CHAPTER 1
INTRODUCTION

Speech is firmly believed as a natural mode of communication among human beings which is endowed with both sound production and perception mechanisms. Speech processing is required without the knowledge of speech production and discernment if a machine is placed in the communication chain.

The speech is recorded by unidirectional microphone and store in a medium such as magnetic tape as analog format which is necessary to convert in a digitized format for digital devices related applications. Computers can readily use it where stored as digital audio samples is known as Pulse Code Modulation (PCM) which is a sample quantizing. It linearly converts all analog signals to digital samples.

The quality of audio is determined by Sample resolution and Sampling rate. The sample resolution is defined by the number of bits per sample and the sizes are 8 bits and 16 bits. The larger sampling size gives the higher quality. The sampling rate or sampling frequency defines the number of samples recorded per second taken from a continuous signal to make a digital or discrete signal. It is represented in terms of Hertz. The higher sampling rate gives the good quality of the audio which collects more data per second than a lower sampling rate which requires more memory.
To facilitate the retrieval of relevant information is important to provide easy access to multimedia documents. Recently, indexing [22], [23], [24], [25] and [26] audio-specific documents such as radio broadcast news or the audio channel of video materials which is mostly consisted of running automatic speech recognizers (ASRs) to extract syntactic or higher level information.

The recorded speech file has been processed in three levels. They are

(i) Acoustic-based information like speaker turns, the number of speakers, speaker genders, speaker identities, other sounds (music, laughs) as well as speech bandwidth or characteristics (studio quality or telephone speech, clean speech or speech over music) can be extracted and added to syntactic information.

(ii) The Information directly linked to the spontaneous nature of speech like disfluencies (hesitations, repetitions, etc.) or emotion is also relevant for rich transcription.

(iii) Linguistic or pragmatic information such as named entity or topic extraction, for instance, particularly interesting for seamless navigation or multimedia information retrieval.

Some types of information extraction are relevant to document structure which is not fall exactly into one category; for example, the detection of sentence boundaries to be based on acoustic cues but also on linguistic ones.

1.1 Basics of Speech Processing

Nowadays man-machine interface seems to be essential for applications such as voice response systems, voice-based person authentication system, information retrieval, and speech coding. The basic speech tasks that are involved in man-
machine interfaces are (a) speech recognition (speech to text conversion), (b) speech synthesis (text to speech conversion), (c) speaker recognition, (d) language identification, (e) speech enhancement and (f) speech coding. The type of speech signal processing depends on the particular speech task.

SS contains knowledge sources at different levels. At segmental level (10-30 ms) speech signal [96], [97] contains information about pitch, formants, position and movement of articulators and shapes of the oral and nasal cavities. At sub-segmental level (3-5 ms), details of glottal pulse shape, open and closed phase of glottis can be observed. At suprasegmental level (>100 ms), characteristics of longer speech segments such as prosody (pitch and duration), stress, prominence, melody, syntax and semantics, emotional state of the speaker and language are present. The co-articulation knowledge consists of the articulatory (production) constraints due to linguistic context. All the information required to perform the basic speech processing task is implicitly present in the speech. The speech processing issue, extract specific features and leads to perform the desired tasks.

The processing speech signal [98], [99] one should be aware of its production and perception characteristics. Speech is produced by exciting time varying vocal tract system with time varying excitation [96], [97]. The schematic diagram of the physiology of speech production is shown in Fig. 1.1.
Fig. 1.1 Physiology of speech production

Speech production mechanism [96], [100] and [97] essentially consists of a vibrating source of sound coupled to a resonating system. Majority of the sounds produced by the larynx acts as the vibrating source and the air column from larynx to the lips, referred to as the vocal tract, acts as the system. The vocal tract system consists of pharynx, oral cavity, and nasal cavity. The production produce some special sounds are called nasal sounds, the nasal tract also plays an important role along with the vocal tract.

The nasal tract begins at the velum and ends at the nostrils, if the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech. But, it is a known fact that no sound can be produced without a supply of force or energy. It is the breathing mechanism, consisting of the lungs and muscles of the chest and abdomen that constitutes the energy supply. The vocal cords are brought together by the use of laryngeal muscles, to form a shelf
across the air way, which leads to the lungs into the trachea. Pressure on the underside of the shelf rises when the edges of cords are held together of pressure rises a certain level, it is sufficient to overcome the resistance offered by the obstruction, and so the vocal cords open.

The ligaments and muscle fibers that make up the vocal cords have a degree of elasticity, and having been forced out of position, they tend to return as rapidly as possible to their initial position. The pressure rises again and the cycle of opening and closing repeated. The major excitation for speech production is due to periodic vibration of vocal folds. This is also known as voiced excitation, since all vowels are produced with this excitation. Other excitations are due to either complete or narrow constriction at different places in the vocal tract system which produces different sound units in response to different excitations by assuming different shapes. Compared to speech production mechanism and speech perception is less understood.

The processing speech signal, acoustic variations (pressure variations) are to be represented in digital domain. A microphone is used to pickup these acoustic variations and convert them into equivalent analog electrical variations. This analog electrical signal is converted to digital signal by using an analog to digital converter, which consists of sampler followed by quantizer and an encoder. The basic issues in digitizing the speech are the influence of transfer characteristics of the microphone, sampling frequency and number of levels in the quantizer. Different features can be extracted from speech signal by analyzing the signal in different ways. The uses of linear prediction (LP) analysis is a set of coefficients are derived for each analysis frame. These coefficients are called linear prediction coefficients (LPCs). The frequency spectrum of the analysis speech frame to be represented by a set of parameters known as spectral coefficients. Another representation of speech signal
is the cepstral domain representation. Cepstrum is defined as the inverse fourier transform of the logarithmic spectrum. The coefficients derived from the cepstrum are called cepstral coefficients. In a similar manner, linear prediction cepstral coefficients (LPCCs) are derived from LP spectrum. Based on the sensitivity of the human ear, features can be extracted with high resolution at low frequency and low resolution at high frequency end.

1.2 Speaker Diarization

Now a days, a rapid increase in the volume of recorded speech is manifested which includes television and audio broadcasts, voice mails, meeting and other spoken documents. There is a growing need to apply automatic human language technologies to allow efficient and effective searching, indexing and accessing of these information sources. Diarization [14], [15], [37, [38], [39] [89], [90] and [40] can be used for helping speech recognition, facilitating the searching and indexing of audio archives and increasing the richness of automatic transcriptions, making them more reliable and potentially helping with other tasks such as summarization, parsing and machine translation.

The speaker diarization method [1], [44], [45], [70], [91], [92] and [46] detects speech start and stop times and identifies the speaker (speaker tracking). So, it is the process of automatically partitioning a conversation involving multiple speakers [4], [5] and [47] into homogeneous segments and grouping together all the segments that correspond to the same speaker as described in Fig. 1.2. The first part of the process is known as speaker segmentation [5] or speaker change detection [41], [42] and [43] while the second one is called as speaker clustering Hence speaker change detection [53], [59], [78] and [94] followed by speaker clustering is known as speaker diarization [54]-[58] and [93].
The task of speaker diarization [6], [80], [81], [82], [83], [84] and [87] no prior information is available regarding the number of speakers involved or their identities. So, it can be considered as a task of identifying the number of speakers and creating a list of speech time intervals for each speaker. It precises, various speaker diarization algorithms [8], [87] that can be categorized into three categories: step by step approaches [61], integrated approaches, and mixed approaches.

Step by step approaches divide the speaker diarization task into number of steps. First finding the speaker change points using distance metrics, then growing the segments [9], [77], [79] during a hierarchical clustering phase and finally determining the number of speakers. In the case of integrated approaches all the steps involved by simultaneously. The mixed strategies are also proposed in, where classical step by step segmentation [7], [60], [62], [74], [85] and [86] and clustering are first applied and then refined using a re-segmentation process during which the segment boundaries, the segment clustering and sometimes the number of speakers are refined.

1.3 Challenging Issues

The present process has various challenging issues are to be solved.

1.3.1 Speaker Diarization

An audio document is segmented into speakers, depending on the environment or the nature of the document:
• to identify the speaker turns and the speaker clusters and to estimate the number of speakers involved in the document, without any priori information.
• to be able to process speech documents as well as documents containing music, silence, and other sounds.
• to be able to process spontaneous speech with overlapping voices of speakers, disfluencies, etc.

The main challenges in the present study are, no priori information is available on the test data. For instance, that the number and the identity of the speakers involved in the conversation are not known. A consequence of this hypothesis is that no reference speaker data are supposed to be available before segmenting an audio signal for instance.

There are three primary domains which have been used for speaker diarization [10], [12], [18] [63], [64], [65] [75] and [19] research and development: broadcast news audio [11] [75] and [76] recorded meetings, and telephone conversations [71], [72], [73]. The data from these domains differ in the quality of the recordings, the amount and types of non speech sources, the number of speakers, the durations and sequencing of speaker turns and the style/spontaneity of the speech. Each domain presents unique diarization challenges. For telephone conversations, only two people are involved and there are few overlapping segments. For broadcast news [48], there are obviously more speakers on the audio documents [49], [50], [51] and [52], but this is mostly prepared speech with a large part of “studio quality voice”. The hardest task definitely corresponds to meeting data with very spontaneous speech, overlapping voices, disfluencies, distant speakers (in case of table microphones) and background noise.
1.4 **Objective of the Thesis**

The main objective is to develop robust techniques for speaker diarization system with focuses on conversations domain. A set of more concrete objectives is defined:

- A diarization system was build to target on meetings domain, the algorithms and techniques examined should not be restricted to only conversations and could be applied to any other domains.

The performance of the resulting system should be competitive with focus on robustness. The main contribution of this research concerns the use of the distribution capturing ability of the AANN for speaker diarization. The proposed method relies on a classical two step based on a detection of speaker turns followed by a clustering process. This work formulates a new algorithm and, it works without any prior knowledge of the identity of speakers.

1.5 **Organization of the Thesis**

The focus of the research work presented in this thesis which is organized into seven chapters. After the introduction part in Chapter 1, Chapter 2 covers the literature review of speaker change detection [20], [21], speaker segmentation and speaker diarization.

Chapter 3 elaborates the acoustic feature extraction techniques.

Chapter 4 presents the mechanism of speaker change detection system.

Chapter 5 explains the Autoassociative neural network and Support vector machine models.

Chapter 6 discusses the results produced by the various proposed methods.

Finally, Chapter 7 devotes the conclusion and future work.