelihood of an utterance given the sequence of these two words, the HMMs for the words are concatenated with an intermediate 'short pause' model and the best path through the 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

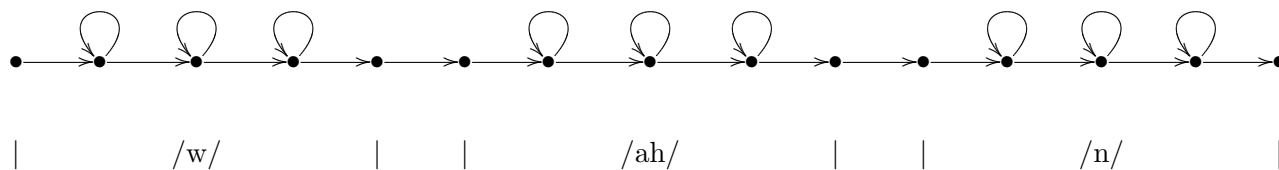


Figure 1.6: Concatenation of phone level HMMs for the phonemes - /w/, /ah/ and /n/ - to get the model of the word 'one'. To find the likelihood of an utterance given the word 'one', the HMMs for the these phonemes are concatenated and the best path through the state transition graph is found. Similarly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

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 our 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 isolated 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 isolated alphabet. On telephone quality speech, Ganapathiraju [62] 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 performance 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 [17]. 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 [29, 30] 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 our acoustic-phonetic knowledge based approach is on the recognition of phonetic features and the correct recognition of phonetic features will lead to correct recognition of phonemes. The recognition system we propose will not be based on processing different frequency bands independently, but we will not be using all the available information at the same time for recognition all the phones. That is, different information (acoustic correlates of phonetic features) will be used for recognition of different features to get partial recognition results (in terms of phonetic features) and at times this information will belong to different frequency bands. We believe that this system is closer to human speech recognition than HMM based systems because the focus is on low level (phone and phonetic feature level) information.

1.7 Overview of the proposed approach

The goal of the landmark based acoustic-phonetic approach to speech recognition is to explicitly target low-level linguistic information in the speech signal by extracting acoustic correlates of the phonetic features. The landmark based approach offers a number of advantages over the HMM based approach. First, because the analysis is carried out at significant landmarks, the method utilizes the strong correlation among the speech frames. This makes the landmark based approach very different from the HMM based approach where every frame of speech is processed assuming independence among the frames. Second, the acoustic measurements in the landmark based approach are made on the basis of knowledge and they are used only for relevant classification tasks which makes the system easy to analyze for errors. HMMs, on the other hand, use all the measurements for all decisions. Third, many coarticulation effects are explicitly taken into account by normalizing acoustic measurements by adjoining phonemes instead of building statistical models for diphones or triphones. In the proposed system, the low level acoustic analysis will be carried out explicitly on the basis of acoustic phonetic knowledge and the probabilistic framework will allow the system to be scaled for any recognition task.

2 Literature Survey

A number of ASR procedures have appeared in the literature that make use of acoustic phonetics knowledge. We would classify these procedures into three broad categories that will make it easy for the reader to contrast these methods with our 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

This is the recognition strategy that we outlined in Section 1.3. The acoustic phonetic approach 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) and 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 [50] used knowledge based acoustic parameters in a fuzzy logic framework to segment the speech signal into the broad classes - vowel, sonorant consonant, 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*, *alveolar* and *velar* for stop consonants. We will use the APs developed by Bitar in our proposed project and find or further optimize APs for some of the phonetic features. A recognition system for isolated or connected word speech recognition was not developed in this work.

Liu [6] 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 substitutions 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

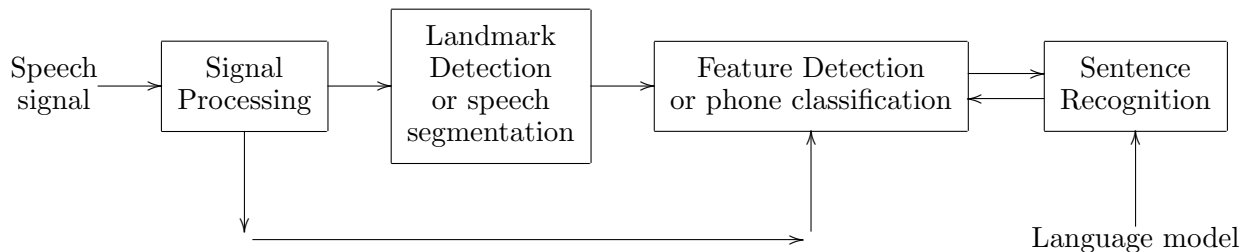


Figure 2.1: Block diagram of acoustic phonetic approach

lexical access was proposed by Stevens et al, [3, 4] as we have discussed in Section 1.2. However, a practical framework for speech recognition was not presented in either of these works.

Salomon [18] used temporal measurements derived from average magnitude difference function (AMDF) 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 al [19] will be used in the proposed project, especially, the temporal measures of periodicity and aperiodicity as well as energy onset and offset will be used to supplement or replace the spectral based measures developed by Bitar [50].

Ali [34] 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 [65]. 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 [6] 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.

Mermelstein [20] 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 recognizing the speech signal, the idea of using convex hull was utilized later by Espy-Wilson [2], Bitar [50] and Howitt [64] in locating sonorant consonants and vowels in the speech signal and we will use it as well in the knowledge based front-end for the proposed system.

2.1.2 Word or sentence recognition systems

The SUMMIT system

The SUMMIT system [36, 37, 38, 39] 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 one of the two ways - (1) acoustic segmentation [8] method finds time instances when the change in the spectrum is beyond a certain threshold and (2) boundary detection methods use statistical context dependent broad class models [41, 40]. 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 [37] and with the improvements made by Halderstadt [38] 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 [38, 39]. We will discuss below this probabilistic framework in some detail because the probabilistic framework we use in our work is similar to it in some ways, although there are significant differences that we discuss 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

$$\hat{W} = \arg \max_W P(W|O) \quad (2.1)$$

Chang [39] 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

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

where U is a sequence of subword units like phones, diphones and triphones. S denotes the segmentation, that is, the length that each unit in the sequence S occupies. 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 or 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 by 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

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

$P(O|WUS)$ 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 proposal. A particular segmentation (sequence of segments) may not use all the features available in the observation sequence O . Therefore, a difficulty is met

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

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

in comparing the term $P(O/WUS)$ for different segmentations. Two different procedures have been proposed to solve this problem - Near-Miss Modeling [39] and anti-phone modeling [37].

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 [38] 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.

We have presented the basic ideas used in the SUMMIT system. Our approach to ASR 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 the system we propose is a phonetic feature based system. That is, phonetic feature models are built in our system instead of phone models. Secondly, although our system uses a similar idea of obtaining multiple segmentations and then carrying further analysis based on the information obtained from those segments, we concentrate on linguistically motivated landmarks instead of analyzing all the front-end parameters extracted from segments and segment boundaries. Third, because we will operate entirely with posterior probabilities of binary phonetic features, we will not need to account for all acoustic observations for each segmentation. Fourth, in our proposed system, 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

Fanty and Cole et al. [42] proposed a neural network based recognizer that can be classified as an acoustic-phonetic approach. 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 [60] 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 [42] was the more advanced version of the FEATURE system [43] developed by Cole et al. 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 like the one we use. 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 we classify this as an acoustic phonetic approach, it should be noted here that this was not an articulatory feature based system.

Fohr et al. [9, 10] proposed 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. 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.

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 front-ends in these systems led to improvement in performance using certain performance criteria.

Bitar and Espy-Wilson [50] 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 the

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 (using digits) by Deshmukh et al. [33] and a similar pattern was observed not only for cross gender testing but also for testing across adults and children.

Hosom [44] 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 [45] 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 [35] database over a baseline HMM/ANN system. When the time-alignment system was used to train the hybrid HMM/ANN for the OGI alphasdigit task [61], 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 unit 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 [46] 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. [47] proposed a method of phoneme classification using 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 [48] 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 the SUMMIT is motivating because it seems 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.

3 Method

In this section we will present the methodology of landmark based ASR using our event-based system (EBS). EBS is characterized by three steps in the recognition process - broad class recognition which results in a set of landmarks, the recognition of place and voicing phonetic features, and lexical access. Multiple hypothesis segmentations allow multiple hypothesis landmark sequences to be extracted, and then each landmark sequence is combined with the APs for place and voicing phonetic features to develop a hypothesis sequence of phonemes. The high level vocabulary or grammar information can be combined with the above procedure for word or sentence recognition. We can express the problem of phoneme recognition as maximizing the posterior probability of landmarks and the corresponding feature bundles (or equivalently, phonemes), given the observation sequence O , that is,

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

where $L = \{l_i\}_{i=1}^M$ is a sequence of landmarks and $U = \{u_i\}_{i=1}^M$ is the sequence of phonemes or bundles of features corresponding to the phoneme sequence.

We segment the speech signal into five broad manner classes - vowel (V), fricative (Fr), sonorant consonant (SC), stop (ST) and silence (SIL) - and obtain a set of landmarks for each broad class as shown in Table 3.1. Let B denote a broad class sequence for an utterance. The meanings of the symbols B , L and U can be explained with the help of Table 3.2 which shows the values of these symbols for the canonical pronunciation /z I r ow/ of the word 'zero'. The sequence of landmarks for an utterance is completely determined by its broad class sequence. Therefore, we can write

$$P(L|O) = P(B_L|O) \quad (3.2)$$

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 . They are only sequences of symbols with no information about the point in time where they occur.

The procedure for obtaining $P(B)$ is presented in Section 3.1 and that for $P(U|OL)$ is presented in Section 3.2. In Section 3.3, the isolated and connected word recognition by EBS is discussed. We will assume while presenting the probabilistic framework that no two consecutive phonemes in the phoneme sequence U have the same broad manner class representation.

3.1 Segmentation using manner phonetic features

Given a sequence of T frames $O = \{o_1, o_2, \dots, o_T\}$, where o_t is the vector of APs at time t , we need to find the most probable sequence of broad classes $B = \{B_i\}_{i=1}^M$ and their durations $D = \{D_i\}_{i=1}^M$. The frame o_t is the set of all the knowledge based acoustic parameters (APs) computed at frame t . Not all the APs at each time frame will be used by EBS but we assume that these are available, so as to develop the probabilistic framework. EBS uses the probabilistic phonetic feature hierarchy shown in Figure 3.1 to segment speech into the five manner classes. The concept of probabilistic hierarchies has appeared before with application to phonetic classification, for example [38, 66], but it has not been used as a uniform framework for speech segmentation as well as phonetic classification. The broad class segmentation problem can be stated mathematically as,

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

Broad Class Segment	Landmark Type	Landmark location
Vowel	Syllabic peak	Maximum value of E[640-2800] in the vowel region
	Vowel onset	Beginning of sonorancy for vowels following fricatives, stops or silence
Stop	Burst	Maximum value of onset around the beginning of stop region
SC	Syllabic dip	Minimum value of E[640-2800] in the SC region
	SC onset	Maximum value of energy offset in the transition region from vowel to SC for intervocalic and post-vocalic SCs
	SC offset	Maximum value of energy onset in the transition region from SC to V for intervocalic and pre-vocalic SCs
Fricative	Fricative onset	Beginning of frication
	Fricative offset	End of frication

Table 3.1: Landmark detection in EBS. This table shows the landmarks extracted for each of the manner classes and the knowledge based acoustic measurements used to obtain the manner landmarks.

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

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

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

$$(3.6)$$

and similarly for the other manner classes. We will use P_t to denote the posterior probability of a feature or a set of features at time t . We have used the fact that the presence of the phonetic feature *sonorant* implies the presence of speech, that is,

$$P_t(\text{syllabic}|\text{sonorant}, O) = P_t(\text{syllabic}|\text{sonorant}, \text{speech}, O) \quad (3.7)$$

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 , we only need to pick the acoustic correlates of the feature in a set of frames $\{t-s, t-s+1, \dots, t+e\}$, using s previous frames and e following frames along with the current frame t . 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 . We can write

$$P_t(V|O) = P_t(\text{speech}|x_t^{\text{speech}})P_t(\text{sonorant}|\text{speech}, x_t^{\text{sonorant}})P_t(\text{syllabic}|\text{sonorant}, x_t^{\text{syllabic}}) \quad (3.8)$$

In general, if we express a broad class b in terms of its underlying N_b phonetic features $\{f_1, f_2, \dots, f_{N_b}\}$,

	/z/	/I/	/r/	/o/	/w/
$U \Rightarrow$	u_1	u_2	u_3	u_4	u_5
	<i>-sonorant</i>	<i>+sonorant</i>	<i>+sonorant</i>	<i>+sonorant</i>	<i>+sonorant</i>
	<i>+continuant</i>	<i>+syllabic</i>	<i>-syllabic</i>	<i>+syllabic</i>	<i>-syllabic</i>
	<i>+strident</i>	<i>-back</i>	<i>-nasal</i>	<i>+back</i>	<i>-nasal</i>
	<i>+voiced</i>	<i>+high</i>	<i>+rhotic</i>	<i>-high</i>	<i>+labial</i>
	<i>+anterior</i>	<i>+lax</i>		<i>+low</i>	
$B \Rightarrow$	Fr	V	SC	V	SC
$L \Rightarrow$	l_1	l_2	l_3	l_4	l_5
	Fricative onset	Vowel onset	SC onset	Vowel onset	SC onset
	Fricative offset	Syllabic peak	Syllabic dip	Syllabic peak	Syllabic dip
			SC offset		SC offset

Table 3.2: An illustrative example of the symbols B , L and U

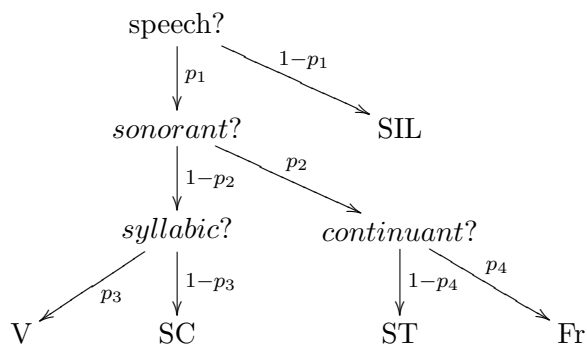


Figure 3.1: Probabilistic Phonetic Feature Hierarchy

we can write

$$P_t(b|O) = \prod_{i=1}^{N_b} P_t(f_i|x_t^{f_i}, f_1, \dots, f_{i-1}) \quad (3.9)$$

Furthermore, we can assume that given the acoustic correlates of the manner phonetic features, the posterior probabilities of the phonetic features are independent across frames, that is, the acoustic correlates of the phonetic features are sufficient to determine the probabilities of the binary manner phonetic features. Therefore, denoting the features for class B_i as $\{f_1^i, f_2^i, \dots, f_{N_{B_i}}^i\}$,

$$P(BD|O) = \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} P_t(B_i|O) \quad (3.10)$$

$$= \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i|x_t^{f_k^i}, f_1^i, \dots, f_{k-1}^i) \quad (3.11)$$

3.1.1 The use of Support Vector Machines (SVMs)

Because the phonetic features are binary valued, the posterior probabilities $P_t(f_k^i|x_t^{f_k^i}, f_1^i, \dots, f_{k-1}^i)$ may be calculated using any binary classifier that can output posterior probabilities for each of the two classes. We use SVMs [21, 22] for this purpose because of some attractive properties for the current task, for example, ability to learn from a small amount of training data and capacity to handle high dimensional data. A brief review of SVMs appears in Appendix C. SVMs have been shown to perform better than Bayesian methods for distinctive feature detection in speech [23, 24] and phonetic classification from hand transcribed segments [25, 26]. Many methods have been suggested to convert SVM outputs to probabilities [52, 53], but we have chosen in our initial experiments to clip the SVM outputs in the range $[-1, +1]$, scale the result down by $1/2$ and translate the outcome to the range $[0, 1]$. This simple scheme works considerably well for speech segmentation as we have shown before [5, 27].

Note that for the recognition of five broad classes, only four binary SVMs are needed, one for each manner phonetic feature, that is, for each node in the phonetic feature hierarchy. Table 3.3 shows the classes that are trained against each other for building the four SVMs. A clear advantage of this system is that to recognize the five broad classes, each class to be recognized does not have to be trained against all of the other classes. For example, the samples of V do not have to be trained against the samples of all the other classes - SC, Fr, ST and SIL. Instead, given the hierarchy, the samples of V are trained against the samples of SC for the SVM that calculates p_3 and the samples of V and SC are trained against the samples of ST and Fr for the SVM that calculates p_2 , and so on. For each binary classifier in Table 3.3, a comparable amount of training data for the two classes is available. Moreover, since all the classifiers are binary, the method overcomes the need to find good multi-class SVMs or other multiclass classifiers. Although a non-probabilistic hierarchy, can be used to limit the number of classifiers to four, such an approach will not allow probabilistic segmentation. Thus, the errors at the phonetic feature level will not be corrected by language constraints.

Table 3.4 shows the APs used by each SVM classifier. Unlike the HMM based approach in general and the statistical methods that build phonetic feature models [46, 48], each classifier in

Phonetic feature	class +1	class -1
Speech	silence	speech
<i>sonorant</i>	sonorant	non-sonorant
<i>syllabic</i>	sonorant con-sonant	vowel
<i>continuant</i>	stop burst	frication noise

Table 3.3: Training of phonetic feature SVMs

EBS uses only the APs that are required for the corresponding phonetic feature. The optimal values of s and e were found for each manner feature by varying these values over a wide range and selecting the values that gave the minimum error on test data. At a frame step size of 5ms, the values $s = 6$ and $e = 3$ were found to be optimal for feature *continuant*. For all of the other classifiers, $s = 0$ and $e = 0$ were found to be optimal, that is, APs from only the current analysis frame are used. This does not imply that no information is used from the rest of the utterance for these classifications. As shown in Table 3.4 some of the APs are normalized by nearest peaks and dips, and some are calculated using F3 (third formant) average throughout the utterance.

3.1.2 Duration approximation

With no duration and language constraints, the class label b_t at time t is hypothesized by

$$\hat{b}_t = \arg \max_b P(b|O)$$

with $b \in \{V, SC, ST, Fr, SIL\}$

The segmentation of the test signal is then found by collapsing consecutive identical class labels. This is a very simple procedure to obtain the most probable segmentation $\{\hat{B}, \hat{D}\}$, but, as we saw in Equations 3.1 and 3.2, the probabilities $P(B|O)$ for different B are more important as far as phoneme and connected word recognition is concerned. We can write

$$P(B|O) = \sum_D P(BD|O) \tag{3.12}$$

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, we make the approximation that is similar to the approximation made by Viterbi decoding in HMM based recognition systems and the SUMMIT system [37],

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

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). The form of Equation 3.14 also enables us to impose certain explicit duration constraints that can reduce

the insertions in the segmentation. Instead of maximizing $P(BD|O)$ for a particular sequence B over all D , the maximization can be carried out only over those D that satisfy the duration constraints. A typical duration constraint is to restrict the duration of vowels, SCs and fricatives to more than 10ms. A probabilistic segmentation algorithm that calculates $\max_D P(BD|O)$ and handles explicit duration constraints will be presented in section 3.1.5.

3.1.3 Priors and probabilistic duation

Probabilistic duration $P(D|B)$ and prior probabilities $P(B)$ can be used if we make a certain set of assumptions. Denote the features for class B_i as $\{f_1^i, f_2^i, \dots, f_{N_{B_i}}^i\}$, the broad class at time t as b_t , and the sequence $\{b_1, b_2, \dots, b_{t-1}\}$ as b^{t-1} . If we make an assumption that the acoustic correlates are sufficient to determine the probabilities of the manner phonetic features, even if the the broad classes of previous frames are provided,

$$P(BD|O) = \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} P_t(B_i|O, b^{t-1}) \quad (3.15)$$

$$= \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i | x_t^{f_k^i}, f_1^i, \dots, f_{k-1}^i, b^{t-1}) \quad (3.16)$$

Now expanding the conditional probability, we get

$$= \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i, x_t^{f_k^i}, f_1^i, \dots, f_{k-1}^i, b^{t-1})}{P_t(x_t^{f_k^i}, f_1^i, \dots, f_{k-1}^i, b^{t-1})}. \quad (3.17)$$

Splitting the priors,

$$P(BD|O) = \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i | f_1^i, \dots, f_{k-1}^i, b^{t-1}) \frac{P_t(x_t^{f_k^i} | f_1^i, \dots, f_k^i, b^{t-1})}{P_t(x_t^{f_k^i} | f_1^i, \dots, f_{k-1}^i, b^{t-1})}. \quad (3.18)$$

Clearly

$$P(BD|O) = \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i | f_1^i, \dots, f_{k-1}^i, b^{t-1}) = P(BD) = P(B)P(D|B) \quad (3.19)$$

Now given the set f_1^i, \dots, f_{k-1}^i or the set f_1^i, \dots, f_k^i , if $x_t^{f_k^i}$ is assumed to be independent of b^t , then

$$P(BD|O) = P(B)P(D|B) \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} \frac{P_t(f_k^i | x_t^{f_k^i}, f_1^i, \dots, f_{k-1}^i)}{P_t(f_k^i | f_1^i, \dots, f_{k-1}^i)}. \quad (3.20)$$

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) Total energy, (5) E[100,400]
<i>sonorant</i>	(1) Probability of voicing [51], (2) First order autocorrelation (3) Ratio of E[0,F3-1000] to E[F3-1000, $f_s/2$], (4) E[100,400]
<i>syllabic</i>	(1) E[640,2800] and (2) E[2000,3000] normalized by nearest syllabic peaks and dips
<i>continuant</i>	(1) Energy onset, (2) Energy offset, (3) E[0,F3-1000], (4) E[F3-1000, $f_s/2$]

Table 3.4: APs used in broad class segmentation. ZCR : zero crossing rate, f_s : sampling rate, F3 : third formant average. E[a,b] denotes energy in the frequency band [aHz,bHz]

Finally we assume

$$\prod_{k=1}^{N_{B_i}} P_t(f_k^i | f_1^i, \dots, f_{k-1}^i) = P(b_t) = \text{constant} \quad (3.21)$$

which is reasonable because given no other information all broad classes at a frame may be equally likely. We have obtained the desired result

$$P(BD|O) = P(B) \prod_{i=1}^M P(D_i|B_i) \times \prod_{i=1}^M \prod_{t=1+\sum_{j=1}^{i-1} D_j}^{D_i+\sum_{j=1}^{i-1} D_j} \prod_{k=1}^{N_{B_i}} P_t(f_k^i | x_t^{f_k^i}, f_1^i, \dots, f_{k-1}^i). \quad (3.22)$$

We have expressed $P(BD|O)$ in terms of the prior probability $P(B)$, class dependent durations $P(B_i|D_i)$ and the posterior probabilities of the manner features, that can be obtained from SVM classifiers.

3.1.4 Initial experiments and results

For preliminary experiments, one SVM was trained for each of the phonetic features and the corresponding positive and negative samples mentioned in Table 3.3. The APs in Table 3.4 were used for classification, and the results were compared to MFCCs simply to show the discriminative power of APs for the same classification task. Linear SVMs were used for the nodes *speech*, *sonorant* and *syllabic*, and Radial Basis Function (RBF) kernels were used for the feature *continuant*. Linear kernels are a good indicator of discriminative ability of the APs, and very insignificant improvements were achieved for the three features - *speech*, *sonorant* and *syllabic* - when non-linear kernels were used. For the feature *continuant*, an RBF kernel performed significantly better than the linear kernel. Table 3.5 compares the classification results for the knowledge based APs and MFCCs. Three different parameter sets were used for the MFCC based experiments - (1) MFCC_E (12 MFCCs and 1 energy), (2) MFCC_E. δ 1 (MFCC_E and the delta coefficients), and (3) MFCC_E. δ 1. δ 2 (MFCC_E. δ 1 and acceleration coefficients). Training was performed on randomly

Classifier	APs	MFCC_E	MFCC_E_δ1	MFCC_E_δ1_δ2
	≤ 5 APs	13 parameters	26 parameters	39 parameters
Silence	93.98	77.83	92.76	94.07
<i>sonorant</i>	93.01	91.63	91.65	93.39
<i>syllabic</i>	76.45	76.77	77.03	78.94
<i>continuant</i>	91.37	93.52	93.84	93.68

Table 3.5: Binary classification results. All results in percentage.

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. We can see that for silence detection and for the feature *sonorant*, the APs perform better than the 13 as well as the 26 MFCC based parameters. Performance is slightly lower but comparable to 39 MFCC based parameters in spite of the fact that the number of APs used is much less than the number of MFCCs. The results also show that there is room for improvement, and since the APs carry strong physical significance, the source of error can be easily found and more accurate APs can be developed. Moreover, APs are substantially more speaker independent than the MFCCs as was shown in [33]. Overall, a segmentation correctness of 79.8% was obtained by EBS on the segmentation task. This result is a considerable improvement over the HMM based segmentation that obtained 69.6% correctness [54]. The experiments in this project were carried out using the SVM Light toolkit [49], which provides very fast training of SVMs, and the NIST scoring package was used for scoring [55].

The comparison of previous work on feature detection is very difficult because of the different test conditions and definitions of features used by different researchers. The result on *sonorancy* feature compares well with Bitar [50] who obtained an accuracy of 94.6% for sonorancy detection on all the 'si' sentences from the TIMIT database. A more exact comparison using identical testing samples and testing conditions will be conducted, if possible, in the project. Our *continuant* result of 91.37% is an improvement over Ali's result [34] on detection of stop consonants of 86%. The high accuracy on stop detection is not surprising because it has been shown by Niyogi [23] that SVMs perform considerably well on stop detection, at least, when compared to HMM based systems. A 76.45% 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 sonorant consonants are correctly spotted. We can compare the result for the feature *syllabic* with Howitt [64] but more detailed results will have to be obtained, especially at the level of landmarks instead of frames to make a fair comparison. Such a comparison will be made in detail in the proposed work.

3.1.5 Probabilistic segmentation algorithm

We propose a Viterbi-like probabilistic segmentation algorithm that takes as input the probabilities of broad manner phonetic features - *sonorant*, *syllabic* and *continuant* - and outputs the probabilities $P(B|O)$ under the assumption of Equation 3.13. The algorithm presented here does not take probabilistic duration into account and an easy modification that allows us to do that is not presented here for brevity. Although we believe this algorithm is simpler and faster (as explained below) than the Viterbi algorithm for the problem at hand, a detailed comparative analysis of the

two algorithms will be carried out in the course of the project. The algorithm has the four steps listed below. We will denote by N the number of broad classes (five in our case) and call them b_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.

1. Location of transition points

Form a sequence of times when the probability of any of the features - *sonorant*, *syllabic* and *continuant* - changes from less than 0.5 to 0.5 or more, or vice versa. Call the set of these times $\Gamma = \{\tau_i\}_{i=1}^L$ where L is the number of such locations. Changing of the posterior probabilities in this way potentially changes the ranking of the broad classes - Fr, V, SC, ST, SIL- in terms of their posterior probabilities. The change of a broad class along a segmentation path will only be allowed at these locations which makes the algorithm more efficient than Viterbi where transitions are allowed at any locations.

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 $B^i = \{b_i\}$ and $D^i = \{\tau_1 - 1\}$. That is for each broad class, we define a path 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

$$\Pi^i = \prod_{t=1}^{\tau_1-1} P_t(b_i|O) \quad (3.23)$$

and use Equation 3.9 to evaluate $P_t(b_i|O)$.

3. Forward computation

for k from 1 to L , (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 $S' = (B', D', \Pi') = (\{B^j \cdot b_i\}, \{D^j \cdot (\tau_{k+1}) - \tau_k\}, \Pi')$, where the \cdot denotes concatenation, and

$$\Pi' = \Pi^j \prod_{t=\tau_k}^{\tau_{k+1}-1} P_t(b_i|O) \quad (3.24)$$

and again using Equation 3.9 to evaluate $P_t(b_i|O)$.

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

end loop 3

end loop 2

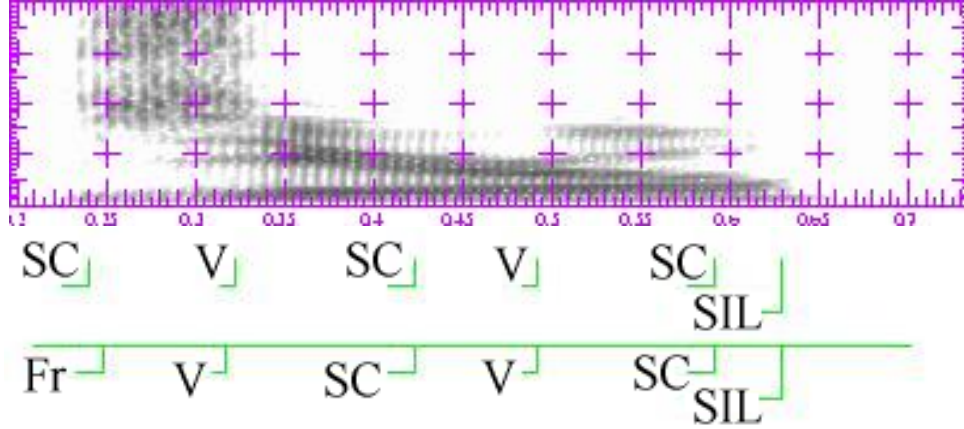


Figure 3.2: A sample output of the probabilistic segmentation algorithm for the digit 'zero'. Two most probable segmentations, SC-V-SC-V-SC and Fr-V-SC-V-SC, obtained by the probabilistic segmentation algorithm are shown in the figure.

- (c) For each path S' in S' , if another path exists with same broad class sequence and greater probability, delete the path S' from S' . This step implements the approximation in Equation 3.13
- (d) Select the N^{best} paths in S' and delete the rest of the paths in S' .
- (e) Assign $S = S'$

end loop 1

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

To impose explicit duration constraints, the most probable path among the paths that satisfy the duration constraints for a given broad class sequence is retained in step (c). The algorithm can be made more computationally efficient by allowing only the transitions to the state for which the probabilities increase instead of allowing the transitions to any possible state. We will explore the computational issues further in the proposed project. Figure 3.2 shows an example of the output of the probabilistic segmentation algorithm for an utterance 'zero' with canonical pronunciation /z I r ow/. The two most probable segmentations obtained from the algorithm - SC-V-SC-V-SC and Fr-V-SC-V-SC are shown in this figure. The correct broad class segmentation corresponding to the canonical pronunciation is Fr-V-SC-V-SC. Therefore, the segmentation obtained with the second highest probability for this case is the correct segmentation.

3.2 Detection of features from landmarks

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 $B = \{V, SC, V\}$ 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, we need to find (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 [11]. 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

$$P_2(/n/|O, L) = P(nasal|l_2, x_{l_2}^{nasal})P(alveolar|nasal, l_2, x_{l_2}^{alveolar}) \quad (3.25)$$

We have made the assumption 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, we may need landmarks from adjoining broad class segments. For example, to find the probability that the SC in a V-SC-V sequence is an /r/ we need the measurement of the third formant (F3) in the adjoining vowels because /r/ is characterized by a sharp decline in F3 relative to the adjoining vowel. Therefore,

$$P_2(/r/|O, L) = P(-nasal|l_2, x_{l_2}^{nasal})P(rhotic|-nasal, l_1, l_2, l_3, x_{l_1, l_2, l_3}^{alveolar}) \quad (3.26)$$

In general, if we represent the bundle of features below the level of broad manner phonetic features for a phoneme u_i by $\{f_{N_{B_i}+1}^i, f_{N_{B_i}+2}^i, \dots, f_{N_i}^i\}$, then, given a sequence of landmarks $L = \{l_i\}_{i=1}^M$ and the observation sequence O , we can write the conditional probability of the sequence of phonemes as

$$P(U|OL) = \prod_{i=1}^M \prod_{k=N_{B_i}+1}^{N_i} P_i(f_k^i | f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, l_{i-1}, l_i, l_{i+1}, x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i}) \quad (3.27)$$

Because many phonemes share the same features, for example, the phonemes /n/, /m/ and /ng/ share the feature *nasal*, the probabilities of all the features are not required to be computed every time the probability of a phoneme is desired. By using landmarks from adjoining manner class segments, EBS explicitly takes contextual effects into account. Therefore, a very important advantage of EBS is that there is no need to build triphone or diphone models. EBS is also more efficient because it does not analyze every frame of a segment for every phoneme.

3.2.1 Initial experiments with place and voicing feature detection

We have carried out some preliminary experiments for place and voicing features for stop consonants and fricatives. Table 3.6 shows the results on binary classification using linear SVM kernels for the TIMIT database. The APs used in this set of experiments have been obtained from [50, 56] but not all of these APs have been incorporated in the current SVM based system. These initial results are encouraging and we believe the results will improve substantially with the incorporation of temporal measures of periodicity and aperiodicity [19] as well as the rest of the APs from Bitar's work [50]. Bitar obtained an accuracy of 92.0% for the feature *anterior* and 95% for the feature *strident* on the TIMIT 'si' sentences and we get comparable values without incorporation of all the APs.

3.3 Framework for isolated and connected 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) [67], and the remaining

Classifier	Number of APs used	Accuracy (%)
Fricative <i>anterior</i> for stridents	8	91.78
<i>strident</i>	5	94.70
<i>labial/alveolar</i> for voiced stops	7	78.41
<i>labial/alveolar</i> for unvoiced stops	7	84.87
<i>voiced</i> for stops	2	83.84

Table 3.6: Place and voicing classification results. All results in percentage.

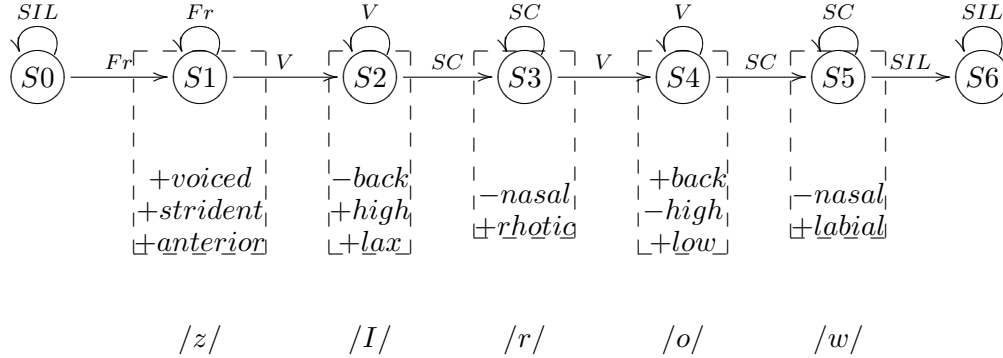


Figure 3.3: A phonetic feature based language model the word 'zero'.

phonetic features can then be estimated from the landmarks obtained from the segmentations. Figure 3.3 shows an FSA based language model for the digit 'zero' and the canonical pronunciation /z I r ow/. The broad manner class representation corresponding to the canonical representation, as mentioned before, is Fr-V-SC-V-SC. In the FSA based language model shown in Figure 3.3, one transition is made for each frame of speech, starting from the initial state S_0 , and the transition probability is equal to the posterior probability of the manner class that labels the transition. Each state corresponds to a bundle of phonetic features, for example, the state S_1 has the phonetic features $\{+voiced, +strident, +anterior\}$ apart from the phonetic features for the broad manner class Fr. Starting with the start state S_0 , 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).

To find the most probable word, the posterior probability of the most probable path among the FSAs of all the words has to be found. Mathematically, finding the best path through FSAs of allowed isolated words can be stated as

$$\hat{U}\hat{L} = \arg \max_{UL \text{ s.t. } P(U)>0} P(UL|O) = \arg \max_{UL \text{ s.t. } P(U)>0} P(L|O)P(U|OL) \quad (3.28)$$

The above mentioned method does not allow the use of a probabilistic language model $P(U)$ because of the posterior framework. Ideally, we would like to have a framework where a probabilistic language model is used and only the relevant acoustic observations are accounted for in each segmentation path.

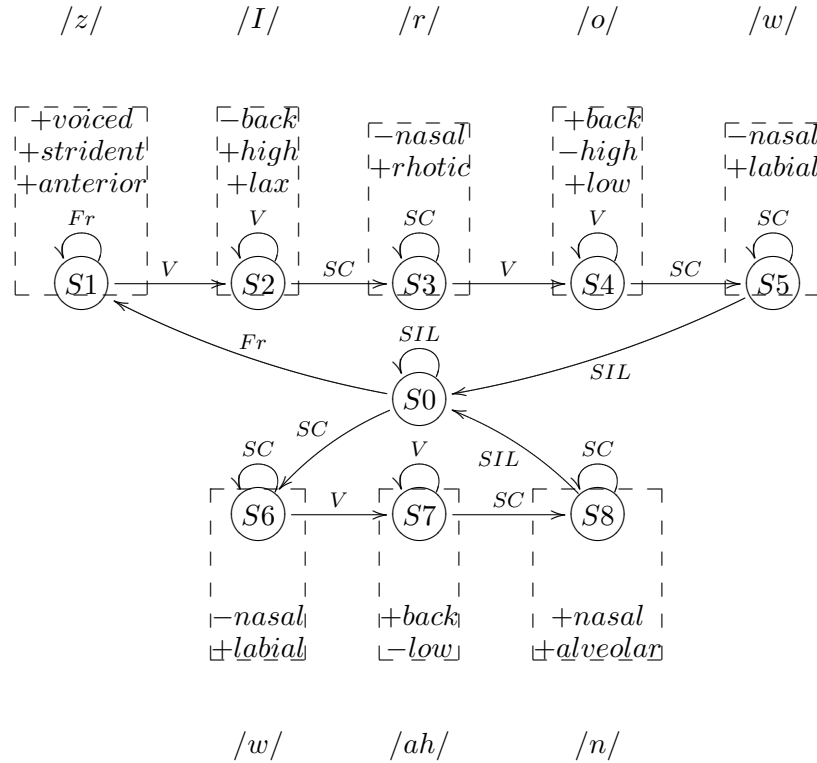


Figure 3.4: A phonetic feature based language model for continuous speech with vocabulary of two words - 'one' and 'zero'.

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. For example, with a vocabulary of two words - 'zero' and 'one', and their pronunciations /z I r ow/ and /w ah n/, respectively, the composite FSA is shown in Figure 3.4, where the FSAs of the two digits are connected by the SILENCE state S_0 . Starting with the start state S_0 , the best path among any sequence consisting of 'one' and 'zero' can be found to obtain the most probable sequence containing the two digits. We have modified the probabilistic segmentation algorithm to carry out constrained segmentation along FSAs, but we omit the description of that algorithm for brevity.

3.3.1 Evolving ideas on the use of probabilistic language model

The posterior framework we have presented for isolated and connected word recognition does not use the prior probabilities $P(U)$. We have certain ideas that may enable us to use a posterior framework along with prior probabilities with certain assumptions. Consider Equation 3.27 where we assumed, given the landmarks $\{l_{i-1}, l_i, l_{i+1}\}$ and the features of u_i below the level of current analysis feature f_k^i , that the place and voicing features of a phoneme u_i are independent of the phoneme sequence $\{u_1, \dots, u_{i-1}\}$ and the landmarks other than $\{l_{i-1}, l_i, l_{i+1}\}$. We can rewrite

Equation 3.27 without assuming this independence as (calling the sequence $\{u_1, \dots, u_{i-1}\}$ as u^{i-1})

$$P(U/OL) = \prod_{i=1}^M \prod_{k=N_{B_i}+1}^{N_i} P_i(f_k^i | f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, L, x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i}, u^{i-1}) \quad (3.29)$$

We can rewrite this equation as

$$P(U/OL) = \prod_{i=1}^M \prod_{k=N_{B_i}+1}^{N_i} P_i(f_k^i | f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, L, u^{i-1}) \frac{P(x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i} | f_k^i, f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, L, u^{i-1})}{P(x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i} | f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, L, u^{i-1})} \quad (3.30)$$

It is straightforward to see that

$$\prod_{i=1}^M \prod_{k=N_{B_i}+1}^{N_i} P_i(f_k^i | f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, L, u^{i-1}) = P(U|L) \quad (3.31)$$

Therefore, if we can pull $P(L)$ from $P(L|O)$ (this is shown in Section 3.1.3), we can get the term $P(LU) = P(U)$ and hence use the prior probabilities. But we must get rid of the term u^{i-1} from the numerator and the denominator and we must reduce L in the numerator and the denominator to the set $\{l_{i-1}, l_i, l_{i+1}\}$ for the above equation to have any practical use because it is not feasible to have a separate model for a phonetic feature for each sequence of phonemes u^{i-1} and sequence of landmarks L . Therefore, the problem is reduced to whether we can make the assumptions

$$P(x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i} | f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, L, u^{i-1}) = P(x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i} | f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, l_{i-1}, l_i, l_{i+1}) \quad (3.32)$$

and

$$P(x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i} | f_k^i, f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, L, u^{i-1}) = P(x_{l_{i-1}, l_i, l_{i+1}}^{f_k^i} | f_k^i, f_{N_{B_i}+1}^i, \dots, f_{k-1}^i, l_{i-1}, l_i, l_{i+1}) \quad (3.33)$$

or whether we can find the APs that satisfy the above constraints. We must find the APs of the place, voicing and the manner features below the level of the broad manner features (call them fine manner features), such that those APs are independent of (1) the fine manner features and the place and voicing features of the preceding phoneme sequence and (2) the landmarks other than the current and the adjoining landmarks. For example, while finding the probability of the feature *nasal* in the sequence V-SC-V, the acoustic correlates of the feature *nasal* must be independent of the narrow manner features and the place features of the adjoining vowels. This can potentially be a significant formal statement for the acoustic-phonetics community. We will investigate this issue further in the proposed project.

3.4 Project Plan

We will try to meet the following objectives in the proposed project in decreasing order in priority

1. Generalization of the probabilistic framework to the case where two or more consecutive phonemes have the same broad class representation.

2. Completion of probabilistic framework for continuous speech recognition (with probabilistic language models), if time permits. The project is mainly aimed at carrying out isolated word and connected word recognition without the use of priors.
3. Incorporation of nasal detector proposed by Pruthi and Espy-Wilson [11] (an implementation project)
4. Testing of the proposed system on various isolated and connected word databases like ISOLET (isolated alphabet) [60], OGI Alphadigits (connected alphabet and digits), TIDIGITS [63], etc.
5. Incorporation of probabilistic duration in the proposed framework, if time permits. Duration is highly dependent on speaking rate. Therefore, we believe that the use of probabilistic duration models is not very significant. Explicit rule based duration constraints can be used in the system as we have explained.

References

- [1] N. Chomsky, N. Halle, "The Sound Pattern of English", MIT Press, 1968.
- [2] C. Espy-Wilson, "A feature-based semivowel recognition system", JASA, vol. 96, pp. 65-72.
- [3] K. N. Stevens, "Implementation of a model for lexical access based on features", ICSLP 1992.
- [4] K. N. Stevens, "Toward a Model for Lexical Access Based on Acoustic Landmarks and Distinctive Features" J. Acoust. Soc. Am. (April, 2002).
- [5] A. Juneja and C Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines", IJCNN 2003, Portland, Oregon
- [6] S. A. Liu, "Landmark Detection for Distinctive Feature Based Speech Recognition", JASA 100(5), pp 3417-, November 1996.
- [7] V. W. Zue, and L. M. Lamel," An Expert Spectrogram Reader: A Knowledge-Based Approach to Speech Recognition", Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1197-1200, 1986.
- [8] J. Glass and V. Zue,"Multi-level acoustic segmentation of continuous speech", Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1988.
- [9] D. Fohr, 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.
- [10] N. Carbonell, D. Fohr, and J. P. Haton,"APHODEX, an acoustic-phonetic decoding expert system", International Journal of Pattern Recognition and Artificial Intelligence 1987.
- [11] T. Pruthi and C. Espy-Wilson,"Automatic Classification of Nasals and Semivowels", 15th International Congress of Phonetic Sciences (ICPhS) 2003, Barcelona, Spain, August 2003.
- [12] F. Jelinek, 'Continuous speech recognition by statistical methods,' Proc. IEEE. 64, No.4, pp.532-556, 1976.
- [13] J. K. Baker, "The dragon system - An overview", IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-23 (1): 24-29, February 1975
- [14] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm", IEEE Trans. Information Theory, IT-13: 260-269, April 1967.
- [15] G. D. Forney, "The Viterbi algorithm", Proc. IEEE, 61: 268-278, March 1973.
- [16] R. P. Lippmann, "Speech Recognition by machines and humans", Speech Communication 22, 1997, 1-15.
- [17] H. Fletcher and J. C. Steinberg, "Articulation testing methods", Bell Syst. Tech. J., vol 88, pp. 806-854, Oct. 1929

- [18] A. Salomon, "Speech event detection using strictly temporal information", Master's Thesis, Boston University, 2000.
- [19] O. Deshmukh, C. Espy-Wilson and A. Salomon, "Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech", submitted to IEEE Trans. on Speech and Audio Processing.
- [20] P. Mermelstein, "Automatic segmentation of speech into syllabic units", J. Acoust. Soc. Am., pp. 880-883, 58 (4), 1975.
- [21] V. Vapnik, "The Nature of Statistical Learning Theory", Springer Verlag, 1995.
- [22] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery", (1998) 121-167.
- [23] P. Niyogi, "Distinctive Feature Detection Using Support Vector Machines", pp 425-428, ICASSP 1998.
- [24] J. Keshet, D. Chazan and B. Bobrovsky, "Plosive Spotting with Margin Classifiers", Eurospeech 2001.
- [25] P. Clarkson, P. J. Moreno, "On The Use Of Support Vector Machines For Phonetic Classification", ICASSP '99. <http://citeseer.nj.nec.com/clarkson99use.html>
- [26] H. Shimodaira, K. Noma, M. Nakai, S. Sagayama, "Support Vector Machine with Dynamic Time-Alignment Kernel for Speech Recognition", Eurospeech 2001
- [27] A. Juneja and C. Espy-Wilson, "Segmentation of Continuous Speech Using Acoustic-Phonetic Parameters and Statistical Learning", in the proceedings of 9th International Conference on Neural Information Processing, Singapore, 2002, Volume 2, Page 726-730 .
- [28] M. Halle and G. N. Clements, "Problem Book in Phonology", Cambridge, MA, MIT Press, 1983.
- [29] J. B. Allen, "How do humans process and recognize speech?", IEEE Trans. on Speech and Audio Proc., 2(4):567-577, October 1994.
- [30] J. B. Allen, "From Lord Rayleigh to Shannon: How do humans decode speech?", <http://auditorymodels.org/jba/PAPERS/ICASSP> .
- [31] L. Rabiner, B. Juang, "Fundamentals of speech recognition", Prentice Hall, 1993.
- [32] HTK documentation, <http://htk.eng.cam.ac.uk/>
- [33] O. Deshmukh, C. Espy-Wilson and A. Juneja, "Acoustic-phonetic speech parameters for speaker independent speech recognition", ICASSP2002, May 13-17, 2002, Orlando, Florida
- [34] Ali, A. M. A., "Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition", Ph.D. Thesis, University of Pennsylvania, 1999.
- [35] "TIMIT Acoustic -Phonetic Continuous Speech Corpus", National Institute of Standards and Technology Speech Disc 1 -1.1, NTIS Order No. PB91 -5050651996, October 1990

- [36] V. Zue, J. Glass, M. Philips, and S. Seneff, "The MIT SUMMIT speech recognition system: A progress report", Proc. DARPA Speech and Natural Language Workshop, pp. 179-189, Philadelphia, Feb. 1989.
- [37] J. Glass, J. Chang, and M. McCandless, "A Probabilistic Framework for Feature-Based Speech Recognition," Proc. ICSLP 96, pp. 2277-2280, Philadelphia, PA, October 1996.
- [38] A. K. Halberstadt, "Heterogenous Acoustic Measurements and Multiple Classifiers for Speech Recognition", MIT Department of Electrical Engineering and Computer Science, November 1998.
- [39] J. Chang, "Near-Miss Modeling: A Segment-Based Approach to Speech Recognition", MIT Department of Electrical Engineering and Computer Science, June 1998.
- [40] S. Lee, "Probabilistic Segmentation for Segment-based Speech Recognition", M.Eng. thesis, MIT Department of Electrical Engineering and Computer Science, May 1998.
- [41] J. Chang and J. Glass, "Segmentation and modeling in segment based recognition", Eurospeech 1997, pages 1199-1202
- [42] M. Fanty, R. A. Cole and K. Roginski, "English Alphabet Recognition with Telephone Speech", Advances in Neural Information Processing Systems, 1992.
- [43] R. Cole, R. Stern, M. Phillips, S. Brill, A. Pilant, and P. Specker, "Feature-based speaker-independent recognition of isolated English letters," in Proc. ICASSP'83,, pp. 731-734, 1983. 172
- [44] Hosom, J. P., "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information", Ph.D. thesis, Oregon Graduate Institute of Science and Technology (now Oregon Health & Science University, OGI School of Science & Engineering), May 2000.
- [45] Massaro D.W., et al. "The Paradigm and the Fuzzy Logical Model of Perception Are Alive and Well." Journal of Experimental Psychology. 122 (March 1993), 115-125.
- [46] L. Deng and D. X. Sun, "A Statistical Framework for Automatic Speech Recognition Using the Atomic Units Constructed From Overlapping Articulatory Features", Journal of the Acoustical Society of America, 95:5 (May 1994), pp. 2702-2719.
- [47] E. Eide, J.R. Rohlicek. H. Gish and S. Mitter, "A linguistic feature representation of the speech waveform", Proceedings ICASSP-93 , 1993,pp.483-486
- [48] K. Kirchhoff, "Robust Speech Recognition Using Articulatory Information", PhD thesis, University of Bielefeld, Germany, July 1999
- [49] T. Joachims, "Making large -Scale SVM Learning Practical", LS8-Report 24, Universita"t Dortmund, LS VIII-Report, 1998.
- [50] N. Bitar, "Acoustic Analysis and Modelling of Speech Based on Phonetic Features", PhD thesis, Boston University, 1997

- [51] ESPS (Entropic Signal Processing System 5.3.1), Entropic Research Laboratory, <http://www.entropic.com>
- [52] J. Drish, "Obtaining Calibrated Probability Estimates from Support Vector Machines", 2001, <http://citeseer.nj.nec.com/drish01obtaining.html>
- [53] J. T. Kwok, "The evidence framework applied to support vector machines", IEEE Transactions on Neural Networks, 11(5):1162-1173, September 2000.
- [54] HMM experiments carried out at Speech Communication Lab by Om Deshmukh, <http://www.ece.umd.edu/omdesh/iconip2002.html>
- [55] Speech Recognition Scoring Package (SCORE) Version 3.6.2, <http://www.nist.gov/speech/tools/>
- [56] A. Juneja and C. Espy-Wilson, "An Event-Based Acoustic-Phonetic Approach for Speech Segmentation and E-Set Recognition", ICPhS 2003, Barcelona, Spain.
- [57] A.M.A. Ali, J. V. Spiegel and P. Mueller, "An Acoustic-Phonetic Feature-based System for the Automatic Recognition of Fricative Consonants", ICASSP-98, vol. II, pp. 961-964, 1998.
- [58] A. M. A. Ali, J. V. Spiegel and P. Mueller, "Automatic Detection and Classification of Stop Consonants using an Acoustic-Phonetic Feature-Based System", XIVth International Congress of Phonetic Sciences, pp. 1709-1712, 1999.
- [59] K.S. Stevens, S. Manuel, and M. Matthies, "Revisiting place of articulation measures for stop consonants: Implications for models of consonant production", Proceedings of the International Congress of Phonetic Sciences (1999).
- [60] ISOLET, Release Version 1.3 (19 August 2002), Center for Spoken Language Understanding, <http://cslu.cse.ogi.edu/corpora/isolet/version.html>
- [61] Alphadigit, Release Version 1.3 (23 August 2002), Center for Spoken Language Understanding, <http://cslu.cse.ogi.edu/corpora/alphadigit>
- [62] A. Ganapathiraju, "Support Vector Machines for Speech Recognition", Ph.D. Dissertation, Department of Electrical and Computer Engineering, Mississippi State University, January 2002.
- [63] Leonard et.al., "A speaker-independent connected-digit database", <http://www ldc.upenn.edu/Catalog/docs/LDC93S10/>
- [64] A. W. Howitt, "Automatic syllable detection for vowel landmarks", PhD thesis, MIT, July 2000.
- [65] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing", J. of Phonetics, 16, pp. 55-76, 1988
- [66] R. Chun, "A hierarchical feature representation for phonetic classification", M. Eng Thesis, MIT, 1996.

- [67] D. Jurafsky and J. H. Martin, "Speech and Language Processing", Prentice Hall, New Jersey 2000

A American English Phonemes

	SYMBOL	EXAMPLE WORD	POSSIBLE PHONETIC TRANSCRIPTION
Stops	b d g p t k dx q	bee day gay pea tea key muddy, dirty bat	BCL B iy DCL D ey GCL G ey PCL P iy TCL T iy KCL K iy m ah DX iy, dcl d er DX iy bcl b ae Q
Affricates	jh ch	joke choke	DCL JH ow kcl k TCL CH ow kcl k
Fricatives	s sh z zh f th v dh	sea she zone azure fin thin van then	S iy SH iy Z ow n ae ZH er F ih n TH ih n V ae n DH e n
Nasals	m n ng nx	mom noon sing winner	M aa M N uw N s ih NG w ih NX axr
Semivowels and Glides	l r w y hh hv	lay ray way yacht hay ahead	L ey R ey W ey Y aa tcl t HH ey ax HV eh dcl d
Syllabic consonants	em en eng el	bottom button washington bottle	b aa tcl t EM b ah q EN w aa sh ENG tcl t ax n bcl b aa tcl t EL
Vowels	iy ih	beet bit	bcl b IY tcl t bcl b IH tcl t

continued on next page

continued from previous page			
	SYMBOL	EXAMPLE WORD	POSSIBLE PHONETIC TRANSCRIPTION
	eh	bet	bcl b EH tcl t
	ey	bait	bcl b EY tcl t
	ae	bat	bcl b AE tcl t
	aa	bott	bcl b AA tcl t
	aw	bout	bcl b AW tcl t
	ay	bite	bcl b AY tcl t
	ah	but	bcl b AH tcl t
	ao	bought	bcl b AO tcl t
	oy	boy	bcl b OY
	ow	boat	bcl b OW tcl t
	uh	book	bcl b UH kcl k
	uw	boot	bcl b UW tcl t
	ux	toot	bcl t UX tcl t
	er	bird	bcl b ER dcl d
	ax	about	AX bcl b aw tcl t
	ix	debit	dcl d eh bcl b IX tcl t
	axr	butter	bcl b ah dx AXR
	ax-h	suspect	s AX-H s pcl p eh kcl k tcl t

B Tables of place and voicing features

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

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

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

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

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

Table B.3: The place features for vowels

C Support Vector Machines

C.1 Structural Risk Minimization (SRM)

SVMs [21, 22] are learning machines for pattern classification and regression tasks based on the principle of structural risk minimization [21]. 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 class labels $\{y_i\}_{i=1}^l$ are produced by a joint probability distribution $P(\mathbf{x}, y)$. For a possible function $f(\mathbf{x}, \alpha)$ that attempts to find the class labels for given vector \mathbf{x} , the expected risk of the function is defined as

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

With a probability η ($0 \leq \eta \leq 1$), the following bound on the expected risk exists [21],

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h) + 1) - \log(\eta/4)}{l}} \quad (\text{C.2})$$

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

$$R_{emp}(\alpha) = \frac{1}{2l} \sum_{i=1}^l |y_i - f(\mathbf{x}_i, \alpha)|. \quad (\text{C.3})$$

The VC dimension h depends on the class of functions $f(\mathbf{x}, \alpha)$ and the empirical risk is defined for 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.

C.2 SVMs

SVMs are maximum margin classifiers. Figure C.1 illustrates the difference between large margin classifiers and small margin classifiers. For linearly separable data, the goal of SVM training for two class pattern recognition is to find a hyperplane

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (\text{C.4})$$

such that the margin $2/\|\mathbf{w}\|$ between the closest training samples with opposite labels is maximized. It is easy to see in Figure C.1 that the classifier in (b) is more robust to noise because a larger amount of noise is required to let a sample point cross a decision boundary. It has been argued by Vapnik [21] that maximization of margin leads to minimization of VC dimension, but no concrete proof exists that SVM training carry out SRM [22]. In general, SVMs select a set of support vectors

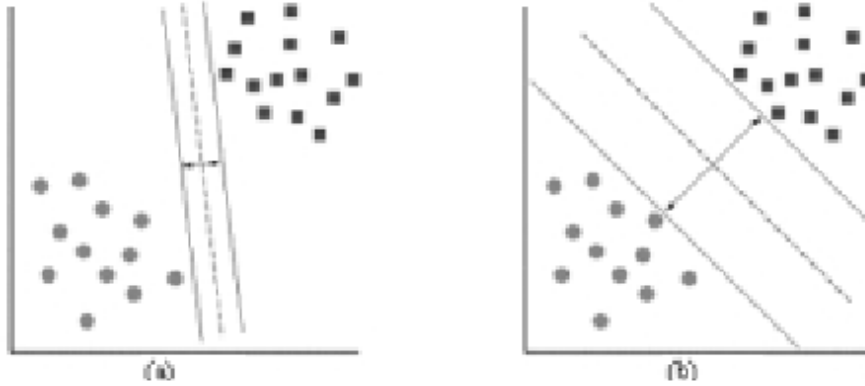


Figure C.1: (a) small margin classifiers, (b) maximum margin classifiers

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

$$f(\mathbf{x}) = \sum_{i=1}^{N_{SV}} y_i \alpha_i K(\mathbf{x}_i^{SV}, \mathbf{x}) - b. \quad (\text{C.5})$$

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 [22]. 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 C.5. Two commonly used kernels are radial basis function (RBF) kernel and linear kernel. For RBF kernel,

$$K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma|\mathbf{x}_i - \mathbf{x}|^2) \quad (\text{C.6})$$

where the parameter γ is usually chosen empirically by cross-validation from the training data. For the linear kernel,

$$K(\mathbf{x}_i, \mathbf{x}) = \mathbf{x}_i \cdot \mathbf{x} + 1 \quad (\text{C.7})$$