MULTIMODAL SPEAKER DIARIZATION

Researcher:

Rehan Ahmad

REG NO. 87-FET/PHDEE/S15

Supervisor:

Prof. Dr. Syed Zubair

Department of Electrical Engineering

Faculty of Engineering & Technology

INTERNATIONAL ISLAMIC UNIVERSITY,

ISLAMABAD

Multimodal Speaker Diarization

Rehan Ahmad

87-FET/PHDEE/S15

Submitted in partial fulfillment of the requirements for the PhD degree in Electronic Engineering at the Department of Electrical Engineering

Faculty of Engineering and Technology

International Islamic University,

Islamabad

Supervisor

Prof. Dr. Syed Zubair August, 2020

DEDICATED TO

My Teachers,

Parents,

Wife,

Brothers and Sister

CERTIFICATE OF APPROVAL

Title of Thesis: Multimodal Speaker Diarization

Name of Student: Rehan Ahmad

Registration No: 87-FET/PHDEE/S15

Accepted by the Department of Electrical Engineering, Faculty of Engineering and Technology, International Islamic University, Islamabad, in partial fulfillment of the requirements for the Doctor of Philosophy degree in Electronic Engineering.

Viva voce committee:

Dr. Suheel Abdullah Malik (Chairman) ____________________
Associate Professor
Department of Electrical Engineering
International Islamic University, Islamabad.

Prof. Dr. Muhammad Amir (Internal Examiner) ____________________
Dean, Faculty of Engineering & Technology
International Islamic University, Islamabad.

Prof. Dr. Ijaz Mansoor Qureshi (External Examiner - I) ____________________
Department of Electrical Engineering
Air University, Islamabad.

Dr. Aamer Saleem Choudhry (External Examiner - II) ____________________
Associate Professor
Hamdard Institute of Engineering & Technology
Hamdard University, Islamabad.

Dr. Syed Zubair (Supervisor) ____________________
Assistant Professor
Department of Electrical Engineering
International Islamic University, Islamabad

May 3, 2018

ABSTRACT

Speaker diarization system identifies the speaker using homogeneous regions in the given audio or audio-visual recordings. It answers the question ‘who spoke when?’. For this system, the data set comprises of multiple speaker recordings such as telephone conversation, broadcast news, meeting recordings etc. It is usually an unsupervised technique where no training data is available, and number of speakers is also unknown. This makes speaker diarization a real challenging problem. Several audio-based speaker diarization systems have been developed in the past which mostly comprises of agglomerative hierarchical clustering technique (AHC) which starts by assuming large number of clusters (speakers) and hierarchically merge them into the optimal number of speakers. In the past decade, researchers mostly have focused on the development of feature embedding techniques to make diarization more robust. So, the first work comprises of development of unsupervised feature embedding based on deep autoencoders.

Due to limitations in audio-based diarization techniques, several multimodal diarization techniques have been proposed that have utilized the speaker’s visual information such as face, head, lips and body movements to identify the active speaker. These multimodal techniques are usually complex and comprise complex audio and visual pipelines. A novel multimodal diarization technique has been proposed here, which utilizes a pre-trained audio-visual synchronization model to find active speakers. Both audio-visual pipelines in the proposed model are relatively simple and matches the unsupervised nature of speaker diarization system. Finally, extending the work of multimodal diarization, speech enhancement model is proposed to further optimize the system’s performance.

LIST OF PUBLICATIONS AND SUBMISSIONS

[1]. **R. Ahmad,** S. Zubair, H. Alquhayz and A. Ditta, “Multimodal Speaker Diarization Using a Pre-Trained Audio-Visual Synchronization Model”, in Sensors, vol. 19, pp. 5163, 2019, doi: 10.3390/s19235163. **(IF=3.031)**

[2]. **R. Ahmad** and S. Zubair, “Unsupervised deep feature embeddings for speaker diarization”, in Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, pp. 3138-3149, 2019, doi: 10.3906/elk-1901-125 **(IF=0.625)**

[3]. **R. Ahmad**, S. Zubair and H. Alquhayz, "Speech Enhancement for Multimodal Speaker Diarization System," in *IEEE Access*, vol. 8, pp. 126671-126680, 2020, doi: 10.1109/ACCESS.2020.3007312. **(IF=4.640)**

ACKNOWLEDGEMENTS

In the name of Allah (Subhanahu Wa Ta'ala), who is the most Gracious and the Merciful. I would like to thank Allah for giving me strength and patience to complete this research work. Peace and blessings of Allah be upon His last Prophet Muhammad (Sallulah-o-Alaihihe-Wassalam) and all his Sahaba (Razi-Allah-o-Anhu) who dedicated their lives for Dawah and spread of Knowledge.

I am truly grateful to my supervisor Dr. Syed Zubair whose inspiration, ideas and efforts make it possible for me to complete my higher studies. He has been a role model for me and many others in teaching, research and other aspects of life. I would also like to thank my mentor Dr. Waqar Qasim who always supported me to learn and spread the knowledge. His supplications gave me lot of strength.

I offer my sincere thanks to my colleagues Dr. Zeshan Aslam Khan, Dr. Naveed Ishtiaq Chaudhry, Engr. Athar Waseem, Engr. Khizer Mehmood, Engr. Baber Khan Jadoon, Engr. Sharjeel Abid Butt for their never-ending support and fruitful and healthy research discussions. I would like to acknowledge the support of International Islamic University Islamabad, Pakistan for providing me full fee waiver during the PhD studies. I am thankful to administration at department, as well as university level for their kind support.

I am really grateful to my father, mother, brothers and sister for their love and support throughout my life. I am also very thankful to my wife for her patience, encouragement, and prayers during every single stage of my PhD degree.

(Rehan Ahmad)

Table of Contents

LIST OF FIGURES

LIST OF ABBREVIATIONS

HMM	Hidden Markov Model
GMM	Gaussian Mixture Model
AHC	Agglomerative Hierarchical Clustering
BIC	Bayesian Information Criteria
SDM	Single Distant Microphone
MDM	Multiple Distant Microphone
SSL	Speech Source localization
SAD	Speech Activity Detection
MFCC	Mel-frequency Cepstral Coefficients
KL	Kullback-Liebler
GLR	Generalized Likelihood Ration
SNR	Signal to Noise ration
FFT	Fast Fourier Transform
DCT	Discrete Cosine Transform
AMI	Augmented Multi-party Interaction
LSTM	Long Short-term Memory
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
RCNN	Recurrent Convolutional Neural Network
SD	Speaker Diarization
ASR	Automatic Speech Recognition
LPS	Log-power Spectrum
IRM	Ideal Ratio Mask

EM	Expectation Maximization
MMSE	Minimum Mean Square Error
MTL	Multi-task Learning
BPTT	Back Propagation Through Time
SVM	Support Vector Machine
PLP	Perceptual Linear Prediction
DER	Diarization Error Rate
DNN	Deep Neural Network
DAE	Denoising Autoencoder
MMSD	Multimodal Speaker Diarization
CCA	Canonical Correlation Analysis
MI	Mutual Information
UIS-RNN	Unbounded Interleaved-state Recurrent Neural Network
ddCRP	Distance-dependent Chinese Restaurant Process
NIN	Network-in-Network
NIST	National Institute of Standards and Technology

LIST OF SYMBOLS

A list of commonly used symbols in this dissertation are given below.

W	Weight matrix
$\boldsymbol{x}$	Input data vector
b	Bias
$\boldsymbol{h}$	Hidden layer output
D	Data
L	Number of model parameters
θ	GMM model
N	Total number of speech frames
E	loss function
$\boldsymbol{\mu}$	Mean vector
Σ	Covariance matrix
r_{ij}	GMM posterior probability
P	Number of pixels
F	Video frame rate
T_{of}	Offset threshold
T_{conf}	Confidence threshold

Chapter 1.

Introduction

This chapter provides an overview of unimodal and multimodal speaker diarization systems. Initially, a background and motivation for this topic is provided. A brief description for the need of multimodal technique is also described. Finally, brief thesis contributions and overview of the thesis is provided.

1.1 Background and Motivation

Speaker diarization is the process of partitioning an audio recording into speakers' homogeneous regions. It answers the question "who spoke when?" in a multi-speaker recording environment. It is usually an unsupervised problem where the number of speakers and speaker-turn regions are unknown. The diarization process automatically determines the speaker-specific segments and group similar ones to form a speaker-specific diary. Its application lies in multimedia information retrieval, speech and speaker recognition and audio processing. Used cases of diarization include analysis of the speakers, speech and their transcription in meeting recordings, TV/talk shows, movies, phone conversations, broadcast news, conferences etc.

Significant effort in speaker diarization techniques started from International competitions by Rich evaluations which was sponsored by National Institute of Standards and Technology (NIST). Early work in speaker diarization started with the telephone data and broadcast news. In late 1990’s and early 2000’s the main aim has been to automatically annotate the TV and radio broadcast transmissions. Later, interest in the meeting recordings has been started from 2002, with several research projects. Some of these projects includes Augmented Multi-party Interaction (AMI), Multimodal Meeting Manager (M4), Swiss Interactive Multimodal Information Management (IM2) etc. Series of diarization systems have been developed on NIST database after year 2002 that includes [1]–[6].

Meeting recordings offer challenging scenarios which eventually affect the diarization process. These is environmental noise, reverberation, spontaneous speech and overlapping speech regions. Similarly, the setup of audio recording equipment, such as single distant microphone (SDM), multi-distant microphone (MDM), lapel and microphone array may also vary. In case of visual recordings, the camera setup may capture individual speaker or group of speakers. All these setups motivate one to develop optimal diarization approaches. For example, in case of microphone array one can use the speech source localization (SSL) technique to find the active speaker. Similarly, in the presence of individual camera setup which captures the face of individual speaker, users usually find motion features to determine the audio-visual synchronization to detect active speaker. Diarization is usually developed base on the give environmental and recording scenarios. Diarization is usually applied in audio domain but the presence of video modality helps to improve the performance of the system. This thesis mainly focuses on the use of multimodal (audio-visual) information to apply the diarization.

1.2 Research Problem Statement

Speaker diarization is usually an unsupervised technique where limited information is available about speakers, their count and the scenario. This becomes more challenging when only audio data is available because it is heavily affected by environmental noise, reverberation and short utterances. Due to advancement in the technology, audio-visual recordings of meetings, broadcast news, TV/ talk shows are available. This has leveraged one to use the visual modality to improve the speaker diarization system.

Number of unsupervised audio-based speaker diarization systems have been developed in the past mostly based on the agglomerative hierarchical clustering technique. Such clustering techniques are usually modeled by HMM/GMM. Presence of video modality gives the advantage of finding the audio-visual relevance by detecting and tracking faces, lip movement etc. Some recent studies in multimodal speaker diarization have applied diarization on both modalities individually and finally fuse their outputs. Similarly, some semi-supervised audio-visual fusion techniques for diarization have also been proposed. This study investigates the use of multimodal data to find the synchronization between the audio and visual modalities to find the active speaker. Such technique helps to get pure clusters which only comprise of single speaker speech frames. Furthermore, the use of feature embedding techniques and speech enhancement have also been investigated to improve the diarization.

1.3 Research Objectives

Objectives of this research is to improve the available speaker diarization systems using multimodal (audio-visual) techniques. This technique would eventually improve its application in areas such as speech recognition, information retrieval, multimodal analysis etc. The general objectives in this research are related to feature embedding, use of audio-visual synchronization model and speech enhancement.

1.4 Research Philosophy

Unimodal diarization techniques are very challenging due to availability of audio only data and different environmental scenarios. While the presence of video modality provides complementary information. By using video modality along with the audio, one can improve the diarization techniques. Similarly, feature extraction and audio denoising also affect the diarization process. Improving any pipeline of the existing diarization system or adding complementary information would eventually improve the diarization process.

1.5 Hypothesis

Formulated research hypothesis of this study is detailed as follows:

- The addition of feature embedding method in diarization pipeline may result in the improvement of the existing technique.
- Use of multimodal technique comprising of audio and video modalities may increase the accuracy of the system. The video modality could help the audio diarization by providing active speaker information.
- Certain environmental factors such as noise etc., degrades the performance of the diarization. So, the use of speech enhancement may eventually improve the performance of the system.

1.6 Research Contributions

Following sections individually describe the research contributions in detail. These contributions comprise of feature embeddings based on deep autoencoders, multimodal speaker diarization system with a pre-trained audio-visual synchronization model and speech enhancement for multimodal speaker diarization.

1.6.1 Feature Embedding

First contribution in this thesis is the unsupervised deep feature embedding technique. This technique is based on deep autoencoders which is trained on the given input data recording. The proposed architecture is trained in a specific way to acquire the compressed domain embedding from the encoder's output. Such technique doesn't require any data other than the available meeting recording. Such embeddings are tested on the popular subset of AMI [7] meeting corpus.

1.6.2 Multimodal speaker diarization

Second contribution to this thesis is novel multimodal speaker diarization system which is based on a pre-trained audio-visual synchronization model. This available model was trained on the large audio and visual streams to find the synchronization between the visible speaker and its respective speech. We proposed a novel audio-visual processing pipeline which utilizes this pre-trained model to find the active speaker and train the speaker specific clusters. Such simple yet effective model performed better than the state-of-the-art audio and multimodal speaker diarization systems.

1.6.3 Speech enhancement for multimodal speaker diarization

Final contribution in this thesis is the analysis of speech enhancement in the proposed multimodal diarization technique. Recently, speech enhancement (denoising) is studied for speech recognition and audio-based speaker diarization system which has provided significant results. Based on these studies we have incorporated the speech enhancement module in multimodal speaker diarization system. This speech enhancement module is based on the LSTM network which is trained on large noisy speech corpus comprising more than 100 noise types. In the proposed multimodal system, speech enhancement improves both audio and video pipelines, which effectively improves the diarization error rate of the system.

1.7 Thesis outlines

The rest of the chapters are organized as follows. Chapter 2 initially describes the general background of speaker diarization system then it provides the detailed literature review of each contribution. Chapter 3 initially describes the dataset, evaluation metrics and finally materials & methods for three main research contributions: feature embedding, multimodal diarization technique and speech enhancement for multimodal speaker diarization. Chapter 4 presents the detailed results and its discussion. Finally, conclusion and future work is presented in Chapter 5.

Chapter 2.

Literature Review

This chapter initially describes the general background of diarization that consists of unimodal and multimodal diarization techniques. Further, it describes the necessary background of the three main contributions for speaker diarization. These contributions consist of feature embedding, multimodal diarization based on an audio-visual synchronization model and speech enhancement.

2.1 General Background of Speaker Diarization

Speaker diarization is the task of segmenting the digital recording in speaker homogenous regions. The output of diarization is useful in automatic speech recognition (ASR), automatic transcription, information retrieval and multimodal analysis. Speaker diarization helps in providing the speaker-specific data. In automatic transcription task, the output of diarization is more easily readable to the humans and also useful to the machines for natural language processing tasks.

There are mainly three lines of research approaches for speaker diarization. The first approach applies the speaker diarization only using audio stream. This is the most common and widely used in the research. The main reason is that the concept of diarization is applicable on the audio, so that each user’s speech could be separated. However, audio only diarization get very challenging with environmental factors and overlapping speech intervals. The second approach performs speaker diarization using synchrony detection. In this approach synchronization between video and audio modalities are determined to apply diarization e.g. synchronization between lip movement and speech is determined. Third approach processes the video and audio modalities separately, fuse both modalities at feature level or the output level. Such technique also utilizes audio-visual tracking tasks for active speakers. Figure 2.1 shows the three lines of inputs for speaker diarization system.

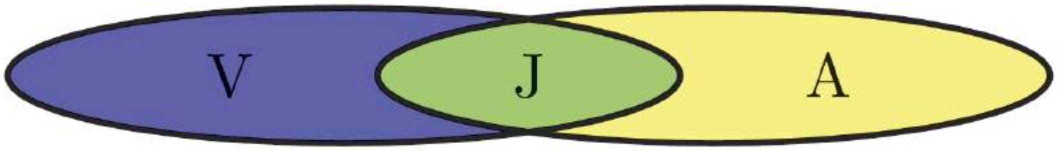


Figure 2. 1 Three lines of inputs: Audio (A), Video (V) and Synchrony (J) for speaker diarization.

In the following sections, first audio based speaker diarization and its necessary components are presented then multimodal speaker diarization system is described.

2.1.1 Audio based speaker diarization

A typical audio based diarization system consists of three main steps. The first one is a preprocessing step that consists of feature extraction such as Mel-frequency cepstral coefficients (MFCCs). Second is speech activity detection (SAD), which removes the silence and nonspeech regions from the speech. Finally clustering and segmentation step, which works iteratively to segment the speaker-change regions and collect the homogeneous segments to make speaker-specific clusters. Audio based speaker diarization systems are usually developed on the basis of one of the two approaches: the bottom-up and the top-down clustering, depicted in Figure 2.2 [8]. The top-down approach usually starts with a single cluster and iteratively increases it to converge into the optimal number of clusters. Similarly, bottom-up approaches start with a large number of clusters and iteratively merge them until they converge into the optimal number of clusters. In both techniques, each optimal cluster represents an individual speaker in the recording. Both techniques are generally based on the Hidden Markov Model (HMM) where each cluster/state is modeled by the Gaussian Mixture Model (GMM). Bottom-up approaches are usually most common and best choice among the researchers, also known as agglomerative hierarchical clustering (AHC). Typically, agglomerative hierarchical clustering (AHC) [9] technique is initialized with a large number of clusters and iteratively merge similar clusters on the basis of the threshold. This threshold is based on one of these metrics: Bayesian information criterion (BIC) [10], Kullback-Liebler (KL) [11] and Generalized Likelihood Ration (GLR) [12]. Among these metrics, Bayesian information criteria is most commonly used in speaker diarization techniques. A typical audio-based

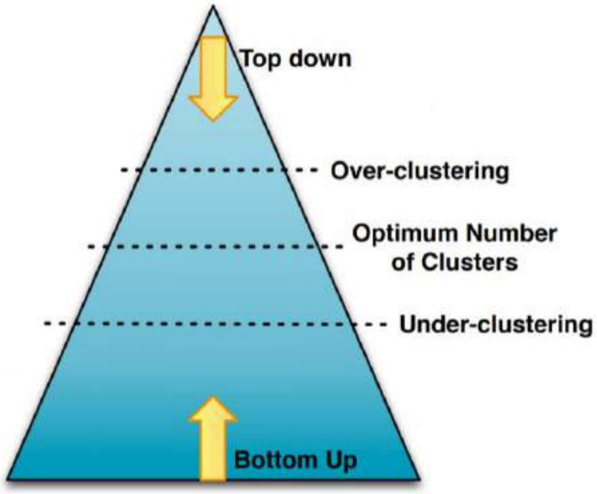


Figure 2. 2 Bottom-up and top-down clustering approaches.

speaker diarization system is presented in Figure 2.3. Beside AHC, some fully supervised techniques have also been proposed.

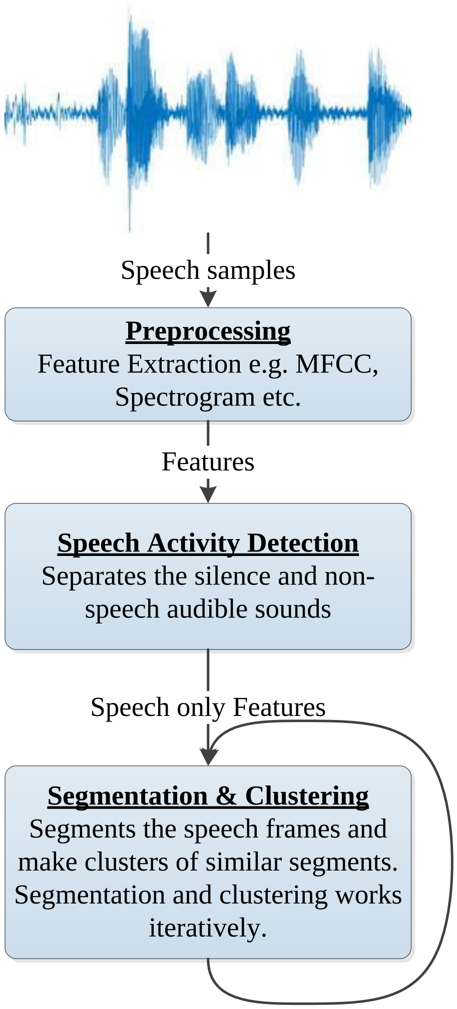

Speech samples
Preprocessing
Feature Extraction e.g. MFCC, Spectrogram etc.
Features
Speech Activity Detection
Separates the silence and non-speech audible sounds
Speech only Features
Segmentation & Clustering
Segments the speech frames and make clusters of similar segments. Segmentation and clustering works iteratively.

Figure 2. 3 Audio based Speaker diarization system.

2.1.2 Feature extraction

The first preprocessing step in any speech processing application is the feature extraction. Mel-frequency cepstral coefficients [13] (MFCC) have been most widely used features for speech processing applications, for example, speech recognition, speaker recognition and speaker diarization. They were introduced in 1980’s by Davis and

Mermelstein. It is known to be the best perception-based features which accurately represents the envelop of the signal. One of the properties of MFCC is that, acquired number of features are uncorrelated. Figure 2.4 shows the feature extraction steps. The process of MFCC extraction consists of following steps:

- Take short frames of the signal spanning over 20ms-30ms with about 50% overlap.
- Compute the power spectrogram of each frame using fast Fourier transform (FFT).
- Apply the Mel-filterbanks on to the power spectrogram and sum their energies.
- Take the logarithm of these energies.
- Apply Discrete Cosine transform (DCT) on these log filterbank energies.
- Select number of coefficients ranging between13-39.

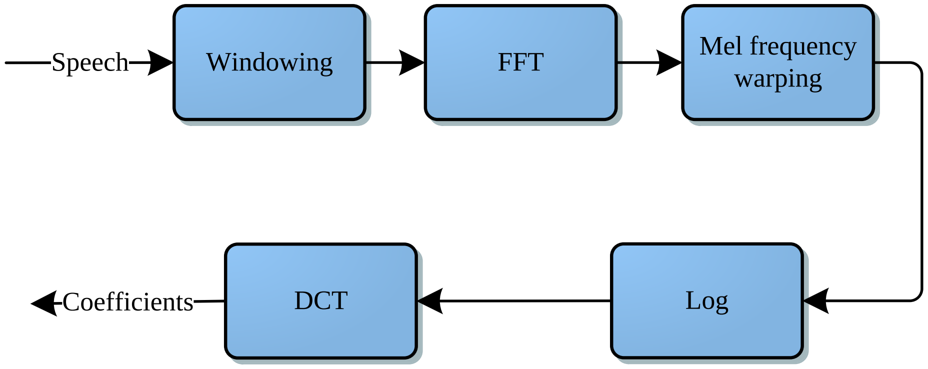


Figure 2. 4 MFCC feature extraction steps.

2.1.3 Speech Activity Detection

Speech activity detection (SAD) is the fundamental and essential part of any speech processing application. It classifies the input speech frames into speech and non-speech segments. The diarization process is significantly affect by the SAD performance. It contributes towards two main error metrics of diarization i.e. false alarm and missed speech. Due to the presence of silence and environmental noise the best performing SAD should filter out both types of segments from the recording. Indeed, the inclusion of noise and silence eventually provides the less discriminant clusters, leading towards speaker confusion in diarization.

SAD is usually applied as a preprocessing step in speaker diarization. Initial approaches for SAD were based on unsupervised technique. However, the presence of the various types of environmental noise factors and room conditions with varying signal-to-noise ratios (SNR) made it significantly challenging problem. A model is usually trained with large speech, silence and noise corpus to apply SAD. Alternatively, in such formal recordings where there is no environmental noise one can go for a simple classifier, which classifies high and low energy frames, representing speech and silence. For developing speech activity detector, MFCC is usually the best choice of features.

2.1.4 Segmentation and Clustering

In most of the speaker diarization systems, segmentation and clustering works together for hierarchical segmentation of speech signals and cluster the ones with. Segmentation usually finds the speaker change regions and clustering process creates and merge the similar clusters. In agglomerative hierarchical clustering, initially large numbers of clusters are initialized, and then speech is randomly divided into these clusters. In the next step, likelihood of each speech frame is computed and the whole speech is re-segmented. Clusters are then trained on the re-segmented speech. Similar clusters are then merged based on the Bayesian formation criteria (BIC) or Kulback-Liebler (KL) divergence. The re-segmentation and clustering process work iteratively until no two clusters remain similar. The resulting clusters represent each individual speaker and their respective speech.

2.1.5 Multimodal speaker diarization

With the availability of audio-visual recordings such as multiparty interactions, meetings, movies, news broadcast, TV shows etc., research in speaker diarization tilted towards multimodal dataset. Diarization task is very challenging when unimodal data is available. Audio-based diarization has lot of complexities due to overlapping speech utterances from various speakers, environmental noise, short utterances and reverberations. To solve the shortcomings and limitations in audio-based diarization approaches, multimodal approaches comprising audio and visual modalities were proposed. Multimodal diarization comprises of using audio and video modalities, jointly solve the problem of active speaker detection and track the same speaker temporally. Such approach usually apply active speaker detection using motion features of lip and face movements [14]–[17]. Furthermore, some audio-visual fusion techniques [16], [18], [19] at the feature level or output level may also be applied. Some multimodal diarization techniques apply diarization on individual modalities and obtain the required outputs on the basis of decision algorithm or by weighted fusion.

The audio and visual modalities provide complementary information having correlations [20] between them, so they are more likely to be robust as compared to audio specific or video specific diarization techniques. The scenario of available recordings varies depending on the recording mediums (cameras and microphones), participant speech turns, silence between speech utterances of different speakers, short speech utterances, overlapping speech and environmental noise. Above discussed visual processing techniques for diarization are less affected by the environmental noise, reverberation etc. as compared to the audio-based approaches.

2.1.6 Evaluation metric

Diarization error rate is usually an evaluation metric for speaker diarization systems. It comprises of four types of errors; missed speech, false alarm, speaker error and overlapping speech error. Among these errors, two are directly addressed by a preprocessing speech activity detection block i.e. missed speech and false alarm. Missed speech is fraction of time when speech is assigned as non-speech and false alarm is the fraction of time when any environmental noise is assigned as a speech. Similarly, speaker error arises when wrong speaker is assigned and overlapping speech error arises when multiple speakers are not assigned correctly.

2.2 Feature Embedding

Feature embedding converts data into a type of feature representation with certain properties, such as it has discriminative features with data samples of different classes. Number of feature embeddings techniques have been proposed for speech recognition, speaker verification, identification and diarization. A noticeable improvement in the performance of diarization systems was achieved using i-vectors [21]–[26]. It models the overall variability of speakers' voices and compress the information into low-dimensional subspace. With the increase in the use of neural networks and deep learning techniques, i-vector based methods were outperformed by d-vector feature embeddings learned by neural networks [8]. Similarly, our propose feature embedding technique is based on a deep neural network architecture.

2.2.1 Related Work

Recent developments in deep learning techniques have made it more convenient to develop models that extract new sets of features either directly from raw datasets [27] or from hand-crafted features [28]–[32]. Recent work in speaker diarization consists of developing methods for the extraction of special features that help in speaker discrimination more robustly. For that purpose, many researchers have evaluated the feature embeddings technique using deep learning models such as long short-term memory (LSTM), deep neural networks (DNNs), and recurrent convolutional neural networks. In [33], Wang et al. proposed the extraction of d-vectors [34] from the LSTM model for speaker diarization purposes. The model uses log-mel-filterbank frames as an input and uses the output frames of the LSTM architecture as a d-vector. Figure 2.5 shows the d-vector extraction process. It then applies diarization based on a spectral clustering algorithm. However, they applied diarization on speech frames that do not contain overlapping speech regions, simplifying the diarization process. Moreover, LSTM was trained for the speaker verification task and the trained network was used to extract feature embeddings.

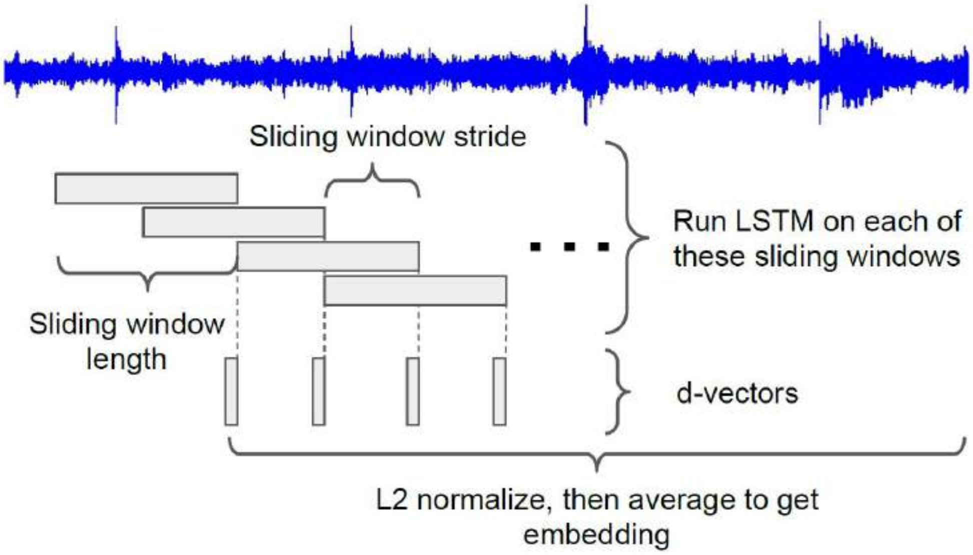

Sliding window stride
Run LSTM on each of these sliding windows
Sliding window length
d-vectors
L2 normalize, then average to get embedding

Figure 2. 5 d-vectors extraction using sliding window and LSTM.

Recently a fully supervised speaker diarization system was proposed by Zhang et al. [35] that utilizes an unbounded interleaved-state recurrent neural network (UIS-RNN) for diarization. The proposed system extracts speaker embedding (d-vectors) from the LSTM model and each speaker is modeled by a parameter-sharing RNN. The RNN model is further integrated with a distance-dependent Chinese restaurant process (ddCRP) to find the number of speakers in an audio recording. UIS-RNN model is presented in Figure 2.6, which shows that the generative process either switches to the same speaker, existing speaker or a new speaker.

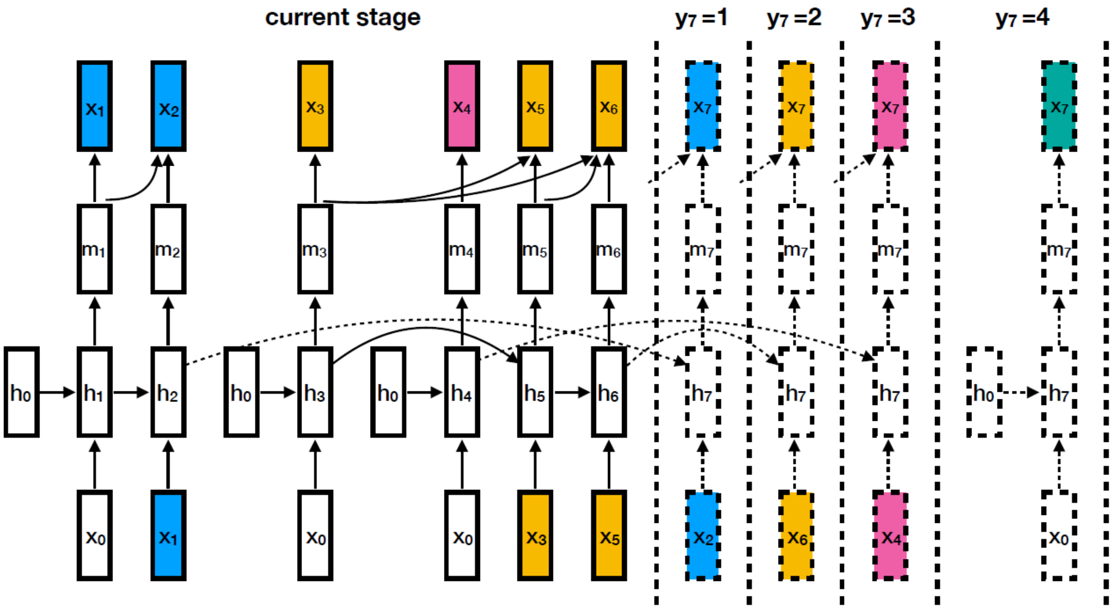
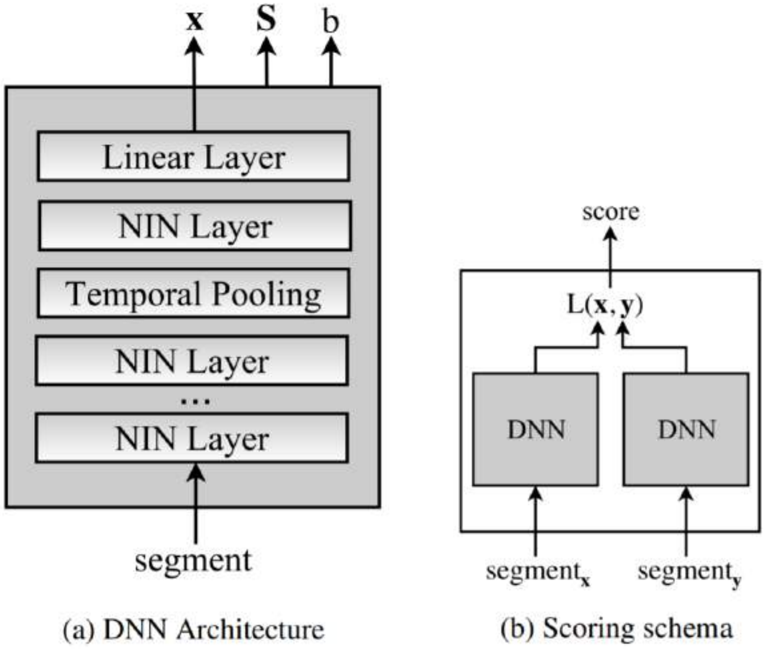


Figure 2. 6 UIS-RNN generative process. Each color indicates label for speaker segments. Each dotted box is next possible generated speaker segment.

Cyrta [36] proposed a recurrent convolutional neural network (RCNN) to extract feature embedding from magnitude spectrograms rather than from MFCC features. The RCNN based architecture was also trained for speaker classification tasks. Similarly, Romero et al. [24] proposed a DNN based feature embedding technique. They replaced the conventional i-vector based features with newly learned embedding from the DNN architecture. The DNN architecture was specially designed based on network-in-network architecture (NIN) [37]. It was trained to jointly learn the discriminative embeddings and a scoring metric to measure the likelihood of segments generated from the same or different speakers. To train the architecture, data was prepared by making pairs of same and different speakers. Figure 2.7 shows the NIN based DNN architecture and scoring method. Rouvier et al. [38] also proposed feature embedding taken from the hidden layers of a DNN. DNN architecture was trained to recognize speakers among a sample of 1000 from training set.

This trained architecture was then used to extract the new features. The i-vector based features were replaced by newly learned features for diarization. Sell et al. [39] described some experiences and lessons learned from the DIHARD diarization challenge. They described several key aspects of state-of-the-art diarization methods, such as feature extraction, feature embeddings (i-vector vs. x-vector), speech activity detection, and training data. Furthermore, the authors described their effective diarization system with wideband data, variational-Bayesian refinement, and single x-vector.



(a) DNN Architecture

(b) Scoring schema

Figure 2. 7 DNN architecture based on Network-in-Network (NIN) and scoring method.

The above discussion shows that neural network based feature embedding, known as d-vector embeddings, have improved speaker diarization performance as compared to i-vector based features. However, d-vector based embedding were extracted from networks trained on speaker verification or classification tasks in supervised settings. Moreover, the network was trained on large datasets, which limits their use for relatively smaller datasets.

Our study proposes an unsupervised feature learning from a deep learning architecture, which closely resembles the original unsupervised speaker diarization pipeline.

2.2.2 Proposed work

Building upon the success of deep learning architectures, our work proposes feature embedding based on autoencoders which is then followed by hierarchical clustering for speaker diarization. In contrast to other deep learning-based approaches, this method is unsupervised and directly matches the unsupervised nature of the speaker diarization system.

2.3 Multimodal Diarization

Diarization tasks are very challenging when unimodal data is available. Audio-based diarization has a lot of complexities due to overlapping speech utterances from various speakers, environmental noise, short utterances and reverberations. Similarly, in video data, speakers may not face the camera, move in a multi-party interaction way or they can be occluded by other speakers. The use of the video modality facilitates one to using lip and face movement detection for diarization. For each available dataset, the configuration of the recording equipment also varies a lot. For example, audio data may be acquired from a far-field microphone array, individual lapel microphones/headsets or single omnidirectional microphones. Similarly, video recordings comprise of individual speaker closeup cameras, cameras covering some group of speakers or a wide camera covering all the available speakers in an event.

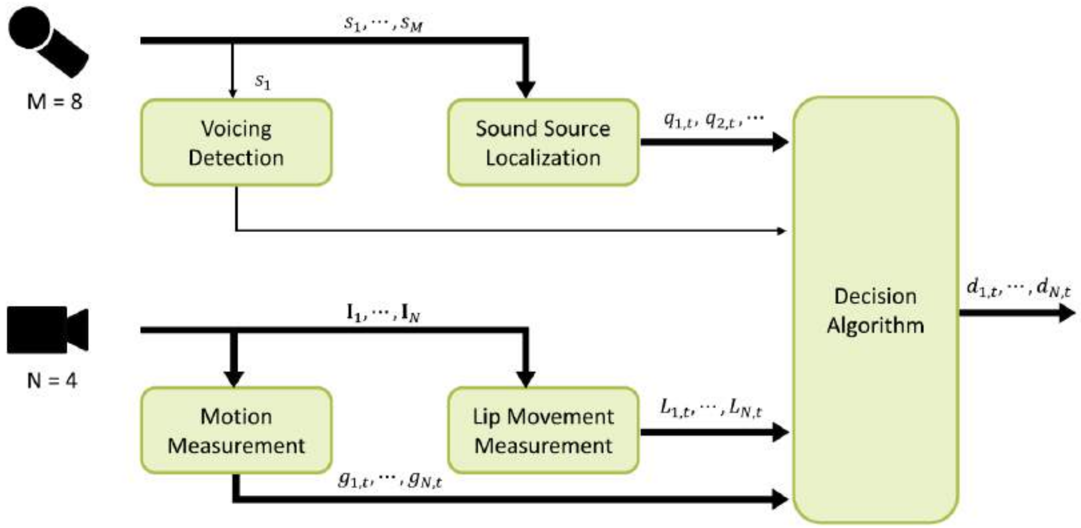
To solve the shortcomings and limitations in audio-based diarization approaches, multimodal approaches comprised of audio and visual modalities were proposed.

Multimodal approaches either use active speaker detection using lip or face movements [14]–[17], [40]–[42] or some audio-visual fusion technique [16], [18], [19] at the feature or output stage after applying diarization on individual modalities. The audio and visual modalities provide complementary information, so they are more likely to be robust as compared to audio-only or video-only diarization techniques. In the last decade, several multimodal diarization techniques have been proposed, e.g., [15], [17], [19], [43]–[47]. The scenario of available recordings varies depending on participant speech turns, silence between speech utterances of different speakers, short speech utterances, overlapping speech and environmental noise. Moreover, the participants may be seated/static or move around.

2.3.1 Related Work

Use of the video modality in speaker diarization is motivated by the fact that audio and video have correlated factors. For example, the lip, face and head movement of an active speaker are highly correlated with his speech. Hence, features extracted from frontal views of speaker faces can be used to discriminate the active speaker. Such visual features are used in speech recognition [48], [49], speech source separation [50], [51] and speaker diarization [52]–[55].

Friedland et al. [56] proposed the use of compressed domain video features for multimodal speaker diarization comprising frame-based visual activity features. These features were computed as a motion vector magnitude. Multimodal fusion was applied for MFCC and video features by a weighted likelihood of Gaussian mixture model. An agglomerative hierarchical clustering technique was used where each cluster was modelled by joint audio and video GMM. In contrast to its simplicity and less computational complexity, this technique might not work in scenarios where speakers move from their position or silent speakers shake their heads while listening. In [57], Garau et al. provided a comparison of two audio-visual synchronization methods. These two methods consist of canonical correlation analysis (CCA) and mutual information (MI) which uses MFCC features along with motion features from face tracks. The MI performed slightly better than CCA. Moreover, it was concluded that lip and chin vertical-movement visual features correlate the most with speech. Similarly, mutual information, which combines acoustic energy and gray-scale pixel’s value variation, was also used by Noulas et al. [43]. A dynamic Bayesian network was used to jointly model the audio and visual features for speaker diarization. Experiments were conducted on meeting recordings consisting of four speakers who face the camera and broadcast news with five people, where only three of them speak. Later, El Khoury et al. [19] proposed audiovisual diarization of people, where individual audio and visual clustering is carried out and fused together using co-occurrence matrices. The audio pipeline consists of MFCC feature extraction followed by SAD and finally segmentation and clustering. Similarly, in the video domain initially shot detection is applied then face detection, people tracking, people clustering and finally face clustering. Audiovisual diarization finally combines both clusters using an association technique. Minotto et al. [15] solved speaker diarization problems through speech source localization (SSL) in the audio domain and face detection and tracking in the video domain. A final decision is made using a supervised support vector machine (SVM) classifier. SSL provides lot of advantage in the analysis because recordings of two or three speakers consist of lots of overlapping speech segments. In [17] Sarafianos et al. applied audio-visual diarization using Fisher linear semi-discriminant analysis. After individual audio and video diarization, audio-visual fusion is applied. Kapsouras et al. [44] proposed to cluster face features and audio features independently and then correlate them based on temporal alignment. The most recent works in diarization [45], [46] mainly focus on the use of the sound source localization (SSL) technique to find active speakers. This technique helps to robustly identify speech overlap regions. Cabañas-Molero et al. [45] proposed to use SSL in the audio domain and motion measurements along with lip movement in the video domain. Both domains are fused together via a decision algorithm. The localization algorithm is evaluated on space volume rather than a discrete point in the space. Figure 2.8 shows block diagram of their multimodal diarization approach.



M = 8
$s_1, \cdots, s_M$
s_1
Voicing Detection
Sound Source Localization
$q_{1,t}, q_{2,t}, \cdots$
Decision Algorithm
$d_{1,t}, \cdots, d_{N,t}$
N = 4
$\mathbf{I}_1, \cdots, \mathbf{I}_N$
Motion Measurement
Lip Movement Measurement
$L_{1,t}, \cdots, L_{N,t}$
$g_{1,t}, \cdots, g_{N,t}$

Figure 2. 8 Multimodal speaker diarization approach with decision algorithm.

Similarly, Gebru et al. [46] proposed multimodal speaker diarization based on spatiotemporal Bayesian fusion, where a supervised localization technique is used to map audio features onto the image. This is achieved by sound source localization in the audio and multiple person visual tracking in the video which are fused via a supervised technique.

In most of the techniques discussed above, either a source localization technique is applied in audio pipeline to locate the active speaker or audio clustering/diarization is applied separately. Similarly, in the video domain face tracking, mouth/lip movement, motion measuring techniques are applied to get diarization results. Finally, audiovisual fusion is applied on the feature level or output level. Both audio and video pipelines require excessive processing to acquire the individual and fusion results. Comparatively, our proposed technique is simple and relies more on a pre-trained SyncNet model to find active speakers. A simple preprocessing pipelines in the audio and video domain is required, which finally ends up in audio-based clustering to acquire diarization. This technique is well suited for formal meeting scenarios where people are static/seated and frontal faces are captured most of the time.

2.3.2 Proposed work

In recently published multimodal diarization approaches, e.g., [45], [46], the focus of the authors is to track active speakers based on speech source localization (SSL). In such approaches, SSL along with video domain processing makes the diarization process computationally intensive. Our proposed technique is comparatively simple and heavily inspired by the work of automatic lip syncing in the wild [58]. In this technique, the author trained an audio-visual convolutional neural network to learn speech and mouth synchronization. The trained model is applicable to determine lip synchronization errors, active speaker detection and lip reading. This study investigates the pre-trained model referred to as SyncNet, to find active speakers in short video segments of closeup camera streams. The focus of this work is to robustly identify active speakers using the pre-trained SyncNet model. Introduced diarization approach comprises of acquiring features in the audio domain and applying GMM-based clustering on those audio frames which were robustly identified as corresponding to an active speaker. The simplicity of this approach is also reflected in the video domain, where face detection is applied, convert the results into short video segments and feed them to an audio-visual pre-trained model to apply inference. Compared to the audio-based diarization techniques consisting of conventional and fully supervised diarization, our results are very significant and proves the validity of such a novel approach. Compared to the complex multimodal technique, proposed approach provides nearly similar results.

2.4 Speech Enhancement

A real application of speaker diarization should address the environmental robustness problems i.e. environmental noise, overlapping speech and reverberations. These three factors significantly affect the performance of speaker diarization system. However, few studies have focused on these issues. Some traditional approaches for speaker diarization proposed to apply speech enhancement module based on Wiener filtering [8]. In recent diarization approaches, Weiner filtering is usually not considered as a suitable choice due to some of its limitations e.g. non-stationary noise tracking problem. Moreover, the resulting enhanced speech also suffers from some artifacts in speech i.e. musical noise [59]. Such noise also degrades the performance of speaker diarization. The emergence of deep learning technique enabled researcher to put great effort in speech enhancement. Most recently, LSTM based speech enhancement is applied as a preprocessing step in audio speaker diarization systems in [60], [61].

This study is an extension of our proposed multimodal speaker diarization system which uses a pre-trained audio-visual synchronization model to find the active speaker. In this study we propose to use LSTM based speech enhancement as a preprocessing step on highly noisy audio recordings. The noisy data set degrades the performance of the output of both audio and visual diarization pipelines. In video pipeline, noisy recordings affect the acquisition of high confidence active speaker segments and in audio pipeline it affects clustering. By applying LSTM based speech enhancement on noisy recordings, it is noticed that enhanced speech improves the performance of such multimodal speaker diarization system.

2.4.1 Related work

Audio based speaker diarization system is highly degraded with the presence of different environmental factors such as noise and reverberation. Recently, deep learning based speech enhancement models were proposed to remove the noise from speech signal. In [62] Narayanan et al. proposed to use ideal ration masks (IRMs) for time-frequency units classification. This mask was used to remove noise from Mel spectrogram before cepstral feature extraction for automatic speech recognition (ASR) system. A significant improvement is noticed for ASR system in term of word error rate. Similarly, in [63] Narayanan et al. used similar time-frequency masking based denoising for speech separation and automatic speech recognition. Their speech separation comprises of two stages. The first one removes additive noise and second one applies a non-linear function to map spectral features to the clean speech. Lu et al. [64] proposed to use deep denoising autoencoder (DAE) for speech enhancement. DAE model was trained on noisy-clean training pairs. This model is further used for filtering out the noise given the noisy speech. In [65], authors formulated speech separation as a binary classification problem using support vector machines (SVM). Furthermore, for discriminative feature extraction pre-trained deep neural network model is used. They presented good results for unseen speakers and background noise. Xu et al. [66] presented regression based speech enhancement. Their model is based on deep neural network (DNN). A large training set is used to learn non-linear mapping from noisy speech to clean speech. More than 100 hours of simulated speech data is used with multi-conditions for training purpose. Gao et al. [67] proposed to use deep neural network (DNN) based speech enhancement with progressive learning framework. Their progressive framework decomposes the problem of noisy to clean speech mapping into subproblems to reduce the system's complexity and enhancing performance. Authors further proposed to use LSTM based model in place of DNN with densely connected progressive learning framework in [68]. Moreover, the new structure is trained with multiple learning targets.

Based on the above mentioned studies, it is evaluated that speech enhancement plays a vital role in enhancing the performance of speech processing applications such as ASR and source separation. However, a few works have used speech enhancement for diarization. In [69], Zhu *et al.* used regression-based DNN to map noisy speech features to clean speech features. Clean speech features were then used for diarization. Furthermore, authors proposed to use perceptual linear prediction (PLP) features which significantly reduced the diarization error rate. Their experiments were conducted on Chinese talk show database i.e. IFLY-DIAR-II. Most recently, Sun et al. [61] proposed to use LSTM-RNN based speech enhancement for audio diarization. This model was previously proposed in [70] which was trained to jointly learn multiple targets i.e. ideal ratio masks (IRM) and log power spectrum (LPS). The model was trained on WSJ0 reading-style speech with more than 100 noise types. Their experiments were conducted on unseen noises, which presents the validity of their technique. Similarly, same model was also used in first DIHARD diarization challenge [60] and it significantly improved the DER.

2.4.2 Proposed work

All the above relevant diarization approaches which utilizes speech enhancement are pure unimodal i.e. audio based. This research proposes to use speech enhancement module in multimodal diarization system where both audio and visual pipelines are get affected. In multimodal diarization system a pre-trained audio-visual synchronization model is employed which provides video segments where visible speaker matches with the speech. In the presence of noise this module is eventually get affected and limited high confidence segments are achieved due to corrupted speech. Similarly, while computing the likelihood of speech from trained clusters the process of assigning speech to each cluster is also affected due to noise. In our third contribution of this thesis, both of these issues have been taken care of using speech enhancement technique.

2.5 Summary

This chapter initially presented a general background of speaker diarization system. This background first described the audio only speaker diarization system and its basic components consisting of feature extraction, speech activity detection, segmentation and clustering. It then described the evaluation metric for speaker diarization i.e. Diarization error rate. Finally, brief overview of multimodal speaker diarization system is presented. After that, introduction, related work, and proposed work of each contribution has been described in detail. Each of these contributions were described independently, consisting of feature embedding, multimodal speaker diarization and speech enhancement. These contributions become part of speaker diarization system.

Chapter 3 is based on the methodological sections. It first describes the dataset and then methodology of each contribution. In the first contribution, methodology of feature embedding technique is described, that is based on deep autoencoder architecture. In the second contribution, a novel multimodal speaker diarization system based on a pre-trained audio-visual synchronization model is described. Finally, in the third contribution, speech enhancement technique based on an LSTM model is proposed. All these methodologies are described in detail, supported by diagrammatic description and mathematical modeling.

Chapter 4 is based on the results and its detailed discussion. All the results and its discussions are also described independently. Chapter 5 finally provides the conclusions of each contribution and future recommendations in speaker diarization system.

Chapter 3.

Methodology

This chapter presents the methodological details of all the proposed research contributions in speaker diarization system. These contributions include feature embedding, multimodal diarization and speech enhancement. Initially, details of the dataset and evaluation metric is presented which is common to all the research contributions, then each methodological section is described separately in detail.

3.1 Dataset

A popular subset AMI meeting corpus [7] consisting of 5.4 h of audio recordings and 5.8 h of audio-visual recordings are used. Former is used in feature embedding technique which is an audio-only diarization and later is used in multimodal diarization techniques. These meetings were recorded in English with mostly non-native speakers. This corpus is available with a range of varying audio recording equipment, e.g. close talk microphones, lapels, far-field and near-field microphone arrays. Similarly, the video recordings consist of individual speaker and room view camera setup.

In the experiments, mix-headset audio recordings are used. This recording file is a beamformed version of headset mic of all the speakers. For multimodal diarization, individual speaker camera recordings are used which are also known as closeup cameras. Figure 3.1 shows closeup camera samples from meeting IS1008a.

The acoustic properties of the recordings also vary due to the different room scenarios. The selected subset of each audio recording contains four speakers. In this corpus, each meeting is recorded in four different sessions and independent recordings are provided. Each meeting has sessions of 15–35 min where each recording file has a meeting ID with small lettering representing the session of that recording. For example, meeting IS1000a shows meeting ID ‘IS1000’ with ‘a’ being the recording of the first session and so on. The manual annotations of each session are also provided to check the validity of the diarization.

Figure 3. 1 AMI IS1008a closeup camera images of individual speakers.

3.2 Evaluation metric

Evaluation metric for speaker diarization is Diarization Error Rate (DER). It is combination of four errors: False alarm (E_{FA}), Missed speech (E_{MS}), Speaker error (E_{SE}) and Overlapping speech error (E_{OS}). E_{FA} is defined as a fraction of time where non-speech is labelled as speech in hypothesis. E_{MS} is the fraction of time when actual speech is labeled as a non-speech in the hypothesis. These two errors directly belong to the speech activity detector. The other two errors are E_{SE} when wrong speaker is assigned and E_{OS} when the reference has multiple speaker and it is not labelled as such in the hypothesis. Finally, DER is sum of all these errors, defined as follows:

$$DER = E_{FA} + E_{MS} + E_{SE} + E_{OS} \tag{3.1}$$

3.3 Feature Embedding technique

The following sections describes all the necessary steps required for feature embedding extraction and its utilization in unsupervised speaker diarization.

3.3.1 Preprocessing

In the preprocessing step the audio recording is converted into the 19-dimensional MFCC features and normalized by zero mean and unit variance. Window lengths of 30 ms and hop-lengths of 10 ms is used in feature extraction. From the available annotations, optimal SAD is applied by setting nonspeech audio samples to zero and then a classifier is trained. For this purpose, support vector machine (SVM) classifier is trained with the top 10% energy frames as speech and lowest 10% energy frames as nonspeech. The trained SVM is then used to classify the rest of the speech frames. This process significantly deceases the SAD error. Furthermore, after excluding the nonspeech MFCC frames, these basic features are used for diarization in the baseline method and used for feature embeddings in the proposed method.

3.3.2 Feature embedding

In this study, proposed feature embedding method is based on deep autoencoder. It is an unsupervised method to learn the new set of features. The autoencoder architecture has two parts: a feature encoding part, which is known as the encoder, and a feature decoding part, which is known as the decoder. A typical single-hidden-layer autoencoder can be represented as follows:

$$\boldsymbol{h} = a(W\boldsymbol{x} + b) \tag{3.2}$$

where $\boldsymbol{x}$ is an input vector for the autoencoder, W and b respectively represent the weight matrix and the bias of the encoder, a is a nonlinear activation function, and $\boldsymbol{h}$ represents the output of the encoder. The encoder output $\boldsymbol{h}$ is then fed into the decoder, which reconstructs the input to generate the output represented by $\widehat{\boldsymbol{x}}$ as follows:

$$\widehat{\boldsymbol{x}} = W'\boldsymbol{h} + b' \tag{3.3}$$

where W' and b' represents the weight matrix and bias of the decoder, respectively. These equations can be extended for any large number of encoder and decoder layers. The proposed architecture consists of symmetric layers at the encoder and decoder sides.

Figure 3.2 shows the proposed architecture of the deep autoencoder. For such an unsupervised feature learning technique, this architecture is trained on the input data X with the same output labels X. The encoder part of the architecture provides the new set of features after the architecture has been trained. Shrinkage architecture [71] has been used to learn the low-dimensional features. To learn such low-dimensional features, five consecutive MFCC frames are grouped together to create the input dimension of $19 \times 5 = 95$, and then the architecture is trained. Motivation behind such feature grouping is that speaker change usually does not occur in five successive frames and each speaker segment contains many successive MFCC frames.

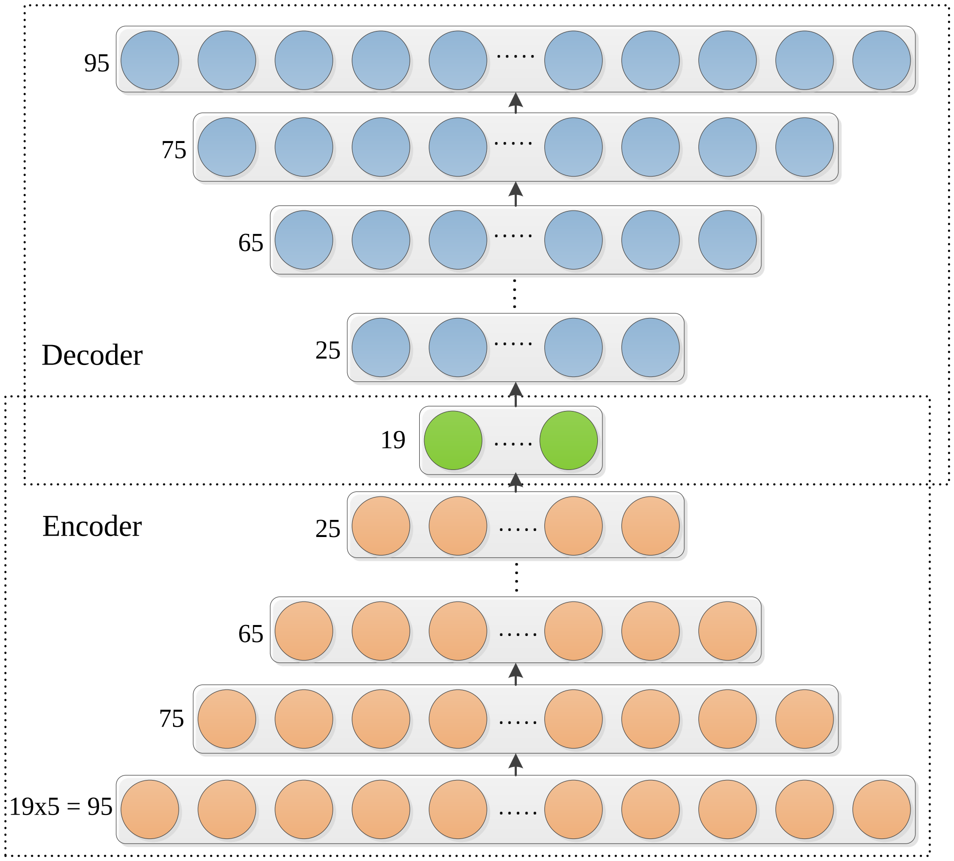


Figure 3. 2 Deep autoencoder architecture

3.3.3 Autoencoder architecture

Feature embedding technique trains deep autoencoders with 13 hidden layers. The architecture is designed such that it shrinks at each layer. The input layer with 95 dimensions to the first hidden layer reduces the nodes by 20 and then each successive hidden layer has ten fewer nodes than the previous one, until it reaches the encoder output with 19 nodes. This architecture was proposed empirically by testing different set of node and layers. It was eventually evaluated that gradual reduction in the number of nodes through each layer provides the reconstruction more robustly. The architecture of encoder layers with number of nodes in each layer is represented as follows:

$$(input)95 \rightarrow 75 \rightarrow 65 \rightarrow 55 \rightarrow 45 \rightarrow 35 \rightarrow 25 \rightarrow 19\ (encoder\ out)$$

Due to the symmetric architecture of the encoder and decoder, decoder layers have the same number of nodes at each layer as the encoder. The decoder architecture tries to reconstruct the encoded information from the encoder part. Nodes of the decoding layer are represented as follows:

$$19\ (encoder\ out) \rightarrow 25 \rightarrow 35 \rightarrow 45 \rightarrow 55 \rightarrow 65 \rightarrow 75 \rightarrow 95\ (out)$$

After the architecture has been trained, the encoder part provides the low-dimensional and newly learned features. To train the architecture, 5 MFCC frames were grouped together, thus making an input vector of 95 dimensions.

3.3.4 Proposed method

The baseline method [72] applies the GMM based segmentation and clustering on MFCC features. An agglomerative hierarchical clustering technique based on GMM is used with initially 16 clusters and 5 Gaussian mixtures in each cluster. For majority vote segmentation, the segment length of 1.5 s is considered. In feature embedding autoencoder architecture input and output are of 95 dimensions. The deep autoencoder is trained with a small batch size of 32 frames, 100 epochs, and Adadelta optimization. As labels of the architecture are the same as the training data, normalized to zero mean and unit variance, mean squared error (MSE) objective function is used. It is represented as follows:

$$MSE = \frac{1}{N} \sum_i (y_i - \widehat{y_i})^2 \quad (3.4)$$

where y_i represents the input speech frame and $\widehat{y_i}$ represents the reconstructed speech frame by the autoencoder model. N represents the total number of speech frames. The

MFCC features were normalized to zero mean and unit variance before applying feature embeddings, and to learn the new features more robustly, hyperbolic tangent (tanh) activation function is used. After the model is trained, the encoder part is used to extract the 19-dimensional features as a representation of 95-dimensional input. The output of encoder is further normalized by zero mean and unit variance and then GMM based segmentation and clustering is applied as discussed for the baseline method. As 5 MFCC frames have already been grouped, segment length is adjusted in GMM training so that total segment length remains 1.5 s.

3.4 Multimodal diarization technique

3.4.1 Audio preprocessing

The audio data set comprises of mix-headset audio recordings that consists of voices from all the speakers in a single wav file. Initially, Mel-Frequency Cepstral Coefficients (MFCC) [73] features are extracted and normalized by zero mean and unit variance. Then, energy-based speech activity detection (SAD) is applied to classify speech and non-speech frames. For that purpose, available annotations are used to make non-speech audio samples to zero. In the SAD block, MFCC features were concatenated with energy features. A support vector machine classifier is applied to classify speech and non-speech frames which is trained on the 10% highest and 10% lowest energy frames. The SAD block provides speech only in MFCC frames and discards non-speech frames. Figure 3.3 shows the audio preprocessing pipeline consisting of MFCC feature extraction and SAD.

3.4.2 Video preprocessing

From the Augmented Multi-party Interaction (AMI) [7] corpus, the available video dataset consists of multiple recordings from cameras mounted in different room places. To capture the face of an individual speaker, closeup cameras mounted on the tabletop is used. This camera configuration is presented in Figure 3.4, where four tabletop cameras are mounted to capture the individual speakers. Face detection is applied on each closeup camera stream and the face-only region is extracted. Afterwards, video frames consisting of silent parts are removed using the output of audio SAD module. Shot detection technique is then applied to track continuous frames which contain faces and split the video frames into each shot where the face detector misses its detection. As in audio-based diarization techniques, segment length is usually defined based on the assumption that each speaker will speak for at least a particular segment time. In conventional audio speaker diarization technique based on HMM/GMM, each cluster is trained on speech frames consisting of at least one segment length duration. In the video part, 2-second segment length is selected to split the video shots into smaller video segments. This can help to identify active speakers in each short video. For each video segment, audio-visual synchronization is determined between the audio and mouth motion in a video. For that purpose, a pre-trained SyncNet model is utilized which find out how much audio speech belongs to the visible speaker.

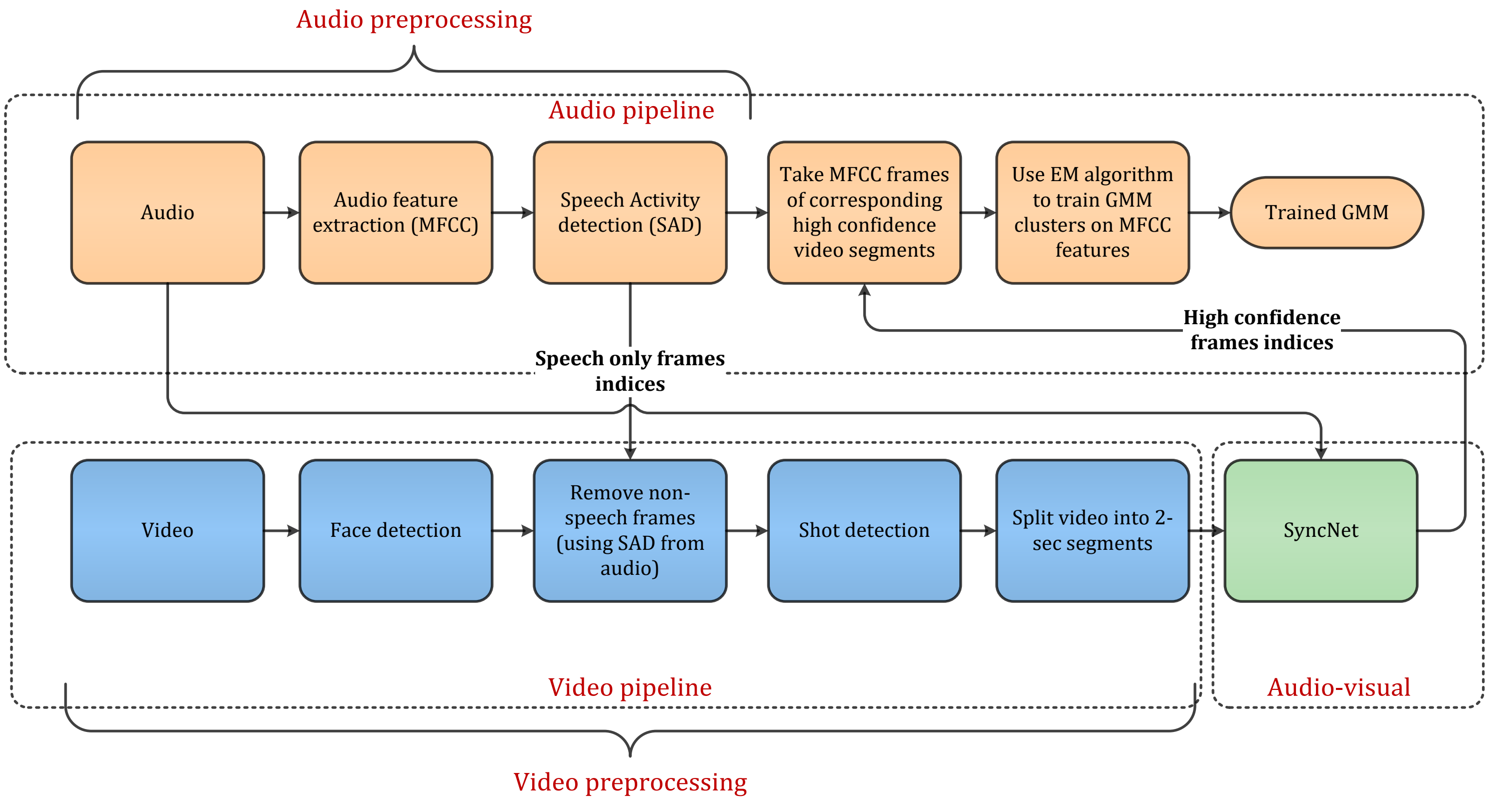

Audio preprocessing
Audio pipeline
Audio
Audio feature extraction (MFCC)
Speech Activity detection (SAD)
Take MFCC frames of corresponding high confidence video segments
Use EM algorithm to train GMM clusters on MFCC features
Trained GMM
High confidence frames indices
Speech only frames indices
Video
Face detection
Remove non-speech frames (using SAD from audio)
Shot detection
Split video into 2-sec segments
SyncNet
Video pipeline
Audio-visual
Video preprocessing

Figure 3. 3 Proposed multimodal speaker diarization system

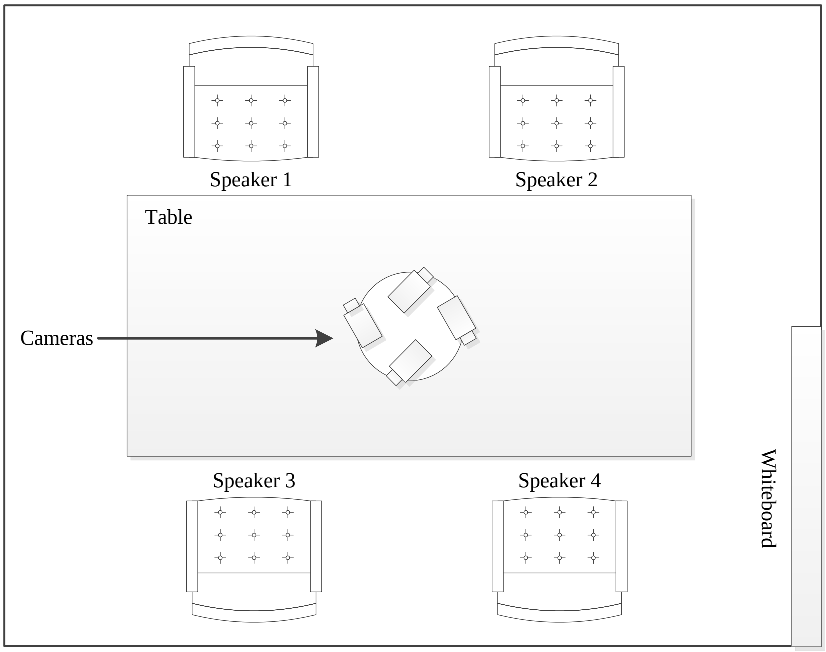


Figure 3. 4 AMI meeting room setup

3.4.3 SyncNet architecture and Inference

The SyncNet architecture [58] is a two-streamed model consisting of audio and visual convolutional neural networks with contrastive loss. This model was trained on several hundred hours of speech from BBC videos that include hundreds of speakers. Audio data with a sampling frequency of 16 KHz is converted into 13-MFCC features at the rate of 100 Hz. The audio part of the network is provided with 0.2 s of speech consisting of 20 MFCC frames, makes 13 × 20 dimensional input. Input to the visual part of the network is the face region which is extracted using a face detection technique. For a 25 Hz video frame rate, five consecutive video frames are grouped, which provides 0.2 s of video segment. For the video network, the input data is of 120 × 120 × 5 dimensional. Figure 3.5 presents the two streamed audio-visual CNN architecture of SyncNet. This architecture takes the output of both the audio and visual last fully connected layer and applies contrastive loss to minimize the distance between genuine audio and corresponding video pairs. This loss is described as follows:

$$E = \frac{1}{2N} \sum_{n=1}^{N} (y_n) d_n^2 + (1 - y_n) \max(margin - d_n, 0)^2 \qquad (3.5)$$

$$d_n = \|v_n - a_n\|_2 \qquad (3.6)$$

where a and v are the outputs of the last fully connected layers, $y \in [0,1]$ is the binary similarity metric between the input and video inputs.

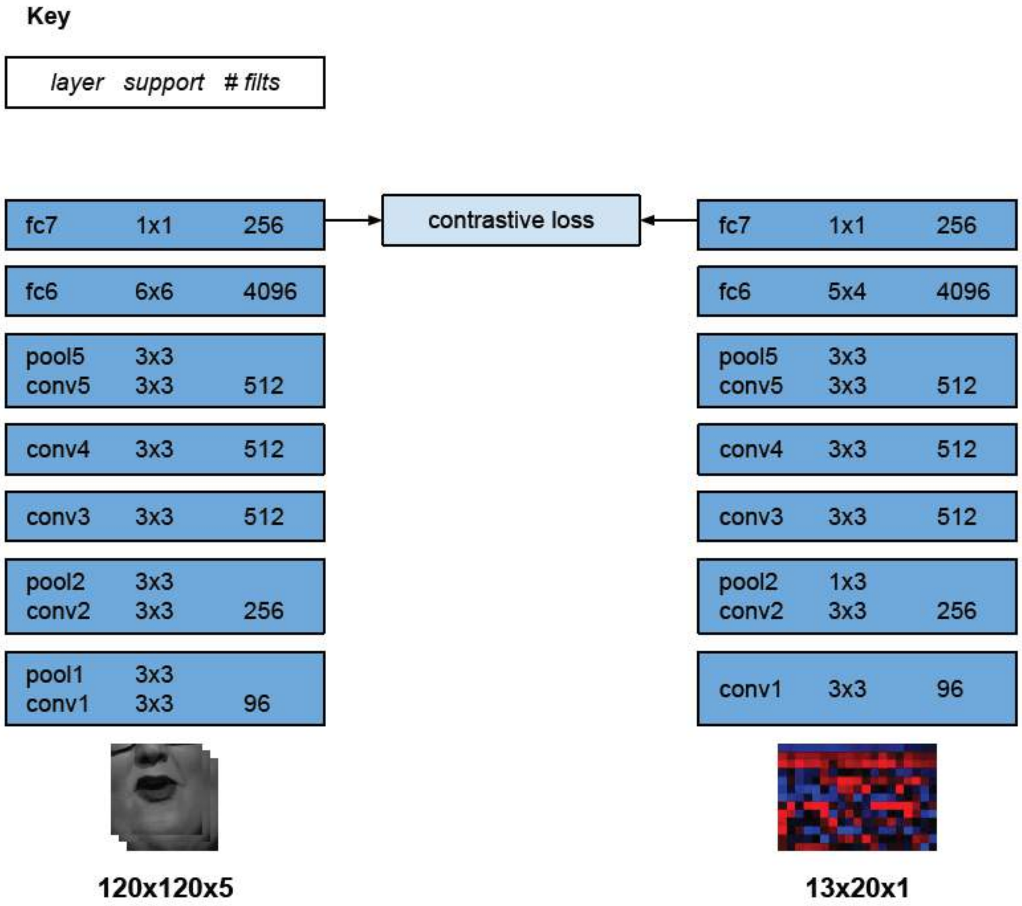


Figure 3. 5 SyncNet’s audio-visual synchronization architecture.

In multimodal speaker diarization system a pre-trained SyncNet model is used to determine the active speaker in each closeup video segment. As discussed in video preprocessing section, video is split into short segments of 2s length each and then SyncNet's inference is applied. The SyncNet model computes two metric values at the output: offset and confidence, which are computed to determine the audio-visual relationship. Segments which have lowest offset and high confidence values determine that the visible speaker is the one who is speaking. In proposed approach, two threshold values are applied which are based on the analysis of those video segments where complete audio belongs to the visible speaker. The first one is the offset threshold, which is defined as $Th_{of} = [0, t_1]$, which only select video segments whose audio-visual offset value is between 0 and t_1 (both inclusive). After shortlisting the video segments by applying the first threshold the confidence threshold $Th_{conf} > t_2$ is applied. It only selects those video segments whose audio-visual matching confidence is greater than t_2. These two types of thresholds hierarchically select only those video segments whose audio matches the visible speaker with high confidence. These segments are named as high confidence video segments. The video frame indices of high confidence video segments provided by SyncNet are used in the audio pipeline to train a GMM cluster on the corresponding MFCC frames. Finally, for each closeup video belonging to one speaker single GMM is trained. Such clusters are named as pure GMM, because they are only trained on high confidence frames.

3.4.4 Complete multimodal diarization pipeline

Audio preprocessing pipeline provides speech only MFCC frames after applying speech activity detection. Video preprocessing pipeline provides short video segments of length 2s each. Following the inference results from the SyncNet architecture, high confidence segments are used to acquire respective MFCC frames. GMM model is trained using Expectation Maximization (EM) with those MFCC frames. So, for each closeup video a GMM model with K mixtures is trained on the corresponding high confidence MFCC frames. Gaussian mixture model can be represented as follows:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \sum_{i=0}^{K} \pi_i N(\boldsymbol{x}, \boldsymbol{\mu_i}, \Sigma_i) \tag{3.7}$$

where $\boldsymbol{\mu_i}$ represents the means vector for each mixture, π_i is the mixture coefficient and Σ_i is the covariance matrix. In expectation step responsibilities are calculated as follows:

$$r_{jc} = \frac{\pi_c N\left(\boldsymbol{x_j}\middle|\boldsymbol{\mu_c}, \Sigma_c\right)}{\sum_{\boldsymbol{i=0}}^{\boldsymbol{K}} \pi_i N(\boldsymbol{x_i}|\boldsymbol{\mu_i}, \Sigma_i)} \tag{3.8}$$

While calculating r_{jc}, j represents the j^{th} datapoint and c represents the mixture number. The maximization step consists of calculating mean vectors, covariance matrix and mixture components as follows:

$$\boldsymbol{\mu_c^{new}} = (1\backslash N_c) \sum_j r_{jc} \, \boldsymbol{x_j} \tag{3.9}$$

$$N_c = \sum_j r_{jc} \tag{3.10}$$

$$\Sigma_c^{new} = \frac{1}{N_c} \sum_j r_{jc} \left(\boldsymbol{x_j} - \boldsymbol{\mu}_c^{new}\right)\left(\boldsymbol{x_j} - \boldsymbol{\mu}_c^{new}\right)^T \tag{3.11}$$

$$\pi_c = \frac{N_c}{n} \tag{3.12}$$

where n is the total number of data points in the dataset.

Such cluster is assumed to be pure and it is most likely to be trained on single speaker. After all the clusters have been trained, which are now speaker-dependent clusters, the likelihood of the rest of the MFCC frames from each cluster is computed. The most likely frames are assigned to that specific cluster. Finally, all the frames are assigned to one of the clusters and the diarization error rate is computed.

In speaker diarization problems the number of speakers are usually unknown, so this is assumed to be equals to the number of available closeup video recordings. In such scenarios where only one recording is available, one can assume the number of speakers equals the number of distinct faces in all the video frames. Proposed diarization technique is designed for such scenarios where all the speakers face the camera.

3.4.5 Experimental setup

Audio pipeline extract 19-dimensional MFCC features with window length of 30ms and hop length of 10ms. MFCC frames thus have sample rate of 100Hz. In SAD block, MFCC features along with energy features were used with same window and hop length. An SVM based classifier is used to train speech and non-speech frames. Further, when audio pipeline gets high confidence video segment information, a GMM with 20 components and diagonal covariance matrix is trained using Expectation Maximization algorithm. For each closeup video, the number of high confidence video segments varies which leads to different number of MFCC frames for each speaker. The specification of the GMM model for each speaker is kept same i.e. 20 components, diagonal covariance.

Given each closeup video with frame rate of 25fps, face detection is applied with subsequent SAD and shot detection. All the available shots are then further segmented into 2 sec chunks to apply SyncNet. This segment length is decided based on the assumption that each speaker may speak for a minimum of 2 sec. Moreover, shots and segments smaller than 7 frames are also discarded because they are too small to provide some reliable audio-visual matching. Two metrics are computed on the output of SyncNet's inference that is offset and confidence, two threshold values were further applied to filter out unsynchronized audio-visual segments. To select high confidence video segments, we choose offset threshold range between 0 and t_1, where $t_1 = 3$. Confidence threshold is applied with value $t_2 = 1.5$. Segments with offsets less than 0 and greater than 3 are discarded. Similarly, if offset value is within the threshold then segments with confidence less than 1.5 are also discarded. After applying these two threshold values and discarding the video segments, it is more likely that in the remaining video segments complete audio only belongs to the visible speaker.

3.5 Speech Enhancement

3.5.1 Audio pipeline

In the audio pipeline, noisy speech is first passed through speech enhancement module to get clean speech. Further, MFCC [73] features are extracted followed by Speech Activity Detection (SAD) block which discards non-speech frames. After acquiring speech only frames, a model is trained to classify speech according to available speakers. For that purpose, high confidence video frames indices which are later mapped to audio frames in audio pipeline are acquired from SyncNet's inference. Finally, a GMM model is trained on those high confidence frames of each speaker. This complete pipeline is an enhanced version of multimodal diarization technique already presented in section 3.4. An updated multimodal diarization with speech enhancement module is presented in Figure 3.6.

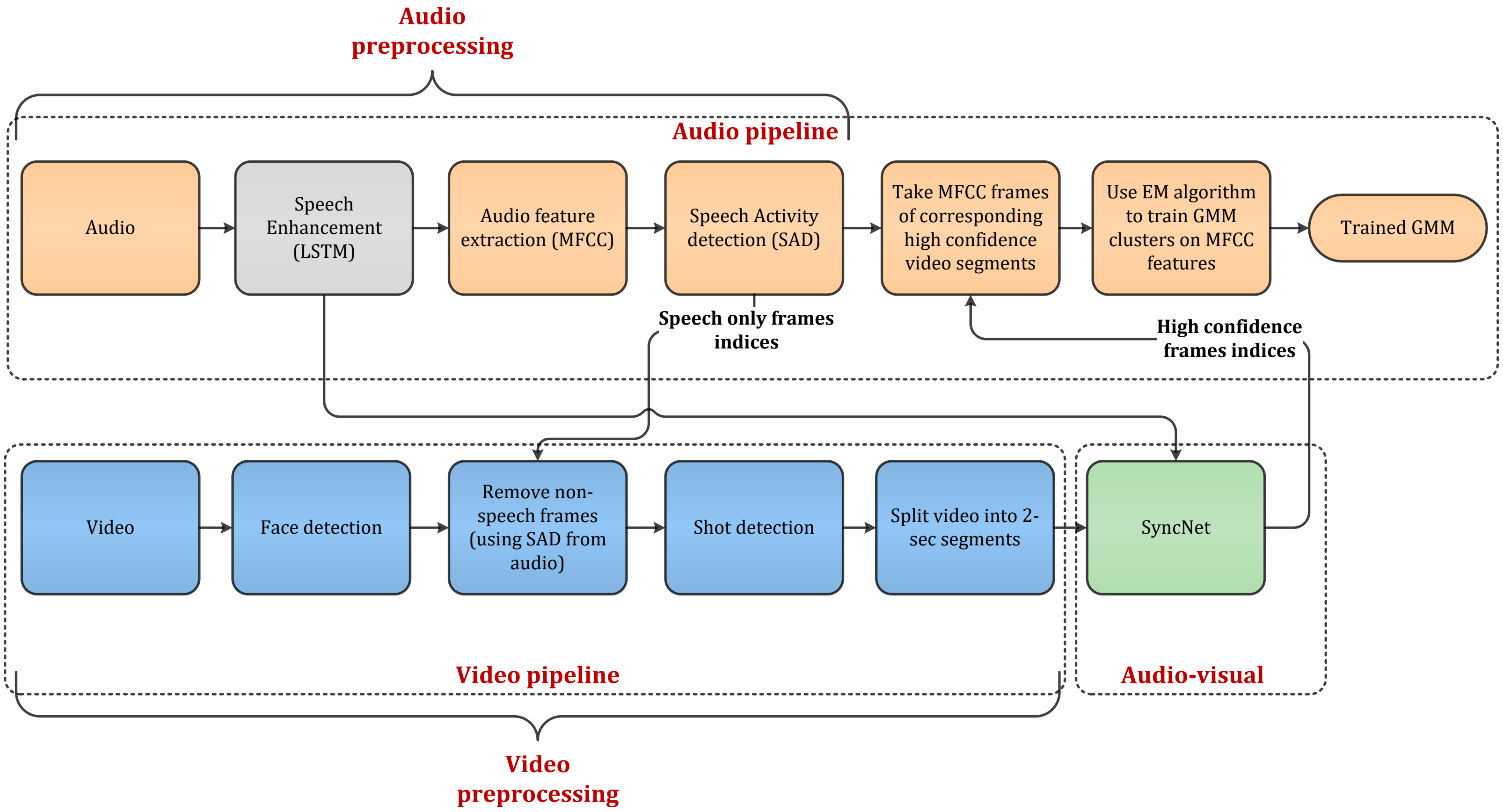

Audio preprocessing
Audio pipeline
Audio
Speech Enhancement (LSTM)
Audio feature extraction (MFCC)
Speech Activity detection (SAD)
Take MFCC frames of corresponding high confidence video segments
Use EM algorithm to train GMM clusters on MFCC features
Trained GMM
Speech only frames indices
High confidence frames indices
Video
Face detection
Remove non-speech frames (using SAD from audio)
Shot detection
Split video into 2-sec segments
SyncNet
Video pipeline
Audio-visual
Video preprocessing

Figure 3. 6 Proposed multimodal speaker diarization system with speech enhancement module.

3.5.2 Speech Enhancement

The speech enhancement module is based on densely connected progressive learning LSTM network. The idea behind progressive learning is that each hidden layer of LSTM network is trained to learn an intermediate target. Figure 3.7 [68] presents the densely connected progressive learning LSTM network with multiple targets (3 targets). These targets are designed to explicitly learn high SNR speech at each layer. Figure 3.7 also presents the minimum mean square error (MMSE) loss function at each target layer as E_1, E_2 and E_3. The weighted loss function for M target layers with multi-task learning (MTL) can be presented as follows:

$$E = \sum_{k=1}^{M} \alpha_k E_k + E_{IRM} \tag{3.13}$$

$$E_k = \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}(\hat{x}_n^0, \hat{x}_n^1, \dots, \hat{x}_n^{k-1}, \Lambda_k) - x_n^k \|_2^2 \tag{3.14}$$

$$E_{IRM} = \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}_{IRM}(\hat{x}_n^0, \hat{x}_n^1, \dots, \hat{x}_n^{k-1}, \Lambda_{IRM}) - x_n^{IRM} \|_2^2 \tag{3.15}$$

where E_k represents mean square error for k^{th} target layer. E_{IRM} is mean square error with ideal ration masks at the final output layer. x_n^k and $\hat{x}_n^k$ are n^{th} D-dimensional reference and estimated log power spectra (LPS) feature vector for k^{th} layer. N represents mini-batch size and Λ_k is the set of weight matrix and bias before k^{th} layer. x_n^0 is the input noisy LPS feature vector with acoustic context and α_k is the weight of k^{th} target layer. Similarly, $\mathcal{F}_{IRM}(\hat{x}_n^0, \hat{x}_n^1, \dots, \hat{x}_n^{k-1}, \Lambda_{IRM})$ and Λ_{IRM} are the corresponding versions of IRM target. The network is optimized with gradient descent and back propagation through time (BPTT) [74] algorithm. Network is trained on about 400 h of Mandarin and English speech having sample rate of 16KHz. Figure 3.8. [68] presents the example of noisy and enhanced speech utterance.

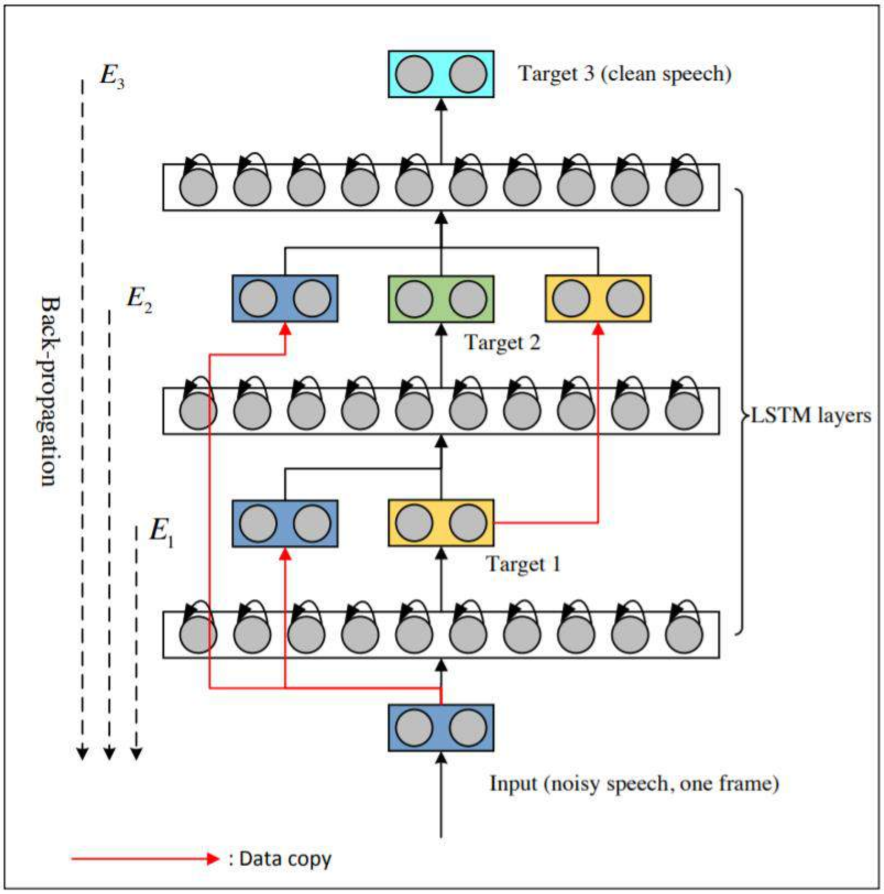
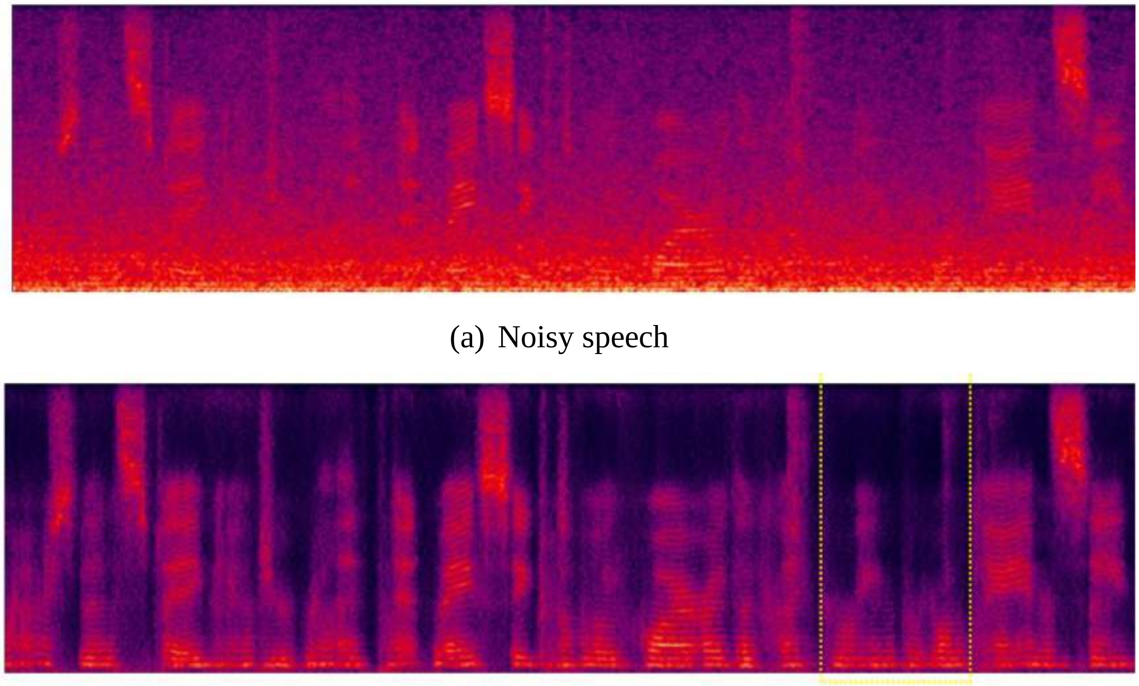


Figure 3. 7 Speech enhancement based on densely connected progressive learning LSTM network.



(a) Noisy speech

(b) Enhanced speech

Figure 3. 8 Spectrograms of speech utterance (a) Noisy and (b) Enhanced.

3.5.3 Complete diarization pipeline with speech enhancement

The video pipeline is similar to one used in multimodal diarization under section 3.4. Similar audio and video dataset is used comprising of mix-head set audio recording and closeup camera video streams of individual speaker. The technique first enhances the speech and then extracts MFCC features, applies SAD and then wait for the audio-visual SyncNet model to provide high confidence video segments information to train a model. Video pipeline processes individual closeup camera stream by applying face detection, SAD and splitting the face video in short video segments. Further, audio-visual SyncNet architecture takes the short video segments with its corresponding audio and finds out audio-visual synchronization. For each short video segment two metrics are computed i.e. offset and confidence. Based on the two thresholds values for each of these metrics, high confidence video segments for each speaker is identified. On the basis of high confidence video segments, their corresponding audio frames are grouped together for each speaker and a GMM model is trained.

In meeting recordings all of the speakers don't usually get equal opportunity to speak within the same time intervals. Hence, the number of audio samples for each speaker varies. To train a model S number of Gaussian mixture models (GMM) are used with each K mixtures, where S represents the number of speakers. Each GMM model is trained with individual speaker's audio samples, whereas remaining audio frames that were not part of the high confidence frames are assigned to the one of the mixture models on the basis of maximum likelihood. For GMM training Expectation maximization (EM) algorithm is used similar to one used in section 3.4.

3.5.4 Experimental setup

Two set of experiments has been conducted, one with the AWGN noise and another with the environmental noise. In the first experiment, audio recordings are corrupted by Gaussian noise with 10dB SNR to create synthetic noisy data set. In the second experiment, environmental noise taken from PNL100 [75] database is used to corrupt the recordings with 10dB SNR. Noisy recordings of both experiments are passed through Wiener filtering [76] and LSTM based speech enhancement. These noisy recordings and their enhanced speech are used in multimodal speaker diarization system for comparison.

In audio pipeline pretrained speech enhancement model is used to remove the noise. 19-MFCC features are extracted with window length of 30ms and hop length of 10ms, having frame rate of $100Hz$. SAD block uses 19-MFCC features along with energy features with same window and hop length. In the final stage of the audio pipeline, high confidence frames are used to train a GMM model for classifying rest of the audio frames. Each GMM model comprises of 20 components with diagonal covariance matrix. An expectation maximization algorithm is used to train this model.

Video pipeline comprises of input at 25fps, applies face detection and then SAD. Further, each face video is converted into short segments of maximum of 2 sec each. On these 2 sec segments, SyncNet inference is applied to find the audio-visual synchronization in term of offset and confidence. All the face video segments are combination of continuous face tracks. So, in this pipeline, we discard any segment that is shorter than 7 frames. As discussed in SyncNet section, two threshold values are applied to select high confidence video segments. For offset value we applied its threshold as $t_1 = 2$, so it selects only those video segments whose offset is between 0 and 2. For the confidence value, applied threshold $t_2 = 1.5$, it selects video segments with confidence threshold greater than 1.5.

3.6 Summary

This chapter initially presented the detailed description of audio and video data set used in this research and then evaluation metric is presented. Furthermore, methodological detail of proposed work is described in three sections. These methodologies comprise of feature embedding technique using deep autoencoders for speaker diarization, multimodal speaker diarization using a pre-trained audio-visual synchronization model and speech enhancement for multimodal speaker diarization. All the mathematical description along with experimentation part is described in detail.

Chapter 4.

Results and Discussion

This chapter presents the results and their detailed discussion for the three experiments already discussed in chapter 3 i.e. feature embedding, multimodal technique and speech enhancement. These three sections are described individually as follows.

4.1 Feature embedding

Feature embedding technique proposed to use deep autoencoder architecture to extract new set of features from encoder output. This architecture was trained on MFCC features of given audio dataset. The feature embedding were used in GMM based agglomerative clustering technique for diarization. Chapter 3 already discussed detailed methodology and its experimentation part. Following sections describe the acquired results in detail.

4.1.1 Results

Table 4.1 shows the computed DER for the baseline method on mix-headset audio recordings. Due to random initialization in the GMM based method, which usually ends up at different local minima, the experiment is conducted on each audio recording 10 times and average DER is computed. The right-most column represents the average DER of each audio recording. Finally, overall average DER for all the audio recordings is computed, which is represented in the last row. It shows that on this subset of 12 audio recordings (5.4 h) the average DER of the baseline method is 44.11%.

Similarly, Table 4.2 shows the DER of the proposed feature embeddings (FE) method. The proposed method’s overall average DER is 41.15%, which gives improvement of 2.96% as compared to the baseline method. Table 4.3 provides a comparison of average DER for each audio recording and presents improvement in the DER for each audio recording. The maximum improvement among the individual recordings has been observed for IS1006b, which is 8.05%. Overall, it has been observed that there is reduction of DER for all the recordings except two, and overall improvement is very significant.

Table 4.1 Diarization error rate (%) of baseline method.

Meeting ID	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run 5**	**Run 6**	**Run 7**	**Run 8**	**Run 9**	**Run 10**	**Average**
IS1000a	36.85	42.21	41.72	42.19	42.35	41.93	42.09	42.17	46.95	42.33	**42.07**
IS1001a	41.62	41.48	41.9	42.61	42.07	41.76	42	42.74	42.52	42.74	**42.14**
IS1001b	48.88	48.34	48.71	48.7	47.57	48.67	48.16	47.6	48.71	47.67	**48.30**
IS1001c	52.3	53.08	53.68	52.08	52.99	53.89	53.11	52.4	52.27	53.09	**52.88**
IS1003b	50.44	51.14	50.27	50.73	50.54	50.17	48.64	50.71	51.15	63.02	**51.68**
IS1003d	68.38	68.81	68.61	67.5	69	67.87	69.04	68.83	68.57	67.82	**68.44**
IS1006b	59.74	49.74	50.04	66.32	49.63	42.63	49.61	59.48	49.76	49.55	**52.65**
IS1006d	66.02	66.95	66.89	67.07	66.9	66.67	66.96	67.13	67.03	66.87	**66.84**
IS1008a	11.87	20.25	20.38	20.53	11.89	11.87	11.96	12.2	20.34	20.43	**16.17**
IS1008b	12.08	12.02	12.73	11.85	11.91	11.74	12.54	11.79	11.87	12.22	**12.07**
IS1008c	41.16	40.46	39.81	40.91	39.64	41.19	40.52	40.71	39.98	41.52	**40.59**
IS1008d	38.23	37.61	26.21	38	37.55	38.1	37.85	37.63	26.23	37.62	**35.50**
Average DER for all the audio recordings											**44.11**

Table 4.2 Diarization error rate (%) of feature embedding method.

Meeting ID	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run 5**	**Run 6**	**Run 7**	**Run 8**	**Run 9**	**Run 10**	**Average**
IS1000a	36.75	39.26	38.08	33.82	37.4	36.8	41.07	36.46	42.39	38.27	**38.03**
IS1001a	41.33	42.62	40.29	40.32	43.16	41.35	42.8	44.2	40.71	43.72	**42.05**
IS1001b	47.81	48.36	48.04	47.69	47.58	48.04	48.19	48.43	47.31	48.3	**47.97**
IS1001c	55.22	52.44	53.46	52.92	55.02	53.71	54.07	52.97	55.05	54.65	**53.95**
IS1003b	49.37	27.13	50.56	50.4	38.23	50.97	50.38	50.85	29.11	49.68	**44.66**
IS1003d	58.99	59.54	60.71	59.06	60.58	69.06	69.5	69.52	60.2	59.79	**62.69**
IS1006b	66.21	42.59	32.96	57.77	44.95	32.18	43.77	50.44	42.2	32.88	**44.59**
IS1006d	66.95	66.89	65.25	65.41	66.84	66.8	65.78	65.79	66.77	66.57	**66.30**
IS1008a	11.41	11.06	11.45	12.61	13.43	11.55	10.97	19.96	13.93	11.71	**12.80**
IS1008b	17.46	12.43	13.19	16.41	13.74	16.27	16.45	13.87	13.03	13.03	**14.58**
IS1008c	40.79	40.26	41.17	25.27	25.93	40.85	40.61	41.41	40.31	40.85	**37.74**
IS1008d	37.91	25.91	27.05	26.42	25.83	25.93	38.04	25.33	26.45	25.57	**28.44**
Average DER for all the audio recordings											**41.15**

Table 4.3 Comparison of average Diarization error rate (%) of baseline and proposed technique.

Meeting ID	**Avg. results of Baseline**	**Avg. results of FE Method**	**Difference (improvement)**
IS1000a	42.07	38.03	4.04
IS1001a	42.14	42.05	0.09
IS1001b	48.30	47.97	0.32
IS1001c	52.88	53.95	-1.06
IS1003b	51.68	44.66	7.01
IS1003d	68.44	62.69	5.74
IS1006b	52.65	44.59	8.05
IS1006d	66.84	66.30	0.54
IS1008a	16.17	12.80	3.36
IS1008b	12.07	14.58	-2.51
IS1008c	40.59	37.74	2.84
IS1008d	35.50	28.44	7.05
Average of all	44.11	41.15	-
Improvement (%)	**2.96**		

4.2 Multimodal diarization

Multimodal speaker diarization technique comprised of using a pre-trained audio-visual synchronization model to find the active speaker in any time interval. Active speaker segments are determined in audio-visual domain and used to train the speaker specific MFCC based GMM clusters. Following sections initially describe the comparison methods consisting of unimodal and multimodal diarization techniques. Finally, proposed method results with each of the comparison methods are presented in detail.

4.2.1 Comparison methods

To compare proposed multimodal technique, first conventional speaker diarization (SD) method consisting of agglomerative hierarchical clustering [9] based on HMM/GMM is used. Such technique initializes with large number of clusters e.g. 16 and then hierarchically merges them based on Bayesian information criterion [10], which eventually ends up with optimal number of clusters (speakers). First method for comparison is conventional speaker diarization (SD) [72] based on GMM based hierarchical clustering which is completely an unsupervised technique.

Second comparison method is fully supervised speaker diarization (SD) described in [77] which employs speech activity detection and speaker change detection [78] based on Bidirectional LSTM [79], neural speaker embedding [80] based on LSTM network and triple loss function. All these modules are combined in speaker diarization pipeline and are jointly optimized with affinity propagation [81] clustering. Each module of this method is fully supervised and trained on about 70% of AMI meeting corpus, while proposed method is completely unsupervised. The subset of AMI corpus used in this approach is either part of training or development set in this competing method, which makes this comparison very challenging for the proposed approach. Results section describes this in detail.

Thirdly, for the completeness of this research comparison to the state-of-the art multimodal technique is presented. This multimodal technique is described in [45] where authors used sound source localization (SSL) technique in the audio domain and motion & lip movement measures in the video domain. The SSL technique is used to detect active speaker and overlapping speech detection in the audio domain. Finally, output from both streams are combined through decision algorithm to acquire diarization results. Results presented comprises of particular scenarios where speakers are seated and do not stand up and move towards the whiteboard or screen.

Comparison to other multimodal techniques are difficult since the scenario of the recordings and proposed technique varies significantly. Specifically, important factors that motivate to develop any diarization technique vary depending on overlapping speech intervals, recording equipment in terms of cameras and mic arrays, available speaker's information and available training data. Some of the recent multimodal diarization techniques [15], [46], [82] employ sound source localization technique in audio pipeline along with motion detection, face detection or mouth/lip movement and finally audio-visual fusion in video pipeline. Proposed techniques in these papers are heavily oriented towards sound source localization and data sets used such as AVDIAR [46], MVAD [15] and AVASM [83] contains large fraction of overlapping speech. However, proposed technique in this thesis doesn't employ any localization technique in audio pipeline, rather it locates active speaker through audio-visual synchronization. Secondly, overlapping speech detection is not considered in this research. Although overlapping speech error is included in computing diarization error rate.

4.2.2 Computational cost

In the proposed system, majority of the cost comes from the video pipeline because audio pipeline consists of MFCC feature extraction, SAD's inference and computing GMM's likelihood. These processes are very quick and requires fewer computations. In video pipeline, face detection is based on dlib's [84] Convolutional neural network (CNN).

The computation time of this face detection module with image size of 288×352 on Nvidia GPU GTX1060 is about 0.03 s. With such resolution this module can process approximately 33 frames per second which is quicker than the frame rate of available recordings i.e. 25 FPS. The processing time of face detector linearly increases with the size of the image. Given number of pixels $P = 101{,}376 = 288 \times 352$, frame rate is $F = 33$. If number of pixels are scaled by an integer number s that is $P * s$, then frame rate would be F/s. The computation time of each frame for four camera streams would be:

$$C_t = 4 * s/F \tag{4.1}$$

Next, computational complex block is SyncNet which is a two streamed (audio and visual) Convolutional neural network. The input to this block is five frames of size 120×120 each. Each short video segment in our case comprises of maximum 2 sec length which is provided to SyncNet for inference. This module processes each short segment in approximately 0.6 sec. The total complexity of this module depends on the number of short video segments. This number varies for each camera recording. Computation time of this module for each camera is:

$$C_{sync} = 0.6 * v_s \tag{4.2}$$

where v_s is the total number of short video segments.

4.2.3 Results and Discussion

Table 4.4 presents diarization error rate for the proposed and comparison method based on Agglomerative hierarchical clustering (AHC) discussed in [72]. Table shows DER for individual recordings, their difference in term of improvement, average DER of all the recordings and finally average improvement. Any negative value in improvement column shows reduction in performance. Conventional method is completely unsupervised technique where actual number of speakers are unknown. Recording scenarios of meeting corpus used in the comparison varies significantly in term of overlapping speech duration, speaker movements to the whiteboard and short utterances. In all the recordings, significant improvement is noticed and maximum error reduction is by IS1003b of about 29.5%. Moreover, on the average results for the whole subset, 13.58 % error reduction is achieved. This is due to fact that the proposed technique creates pure audio clusters with the help of high confidence video frames acquired from audio-visual SyncNet model. Such technique significantly reduces the speaker error which assigns wrong speaker to the audio segments.

Second comparison to proposed method is fully supervised speaker diarization technique [77] which is very challenging for the proposed methodology. One of the recording IS1003b is not included in the comparison because it was not part of any training, testing or development set. Table 4.5 shows DER for individual recording with their gain or reduction in improvement. Third column represents that the recording is either part of training (Train) or development (Dev) set. In this method recordings of the development set are used for hyperparameter optimization for speech activity detector, speaker change detector and speaker diarization pipelines. When compared to the proposed method we noticed DER improvement in most of the recordings. While, in the training subset we see both increase and decrease in DER. A noticeable improvement from the training set is gained for IS1000a recording while maximum impairment is by IS1006b. Overall, average improvement is 1.4 % for the proposed multimodal method. Maximum improvement gain is 17.2% for IS1000a.

Table 4.4 Comparison of Diarization error rate (%) score with conventional speaker diarization (SD).

Meeting ID	**Conventional SD** [72]	**Proposed Multimodal**	**Difference (Improvement)**
IS1000a	42.079	29.313	12.766
IS1001a	42.144	37.573	4.571
IS1001b	48.301	35.709	12.592
IS1001c	52.889	24.389	28.5
IS1003b	51.681	22.169	29.512
IS1003d	68.443	48.655	19.788
IS1006b	52.65	42.861	9.789
IS1006d	66.849	58.497	8.352
IS1008a	16.172	10.946	5.226
IS1008b	12.075	12.715	-0.64
IS1008c	40.59	22.217	18.373
IS1008d	35.503	21.376	14.127
Average	44.11	30.535	-
Average improvement	**13.58**		

Finally, table 4.6 presents the result of multimodal technique where 5.8 h of particular subset of IS recordings are taken in which all the speakers are seated. The reason to choose the subset which comprises of static/seated speakers is that the proposed technique does not employ any localization technique. The results clearly show that the proposed technique performs nearly same to the state-of-the art multimodal approach with just

0.56 % impairment. It shows that such technique is as effective as any such complex diarization approach.

Table 4.5 Comparison of Diarization error rate (%) score with fully supervised speaker diarization (SD) system.

Meeting ID	**Fully supervised SD** [77]	**Recording set**	**Proposed Multimodal**	**Difference (Improvement)**
IS1000a	46.55	Train	29.313	17.237
IS1001a	43.31	Train	37.573	5.737
IS1001b	26.77	Train	35.709	-8.939
IS1001c	25.74	Train	24.389	1.351
IS1003d	59.56	Train	48.655	10.905
IS1006b	29.87	Train	42.861	-12.991
IS1006d	51.06	Train	58.497	-7.437
IS1008a	13.84	Dev	10.946	2.894
IS1008b	14.97	Dev	12.715	2.255
IS1008c	22.26	Dev	22.217	0.043
IS1008d	26.25	Dev	21.376	4.874
Average	32.74	-	31.29	-
Average improvement	**1.44**			

Table 4.6 Comparison of Diarization error rate (%) score with Multimodal speaker diarization (MMSD) system.

Meeting ID	**MMSD** [45]	**Proposed** **Multimodal**	**Difference** **(Improvement)**
IS	21.68	22.24	-0.56

4.3 Speech Enhancement

Speech enhancement technique is used to in proposed multimodal speaker diarization system which effectively increased the accuracy of the system. This enhancement technique is based on densely connected progressively learning LSTM network.

4.3.1 Results

Table 4.7 presents the results for multimodal speaker diarization system for AWGN based noisy speech and its enhancement via Wiener filtering and LSTM model. The results indicate that in the presence of AWGN noise, Wiener filtering performs better than the LSTM based model with average improvement of just 0.76%. This is due to the fact that Wiener filtering technique tracks stationary noise more robustly. However, both models significantly improved the DER of the system. On average, LSTM model improved the DER by 2.07% and Wiener filter by 2.83%. Any negative value in the improvement column indicates the degradation in the performance.

Table 4.8 presents the results of the second experiment with realistic environmental noise. Both noisy and enhanced speech provided by the two models i.e. Wiener and LSTM are presented in this table. The improvement column represents that the performance of LSTM based is very significant, while the performance of Wiener filtering is poor in case of realistic environmental noise. On average LSTM model provided 11.6% improvement and Wiener filtering degraded the performance by 1.23% in diarization error rate of the system. Maximum improvement provided by an LSTM system is in the case of IS1008b recording i.e. 28.22%.

Finally, Table 4.8 presents the training MFCC frames for each speaker acquired from SyncNet inference. These are only high confidence frames acquired by applying the two thresholds. The results show that in the presence of the environmental noise the SyncNet inference does not correctly recognize the audio-visual synchronization. Hence, a smaller number of synchronized frames are acquired, which results in smaller training samples. When speech enhancement technique has been applied, a good number of high confidence frames are acquired for each speaker. This table only presents the samples of three recordings, other recordings follow the similar behavior.

Table 4.7 Diarization error rate (%) for AWGN noisy speech and its enhancement.

Meeting ID	**Noisy Speech**	**LSTM Denoising**	**Wiener Denoising**	**Improvement (Noisy-LSTM)**	**Improvement (Noisy-Wiener)**
IS1000a	42.65	38.65	34.05	4.00	8.6
IS1001a	45.67	45.52	38.25	0.15	7.42
IS1001b	42.89	42.24	39.99	0.65	2.90
IS1001c	39.96	35.12	38.00	4.84	1.96
IS1003b	25.82	25.46	26.83	0.36	-1.01
IS1003d	53.80	52.06	52.11	1.74	1.69
IS1006b	49.38	49.58	47.53	-0.20	1.85
IS1006d	65.35	60.22	60.07	5.13	5.28
IS1008a	17.07	13.22	14.18	3.85	2.89
IS1008b	16.39	14.77	18.31	1.62	-1.92
IS1008c	33.25	31.25	29.15	2.00	4.10
IS1008d	24.09	23.44	23.93	0.65	0.16
Average	**38.03**	**35.96**	**35.20**	**-**	**-**

Table 4.8 Diarization error rate (%) for Environmental noisy speech and its enhancement.

Meeting ID	**Noisy Speech**	**LSTM Denoising**	**Wiener Denoising**	**Improvement (Noisy-LSTM)**	**Improvement (Noisy-Wiener)**
IS1000a	54.15	33.61	57.8	20.54	-3.65
IS1001a	41.11	40.66	42.04	0.45	-0.93
IS1001b	52.71	43.51	49.14	9.2	3.57
IS1001c	50.61	38.2	51.06	12.41	-0.45
IS1003b	50.38	31.7	45.44	18.68	4.94
IS1003d	61.74	52.82	63.82	8.92	-2.08
IS1006b	61.22	54.41	67.92	6.81	-6.7
IS1006d	69.26	65.44	68.94	3.82	0.32
IS1008a	37.1	14.92	51.28	22.18	-14.18
IS1008b	48.98	20.76	44.93	28.22	4.05
IS1008c	32.64	25.63	30.22	7.01	2.42
IS1008d	24.84	23.84	26.92	1	-2.08
Average	**48.72**	**37.12**	**49.95**	**-**	**-**

Table 4.9 Training samples (MFCC frames) for noisy and enhanced speech for each speaker acquired from SyncNet inference.

	Noisy recordings (Environmental)				**LSTM Speech Enhancement**			
Meeting ID	**Speaker 1**	**Speaker 2**	**Speaker 3**	**Speaker 4**	**Speaker 1**	**Speaker 2**	**Speaker 3**	**Speaker 4**
IS1000a	3384	4120	508	5864	4228	13700	2960	4920
IS1003d	3296	2688	2160	1084	13296	12112	11984	2804
IS1008a	200	5068	6156	4932	2064	7056	9952	6292

4.4 Summary

This chapter described the results of each contribution in this thesis. The following paragraph describes the summary of the technique and results in detail.

The first section presented the results for feature embedding technique which is based on deep autoencoders, followed by agglomerative hierarchical clustering (AHC) for diarization. Such embeddings are compared with conventional speech features i.e. MFCC. Moreover, feature embeddings are acquired by grouping five consecutive frames of MFCC. The experiments showed that those feature embeddings improved the DER of the speaker diarization system as compared to MFCC features. In particular, the acquired features significantly improved the accuracy by reducing the average DER.

The second section presented results of multimodal diarization technique which utilized a pre-trained SyncNet model. The audio pipeline applies feature extraction, speech activity detection and finally clustering technique for speaker diarization. Only high confidence MFCC frames acquired through SyncNet inference are used to train GMM models and then rest of the audio frames are clustered on the basis of maximum likelihood. The video pipeline apply face detection to crop face only region, remove silence frames, apply shot detection and finally split the video into 2-sec short segments. These short video segments are then provided to pre-trained SyncNet model which runs the inference on them. On the output results of this inference we apply two thresholds to select those video segments which are confident enough on the active speaker in the video and its respective audio domain. When compared with the audio based diarization techniques effectiveness of such novel multimodal diarization technique is noticed. The main advantage of our proposed method is in providing pure clusters which are trained on frames belonging to the single speaker. While in conventional diarization approach which uses agglomerative hierarchical clustering technique, there is greater chance of impurity in term of merging clusters that have voice of multiple speakers. Our technique robustly prevents any such impurity, that is based on the threshold selections for offset and confidence metrics in video domain. Beside this, our proposed technique is fully unsupervised and doesn’t even require any out of domain data for training purpose. Furthermore, proposed technique is applicable on meeting recordings, TV/ talk shows and movies where speakers face the camera most of the time. Finally, it is also concluded that such technique is similar in accuracy to one of the state-of-the-art multimodal approach.

Finally, the third section presented the results for speech enhancement module which is based on LSTM model. This model is trained to jointly learn IRM and LPS to enhance the noisy speech. Previously proposed techniques focused to use speech enhancement in audio-based speaker diarization system. Whereas the current study specifically focused on the use of speech enhancement module for multimodal speaker diarization system. This multimodal technique is based on the use of audio-visual synchronization model to find active speaker. So, enhancement speech eventually affects both audio and video pipelines of the multimodal system. Finally, it was evaluated that for high noisy audio recordings the speech enhancement significantly improves the diarization error rate of multimodal speaker diarization system.

Chapter 5.

Conclusion and Future Work

This chapter draws the conclusion of each study based on their materials, methods and results. These studies consist of feature embedding method, multimodal speaker diarization system and speech enhancement for multimodal speaker diarization. This chapter also describe future research recommendations, possibilities and gaps.

5.1 Conclusion

5.1.1 Feature embedding

Following points concludes the feature embedding technique which is based on deep autoencoders. These embeddings were utilized in an audio based unsupervised speaker diarization system which is based on agglomerative hierarchical clustering.

- Proposed feature embedding method is completely unsupervised and do not require data other than given recordings for diarization. Such technique matches with the unsupervised nature of speaker diarization system.
- These embedding are extracted by training deep autoencoders in a way that five consecutive speech frames are grouped together to learn low dimensional features. The proposed architecture also acts as non-linear PCA and acquired features are in compressed domain.

- Compared with the benchmark MFCC features, feature embedding performed significantly better in term of DER. On the popular subset of AMI corpus, average diarization error rate of feature embedding technique is 2.96% better than the MFCC features, which validates the effectiveness of the proposed technique. However, maximum improvement of 8.05% is noticed in one of the recording.

5.1.2 Multimodal speaker diarization

A novel multimodal diarization technique is proposed, based on a pre-trained audio-visual synchronization model. Both audio and video processing techniques have been proposed in this research to apply diarization. Following points concludes this research.

- Using a pre-trained model to find the audio-visual synchronization for speaker diarization improves the accuracy as compared to audio-based speaker diarization.
- Such technique which utilizes a pre-trained model has less computation complexity as compared to other state-of-the-art multimodal techniques.
- By applying threshold on the acquired synchronized segments through SyncNet, high confidence frames were acquired. Clusters trained on these high confidence frames have more probability of being pure clusters because they contain speech frames of only single speaker.
- This simple yet effective multimodal technique is completely unsupervised and matches the nature of unsupervised speaker diarization system.
- The performance of the system has been tested with several benchmarks i.e. unsupervised audio speaker diarization, fully supervised audio speaker diarization and multimodal speaker diarization systems.

- Compared to the unsupervised audio speaker diarization, the performance of multimodal speaker diarization is very significant i.e. average DER improvement of 13.58%.
- Fully supervised diarization provides a really challenging comparison because each pipeline of the benchmark is fully supervised with a subset of AMI corpus. Compared to this benchmark, proposed method even supersedes with average DER improvement of 1.44%.
- Compared to the SOTA multimodal speaker diarization system proposed system has almost equivalent performance with just 0.56% average impairment on complete IS dataset.

5.1.3 Speech Enhancement

LSTM based speech enhancement model was proposed to use in multimodal speaker diarization system.

- Speech enhancement model was trained to learn multiple targets through progressive learning approach with more than 100 noise types.
- This model enhances the noisy speech input which was further used in audio and video pipelines.
- The enhanced speech helps in better acquisition of audio-visual synchronized speech through SyncNet architecture. Moreover, enhanced speech provides better likelihood of speech frames from each cluster. Thus, it reduces the diarization error rate significantly when input speech contains the Gaussian and realistic environmental noises.

- Experiments were performed with two types of noises i.e. AWGN and environmental noise, taken from PL100 dataset and two types of denoising i.e. Wiener filtering and LSTM model. The results show that the speech enhancement technique using LSTM based model significantly improves the performance of multimodal speaker diarization system as compared to the Wiener filtering. Moreover, Wiener filtering may perform better than LSTM model in case of AWGN noise but not in the case of realistic environmental noise. On average LSTM model improved the DER of multimodal speaker diarization system by 2.07% for AWGN and 11.6% for environmental noise.

5.2 Future recommendations

Following are some future recommendations and research directions for multimodal speaker diarization system which could help to further improve the system.

- In proposed multimodal speaker diarization system, one can exploit speech source localization (SSL) technique to find active speaker along with the audio clustering. This would help to reduce the speaker assignment error. For this purpose, AMI corpus can be utilized which provides the audio recordings of microphone array consisting of eight channels.
- Using speech source localization, one can also assign overlapping speech segments to multiple speakers.
- One can also exploit the audio-video features fusion technique for multimodal speaker diarization systems to reduce the complexity of the system. For that

purpose, models similar to bimodal deep architectures [85] can be further exploited.

- One possible direction is the development of end to end deep learning model for multimodal speaker diarization.

BIBLIOGRAPHY

[1] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” *2003 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2003*, pp. 411–416, 2003.

[2] X. Anguera, C. Wooters, and J. Hernando, “Automatic cluster complexity and quantity selection: Towards robust speaker diarization,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) Springer, Berlin, Heidelb.*, vol. 4299 LNCS, pp. 248–256, 2006.

[3] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 4625 LNCS, pp. 509–519.

[4] G. Friedland *et al.*, “The ICSI RT-09 Speaker Diarization System,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 2, pp. 371–381, Feb. 2012.

[5] X. Anguera, C. Wooters, and J. M. Pardo, “Robust speaker diarization for meetings: ICSI RT06s evaluation system,” *INTERSPEECH 2006 9th Int. Conf. Spok. Lang. Process. INTERSPEECH 2006 - ICSLP; Pittsburgh, Pennsylvania, USA*, vol. 4, no. October, pp. 1674–1677, 2006.

[6] C. Wooters, J. Fung, B. Peskin, and X. Anguera, “TOWARDS ROBUST SPEAKER SEGMENTATION : THE ICSI-SRI FALL 2004 DIARIZATION SYSTEM International Computer Science Institute , Berkeley , CA , U . S . A . Polytechnical University of Catalonia (UPC), Barcelona , Spain,” *Fall 2004 Rich Transcr. Work.*,

p. 23, 2004.

[7] J. Carletta *et al.*, “The AMI Meeting Corpus: A Pre-announcement Machine Learning for Multimodal Interaction,” in *Machine Learning for Multimodal Interaction, Bethesda, MD, USA, 1-4 May*, 2006, vol. 3869, pp. 28–39.

[8] X. Anguera *et al.*, “Speaker Diarization : A Review of Recent Research,” *Lang. Process.*, vol. 1, no. August, pp. 1–15, 2010.

[9] K. J. Han and S. S. Narayanan, “Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH Brisbane, Australia, 22–26 September*, 2008, pp. 20–23.

[10] S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion,” *Proc. DARPA Broadcast News Transcr. Underst. Work.*, vol. 6, pp. 67–72, 1998.

[11] J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, “Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2006, vol. 5.

[12] W. H. Tsai, S. S. Cheng, and H. M. Wang, “Speaker clustering of speech utterances using a voice characteristic reference space,” in *8th International Conference on Spoken Language Processing, ICSLP 2004*, 2004, pp. 2937–2940.

[13] S. B. Davis and P. Mermelstein, “Comparison of Parametric Representations for

Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4. pp. 357–366, 1980.

[14] H. Bredin and G. Gelly, "Improving Speaker Diarization of TV Series using Talking-Face Detection and Clustering," *Proc. 2016 ACM Multimed. Conf. - MM '16*, pp. 157–161, 2016.

[15] V. Peruffo Minotto, C. Rosito Jung, and B. Lee, "Multimodal Multi-Channel On-Line Speaker Diarization Using Sensor Fusion Through SVM," *IEEE Trans. Multimed.*, vol. 17, no. 10, pp. 1694–1705, Oct. 2015.

[16] H. Bredin, C. Barras, and C. Guinaudeau, "Multimodal person discovery in broadcast TV at MediaEval 2016," *CEUR Workshop Proc.*, vol. 1739, pp. 2–4, 2016.

[17] N. Sarafianos, T. Giannakopoulos, and S. Petridis, "Audio-visual speaker diarization using fisher linear semi-discriminant analysis," *Multimed. Tools Appl.*, vol. 75, no. 1, pp. 115–130, 2016.

[18] X. Bost, G. Linares, and S. Gueye, "Audiovisual speaker diarization of TV series," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015-Augus, pp. 4799–4803, 2015.

[19] E. El Khoury, C. Sénac, and P. Joly, "Audiovisual diarization of people in video content," *Multimed. Tools Appl.*, vol. 68, no. 3, pp. 747–775, 2014.

[20] N. Seichepine, S. Essid, C. Fevotte, and O. Cappe, "Soft nonnegative matrix co-

factorizationwith application to multimodal speaker diarization," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 3537–3541, 2013.

[21] N. Dehak, P. Kenny, and R. Dehak, "Front-end factor analysis for speaker verification," *Audio, Speech, and*, vol. 19, no. 4, pp. 1–12, May 2010.

[22] S. Madikeri, I. Himawan, P. Motlicek, and M. Ferras, "Integrating online i-vector extractor with information bottleneck based speaker diarization system," in *Annual Conference of the International Speech Communication Association, INTERSPEECH, Dresden, Germany, 6–10 September*, 2015, vol. 2015-Janua, pp. 3105–3109.

[23] Y. Xu, I. Mcloughlin, and Y. Song, "Improved i-vector representation for speaker diarization," *Circuits, Syst. Signal Process.*, vol. 35, no. 9, pp. 3393–3404, 2015.

[24] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, New Orleans, LA, USA, 5–9 May*, 2017, pp. 4930–4934.

[25] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings, South Lake Tahoe, NV, USA, 7–10 December*, 2014, pp. 413–417.

[26] S. Meignier, P. Del, Y. Est, and L. Mans, "Recent Improvements on ILP-based Clustering for Broadcast News Speaker Diarization," no. June, pp. 187–193, 2014.

[27] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks - FEATURES LEARNT BY CONVOLUTIONAL NETWORKS,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2014, vol. 8689 LNCS, no. PART 1, pp. 818–833.

[28] A. Jati and P. Georgiou, “Speaker2Vec: Unsupervised learning and adaptation of a speaker manifold using deep neural networks with an evaluation on speaker segmentation,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, vol. 2017-Augus, pp. 3567–3571.

[29] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in *2016 IEEE Workshop on Spoken Language Technology, SLT 2016 - Proceedings*, 2017, pp. 165–170.

[30] S. Dey, T. Koshinaka, P. Motlicek, and S. Madikeri, “DNN Based Speaker Embedding Using Content Information for Text-Dependent Speaker Verification,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 5344–5348, 2018.

[31] H. Delgado, X. Anguera, C. Fredouille, and J. Serrano, “Fast Single- and Cross-Show Speaker Diarization Using Binary Key Speaker Modeling,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 23, no. 12, pp. 2286–2297, 2015.

[32] S. Madikeri, D. Imseng, and H. Bourlard, “Improving Real Time Factor of Information Bottleneck-based Speaker Diarization System,” *Idiap*, 2015.

[33] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, “Speaker diarization with LSTM,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 5239–5243, 2018.

[34] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April, pp. 4879–4883.

[35] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, “Fully Supervised Speaker Diarization,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 6301–6305, Oct. 2019.

[36] P. Cyrta, T. Trzciński, and W. Stokowiec, “Speaker diarization using deep recurrent convolutional neural networks for speaker embeddings,” in *Advances in Intelligent Systems and Computing*, 2018, vol. 655, pp. 107–117.

[37] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, “Acoustic modelling from the signal domain using CNNs,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, vol. 08-12-Sept, pp. 3434–3438.

[38] M. Rouvier, P. M. Bousquet, and B. Favre, “Speaker diarization through speaker embeddings,” in *2015 23rd European Signal Processing Conference, EUSIPCO 2015*, 2015, pp. 2082–2086.

[39] G. Sell *et al.*, “Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural dihard challenge,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, vol. 2018-Septe, pp. 2808–2812.

[40] N. Le and J.-M. Odobez, “Learning Multimodal Temporal Representation for Dubbing Detection in Broadcast Media,” *Proc. 2016 ACM Multimed. Conf. - MM ’16*, pp. 202–206, 2016.

[41] B. G. Gebre, P. Wittenburg, T. Heskes, and S. Drude, “Motion history images for online speaker/signer diarization,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014, pp. 1537–1541.

[42] F. Bechet *et al.*, “Multimodal understanding for person recognition in video broadcasts,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. September, pp. 607–611, 2014.

[43] A. Noulas, G. Englebienne, and B. J. A. Kröse, “Multimodal Speaker diarization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 79–93, Jan. 2012.

[44] I. Kapsouras, A. Tefas, N. Nikolaidis, G. Peeters, L. Benaroya, and I. Pitas, “Multimodal speaker clustering in full length movies,” *Multimed. Tools Appl.*, vol. 76, no. 2, pp. 2223–2242, 2017.

[45] P. Cabañas-Molero, M. Lucena, J. M. Fuertes, P. Vera-Candeas, and N. Ruiz-Reyes, “Multimodal speaker diarization for meetings using volume-evaluated SRP-PHAT and video analysis,” *Multimed. Tools Appl.*, vol. 77, no. 20, pp. 27685–27707, Oct.

2018.

[46] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1086–1099, 2018.

[47] M. Ferràs, S. Masneri, O. Schreer, and H. Bourlard, "Diarizing large corpora using multi-modal speaker linking," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. September, pp. 602–606, 2014.

[48] Y. Komai, Y. Ariki, and T. Takiguchi, "Audio-visual speech recognition based on AAM parameter and phoneme analysis of visual feature," in *Pacific-Rim Symposium on Image and Video Technology, Gwangju, South Korea, 20-23 November*, 2011, no. PART1, pp. 97–108.

[49] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," in *Proceedings of the IEEE*, 2003, vol. 91, no. 9, pp. 1306–1325.

[50] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[51] Z. Barzelay and Y. Y. Schechner, "Onsets coincidence for cross-modal analysis," *IEEE Trans. Multimed.*, vol. 12, no. 2, pp. 108–120, Feb. 2010.

[52] J. W. Fisher, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information*

Processing Systems, Vancouver, BC, Canada, 3–8 December, 2001, pp. 772–778.

[53] M. R. Siracusa and J. W. Fisher, “Dynamic dependency tests for audio-visual speaker association,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Honolulu, HI, USA, 15–20 April*, 2007, vol. 2, pp. II-457-II–460.

[54] A. K. Noulas and B. J. A. Krose, “On-line multi-modal speaker diarization,” in *9th International Conference on Multimodal Interfaces, ICMI’07, Nagoya, Aichi, Japan, 12–15 November*, 2007, pp. 350–357.

[55] H. J. Nock, G. Iyengar, and C. Neti, “Speaker localisation using audio-visual synchrony: An empirical study,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2728, pp. 488–499, 2003.

[56] G. Friedland, H. Hung, C. Yeo, and U. C. Berkeley, “MULTI-MODAL SPEAKER DIARIZATION OF REAL-WORLD MEETINGS USING COMPRESSED-DOMAIN VIDEO FEATURES,” in *International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19-24 April*, 2009, pp. 4069–4072.

[57] G. Garau, A. Dielmann, and H. Bourlard, “Audio-visual synchronisation for speaker diarisation,” in *11th Annual Conference of the International Speech Communication Association, INTERSPEECH, Makuhari, Japan, 26–30 September*, 2010, pp. 2654–2657.

[58] J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in *Asian conference on computer vision, Taipei, Taiwan, 20–24 November*, 2017, vol.

10117 LNCS, pp. 251–263.

[59] J. Benesty, J. Chen, Y. Huang, and I. Cohen, “Spectral Enhancement Methods,” in *In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing, vol 2. Springer, Berlin, Heidelberg*, 2009, pp. 1–30.

[60] L. Sun *et al.*, “Speaker diarization with enhancing speech for the first DIHARD challenge,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, vol. 2018-Septe, pp. 2793–2797.

[61] L. Sun *et al.*, “A Novel LSTM-Based Speech Preprocessor for Speaker Diarization in Realistic Mismatch Conditions,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April, pp. 5234–5238.

[62] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 7092–7096.

[63] A. Narayanan and D. L. Wang, “Investigation of speech separation as a front-end for noise robust speech recognition,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, no. 4, pp. 826–835, Apr. 2014.

[64] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 436–

440.

[65] Y. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[66] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[67] T. Gao, J. Du, L. R. Dai, and C. H. Lee, “SNR-based progressive learning of deep neural network for speech enhancement,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 3713–3717, 2016.

[68] T. Gao, J. Du, L. R. Dai, and C. H. Lee, “Densely Connected Progressive Learning for LSTM-Based Speech Enhancement,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April, pp. 5054–5058.

[69] W. Zhu, W. Guo, and G. Hu, “Feature mapping for speaker diarization in noisy conditions,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 5445–5449.

[70] L. Sun, J. Du, L. R. Dai, and C. H. Lee, “Multiple-target deep learning for LSTM-RNN based speech enhancement,” in *2017 Hands-Free Speech Communications and Microphone Arrays, HSCMA 2017 - Proceedings*, 2017, pp. 136–140.

[71] Y. Xu *et al.*, “Unsupervised Feature Learning Based on Deep Models for

Environmental Audio Tagging," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1230–1241, Jun. 2017.

[72] E. Gonina, G. Friedland, H. Cook, and K. Keutzer, "Fast speaker diarization using a high-level scripting language," in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Waikoloa, HI, USA, 11–15 December*, 2011, pp. 553–558.

[73] S. Molau, M. Pitz, R. Schlüter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, , Salt Lake City, Utah, USA, 7-11 May*, 2001, vol. 1, pp. 73–76.

[74] M. P. Cuéllar, M. Delgado, and M. C. Pegalajar, "An application of non-linear programming to train Recurrent Neural Networks in Time Series Prediction problems," in *ICEIS 2005 - Proceedings of the 7th International Conference on Enterprise Information Systems*, 2005, pp. 35–42.

[75] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, 2010.

[76] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series with engineering applications*. Technology Press of the Massachusetts Institute of Technology, 1964.

[77] R. Yin, H. Bredin, and C. Barras, "Neural speech turn segmentation and affinity

propagation for speaker diarization," in *Annual Conference of the International Speech Communication Association, INTERSPEECH, Hyderabad, India, 2–6 September*, 2018, vol. 2018-Septe, pp. 1393–1397.

[78] R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast TV using bidirectional long short-term memory networks," in *Annual Conference of the International Speech Communication Association, INTERSPEECH, Stockholm, Sweden, 20–24 August*, 2017, pp. 3827–3831.

[79] A. Graves, N. Jaitly, and A. R. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU, Olomouc, Czech Republic, 8–12 December*, 2013, pp. 273–278.

[80] H. Bredin, "TristouNet: Triplet loss for speaker turn embedding," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 May*, 2017, pp. 5430–5434.

[81] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science (80-.).*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[82] I. D. Gebru, S. Ba, G. Evangelidis, and R. Horaud, "Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model," in *IEEE International Conference on Computer Vision*, 2016, vol. 2016-Febru, pp. 702–708.

[83] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-Localization of Audio Sources in Images Using Binaural Features and Locally-Linear Regression," *IEEE*

Trans. Audio, Speech Lang. Process., vol. 23, no. 4, pp. 718–731, 2015.

[84] D. E. King, “Dlib-ml: A Machine Learning Toolkit,” *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[85] J. Ngiam, A. Khosla, and M. Kim, “Multimodal deep learning,” in *28th international conference on machine learning*, 2011, pp. 689–696.