CHAPTER 4
SPEAKER CHANGE DETECTION

4.1 Introduction

Speech Activity Detection here after (SAD), identifies audio regions contains speech from any type of the speakers present in the audio recording. In the non-speech regions may contain silence, laughing, music, room noise, or background noise. The inclusion of non-speech [88] frames into the clustering process makes it difficult to correctly differentiate between two speaker models. These are broadly classified into four categories.

- Energy or spectrum-based speech detection.
- Model based speech or non-speech detection.
- Hybrid speech or non-speech detection.
- Multi-channel speech activity detection.

The energy-based SAD is mostly used for telephone audio since non-speech tends to be silence or slowly varying noise sources. For meeting audio, the non-speech can be from a variety of noise sources, like paper shuffling, coughing, laughing, etc. and energy-based methods do not work well for distant microphones. Due to the limitation of energy-based approach, in general, model based speech/non-speech detectors are frequently used in many speaker diarization systems as they are able to characterize various acoustic phenomena. The simplest
system uses just two models for speech and non-speech. A more complex system is described with four speech models that includes gender/bandwidth combinations. Noise and music are explicitly modeled and the systems comprise of five classes: speech, music, noise, speech + music, and speech + noise. The speech + music and speech + noise models are used to minimize the false rejection of speech occurring in the presence of music or noise, and this data is subsequently reclassified as speech. The classes can be broken down further, there are five models for non-speech (music, laughter, breath, lip-smack, and silence) and three for speech (vowels and nasals, fricatives, and obstruents). The model-based approach has its own limitation: its models need to be trained with pre-labeled data using training set. This requires the data to be annotated with class labels and this process takes much effort. Moreover, depending on the complexity of the models, there might not be enough data to build these models. The performance of these models on unseen data (which in statistical machine learning is known as the term generalization) is also an important issue especially in the case where testing data is substantially different from development data. To mitigate these problems, the hybrid approach is proposed. This approach comprises of two stages: the first stage is a simple energy-based detector, the second stage is a model-based detector in which the models are trained on the test data itself, hence no training data is required.

The derivative filter in combination with a finite state machine (FSM) to detect speech and non-speech regions. These initial labels are then used to build two SVMs for speech/non-speech; the system iteratively segments and trains both models until the overall likelihood stops increasing. The drawback of using energy however is that it is not possible to use this approach when the audio contains fragments with high energy levels that are non-speech. A more complex system is described, which are able to detect audible non-speech. In this system, speech, silence or non-speech sounds regions are first detected by pre-trained models on
conversation. Those regions with high confidence scores are then split to three classes: speech, non-speech with low energy, non-speech with high energy and high zero crossing rate. Three models are built up iteratively, and the audio is re-segmented a number of times.

In recent years, with the increasing availability of multi-channel audio, there have been a number of related efforts toward multi-speaker speech activity detection. Which performs a systematic analysis of features for classifying multichannel audio into four sub-classes: local channel speech, crosstalk speech, local channel and crosstalk speech and non-speech. They look at the frame-level classification accuracy for each class with the various features selected for analysis. A key result from this work is that, from among the twenty features examined, the single best performing feature for each class is one derived from cross-channel correlation. This fact shows evidence of the importance of cross-channel information for multi-channel detection task. This scheme is later used in a multi-speaker speech activity detection system that models vocal interaction between meeting participants with joint multi-participant models.

Speaker change can be considered as an event in the multispeaker speech data. We develop a classification model for the detection of speaker change events in continuous speech. The event is characterized by the end of speaking by the current speaker and the start of speaking by a different speaker. Therefore a speech data around a speaker change point includes the data of two speakers. The patterns extracted from the speech data around a speaker change point is considered as a positive example. The speech data between two consecutive speaker change points include the data one speaker only. Therefore the patterns extracted from the speech data between two consecutive speaker change points are considered as negative examples. The positive and negative examples can be used to train a classification model for detection of speaker change points [69].
The main issue in the classification based approach for speaker change detection is the duration of the speech signal to be considered for pattern extraction. Let \( t_{i-1}, t_i, \) and \( t_{i+1} \) be the \((i-1)^{th}, i^{th}\) and \((i+1)^{th}\) speaker change points in multispeaker conversation. The duration of the \(i^{th}\) speaker turn is given by \(d_i = t_{i+1} - t_{i}\). The information necessary for identifying \(t_i\) as a speaker change point is present in the speech data between \(t_{i-1}\) and \(t_{i+1}\) as this segment contains the data of two speakers. As the speaker turn durations vary, it is difficult to determine a suitable length of the window of the speech signal to be processed for detection of speaker turns with different durations.

A short window may not have enough data to capture the speaker change information of long speaker turns. A very long window may include the data of more than one speaker turn, and therefore may not be suitable for detection of short speaker turns. We study the effect of window size on the performance of the classification based approach to speaker change detection. The window of a chosen size includes a number of short-time analysis frames extracted from the speech signal in the window. Therefore the dimensions of the pattern vector derived for a window by concatenating the feature vector of frames in the window is very high, typically in the range of 150-800. We consider the Support Vector Machines (SVM) model for binary classification of large dimensional pattern vectors extracted from the windows of the speech signal in multispeaker speech data. For training the model, the positive examples are obtained by processing the fixed length window around the manually marked speaker change points. The negative examples are obtained by processing the fixed length windows of the signal between the manually marked speaker change points.
4.2 Sliding Window (SW) Method

The sliding window method is used for detection of speaker change points using the trained Support Vector Machine (SVM). In the sliding window method, a window of fixed length is processed to obtain a test pattern. Then the window is sided by a frame. The test patterns obtained using the sliding window methods are classified using the trained SVM to given the speaker change hypothesis.

The input to the speaker change detection system is a continuous speech signal of multispeaker speech data as in audio recording of a conversation or broadcast news. The multispeaker speech adapt typically consist of many silence regions due to the pauses while speaking. It is necessary to remove the pauses from the speech signal so that the fixed length windows around the speaker change points include the data of two speakers.

This research has practiced a Support Vector Machines (SVMs) for detection of silence in the continuous speech signal. The manually marked silence regions are processed to extract the positive examples of silence. The manually marked speech regions are processed to extract negative examples of silence.

4.3 SCD system

A pattern vector is obtained by concatenating three frames in a silence region or in a speech region. The sliding window method with a window width of three frames is used to detect the silence regions in the continuous speech signal using the SVM trained with the positive and negative examples of silence. The block diagram of the speaker change detection system is given in figure 4.1.
The continuous speech signal after the detection and removal of silence regions is given as input to the speaker change detection SVM.

For a chosen length of window of frames, the sliding window method [79] is used to derive the test patterns. The test patterns with the positive output of the SVM are hypothesized as the speaker change points. As the chosen window length is not suitable for different durations of speaker, turns several hypothesis of spurious. The work considers two methods for reducing the number of false alarms. In the first method, at threshold of five frames is used on the duration of speaker turns. When there is multiple hypotheses in a window of five frames, the hypotheses with the maximum output is retained and the other hypotheses are removed. Thus the SVM output is smoothed to eliminate the redundant hypothesis with very short speaker turn durations. For further reduction of the number of false alarms, the research evaluates the performance of the speaker change detection on validation data set. The false hypotheses for the negative examples in the validation
data set are identified. These false hypotheses are used as the negative examples in training a SVM for reducing the number of false alarms. The positive examples used in training the speaker changes detection SVM are also used as the positive examples for the false alarms reduction. The false alarm reduction is helpful in further discrimination of correct hypothesis and false hypotheses given by the speaker change detection SVM.