A Study on Acoustic-Phonetic Boundary Detection in Continuous Speech

A THESIS

submitted by

Venkatesh Keri

for the award of the degree

of

Master Of Science (by Research)
in
Computer Science & Engineering

LANGUAGE TECHNOLOGIES RESEARCH CENTER

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD - 500 032, INDIA

May 2011

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

This is to certify that the work contained in this thesis titled **A Study on Acoustic-Phonetic Boundary Detection in Continuous Speech** submitted by **Venkatesh Keri** for the award of the degree of Master of Science (by Research) in Computer Science & Engineering is a bonafide record of research work carried out by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Date

Dr. Kishore Prahallad

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to Dr. Kishore Prahallad, my advisor for his guidance, encouragement and support throughout my duration as an MS student at IIIT-H.

I express my sincere gratitude to Prof. Raj Reddy and Prof. B.Yegnanarayana for their valuable advices and inputs during this research.

I am grateful to Prof. Rajeev Sangal, Director of Language Technologies Research Center, IIIT-H for providing an excellent environment for work with ample facilities and academic freedom.

I am very grateful for having had the opportunity to study among my colleagues: Sachin Joshi, Santhosh Yuvaraj, Anand Arokia Raj, Satish Chandra Pammi, Gopalakrishna, Vijayaditya, Gautam Mantena, Ramakrishna Raju, Lakshmikanth and Bhaskar. In particular, I thank Veera Raghavendra and Srinivas Desai - for all the support, fruitful discussions and fun times together.

Needless to mention that without the love and moral support of my family, this work would not have been possible.

Venkatesh Keri

ABSTRACT

Keywords: Unconstrained acoustic-phonetic boundary detection, discriminative classification, Artificial Neural Networks, excitation based features.

Acoustic-phonetic speech segmentation is the process of detecting the acoustic-phonetic boundaries and labeling each segment with a phonetic symbol in the spoken utterance.

In the scope of this thesis, we address issues related to acoustic-phonetic boundary detection. The first part of this thesis is focused on the different approaches for acoustic-phonetic boundary detection. Previous works indicate that, acoustic-phonetic boundary detection approaches can be broadly classified into two types, i.e., signal processing based approaches and classification based approaches. Signal processing based approaches typically rely on peak-picking algorithms on temporal trajectories of signal energy, sub-band energy etc.,. So, these approaches can operate independent of language and transcription. However, signal processing approaches are sensitive to parameters used and are less robust than supervised classifiers. On the other hand, supervised classifiers require hand labeled data to train classification models. Apart from the above approaches, there are also approaches which combine both acoustic-phonetic boundary detection and acoustic-phonetic segment labeling using hidden Markov model (HMM) forced alignment based approach which can be termed as constrained acoustic-phonetic speech segmentation approach. It should be noted that, signal processing based approaches are found to be highly sensitive to the parameters used and hence are less robust. Supervised classifier based approaches are limited by requirement of large amount of manually segmented data to train the classifier. Constrained acoustic-phonetic speech segmentation approaches are limited by requirement of a good transcription for both training and testing, which is very difficult to obtain. In order to over come these limitations, we are proposing a unsupervised discriminative classifier approach which does not require manually segmented data and phonetic transcription. The input to the discriminative classifier is obtained from boundaries automatically detected by signal processing based boundary detection approaches. We show that unsupervised classification based approaches perform better than signal processing based approaches and HMM based approaches.

The second part of the thesis focuses on significance of linear prediction residual for acoustic-phonetic boundary detection in continuous speech. Alternate features extracted from LP residual (excitation based features) are explored to improve the performance of boundary detection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Segmentation of speech signal

Speech segmentation can be described as a process of identifying boundaries in the speech signal and labeling each of the speech segments between two adjacent boundaries with a symbol. This process of identifying boundaries and labeling, can be addressed at various levels of details such as:

- Speech / non-speech segmentation: This is a task of detecting the begin and end of speech in the audio signal. Apart from speech, an audio signal may also contain non-speech data such as music, noise, silence etc. It is important to segment the audio data into speech / non-speech segments, as it acts as pre-processing step for many speech systems such as speech recognition, speaker recognition, etc.,. Output of this process is the boundaries between the speech and non-speech segments in an audio signal.

- Phrase segmentation: Phrase is a group of words functioning as a single unit in a sentence. This process is the task of detecting acoustic phrases in a speech signal. It is often observed that pauses in the speech signal are good indicators of acoustic phrasal boundaries. However, not all acoustic pauses correspond to acoustic phrasal boundaries. Hence, the task here is to detect acoustic pauses in a spoken utterance which correspond to phrasal boundaries.

- Voiced / unvoiced segmentation: This is a task of identifying the voiced (quasi-periodic signal) and unvoiced (non-periodic signal) regions in the speech signal. The task here is to detect the boundaries between voiced and unvoiced segments and label these segments as voiced or unvoiced.

- Word segmentation: Word is one of the smallest meaningful unit, comprising of sequence of phones or syllables. It is a task of detecting the boundaries of a word in a speech signal.
- Syllable segmentation: Syllable is a unit of organization for a sequence of phonemes which should at least have one voiced phone. It is a task of detecting the boundaries of syllables in a speech signal and labeling these speech segments with its respective syllables.
- Acoustic-Phonetic segmentation: Phone is a smallest, acoustic unit of pronunciation. It is the link between the speech signal which is continuous and the phoneme which is just a discrete, abstract cognitive concept bound by language constraints. Acoustic-phonetic segmentation is a task of detecting the boundaries of phones in a speech signal and labeling those speech segments with phones.

1.1 Acoustic-phonetic segmentation

In order to perform acoustic-phonetic segmentation of speech automatically using machine, we have to first understand the processes involved in manual phonetic segmentation of speech done by humans. As described in motor theory by Liberman et. al. [4] and multi-cue model by [5], human beings are capable of recognizing phones using only printed spectrograms that display the speech signal along the dimensions of time, frequency and amplitude axes. This process of manually segmenting the speech into acoustic-phonetic units by humans is termed as *human spectrogram reading.* These techniques employed in spectrogram reading are useful in designing the algorithms for machines to perform speech segmentation. The seminal paper on spectrogram reading [29] describes the manual segmentation approach of Victor Zue, and the analysis of his approaches. It was found

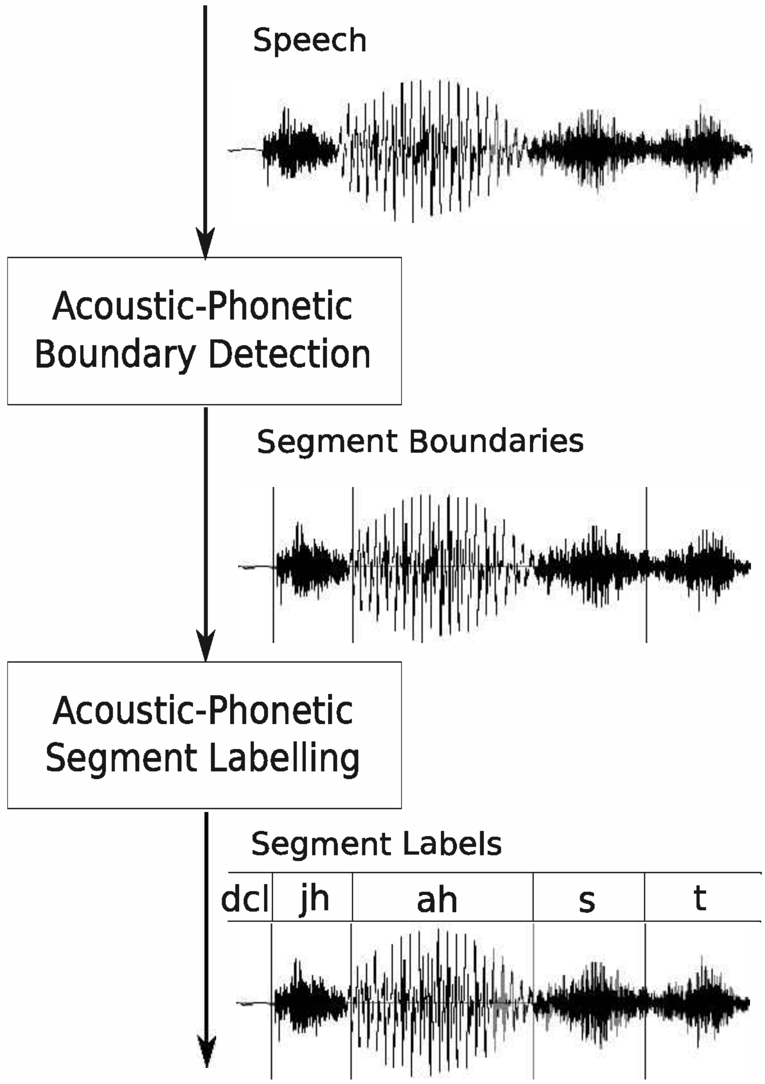
Speech
Acoustic-Phonetic Boundary Detection
Segment Boundaries
Acoustic-Phonetic Segment Labelling
Segment Labels
dcl
jh
ah
s
t

Figure 1.1: *Block Diagram of acoustic-phonetic segmentation of speech.*

that he was able to identify more than 97% of all phonetic segments in continuous speech. Victor Zue used a two-pass method for reading spectrograms; in the first pass, he identified the boundaries of acoustic-phonetic segments and in the second pass he identified the phones in each segment with some segment boundary adjustments. Boundary identification was done primarily by locating the points of spectral change and shape changes in intensity. Some boundaries were determined based on relative local duration; for example, two adjacent stop closures can be identified as two phones even without change in a spectral information, because the combination of both stop closure is significantly longer than the duration of a single stop closure. Changes in formant frequencies (such as dip in the first frequency or a decrease in formant amplitude) were used by him to identify transition within a sonarent region. In some cases, such as liquid-vowel transitions, no boundary was marked until the second pass. Once the speech has been initially segmented, each segment was assigned a phonetic label. Label assignment was done based on (a) knowledge of unique spectral patterns for a phone, (b) knowledge of co-articulatory effects, and (c) constraints imposed by English phonology. Even for highly complex sounds such as plosives, he was able to identify and classify plosives with great accuracy based on the characteristic patterns of the manner and place of articulation.

Just like manual acoustic-phonetic speech segmentation performed by Victor Zue, automatic acoustic-phonetic speech segmentation without using phonetic transcription, can also be divided into two phases:

- Acoustic-phonetic boundary detection: In this phase, the boundaries of acoustic-phonetic units in speech signal are automatically detected. This phase is independent of language, phonetic transcription and phone-set.

- Acoustic-phonetic segment labeling: In this phase, each segment is labeled with a phone symbol. The choice of the phone set may be dependent on language and the text transcription if available for the utterance.

1.2 Issues in acoustic-phonetic boundary detection

In this thesis, our work is mainly focused on the first phase of acoustic-phonetic speech segmentation, i.e., approaches for acoustic-phonetic boundary detection in a speech signal without using any phonetic transcription. The issues involved in this process are described below.

1. **Effect of phone duration variability on acoustic-phonetic boundary detection:** At the physical level, the rate of speech is governed by the inertia of the articulators. The body of the tongue moves relatively slowly, and the rate of sonarent phones is limited by the rate at which the tongue moves. The lips and tongue can move faster and so plosive sounds occur over a much shorter time interval. In addition to durational variation due to phone differences, vowel duration may change by a factor of eight, depending on speaking rate, syntax and stress. The factors that influence the phone duration while speech production and speech perception result in fairly complex models. A preliminary model proposed by Klatt for speech synthesis had seven factors that influenced the duration structure of sentence and these factors were accounted by eight rules. A simpler model proposed by Van Santen is able to account for 86% of the variance of vowel durations only in a large corpus of manually segmented speech [6]. This concludes that phone duration structure in different conditions of speech is highly complex and difficult to understand and build models.

2. **Effect of co-articulation on acoustic-phonetic boundary detection:**

   Co-articulation is the effect that one phone has on its neighboring phone, which is manifested as a smooth change in formant frequencies from one phone to the next as shown in Fig 1.2. This smooth transition between phones is one of the main factors that makes it difficult to determine the exact location of phonetic boundary. Ohman [7], proposed a model to handle co articulation in VCV utterances using a vocal tract shape based information. Even though this model was successful on VCV utterances, it had its limitations in handle co-articulation effects in CVC utterances. In the model proposed by Lofqvist, speech segments have a overlapping "dominance functions" that control the articulators, with one dominance function per articulator. These dominance functions can differ in time offset, duration and

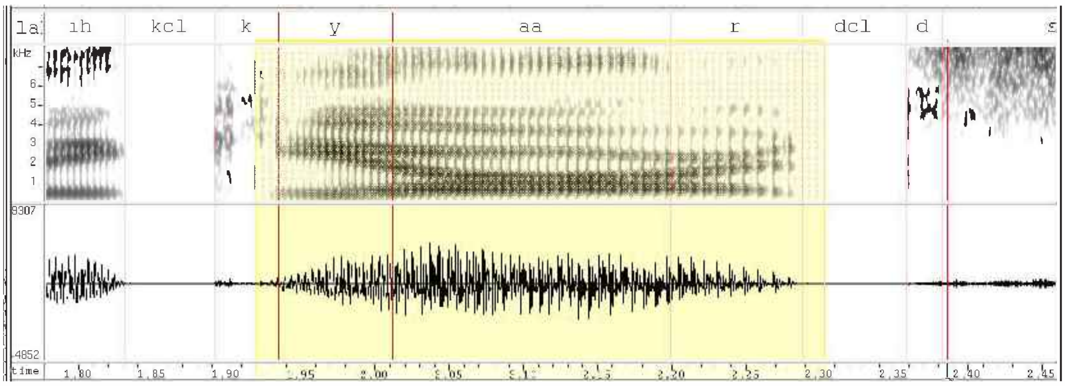

Figure 1.2: *The co articulation effect on two boundaries in the word "yard" i.e., boundary between /y/ & /aa/ and /aa/ & /r/ can be observed in shaded region of the above spectrogram.*

magnitude, giving relatively more or less weight to articulators associated with a given speech segment. Although this model is successful in modeling visual speech, it is not obvious how this model could be used directly in current speech segmentation systems, in which the articulators are the best parameters that one can use. There were several such studies and they concluded that co-articulatory patterns are not explained adequately by any of the theories or models. This concludes that co-articulation is highly complex and difficult to understand and build models.

3. **Significance of source features for acoustic-phonetic boundary detection:** As described in production and perception models, present techniques use acoustic features like vocal-tract or perceptual based features only throwing away the other information in the speech signal. So, as described in Victor Zue's analysis, new features have to be explored in-order to improve the performance of acoustic-phonetic boundary detection.

1.3 Issues addressed in this thesis

The focus of this thesis is acoustic-phonetic boundary detection in a speech signal. These approaches mainly rely on the acoustics only and is not constrained by the phonetic transcription of an utterance, and hence the task is to find the acoustic event in the speech signal that marks the change in the phonetic boundaries. These approaches can be broadly classified into supervised and unsupervised approaches. These methods are thus language independent.

Unsupervised approaches such as [8], [9] use some form of peak-picking algorithm to detect the acoustic-phonetic boundaries. This does not have any training phase and hence no training data is required. Unlike unsupervised approaches, supervised approaches such as [10] a discriminative classifier, a model is trained using some training data and is used to detect the boundary frames of a test speech signal without using any phonetic transcription. Following are different modules that are addressed in this work:

1. **A supervised discriminative classifier approach** Youngjoo Suh and Youngjik Lee, proposed a discriminative classifier approach using multi-layer perceptron on a single speaker database, which consisted of three phases: The pre-processor utilizes a sequence of 44 order feature parameters for each frame of speech and manually labeled data to prepare the input data to the next phase. Multi-layer perceptron (MLP) has an input layer with 176 nodes, one hidden layer and an output layer with one node and the value at the output node gives an estimate of the phonetic boundary for the present frame. Post-processor decides the positions of phonetic boundaries using MLP output value. In Suh and Lee [10], this approach was applied on a single speaker Korean database. In the present work, we adapt a similar approach to a multi-speaker database, and compare it with signal processing and forced-alignment based approaches for acoustic-phonetic boundary detection.

2. **Comparison of different approaches** Other than the supervised/unsupervised acoustic-phonetic boundary detection approaches, there are some other approaches which can be used for boundary detection, but require phonetic transcription along with the speech signal to perform the task. HMM, DTW based speech segmentation approaches are some of such approaches which segment the speech signal into phonetic segments, hence obtaining the acoustic-phonetic boundaries. In order to check the performance of the above adapted approach [11], a comparative study on different approaches such as unsupervised group-delay function (GDF) based boundary detection approach and HMM based speech segmentation approach and the results are provided in chapter 3.

3. **Unsupervised discriminative classifier approach** As the approach proposed by Youngjoo Suh and Youngjik Lee [10] is a supervised approach, it requires manually labeled data to train the MLP classifier. But it is difficult and tedious to obtain a precise and sufficient amount of data to train. Hence we propose an unsupervised approach, to train the discriminative classifier model to perform the acoustic-phonetic boundary detection. Initially, an unsupervised peak-picking approach is applied on the training data to obtain an initial acoustic-phonetic boundaries. Discriminative artificial neural networks (ANN) classifier was trained using automatically obtained boundaries by signal processing approaches.

4. **Significance of excitation features for acoustic-phonetic boundary detection**

   Present state-of-art approaches mostly use perceptually motivated features like MFCC or Filter based features such as LPCC or filter-bank based approaches which basically try to model the formant tracking information. But these features fail at some situations such as boundaries between unvoiced stop bursts / unvoiced fricatives (eg., /k-s/), two unvoiced fricatives (eg.,

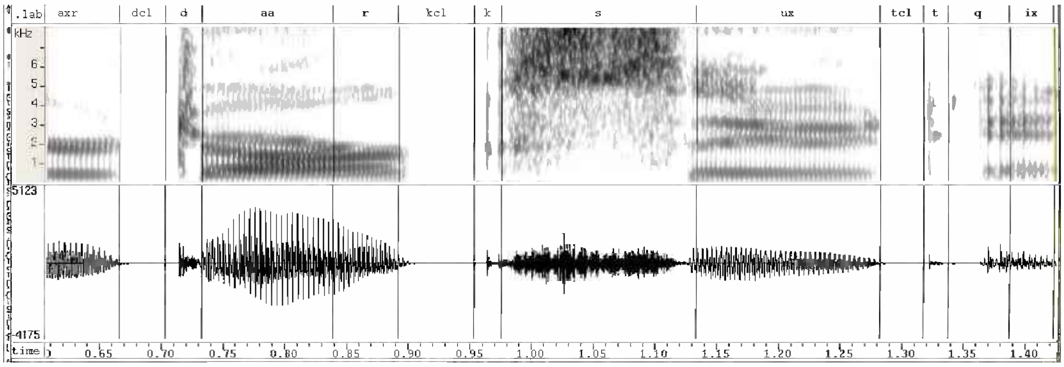

Figure 1.3: *An example of boundary where either of the adjacent phones do not have formant tracks such as between /k/ and /s/, where perceptual and vocal-tract based features fail to detect the boundary.*

/ch-s/) etc., where there is no formant track information at all as shown in Fig 1.3. So, there is a need to explore different features which are independent of formant tracking for the task of speech segmentation. Most of the presently used acoustic features try to capture perceptual or vocal-tract characteristics completely throwing away the source features like LP residual based features. Even though, Markel et. al [12] shows that LP residual spectrum is almost flat, some of the previous works on speaker verification [13], speech coding etc., showed that LP residual has potential to improve the performance in their respective works. This motivated us to explore LP residual based features for the task of acoustic-phonetic boundary detection in speech signal.

1.4 Contributions

The contributions of this thesis can be summarized as follows:

- A framework for supervised and unsupervised ANN based, acoustic-phonetic boundary detection approach is developed.

- A comparative study between some of the start-of-art approaches which fall under different categories such as supervised / unsupervised, force-aligned etc in comparison with the above described approaches are reported. This study showed that, supervised ANN based approach performed almost as good as supervised, forced-aligned approaches.

- Significance of source characteristics in speech for the task of acoustic-phonetic boundary detection in speech signal is analised. This analysis helped in proposing that, excitation based features can be used as complimentary evidence along with filter based features to obtain a better segmentation performance.

1.5 Organization of thesis

The rest of the thesis is organized as follows:

In **chapter 2**, an overview of different acoustic-phonetic boundary detection approaches such as manual and automatic approaches with their limitations were described in detail. Different automatic acoustic-phonetic boundary detection approaches, forced-alignment based approaches, both supervised and unsupervised approaches were discussed in detail.

In **chapter 3**, a detailed description of the supervised classification based acoustic-phonetic boundary detection and the proposed unsupervised approach were described. This chapter ends with the comparison and analysis of the different baseline approaches with the above two approaches is reported.

In **chapter 4**, we have described the motivation and the procedure for exploring the excitation based features for the task of acoustic-phonetic boundary detection. This chapter concludes with the comparison and analysis of different features such as LPCC and HECC using supervised ANN based acoustic-phonetic boundary detection approach.

Finally in **chapter 5**, the conclusions that can be drawn from the thesis are outlined along with the limitations of the work in the thesis and the possible directions for future work.

CHAPTER 2

Approaches for acoustic-phonetic boundary detection

In this chapter, we will report different approaches for acoustic-phonetic boundary detection in continuous speech. Broadly, boundary detection in continuous speech can be classified into manual and automatic approaches based on the amount of human intervention in doing the task. If the task is performed solely by the human annotators, then it is called as *manual acoustic-phonetic boundary detection.* On the other hand if the task is performed automatically performed by machines without any human intervention, then it is called as *automatic acoustic-phonetic boundary detection.*

In this review chapter, several approaches for manual and automatic acoustic-phonetic boundary detection are described along with their limitations. Finally, we will describe the need for new approaches and features for acoustic-phonetic boundary detection.

2.1 Manual acoustic-phonetic boundary detection

Manual acoustic-phonetic boundary detection is a process of identifying the phonetic boundaries in speech by manually examining cues from spectrogram, energy and pitch. This process is described as human spectrogram reading. It is often observed that, no two human annotators can identify the phonetic boundaries exactly same. As a result, manual speech segmentation is usually reported as inter-labeler agreement, with one set of manual boundaries chosen as nominally correct, and the other set of boundaries measured in relation to the first set. Following are a few studies on manual boundary detection:

Cosi et al. [14] reported a manual segmentation performance of about 6 msec mean deviation, 55% agreement within 5 msec, and 93.5% agreement within 20 msec on 10 Italian continuous speech utterances sampled at 16 kHz sampling rate.

Ljolije et al. [15] evaluated a manual segmentation task on 100 Italian utterances from two human transcribers and found 80.8%, 92.9% and 96.8% agreement within 10, 20 and 30 msec respectively.

Wesenick and Kipp [16] reported an average agreement levels of 63%, 73%, 87%, and 96% within 0 msec (perfect correspondence), 5 msec, 10 msec and 20 msec respectively. Annotators used in this study were all graduate students in phonetics, and all had received an intensive training session. As a part of this training, a number of conventions were established to ensure consistency among the annotators. One such rule was to always set a segmentation boundary where the values of speech changed from negative to positive. This is one of the reasons for the best performance reported for human consistency on the task of phonetic segmentation.

Leung and Zue [17] reported an agreement of 80%, 87% and 93% within 10 msec, 15 msec and 20 msec on a database of five phonetically balanced sentences, recorded at 16 kHz and annotated by two human annotators.

Cole et al. [18] reported an inter-annotator agreement for five languages i.e., English, German, Mandarin, Spanish and Hindi from OGI Multi-lingual speech corpus. There were two major changes when compared to previous works: (a) It was done on 8 kHz telephone-band speech to check the channel effects. Inter-annotator agreement by native US English annotator on US English data was found to be 79% within 10 msec which is marginally lower than the value reported by Leung. Hence showing that the channel conditions have minimal effect on manual segmentation. (b) A part of the annotation was performed by non-native annotators to check the annotation consistency between native and non-native annotators. To perform this analysis, a German database was annotated by two native and two non-native annotators. The inter-annotator agreement between native annotators was about 63% and 79% agreement within 5 msec and 10 msec respectively, and between two non-native annotators it was about 69% and 81% within in 5 msec and 10 msec which is comparable to the former.

Hosom et al. [3] reported an inter-annotator agreement of about 81.7% and 93.5% agreement within 10 msec and 20 msec on 50 TIMIT sentences annotated by two annotators. For evaluation, they (a) merged glottalized sounds such as /q/ (a glottalized sounds are the sounds produced with some important event (a movement or a closure) of the glottis.) (b) did not evaluate boundaries between stop closures and silences (as any such boundary is placed arbitrarily). These results correspond well with the previous works by Cosi, Ljolije, Leung and Cole.

In order to check how consistency of annotation by the same annotator and to measure properly how fast it is possible to work, the manual segmentation and labeling of about two minutes of the speech data was performed by Kvale [19]. Manual annotation was done two times by the same annotator with a gap of three months between the two annotations. Kvale reported that the agreement between the two sets of segment boundaries was 63% and 96.5% within 5 msec and 20 msec respectively and there were very minimal labeling errors of about 0.5%. On the other hand, it took 135 minutes to segment and label 748 segments, i,e. 5.5 phones per minute. A summary of performances reported by the above works is shown in Table 2.1.

In summary, there is a fairly consistent agreement among human annotators for continuous speech, even across language and channel conditions. Agreement between two annotators and between two annotations of the same annotator are

Table 2.1: *Comparison of different manual acoustic-phonetic boundary detection performances.*

Previous	Language	No. of	No. of	Agreement % with $\tau(ms) \leq$						
Works	(samp. rate kHz)	Utts.	Annotators	0	5	10	15	20	25	30
Leung [17]	English (16)	5	2	-	-	80	87	93	-	-
Cosi [14]	Italian (16)	10	3	-	55	-	-	94	-	-
Wesenick [16]	German (16)	64	3	63	73	87	93	96	-	99
Ljolije [15]	Italian (16)	100	2	-	-	81	-	93	-	97
Cole [18]	English (8)	10	2	-	67	79	-	-	-	-
	Mandarin (8)	10	2	-	71	83	-	-	-	-
	Spanish (8)	10	2	-	53	71	-	-	-	-
	German (8)	10	2	-	63	79	-	-	-	-
	German (8)	10	2 (NN)	-	68	81	-	-	-	-
	Hindi (8)	10	2 (NN)	-	67	79	-	-	-	-
Hosom [3]	English (16)	50	2	-	61	82	89	94	95	97
Kvale [19]	English (16)	2 min	1	-	63	88	94	97	98	-

almost similar (but not same). This highlights the point that it is not possible to obtain perfect annotation, even by the same annotator. Hence this shows that manual boundary detection is a tough and time consuming job. There is an average agreement of 94.17% within 20 msec for the measured manual agreements , with a maximum of 97% within 20 msec for highly trained specialists using a set of rigorous and well defined conventions. One can obtain highly accurate phonetic segmentation using manual annotation, but it has many limitations as described below.

2.1.1 Limitations

- The manual intervention in this process is extremely time consuming and tedious which drastically increases the time required to obtain the phonetic boundaries.
- As mentioned by Cosi et al. [14], Torkkola [20] and Van Erp and Bover [21], it is almost impossible to reproduce the manual segmentation results, due to

the variability of human visual and acoustical perceptual capabilities. Also, it is difficult of establish a clear common labeling strategy. Consequently, the manual segmentation procedure is implicitly inconsistent.

- This approach has to be done by highly trained human annotators who are difficult and costly to obtain.
- If a new database have to be segmented, then all the processes of this approach have to be repeated by human annotators.

2.2 Automatic acoustic-phonetic boundary detection approaches

As mentioned earlier, if the acoustic-phonetic boundary detection is performed automatically by a machine without any human intervention, then it is called as *automatic phonetic speech segmentation*. Boundary detection approaches can be broadly classified into three types. They are signal processing based approaches, classifier based approaches and force-alignment based approaches. This section deals with different types of approaches for automatic acoustic-phonetic boundary detection in continuous speech and provides an overview of some of the previous works in each of these types.

2.2.1 Signal processing based approaches

These approaches generally use the combination of signal processing techniques and peak-picking methods to perform the task of acoustic-phonetic boundary detection. Mostly these approaches fall in unsupervised category as they don't require any manually labeled data. Some of these approaches are discussed below:

Wilpon et al. [22] has proposed an unconstrained segmentation procedure which is based on measuring the spectral variations over time by computing the spectral variation contour. First step in obtaining spectral variation contour is by computing the distortion between consecutive frames on some spectral representations such as linear prediction coefficients (LPC) or cepstral coefficients using Itakura-Saito distance or Euclidean distance. Second, spectral variation contour is obtained by computing the stationarity of the signal at each frame as the average of the distortion over the frames around it.

Sharma and Mammone [23], proposed a "Blind" speech segmentation procedure which allows a speech signal to be segmented into sub-word units without the knowledge of any transcription. This procedure involves two phases. First is to finding the optimal number sub-word segments in the speech signal. Second is to find the optimal location of sub-word segment boundaries. So, initially they estimate the range of the number of segments $[K_{min}, K_{max}]$. Number of syllables in a speech segment was considered as minimum number of segments in speech signal (K_{min}). Number of syllables in speech signal was found by using **Convex Hull method**. Maximum number of segments in speech signal (K_{max}) is estimated by using a spectral variation function (SVF). Then for every k in this range, they have performed level building dynamic programming (LBDP) based speech segmentation for k-segments and computed k-cluster optimality criterion (Q_k) using normal decomposition method. The optimal number of sub words K_0 is found as: $K_0 = argmax_{k=Kmin}^{k=Kmax} Q_k$. Finally, segment boundaries are obtained by performing DP-based segmentation for K_0 sub words.

Anna Esposito and Guido Aversano [9], proposed a segmentation algorithm, that carries out the phone-level segmentation using a bottom-up approach. This is based on the detection of spectral instability in multiple frequency bands. According to results reported in this work, the algorithm gives better performance than other methods of the same class of approaches. A part of their work was on the analysis of the relations between segmentation performance and their dependence from the algorithm's parameters. They found that these approaches are very much dependent on the parameters for better performance.

Dusan and Rabiner [24], developed an approach which extracts 10 dimensional mel-filter cepstral coefficients (MFCC) for every 10 msec and computes a spectral transition measure (STM) to capture the spectral rate of change in time. Since the spectral rate of change usually displays peaks at the transition between phones, such a measure is used to detect the phone boundaries. Once STM is obtained at every frame, next step is to detect the boundaries using peak picking method and remove the spurious peaks using post-processing methods. They reported their analysis on TIMIT train-set and the algorithm has deleted about 15.4% and inserted about 28.2% of boundaries. They also reported that of the correctly detected boundaries i.e., 84.6%(100 - 15.4), there was an agreement of 70%, 89% and 95% with in 10 msec, 20 msec and 30 msec respectively.

Estevan et al. [25], proposed an unsupervised maximum marginal clustering (MMC) based approach for phonetic speech segmentation. MMC algorithm is applied on a sliding window of width 18 mel-filter cepstral coefficients (MFCC) vectors. Based on the similarity of the MFCC vectors in this window, each MFCC vector is clustered into two clusters, ignoring the time order of MFCCs. Then Euclidean distance between the mean of two clusters for each window is computed to obtain a contour in time. Then a peak-detection algorithm is applied to obtain the phonetic boundaries. They reported a correct detection rate (CDR) of 67.9% and under segmentation rate of 1.4% within 20 msec on TIMIT test-set.

Ladan Golipour and Douglas O'Shaughnessy [8], presented an unsupervised and unconstrained approach for phone-level speech segmentation. Their goal was to recognize the locations of main energy changes in frequency over time using STFT of speech signal and sub-band analysis, which can be described as phone-level boundary. They employed the modified group-delay function to achieve a more clear representation of the locations of boundaries and smooth out undesired fluctuations of the signal. They reported a performance of 87% correct boundary detection and an error of 6.6 % extra boundaries on a part of TIMIT test-set.

2.2.1.1 Limitations

- One of the key limitation to these signal processing approaches is that, the most of the signal processing based approaches are highly sensitive to the parameters used and hence are less robust, as observed by [9]. For example, the delta energy contour of a speech signal have many peaks, of which some are boundaries and the rest are spurious peaks. A small change in the peak cut off threshold, can change the performance of boundary detection drastically.

2.2.2 Force-alignment based approaches

Force-alignment based acoustic-phonetic boundary detection is an approach wherein the speech signal is aligned with its corresponding phonetic transcription. So these approaches are constrained by the phonetic transcription of an utterance to obtain the boundaries and hence require phonetic transcription. These approaches perform acoustic-phonetic speech segmentation, i.e., both acoustic-phonetic boundary detection and acoustic-phonetic segment labeling at the same time. Hence these approaches differ from other approaches in two ways. Firstly, it requires phonetic transcription of the utterance as the input for alignment and secondly, it will give the segment boundaries as well as the segment labels in the same step instead of two different steps as shown in Fig 1.1. In these type of approaches, the number, the identity, and the order of the phonemic units are known a-priori. The accuracy or performance of these approaches mainly depend on the choice of utterance transcription available, and segmentation procedure employed. The rest of the section deals with these variations and their limitations.

2.2.2.1 Utterance transcription

Even though the segmentation of the speech data is a task performed at phonetic level (Barry and Fourcin, 1992), the transcription of an utterance that is submitted to the system can be one of the following levels (Barry and Fourcin, 1992; Kvale 1993):

1. *A citation, canonical or standard phonemic transcription:* This represents the concatenation of the standard pronunciations of the words contained in the uttered sentences. It is a sequence of phones which considers neither assimilations nor coarticulations effects in speech signal.

2. *A phonotypical or text-to-phone phonemic transcription:* This is obtained by the concatenation of the standard pronunciation, with the consideration of the context of the words. This takes care the word junction phone variation.

3. *An auditory phonemic transcription:* As a word can have multiple pronunciation, using only standard pronunciation may not obtain the exact transcription. This transcription is obtained by listening carefully to the speech signal. It still consists of the phones which are actually perceived by the listener.

4. *An audio-visual or manual-labeling phonetic transcription:* This transcription however is a sequence of acoustically motivated sound segments, called phones. These phones have relatively stable changing acoustic-phonetic properties. These are identified by an audio-visual (spectrogram) inspection of the speech signal manually. This can be considered as the most precise representation of the speech signal.

Table 2.2: *Describes different types of transcription for a given orthographic representation, resources required for each type and an example.*

Transcription Type	Example	Resources Required
Orthographical	"romantic gift"	-
Canonical	/r/ /ow/ /m/ /ae/ /n/ /t/ /ih/ /k/ /g/ /ih/ /f/ /t/	Pronunciation dictionary
Phonotypical	/r/ /ow/ /m/ /ae/ /n/ /t/ /ih/ /g/ /ih/ /f/ /t/	Context based letter to sound rules
Auditory	/r/ /ax/ /m/ /ae/ /nx/ /ix/ /k/ /g/ /ih/ /f/ /t/	Manual annotation
Manual-labelling	[r] [ax] [m] [ae] [nx] [ix] [kcl] [g] [ih] [f] [tcl] [t]	Manual annotation

Table 2.2 shows different types of transcriptions that may arise from a particular orthographical representation, resources required for each type and an example for each type. Examples for each type are extracted from the TIMIT database [26]. From Table 2.2 we can observe that, descending order of difficulty to obtain transcription is as follows: Manual-labeling ¿= Auditory ¿ Phonotypical ¿ Canonical ¿ Orthographical. Hence among the phone level transcriptions, canonical is easiest and manual-labeling is toughest to obtain.

2.2.2.2 DTW-based approaches

Dynamic time warping (DTW), is a dynamic programming algorithm that aligns two sets of features in time using set of reference features and a distant metric, so that the error between the two features is minimized.

Svendsen and Soong [27] used DTW to align the input speech with speaker independent phonetic templates obtained from spectral average of different speakers phonetic templates. They reported an agreement of 32%, 72% and 92% wthin 15 msec, 30 msec and 45 msec.

Falavigna and Omologo [28] also aligned input speech with phonetic prototypes, but used spectral variation function to emphasize changes in signal. They reported an agreement of 61% within 20 msec.

Leung and Zue [17] developed a system for automatic alignment of phonetic transcriptions with continuous speech. The speech signal is first segmented into six broad phonetic classes using a non-parametric pattern classifier. Second, using a knowledge-based dynamic programming algorithm, the sequence of broad classes is aligned with the phonetic transcriptions. These broad classes provide reliable segments for more detailed segments and refinement of boundaries. Finally, this initial time alignment serves as anchor points for subsequent detailed phonetic alignment utilizing a set of heuristic rules. The system was evaluated on three speakers (2 male and 1 female) read speech. Results were approximately 75% and 90% agreement within 10 msec and 20 msec respectively.

In summary, this approach requires a TTS or a speech templates with boundaries. Inspite of all these, the results were not as good as manual boundaries.

2.2.2.3 HMM-based approaches

Rapp noted that because "the task if phone alignment can be considered as simplified speech recognition, it is natural to adapt a successful paradime of automatic speech recognition, namely HMMs for phonetic segmentation" [29]. So, automatic phonetic segmentation can be performed within the HMM framework by constraining the grammar network in an HMM phone recognizer to only recognize the given phone sequence. "Recognition" is obviously performed with perfect accuracy, but in doing the recognition search we also determine the most likely state sequence, and this gives us the phone boundaries. The phone boundaries are provided as the time instances of the model transitions by tracing back the optimal path found by the Viterbi algorithm. Often this operation is called **forced alignment**. A number of studies have investigated using a state of art general purpose speaker independent / dependent speech recognizer to perform the alignment. Some of these approaches are described below.

Ljolije and Riley [15], built a three-state HMM system that has different types of phonetic models, depending on the availability of training data such as triphone models, quasi-triphone models and monophone models. The HMM uses full-covariance Gaussian probability density function to estimate the state emission probability, a Gamma distribution duration model, and 10 msec frame shift. Two types of models were trained: first using the manual alignments in the TIMIT database, and second using both manual alignments and Viterbi re-estimation of the alignments. In either cases, they found 80% agreement within 15 msec.

Dalsgaard, Andersen, Berry and Jorgensen [30],[31], [32] used a self-organizing neural network (SONN) to estimate the probabilities of distinctive phonetic features. For example, phone /s/ was defined by the vector [front back mid round dent velar frication], where each of the distinctive features may have the value +1, 0 or -1 depending on whether it is present, not relevant or absent. These distinctive features were subjected to principle component analysis to determine the most relevant features for phonetic classification. These principle components were used to model phonetic likelihoods with Gaussian probability density functions, and then Vitterbi search was applied to these likelihoods to align the speech. When evaluated on EUROM0 corpus using 15 principle components this system yielded an agreement of 66.5%, 77.5%, and 52.0% with 20 msec on Danish, English and Italian languages respectively.

Brugnara et al. [33], [34], [1] developed an HMM force-alignment system which was trained using the TIMIT manual labels as an initial segmentation. They used spectral variation features in addition to the standard cepstral-domain features, which resulted in a 2% relative reduction in error. They evaluated this system on TIMIT system on TIMIT test-set, and reported an agreement of 75.3%, 84.4% anf 88.9% within 10 msec, 15 msec and 20 msec respectively. They also compared this system to an identical system trained without initialization from the manual alignment information, and found that the system trained using the manual alignments had a 50 % reduction in error compared to the system trained without manual alignment.

Kvale [19], developed a two phase segmentation algorithm. First segment the speech into acoustically similar segments by using sequence-constrained vector quantization (SCVQ). SCVQ is a special case of VQ which is constrained so that all the vectors in a cluster are contiguous in time. Using this method, they generated about 2.5 times as many segment boundaries as phonetic boundaries. In second phase, a three state mono phone single mixture HMM with skip state is trained on each phone using EUROM0 corpus (16 kHz read speech male and female speakers in different languages). Then, during HMM segmentation, Viterbi search was constrained by the condition that a state transition is only allowed at the hypothesized segment boundaries. Results on the English speaker showed an agreement of 82.3% within 20 msec. This system yielded an agreement of 86.1%, 82.3%, 84.5% and 86.4% with 20 msec on Danish, English, Italian and Norwegian languages respectively from EUROM0 corpus.

Rapp [29] trained a forced-alignment system for German using HTK toolkit [35]. He used a 10 msec frame shift and reported an agreement of 84% within 20 msec on Kiel Read Speech Corpus..

Pauws, Kamp and Willems [36] trained an HMM system using a three-step process in order to avoid the initialization of their training with manual alignments. Their system was trained and evaluated on 8 kHz Dutch isolated word database containing 827 words spoken by a single speaker. In the first step, the speech was segmented into three broad phonetic classes, namely silence, voiced and unvoiced, using energy in different bands, the zero crossing rate, and the spectral slope. This step had an agreement of 82.05% with in 20 msec of the manual broad phonetic boundaries. Given this segmentation, the next step was to use sequence-constrained vector quantization (SCVQ) within each broad phonetic class to align the phones which resulted in an agreement of 70.37% within 20 msec. In the final step, HMM was trained to recognize each phone, with the initial segmentation taken from the second-step results. It used frame shift of 5 msec, six states per phone for all phones except bursts, which had 2 states per phones, hence enforcing a minimum duration of 10 msec for bursts and 30 msec for others. Performance of this system was 89.5% agreement within 20 msec. In a comparative study, this hierarchical system was compared with a forced-alignment HMM system that was initialized with manual segmentation and another system that was initialized randomly or a linear system with equal-duration segmentation. These systems had an agreement of 96.0% and 76.14% within 20 msec. It should be noted, however, that as all these systems are trained and tested on single speaker data, they cannot be compared with speaker-independent alignment systems.

Wesenick, and Kipp [37] implemented an HMM system to use in cases where only word-level transcription is available. This system performed simultaneous alignment of the canonical dictionary pronunciation and several pronunciation variants. The HMM system used context-independent models with between three and six states per phone and 10 msec frame shift. The HMM system was trained and evaluated on the PHONODAT-II corpus of German speech, and was initialized with manually-aligned data. The post-processing refinement adjusted the boundaries within a 10 msec window using simple time-domain techniques gave an agreement of 84% in 20 msec.

Wightman and Talkin [38] developed an HMM-based system called the "the Aligner", with the acoustic model training and Viterbi search implemented using HTK Toolkit [35]. It uses a 10 msec frame shift and five Gaussian mixtures per state state to estimate the observation emission likelihoods. The system was trained using TIMIT labels as an initial segmentation. In evaluation of their systems, they did not use the TIMIT phonetic sequence directly, but they mapped the forced-alignment phones to the TIMIT phone sequence. Performance of this system on the TIMIT test-set using manually obtained transcript was approximately 80% agreement within 20 msec.

Pellom [39], used a HMM forced aligmnet based segmentation approach with variety of enhancement algorithms for segmenting the noisy speech. The system used a 5 msec frame shift, 5 state mono phone HMMs, 16 GMMs per state, Gamma distribution transition probabilities and gender dependent models. When phone-level transcriptions are not available, the system generates pronunciation using CMU dictionary and word junction modeling. The system was trained on TIMIT train-set that had been down-sampled to 8 kHz and evaluated on the TIMIT test-set (8 kHz clean speech), the NTIMIT corpus (telephone-band speech) and the CTIMIT (cellular-band speech) using various noise reduction techniques. He reported an agreement of 85.9%, 74.9% and 63.7% within 20 msec for TIMIT, NTIMIT and CTIMIT respectively.

Hosom [3] described a baseline forced-alignment system and a proposed system with several modifications to this baseline system for speaker-independent phone alignment. The baseline system was an HMM/ANN hybrid which computes probability estimates of observations using an Artificial Neural Network (ANN) instead of a Gaussian Mixture Model (GMM). They used a 13 dimensional mel-frequency cepstral coefficients with their delta coefficients. Depending on the nature of the phone, number of states for each phone were decided. So, they used 451 states to represent these 61 phonetic and sub-phonetic units. While training of ANN for HMM/ANN hybrid, the probability of an observation given a state was estimated using a 3-layer ANN trained for each state. The ANN had as input features a context window of five frames, with frames at -60, -30, 0, 30, and 60 ms relative to the center frame. The network thus had an input layer of 130 nodes (13 + 13 features per frame and 5 frames), a hidden layer of 300 nodes, and an output layer of 451 nodes. The proposed system implements three modifications to the baseline system: (1) The feature set includes, in addition to the baseline systems cepstral features and normalized log energy (computed with a 100-ms window), four additional energy-based feature streams; (2) The system uses, in addition to probabilities of each phone-based state given an observation, probabilities of a state transition given that observation; and (3) Instead of computing context-dependent phone probabilities directly, the system computes the probabilities of distinctive phonetic features (such as manner, place and height of a phone). The probabilities of these features are then combined to obtain phone probabilities using Massaro's fuzzy-logic model of perception (FLMP). Performance of the baseline system on the test partition of the TIMIT corpus is 91.48% within 20 ms, and performance of the proposed system on this corpus is 93.36% within 20 ms.

Toledano et al. [40], proposed a statistical correction procedure for HMM based phonetic aligner to compensate for the systematic errors produced by context-dependent HMMs and the use of speaker adaptation techniques is considered to increase the segmentation precision. A general framework is proposed for the local refinement of boundaries, and the performance of several pattern classification approaches. This resulting system was able to increase the performance of a baseline HMM segmentation from 27.12%, 79.27%, and 97.75% of agreement within 2 msec, 20 msec, and 50 msec respectively to 65.86%, 96.1%, and 99.31% in speaker-dependent mode.

Mporas et al. [41], has proposed fusion scheme for combining multiple phonetic boundary predictions which are obtained through various segmentation engines. They have used 112 HMM based speech segmentation engines, which differ in the setup of HMMs and speech parameterization techniques. Best performing on TIMIT corpus among these was a three state, two Gaussian per state, context-independent HMM based segmentation using Human Factor Cepstral Coefficients (HFCC) with an agreement of 68.38% and 79.41% within 15 msec and 20 msec respectively. This present fusion technique in combination with support vector regression (SVR), improved to 82.28% and 88.18% respectively.

In summary, the reported systems represent numerous refinements on the standard HMM procedure, but in all cases the basic process remains the same, namely estimating phonetic likelihoods at each frame, and then searching through these likelihoods with a constrained Viterbi search to determine the phonetic alignment. Direct comparison of the results from these systems is not possible, as even the system evaluated on TIMIT corpus, there are many variations such as HMM setup, frame size, frame shift, number of phones used, sampling rate of the speech signal, etc. If, however, we assume that the performance difference due to these variations in tuning of the parameters, features used etc., Thus, we can conclude that the performance of HMM systems on TIMIT corpus ranges between 80% and 94% within 20 msec. Apart from these variations, all the HMM systems require phonetic transcription (manual / canonical) and some times even the manual segmentation to train the systems.

2.2.2.4 Limitations

1. The main limitation of TTS/DTW based algorithms is that they are very much depend on the quality of the synthesizer. A good synthesizer requires good segmented speech data and hence the approach itself is controducting.

2. One of the main reason limitation of HMM systems is that it requires a good transcription for both training and testing, which is very difficult to obtain.

2.2.3 Classification based approaches

In the present approach, acoustic-phonetic boundary detection is performed by using pattern classification techniques. Speech signal can be represented as the sequence of frames of which some are boundaries and the rest are non-boundaries. The core idea of this approach is to classify all frames in the speech signal into two classes i.e., boundary / non-boundary frames. Let x is a frame in the speech signal, B denote the boundary class and NB denote the non-boundary class, then this approach estimates $p(B/x)$ and $p(NB/x)$. If $p(B/x) > p(NB/x)$ then x is a boundary frame else x is a non-boundary frame. These approaches fall in the supervised / unsupervised unconstrained approaches where the segmentation is carried out using some models trained without using phone transcription. There are a few approaches in this area, apart from the work reported by Suh et. al. [10], [2].

Youngjoo Suh and Youngjik Lee [10], proposed an approach using multi-layer perceptron, which consisted of three parts: preprocessor, MLP-based phonetic segmentor and post processor. The preprocessor utilizes a sequence of 44 order feature parameters for each frame of speech, based on the acoustic-phonetic knowledge and manually labeled data. The MLP has an input layer with 176 nodes, one hidden layer and an output layer with one node. The output value from output node decides whether the current frame is a phone boundary or not. Post processing decides the positions of phone boundaries using output of MLP. They reported an agreement of 84%, 87 % within 5 msec and 15 msec and an insertion of 9% on a single speaker Korean read speech database.

Keshet et al. [2] proposed a support vector machine (SVM) based discriminative learning procedure. This approach used manually segmented boundaries and their labels to train the data. The alignment function was devised to map the input acoustic and symbolic (phone) representations of the speech utterance along with the target alignment (phone start times) into an abstract vector space. A specific mapping into the abstract vector-space which utilizes standard speech features (e.g. spectral distances) as well as confidence outputs of a frame wise phone classifier was employed. Building on techniques used for large margin methods for predicting whole sequences, our alignment function distills to a classifier in the abstract vector-space which separates correct alignments from incorrect ones. This system had an agreement of 80.0% and 92.3% within 10 msec and 20 msec.

2.2.3.1 Limitations

- One of the key limitation to this approach is that it require good amount of manually segmented data to train the discriminative classifier models.

2.3 Need for new approach and features

Among all the approaches discussed in this chapter, HMM based approaches out perform others. But there are three major limitations in such HMM based systems: firstly, these HMM based systems require exact, manually obtained phonetic transcription which is very difficult to obtain. On the other hand, the canonical phone transcript can be obtained very easily using dictionary look-up, but the performance will go down because of the phonetic mismatch between the transcription and the speech signal. A simple analysis was performed to check the degree of mismatch between the canonical and manually obtained phonetic transcription of TIMIT train and test data, in order to understand the gravity of the problem. This can be analyzed by aligning both the transcripts using simple string matching based dynamic programming algorithm and capture the number of phone insertions, deletions and substitutions by canonical phone transcript over manually labeled phone transcript. Table 3.1 shows that while canonical phone transcript matches with manual phone transcript 84% of times, but there is a phone error rate (PER) of about 28% over the latter on both TIMIT train and test sets. Second, when dealing with non-native speech data, non-native pronunciation will be different and thus have phone insertions, deletions and substitutions due to mis-pronunciations. In addition to this, non-native speech may contain even the phones which are not present in the native English language. This puts the barrier on using the native HMM models itself for non-native speech segmentation. Third, in order to overcome the previous issue, and still use HMM

Table 2.3: *Accuracy, Phone Error Rate, Substitutions, Insertions, and Deletions of phones by canonical phone transcript over exact phone transcript using TIMIT corpus.*

Data	**Acc.**	**PER**	**Sub.**	**Ins.**	**Del.**
Train	83.9%	27.1%	6685	5625	1612
Test	83.7%	27.7%	19146	16225	4205

models, one has to build a non-native HMM models which requires huge amount of non-native speech data for training.

In order to overcome the above issues, our proposal is the following. Signal processing based approach can be used for acoustic-phonetic boundary detection in continuous speech without using transcription. It can be seen from the previous studies that the parameters and thresholds used in such approaches are sensitive and can effect the performance to a large extent when used on other databases. To overcome this issue, we propose to use classification based approaches by training a classifier on the boundaries obtained by signal processing approaches. Thus bu combining the signal processing and classification based approaches, we have avoided the need for transcription as well as manually annotated boundaries.

Another limitation of the existing approaches is that the spectral features such as mel-filter cepstral coefficients (MFCC), filter-banks, linear prediction cepstral coefficients (LPCC) etc., are used as features for acoustic-phonetic boundary detection. But these spectral features fail at some situations such as boundaries between unvoiced stop bursts / unvoiced fricatives (eg., /k-s/), two unvoiced fricatives (eg., /ch-s/) etc., where there are no formant track information at all. So, there is a need to explore different features which are independent of formant tracking for the task of speech segmentation.

2.4 Summary

In this chapter an overview of different approaches such as manual and automatic methods for acoustic-phonetic boundary detection of continuous speech are discussed. We can observe from the manual boundary detection limitations that even though they perform better than automatic approaches, they are very costly, not reusable and inconsistent, which conforms that automatic techniques are must for acoustic phonetic boundary detection. Automatic boundary detection approaches can be broadly categorized into signal processing, model and forced-alignment based approaches. Main advantage of signal processing based approaches is that they neither require any kind of transcription nor any training, but they are totally dependent on the threshold parameters which makes these approaches little unstable to use on different databases. This instability can be rectified by using classifier based approaches, which also do not require phonetic transcription but requires manually labeled data to train the models which is a bottle neck at training phase. Force-alignment based approaches such as DTW, requires a Text-to-Speech system or recorded prompt of the utterance and its manual segment boundaries that has to be segmented in order to perform the segmentation of the multiple repetitions of the same utterance and so its usage is very restricted. Among the automatic boundary detection approaches, high performance was obtained by HMM based force-alignment approaches but it requires manually labeled phonetic transcription which is the main bottle neck for this approach both at training as well as testing phase. In order to overcome the limitations such as instability for signal processing based approaches, manual boundaries for classifier based approaches and manual phonetic transcription for force-alignment based approaches, we are exploring for a new approach.

CHAPTER 3

Classification based acoustic-phonetic boundary detection

Acoustic-phonetic speech segmentation is a process of boundary detection and segment labeling. In boundary detection phase, the boundaries of acoustic-phonetic units in speech signal are automatically detected independent of language, transcription and phone set. In labeling phase, each segment is labeled with a phone symbol. Traditional approaches of acoustic-phonetic boundary detection use signal processing methods or employ supervised classifiers. Signal processing based approaches typically rely on peak-picking algorithms on temporal trajectories which operate independent of language and transcription . However, signal processing approaches are sensitive to parameters used and are less robust than supervised classifiers. On the other hand, supervised classifiers require hand labeled data to train classification models. The process of obtaining hand-labeled data is not only difficult but also laborious. In this chapter, unsupervised classification approach for acoustic-phonetic boundary detection approach is proposed by combining signal processing and classification based approaches. **In this approach, first step is to obtain acoustic-phonetic boundaries using signal processing based methods. These boundaries are then used to train boundary detection classifier.**

A comparison of unsupervised classifier, supervised classifier, signal processing and HMM based forced-alignment approaches are performed. The result show that the unsupervised classifier performs better than signal processing and HMM based forced alignment approaches, while supervised classifier out performs others.

3.1 Signal processing based acoustic-phonetic boundary detection

Signal processing approaches generally use the combination of signal processing techniques and peak-picking methods to perform the task of acoustic-phonetic boundary detection. Mostly these approaches fall in unsupervised category as they don't require any manually labeled data. Rest of the section describes the two signal processing based approaches i.e., mean spectral smoothing (MSS) and group-delay function (GDF) based approaches in detail.

3.1.1 Mean spectral smoothing based approach (MSS)

Most of the signal processing approaches are based on predicting acoustic changes that occur at the phonetic boundaries. These acoustic changes at the phone boundaries can be interpreted as difference in the distribution of the two adjacent phones. Difference between the two adjacent speech segment distributions will be high when the difference is measured at phone transitions and will be least when they are measured within same phone.

In order to implement this algorithm, we have to choose a statistical parameter that can capture the distribution with minimal amount of data, a measure to compute the difference between the distributions, that is compatible with the chosen statistical parameter. Finally we have to choose the feature on which this algorithm has to be applied to get a better performance. As *mean* is a first order moment, it required less amount of data to represent a distribution when compared to other higher order moments such as standard deviation, skewness etc., we are choosing *mean* as the statistical parameter to capture the distribution. Regarding measure to compute the difference between the distributions, there are many measures such as Euclidean distance, City distance etc., but when performance of this algorithm was tested on a small subset of TIMIT train-set, Euclidean distance was out performing others. Hence we choose to use Euclidean distance as the measure for computing the distortion between distributions. We choose Mel-Filter Cepstral Coefficients (MFCC generated using frames of 10 msec frame size and 5 msec frame shift) as a feature for algorithm, as it is one of the highly used acoustic feature in many speech processing systems.

Following is the pseudo code of MSS approach:

- **Step 1:** Compute the mean MFCC of the n frames to the left ($\boldsymbol{\mu}_L$) and right ($\boldsymbol{\mu}_R$) of i^{th} frame in the speech signal.

$$\boldsymbol{\mu}_L(i) = \frac{1}{n}\sum_{j=i-n}^{i} \boldsymbol{x}_j \quad and \quad \boldsymbol{\mu}_R(i) = \frac{1}{n}\sum_{j=i}^{i+n} \boldsymbol{x}_j \tag{3.1}$$

  Where $\boldsymbol{x}_j$ is a 13 dimensional MFCC vector of j^{th} frame in a speech signal and n is the number of frames to the left and right which is taken as five here.

- **Step 2:** Compute the Euclidean distance between the left and right distribution means of i^{th} frame in the speech signal

$$\boldsymbol{D}(i) = \sqrt{\sum_{k=1}^{K} [\mu_L(i,k) - \mu_R(i,k)]^2} \tag{3.2}$$

  Where K is the dimensionality of the mean vectors which is same as that of dimension of MFCC vectors.

- **Step 3:** Repeat step 1 and 2 for all the frames in the speech signal to obtain D for all i.

- **Step 4:** Once D is computed for all i we obtain a distribution distortion contour as a function of time. All the peaks in this contour correspond to highly distorted frames and hence most likely phonetic boundaries. So, we have to apply a peak picking algorithm to get the segment boundaries. This peak picking is performed by computing first derivative of D and all the frames at which it is changing from positive to negative as the difference in the distributions and hence the phone boundaries.

$$\boldsymbol{\nabla} D = \frac{d}{d_i}\boldsymbol{D} \tag{3.3}$$

$$\boldsymbol{B}(i) = \begin{cases} boundary & \text{if } \boldsymbol{\nabla} D(i-1) > 0 \ \& \ \boldsymbol{\nabla} D(i) \leq 0 \\ non - boundary & \text{Otherwise} \end{cases} \tag{3.4}$$

3.1.2 Group-delay function based approach (GDF)

GDF-based boundary detection is also an example of signal processing based approach which focuses on acoustic cues to detect the transient behavior at the phone boundaries as shown in the Fig 3.1. Brief description of this work is presented here and more details can be found in [8].

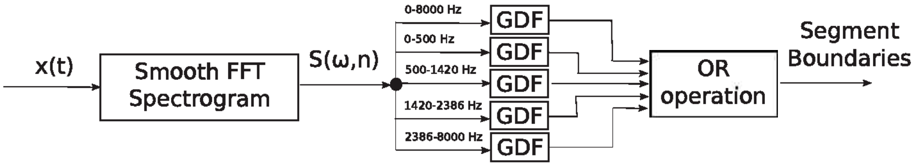

Figure 3.1: *Block diagram of GDF based phonetic speech segmentation.*

- Speech signal is divided into frames with a frame size of 8 ms and frame shift of 4 ms and a 512 point FFT is applied on each frame to obtain $X(w, n)$.
- Power Spectrum $P(w, n)$ is computed using real and imaginary components of $X(w, n)$ and it is smoothed using median filtering to obtain $S(w, n)$.
- Compute the gradient of $S(w, n)$ to obtain a measure for the change in the energy of each frequency during the time of utterance. These energy changes are summed over in 5 different frequencies i.e., 0-8000Hz, 0-500Hz,

500-1420Hz, 1420-2386Hz and 2386-8000Hz to obtain different $Y(n)$ for each band.

- As modified group-delay function was used to derive significant information such as peaks in the spectral envelop, it is applied on each $Y(n)$ separately to obtain boundaries using equation (3.5).

$$\tau_{Y(n)} = sign\left|\frac{Y_R(n)Z_R(n) + Y_I(n)Z_I(n)}{S(n)^2}\right|^{\alpha} \tag{3.5}$$

- An "OR" operation is performed on the boundaries obtained by different bands to obtain final boundaries.

3.2 HMM based acoustic-phonetic speech segmentation

In this section, we are describing the forced-alignment based acoustic-phonetic boundary detection approach. These approaches are termed as acoustic-phonetic speech segmentation approach as these approaches obtain the segment boundaries and segment labels simultaneously. Forced-alignment is the technique employed by most of the HMM based speech segmentation approaches. These approaches require phonetic transcription to perform speech segmentation which is combined process of obtaining segment boundaries and segment labels. There are two kinds of phonetic transcription s that can be used for HMM based speech segmentation.

- *Manual phonetic transcription :* This type of transcription is obtained by manually transcribing the speech signal by a human annotator.

- *Canonical phonetic transcription :* This type of transcription is obtained by using word pronunciation dictionary lookup for each word in the sentence and concatenating them.

Depending on the type of transcription used for segmentation, we can classify HMM based segmentation approaches into following two types:

3.2.1 HMM-based segmentation using manual phonetic transcription

In this approach, manual phonetic transcription is forced aligned with speech signal to obtain segment boundaries. As these approaches use manual phonetic transcription for training and segmenting, there will be no boundary deletions and insertions. But, obtaining manual phonetic transcription is tedious and expensive task. Previous works by Brugnara et. al. [1] and Hosom [3] are some of the examples of such works, where they have used manual boundaries and manual phonetic transcription to train the HMM models on TIMIT train-set and segmented the TIMIT test-set using manual phonetic transcription .

3.2.2 HMM-based segmentation using canonical phonetic transcription

As the manual transcription is expensive, the alternative is to use canonical phonetic transcription for speech segmentation. Problem with such transcription is that one cannot get exact phonetic sequence of the speech utterance and hence there will be insertions, deletions and substitutions of phones in the phonetic transcription . These phone insertion, deletion and substitutions will have a direct effect on the boundary detection as HMM forced-alignment based approach use canonical phonetic transcription . So, before analyzing the speech segmentation

Table 3.1: *Accuracy, Phone Error Rate, Substitutions, Insertions, and Deletions of phones by canonical phone transcript over exact phone transcript using TIMIT corpus.*

Data	**Acc.**	**PER**	**Sub.**	**Ins.**	**Del.**
Train	83.7%	27.7%	19146	16225	4205
Test	83.9%	27.1%	6685	5625	1612

performance, we have to analyze the similarity between the *manual* and *canonical* phone transcripts and errors in *canonical* phone transcription . This can be analyzed by aligning both the transcripts using simple dynamic programing and computing the number phone insertions, deletions and substitutions by *canonical* phone transcription over *manual* phone transcript Table 3.1 shows that 83.7% of the *canonical* phone transcription matches with *manual* phone transcription and former has a phone error rate (PER) of 27.7% over the latter.

In order to evaluate the acoustic-phonetic boundary detection approaches with HMM-based approaches using canonical phonetic transcription , two HMM based approaches were developed on the same data and using the same phone set where ever needed, so that they can be directly compared and analyzed. Following are the detailed descriptions of HMM training, and two HMM based approaches:

3.2.2.1 Training HMM Models

The main advantage of using HMM models for speech segmentation is that it is built using extensive knowledge and infrastructure of speech recognition. Just as in speech recognition, HMMs for speech segmentation are also trained using the standard expectation-maximization (EM) algorithm. State sequence Θ is generated from *canonical phone transcript* and observation sequence O is obtained by parameterizing the speech signal. Speech parametrization is performed by computing a feature vector for every 5 ms using a 10 ms Hamming window and a pre-emphasis coefficient of 0.97. The feature vector used for HMM-based segmentation is a 12 Mel-Frequency Cepstral Coefficients (MFCCs) with Cepstral Mean Normalization (CMN) and normalized log energy, as well as their first and second order differences yielding a total of 39 components. To compute the likelihood function, state sequence Θ is considered as hidden data. Thus in order to obtain a maximum likelihood estimate $\bar{\lambda}$ of the model parameters, we must calculate the conditional expectation of the likelihood given a current set of parameters λ. Objective function $Q(\lambda, \bar{\lambda})$ has to be maximized in successive iterations:

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} P(O, \theta | \lambda) \log P(O, \theta | \bar{\lambda}) \tag{3.6}$$

Even though both speech recognition and speech segmentation use HMMs, however, it is important to realize that the goals of both the tasks are different. Hence these differences are reflected in the topology of HMM models. Past research [40][42] have indicated that context-independent models are preferred over context-dependent models and almost no improvement beyond two Gaussian per state for speech segmentation task. HMM topology of each phone is context-independent, 5 state sequential, left-to-right without any skip-state and observation probability distribution of each state is characterized by 2 mixture Gaussian. HMM based segmentation is implemented using HTK toolkit [35].

3.2.2.2 Speech segmentation in force-alignment mode (HMM-FA)

It is an approach, which assumes that phonetic transcription of the speech signal is known. In the present case as we are not using manually labeled phonetic transcription , we are generating the canonical phonetic transcription by concatenating the pronunciation of each word from canonical pronunciation dictionary. Figure 3.2 shows an example of concatenated phone state sequence used for HMM forced-alignment based phonetic speech segmentation. The corresponding concatenated phone HMM models are force-aligned with the parametrized observation sequence of speech signal to compute $[\log(P(O|\theta_t)]$ for every t and Viterbi search is used to find the optimal segment boundaries.

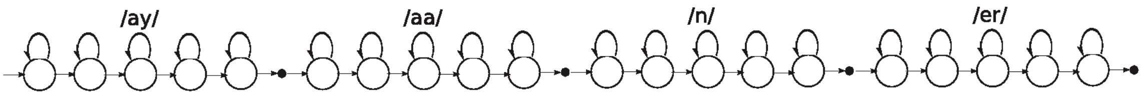

Figure 3.2: *Topology of phone state sequence used in HMM-FA based phonetic speech segmentation.*

3.2.2.3 Speech segmentation in phone-loop mode (HMM-PL)

It is an unconstrained model based approach, which does not require any phone transcription . Assuming all phones are equally likely, the topology of the this approach is shown in Fig 3.3. HMM models are obtained using the same training procedures described in training HMM models for HMM-FA based speech segmentation. Log likelihood of all states for each observation vector is computed as $[\log \max_{t=1}^{T} P(O|\theta_t)]$ (where T is total number of phone models) and Viterbi search is used to find the optimal segment boundaries.

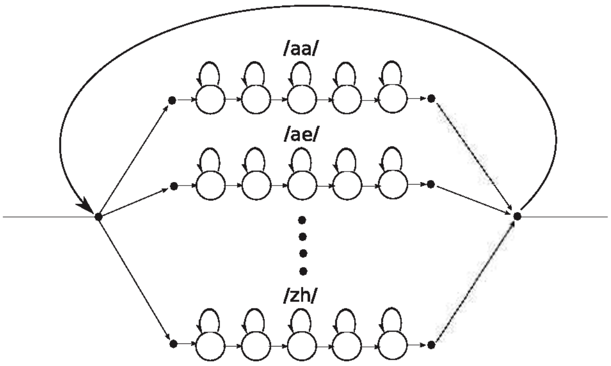

Figure 3.3: *Topology of phone state sequence used in HMM-PL based phonetic speech segmentation.*

Even though the above two approaches i.e., HMM-FA and HMM-PL based approaches are speech segmentation approaches which output the segment boundaries and segment labels at the same time, we have compared the other boundary detection approaches by only considering the segment boundaries.

3.3 Discriminative classification based acoustic-phonetic boundary detection

Signal processing based approaches search for boundaries without using any transcription where as HMM based approaches, search for boundaries using acoustic-phonetic transcription . But both the approaches search only for acoustic-boundaries in a speech signal. On the other hand, classification based approaches search for both boundary as well as non-boundary regions without using any transcription .

As described in the above approaches, present algorithm is also based on acoustic changes as a criterion for phonetic boundary detection without using any transcription. We have developed a framework wherein manually or automatically obtained segment boundaries (such as MSS, GDF) are used to train a boundary / non-boundary classifier without any trial and error analysis as in signal processing based approach. As this approach involves training a classifier using some machine learning technique, it can learn the different segment boundary patterns from multiple instances in the training data.

To train the boundary / non-boundary classifier an appropriate machine learning technique has to be selected. There are many ML techniques which can be used to train a classifier such as artificial neural networks (ANN), support vector machines (SVM), classification and regression tree (CART) etc,., But in the scope of this chapter we choose to use artificial neural networks (ANN)

Artificial Neural Network (ANN) models consist of interconnected processing nodes, where each node represents the model of an artificial neuron, and the interconnection between two nodes has a weight associated with it. ANN models with different topologies perform different pattern recognition tasks. For example, a feed-forward neural network can be designed to perform the task of pattern mapping, whereas a feedback network could be designed for the task of pattern association. A multi-layer feed forward neural network is used in this work to build a classifier which can classify each frame in the speech signal into boundary / non-boundary frames using acoustic features as input.

Figure 3.4 shows the block diagram of supervised ANN based acoustic-phonetic boundary detection. This is basically divided into three stages: first step is to prepare the data to train the classifier, next step is to train an ANN classifier and final step is to use this boundary detection framework to segment the speech using ANN classifier.

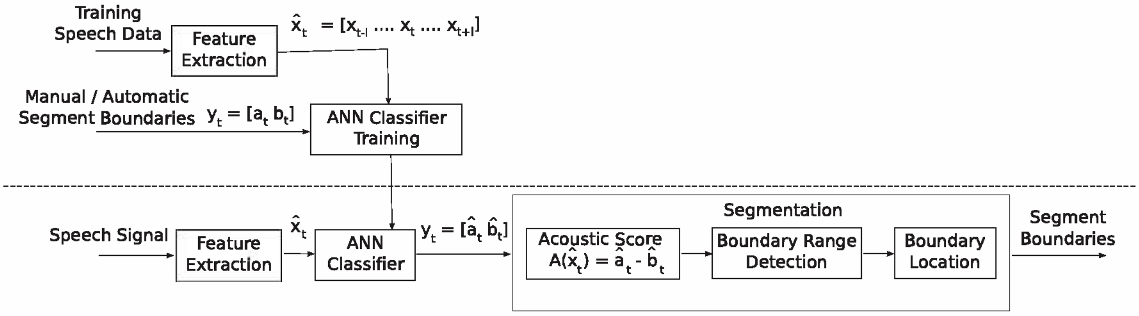

Figure 3.4: *Block diagram of ANN based acoustic-phonetic boundary detection.*

3.3.1 Preparation of input / output data

Most of the current boundary detection approaches use features that represent spectral information usually wrapped to emphasize perceptually relevant aspects such as Mel-Filter cepstral coefficients (MFCC). For a given speech signal, 13 dimensional MFCC are extracted with frame size 10 msec and frame shift 5 msec.

In order to train an ANN classifier, class labels (boundary / non-boundary) for every frame are required. Class labels are obtained from the manually/automatically marked boundaries in the TIMIT train-set. Using manual boundaries, all the feature vectors at the phone boundary frames are labeled as examples of boundary class. As all the frames between any two adjacent boundaries are non-boundary frames, there will be huge imbalance between the number of frames of each class in the training data, which could bias the classifier. In order to overcome this imbalance in the training data for two classes, for every pair of adjacent boundaries, only the frame which is in the middle of the two adjacent boundaries are selected as examples of non-boundary class.

Once the labels are assigned to the frames, the next step is to decide input and output vector format to train the classifier. Let $\boldsymbol{x}_t$ denote the feature vector extracted at frame t, then the input to ANN is an augmented feature vector $\hat{\boldsymbol{x}}_t = [\boldsymbol{x}_{t-l}, .., \boldsymbol{x}_t, .., \boldsymbol{x}_{t+l}]$. The value of l denotes the number of neighboring feature vectors appended to $\boldsymbol{x}_t$. Given $\hat{\boldsymbol{x}}_t$, the corresponding class label $\boldsymbol{y}_t = [a_t \; b_t]$ is created, where a_t and b_t are boundary and non-boundary evidence respectively for frame t whose values depend on the output target function used for training the network.

In our work, l is taken as 5, so each feature vector is a concatenation of 11 frames. Total dimension for each input vector to ANN is 143 (11 frames x 13 coefficients). Hidden and output layer target function used for training an ANN classifier is tangential (N) function. As output layer target function is tangential, output vector $\boldsymbol{y}_t$ is [1 -1] and [-1 1] for boundary and non-boundary frames respectively.

3.3.2 Training a classifier

An ANN is trained to classify the MFCC of speech signal into boundary / non-boundary class labels, i.e., if $G(\hat{\boldsymbol{x}}_t)$ denotes the ANN mapping of $\hat{\boldsymbol{x}}_t$, then the error of mapping is given by $\epsilon = \sum_t ||\boldsymbol{y}_t - G(\hat{\boldsymbol{x}}_t)||^2$. $G(\hat{\boldsymbol{x}}_t)$ is defined as

$$G(\hat{\boldsymbol{x}}_t) = \widetilde{g}(g(\boldsymbol{w}^{(2)} g(\boldsymbol{w}^{(1)} \hat{\boldsymbol{x}}_t))), \tag{3.7}$$

where

$$\widetilde{g}(\vartheta) = \vartheta, g(\vartheta) = \alpha \;\; \tanh(\beta \, \vartheta). \tag{3.8}$$

Here $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}$ represents the weight matrices of hidden, and output layers of ANN respectively. The values of the constants α and β used in *tanh* function are 1.7159 and 2/3 respectively. A generalized back propagation learning is used to adjust the weights of the neural network so as to minimize ϵ, i.e., the mean squared error between the desired and the actual output values. Selection of initial weights, architecture of ANN, learning rate, momentum and number of iterations are some of the optimization parameters in training an ANN [43]. Once the training is complete, we get a weight matrix that represents the function between the spectral features of a speech signal and their class evidences. Such a weight matrix can be used to classify a feature vector from the speech signal into boundary / non-boundary class labels.

3.3.3 Supervised and unsupervised classifiers

As discussed above, ANN classifier can be trained using manually or automatically generated acoustic-phonetic boundaries. In the rest of the thesis, the approach using manual boundaries to train the classifier is termed as supervised classification approach (SC) and the approach using automatic boundaries for training the classifier is termed as unsupervised classification approach (UC). Automatic boundaries are typically obtained from signal processing based methods such as MSS / GDF. These signal processing methods do have errors in their boundary detection. Hence a unsupervised classifier (UC) is typically trained using noisy training data.

3.3.4 Detection of acoustic-phonetic boundaries from classifier output

The proposed framework for phonetic speech segmentation can be divided into three main phases, i.e., *computing frame-level acoustic score*, *detection of boundary* *regions* and *location of boundary.*

3.3.4.1 Frame-level Acoustic Scores

Feature vectors are extracted from the speech signal for each frame t as described in Section 3.3.2. The corresponding augmented vector $\hat{\boldsymbol{x}}_t$, is given to the ANN classifier to obtain boundary/non-boundary evidences. Let $\hat{a}_t$ and $\hat{b}_t$ are the predicted boundary and non-boundary evidences respectively by the ANN classifier. Sign of $\hat{a}_t$ and $\hat{b}_t$ can be both positive, both negative and one positive and other negative and their values can range between -1 and +1, unlike a_t and b_t used while training which always have opposite signs and values can be only +1 or -1. Different cases and their implications are as follows:

- Case 1: Both $\hat{a}_t$ and $\hat{b}_t$ are positive

  As both the evidences are positive, it implies that the frame $\hat{\boldsymbol{x}}_t$ is both boundary as well as non-boundary. This can be resolved by using the actual values of $\hat{a}_t$ and $\hat{b}_t$. If $|\hat{a}_t| > |\hat{b}_t|$, i.e., predicted evidence of $\hat{\boldsymbol{x}}_t$ being classified as boundary is greater than classified as non-boundary, hence it can be classified as boundary, else it can be classified as non-boundary.

- Case 2: Both $\hat{a}_t$ and $\hat{b}_t$ are negative

  As both the evidences are negative, it implies that the frame $\hat{\boldsymbol{x}}_t$ is neither boundary nor non-boundary. This can be resolved by using the actual values of $\hat{a}_t$ and $\hat{b}_t$. If $|\hat{a}_t| < |\hat{b}_t|$, i.e., predicted evidence of $\hat{\boldsymbol{x}}_t$ being classified as boundary is greater than classified as non-boundary, hence it can be classified as boundary, else it can be classified as non-boundary.

- Case 3: $\hat{a}_t$ is positive and $\hat{b}_t$ is negative

  As $\hat{a}_t$ is positive and $\hat{b}_t$ is negative, it implies that the frame $\hat{\boldsymbol{x}}_t$ can be classified as boundary. The actual values of $\hat{a}_t$ and $\hat{b}_t$ are still useful as they can be used as confidence measure of $\hat{\boldsymbol{x}}_t$ being classified as boundary.

- Case 4: $\hat{a}_t$ is negative and $\hat{b}_t$ is positive

  As $\hat{a}_t$ is negative and $\hat{b}_t$ is positive, it implies that the frame $\hat{\boldsymbol{x}}_t$ can be classified as non-boundary. The actual values of $\hat{a}_t$ and $\hat{b}_t$ are still useful as they can be used as confidence measure of $\hat{\boldsymbol{x}}_t$ being classified as non-boundary.

In order to incorporate all the above cases in a simple form, equation 3.9 is used where $A(\hat{\boldsymbol{x}}_t)$ represents the *frame level acoustic score.* In general, range of $A(\hat{\boldsymbol{x}}_t)$ is between -2 and +2. For cases 1 and 2, it ranges between -1 and +1, for case 3, it ranges between +1 and +2 and for case 4, it ranges between -2 and -1. Once $A(\hat{\boldsymbol{x}}_t)$ is computed for all $\hat{\boldsymbol{x}}_t$, we obtain an acoustic score contour (A).

$$A(\hat{\boldsymbol{x}}_t) = \hat{a}_t - \hat{b}_t \tag{3.9}$$

3.3.4.2 Detection of Boundary Region:

Most of the times, it is not possible to pin point the boundary at once. So, in order to over come this issue, we have employed a two phase method in which first phase is used to roughly detect the region where the boundary can be detected and in second phase exact position of boundary marked. The main task of this phase is to identify the regions in the speech signal where the boundaries are likely to occur using an acoustic score contour (A). Once we obtain the *acoustic scores*, next step is to obtain the frame level classification which can be obtained by equation 3.10:

$$C(t) = \begin{cases} boundary & \text{if } A(\hat{\boldsymbol{x}}_t) > 0 \\ non-boundary & \text{Otherwise} \end{cases} \quad \forall t \tag{3.10}$$

where $C(t)$ stands for class label of t^{th} frame in the speech signal. But some unexpected patterns can be found in C such as t^{th} frame is non-boundary and $t-1^{th}$ and $t+1^{th}$ frames are boundaries, which means that the duration of that phone segment is just 5 msec which is highly unlikely. This is mostly because of incorrect classification by ANN classifier. This can be seen as spurious peaks and valleys in *acoustic score contour (A).* In order to remove these spurious peaks in A, an n-point linearly weighted mean smoothing is applied to smooth out the spurious peaks and valleys. In this case we have applied 5-point linear weighted mean smoothing to obtain a smooth acoustic score contour $\hat{A}$.

$$\hat{A}(\hat{\boldsymbol{x}}_t) = \frac{\sum_{i=1}^{n} f(i) A(\hat{\boldsymbol{x}}_{t-((n+1)/2)+i})}{\sum_{i=1}^{n} w(i)} \tag{3.11}$$

where

$$f(i) = \begin{cases} i & \text{if } i \le (n+1)/2 \\ n-i+1 & \text{if } i > (n+1)/2 \end{cases} \tag{3.12}$$

Once we have the *smoothed acoustic score contour* $(\hat{A})$, next step is to obtain *boundary regions* This is the segment of speech signal, where there is a high likelihood of phonetic boundary or the phone transition region. The region of consecutive boundary frames without any non-boundary frame is defined as *boundary region*, which can be interpreted as a part of *smooth acoustic contour* where $\hat{A}(\hat{\boldsymbol{x}}_t) > 0 \ \forall \ t$. If $\imath_q$ and $\jmath_q$ denote the begin and end frames of a boundary region q, it has to satisfy the condition $B(\imath_q, \jmath_q) = 1 \ \ and \ \ B(\imath_q - z1, \jmath_q + z2) < 1$, where $z1$ and $z2$ are positive integers, and $B(\imath_q, \jmath_q)$ is defined in the following equation

$$B(\imath_q, \jmath_q) = \frac{\sum_{t=\imath_q}^{\jmath_q} \hat{A}(\hat{\boldsymbol{x}}_t)}{\sum_{t=\imath_q}^{\jmath_q} \left|\hat{A}(\hat{\boldsymbol{x}}_t)\right|} \tag{3.13}$$

3.3.4.3 Location of Boundary

After the boundary regions are detected, exact location of boundary in each of these regions has to be located. Once the number of *boundary regions* are decided, number of boundaries are also fixed as only one boundary from each of these *boundary regions* are marked as phonetic boundary. Among the frames in each region q, the frame with highest acoustic score is marked as boundary frame using equation (3.14), where $\hat{i}_q$ is the index of the boundary frame from region q.

$$\{\hat{i}_q\} = \arg\max_t \{\hat{A}(\hat{\boldsymbol{x}}_t)\}_{t=\imath_q}^{\jmath_q} \tag{3.14}$$

3.4 Evaluation criteria

The performance of speech segmentation is evaluated using the following five metrics, essentially by comparing the predicted boundary with the manually marked boundary in the speech signal. If ζ_i denote the time stamp of the manually marked boundary i in the speech signal, then a region of tolerance ϵ_i is defined as $(\zeta_i - (\zeta_i - \zeta_{i-1})/2) \leq \epsilon_i \leq (\zeta_i + (\zeta_{i+1} - \zeta_i)/2)$. For every i, if there exists a predicted boundary $\hat{i}$ with its time stamp denoted by $\zeta_{\hat{i}}$, such that $\zeta_{\hat{i}}$ is within ϵ_i, then $\hat{i}$ is considered as correct boundary. If there are more than one predicted boundary within ϵ_i then one of the predicted boundaries which is nearest to ζ_i is considered as correct boundary, and the rest are considered as inserted boundaries. If there is no predicted boundary within ϵ_i, then i is considered as deleted boundary.

RMS Error: It is the root mean square of the deviations between the manual and its nearest correct boundaries.

Agreement Percentage (AGR): It is the percentage of correct boundaries with a tolerance (absolute deviation) of less than τ ms over the total number of correct boundaries.

Boundary Error Rate (BER): It is defined as the summation of insertion (INS) and deletion (DEL) percentages. Here, the INS percentage is computed as number of insertions over the total number of manual boundaries, and the DEL percentage is computed as number of deletions over the total number of manual boundaries.

Performance of boundary detection is better when RMS, DEL, INS & BER are low and AGR is high.

3.5 Results and discussions

In this section, we describe database used, different experiments performed to analyze the performance of the supervised and unsupervised ANN based acoustic-phonetic boundary detection approaches.

Database used:

The TIMIT corpus (a joint effort between MIT, Texas Instruments and SRI) contains read speech from 630 speakers from eight dialect regions of the United States. The sentences were designed to be phonetically rich and were recorded with a Sennheiser noise canceling, head-mounted microphone in a quiet environment. The speech was digitized at 16 kHz with 16-bit resolution. The corpus contains waveform data, text transcription s, canonical pronunciation dictionary and manually segmented and phonetic labels. TIMIT corpus has been labeled manually with 60 phones (excluding pause) out of which only 54 phones are used in the pronunciation dictionary, as remaining six phones are rarely used allophones. Hence we are using only 54 phones and rest of the six phones are mapped to their respective alternative allophones. Each speaker recorded ten utterances of which two sentences were common across all the speakers, finally containing 6300 utterances in the database. Of these 630 speakers, 460 speakers are used for training and 168 are used for testing. But, all the experiments in this chapter are trained on 3696 (TIMIT train-set) files from the training partition of the TIMIT corpus (excluding "SA" files) and the results are reported on 1344 (TIMIT test-set) files from the testing partition of the TIMIT corpus (excluding "SA" files) as used in [3] [1] [2], so that our results can be comparable to that of previous works.

Performance of different approaches:

In this section we are comparing different approaches that are described till now. Table 3.2(a) shows the performance of HMM-based speech segmentation approaches using manual phonetic transcription built by Brugnara et. al. [1] and Hosom [3] and a classification approach by Joseph et. al [2] on TIMIT test-set. As these approaches use manual phonetic transcription for training and segmenting, there will be no boundary deletions and insertions. So, when AGR scores are compared, except for $\tau < 10ms$, SC and UC without using any phonetic transcription are performing better than [1], and almost as good as [2], [3].

Table 3.2(b) shows that SC performs better than other approaches. It is understandable as SC is trained using manual training data. We can also observe that, UC performed better than MSS and GDF, even though the initial boundaries to prepare ANN data for SC is obtained from MSS. This shows that, the ANN training was able to discriminate between the correct and wrong boundaries generated by MSS inorder to build a better model.

From Table 3.2, we can also observe that SC out performs HMM-PL and HMM-FA and UC out performs HMM-FA.

Table 3.2 shows that ANN based approach not only out performed GDF, HMM-PL and HMM-FA based approaches, but also performs as good as constrained approaches using *exact* phone transcription [1][2] except for $\tau \leq 10ms$. When performances of SC (supervised ANN based approach), MSS (unsupervised heuristic approach) and UC (unsupervised ANN based approach) are compared, it is obvious that SC over performed MSS and UC, as it is trained using the manually segmented data. But more interesting part of this study is that inspite of using the erroneous training data obtained from MSS, UC has improved the acoustic-phonetic boundary detection BER by 3.9% over MSS performance. This shows the significance of classification based approaches for acoustic-phonetic boundary detection. It also shows the feasibility of unsupervised classification approaches for acoustic-phonetic boundary detection.

Table 3.2: *Agreement percentage (AGR) for different tolerance (τ) values, mean deviation (RMS), deletion percentage (DEL), Insertion percentage (INS) and boundary error rate (BER) of MSS, GDF, SC, UC, [1], [2], [3], HMM-FA and HMM-PL. In this work, we have considered BER as the first priority for evaluation and then AGR and RMS.*

	Approach	**Training**		**Testing**	**AGR % with $\tau \leq$**				**RMS(ms)**	**DEL**	**INS**	**BER**
		Manual Boundaries	**Phonetic Transcription**	**Phonetic Transcription**	$10ms$	$20ms$	$30ms$	$40ms$				
a	Brugnara [1]	Yes	Manual[1]	Manual[1]	74.6	**88.8**	94.1	96.8	-	-	-	-
	Joseph [2]	Yes	Manual[1]	Manual[1]	80.0	**92.3**	96.4	98.2	-	-	-	-
	Hosom [3]	Yes	Manual[1]	Manual[1]	79.30	**93.36**	96.74	98.22	-	-	-	-
b	MSS	NA[2]	NA[2]	No	49.24	**89.62**	96.41	98.46	13.1	16.80	16.71	**33.51**
	GDF	NA[2]	NA[2]	No	42.97	**88.56**	96.31	98.27	11.4	24.51	12.3	**36.81**
	SC	Yes	No	No	59.10	**92.18**	97.39	99.06	9.3	13.91	7.81	**21.72**
	UC	No	No	No	49.83	**89.80**	96.41	98.44	11.3	23.28	6.33	**29.61**
c	HMM-FA	No	Canonical[3]	Canonical[3]	55.85	**82.51**	94.76	98.19	15.2	10.75	19.97	**30.72**
	HMM-PL	No	Canonical[3]	No	51.82	**81.71**	94.89	98.16	15.7	17.33	9.75	**27.08**

[1] Manual phonetic transcription refers to the transcription obtained manually by a human annotator.

[2] NA refers to no training is required.

[3] Canonical phonetic transcription refers to the transcription obtained automatically using pronunciation dictionary.

3.6 Summary

In this chapter, we have described different signal processing methods, such as MSS and GDF based approaches, HMM methods such as HMM-FA and HMM-PL based approaches and classification methods such as SC and UC based approaches. When performances of MSS, GDF, SC and UC are compared, it is obvious that SC over performed MSS, GDF and UC. But the more interesting part of this study is that inspite of using the erroneous data obtained from simple MSS for training UC approach, BER of UC approach is reduced by 3.8% over MSS. This implies that ANN was able to identify and nullify the effect of some of the error patterns in the MSS output while discriminative training.

A comparative study was employed between acoustic-phonetic boundary detection without using any transcription and the HMM force-alignment based works [1], [2], [3] which used manual phonetic transcription on the same TIMIT train and test. Our observation was that SC performed as good as those HMM based approaches and UC performed better than one of the systems. As manual phonetic transcription is expensive, comparison with HMM-based approach using canonical phonetic transcription were employed. Among, HMM-FA and HMM-PL, HMM-PL out performed. This is mainly because of boundary insertions, which is directly reflected from the observation made from table 3.1 that canonical phone transcription has more number of phone insertions. We observed that SC is better than HMM-FA and HMM-PL, but UC was better than only HMM-FA. Finally, our experiments on INE and TEL databases show that the SC and UC based approach trained on TIMIT database could be used to segment non-native English and Telugu speech data. Hence showing that when manual phone transcription is not known, it may be better to use a SC based approach and if the manual boundaries are not present, then it is better to use a UC based approach.

CHAPTER 4

Significance of excitation based features for acoustic-phonetic boundary detection

This work investigates the significance of excitation based features for the task of boundary detection in continuous speech. In this work, we have compared the boundary detection performance of excitation based features with filter based features (linear prediction cepstral coefficients). It is typically known that excitation based features are useful to detect voiced-unvoiced or unvoiced-voiced segment boundaries. Our experiments and analysis done in this work demonstrate that excitation based features contain information about voiced-voiced and unvoiced-unvoiced segment boundaries along with voiced-unvoiced and unvoiced-voiced segment boundaries.

4.1 Significance of LP residual

The excitation-filter model of speech production consists of a filter that is excited by either a quasi periodic train of impulses or a random noise. Linear Prediction (LP) analysis is one of the most common techniques used to estimate the parameters of the filter. In LP analysis, the sample $s(n)$ is estimated as a linear weighted sum of past p samples. The predicted sample $\hat{s}(n)$ and its error $r(n)$ is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} \alpha_k s(n-k), \tag{4.1}$$

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} \alpha_k s(n-k) \tag{4.2}$$

where p is the order of the prediction, α_k, $1 \leq k \leq p$ is a set of LP coefficients (filter parameters) which characterizes the vocal tract and $r(n)$ is also called as LP residual (excitation) which characterizes the vocal folds.

In tasks such as speech recognition, speaker recognition and speech segmentation, it is well known that the filter parameters are widely used [3], [10], [44], [45]. On the other hand, it is often assumed that $r(n)$ is an uncorrelated noise, possess a flat spectrum and hence may not be useful for the above tasks [12]. Some researchers often summarize the whole residual in just one number representing pitch [46]. But it is shown that residual carries complimentary information to filter parameters [47][13].

It is important to understand that the information contained in the residual depends on LP order p. In practice, LP order of $12 - 16$ is assumed to be good enough to capture the shape of the vocal tract for speech signal sampled at 16 KHz [48]. If p is small (≤ 4), then the LP residual contains most of the information present in the speech signal, and if p is very large, then the LP coefficients contains most of the information present in the speech signal [49]. It is this property of LP model which makes the LP residual a complementary information to LP coefficients.

Our work provides a study on excitation based features for acoustic-phonetic boundary detection in continuous speech. To our knowledge such investigation has not been known so far. Fig 4.1(b) is a narrow-band spectrogram of residual which shows that residual is a good feature for detecting voiced/unvoiced regions and pitch estimation. But the boundaries between stop release, voiced pair (/b/-/r/, /t/-/aa/, /t/-/ih/); vowel, semi-vowel pair (/r/-/ih/, /aa/-/r/); stop release, fricative pair (/k/-/s/) etc., are not clear. On the other hand Fig 4.1 (c) is a wide-band spectrogram of residual. Though, it does not contain any pitch information, boundaries such as voiced, nasal pair (/ax/-/n/, /n/-/ao/, /er/-/n/, /n/-/ix/); stop release, voiced pair (/b/-/r/, /t/-/aa/, /t/-/ih/); stop closure, voiced/unvoiced pair (/tcl/-/t/, /s/-/tcl/, /l/-/tcl/) can be visually perceived in the spectrogram.

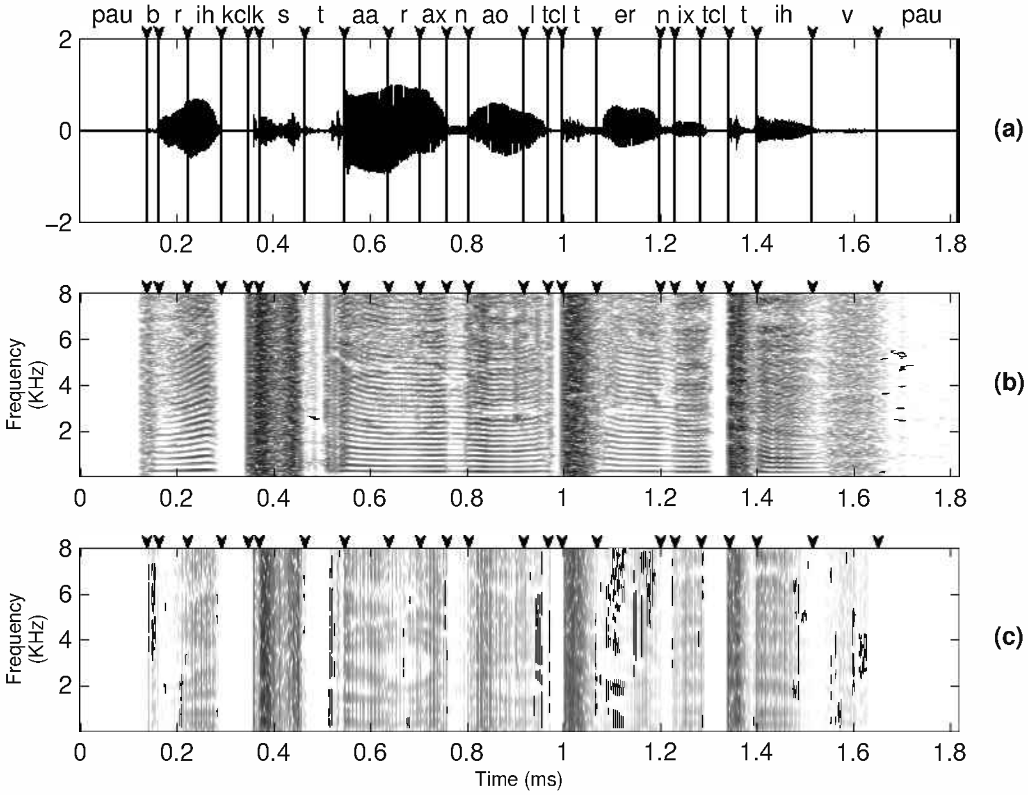

Figure 4.1: *(a) Speech signal, (b) Narrow-band spectrogram with 30 ms frame size and 1 ms frame shift and (c) Wide-band spectrogram with 4 ms frame size and 1 ms frame shift of* 16^{th} *order LP residual and all are marked with manual phone boundaries.*

To study whether the visually observed segment boundaries could be automatically detected from a residual, we computed the temporal changes in the spectrogram as follows. Let $S(\eta,\omega)$ and $X(\eta,\omega)$ denote the spectrogram (two dimensional signal) of a speech signal $\boldsymbol{s}(\eta)$ and residual signal $\boldsymbol{r}(\eta)$ respectively which are smoothed using 4-by-4 median filter. Here $1 \leq \eta \leq F$, where F is the total number of frames.

$$s'(\eta) = \sum_{\omega} \left| \frac{\partial}{\partial_{\eta}} \log S(\eta,\omega) \right| \tag{4.3}$$

$$r'(\eta) = \sum_{\omega} \left| \frac{\partial}{\partial_{\eta}} \log X(\eta, \omega) \right| \tag{4.4}$$

Fig 4.2 shows the spectrogram and temporal changes of speech signal $\boldsymbol{s}(\eta)$ and its residual $\boldsymbol{r}(\eta)$, where peaks in $s'(\eta)$ obtain the maximum temporal changes in those regions and hence segment boundaries. We can observe that most of the peaks that are present in $s'(\eta)$ obtained from a speech signal, are also present in $r'(\eta)$ obtained from a residual signal. This indicates that the residual signal does contain some information about segment boundaries. Hence, a detailed study is conducted to explore the significance of LP residual for acoustic-phonetic boundary detection.

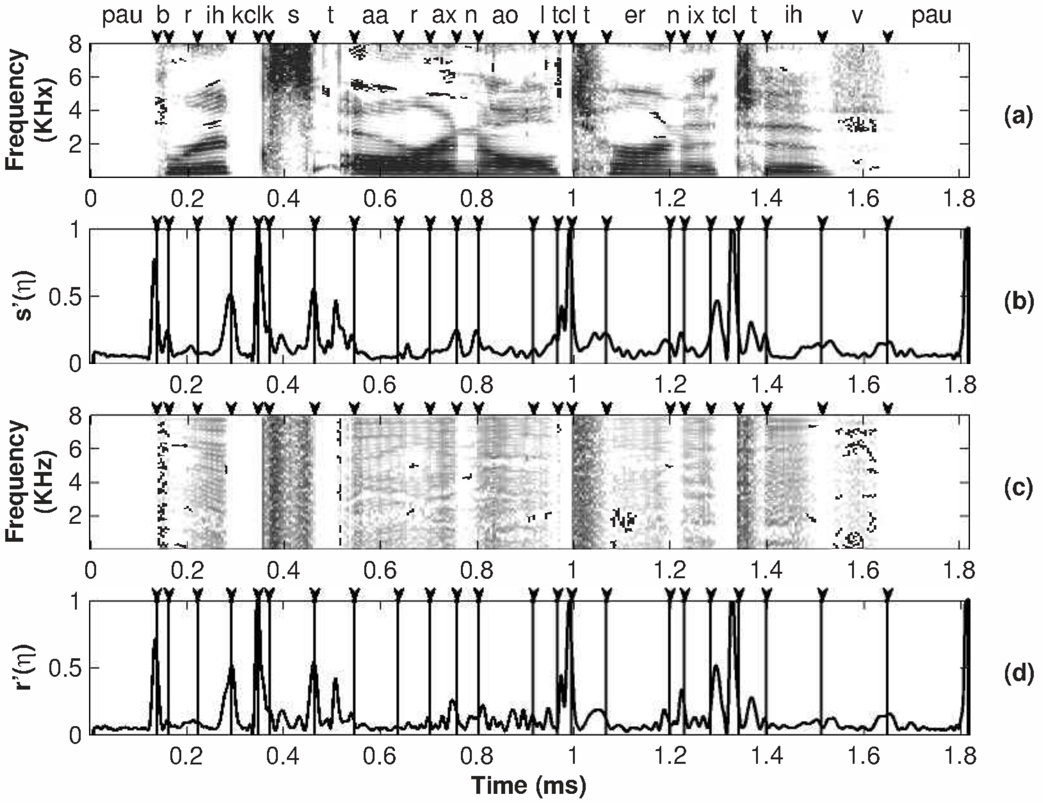

Figure 4.2: *(a) Spectrogram of speech signal, (b) $s'(\eta)$ obtained from speech signal, (c) Spectrogram of residual signal, (d) $r'(\eta)$ obtained from residual signal and all are marked with manual phone boundaries.*

4.2 Extraction of features from speech signal

A pre-emphasized speech signal is decomposed into a sequence of overlapping frames with 10 ms frame size (N samples) and an overlap of 5 ms to obtain filter parameters α_k and excitation parameters $r(n)$ using equations (4.1) & (4.2).

4.2.1 Filter based features

One of the important filter based representation of speech is linear prediction cepstral coefficients (LPCC). The cepstrum of a signal is computed by taking a Fourier (or similar) transform of the log spectrum. In the case of linear prediction cepstral coefficients, the required spectrum is the linear prediction spectrum which can be obtained from the Fourier transform of the filter coefficients. However, it can be shown that the required cepstra can be more efficiently computed using a simple recursion given below:

$$c(m) = \begin{cases} ln\ \sigma^2, & m = 0 \\ a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, & 1 \leq m \leq p \\ \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, & m > p \end{cases} \tag{4.5}$$

where σ^2 is the gain term in the LPC model, m is number of cepstral coefficients, and p is LP-order.

4.2.2 Excitation based features

LP residual is represented using Hilbert envelope of the residual signal. Significant excitation instants in LP residual correspond to glottal closure instants (GCI) which exists only in voiced segments. Hence such information could be helpful in identifying a boundary between voiced and unvoiced phones with better accuracy.

However, GCI in a residual have large error around the instants which can be reduced by obtaining the Hilbert envelope of the residual as shown in Fig 4.3 [50]. The analytic signal $r_a(n)$ corresponding to $r(n)$ is given by

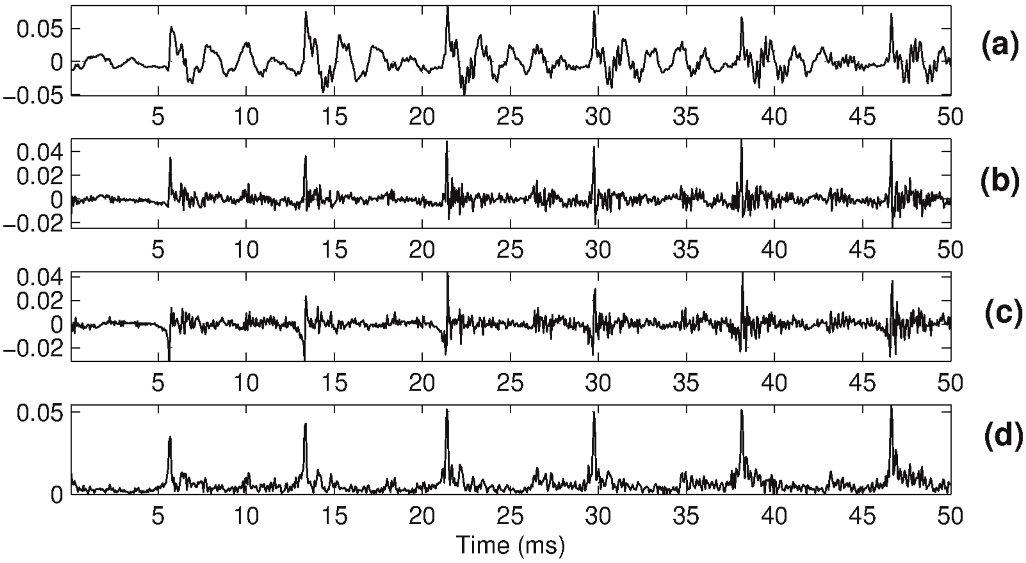

Figure 4.3: *(a) Speech signal $s(n)$, (b) LP residual $r(n)$, (c) Hilbert Transform of residual $r_h(n)$, (d) Hilbert envelope of residual $h_e(n)$.*

$$r_h(n) = IDFT[R_h(\omega)] \tag{4.6}$$

where

$$R_h(\omega) = \begin{cases} -jR(\omega), & 0 \leq \omega < \pi \\ jR(\omega), & -\pi \leq \omega < 0 \end{cases} \tag{4.7}$$

Here $R(\omega)$ is the Fourier transform of $r(n)$ obtained by using equation (4.9) and IDFT denotes the inverse discrete Fourier transform. Hilbert envelope $h_e(n)$ of the analytic signal $r_a(n)$ is given by equation (4.8). Once the Hilbert envelope is obtained, amplitude spectrum is estimated by applying Fourier transform on $h_e(n)$ using (4.9) and its cepstrum is estimated by using (4.10) on the obtained spectrum, where

$$h_e(n) = |r_a(n)| = \sqrt{r^2(n) + r_h^2(n)} \tag{4.8}$$

$$|H_e(\omega)| = \left| \sum_{n=0}^{N-1} h_e(n) e^{-j2\pi n\omega/N} \right| \tag{4.9}$$

$$v(m) = \frac{1}{N} \sum_{\omega=0}^{N-1} ln|H_e(\omega)| e^{j2\pi n\omega/N}, \quad \forall m \in [1, M] \tag{4.10}$$

Fig 4.4 shows the spectrograms of different features that are used in our experiments. In all the experiments, a 16^{th} order LP analysis is performed to extract 17 dimensional LPCC. Hilbert envelope of 16^{th} order LP residual is used to compute 15 dimensional HECC.

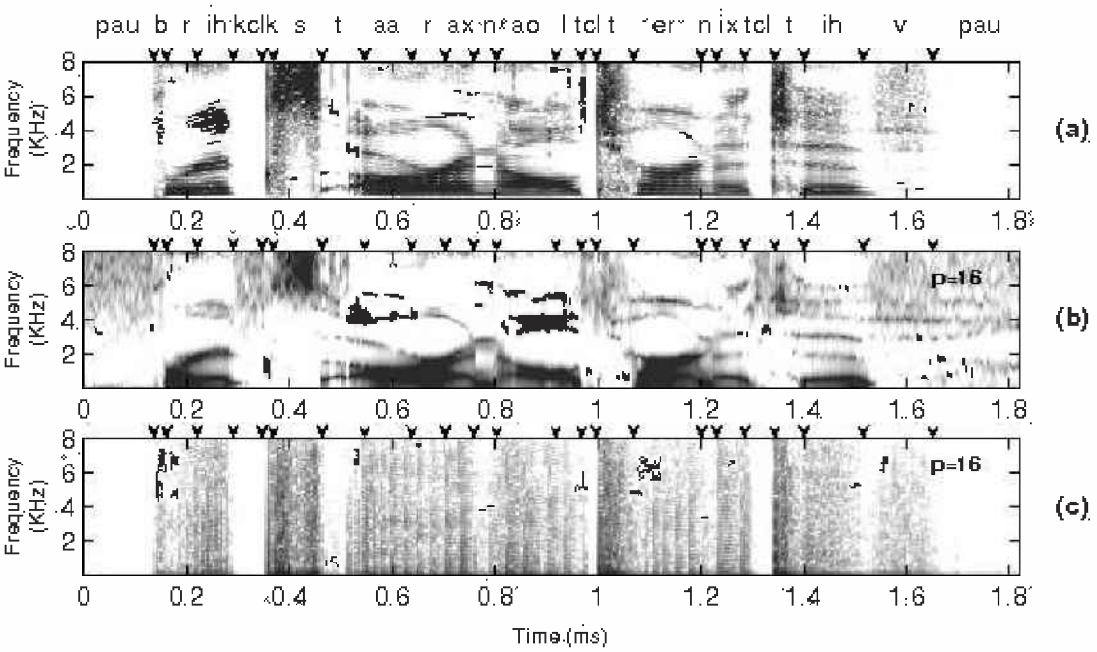

Figure 4.4: *Spectrogram of (a) Speech signal, (b) LP-Spectrum, (c) Hilbert envelope of* 16^{th} *order LP residual and all spectrograms are marked with manual phone boundaries.*

4.3 Acoustic-phonetic boundary detection

Acoustic-phonetic boundary detection can be performed using signal processing based approaches, HMM-based approaches or Classification based approaches as described in the previous chapter. Of all these approaches, classification based approaches out perform other approaches. Among these classification based approaches, supervised classification based approach (SC) performed better than unsupervised classification based approach (UC). Hence, we are employing supervised classification based approach (SC) to perform the experiments on different features. As described in the previous chapter, classification for SC was performed using artificial neural networks (ANN). Training phase of SC involves two phases: 1) Preparation of input / output data; and (2) Training ANN classifier using the above data. Boundary detection using SC involves the following three phases: (1) Obtaining acoustic scores of all frames in the speech signal using ANN classifier output; (2) Detection of boundary regions in speech signal; and (3) Location of boundaries in the speech signal. For details, please refer to the section 3.3.

4.4 Results and discussion

All our experiments are conducted on TIMIT corpus, which is recorded in a clean environment at 16 KHz sampling rate and has been labeled manually using 61 phones. Excluding "SA" files, this corpus has 3696 training files and 1344 testing files [3]. In all the experiments reported in this work, the value of l is fixed to be 5, i.e., a context of five frames to the left and right are used to create the augmented feature vector $\hat{\boldsymbol{x}}_t$. Table 4.1 shows the performance of acoustic-phonetic boundary detection for different features referred in section 4.2. Here LPCC and HECC are obtained from 16^{th} order LP analysis. The performance of acoustic-phonetic boundary detection is measured using boundary error rate (BER) and agreement percentage (AGR). For details, please refer to the section 3.4.

Table 4.1: *ANN configurations and performance of boundary detection for different feature sets. + in* 1^{st} *row denotes the concatenation of two feature sets. In* 3^{rd} *row L denotes linear activation and N denotes nonlinear tangent activation and the value indicate the number of perceptrons in the particular layer of ANN.* $5^{th} - 9^{th}$ *rows show the agreement percentage of correctly predicted boundaries using different features for different tolerance* (τ) *values.*

		LPCC	**HECC**	**LPCC+HECC**
Feature Dimention		17	15	32
ANN Configuration		187L 37N 2N	165L 33N 2N	352L 70N 2N
RMS (ms)		11.7	12.9	9.9
AGR	10ms	54.19	48.57	56.32
	20ms	89.69	86.24	91.26
	30ms	96.55	94.61	96.91
	40ms	98.66	97.69	98.76
	50ms	99.38	98.90	99.48
DEL %		20.55	17.95	13.26
INS %		7.92	11.82	8.36
BER %		28.47	29.78	21.62

Table 4.2: *Performance on C0's of LPCC, RCC and HECC of* 16^{th} *order LP analysis. ANN configuration used to train these features is 11L 5N 2N.*

	LPCC	**HECC**
RMS (ms)	13.7	16.4
AGR % ($\tau = 20ms$)	85.62	79.05
BER %	40.48	41.33

4.4.1 Usefulness of residual in detecting voiced-voiced and unvoiced-unvoiced boundaries

C0 of LPCC and HECC basically represent the log energy of the speech and the residual signals respectively. Fig 4.5 shows that C0 contours of LPCC and HECC of 16^{th} order LP analysis, which look similar except that C0 of LPCC is smoother than that of HECC. Table 4.2 shows that the performance of acoustic-phonetic boundary detection using only C0 of LPCC and HECC. From this table we can observe that even though there is an observable difference in RMS and AGR% between LPCC and HECC, there is almost no change in BER%. When the re-

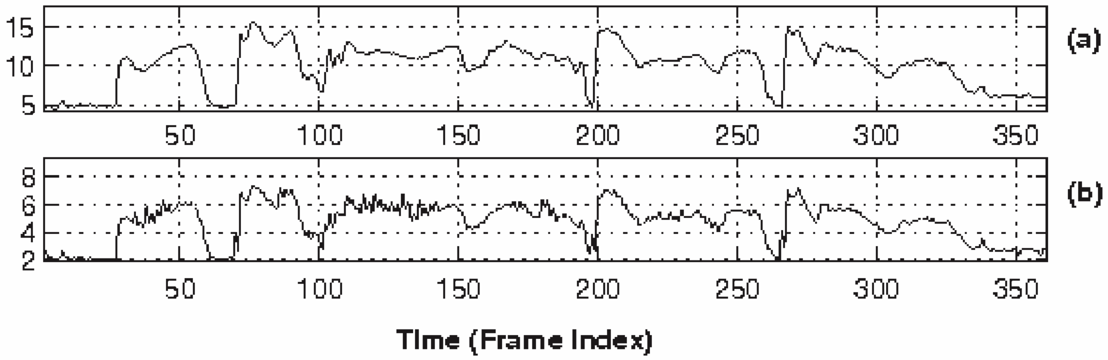

Figure 4.5: *C0 contour of (a) LPCC, (b) HECC, of* 16^{th} *order.*

Table 4.3: *Grouping of all phones into different broad phonetic classes.*

Class	Class Description	Phones	Voiced / Unvoiced
NAS	Nasal	eng, m, n, ng, nx	Voiced
SV	Semivowels	l, r, w, y	Voiced
VOW	Vowels	aa, ae, ah, ao, aw, ax, ax-h, axr, ay, eh, el, em, en, er, ey, ih, ix, iy, ow, oy, uh, uw, ux	Voiced
VC	Voiced Closure	bcl, dcl, gcl	Voiced
VF	Voiced Fricative	dh, hv, jh, v, z, zh	Voiced
VS	Voiced Stop	b, d, dx, g	Voiced
UVC	Unvoiced Closure	kcl, pcl, tcl	Unvoiced
UVF	Unvoiced Fricative	ch, f, hh, s, sh, th	Unvoiced
UVS	Unvoiced Stop	k, p, q, t	Unvoiced

spective BER performance of LPCC and HECC in Table 4.2 and Table 4.1 are compared, the improvement of 11.55% (40.48% to 28.47%) by LPCC and improvement of 12.01% (41.33% to 29.78%) by HECC suggest that the contribution of higher order cepstral coefficients of LPCC and HECC are very similar.

In most of the previous works, energy of the residual, which is proportional to C0 is used to detect the boundaries between voiced/unvoiced regions. When results in Table 4.2 and Table 4.1 are compared the changes in BER of HECC from 41.33% to 29.78%, AGR% (within 20 msec) of HECC from 79.05% to 86.24% and RMS of HECC from 16.4 msec to 12.9 msec show that the higher order cepstral coefficients of HECC contain information required for acoustic-phonetic boundary detection.

Table 4.4: *Boundary deletion percentages of each class pair (CPDEL) is computed for LPCC & HECC. Shaded cells correspond to class pairs for which HECC performed better. Of these shaded cells, class pairs with $\star$ corresponds to either voiced-voiced or unvoiced-unvoiced and class pairs with $\ddagger$ corresponds to either voiced-unvoiced or unvoiced-voiced.*

Class		**NAS**	**SV**	**VOW**	**VC**	**VF**	**VS**	**UVC**	**UVF**	**UVS**
NAS	# Class Pairs	72	288	1804	352	313	42	566	362	102
	LPCC	86.11	30.56	9.26	44.32	26.84	47.62	38.52	2.49	18.63
	HECC	75.00*	33.33	12.80	46.88	19.49*	26.19*	56.18	5.25	26.47
SV	# Class Pairs	120	97	3915	219	115	40	152	189	51
	LPCC	10.83	40.21	40.23	6.85	4.35	17.50	2.63	1.59	11.76
	HECC	20.83	32.99*	37.29*	5.02*	10.43	25.00	3.95	6.35	3.92‡
VOW	# Class Pairs	3499	2444	1286	1639	2016	599	2574	2255	413
	LPCC	11.23	60.72	47.51	3.60	6.15	25.21	2.95	1.06	19.85
	HECC	12.00	40.75*	41.29*	2.62*	15.72	22.54*	2.56‡	6.56	17.43‡
VC	# Class Pairs	46	58	65		357	1795		58	123
	LPCC	30.43	6.90	10.77		7.84	16.27		5.17	0.81
	HECC	36.96	31.03	24.62	NA	5.88*	9.86*	NA	5.17‡	0.00‡
VF	# Class Pairs	71	162	2162	177	102		201	155	48
	LPCC	18.31	8.64	16.65	14.69	21.57		8.46	37.42	4.17
	HECC	8.45*	17.90	25.21	11.86*	19.61*	NA	6.47‡	31.61‡	0.00‡
VS	# Class Pairs		411	2196		38			52	
	LPCC		36.25	45.31		63.16			63.46	
	HECC	NA	35.77*	42.99*	NA	73.68	NA	NA	61.54‡	NA
UVC	# Class Pairs	47	61	41		91	151		556	3254
	LPCC	46.81	14.75	7.32		3.30	2.65		5.94	4.21
	HECC	29.79‡	29.51	17.07	NA	6.59	2.65‡	NA	3.42*	4.27
UVF	# Class Pairs	27	309	2453	134	37		774	112	51
	LPCC	14.81	1.62	1.26	3.73	18.92		8.01	33.93	7.84
	HECC	11.11‡	7.12	4.77	0.75‡	16.22‡	NA	4.01*	34.82	0.00*
UVS	# Class Pairs		847	3043					304	31
	LPCC		5.31	14.79					76.32	22.58
	HECC	NA	12.99	16.60	NA	NA	NA	NA	73.03*	12.90*

In order to investigate the different types of boundaries captured by excitation based features other than voiced-unvoiced / unvoiced-voiced boundaries, a detailed analysis of boundary deletion is performed. To analyze the boundaries between each phone pair, a matrix of 61 phones against 61 phones is required, which is a large matrix and hence difficult to comprehend. In order to overcome this problem, we have grouped phones into nine broad phonetic classes as shown in Table 4.3. Thus the boundary deletions are computed only for 81 class pairs instead of 3721 (61 X 61) phone pairs. Table 4.4 shows the boundary deletion percentages of each class pair (CPDEL) for LPCC and HECC. CPDEL is computed using the equation (4.11). A smaller value of CPDEL indicates a better acoustic-phonetic boundary detection performance.

$$CPDEL = \frac{\#\ Deleted\ Class\ Pairs}{\#\ Class\ Pairs}\ X\ 100 \tag{4.11}$$

Out of these 81 class pairs, 16 class pairs in this table are marked as "NA" (Not Analyzed) as the number of examples for these class pairs is less than 20. So, only the remaining 65 class pairs were analyzed. Of these 65 class pairs, the acoustic-phonetic boundary detection performance of 36 class pairs (marked with ⋆ and ‡) was better using excitation based features (HECC) than filter based features (LPCC). The acoustic-phonetic boundary detection performance of the remaining 29 class pairs was better using LPCC. In this table there are nine cells which indicate the performances between same class pairs such as NAS-NAS, VF-VF, UVC-UVC etc. We can observe that of these nine same class pairs, five perform better using HECC (NAS-NAS, SV-SV, VOW-VOW, VF-VF, UVS-UVS), one perform better using LPCC (UVF-UVF) and the remaining three are NA (VC-VC, VS-VS, UVC-UVC). Of the 36 class pairs for which excitation based features (HECC) performed better than LPCC, 14 class pairs (markers with ‡) are voiced-unvoiced / unvoiced-voiced boundary class pairs and the remaining 22 class pairs (marked with ⋆) are voiced-voiced / unvoiced-unvoiced boundary class pairs. The above observations show that excitation based features (HECC) has potential to detect the boundaries even in voiced-voiced / unvoiced-unvoiced class pairs.

4.4.2 Complimentary nature of residual features

From the Table 4.4 we can observe that, even though LPCC performs better than HECC, when both are combined at feature level (LPCC+HECC) and an ANN network is trained using the concatenated feature vectors, then resulting models perform better than both LPCC and HECC i,e., the performance of LPCC+HECC is 21.62% BER, 91.26% AGR% (within 20 msec), and RMS is 9.9%. Hence this show that excitation (HECC) and filter (LPCC) based features are complimentary in nature for the task of acoustic-phonetic boundary detection.

4.5 Summary

In this work, we have investigated the significance of excitation based features represented by linear prediction residual for the task of acoustic-phonetic boundary detection. We have shown that the features extracted from LP residual contain useful information about the phone boundaries in speech signal. It was also observed that the excitation based features contain information about voiced-voiced and unvoiced-unvoiced segment boundaries along with voiced-unvoiced and unvoiced-voiced segment boundaries. Moreover, the evidence provided by the LP residual features were found to be of complementary in nature to vocal tract features. Further investigations has to be performed for extraction of better features from residual signal and better ways of combining the filter and excitation based features.

CHAPTER 5

Conclusion

Acoustic-phonetic boundaries of a speech signal can be detected using manual or automatic approaches. As mentioned in chapter 2, manual approaches have the following drawbacks: (1) Highly accurate, but tedious and time consuming; (2) Agreement between two annotations or between two annotators of the same signal are almost same but not same; These drawbacks of manual boundary detection, drives the need for automatic approaches. The automatic approaches can be broadly divided into (1) Signal processing approaches: These approaches do not require either the training data or the phonetic transcription. But these approaches are highly sensitive to parameters used. (2) Forced-alignment based approaches: These approaches require phonetic transcription and a training data to train the models like HMMs. These HMM based approaches obtain high accuracy boundary detection but they are dependent on the correctness of phonetic transcription. These approaches may also require manual boundaries to train the models. (3) Classification based approaches: These approaches obtain acoustic boundaries by employing boundary / non-boundary classifier on each frame to detect the boundary regions and apply a post processing techniques on these boundary regions to obtain location of boundaries. There are a few such approaches and even these approaches are supervised and hence require manual boundaries and sometimes even manual transcription. The limitation of this approach is that it requires a good amount of manually segmented data to train the classifier. In this thesis, we have explored signal processing methods such as mean spectral smoothing (MSS) and group-delay function (GDF) based approaches for acoustic-phonetic boundary detection. Apart from this, we have also explored HMM forced-alignment (HMM-FA) and HMM phone-loop (HMM-PL) based approaches using canonical transcription. We have developed a framework wherein manually or automatically obtained segment boundaries are used to train a boundary / non-boundary classifier. Once the classifier is trained, (1) Frame level acoustic scores are computed using classifier output; (2) Boundary regions are detected and ; (3) Boundaries are located. Supervised classification approach (SC) use manual boundaries to train the classifier. Unsupervised classification approach (UC) use the automatic boundaries generated by signal processing approaches such as MSS and GDF to train the classifier. A comparative study of all these approaches show that the classification based approaches outperform other approaches. Among these classification based approaches, SC performed better than UC. This can be justified as SC is trained on clean data where as UC is trained on noisy data.

Most of the present systems use traditional filter based features such as MFCC, filter-banks, LPCC etc., for acoustic-phonetic boundary detection ignoring the excitation based features. So, we have explored the significance of excitation based features for acoustic-phonetic boundary detection. This was motivated from the previous work by Markel et. al [12], which shows that even at comparatively high LP orders, the spectral flatness of the residual signal is not zero, concluding that the LP residual obtained using optimal LP-order has some information that can be used for speech processing. This observation was exploited and put to use by many researchers for the task such as voice activity detection (VAD), speaker recognition etc. Some of our spectrogram visualizations and observations hinted that there are cues even for phonetic boundary information in the residual. In this process, experiments were conducted on linear prediction cepstral coefficients (LPCC) and Hilbert-envelop cepstral coefficients (HEC) using supervised ANN based acoustic-phonetic boundary detection. We have shown that the features extracted from LP residual contain useful information about the phone boundaries in speech signal. It was also observed that the excitation based features contain information about voiced-voiced and unvoiced-unvoiced segment boundaries along with voiced-unvoiced and unvoiced-voiced segment boundaries. Moreover, the evidence provided by the LP residual features were found to be of complementary in nature to vocal tract features.

5.1 Future work

- **Robustness:** A large amount of work has to be done to make the classification based approaches and excitation based features more robust, so as to improve the performance not only with the clean speech, but also in highly coarticulated speech, and noisy speech.
- **Unsupervised boundary detection:** The performance of unsupervised approach is still not as good as SC. Nothing or very little can be done to improve the signal processing approach output, but major improvement has to come from the second phase, where it should be able to discard the erroneous classification samples and use only the correct once.
- **Phonetic labeling of speech segments:** Once the speech is segmented into phonetic segments, the next task is to label the segments with appropriate phonetic labels. This task can be basically defined as the task of classifying the segments into phonetic units.

REFERENCES

[1] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden markov models," *Speech Commun.*, vol. 12, pp. 357–370, August 1993.

[2] K. Joseph, S. Shai, S. Yoram, and C. Dan, "Phone alignment based on discriminative learning," in *Proceedings of Interspeech 2005*, Lisbon, Portugal, 2005.

[3] John-Paul Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Commun.*, vol. 51, pp. 352–368, April 2009.

[4] Alvin M. Liberman and Ignatius G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1 – 36, 1985.

[5] Cole A. Ronald and Scott Brian, "Toward a theory of speech perception," *Psychological Review*, vol. 81, no. 4, pp. 348 – 374, July 1974.

[6] J. P. H. van Santen, "Contextual effects on vowel duration," *Speech Commun.*, vol. 11, pp. 513–546, December 1992.

[7] S. E. G. O.hman, "Coarticulation in vcv utterances: Spectrographic measurements," *Journal of The Acoustical Society of America*, vol. 39, 1966.

[8] Douglas O'Shaughnessy Ladan Golipour, "A new approach for phoneme segmentation of speech signals," in *Proceedings of Interspeech*, Antwerp, Belgium, August 2007.

[9] Gerard Chollet, Anna Esposito, Marcos Faundez-Zanuy, and Maria Marinaro, *Nonlinear Speech Modeling and Applications: Advanced Lectures and Revised*

Selected Papers (Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[10] Youngjoo Suh and Younguk Lee, "Phoneme segmentation of continuous speech using multi-layer perceptron," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 3-6 1996, vol. 3, pp. 1297 –1300 vol.3.

[11] Keri Venkatesh and Prahallad Kishore, "A comparative study of constrained and unconstrained approaches for segmentation of speech signal," in *Proceedings of Interspeech 2010*, Makuhari, Japan, 2010.

[12] John E. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1982.

[13] Philippe Thévenaz and Heinz Hügli, "Usefulness of the lpc-residue in text-independent speaker verification," *Speech Commun.*, vol. 17, pp. 145–157, August 1995.

[14] P Cosi, D Falavigna, and M Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies," *European Conference on Speech Communication and Technology*, 1991.

[15] Ljolje Andrej, Hirschberg Julia, and Jan P H van, "Automatic speech segmentation for concatenative inventory selection," in *Proceedings of Speech Synthesis, Springer-Verlag*, New York, USA, 1997.

[16] Maria B. Wesenick and Andreas Kipp, "Estimating the quality of phonetic transcriptions and segmentations of speech signals," in *Proceedings of the ICSLP*, 1996, pp. 129–132.

[17] Hong Leung and V. Zue, "A procedure for automatic alignment of phonetic transcriptions with continuous speech," in *Acoustics, Speech, and Signal Pro-*

cessing, IEEE International Conference on ICASSP '84., mar 1984, vol. 9, pp. 73 – 76.

[18] Ronald Cole, Beatrice T. Oshika, Mike Noel, Terri Lander, Terri L, and Mark Fanty, "Labeler agreement in phonetic labeling of continuous speech," in *In Preceedings of ICSLP*, Yokohama, Japan, Spetember 1994.

[19] Knut Kvale, *Segmentation and labelling of speech*, Ph.D. thesis, Institutt for Teleteknikk, Trondheim, 1995.

[20] K. Torkkola, "Automatic alignment of speech with phonetic transcriptions in real time," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, apr 1988, pp. 611 –614 vol.1.

[21] A. Van Erp and L. Boves, "Manual segmentation and labelling of speech," in *In Proceedings of Speech-88*, 1988, pp. pp. 1131–1138.

[22] J. Wilpon, B. Juang, and L. Rabiner, "An investigation on the use of acoustic sub-word units for automatic speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, apr 1987, vol. 12, pp. 821 – 824.

[23] Manish Sharma and Richard J Mammone, ""blind" speech segmentation: automatic segmentation of speech without linguistic knowledge," in *In Proceedings of ICSLP*, 1996.

[24] Sorin Dusan and Lawrence Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *Proceedings of Interspeech 2006*, Pittsburgh, PA, USA, 2006.

[25] Y.G. Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, april 2007, vol. 4, pp. IV–937 –IV–940.

[26] A. Vorstermans, J. P. Martens, and B. Van Coile, "Automatic segmentation and labelling of multi-lingual speech data," *Speech Communication*, vol. 19, no. 4, pp. 271 – 293, 1996.

[27] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, apr 1987, vol. 12, pp. 77 – 80.

[28] D Falavigna and M. Omologo, "A dtw-based approach to the automatic labeling of speech according to the phonetic transcription," in *In Proceedings of the European Signal Processing Conference*, Barcelona, Spain, 1990, pp. pp. 1139–1142.

[29] S. Rapp, "Automatic phonetic transcription and linguistic annotation from known text with hidden markov models / an aligner for german," in *In Proceedings of ELSNET Goes East and IMACS Workshop*, Moscow, Russia, 2000.

[30] P. Dalsgaard, O. Andersen, and W. Barry, "Multi-lingual label alignment using acoustic-phonetic features derived by neural-network technique," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, apr 1991, pp. 197 –200 vol.1.

[31] Paul Dalsgaard, "Phoneme label alignment using acoustic-phonetic features and gaussian probability density functions," *Computer Speech and Language*, vol. 6, no. 4, pp. 303 – 329, 1992.

[32] P. Dalsgaard, O. Andersen, W. Barry, and R. Jorgensen, "On the use of acoustic-phonetic features in interactive labelling of multi-lingual speech corpora," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, mar 1992, vol. 1, pp. 549 –552 vol.1.

[33] Falavigna D. Guiliani D. Gretter R. Angelini B., Brugnara F. and Omologo M., "Automatic segmentation and labeling of english and itelian speech

databases," in *In Proceedings of Eurospeech*, Berlin, Germany, september 1993, pp. pp. 653–656.

[34] Falavigna D. Brugnara F. and Omologo M., "A hmm-based system for automatic segmentation and labeling of speech," in *In Proceedings of ICSLP*, Banff, Alberta, Canada, October 1992, pp. 803–806.

[35] Steve Young, *The HTK Book*, Cambridge, 2009.

[36] Steffen Pauws, Yves Kamp, and Lei Willems, "A hierarchical method of automatic speech segmentation for synthesis applications," *Speech Communication*, vol. 19, no. 3, pp. 207 – 220, 1996.

[37] Wesenick M.-B. Kipp, A. and F. Schiel, "Automatic detection and segmentation of pronunciation variants in german speech corpora," in *In Proceedings of ICSLP*, Philadelphia, PA., October 1996, pp. 106–109.

[38] C. W. Wightman and D. T. Talkin, "The aligner: Text-to-speech alignment using markov models," in *Proceedings of Speech Synthesis, Springer-Verlag*, New York, USA, 1997.

[39] Pollom B. L. and Hansen J. H. L., "Automatic segmentation and labeling of speech recorded in unknown noisy channel environments," in *Proceedings of the ESCA-NATO Workshop on Robust Speech Recognition for unknown communication channels*, 1997.

[40] D.T. Toledano, L.A.H. Gomez, and L.V. Grande, "Automatic phonetic segmentation," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 617 – 625, nov 2003.

[41] Iosif Mporas, Todor Ganchev, and Nikos Fakotakis, "Speech segmentation using regression fusion of boundary predictions," *Computer Speech and Language*, vol. 24, no. 2, pp. 273 – 288, 2010.

[42] Iosif Mporas, Todor Ganchev, and Nikos Fakotakis, “Speech segmentation using regression fusion of boundary predictions,” *Computer Speech & Language*, May 2009.

[43] B. Yegnanarayana, *Artificial Neural Networks*, Prentice-Hall of India Pvt.Ltd, 2004.

[44] S Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[45] Petr Horák and Betty Hesounová, “Automatic speech segmentation with the application of the czech tts system,” in *Proceedings of the Third International Workshop on Text, Speech and Dialogue*, London, UK, 2000, TDS ’00, pp. 201–206, Springer-Verlag.

[46] A. I. C. Monaghan and D. R. Ladd, “Speaker-dependent and speaker-independent parameters in intonation,” in *Proceedings of ESCA, Speaker Characterisation in Speech Technology*, Edinburgh, 1990.

[47] B. Yegnanarayana, K. Sharat Reddy, and S.P. Kishore, “Source and system features for speaker recognition using aann models,” *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 409–412, 2001.

[48] Lawrence Rabiner and Biing H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, united states ed edition, April 1993.

[49] A. H. Gray and J. D. Markel, “A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis,” *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 22, no. 3, pp. 207–217, 1981.

[50] K.S. Rao, S.R.M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *Signal Processing Letters, IEEE*, vol. 14, no. 10, pp. 762 –765, October 2007.