SPEAKER SEGMENTATION USING EXCITATION SOURCE FEATURES

A THESIS

submitted by

DHANANJAYA. N

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

JUNE 2004

ACKNOWLEDGEMENTS

First and foremost, I express my sincere gratitude to Prof. B. Yegnanarayana for his invaluable guidance all through my research work. I thank him for creating a stimulating and unrestrained environment for research, which is very much essential for a young aspirant researcher. His extraordinary patience, unmatched enthusiasm and never say die attitude has been a constant source of inspiration for me to continue in the field of research.

I thank Dr. C. Chandra Sekhar for all the invaluable advise and constructive criticism. His critical views, keenness to excel, and a very friendly attitude are some of the characters I would like to imbibe.

I thank Dr. Hema A. Murthy for all the thought provoking discussions, both technical and non-technical. Being one of the most active persons on the campus, she has really been an inspirational character for me.

I thank Prof. C. Pandurangan and Prof. S. Raman, chairmen of the department during my MS research for providing me with constant support and excellent facilities to carry out my research. I am highly indebted to Prof. S. Raman for all the help he has extended for my continuation in the field of research.

I thank my GTC members, Dr. Deepak Khemani and Prof. K. N. Bhat for their useful advise and for sparing their valuable time to evaluate the progress of my research work.

I thank Dr. S. Rajendran for the constant support and encouragement in terms of advice and resources. I am grateful to him for the confidence he has shown in me and

providing me an opportunity to learn some of the basics in system administration.

I am grateful to all the faculty and non-teaching staff within the department as well as outside the department for all the help I have received during my stay at IIT Madras.

I thank all my lab members Surya-sir, Prasanna-sir, Rao-sir, Nayeem-ji, Puttu-sir, CKM-sir, Pal-sir, Anil-bhai, motu-Suresh (dilwala), Satish-pyare, SR-Chaitu, feelings-Murthy, Anand (smart boy), Leena, Shahina-ji, Anitha, Anvita, Vinod, Venky (ultimate torture m/c), Sharat, Gupta, Kiran, Jinu, Ram, Shiva and Guru for creating a very friendly and conducive atmosphere within and outside the lab. I cherish the moments of fun we all shared and thank everyone for all the cooperation, understanding and help I have received.

My special thanks to Rao-sir, Guru, Prasanna-sir and Surya-sir without whose help and brotherly care this thesis would not have come to this stage. I thank Kumari-akka, Anitha-akka and Nirmala-akka for all the care, affection and tasty food, without which staying away from home for such a long period would have been difficult.

I would like to thank all my teachers and friends who have helped me reach this stage in life. I am grateful to my teacher and guide during my B.E. Dr. Ananthashayana who kindled my interests in signal processing, without which I would probably not have entered this field.

Finally, I would like to dedicate this work to my parents (Akka-Anna), beloved sister (Rekha), darling niece (chinnu) and dearest friend Guru for all the confidence, love and affection they have showered upon me.

- Dhananjaya. N

ABSTRACT

Keywords: *Speaker segmentation; speaker change detection; multispeaker speech; speaker turn; speaker separation; speaker tracking; 2-speaker detection; segmentation cost; autoassociative neural network; within-speaker to across-speaker dissimilarity ratio.*

Speaker segmentation involves detection of speaker changes in a multispeaker speech signal. Speaker segmentation or speaker change detection is the first step in applications like 2-speaker detection, transcription of multispeaker speech, forensic investigations and audio indexing, which involve processing of multispeaker data. In this research work we address the issue of detecting speaker changes in a casual 2-speaker conversation, which contains short speaker turns. The existing approaches for speaker segmentation depend mainly on vocal tract characteristics of speakers, to detect a speaker change. They rely on the dissimilarity of distributions, of the feature vectors, estimated from two adjacent windows of speech. This requires significant amount of data ($>$ 5 sec) in each of the windows, and hence the existing approaches are best suited for applications that handle multispeaker data with long speaker turns (e.g. transcription of broadcast news). This statistical approach to a point phenomenon (speaker change) fails when the given conversation involves short speaker turns ($<$ 5 sec). Casual conversations contain a large number of short speaker turns, and an automated, accurate segmentation of these conversations is crucial for tasks handled by forensic and intelligence agencies. In this thesis, we explore the possibility of using excitation source information as an alternate feature for speaker segmentation. The excitation source signal energizing the vocal tract system during the production of voiced speech, has significant information for characterizing a speaker. Linear prediction (LP) residual, obtained by removing the vocal tract information from the speech signal, is a good approximation of the excitation source signal. An autoassociative neural network (AANN) model can be used to capture the higher order correlations among samples of the LP residual signal, at a subsegmental level (less than a pitch period). Within-speaker to across-speaker dissimilarity (WAD) ratio is proposed as a new measure to evaluate the ability of a given feature set in characterizing a speaker. Excitation features are shown to have lesser dissimilarity across sounds within a speaker, as compared to the vocal tract features. Hence excitation source features are better suited for characterizing a speaker from limited amount of voiced speech ($<$ 5 sec). The ability of an AANN model to capture the subsegmental excitation characteristics of a speaker from limited data, is used to propose a new approach for speaker segmentation. The proposed approach using excitation source features works better than the commonly used approach based on vocal tract features, in segmenting casual 2-speaker conversations. The 2-speaker detection task, which is of significance in forensic applications, is considered as an application of speaker segmentation. The performance of a 2-speaker detection system is shown to improve when speaker segmentation is performed on the training and test conversations, as compared to the case when no segmentation is done.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

1sp	- 1-speaker detection
2sp	- 2-speaker detection
AANN	- Autoassociative Neural Network
BIC	- Bayesian Information Criterion
EER	- Equal Error Rate
FAR	- False Acceptance Rate or False Alarm Rate
GC	- Glottal Closure
GMM	- Gaussian Mixture Model
KL	- Kullback-Leibler
LLR	- Log-Likelihood Ratio
LP	- Linear Prediction
LPC	- Linear Prediction Coefficients
LPCC	- Linear Prediction Cepstral Coefficients
LSP	- Line Spectral Pairs
MDR	- Missed Detection Rate
MFCC	- Mel-Frequency Cepstral Coefficients
ML	- Maximum Likelihood
NIST	- National Institute of Standards and Technology
NN	- Neural Network
PAP	- Periodic-Aperiodic
SCD	- Speaker Change Detection

SNR	- Signal to Noise Ratio
TOC	- Table of Content
VQ	- Vector Quantization
WAD	- Within-speaker to Across-speaker Dissimilarity

CHAPTER 1

SPEAKER SEGMENTATION - AN INTRODUCTION

The source of production of sound is an important criterion for classifying sounds into broad categories. Speech produced by human beings is one such important category of sounds. A continuous stream of audio can contain sounds produced by different sources, including that produced by humans. The process of dividing an audio stream into smaller homogeneous segments, based on a specific criterion, is termed as audio segmentation [1–3]. The segmented regions of speech in turn can contain more than one speaker speaking over non-overlapped periods of time. Such a multispeaker speech can be further segmented using speaker identity as a criterion. A conversation between two or more speakers is the most natural form of multispeaker speech, and the task of segregating speakers in a conversation is studied in this thesis.

1.1 What is Speaker Segmentation?

Speaker segmentation involves processing multispeaker speech signals. A *multispeaker speech* signal contains speech from two or more speakers, speaking one after the other over non-overlapped intervals of time. Given a multispeaker speech signal, the task of identifying all instants of time where there is a change in speaker, is termed as *speaker segmentation* or *speaker change detection*. The *speaker changes* are also referred to as *speaker boundaries* or *speaker transitions*. Speaker segmentation results in the multispeaker speech divided into smaller segments of speech termed as speaker turns. A *speaker turn* is defined as a segment of speech wherein a single speaker is talking, from the time he has taken over from another speaker and till the time he is taken over by any other speaker. *Speaker tracking* is a problem closely related to the speaker segmentation task. It involves speaker change detection followed by a speaker separation task. *Speaker separation* is the task of segregating segments of speech, with each segment containing only one speaker, into as many groups as the number of speakers in the conversation. Speaker tracking here refers to tracing a speaker within a conversation and it should not be confused with tracking a speaker in space (*source localization* problem).

1.2 Need for Speaker Segmentation

Speaker segmentation is the first step in most of the applications that involve processing multispeaker speech. Some of the important applications that handle multispeaker data are as follows:

- Forensic applications
  - Speaker tracking
  - 2-speaker (2sp) detection
  - Recognition of conversational speech
- Audio indexing
  - Transcription of broadcast news
  - Speaker based indexing

The volume of voice traffic over various channels of communication is so high that storage of all this data for analysis at a later stage is not viable. Even if it is stored, manual analysis of this huge data is tedious. It has been found that these conversations carry significant evidence for the prevention of crimes, if only they are detected in advance. Intelligence agencies and forensic experts have interests in automating this analysis of conversations. Some of the information of interest for a forensic expert are as follows:

- How many speakers are involved in the conversation?
- Is there a familiar or known voice?
- What is the topic of discussion?
- Who said what and when?
- What is the background environment in which a speaker is talking?

The accuracy of these information is of high significance, and hence there is a need for reliable techniques to analyze multispeaker speech. Tracking a speaker within a conversation, detection of a common speaker between two different conversations and identification of a known speaker in a conversation are tasks of significance in forensic applications [4] [5]. A speaker segmentation followed by speaker separation can simplify these tasks. The 2-speaker detection problem and the role of speaker segmentation in detecting speakers in a conversation will be explained in greater detail in Chapter 5.

The volume of data stored in the form of audio and video documents is increasing, due to increasing storage capacity, decreasing storage cost and increasing computing power. Some examples of audio documents are broadcast news archives, interviews, discussions over a meeting, voice mails, dialogues from a movie, lectures and presentations, important speeches and music albums. Now the challenge is to automatically analyze, label and index the audio documents for efficient organization and easy retrieval. Fig. 1.1 shows some of the tasks involved in audio indexing [2,3,6–9].

Transcription of multispeaker speech is an important task in audio indexing applications [7,10]. Conventional speech recognizers work best on speech from a single speaker and the performance degrades if the given speech is a conversation involving two or more speakers. Speaker adaptive speech recognizers [11–13] are used in such a scenario, provided the speech regions of the conversant speakers are separated out.

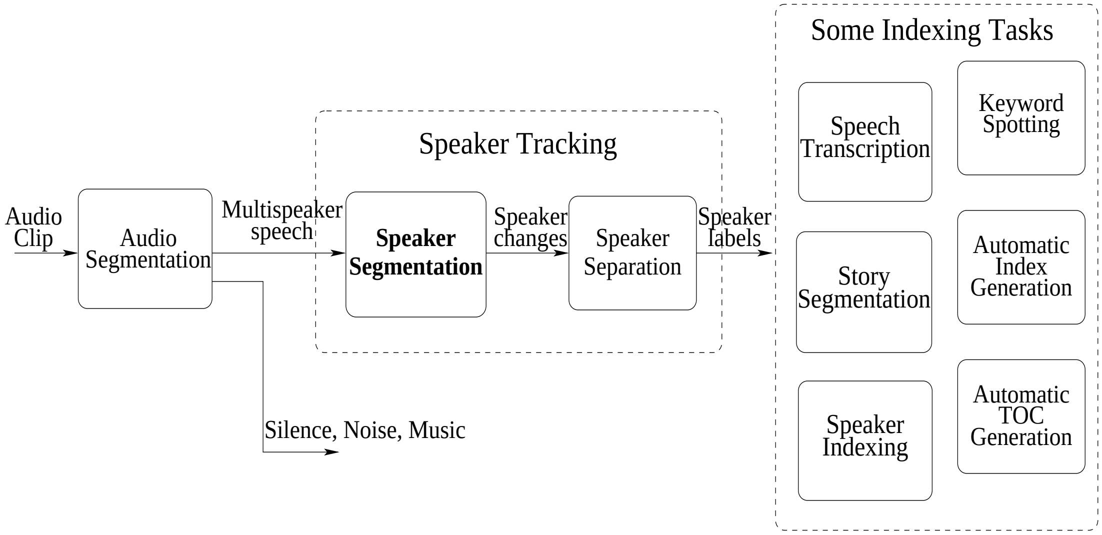

Fig. 1.1: An overview of some of the audio indexing tasks.

1.3 Issues in Speaker Segmentation

The two major issues in speaker segmentation are as follows:

1. Nature of multispeaker data

2. Robustness of the features

1.3.1 Nature of multispeaker data

The multispeaker speech data can be from a wide variety of sources. It can be a casual conversation, a discussion over a formal meeting, a broadcast news or an audio clip from a movie, to mention a few. Some of the important issues in segmenting different types of multispeaker speech are:

- Amount of data available for analyzing speaker changes
- Lack of *a priori* information
- Overlapped speech, laughter and other sounds
- Background noise, speech or music

Formal conversations like interviews and television or radio news bulletins have speakers normally talking for longer durations, whereas casual conversations have frequent speaker changes with several monosyllabic sounds, laughter and simultaneous speaking. Hence, the amount of data available on either side of a speaker change varies with the type of conversation. This research work addresses issues in detecting speaker changes in casual conversations.

Two important *a priori* information about a conversation are the number of speakers involved in the conversation and the identity of the speakers (i.e., reference data for the speakers). The focus of this thesis is on the segmentation of 2-speaker conversations. Though the number of speakers is known in advance, there is no reference data available for the speakers, *a priori*.

In a casual conversation, the probability of speakers speaking simultaneously, laughter and other monosyllabic sounds of gesture, is high. Identification of such regions may be essential in applications like 2-speaker detection, as *impure data* (speech data of a speaker corrupted by traces of a second speaker) can degrade the performance of the verification task [14]. The detection or analysis of such regions of speech in a conversation is not handled in this work.

The channel over which the conversations are recorded, background noise and the possibility of background music (in case of audio) contribute to the complexity of the problem [15]. Channel normalization [16–18] may be essential to avoid detecting channel changes as speaker changes. The data considered in this work are telephonic conversations between two speakers, recorded at a sampling rate of 8 kHz. The speech data is clean and it is assumed that there is no background music. The channel issues are not studied in this work.

1.3.2 Robust features

Speech is a composite signal which has information about the message, the speaker identity and the language [19–21]. It is difficult to isolate the speaker specific features alone from the signal. Also, speaker segmentation of casual conversations requires that speakers be characterized from limited data. The variability of the chosen features within a speaker is a major issue in speaker characterization problems. Features capable of characterizing a speaker from limited data are explored in this thesis.

1.4 Motivation for Speaker Segmentation

The ability of humans to detect speaker changes almost effortlessly signifies that there is ample evidence for automatic detection of speaker boundaries. The speaker characteristics in a speech signal can be attributed to the structure of the vocal tract, the rate and manner of vibration of the vocal folds, and the idiosyncrasies of the speaker [19,21]. The vocal tract characteristics are usually referred to as spectral features and are widely used for speaker characterization [17,22–24]. These spectral features require large amount of speech data for characterizing a speaker. Most approaches for speaker segmentation use spectral features and focus on multispeaker speech which have long speaker turns (e.g. broadcast news). These approaches fail for casual conversations which typically have short speaker turns. The characteristics of the vocal folds, which are referred to as the excitation source features are usually neglected, except the rate of vibration (pitch). The excitation features at the subsegmental level (less than a pitch period) have significant speaker information [25]. Speaker characterization using subsegmental excitation features require lesser amount of speech data, as compared to the vocal tract features [26]. Hence we explore the possibility of using the excitation source features at the subsegmental level for speaker segmentation.

1.5 Organization of the Thesis

Chapter 2 gives an overview of the existing approaches for speaker change detection. The drawbacks of the existing approaches and hence the need for alternate approaches are discussed in the chapter. Chapter 3 discusses the possibility of using excitation source features for speaker segmentation, as an alternative to the commonly used vocal tract features. The ability of autoassociative neural network models to capture the excitation source features at a subsegmental level is also discussed. A new parametric measure to evaluate the suitability of a chosen set of features for characterizing a speaker from limited amount of data is introduced in this chapter. A new approach for speaker change detection using the excitation source features and neural network models is proposed in Chapter 4. The performance of the proposed approach is analyzed and compared with the existing approach using vocal tract features. The usefulness of speaker segmentation in enhancing the performance of 2-speaker detection task is illustrated in Chapter 5. Chapter 6 summarizes the research work carried out as part of this thesis, highlights the contributions of the work and discusses the directions for further research.

CHAPTER 2

APPROACHES FOR SPEAKER SEGMENTATION

The basic principle of the widely used approach to speaker segmentation and a brief review of some of the work done on speaker segmentation are given in this chapter. The limitations of the existing methods in processing conversational speech, that has short speaker turns, is brought out. The significance of short speaker turns and hence the need for alternate features and techniques for speaker change detection is emphasized.

2.1 Speaker Segmentation using Vocal Tract Features

The widely used approach for speaker change detection works on the principle outlined in Fig. 2.1. The input multispeaker speech is converted into a sequence of acoustic features. Short time (10-30 ms) spectral analysis of speech [21] with overlapped windows (5-15 ms) is performed to convert the multispeaker speech into a sequence of acoustic features. A dissimilarity value is computed between a pair of adjacent windows (5-10 sec). A sequence of dissimilarity values is computed by moving the pair of windows by a constant shift. A large dissimilarity hypothesizes a speaker change. As it is difficult to set an absolute threshold on the dissimilarity values, the locations of peaks in the dissimilarity sequence are hypothesized as the speaker changes. This principle is often referred to as metric-based approach, as it uses a dissimilarity metric to compare two sets of features from two adjacent windows of speech.

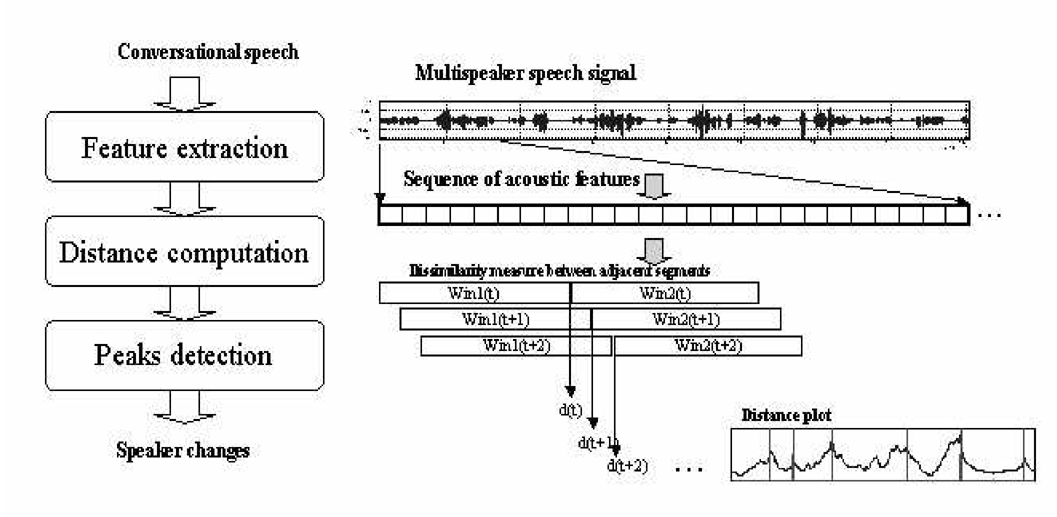

Fig. 2.1: Metric-based approach for speaker segmentation using vocal tract features.

2.2 Review of Existing Approaches

The metric-based approach outlined in Section 2.1 is widely used approach for speaker segmentation [4, 7, 15, 27–29]. Vocal tract characteristics of a speaker are commonly used as acoustic features for change detection. Cepstral coefficients [21] and line spectral pairs (LSP) [23] are the widely used vocal tract features. A variety of dissimilarity measures (also called distance or distortion measures) can be employed to compare two sets of features [21, 30–37]. The most commonly used metrics are Bayesian information criterion (BIC), log-likelihood ratio (LLR) and Kullback-Leibler (KL) divergence. Transcription of broadcast news archives is the application widely considered for the speaker segmentation task.

Speaker separation is an essential task in most of the applications considered for speaker segmentation [4, 7, 15, 28, 29]. The speaker separation task is treated as a clustering problem (unsupervised classification). The first step is to split the given conversation into smaller segments of speech such that each segment contains only one speaker. This is done either by detecting the speaker changes or by uniformly splitting the data into small segments of the same size, say one second. These segments of speech are then clustered in an unsupervised manner so as to form one cluster for each speaker. A range of clustering methods [38–44] have been studied. The simplest but widely used is the agglomerative clustering, which combines two closest segments at a time. At the end of clustering process, each cluster is hypothesized to represent one speaker. The stopping criterion for clustering has been a major problem when the number of speakers is not known *a priori* [15, 45]. Vector quantization (VQ) based speaker clustering is used for speaker change detection in [46]. Different clustering strategies have been used to solve the speaker separation problem without performing an exclusive change detection before clustering [47, 48].

As a continuation of the above two steps, a model is built for each speaker using the clustered data [15]. The feature vectors are reclassified using these speaker models. Gaussian mixture models (GMM) [49] are commonly used for this purpose [50]. Neural network (NN) models have also been used for modeling speakers [51].

Gish *et al.* [4] have used the generalized log-likelihood ratio as a dissimilarity measure to compare two distributions estimated using maximum likelihood estimation. Mel-frequency cepstral coefficients (MFCC) are used as the features. The conventional agglomerative technique is used for clustering the segmented data. Tracking the commands given by an air traffic controller to the pilots is studied as an application. As the percentage of controller's speech is high compared to that of individual pilots, the largest cluster is labeled as that of the controller. Bayesian information criterion (BIC) was used by Chen and Gopalakrishna [27] to define a dissimilarity measure to compare two sets of features, based on the comparison of their parametric statistical models. The use of BIC for speaker change detection is discussed in the next section. An approach similar to [4] is used by John Makhoul et al. [7] for segmenting speakers in their system *'Rough'n'Ready'* for audio indexing. Transcription of multispeaker speech (mostly broadcast news) for automatic indexing and retrieval is the application studied. Delacourt *et al.* [29] have discussed a 2-pass algorithm for speaker segmentation. In the first pass, speaker changes are hypothesized using the Kullback-Leibler (KL) distance, which are validated in the next pass using BIC measure. The threshold on distance values for speaker change hypotheses is found to be sensitive to the type of data under consideration. Different distance measures and clustering techniques have been studied and compared by Johnson [28], to separate speakers in news broadcasts for the purpose of speech transcription. The existing techniques make an assumption that the speaker turns are long enough ($>$ 5 sec). This affects the segmentation performance severely in case of casual conversations which contain a large number of short speaker turns ($<$ 5 sec).

2.3 BIC for Speaker Change Detection

This is a metric-based approach which uses the Bayesian information criterion to define a dissimilarity metric and will be referred to as BIC approach. The BIC is a maximum likelihood criterion penalized by the model complexity (the number of model parameters), widely used for model selection [43, 52]. It can be used for change detection in a sequence of acoustic features [27]. If $\mathcal{Z} = \{\boldsymbol{x}_k\}, \ k = 1, \ldots, N_Z$ is a sequence of feature vectors, then the possibility of a speaker change at $k = N_X$ $(N_X < N_Z)$ can be examined by testing the two hypotheses:

- $\mathcal{H}_0$: the entire sequence $\mathcal{Z}$ is generated by a single speaker and is thus assumed to be represented by a single multivariate Gaussian process $\mathcal{M}_Z(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$.

- $\mathcal{H}_1$: the sequence $\mathcal{X} = \{\boldsymbol{x}_k\},\ k = 1, \ldots, N_X$ belongs to one speaker and the sequence $\mathcal{Y} = \{\boldsymbol{x}_k\},\ k = N_X + 1, \ldots, N_Z$ (where $N_Z = N_X + N_Y$) belongs to a different speaker, represented by two separate Gaussian processes $\mathcal{M}_X(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\mathcal{M}_Y(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$, respectively.

Here $(\boldsymbol{\mu}_X,\ \boldsymbol{\Sigma}_X)$, $(\boldsymbol{\mu}_Y,\ \boldsymbol{\Sigma}_Y)$ and $(\boldsymbol{\mu}_Z,\ \boldsymbol{\Sigma}_Z)$ are the maximum likelihood estimates of the mean vectors and covariance matrices of the three multivariate Gaussian processes $\mathcal{M}_X$, $\mathcal{M}_Y$ and $\mathcal{M}_Z$, respectively. The difference in the BIC values of the two hypotheses (d_{BIC} or ΔBIC) is given by

$$d_{BIC} = R - \tau P \tag{2.1}$$

where

$$R = \frac{N_X}{2} log|\boldsymbol{\Sigma}_X| + \frac{N_Y}{2} log|\boldsymbol{\Sigma}_Y| - \frac{N_Z}{2} log|\boldsymbol{\Sigma}_Z| \tag{2.2}$$

is the log-likelihood ratio, $P = \frac{1}{2}(p + \frac{1}{2}p(p+1)) \times logN_Z$ is the penalty factor, where p is the dimension of the feature vector and τ is a constant which acts as a threshold for decision making. A positive value of ΔBIC indicates that two multivariate Gaussian models best fit the sequence $\mathcal{Z}$, which in turn means that a speaker change is hypothesized at $k = N_X$. The threshold parameter τ is sensitive to channel and other variabilities in the speech data and requires fine-tuning for the type of data under consideration.

2.4 Need for Alternate Approaches

The existing approaches, as outlined in the previous sections, use vocal tract features and employ a statistical approach to detect a point phenomenon (speaker change).

The limitations of the existing approaches to handle conversational speech and the significance of short speaker turns will be discussed in this section.

2.4.1 Limitations of the existing approaches

The existing approaches for speaker change detection rely on features and dissimilarity measures which are widely used in speaker verification applications [22,23,30,37]. The vocal tract features vary significantly for different sounds within a speaker and hence a large number of examples for each type of sound are necessary for representing a speaker. Additionally, the dissimilarity measures compare distributions of feature vectors from two segments of speech. The estimation of these multivariate probability distributions require large amount of data and the performance of the text-independent speaker verification task degrades significantly as the amount of data reduces below 30 sec [14]. The existing approaches for speaker change detection use a window size of around 5 sec, and consider applications (e.g. transcription of broadcast news) which handle multispeaker data with long speaker turns. As the window size reduces below 5 sec, the reliability of the dissimilarity scores for speaker change detection reduces, due to inaccurate estimations of the distributions. The evidence (large dissimilarity) at a genuine speaker change deteriorates, and at the same time several spurious speaker changes may be hypothesized. A casual conversation typically has a large number of short speaker turns and hence the existing approaches are not suitable for this kind of data.

The limitations of the existing approaches can be readily seen from Figs. 2.2 and 2.3. A short segment of a broadcast news clip with only one speaker change is shown in Fig. 2.2(a). The Figs. 2.2(b) to (e) show the ΔBIC plots for window sizes of 3 sec, 1 sec, 0.5 sec and 0.1 sec, respectively. It is evident from the plots that the evidence at the speaker boundary deteriorates and several spurious speaker changes may be hypothesized as the window size is reduced. The Fig. 2.3 shows the evidence provided by the ΔBIC plots for a 2-speaker conversation containing short speaker turns. As there are speaker turns of the order of less than a second, the ΔBIC plots are computed for window sizes 0.1 sec, 0.2 sec, 0.5 sec and 1 sec (Figs. 2.3 (b) to (e), respectively). It can be seen from the plots that the existing approach fails completely for short speaker turns, due to limited or corrupted (speech from both the speakers) data.

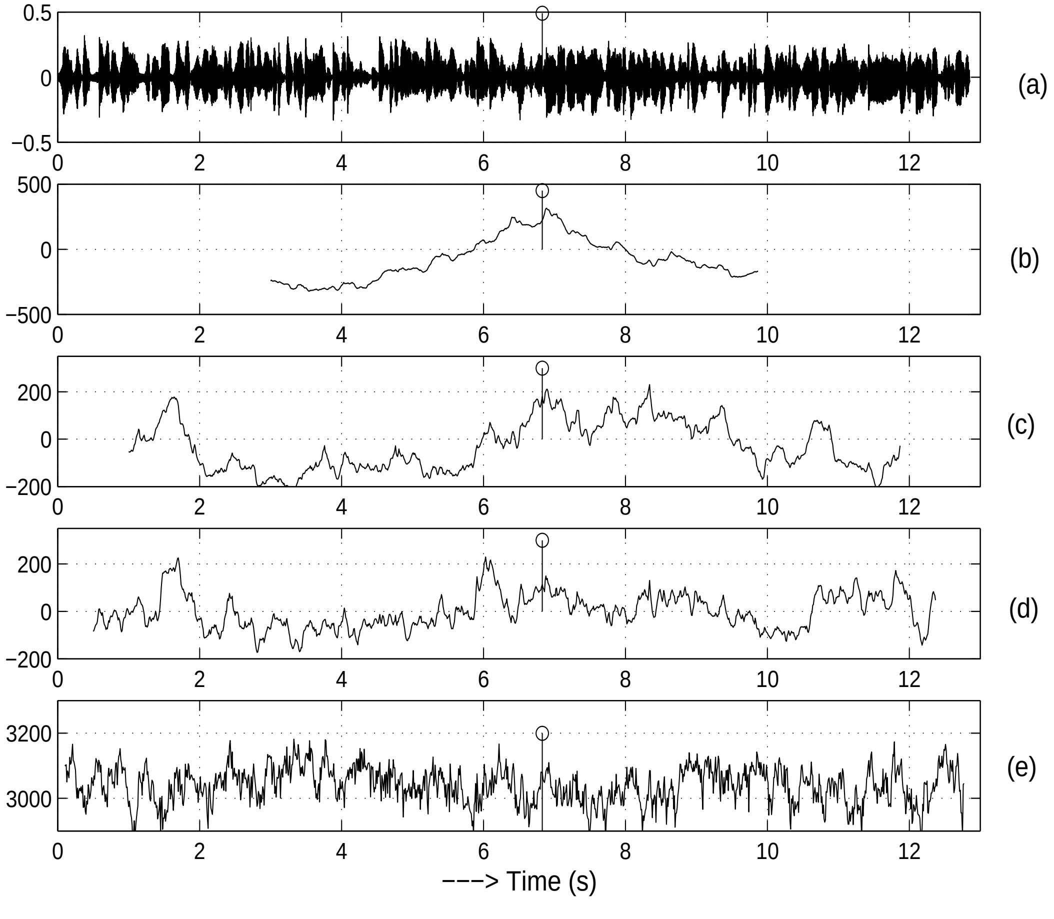

Fig. 2.2: (a) Waveform of a 2-speaker speech signal with long (> 5 sec) speaker turns. The ΔBIC plots for windows of size (b) 3 sec, (c) 1 sec, (d) 0.5 sec and (e) 0.1 sec. True speaker change is marked in all the subplots.

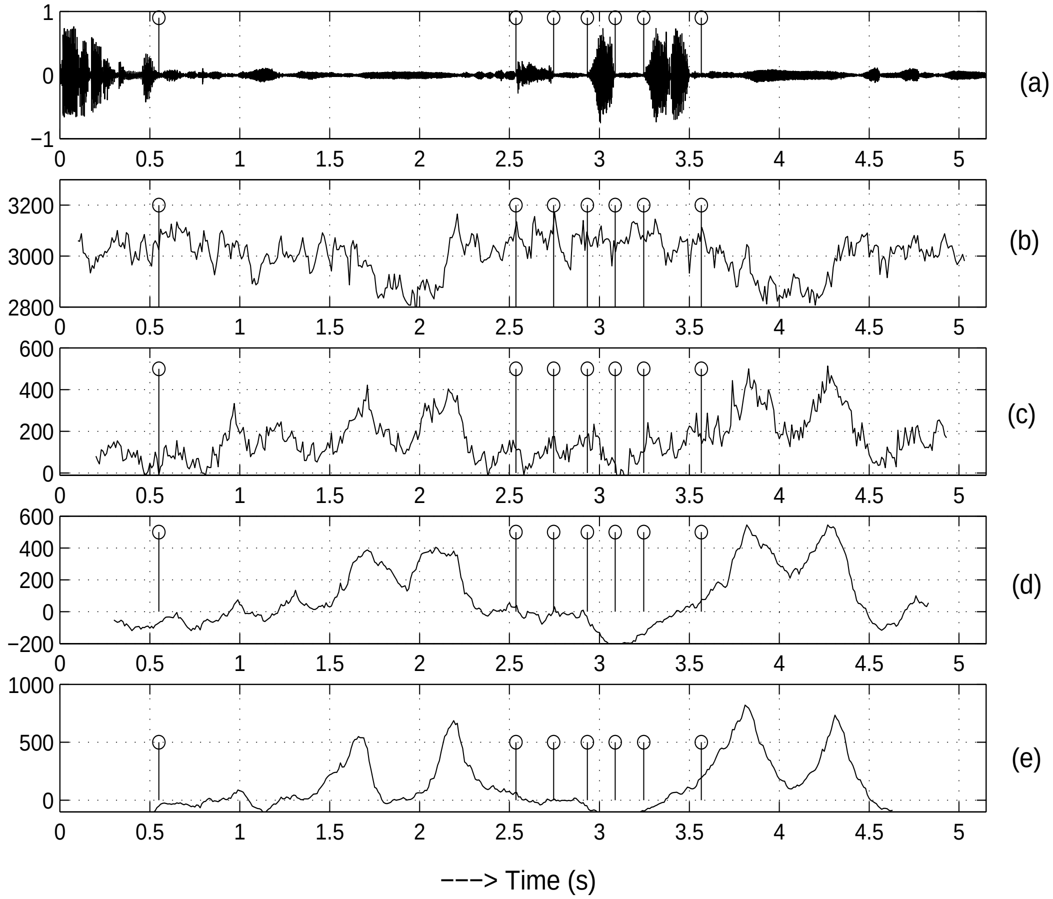

Fig. 2.3: (a) Waveform of a 2-speaker speech signal with short (< 5 sec) speaker turns. The ΔBIC plots for windows of size (b) 0.1 sec, (c) 0.2 sec, (d) 0.5 sec and (e) 1 sec. True speaker changes are marked in all the subplots.

Summarization of the drawbacks :

The limitations of the existing approaches can be summarized as follows:

- Dependence on vocal tract features alone
- Statistical approach to a point phenomenon
    - Dissimilarity measures rely on probability distributions
    - Gaussian distributions are assumed
    - Require large speaker turns for better estimation of the distributions,

  - Fail for conversational speech which have a large number of short speaker turns
  - Setting a threshold on the dissimilarity measure for validating a speaker change is a difficult problem

- Applications under consideration downplay the importance of short speaker turns

2.4.2 Significance of short speaker turns

The distribution of the duration of speaker turns varies significantly with the type of multispeaker speech under consideration. Broadcast news data typically has speaker turns of long durations (> 5 sec) and hence short speaker turns may not be of much significance. Also, one can afford a greater flexibility in the accuracy of the detection of the speaker change locations. However, the importance of short speaker turns cannot be downplayed in applications like forensic investigations, which handle casual conversations.

In order to study the frequency distribution of the duration of speaker turns, ten different 2-speaker conversations, each 5 minute of duration, are considered. A total of 880 speaker changes (manually marked) are present in approximately 3000 seconds of the conversational speech. The frequency distribution of the duration of speaker turns and its cumulative distribution are given in Figs. 2.4(a) and (b), respectively. The speaker turn durations are computed by considering only voiced speech (after the removal of silence and unvoiced regions) and speaker turns greater than 10 seconds are clamped to 10 seconds. It can be seen that around 60% of the speaker turns are of one second or lesser duration. The maximum amount of data available for detecting a speaker change is dictated by shorter of the two speaker turns on either side of the

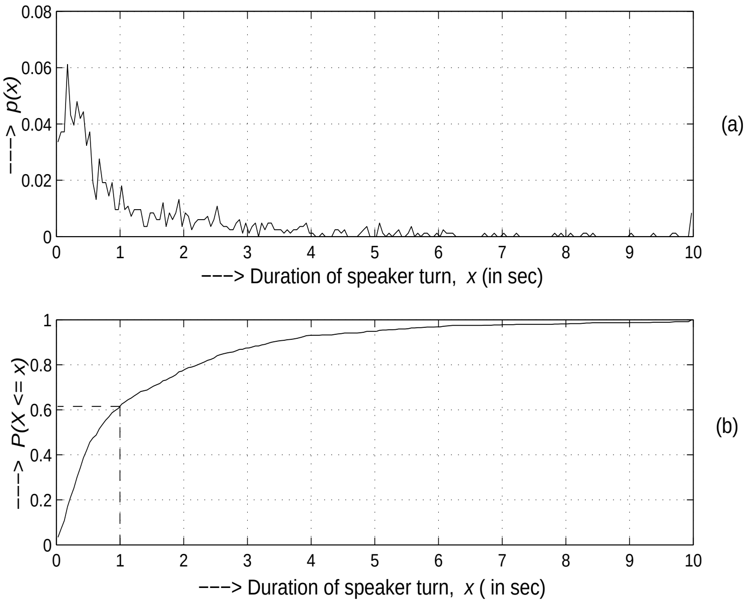

Fig. 2.4: (a) Frequency distribution of speaker turn duration. (b) Cumulative frequency distribution of the speaker turn duration.

change. The frequency distribution of the duration of the shorter speaker turn and its cumulative distribution are shown in Figs. 2.5(a) and (b), respectively. It is seen that around 90% of the speaker changes are bounded, on atleast one side, by speaker turns of one second or lesser duration. This shows the abundance of short speaker turns and the need for techniques to detect speaker changes with limited amount of data.

The above analysis considers only the frequency of speaker turns of various duration and the percentage coverage with respect to the total number of speaker changes. The curve in Fig. 2.6 gives the percentage of conversation time (normalized between 0 and 1) covered by speaker turns less than or equal to a given duration. While more than 50% of the conversation time is covered by speaker turns of 3 seconds or lesser duration, around 20% of the total conversation time is due to turns of one second or lesser duration.

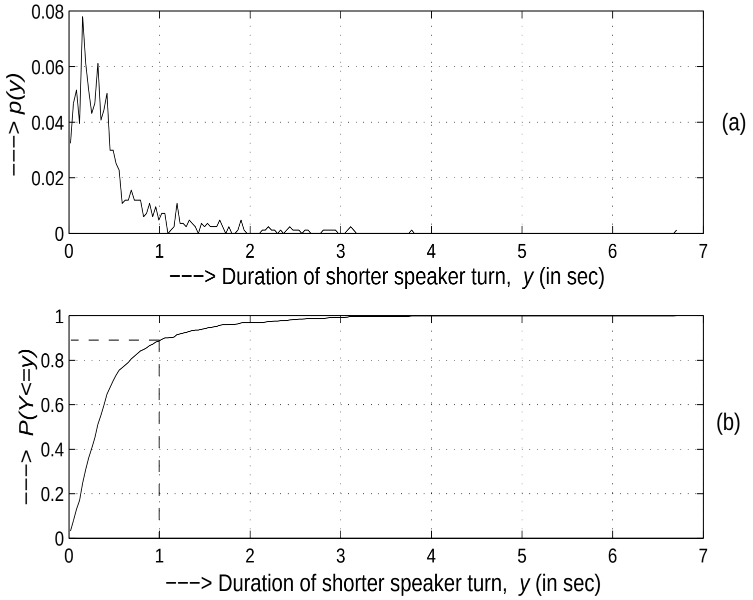

Fig. 2.5: (a) Frequency distribution of the duration of shorter turn around a speaker change. (b) Cumulative distribution of (a).

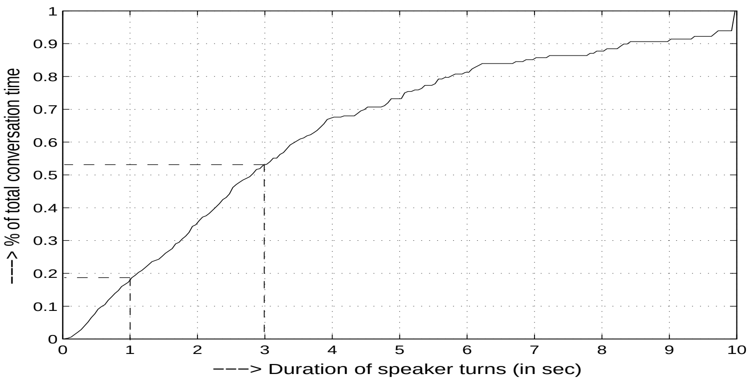

Fig. 2.6: Percentage of the conversation time (cumulative) covered by speaker turns of varying durations.

2.5 Summary

In this chapter, the general principle of the commonly used approach for speaker segmentation was outlined. A brief review of some of the work done on speaker segmentation and the use of ΔBIC for speaker change detection was also discussed. The limitation of the BIC approach which uses vocal tract features, in detecting speaker changes in a multispeaker data with short speaker turns, was brought out. In the next chapter, the possibility of using the excitation source features for change detection will be discussed.

CHAPTER 3

SIGNIFICANCE OF EXCITATION SOURCE FEATURES IN SPEAKER CHARACTERIZATION

Most of the problems related to speech are *pattern recognition* problems, which have a *representation problem* followed by a *comparison problem.* A good representation can trivialize the comparison task. Speech is a composite signal carrying in it information regarding the message, the language and the speaker. Depending on the application of interest, a set of parameters known as *features*, is extracted to characterize the required information. As the information about the speech, speaker and language are tightly integrated into the speech signal, separating them is a challenging task. The speaker characteristics present in the signal can be attributed to the anatomical and the behavioral aspects of the speech production mechanism. The representation and extraction of the behavioral characteristics is a difficult task, and usually requires large amount of data. The speech production mechanism in human beings can be modeled by a two-stage system [19–21]. A time-varying excitation system energizes a time-varying vocal tract system to generate a quasi-periodic and quasi-stationary speech signal. As outlined in Chapter 2, the existing approaches bank only on the vocal tract features for speaker change detection. In this chapter, the possibility of using the characteristics of the excitation source system for speaker segmentation, is explored. A brief discussion on the various excitation source features is given in

Section 3.1. The significance of excitation source in characterizing a speaker's voice from perception point of view is discussed in Section 3.2. The use of neural network models to capture the excitation characteristics at a subsegmental level (less than a pitch cycle) is illustrated in Section 3.3. A new metric to evaluate the effectiveness of a feature set in characterizing a speaker from limited amount of data, is introduced in Section 3.4. Also, the effectiveness of the excitation source features as compared to that of the vocal tract features, in characterizing a speaker from limited data, is studied in this section.

3.1 Excitation Source Features

The vocal tract system can be modeled as a time-varying all-pole filter using segmental analysis [21]. The segmental analysis corresponds to the processing of speech as short (10-30 ms) overlapped (5-15 ms) windows. The vocal tract system is assumed to be stationary within the window and is modeled as an all-pole filter of order p using linear prediction (LP) analysis [21,53]. The LP analysis works on the principle that a sample value in a correlated, stationary sequence can be predicted as a linear weighted sum of the past few (p) samples. If $s(n)$ denotes a sequence of speech samples, then the predicted value at the time instant n is given by,

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \ s(n-k) \tag{3.1}$$

where $\{a_k\}, \ k = 1, 2, ..., p$ is the set of linear predictor coefficients (LPC) and p is the order of the LP filter. The error at time n and the sum of squared errors E are given by,

$$r(n) \ = \ s(n) \ - \ \hat{s}(n) \tag{3.2}$$

$$E \;=\; \sum_{n} r^2(n) \tag{3.3}$$

The cost function E is minimized with respect to $\{a_i\},\ i = 1, 2, ..., p$ over the interval $-\infty \leqslant n \leqslant \infty$ (autocorrelation formulation) as,

$$\partial E/\partial a_i \;=\; 0 \qquad\qquad 1 \leqslant i \leqslant p \tag{3.4}$$

This minimization leads to a set of normal equations,

$$\sum_{k=1}^{p} a_k\ R(i-k) = -R(i) \qquad\qquad 1 \leqslant i \leqslant p \tag{3.5}$$

where

$$R(i) = \sum_{n=-\infty}^{\infty} s(n)\ s(n+i) \qquad\qquad -\infty \leqslant i \leqslant \infty \tag{3.6}$$

is the autocorrelation signal. The solution of these normal equations gives the values of the predictor coefficients $\{a_k\},\ k = 1, 2, ..., p.$ The error signal $r(n)$ obtained by inverse filtering the speech signal is referred to as the LP *residual.* The smooth variations (highly correlated) in the speech signal are captured by the LPCs and are attributed to the vocal tract characteristics. The complex poles of the LP filter occur as conjugate pairs, and each pair represents a resonator cavity, with a maximum response at a frequency (called as *resonant frequency*) where the poles are located on the z-plane [54–56]. The vocal tract can be considered as a cascade of resonator cavities with different shapes and sizes [21]. The resonant frequencies of these cavities are referred to as *formants.* The LP residual signal has large error values at regular intervals and can be attributed to the periodic impulses of excitation. Hence the LP residual is a good approximation to the excitation source signal and can be used further to extract the excitation source characteristics. A segment of voiced speech (windowed), frequency response of the inverse filter and the corresponding LP residual are shown in Fig. 3.1.

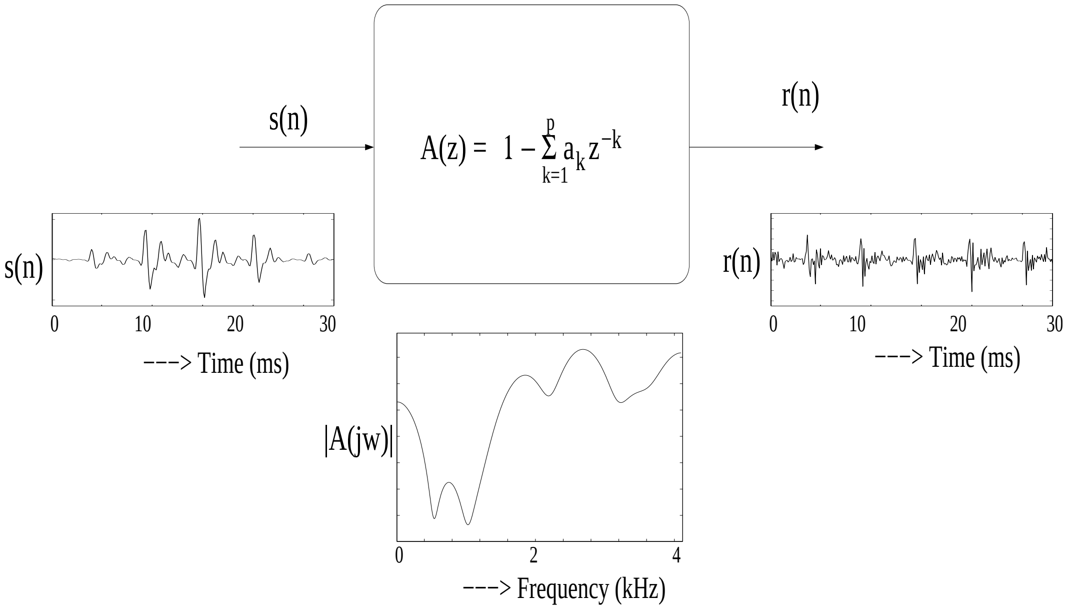

Fig. 3.1: Inverse filtering of speech to obtain the LP residual.

The excitation source features can be classified into three broad categories based on the size of the window used to analyze the speech signal or the excitation source signal. They are:

- Subsegmental features
- Segmental features
- Suprasegmental features

The *subsegmental features* are extracted over a very short (1-5 ms) analysis window, typically less than a pitch period. Parameters modeling the shape of the *glottal flow derivative* waveform (derivative of the *glottal volume velocity* waveform) can be used to characterize a speaker's voice [57–61]. An approximation to the glottal flow derivative can be obtained using the LP residual signal. However, accurate estimation of these parameters is a tough task and are highly susceptible to noise. Alternatively, an *autoassociative neural network* (AANN) model can be used to capture the excitation characteristics of a speaker present in the LP residual signal [25, 26, 62].

The *segmental features* are extracted over a short (10-30 ms) analysis window, typically comprising a few pitch cycles. The term segmental features arises due to the popular segmental analysis of speech over a short $(10-30$ ms) interval of time during which the signal is assumed to be stationary [21]. *Pitch* or *fundamental frequency (f_0)* and *periodic-aperiodic* (PAP) *ratio* are two important features extracted at segmental level. Pitch refers to the periodicity in the signal and is often expressed or measured in terms of the fundamental frequency f_0 [20]. The periodic and aperiodic components of the excitation signal can be separated out using PAP decomposition [63, 64]. The ratio of the energies of the periodic and aperiodic components is termed as the PAP ratio.

Suprasegmental features mainly refer to the behavioral aspects (speaking habits) of a speaker and are typically extracted over a large (>100 ms) analysis window. The temporal variation of any of the segmental or subsegmental features can be considered as a suprasegmental feature. *Intonation* (pitch contour), syllable durations and speaking rate are some of the important suprasegmental features. *Jitter*, the maximum perturbation in durations (pitch) of successive signal periods, and *shimmer*, the maximum perturbation in the peak values of successive signal periods, are also used to characterize a voice [63, 64]. Efficient representation and accurate estimation are the major issues in using suprasegmental features for speaker characterization.

The excitation source characteristics of a speech signal can be summarized as follows:

- Subsegmental features
    - Parameters characterizing the glottal flow derivative waveform
    - Excitation characteristics present in LP residual, as captured by an AANN

model

- Segmental features
  - Pitch or fundamental frequency (f_0)
  - Periodic-aperiodic (PAP) ratio
- Suprasegmental features
  - Intonation
  - Syllable duration
  - Speaking rate
  - Jitter
  - Shimmer

The focus in this thesis work is on the subsegmental excitation features. The ability of an AANN model to capture the subsegmental excitation features in the LP residual signal will be discussed in Section 3.3.

3.2 Perceptual Significance of the Excitation Source

The objective of this study is to understand the role played by the excitation source signal in characterizing a speaker's voice from a listener's point of view. Though the vocal tract system and the excitation system of the human speech production mechanism are tightly coupled, a reasonable approximation of the vocal tract can be obtained by LP analysis of the speech signal. A 12^{th} order LP analysis (a window size of 20 ms and a shift of 10 ms) is used to separate the vocal tract information (LPCs) from the excitation source information (LP residual) [21]. The speech signal can be resynthesized by exciting the sequence of LP filters using blocks of residual. In many applications this residual error signal is discarded, after capturing the pitch (f_0) information. Now if speech is synthesized using a random excitation signal, but with the same LPCs, the voicing characteristics of the speaker are totally lost. The synthesized speech sounds like a whispered speech and it is difficult to identify the speaker by listening, though the message may be interpreted. This signifies the importance of the excitation source information in characterizing a speaker's voice. If speech is synthesized using a train of impulses with a periodicity of $1/f_0$, the speech sounds artificial and many of the speaker-specific characteristics are lost. This signifies the importance of excitation characteristics other than the rate of vibration of the vocal folds (pitch). The high signal to noise ratio (SNR) regions in the LP residual, at the *glottal closure* (GC) events, can be approximated by the peaks in the magnitude of the analytic signal (Hilbert transform of the residual signal). Listening to speech synthesized separately by emphasizing and deemphasizing the GC regions in the LP residual show that the GC regions contribute significantly towards retaining the speaker characteristics. The various excitation signals and the corresponding resynthesized speech signals are shown in Fig. 3.2.

3.3 Speaker Characterization using Excitation Features

The subsegmental features embedded in the LP residual signal can be captured using an autoassociative neural network model [25, 26, 62]. A five layer AANN model with nonlinear hidden layers and a linear output layer is used, as shown in Fig. 3.3. The structure of the AANN model is denoted as P_1 L P_2 N P_3 N P_4 N P_5 L, where P_1 to P_5 are the number of nodes in each of the five layers, $P_1 = P_5$, L (linear) and N

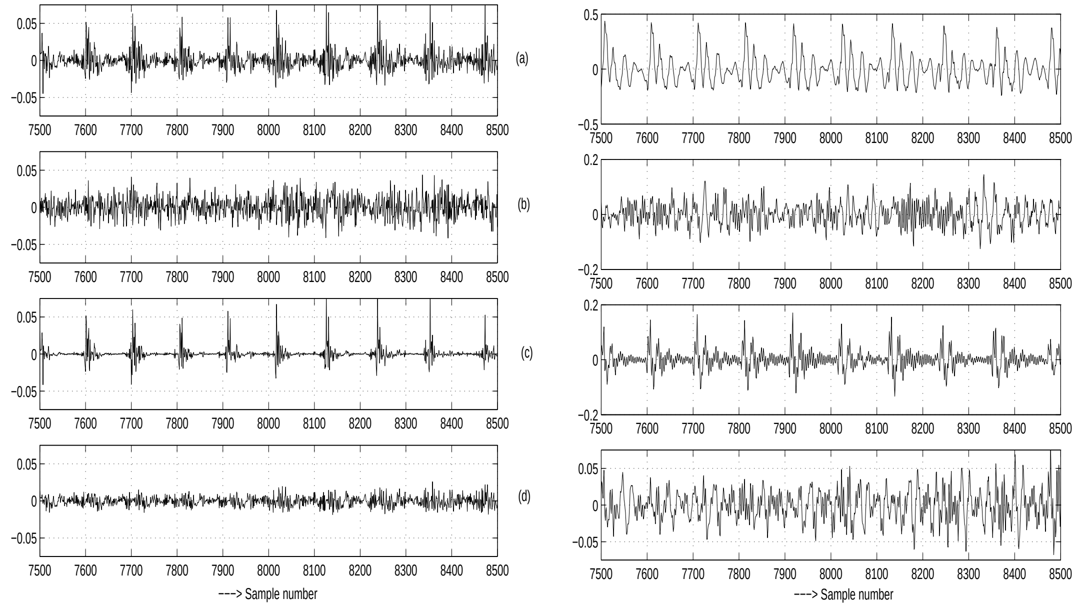

Fig. 3.2: Different types of excitation signals and the corresponding synthesized speech signals. (a) LP residual, (b) random noise excitation, (c) residual with non-glottal closure regions suppressed, (d) residual with glottal closure regions suppressed, (e) to (h) speech signals synthesized using excitation signals shown in (a) to (d) respectively.

(nonlinear) represent the type of activation function used by the nodes in that layer. Let $r(n)$ be the given LP residual signal of length N_r, after removal of silence and unvoiced portions. A rectangular window of size d (samples) is slided over the signal $r(n)$ with a shift of one sample, to obtain a sequence of frames $\{\boldsymbol{x}(n)\},\ n = 1, \ldots, N_r$ as shown in Fig. 3.4.

Typically, the length of the frames d ($d = P_1 = P_5$), corresponds to a duration of 2 to 5 ms (less than one pitch period). The magnitude normalized frames are then presented to the neural network one after the other, in the same sequence. If $\boldsymbol{x}(n)$ and $\boldsymbol{y}(n)$ represent the input and output of the AANN model, the error in mapping at the time instant n is computed as $\boldsymbol{u}(n) = \boldsymbol{x}(n) - \boldsymbol{y}(n)$. The weights of the network, initially set to small (≈ 0) random values, are adjusted using *generalized delta learning rule* and *error backpropagation* [41,42], on a pattern-by-pattern basis. The network is now said

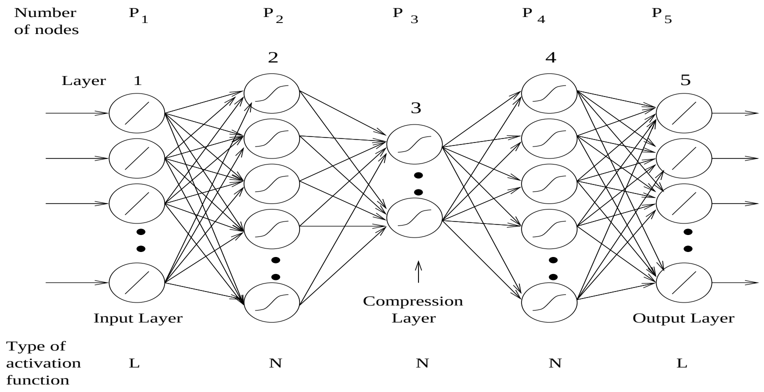

Fig. 3.3: A five layer autoassociative neural network of structure P_1 L P_2 N P_3 N P_4 N P_5 L, where L (linear) and N (nonlinear) denote the type of activation functions.

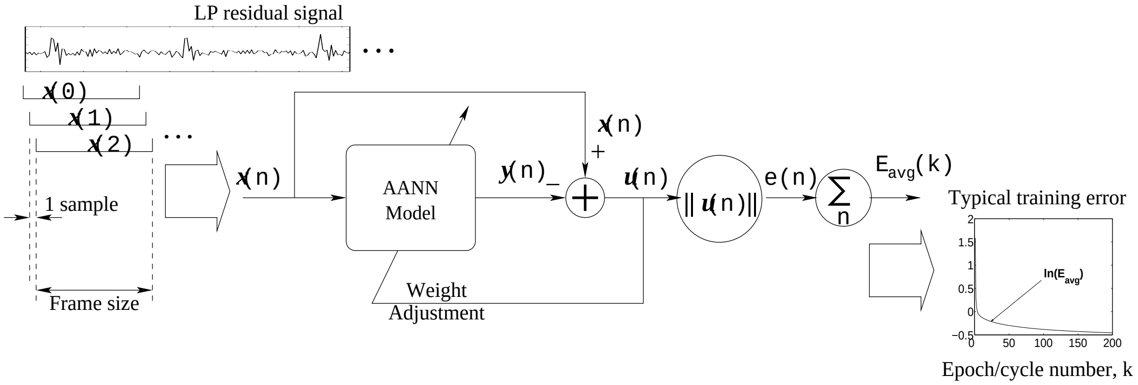

Fig. 3.4: Training an AANN model to capture the subsegmental features in the LP residual.

to have learned over one cycle or *epoch*. The error at the end of an epoch is computed as $E_{avg}(k) = \frac{1}{N_r}\sum_{n=1}^{N_r} e^2(n)$, where k is the epoch number, $e(n) = \frac{1}{d}\sum_{m=1}^{d} u_m^2(n)$ is the mean squared error at time instant n, and $u_m(n)$ is the m^{th} component of the error frame $\boldsymbol{u}(n)$. The learning process is repeated for a predetermined number of epochs or until the accumulated error (or a change in it as compared to the previous epoch) is significantly low, depending upon the complexity of the task. The goal of the learning process is to minimize the cost function E_{avg}. The *training error* E_{avg} accumulated over each epoch typically shows a decreasing trend, as shown in Fig. 3.4, signifying the learning ability of the network. The AANN model now characterizes or represents the subsegmental excitation features present in the LP residual signal.

Once an AANN model has captured the subsegmental features in the LP residual, it can be used to find the similarity or dissimilarity of characteristics of any given LP residual signal with respect to that used while training. The given LP residual signal is tested against the trained AANN model as shown in Fig. 3.5, in a way similar to that during training. The mean squared error $e(n)$, obtained for each frame represents the amount of dissimilarity between the current frame and the set of frames used for training the network. Larger the error, farther are the frame characteristics from those used while training. This frame error is converted into a normalized (0 to 1) similarity measure known as the *confidence score*, given by $c(n) = exp(-e(n))$. The average frame confidence $C_{avg} = \frac{1}{N_t}\sum_{n=1}^{N_t} c(n)$, gives the similarity between the training sequence and the test sequence, where N_t is the number of frames in the test signal. Fig. 3.6 shows the typical confidence scores for a small portion of the residual signal. It can be seen that the confidence scores are high around the glottal closure regions

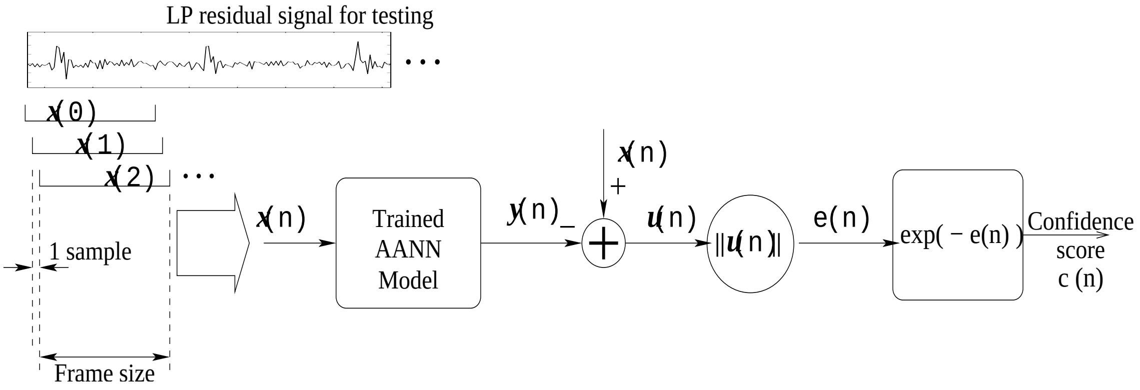

Fig. 3.5: Comparing the characteristics of a given LP residual signal against that captured by a trained AANN model.

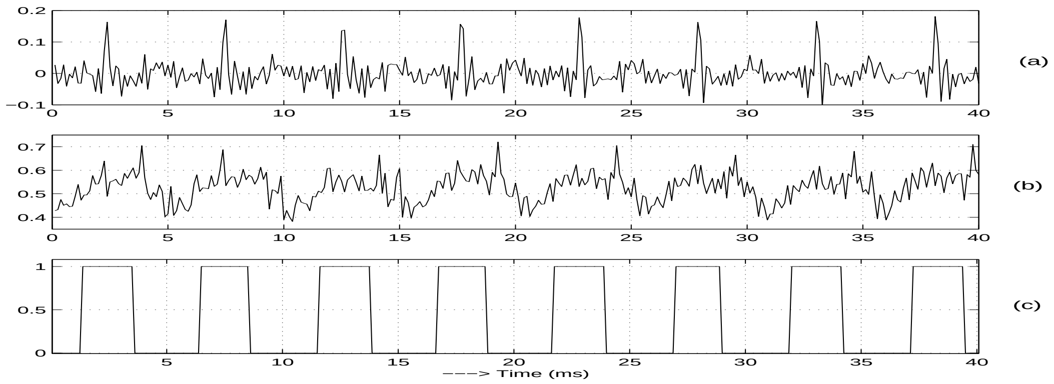

Fig. 3.6: Frame selection for training AANN models. (a) LP residual of a voiced speech segment. (b) Confidence scores of an AANN model for the input in (a). (c) Weight function derived from the energy of the LP residual, for frame selection.

due to high SNR of the residual signal around GCs. Owing to this, the training and testing processes are modified to use frames only around the GC instants.

3.4 Within-speaker to Across-speaker Dissimilarity of Sounds

Speaker segmentation of conversational speech requires that the speaker characteristics be captured with minimum amount of speech data. The choice of feature vector and the comparison strategy play a significant role on the performance of the segmentation task. It is desired that the features characterizing a speaker are invariant to the type of sound unit. Given an alphabet of N sounds $V = \{v_i\},\ i = 1, \ldots, N$, and a set of M speakers $S = \{s_j\},\ j = 1, \ldots, M$, the within-speaker to across-speaker dissimilarity (WAD) ratio for a feature set can be defined at different levels.

The WAD ratio for the sound v_i of speaker s_j is defined as

$$\alpha(v_i, s_j) = \frac{\frac{1}{N-1}\sum\limits_{\substack{k=1\\k\neq i}}^{N} d((v_i, s_j), (v_k, s_j))}{\frac{1}{M-1}\sum\limits_{\substack{k=1\\k\neq j}}^{M} d((v_i, s_j), (v_i, s_k))} \tag{3.7}$$

where $d((v_i, s_j), (v_k, s_j))$ can be any dissimilarity measure between the (sound, speaker) pairs (v_i, s_j) and (v_k, s_j). The WAD ratio for a speaker s_j is defined as

$$\beta(s_j) = \frac{\frac{1}{N}\frac{1}{N-1}\sum\limits_{i=1}^{N}\sum\limits_{\substack{k=1\\k\neq i}}^{N} d((v_i, s_j), (v_k, s_j))}{\frac{1}{N}\frac{1}{M-1}\sum\limits_{i=1}^{N}\sum\limits_{\substack{k=1\\k\neq j}}^{M} d((v_i, s_j), (v_i, s_k))} \tag{3.8}$$

The WAD ratio for a given feature set is defined as

$$\gamma = \frac{\frac{1}{M}\frac{1}{N}\frac{1}{N-1}\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\sum\limits_{\substack{k=1\\k\neq i}}^{N} d((v_i, s_j), (v_k, s_j))}{\frac{1}{M}\frac{1}{N}\frac{1}{M-1}\sum\limits_{j=1}^{M}\sum\limits_{i=1}^{N}\sum\limits_{\substack{k=1\\k\neq j}}^{M} d((v_i, s_j), (v_i, s_k))} \tag{3.9}$$

A value of $\alpha(v_i, s_j)$ less than one signifies that the sound v_i is a good candidate for characterizing a speaker using the given feature set. A value of $\beta(s_j)$ less than one signifies that the given feature set is good at discriminating the speaker s_j from other speakers. A value of γ less than one signifies that the given feature-set is good at discriminating speakers from one another. Given two feature sets, the feature set which gives a lesser value of γ is better suited for characterizing a speaker from limited amount of data. An ideal value of $\gamma = 0$ makes the feature an ideal one for speaker characterization.

In order to study the suitability of the vocal tract features (LPCC) and the subsegmental excitation characteristics for speaker characterization from limited data, the WAD ratios for both the features are computed separately. The dataset contains, isolated utterances of N=5 voiced sounds (vowels /a/,/i/,/u/,/e/ and /o/), collected from $M = 5$ different speakers. LPCCs (19 dimensional) obtained from a 12^{th} order LP analysis were used as the vocal tract features, and ΔBIC was used as the dissimilarity measure. The excitation source features in the LP residual signal is captured using a 5-layer AANN model ($40L$ $60N$ $12N$ $60N$ $40L$) as described in Section 3.3. A model is generated for each sound of every speaker. A confidence score $c(v_i, v_j)$ obtained by testing a sound v_i against the model of another sound v_j gives the similarity between the two sounds. The dissimilarity between the two sounds is computed as $d(v_i, v_j) = 1 - c(v_i, v_j)$. The WAD ratios $\alpha(v_i, s_j)$ of different (sound, speaker) pairs for the LPCC features and the excitation source features are given in Tables 3.1 and 3.2 respectively. It can be seen from the tables that some speakers are better characterized by their vocal tract features (speaker s_1), while some others (speakers s_4 and s_5) by their excitation source characteristics. The LPCC features give an overall WAD ratio of $\gamma = 1.3713$, while it is $\gamma = 0.9656$ for the excitation source features. This shows that the excitation source features have lesser dissimilarity within a speaker as compared to that across speakers. Hence excitation source features are better suited for segmenting multispeaker speech with short speaker turns.

Table 3.1: The WAD ratios $\alpha(v_i, s_j)$ of sounds for different speakers, for vocal tract features (LPCCs).

	Sound				
	v_1	v_2	v_3	v_4	v_5
Speaker	$/a/$	$/i/$	$/u/$	$/e/$	$/o/$
s_1	0.47	0.51	0.76	0.61	0.42
s_2	1.24	1.28	3.94	1.06	1.23
s_3	0.91	1.31	2.28	0.83	0.83
s_4	1.37	1.67	2.71	1.44	1.41
s_5	1.32	1.29	2.70	1.21	1.48

Table 3.2: The WAD ratios $\alpha(v_i, s_j)$ of sounds for different speakers, for excitation source features.

	Sound				
	v_1	v_2	v_3	v_4	v_5
Speaker	$/a/$	$/i/$	$/u/$	$/e/$	$/o/$
s_1	1.72	1.21	0.61	1.83	1.88
s_2	1.09	0.95	0.44	1.57	1.06
s_3	1.26	1.27	0.43	1.92	1.07
s_4	0.66	0.69	0.11	0.71	0.34
s_5	0.82	0.65	0.38	0.87	0.59

3.5 Summary

Some of the important excitation source features at subsegmental, segmental and suprasegmental levels were listed in this chapter. The significance of excitation source features in characterizing a speaker's voice, and the ability of AANN models to capture the subsegmental features in the excitation signal were discussed. It was noted that the WAD ratio for the subsegmental excitation features is lesser than that for the vocal tract features, for small segments (syllables) of speech. This shows that the excitation features have lesser dissimilarity within a speaker, and hence are better suited for segmenting multispeaker speech with short speaker turns. A new approach for speaker segmentation using the subsegmental excitation features will be proposed in the next chapter.

CHAPTER 4

SPEAKER SEGMENTATION USING EXCITATION SOURCE FEATURES

In the previous chapter, it was shown that excitation source information can play a significant role in characterizing a speaker from limited amount of data. The ability of autoassociative neural network models to capture the excitation characteristics of a speaker from the LP residual signal was also discussed. As suggested by the WAD ratio, the excitation source features are better suited for detecting speaker changes in a casual conversation, which is likely to contain short speaker turns. In this chapter, a new approach for speaker segmentation is proposed using neural network models and the excitation source information. The principle on which the proposed approach works is discussed is Section 4.1. The approach for speaker segmentation using subsegmental excitation features is presented in Section 4.2. A comparison of performance of the proposed approach and the metric-based (BIC) approach using vocal tract features, is given in Section 4.3.

4.1 Basic Principle of the Proposed Approach

The proposed approach for speaker segmentation relies on the principle that the excitation source features have lesser dissimilarity within a speaker, as compared to the dissimilarity across speakers. This means that smaller amount of data is sufficient to model a speaker using the excitation features, as compared to the vocal tract features.

Given a multispeaker speech signal, if one of the speakers can be modeled, then segments of speech corresponding to the modeled speaker can be separated from segments of the remaining speakers. The process can be repeated until only one speaker is left out. Though this principle applies to a multispeaker conversation with any number of speakers, the focus of this thesis is on 2-speaker conversations. As there is no *a priori* information about the identities of the speakers involved in the conversation, the automation of the above process becomes essential. Two major issues to be addressed here are:

- Sufficiency of data for speaker modeling
- Automatic detection of single speaker regions for speaker modeling

4.1.1 Sufficiency of data for speaker modeling

The amount of speech data required to characterize a speaker plays an important role in devising a method for speaker segmentation. The evidence for speaker change detection will be reliable, if the modeling of the speaker is good. Speaker verification experiments using excitation source information show that around 5 seconds of voiced data is optimal for training [26]. The performance almost remains constant above 5 seconds, while it degrades gradually for lesser amount of training data. In the case of speaker segmentation, one needs to discriminate only between the two conversing speakers, unlike the speaker verification task where a speaker has to be discriminated against a number of speakers. Hence lesser amount of training data may be sufficient for speaker segmentation. In order to study the availability of evidence for speaker change detection and sufficiency of the training data, speaker models are built from varying amounts of training data (5 sec, 2 sec, 1 sec, 0.5 sec and 0.25 sec). The data for training the models are manually marked from a 2-speaker conversation. The entire conversation is tested against each of the models. Regions belonging to the modeled speaker are expected to give a higher confidence score, as compared to the regions of the other speaker. Fig. 4.1 shows the evidence obtained by models trained with different amounts of data. While the evidence for speaker change detection deteriorates with the reduction in the amount of training data, it is seen that there is considerable evidence even for 0.5 sec of training data. The actual decision on the amount of training data to be used depends on the availability of contiguous segments of speech from a single speaker and their automatic detection. It is a compromise between the strength of evidence for speaker change detection and the chances of automatically detecting such contiguous segments. About one second of training data is reasonable in terms of both

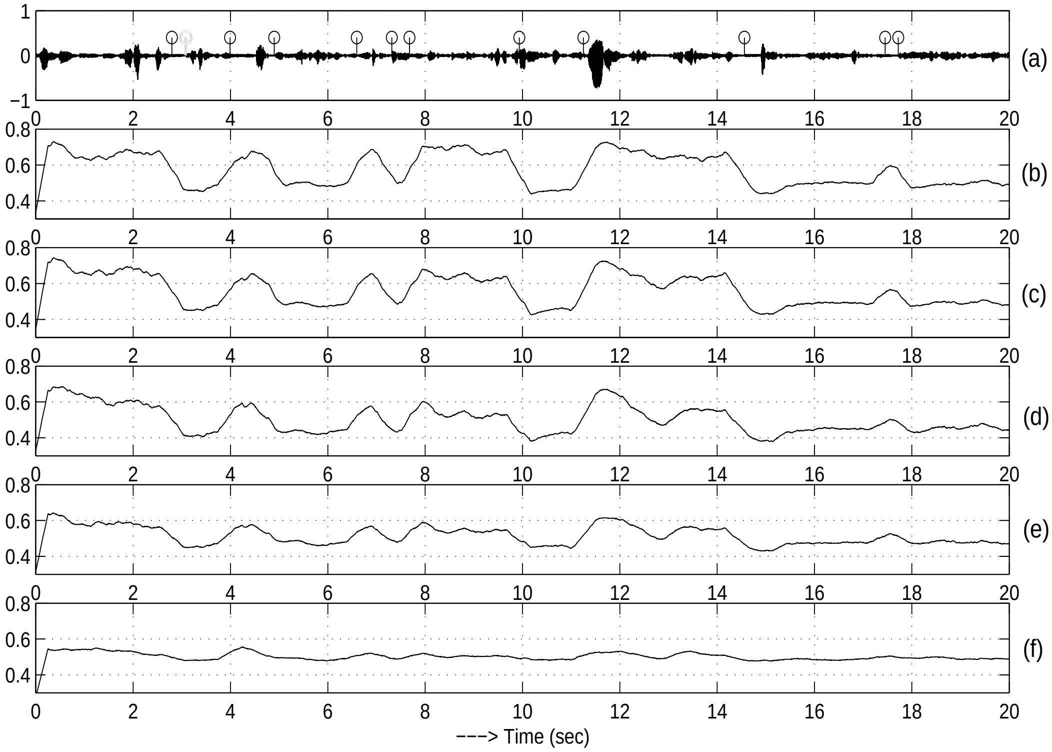

Fig. 4.1: (a) Waveform of a 2-speaker speech signal with the actual speaker changes marked by vertical poles. Evidence (confidence scores) obtained by models built from (b) 5 sec, (c) 2 sec, (d) 1 sec, (e) 0.5 sec and (f) 0.25 sec of training data.

sufficiency and availability of data. The automatic identification of regions containing a single speaker is discussed in the following section.

4.1.2 Automatic detection of single speaker regions

In a casual conversational speech it is not guaranteed that a randomly chosen segment of 1 sec data (voiced) contains only one speaker. In order to circumvent this problem, M (about 10) models are built from M adjacent segments (1 sec) of speech, with an overlap of half a second. The entire conversation is tested against each of the M models to obtain the confidence scores. The confidence plots are smoothened using a moving average window [55]. Typically a window size of 0.5 sec is used. The similarity between any two models can be measured in terms of the cross-correlation coefficient [54, 65] between the two mean-subtracted confidence plots. The cross-correlation coefficient between any two signals $\mu_1(n)$ and $\mu_2(n)$ is given by

$$\rho = \frac{\sum\limits_{n=-\infty}^{\infty} \mu_1(n)\mu_2(n)}{\left[\sum\limits_{n=-\infty}^{\infty} \mu_1^2(n) \sum\limits_{n=-\infty}^{\infty} \mu_2^2(n)\right]^{1/2}} \tag{4.1}$$

In order to automatically detect the training segments that contain speech from only one speaker, the following hypotheses should be true:

- If the value of the cross-correlation coefficient is high and positive (close to +1), then the two models belong to the same speaker. Also, there is a high probability that the training data of each model contains only one speaker.
- If the value of the cross-correlation coefficient is high and negative (close to -1), then the two training segments are pure, but belong to different speakers.
- The chances of any two impure training segments giving a high correlation value are low.

Fig. 4.2 shows the smoothened, mean-subtracted confidence plots for ten models (M=10), each model trained with one second of voiced data from a single conversation. The values of the cross-correlation coefficient computed among the ten confidence plots shown in Fig. 4.2 are given in Table 4.1. The confidence plots and the tabulated correlation coefficient values, clearly show that all the three hypotheses made earlier hold good. The models M_1, M_2, M_8 and M_9 show a high degree of similarity and belong to the same speaker. Similarly, the models M_4 and M_5 have a high degree of similarity and belong to the second speaker. It can be seen that the evidences from models of different speakers have reversed trends, and hence yield negative cross-correlation

Table 4.1: Cross-correlation coefficient values between confidence score plots of 10 models generated from adjacent, overlapped segments of a male-male conversational speech data. The highest value among the non-diagonal entries in each row is highlighted.

Models	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}
M_1	1.00	0.56	0.06	-0.18	-0.12	-0.06	0.49	0.54	**0.61**	0.38
M_2	**0.56**	1.00	0.26	-0.05	-0.00	0.05	0.41	0.38	0.45	0.45
M_3	0.06	0.26	1.00	**0.54**	0.35	0.38	0.25	-0.07	-0.01	0.28
M_4	-0.18	-0.05	0.54	1.00	**0.62**	0.46	0.09	-0.31	-0.29	0.02
M_5	-0.12	-0.00	0.35	**0.62**	1.00	0.49	0.09	-0.22	-0.22	0.03
M_6	-0.06	0.05	0.38	0.46	**0.49**	1.00	0.24	-0.19	-0.18	0.15
M_7	0.49	0.41	0.25	0.09	0.09	0.24	1.00	0.36	0.42	**0.50**
M_8	0.54	0.38	-0.07	-0.31	-0.22	-0.19	0.36	1.00	**0.65**	0.30
M_9	0.61	0.45	-0.01	-0.29	-0.22	-0.18	0.42	**0.65**	1.00	0.47
M_{10}	0.38	0.45	0.28	0.02	0.03	0.15	**0.50**	0.30	0.47	1.00

coefficient values. Models M_3, M_6, M_7 and M_{10} are trained from impure segments and hence have a relatively lower correlation values with other models. The pair of models with the highest absolute value of the cross-correlation coefficient (0.65 between M_8 and M_9 from Table 4.1) are hypothesized to be trained from pure segments (a pure segment is one that contains speech from only one speaker) of speech. The sign of the cross-correlation coefficient value suggests whether they belong to the same speaker or not.

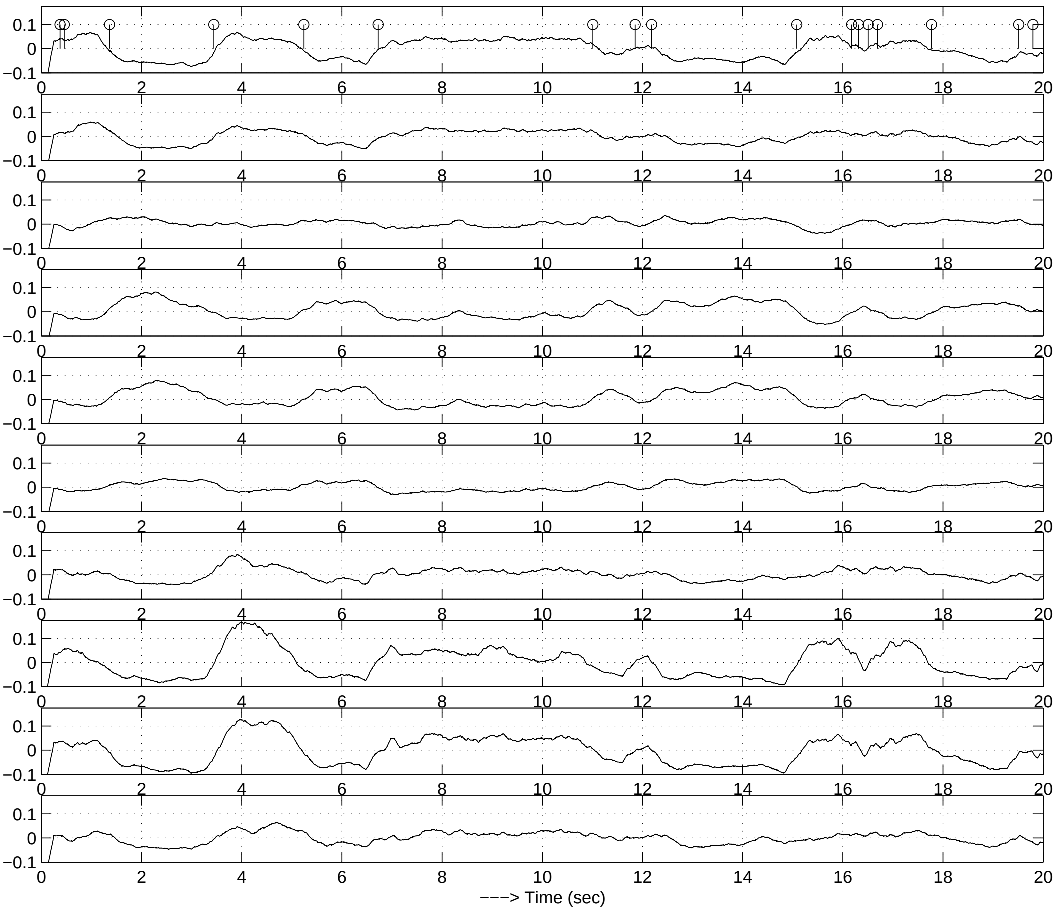

Fig. 4.2: Smoothened, mean-subtracted confidence score plots of 10 models (M_1 to M_{10} from top to bottom) generated from adjacent, overlapped segments of a male-male conversational speech data. The manually marked speaker changes are shown in the first subplot.

4.2 Proposed Approach for Speaker Segmentation

The previous section shows that the models trained with one second of single speaker data, provide significant evidence for speaker change detection. Automatic detection of training segments containing only one speaker was also discussed. A new approach for speaker segmentation using excitation source features and neural network models is proposed in this section. As the speaker separation problem is closely associated with the speaker change detection problem, a simple agglomerative clustering is proposed for speaker separation.

4.2.1 Speaker change detection

The proposed method for speaker segmentation has two phases, model generation and change detection.

4.2.1.1 Model generation phase

The issues in model generation were discussed in Section 4.1. A summary of the steps involved in the generation of models is given in this section. An AANN model is trained from approximately 1 sec of contiguous voiced speech which is hypothesized to contain only one speaker. In a casual conversational speech it is not guaranteed that a randomly chosen segment of 1 sec (voiced speech) contains only one speaker. In order to circumvent this problem, M (about 10) models are built from M adjacent speech segments of one second, with an overlap of half a second. The entire conversation is tested against each of the models to obtain M confidence score plots. The values of the cross-correlation coefficient between all possible pairs of confidence score plots are computed. The confidence scores are smoothened using a moving average window (0.5 sec) and mean subtracted before computing the cross-correlation coefficient values.

Out of M models, N (2 to 4) model that give the highest values of cross-correlation coefficient with one another are chosen. The entire process of model building and selection is depicted in Fig. 4.3.

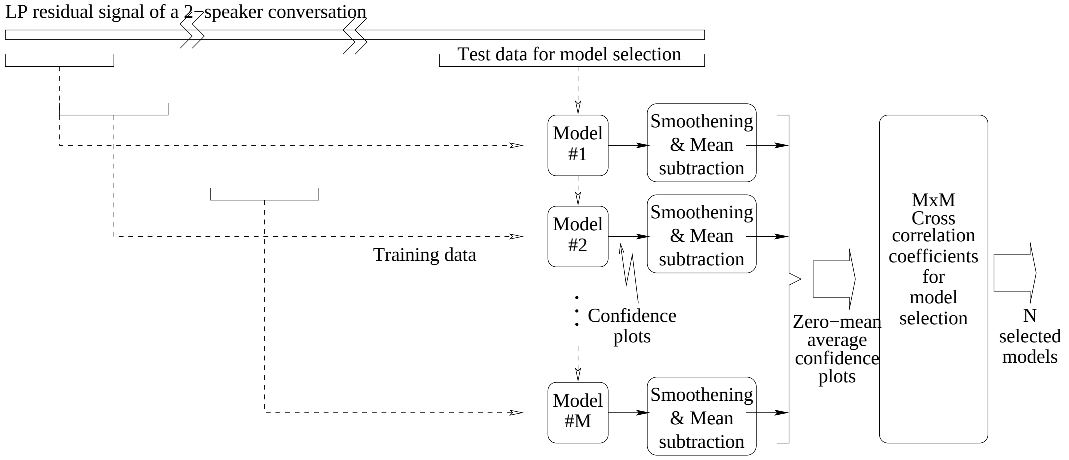
LP residual signal of a 2−speaker conversation
Test data for model selection
Training data
Model #1
Model #2
Model #M
Smoothening & Mean subtraction
Smoothening & Mean subtraction
Smoothening & Mean subtraction
Confidence plots
Zero−mean average confidence plots
MxM Cross correlation coefficients for model selection
N selected models

Fig. 4.3: Model building and selection phase in speaker change detection.

4.2.1.2 Change detection phase

The change detection phase has two steps - combining evidence from multiple models and detection of peaks in the combined confidence score plots.

Combining evidence from multiple models :

This phase involves combining evidence from the N confidence score plots chosen in the previous step. The evidence provided by each of the N models can be of varying degree. A minimum of one pair of models ($N = 2$) with single speaker data is essential for the reliability of the evidence. While there may be more evidence in some cases, it need not be true with all conversations. Also, combining evidence from multiple models in a constructive manner is a tough problem in itself. Hence evidence from only two models, with the highest degree of similarity, are used for speaker segmentation.

Let $c(n)$ denote the sequence of raw confidence scores obtained by testing the given conversation against a model. An average confidence signal

$$\mu(n) = c(n) * h_m(n) \tag{4.2}$$

is computed, where '$*$' denotes convolution operation [56] and

$$h_m(n) = \begin{cases} \frac{1}{N_A} & -\frac{N_A}{2} < n \leqslant \frac{N_A}{2}, \\ 0 & elsewhere \end{cases} \tag{4.3}$$

where $N_A = T_A \times f_s$ is the length of the analysis window (in number of samples), T_A is the duration of the analysis window (in sec) and f_s is the sampling rate of the speech signal. An absolute delta-mean sequence is computed as

$$\Delta\mu(n) = |\mu(n) * h_d(n)| \tag{4.4}$$

where

$$h_d(n) = \begin{cases} -1 & n = -\frac{N_A}{2}, \\ 1 & n = \frac{N_A}{2}, \\ 0 & elsewhere. \end{cases} \tag{4.5}$$

The $\Delta\mu$ value at any instant represents the change in average confidence values between two adjacent window segments of length N_A around that instant. Hence a peak in the $\Delta\mu$ plot can be hypothesized as a speaker change. If $\Delta\mu_1(n)$ and $\Delta\mu_2(n)$ are the confidence plots obtained from the chosen two models, they can be combined using product rule (AND logic) as

$$\Delta\mu_{AND}(n) = \sqrt{\Delta\mu_1(n)\ \Delta\mu_2(n)} \tag{4.6}$$

They can also be combined using sum rule (OR logic) as

$$\Delta\mu_{OR}(n) = \frac{1}{2}\left(\Delta\mu_1(n) + \Delta\mu_2(n)\right) \tag{4.7}$$

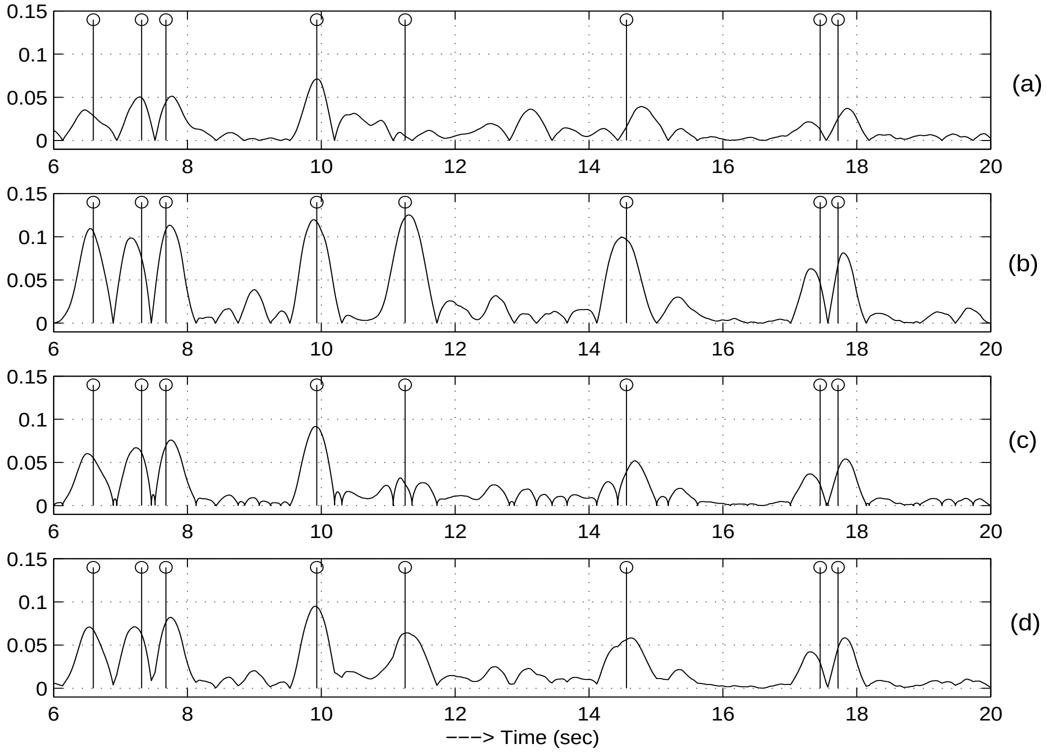

Fig. 4.4: Combining evidence at score level. (a) $\Delta\mu_1(n)$, (b) $\Delta\mu_2(n)$, (c) $\Delta\mu_{AND}(n)$, and (d) $\Delta\mu_{OR}(n)$. The vertical lines indicate actual speaker changes.

Fig. 4.4 shows the individual evidence and the combined scores using both methods. It can be seen that the evidence combined using product rule has more ripples and destroys the evidence in some cases. Instead of combining the evidence at the score level, the individual confidence sequences can also be processed independently and the decisions can be combined at a later stage using AND or OR logic. The choice of the combining logic is a compromise between a high detection rate and a low false alarm rate. The OR logic favours a high detection rate, while the AND logic favors a low false alarm rate. A majority based logic can be employed if evidence from more than two models are considered. The parameters that affect the performance of the speaker segmentation task are the duration of the analysis window T_A and the decision logic used to combine the individual evidence. The evidence available when different lengths of the analysis window are used, is shown in Fig. 4.5. It can be seen that the strength

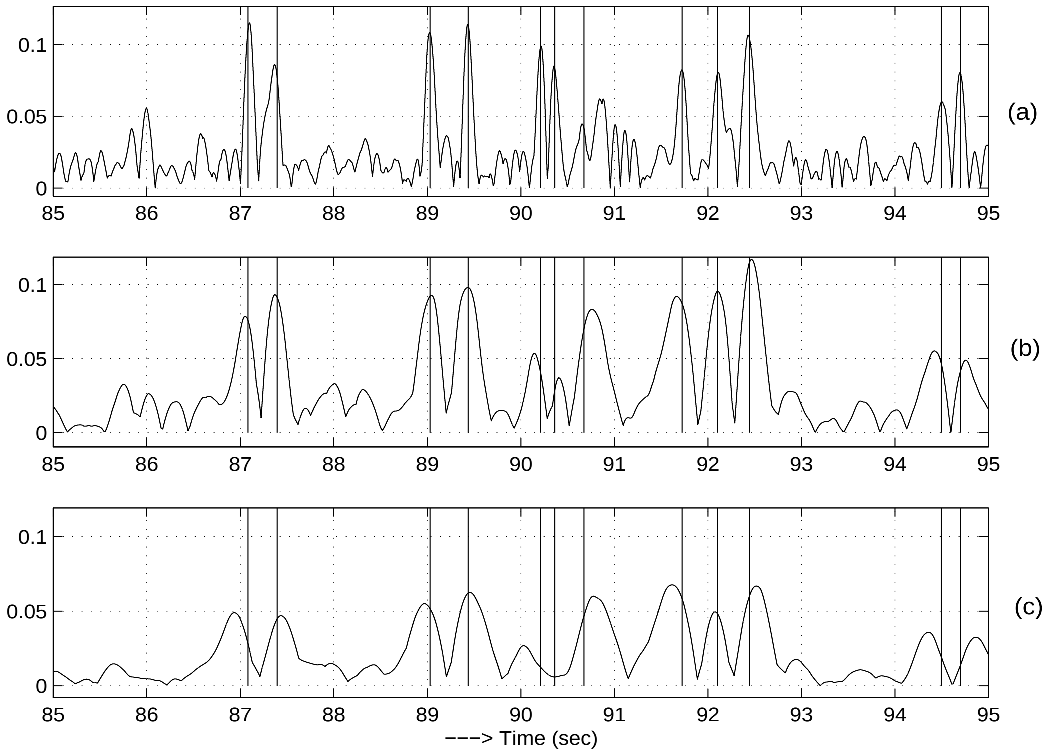

Fig. 4.5: Evidence for varying lengths of the analysis window T_A. The $\Delta\mu$ plots for window sizes of (a) 100 ms (b) 250 ms and (c) 500 ms. Manually marked speaker changes are shown as vertical lines in all the plots.

of the peaks (evidence) reduces slightly as the duration of analysis is increased, but at the same time there is a significant reduction in the number of spurious peaks. The effect of these parameters on the performance of the speaker segmentation task is discussed in Section 4.3.3.

Peak detection and validation :

A peak in a $\Delta\mu$ sequence corresponds to a significant change in the excitation characteristics of the speech signal. Hence a peak in a $\Delta\mu$ sequence is hypothesized as a speaker change. Larger the peak strength, greater is the probability of the hypothesized speaker change being a genuine one. As it is difficult to set an absolute threshold on the peak strength, all peaks in the $\Delta\mu$ sequence are detected in the first step and then validated to eliminate spurious peaks. A simple differential operator,

$$w(n) = \begin{cases} -\frac{2}{N_A} & -\frac{N_A}{2} \leqslant n < 0, \\ \frac{2}{N_A} & 0 < n \leqslant \frac{N_A}{2}, \\ 0 & elsewhere \end{cases} \tag{4.8}$$

is applied on the $\Delta\mu$ sequence to detect the peaks. The positive zero crossings of the signal $y(n) = \Delta\mu(n) * w(n)$ are hypothesized as speaker changes. In the next step, speaker change hypothesis at a peak is validated to reduce the number of false alarms. The hypothesis is true, if the peak strength is greater than $\lambda = (m - p\sigma)$, where m is the average peak strength, σ is the average deviation of the peak strengths from m over the entire conversation and p is a constant which controls the dynamic threshold λ. Fig. 4.6 shows the results of the peak detection and validation process for a dynamic threshold of $\lambda = (m - 0.5\sigma)$. The effect of varying the parameter λ on the performance of the change detection task is discussed in Section 4.3.3.

4.2.2 Speaker separation

Speaker change detection results in a sequence of segments, each of which is hypothesized to contain a single speaker. An agglomerative clustering [28, 38] is performed to separate the segments into two groups. The algorithm starts with N_{SC} clusters with one segment in each cluster. The number of clusters is reduced iteratively, by combining two segments at a time which are most similar. Absolute difference between the average confidence scores of two segments is used as the dissimilarity measure. The clustering process is continued until only two groups are left. The two groups are now hypothesized to represent each of the speakers involved in the conversation. The results of the speaker separation process are shown in Fig. 4.7. The speaker separation

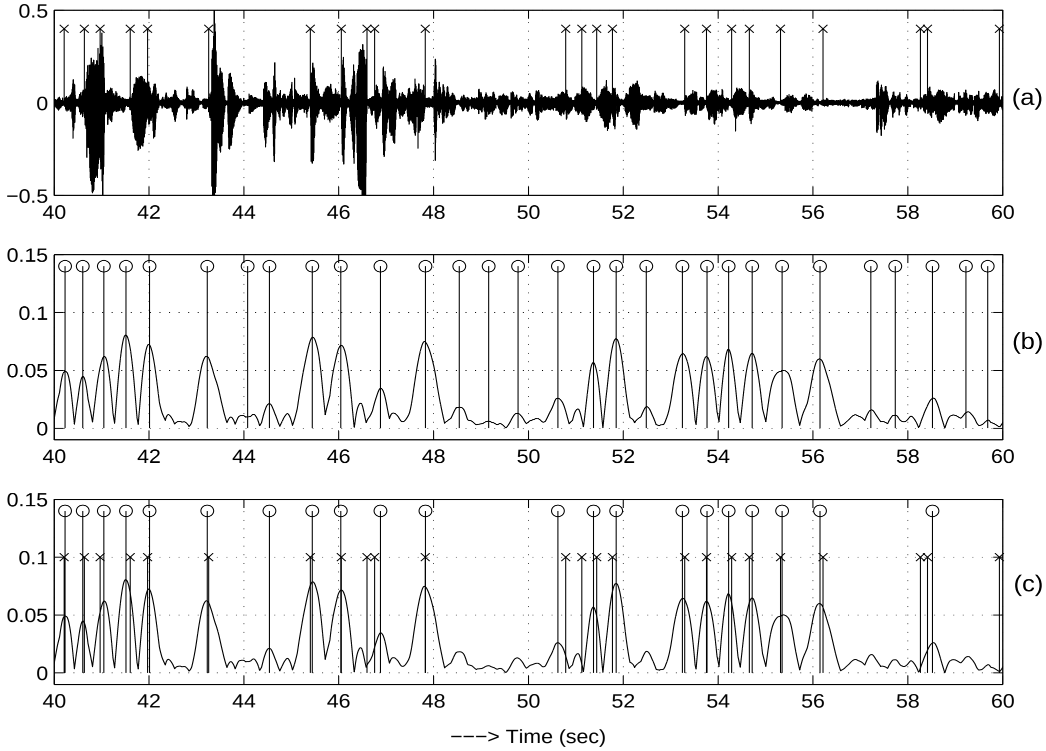

Fig. 4.6: Speaker change hypothesis and validation. (a) The speech signal with actual speaker changes marked manually. (b) The $\Delta\mu$ plot (combined using sum rule) with speaker changes hypothesized as a first step. (c) The final validated speaker changes. The actual speaker changes are marked using poles with a cross at the top.

process validates a speaker change hypothesized in the previous step. If the two segments on either side of a hypothesized speaker change are assigned to the same group, the speaker change hypothesis is said to be invalidated. When speaker separation is the primary objective of the segmentation task, oversegmentation is preferred during the chance detection phase, since a missed speaker change is more expensive than a false hypothesis.

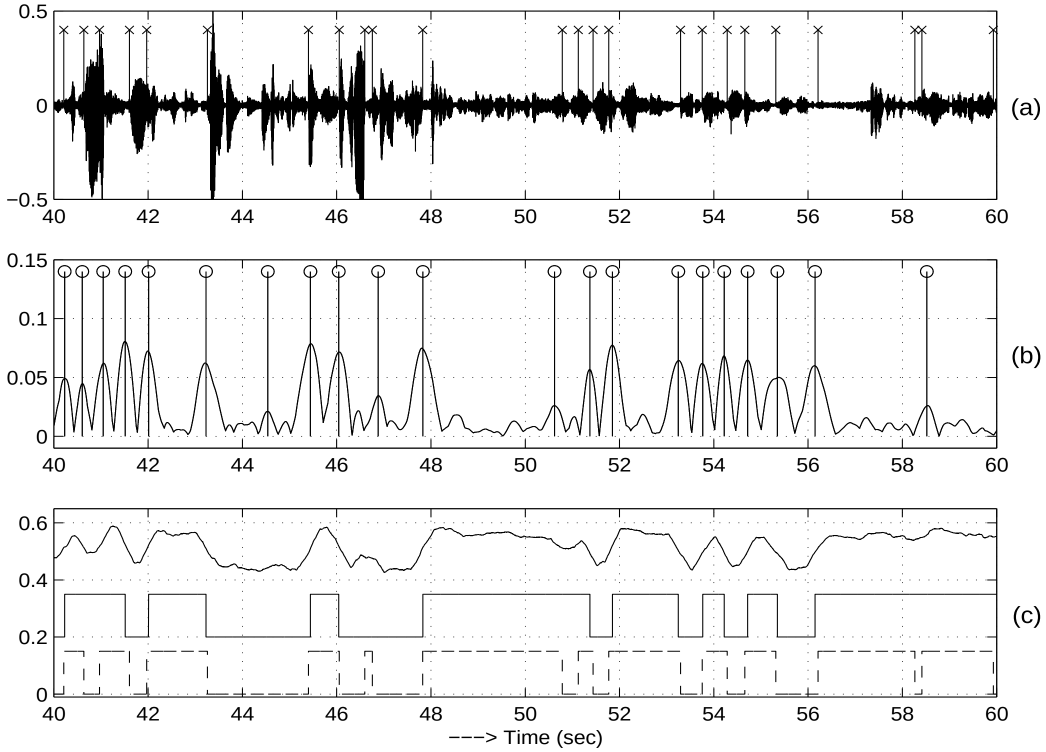

Fig. 4.7: Results of speaker separation task. (a) 2-speaker speech signal (vertical poles with a cross at the top indicate actual speaker changes). (b) $\Delta\mu$ plot (vertical poles with a circle at the top indicate hypothesized speaker changes). (c) Combined average confidence plot. The binary signal (solid) gives the decision of the speaker separation process. The actual speaker separation decision (dashed plot) is given as reference.

4.3 Performance Evaluation of the Proposed Approach

The metrics used to evaluate the performance of the speaker segmentation and separation tasks are discussed in Section 4.3.1. The dataset used for the performance studies is described in Section 4.3.2. The performance of the proposed approach for speaker segmentation and separation is discussed in Section 4.3.3. A comparison with the metric-based system using vocal tract features is also given in this section.

4.3.1 Performance evaluation metrics

The performance of the speaker change detection process is evaluated using the false alarm rate (FAR) and the missed detection rate (MDR). These are defined as below:

$$P_{FAR} = \frac{N_{FA}}{N_{ACT} + N_{HYP}} \times 100\% \tag{4.9}$$

$$P_{MDR} = \frac{N_{MISS}}{N_{ACT}} \times 100\% \tag{4.10}$$

where

N_{ACT} is the number of actual speaker changes (manually marked),

N_{HYP} is the total number of speaker changes hypothesized or detected,

N_{FA} is the number of false acceptances (a hypothesized change being wrong),

N_{MISS} is the number of actual speaker changes missed out.

There are other possible definitions of FAR, $P_{FAR1} = \frac{N_{FA}}{N_{ACT}} \times 100\%$ and $P_{FAR2} = \frac{N_{FA}}{N_{HYP}} \times 100\%$. While P_{FAR1} is not restricted to an upper limit of 100%, the definition of P_{FAR2} does not take care of the number of actual speaker changes. Hence the number of false alarms is weighed against a sum of the $N_{ACT} + N_{HYP}$. An ideal system should give an FAR of 0% and an MDR of 0%. The accuracy of the speaker changes detected, measured in terms of the deviation of the hypothesized changes from the manually marked ones, is also an important factor while evaluating the performance of a segmentation algorithm.

The performance of the speaker separation process is measured in terms of the *segmentation cost* function [5] given by

$$C_{seg} = 1 - T_c/T_t \tag{4.11}$$

where T_c is the duration of voiced speech that is correctly segmented and T_t is the total duration of the voiced speech in the conversation. The cost function is normalized by a factor C_{def}, which is the minimum segmentation cost that can be obtained even without processing the conversation (by assigning the entire conversation to either of the speaker).

$$C_{def} = min \left\{ \begin{array}{c} C_{seg} | Entire\ speech\ is\ assigned\ to\ speaker\ 1, \\ C_{seg} | Entire\ speech\ is\ assigned\ to\ speaker\ 2 \end{array} \right\} \tag{4.12}$$

$$C_{norm} = C_{seg} / C_{def} \tag{4.13}$$

A good system should give a C_{norm} value close to zero, and a value close to one is as good as not processing the conversation.

4.3.2 Dataset for performance studies

A total of 10 different 2-speaker conversations, each of 5 min duration are used to evaluate the performance of speaker segmentation system. The 10 conversations constitute 5 female-female conversations, 3 male-male conversations and 2 male-female conversations. The data set has a total of 880 (N_{ACT}) actual speaker changes (manually marked). The data is part of the NIST-2003 database for speaker verification tasks and are casual telephonic conversations recorded at the switchboard [5]. The data is recorded at a sampling rate of 8 kHz. The data has approximately 3000 sec of conversational speech, which contains around 1149 sec (T_t) of voiced speech.

4.3.3 Results of speaker segmentation and separation

The performance of the proposed approach for speaker segmentation depends on the following parameters:

- Analysis window length T_A

- The threshold λ for validating speaker change hypotheses
- The logic used to combine evidence
  - Score-level combination
  - Decision-level combination

The effect of these parameters on the performance of speaker segmentation is discussed in this section. Also, the effect of the analysis window length T_A and the combination logic used to combine scores, on the performance of speaker separation is discussed. The performance of speaker change detection after the speaker separation process is studied. A comparison of the proposed approach with the metric-based approach (described in Chapter 2) is given towards the end of the section.

4.3.3.1 Accuracy of the speaker change hypotheses

The accuracy of the hypothesized speaker changes is measured in terms of the deviation from the manually marked speaker changes. The length of the analysis window T_A controls the maximum tolerance in accuracy (deviation from the manual markings). The comparison being done between two adjacent windows of size T_A each, a tolerance T_d greater than $T_A/2$ is not logical. Here, the error in the manually marked speaker changes is assumed to be zero. If T_{err} is assumed to be the average error in manual markings, then the tolerance for automatic detection should be within $T_{max} = T_A/2 + T_{err}$. Assuming that the average error in manual markings is close to zero, we set an upper limit on the maximum tolerance at $T_{max} = T_A/2$.

4.3.3.2 Effect of analysis window length T_A on speaker segmentation

The performance of the proposed speaker change detection algorithm is given in Table 4.2 for different lengths of the analysis window T_A (100 ms, 250 ms and 500 ms).

In each case, the performance is evaluated for different tolerance limits, varying upto a maximum tolerance of $T_A/2$. It is seen that the missed detection rate for a common tolerance of 50 ms is less for an analysis window size of 100 ms. However, the number of false hypotheses is high for smaller lengths of the analysis window. The larger window lengths mask some of the finer variations in the features within the window and hence can afford a larger tolerance. This results in a reduction in the number of misses for the maximum allowable tolerance, as the size of the analysis window is increased. It is seen that a best MDR of 13.52% can be obtained at an FAR of 33.06%, with a maximum allowable tolerance of 250 ms.

Table 4.2: Performance of the speaker change detection task for different lengths of the analysis window T_A. (Score-level combination using sum rule, peak validation threshold $\lambda = \mu - 0.5\sigma$, $N_{ACT} =$ 880).

T_A	Tolerance	N_{HYP}	N_{FA}	N_{MISS}	$P_{FAR}(\%)$	$P_{MDR}(\%)$
100 ms	10 ms	3188	2938	631	72.2	77.39
	20 ms	3188	2693	383	66.2	43.52
	50 ms	3188	2541	218	62.46	24.7
250 ms	25 ms	1824	1076	633	48.82	71.93
	50 ms	1824	1397	448	51.66	50.91
	125 ms	1824	1219	238	45.08	27.05
500 ms	50 ms	1428	1200	647	51.99	73.52
	100 ms	1428	1019	447	44.15	50.08
	250 ms	1428	763	119	33.06	13.52

4.3.3.3 Effect of validation threshold λ on speaker segmentation

The performance of the speaker change detection task depends on the dynamic threshold λ, used to validate the speaker changes hypothesized at the first level. Table 4.3 compares the performance for different thresholds against the case where no validation is done. It can be seen from the table that setting the threshold is a compromise between a low MDR and a low FAR.

Table 4.3: Performance of the speaker change detection task for different thresholds in validating peaks. (Analysis window of duration $T_A = 500$ ms, detection accuracy $T_d = 250$ ms, $N_{ACT} = 880$).

Validation threshold λ	N_{HYP}	N_{FA}	N_{MISS}	$P_{FAR}(\%)$	$P_{MDR}(\%)$
No validation	1680	988	105	38.59	11.93
$\mu - 0.5 * \sigma$	1428	763	119	33.06	13.52
$\mu - 0.25 * \sigma$	966	367	200	19.88	22.73
μ	615	156	375	10.43	42.61

4.3.3.4 Effect of different combining strategies on speaker segmentation

The effect of using different strategies to combine evidence from multiple models, on the speaker change detection performance, is shown in Tables 4.4 and 4.5. Table 4.4 corresponds to combining the evidence at score level. It can be seen that the number of misses is more using the product rule, which is analogous to AND logic. The number of false alarms has also increased, which is probably counter-intuitive for AND logic. The increase is due to the fact that the evidence is combined at the score level as can be verified from Fig. 4.4.

Table 4.4: Performance of the speaker change detection task for different strategies of combining the scores. (Analysis window of duration $T_A = 100$ ms, peak validation threshold $\lambda = \mu - 0.5\sigma$, detection accuracy $T_d = 50$ ms, $N_{ACT} = 880$).

Rule for score-level combination	N_{HYP}	N_{FA}	N_{MISS}	$P_{FAR}(\%)$	$P_{MDR}(\%)$
Product rule	3444	2824	249	65.31	28.3
Sum rule	3188	2541	218	62.46	24.7

Table 4.5: Performance of the speaker change detection task by combining decision based on individual evidence. (Analysis window size $T_A = 100$ ms, peak validation threshold $\lambda = \mu - 0.5\sigma$, detection accuracy $T_d = 50$ ms, $N_{ACT} = 880$).

Logic for decision-level combination	N_{HYP}	N_{FA}	N_{MISS}	$P_{FAR}(\%)$	$P_{MDR}(\%)$
AND	1603	1252	522	50.42	59.3
OR	4715	3967	160	70.9	18.18

The performance by combining the decisions after processing the individual evidence separately is given in Table 4.5. The decisions are said to be corroborative if they are within a tolerance limit of T_{tol}. Comparing results from Table 4.4 and Table 4.5, it can be seen that the OR logic reduces the number of misses, increasing the false alarms at the same time. The AND logic increases the number of misses severely, while reducing the false alarms.

4.3.3.5 Performance of the speaker separation task

The speaker separation process prefers a few additional false alarms rather than missing out a genuine speaker change. Hence when the goal of the speaker segmentation task is to separate the segments of the conversant speakers, oversegmentation is always preferred at the change detection stage. The performance of the speaker separation task for different lengths of the analysis window is given in Table 4.6. The length of the analysis window seems to have little effect on the speaker separation performance as compared to the change detection performance (Table 4.2). The C_{seg} values shows that around 92 to 94% of the overall speech is assigned to the correct speaker. However, a C_{def} value of 0.3975 specifies that around 60% of the conversation can be correctly assigned even without any processing. Hence by using C_{def} as a reference, the C_{norm} values suggest that an accuracy of around 81 to 85% is achieved. The evidence from the two selected models is combined at the score level. The results in Table 4.7 show that the performance of the speaker separation task does not vary much. This may be due to the fact that only two models are considered and a high degree of similarity exists in the evidence provided by the two models.

Table 4.6: Performance of the speaker separation task for varying sizes of the analysis window. (Peak validation threshold $\lambda = \mu - 0.5\sigma$, sum rule for combining scores, $T_t =$ 1149 sec, $C_{def} =$ 0.3975).

Analysis window size (T_A)	T_c	C_{seg}	C_{norm}
100 ms	1079.4	0.0601	0.1511
250 ms	1071.7	0.0673	0.1693
500 ms	1063.1	0.0748	0.1883

Table 4.7: Performance of the speaker separation task for different strategies for combining evidence at score-level. (Analysis window size $T_A =$ 100 ms, peak validation threshold $\lambda = \mu - 0.5\sigma$, sum rule for combining scores, $T_t =$ 1149 sec, $C_{def} =$ 0.3975).

Score combination logic	T_c	C_{seg}	C_{norm}
Product rule	1084.4	0.0562	0.1414
Sum rule	1079.4	0.0601	0.1511

4.3.3.6 Validation of speaker change hypotheses by speaker separation

The speaker changes hypothesized by the speaker segmentation process act as the starting point for speaker separation. The segments of speech bound by the speaker change hypotheses are grouped into two categories during speaker separation. In the process, a speaker change hypothesis may either be validated or invalidated. The performance of the speaker change detection task reevaluated after the speaker separation task is given in Table 4.8. The performance of the speaker segmentation task before speaker separation is also given for comparison. It can be seen that there is an increase in the number of misses, but at the same time there is a reduction in the number of false alarms. This is mainly due the elimination of weak and spurious speaker change hypotheses by the speaker separation process. It can seen that a missed detection rate of around 40 to 50% (P_{MDR} of 38.64 and 50.45 from Table 4.8), corresponds to an inaccuracy of 6 to 7.5% in speaker separation (C_{seg} values of 0.0604 and 0.0748 from Table 4.6). This shows that the segmentation cost C_{seg} is heavily biased towards the longer speaker turns. While it is a good measure for the speaker separation process, it is not best suited for evaluating the performance of speaker segmentation.

Table 4.8: Effect of speaker separation on the performance of speaker change detection. (Peak validation threshold $\lambda = \mu - 0.5\sigma$, sum rule for combining scores, detection accuracy of $T_A/2$, $N_{ACT} = 880$).

Speaker segmentation	Analysis window size (T_A)	N_{HYP}	N_{FA}	N_{MISS}	$P_{FAR}(\%)$	$P_{MDR}(\%)$
Before speaker separation	100 ms	3188	2541	218	62.46	24.7
	250 ms	1824	1219	238	45.08	27.05
	500 ms	1428	763	119	33.06	13.52
After speaker separation	100 ms	859	325	340	18.69	38.64
	250 ms	535	92	419	6.5	47.61
	500 ms	429	34	444	2.6	50.45

4.3.3.7 Comparison of the proposed approach and the metric-based approach

The performance of the proposed approach for speaker segmentation and the metric-based (BIC) approach (described in Chapter 2) is given in Table 4.9. An analysis window of length 500 ms, a peak validation threshold of $\lambda = \mu - 0.5 * \sigma$ and a 250 ms tolerance on the accuracy of detected changes are used for both methods. It is seen that the proposed approach for speaker segmentation using excitation source information performs significantly better than the metric-based approach that uses vocal tract information. Also, the speaker separation performance by the proposed approach is better than the metric-based approach as shown in Table 4.10.

Table 4.9: Performance of the speaker segmentation task for the proposed approach and the metric-based approach. (Analysis window size $T_A =$ 500 ms, peak validation threshold $\lambda = \mu - 0.5\sigma$, detection accuracy of $T_A/2$, $N_{ACT} =$ 880

Speaker segmentation using	N_{HYP}	N_{FA}	N_{MISS}	$P_{FAR}(\%)$	$P_{MDR}(\%)$
Excitation source features	1428	763	119	33.06	13.52
Vocal tract features	2072	1825	631	61.82	71.7

Table 4.10: Performance of the speaker separation task for the proposed approach (using excitation features) and the metric-based approach (using vocal tract features). (Analysis window size $T_A =$ 500 ms, peak validation threshold $\lambda = \mu - 0.5\sigma$, $T_t =$ 1149 sec, $C_{def} =$ 0.3975)

Speaker separation using	T_C	C_{seg}	C_{norm}
Excitation source features	1063.1	0.0748	0.1883
Vocal tract features	838.66	0.2701	0.6795

4.4 Advantages and Limitations of the Proposed Approach

Some of the important advantages and limitations of the proposed approach can be summarized as follows:

Advantages:

- The dissimilarity of the excitation features within a speaker is less as compared to the dissimilarity with sounds of other speakers. This makes the excitation source features a better option for detecting speaker changes in conversations with short speaker turns.
- The excitation features are less affected by channel variations, as compared to the vocal tract features. Hence the excitation features may give consistent segmentation performance, irrespective of the channel of recording.

Limitations:

- The proposed method requires around one second of contiguous speech (voiced) containing only one speaker. Atleast two such segments are essential for automatic detection of such regions. The performances of the speaker segmentation and separation tasks can degrade if the the above requirements are not met.
- The proposed method requires some adaptation time and data for every conversation it tries to segment. Hence it is not suited to process conversations on a real-time basis. But once the models are generated, the conversations may be segmented on a real-time basis.

4.5 Summary

A new approach for speaker segmentation using the excitation source features was proposed in this chapter. A simple agglomerative clustering was used to separate the segments belonging to the two speakers. The various analysis discussed in Section 4.3.3 on the performance of the speaker segmentation and separation tasks are tabulated in Table 4.11. The performance of the speaker segmentation task depends on the analysis window length, validation threshold and the decision strategy used to combine evidence. The choice of these parameters is a compromise between a low missed detection rate and a low false alarm rate. The results have shown that the proposed approach based on excitation source features performs significantly better than the existing metric-based approach that use vocal tract features. Some of the directions for future work based on the proposed approach will be discussed in Chapter 6. Chapter 5 discusses the application of speaker segmentation and separation in the 2-speaker detection task.

Table 4.11: Summary of analysis on the performance of speaker segmentation and separation tasks.

- Performance of speaker segmentation for different lengths of the analysis window.
    - Section 4.3.3.2, Table 4.2
- Performance of speaker segmentation for different values of the validation threshold.
    - Section 4.3.3.3, Table 4.3
- Performance of speaker segmentation for different strategies in combining evidence.
    - Combination of evidence at the score level.
        - Section 4.3.3.4, Table 4.4
    - Combination of evidence at the decision level.
        - Section 4.3.3.4, Table 4.5
- Performance of speaker separation for different lengths of the analysis window.
    - Section 4.3.3.5, Table 4.6
- Performance of speaker separation for different strategies in combining evidence.
    - Section 4.3.3.5, Table 4.7
- Reevaluation of the performance of speaker segmentation after speaker separation.
    - Section 4.3.3.6, Table 4.8
- Comparison of the proposed approach with the metric-based approach.
    - Section 4.3.3.7, Tables 4.9 and 4.10.

CHAPTER 5

APPLICATION OF SPEAKER SEGMENTATION IN 2-SPEAKER DETECTION

In Chapter 3, it was shown that the excitation features at the subsegmental level can characterize a speaker better than the vocal tract features, from limited data. In the previous chapter, a new approach for segmentation and separation of speakers in a conversation, was proposed. The 2-speaker detection task (also referred to as the 2sp task), which is of significance in forensic applications, is considered in this chapter as an application of the speaker segmentation and separation tasks. The objective here is to demonstrate the impact of speaker segmentation on the performance of the 2sp task. A detailed study on various issues in the 2sp task is not within the scope of this chapter.

5.1 2-speaker Detection Task - An Overview

The 2-speaker detection task involves detecting or verifying a speaker within a conversation. There could be several variations to the definition of the task. In this thesis, the 2sp task defined by NIST (National Institute of Standards and Technology, USA) as part of the annual speaker verification evaluation [5, 66], is considered. The NIST 2sp task has two phases, a training phase and a testing phase. The training phase involves building a model for a speaker from three different conversations (each of five minutes duration) of the speaker with three different speakers. The common speaker has to be identified before building a model for the speaker. The definition of the task ensures that only one speaker is common among the three conversations available for training. The second speaker in each of the conversations does not repeat in any of the other two conversations. Also, it is given that each of the speakers in a conversation contributes to atleast 20% of the overall duration, with an average contribution of 50%. During the testing phase, a single conversation (of five minutes duration) is to be tested against the model of a claimant speaker. It is to be verified if the claimant speaker is one of the speakers in the test conversation. Fig. 5.1 gives an overview of the 2sp detection task.

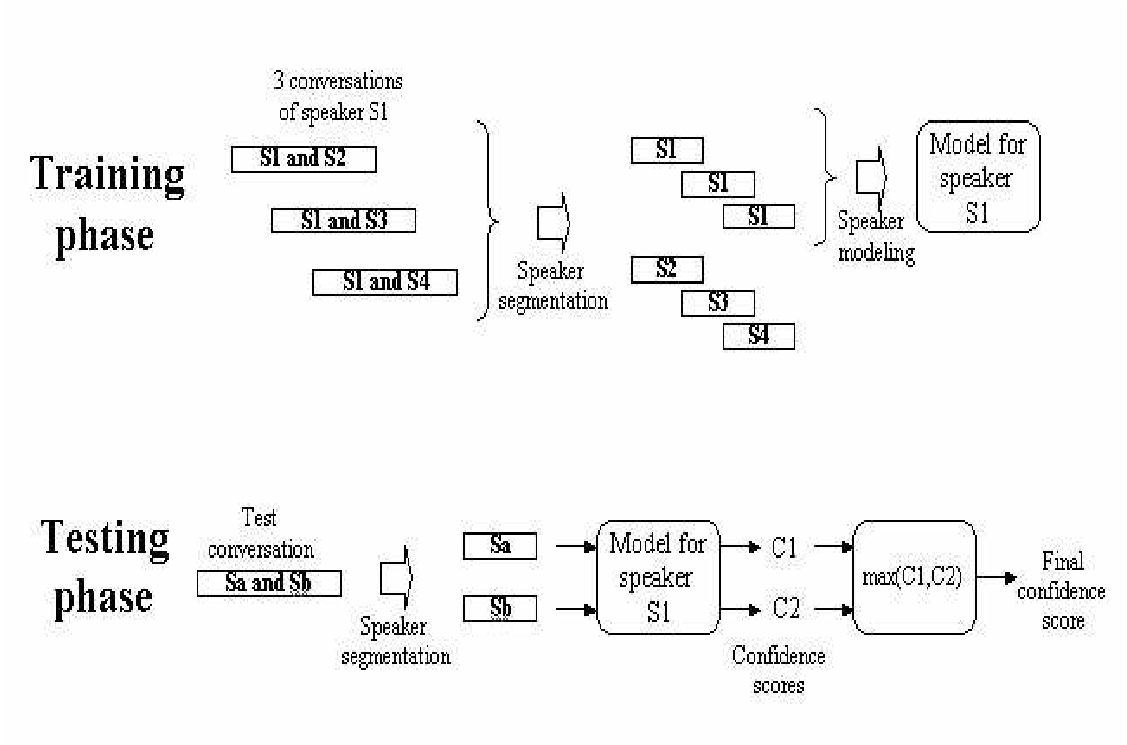

Fig. 5.1: Overview of the 2sp detection task.

5.2 2-speaker Detection Without and With Speaker Segmentation

In order to study the importance of the speaker segmentation and separation tasks, the performance of the 2sp task is studied under two cases - (a) without speaker segmentation and (b) with speaker segmentation. In both the cases, a strategy similar to a 1-speaker detection task is employed. Hence, the general strategy for 1-speaker detection will be discussed before the actual 2sp tasks.

5.2.1 1-speaker detection strategy

The 1-speaker detection task (also referred to as 1sp task), in general, has two major phases - speaker modeling (training) phase and speaker verification (testing) phase. In the training phase a parametric model is developed to characterize a speaker from the training data provided for the speaker. The vocal tract features are found to characterize a speaker better than the excitation source features, when large amount ($>$ 30 sec) of training data is available [14,26]. Hence, 19-dimensional LPCCs obtained using a 12^{th} order LP analysis of speech (20 ms window size and 10 ms shift) are used as feature vectors. A five layer AANN model with a structure $19L38N12N38N19L$ is used to capture the distribution of the vocal tract features [67]. During the testing phase, the LPCC features extracted from the test utterance are presented to the claimant model and an average confidence score is obtained [62]. The average confidence score indicates the probability that the test speaker is the same as the claimant speaker. A threshold is set on the confidence score to decide the authenticity of the claim. The 1sp system used in these experiments is based on the speaker verification systems detailed in [62, 68–70].

To evaluate the performance of a speaker verification system, several test utterances are considered and each test utterance is tested against several claimants. The average confidence scores, obtained by testing different test utterances against different claimant speakers, can be in different ranges due to the variability in the type of data used for training and testing. Hence, there is a need for normalizing these confidence scores to a common range, so that a common threshold be set, independent of the test and claimant speaker pair, and independent of the type of data used for training or testing. Model normalization, followed by a test utterance normalization is performed on these confidence scores, to reduce the effect of variability in training data and test data respectively [68, 71]. If a confidence score is above the common threshold, the test and the claimant speakers are declared to be the same.

5.2.2 2-speaker detection without speaker segmentation

The definition of the 2sp task ensures that only one speaker is common among the three conversations available for training. Also, the second speaker in each of the conversations does not repeat in any of the other two conversations. Hence, the overall training data contains four speakers and the target speaker has almost three times the data of any other speaker. A single AANN model is trained using all the training data without separating the speakers in conversation. The hypothesis is that the AANN model captures the distribution of features corresponding to all four speakers, but the target speaker who has more data than any other speaker is modeled better. During testing, the entire conversation is tested against a specified target speaker model. The confidence scores of only those frames (or feature vectors) which are above the mean confidence score are considered to compute the final average confidence score. The hypothesis is that, if the target speaker is one of the speakers in the test conversation, the frames of the target speaker give higher confidence scores than that of the other speaker. At the same time, if the target speaker is neither of the test speakers, all frames of the test conversation should give low scores. The drawback of this approach is that the training patterns for the AANN models are corrupted with data from other speakers and can inhibit the ability of the AANN model to characterize a speaker. Also, if the claimant happens to be one of the three non-target speakers in the training data, the performance of the 2sp task can degrade.

5.2.3 2-speaker detection with speaker segmentation

Speaker segmentation followed by a speaker separation task is performed on the training conversations as well as the test conversation, based on the approach proposed in the previous chapter. An analysis window length of $T_A = 100$ ms, a peak validation threshold $\lambda = \mu - 0.5\sigma$ and sum rule for combining evidence are used for speaker segmentation and separation. The speaker separation process splits a conversation into two clusters, each representing one speaker. The common speaker from the three training conversations is determined by performing an agglomerative clustering, which starts with six initial clusters and ends with four clusters. As there is considerable amount of data in each of the clusters, LPCCs are used as features and ΔBIC is used as the dissimilarity measure. The largest cluster is hypothesized as the common or target speaker. During testing, a claim is tested against both the speakers separated from a test conversation separately, and the largest confidence score is retained.

5.3 Performance of 2-speaker Detection Systems

The performance of the system is evaluated in terms of equal error rate (EER). The EER is that value of false acceptance rate (FAR) or missed detection rate (MDR) at which both are equal. A false acceptance is a case when an impostor is declared as a genuine speaker, and a missed detection is a case when a genuine speaker is declared as an impostor. The false acceptance rate and missed detection rate vary as the threshold on the normalized confidence scores is varied.

The performance of the 2sp system is evaluated on a small set of 50 target speakers. A total of 150 conversations for training speaker models (3 per target speaker) and 50 test conversations (one each per target speaker) are considered. Eleven claims are made against each of the test conversations, of which only one is genuine. A claim or a claimant corresponds to an already modeled speaker, and a genuine claim is one where the claimant speaker is one of the conversants in the test conversation. Thus a total of 550 claims are verified, of which 50 are genuine claims and the remaining 500 are impostor claims. Apart from the EER, the performance of the 2sp task is measured using a rank based method. The eleven confidence scores obtained for the eleven claims against a single test conversation, are ranked against each other. The number of times (out of the total 50 tests) a genuine claim wins over the remaining 10 impostor claims is used as a performance measure. The performance of the 2sp task without and with speaker segmentation is given in Table 5.1. The table clearly shows that the separation of the two speakers in a conversation brings about a significant improvement in the performance of the 2-speaker detection task.

5.4 Summary

The 2-speaker detection task was considered as an application in order to illustrate the usefulness of the speaker segmentation and separation tasks. Based on the principle of a 1-speaker verification system, two different 2sp systems were designed, one without speaker segmentation and the other with speaker segmentation. The performance

Table 5.1: The performance of the 2sp task without and with speaker segmentation.

2sp task	Number of genuine first ranks (out of 50)	EER (%)
Without speaker segmentation	26	33
With speaker segmentation	39	24

of the 2sp task was shown to improve when speaker segmentation and separation were performed before the verification task, thus signifying the importance of speaker segmentation and separation.

CHAPTER 6

SUMMARY AND CONCLUSIONS

The existing methods for speaker segmentation provide a statistical solution to detect a point phenomenon (speaker change). They rely on the vocal tract features, which have a high degree of dissimilarity among sounds of the same speaker as compared to the same sounds in different speakers. This limits the existing approaches from being used on casual conversations that contain short (< 5 sec) speaker turns.

An alternate approach for detecting speaker changes using excitation source features has been proposed. AANN models were used to capture the excitation features in the LP residual signal. The proposed method for speaker segmentation using excitation source features was found to perform better than the system using vocal tract features. Use of speaker segmentation in 2-speaker detection task was shown to improve the performance of the task.

6.1 Contributions of the Work

The contributions of the research work carried out as part of this thesis can be summarized as follows:

- The limitations of the existing approach using vocal tract features in segmenting conversations with short speaker turns are brought out.
- The within-speaker to across-speaker (WAD) ratio is introduced as a metric to evaluate the ability of a feature set in characterizing a speaker from limited

amount of data.

- The excitation features are shown to be better than the vocal tract features for characterizing speakers from limited data.
- An alternate approach for speaker segmentation of conversational speech using excitation source features is proposed. The proposed approach was shown to perform better than the existing approach based on vocal tract features.
- An approach for speaker separation task, which uses excitation source features for combining segments is proposed.
- The ability of speaker segmentation in improving the performance of the 2-speaker detection task is demonstrated.

6.2 Scope for Further Research

Some of the directions for continuing the research carried out as part of this thesis are as follows:

- The proposed approach for speaker change detection can be easily extended to process multispeaker speech with more than two speakers.
- The current work was focused on detecting speaker changes in casual telephone conversations. The performance of the proposed algorithm can be studied for different types of data, with different channel and noise considerations.
- The proposed algorithm requires some amount of data from the conversation for the generation of models. This can be a limitation in processing short clips of a conversation. Also, the generation of the models take significant time making the approach unsuitable for real-time processing of multispeaker speech. However,

once the models are trained, the conversations can be processed in almost real-time. It is observed that the models trained with data less than one second also contain evidence for detecting speaker changes. Reduction in the amount of training data, increase in the number of models for combining evidence, and efficient methods to combine evidence from multiple models may help reduce the overhead involved in training the models.

BIBLIOGRAPHY

[1] L. Lu, H. Jiang, and H. J. Zhang, "A robust audio classification and segmentation method," in *Proc. 9th ACM Multimedia*, (Ottawa, Ontario, Canada), pp. 203–211, Oct. 2001.

[2] T. Zhang and C. C. J. Kuo, "Heuristic approach for generic audio data segmentation and annotation," in *Proc. 7th ACM Multimedia*, (Orlando, Florida, USA), pp. 67–76, Nov. 1999.

[3] C. Montacie and M. J. Caraty, "A silence/noise/music/speech splitting algorithm," in *Proc. Int. Conf. Spoken Language Processing*, vol. 4, (Sydney, Australia), pp. 1579–1582, Dec. 1998.

[4] H. Gish, M. Siu and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, vol. 2, (Toronto, Canada), pp. 873–876, May 1991.

[5] "The NIST year 2003 speaker recognition evaluation plan," in *Proc. NIST Speaker Recognition Workshop*, (Baltimore, Maryland, USA), June 2003.

[6] M. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. Stevens, and H. Wactlar, "Informedia digital video library," *Commun. ACM*, vol. 38, no. 4, pp. 57–58, 1995.

[7] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proc. IEEE*, vol. 88, pp. 1338–1353, Aug. 2000.

[8] F. Kubala, S. Colbath, D. Liu, A. Srivastava, and J. Makhoul, "Integrated technologies for indexing spoken language," *Commun. ACM*, vol. 43, pp. 48–56, Feb. 2000.

[9] D. Roy and C. Malamud, "Speaker identification based text to audio alignment for an audio retrieval system," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, (Munich, Germany), pp. 1099–1102, Apr. 1997.

[10] L. Nguyen, S. Matsoukas, J. Davenport, D. Liu, J. Billa, F. Kubala, and J. Makhoul, "Further advances in transcription of broadcast news," in *Proc. Eurospeech*, (Budapest, Hungary), pp. 667–670, Sept. 1999.

[11] C. Leggetter and P. Woodland, "Flexible speaker adaptation for large vocabulary speech recognition," in *Proc. Eurospeech*, (Madrid, Spain), pp. 1155–1158, Sept. 1995.

[12] D. Pye and P. Woodland, "Experiments in speaker normalization and adaptation for large vocabulary speech recognition," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, (Munich, Germany), pp. 1047–1050, Apr. 1997.

[13] L. Neumeyer, A. Sankar, and V. Digalakis, "A comparative study of speaker adaptation techniques," in *Proc. Eurospeech*, (Madrid, Spain), pp. 1127–1130, Sept. 1995.

[14] A. Martin and M. Przybocki, “1-speaker detection,” in *Proc. NIST Speaker Recognition Workshop*, (Vienna, Virginia, USA), May 2002.

[15] L. Lu and H.-J. Zhang, “Speaker change detection and tracking in real-time news broadcasting analysis,” in *Proc.* 10^{th} *ACM Multimedia*, (Juan-les-pins, France), pp. 602–610, Dec. 2002.

[16] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[17] H. Gish and M. Schmidt, “Text-independent speaker identification,” *IEEE Signal Processing Magazine*, vol. 11, pp. 18–32, Oct. 1994.

[18] R. J. Mammone, X. Y. Zhang, and R. P. Ramachandran, “Robust speaker recognition: A feature-based approach,” *IEEE Signal Processing Magazine*, vol. 13, pp. 58–71, Sept. 1996.

[19] K. N. Stevens, *Acoustic Phonetics.* Cambridge, England: The MIT Press, 1999.

[20] D. O’Shaughnessy, *Speech Communication: Human and Machine.* Massachusetts, USA: Addison-Wesley Publishing Company, 1987.

[21] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ, USA: Prentice-Hall Inc., 1993.

[22] D. O’Shaughnessy, “Speaker recognition,” *IEEE ASSP Magazine*, vol. 3, pp. 4–17, Oct. 1986.

[23] J. P. Campbell, “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[24] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Speech and Audio Processing*, vol. 29, pp. 254–272, Apr. 1981.

[25] K. S. Reddy, “Source and system features for speaker recognition,” Master’s thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, Sept. 2001.

[26] C. S. Gupta, “Significance of source features for speaker recognition,” Master’s thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, Apr. 2003.

[27] S. Chen and P. Gopalakrishna, “Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, (San Mateo,CA:Morgan Kaufmann), pp. 127–132, Feb. 1998.

[28] S. Johnson, “Speaker tracking,” Master’s thesis, Cambridge University Engg. Dept., UK, Dec. 1997.

[29] P. Delacourt and C. J. Wellekens, “DISTBIC: A speaker-based segmentation for audio data indexing,” *Speech Communication*, vol. 32, pp. 111–126, 2000.

[30] M. Basseville, “Distance measures for signal processing and pattern recognition,” *”Signal Processing”*, vol. 18, pp. 349–369, Dec. 1989.

[31] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. 24, pp. 380–391, Oct. 1976.

[32] R. M. Gray, A. Buzo, A. H. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. 28, pp. 367–376, Aug. 1980.

[33] F. Itakura and T. Umezaki, "Distance measure for speech recognition based on the smoothed group delay spectrum," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, (Dallas, TX, USA), pp. 1257–1260, Apr. 1987.

[34] T. Kailath, "The divergence and Bhattacharya distance measures in signal selection," *IEEE Trans. Commun.*, vol. 15, pp. 52–60, 1967.

[35] D. Kazakos and P. Papantoni-Kazakos, "Spectral distance measures between Gaussian processes," *IEEE Trans. Automatic Control.*, vol. 25, pp. 950–959, Oct. 1980.

[36] D. Mansour and B. H. Juang, "A family of distortion measures based up on projection operation for robust speech recognition," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, (New York, NJ, USA), p. Apr., 1988.

[37] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York, USA: Academic Press, 1972.

[38] N. A. J. Hastings and J. B. Peacock, *Cluster Analysis.* New York, USA: Halsted Press, 1980.

[39] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," in *Proc. DARPA Speech Recognition Workshop*, (San Mateo, CA:Morgan Kaufmann, USA), pp. 108–111, Feb. 1997.

[40] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York, NJ, USA: John Wiley, 1973.

[41] B. Yegnanarayana, *Artificial Neural Networks.* New Delhi: Prentice-Hall of India, 1999.

[42] S. Haykin, *Neural networks: A Comprehensive Foundation.* New Jersey, USA: Prentice-Hall International, 1999.

[43] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* New York, NJ, USA: John Wiley & Sons, 1984.

[44] N. A. J. Hastings and J. B. Peacock, *Statistical Distributions.* New York, NJ, USA: John Wiley & Sons, 1974.

[45] R. C. Dubes, "How many clusters are best? - an experiment," *Pattern recognition*, vol. 20, no. 6, pp. 645–663, 1987.

[46] K. Mori and S. Nakagawa, "Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, (Orlando, Florida, USA), May 2002.

[47] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Pacheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 71–77, Jan. 1998.

[48] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. J. McLaughlin, and M. A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proc. Int. Conf. Spoken Language Processing*, vol. 7, (Sydney, Australia), pp. 3193–3196, Dec. 1998.

[49] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, no. 10, pp. 19–41, 2000.

[50] M. Nishida and T. Kawahara, "Speaker model selection using Bayesian information criterion for speaker indexing and speaker adaptation," in *Proc. Eurospeech*, (Geneva, Switzerland), Sept. 2003.

[51] D. K. Roy, "Speaker indexing using neural network clustering of vowel spectra," *Int. Journal of Speech Technology*, vol. 1, no. 2, pp. 143–149, 1997.

[52] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[53] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[54] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach.* New Delhi: TATA McGraw-Hill, 2003.

[55] J. G. Proakis and D. G. Manolakis, *Digital Signnal Processing.* New Delhi: Prentice-Hall of India, 2000.

[56] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Englewood Cliffs, New Jersey, USA: Prentice Hall, 1975.

[57] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Communication*, pp. 167–184, 1982.

[58] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Speech and Audio Processing*, vol. 27, pp. 309–319, Aug. 1979.

[59] D. G. Childers and C. Ahn, "Modeling the glottal volume-velocity waveform for three voice types," *Journal of the Acoustical Society of America*, vol. 97, pp. 505–519, Jan. 1995.

[60] G. Fant, J. Liljencrants, and Q. G. Lin, "A four-parameter model of glottal flow," Q. Prog. Stat. Rep., Speech Trans. Lab., R. Inst. Technol., Stockholm, Sweden, vol. 4, pp. 1-17, 1985.

[61] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 569–586, Sept. 1999.

[62] B. Yegnanarayana, K. S. Reddy, and S. P. Kishore, "Source and system features for speaker recognition using AANN models," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, vol. 1, (Salt Lake city, Utah, USA), pp. 409–412, May 2001.

[63] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 1–11, Jan. 1998.

[64] C. d'Alessadro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," vol. 6, pp. 12–23, Jan. 1998.

[65] A. Papoulis, *Probability, Random variables and Stochastic Processes.* New York, NJ, USA: McGraw-Hill, 1965.

[66] Bonastre. J. F and Meignier. S and Merlin. T, "Speaker detection using multi-speaker audio files for both enrollment and test," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, (Hong Kong), pp. 77–80, Apr. 2003.

[67] B. Yegnanarayana and S. P. Kishore, "AANN - an alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, Apr. 2002.

[68] B. Yegnanarayana and et. al., "IIT Madras - Speaker recognition system," in *Proc. NIST Speaker Recognition Workshop*, (Baltimore, Maryland, USA), June 2003.

[69] M. S. Ikbal, "Autoassociative neural network models for speaker verification," Master's thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, Dec. 1999.

[70] S. P. Kishore, "Speaker verification using autoassociative neural network models," Master's thesis, Dept. of Computer Science and Engineering, IIT Madras, Chennai, Dec. 2000.

[71] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital signal processing*, no. 10, pp. 42–54, 2000.

LIST OF PUBLICATIONS

Presentations in Conferences :

1. Guruprasad. S, Dhananjaya. N and B. Yegnanarayana, "AANN Models for Speaker Recognition Based on Difference Cepstrals," in *Proc. Int. Joint Conf. Neural Networks*, (Portland, OR, USA), pp. 692-697, July 2003.

2. Dhananjaya. N, Guruprasad. S, and B. Yegnanarayana, "Speaker Segmentation Based on Subsegmental Features and Neural Network Models," Communicated to *Int. Conf. Neural Information Processing*, to be held in Science City, Calcutta, during Nov. 22-25, 2004.