SPEAKER DIARIZATION IN BROADCAST NEWS

BY

MS. PANTID CHANTANGPHOL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING (INFORMATION AND COMMUNICATION TECHNOLOGY FOR EMBEDDED SYSTEM)

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2020

SPEAKER DIARIZATION IN BROADCAST NEWS

BY

MS. PANTID CHANTANGPHOL

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING (INFORMATION AND COMMUNICATION TECHNOLOGY FOR EMBEDDED SYSTEM)

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THAMMASAT UNIVERSITY

ACADEMIC YEAR 2020

THAMMASAT UNIVERSITY

SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THESIS

BY

MS. PANTID CHANTANGPHOL

ENTITLED

SPEAKER DIARIZATION IN BROADCAST NEWS

was approved as partial fulfillment of the requirements for
the degree of Master of Engineering (Information and Communication Technology for Embedded Systems)

on September 25, 2020

Chairperson	(Jessada Karnjana, Ph.D.)
Member and Advisor	(Assistant Professor Sasipom Usanavasin, Ph.D.)
Member	(Natsuda Kaothanthong, Ph.D.)
Member	(Associate Professor Takahiro Shinozaki, Ph.D.)
Director	(Professor Pruettha Nanakom, D.Eng.)

Thesis Title	SPEAKER DIARIZATION IN BROADCAST NEWS
Author	Ms. Pantid Chantangphol
Degree	Master of Engineering (Information and Communication Technology for Embedded Systems)
Faculty/University	Sirindhorn International Institute of Technology/ Thammasat University
Thesis Advisor	Associate Professor Sasiporn Usanavasin, Ph.D.
Thesis Advisor	Jessada Karnjana, Ph.D.
Thesis Advisor	Associate Professor Takahiro Shinozaki, Ph.D.
Academic Years	2020

ABSTRACT

Speaker Diarization is a multimedia indexing technology that makes use of audio information to answer the question “Who spoke when?” This thesis presents a step-by-step of speaker diarization system implemented in python-based that is evaluated using the Diarization Error Rate (DER) metric.

The proposed system, designed for segmenting audio recordings of broadcast news, provides implementations of a combination feature extraction based on Dense Convolutional Network (DenseNet) for segmenting the speech according to speaker id with various background noise. This clustering algorithm offer lower DER as well as a computational advantage compared to the other classifier. The proposed speaker diarization achieves a favorable performance on Hollywood movie dataset, AVA speech dataset, CALLHOME American English, 2003 NIST Rich Transcription and the 2000 NIST Speaker Recognition Evaluation compared to the supervised speaker diarization of the speaker diarization with LSTM.

Keywords: Speech activity detection, Convolutional Neural Network, DenseNet, Deep learning, Feature combination, Fusion Mode

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Sirindhorn International Institute of Technology (SIIT) for the past two years I have enjoyed.

I would like to thank my supervisor, Assistant Professor Sasiporn Usanavasin, Jessada Karnjana and Associate Professor Takahiro Shinozaki for the constant support and guidance over the past year and a half. They were always available to clear any doubts or queries and provided constructive criticism about my approach and reports.

Special thanks are reserved for my friends who made the two years memorable. I would also like to thank my classmates and colleagues for being there to discuss ideas and for making grate conversation.

Last, but not the least, I would like to thank my family for giving me tons of support, more than I could have asked for.

Ms. Pantid Chantangphol

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

Symbols/Abbreviations	Terms
ASR	Automatic Speech Recognition
BIC	Bayesian Inference Criterion
DenseNet	Dense Convolutional Network
DER	Diarization Error Rate
DFT	Discrete Fourier Transform
DS theory	Dempster-Shafer theory
FASR	false alarm speech rate
GMMs	Gaussian Mixture Models
HAC	Hierarchical Agglomerative Clustering
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
HRI	Human-Robot Interaction
ILP	Integer Linear Program
KL divergence	Kullback–Leibler divergence
LFCC	Linear Frequency Cepstral Coefficients
Log-mel	Log-mel spectrogram
LPCCs	Linear Prediction Cepstral Coefficients
MFCC	Mel-frequency cepstral coefficients
MSR	Missed Speech Rate
NCLR	Normalized Cross Likelihood Ratio
PLP	Perceptual Linear Prediction
ResNet	Residual neural network
SAD	Speech Activity Detection
SAM	Subtitle-Aligned Movie
STFT	Short-Time Fourier Transform

SVMs	Support Vector Machines
UBM	Universal Background Model

CHAPTER 1
INTRODUCTION

The number of multimedia uploads on the internet every day is ever-increasing, especially after smart phones and smart devices have gained popularity in the recent years. Multimedia powerhouses and online search engines get trillions of searches queries every day. Search engines now also facilitate searching through audio, image and video databases. Keeping track of all the data and storing it efficiently for possible future access is becoming more and more important. Hence indexing techniques in multimedia formats are getting attention. Audio segmentation is one such indexing technique where the content in a collection of multimedia recordings is organized based on the semantic information provided through their audio data.

Speaker diarization is one such audio indexing problem where the question being asked to the machine is “Who spoke when?” Speaker diarization acts as a precursor to many other speech technologies. For instance, performing automatic meeting transcriptions, automatic stenographers, Dictaphones all would function faster and more efficiently if the machine knows the active speaker at any given instant.

When we were in the meeting and discussion, we need to use speech to clarify our thinking to others. And we need secretary to do the meeting records. If computers can generate the meeting record automatically, it will save the time. Speaker diarization is one audio partitioning problem where answer the question “Who spoke when?” in the audio file. Speaker diarization will be a pre-processing to other speech technologies. The speech technology will faster and more efficiently if the system knows the active speaker at the moment. The Speaker diarization have been the topic of choice of many researchers in recent years. Speaker diarization is mainly applied to enable smooth Automatic Speech Recognition (ASR) results. It also makes smooth communication between computers and their users as an attempt to give computers the ability to recognize the speaker and offer the meeting record.

2.1 State of problem

It has often been said that “Speech is the mirror of the mind”. This statement means the speech will reflect the hidden thinking of the human. It is not the only speech that is the key of communication. Pose, Facial expression, and actions are also essential factors for communication. But the speech is a higher importance factor because it is easily understanding.

In the meeting scenario, humans can understand the others’ thinking by speaking. If we find the practical solutions for automatic recognition of speech, we can generate the meeting automatically. In recent year, the accuracy of automatic speech recognition will be increased. The hearing ability of computer will be equal to human perception. But the computer cannot generate the text from speech in the discussion because the computer does not know the speaker. It does not know which part of the speech comes from different speaker. It needs to do speech partitioning to answer the question “who spoke when” in this meeting. It was called “Speaker diarization”.

In the first place, the speaker diarization was a research topic for the text mining in the tv drama. After the speech technology and automatic speech recognition have developed, the research works on speaker diarization are interesting. In the past, a lot of research works focus on speaker diarization with speech analysis technology, which is energy based. Recently, recurrent neural network architecture is used for speaker diarization to improve the accuracy of diarization error rate (DER). After that, convolutional neural network technology has become a tool to implement the algorithm for speech processing applications including facial speaker diarization. To achieve better performance, some researchers have adjusted their approaches by using fewer feature points or by classifying a neutral image with a neural network before the classification phase. If we can solve the problem by identifying a set of essential feature points and by using genuine feature detection techniques with limitation resources, the overall performance can be increased.

The research problem of speaker diarization has been applied in various domains – Telephone conversations, meeting recordings and broadcast news archiving. Although the problem remains the same, the design criteria for each of these are different due to the nature of the audio recording and the variability they possess. The

peculiarities of a typical recording from each of three domains are comparatively mentioned below.

Table 1.1 Analytic criteria of speaker diarization systems in various domains.

	Broadcast news	Meeting conversations	Telephone conversations
Duration of uninterrupted speech	Long	Short	Short
Neglible speaker overlap	Negligible speakers overlap	Higher speakers overlap	Moderate speakers overlap
Variety of background noise	Presence of music, jingles, variety of background noise	Uniform background conditions	Uniform background conditions
Dominant speaker	Yes	No	No
Number of speakers	Unknown	Unknown	Known (two person)

1.2 Purpose of study

The purpose of this study is to introduce a speaker diarization in broadcast news domain based on convolutional neural network to develop a model to partitioning and identifying speaker in the audio files in term of the start and stop time with speaker id. And to determine the suitable speech processing techniques and artificial intelligence techniques for speaker diarization. The focus of this thesis is on solving speaker diarization for the broadcast news domain. A system has been developed in Python that was tested on five datasets, which are the Hollywood movie dataset, AVA speech and active speaker dataset, CALLHOME American English, 2003 NIST Rich Transcription and 2000 NIST Speaker Recognition Evaluation. The system has 3 sub processes, which are the speech activity detection, the speaker change detection and the speaker clustering.

1.3 Significance of study

The finding of this study will have the potential to significantly improve speaker diarization system performances. It will demonstrate the value of speech processing and artificial intelligence to solve a challenging problem of speaker diarization. Nevertheless, if we build a humanoid robot that can interact with a human, the speaker diarization problems must be solved. So, a computer for real-world applications should

have the ability to recognize speech from different speaker. The medical robot, which is designed to provide care for the disabled or unhealthy person, should recognize the patient's current information, in order to adjust personalized strategies for patient care and patient interaction. In the case of educational tools, the speaker diarization can identify students' understanding and emotion of the entire distance learning process. For personal robots, the speaker diarization is important for human-robot interaction (HRI) to recognize the user’s speech and handle user’s reaction properly. The facial expression recognition will be improving the quality of human-computer interaction (HCI) by making the interaction between the system and human to be like human-human interaction. That can create more productive and neutral working relationships.

CHAPTER 2
REVIEW OF LITERATURE

In recent years, speaker diarization is an interesting topic that has attracted interest from researchers. Research on speaker diarization involves many fields such as image processing, video processing, and speech processing. It also has various applications in the areas such as a robot, smart car, smart home, intelligent assistant, educational toy, mobile-based service, and other interactive systems. In the interaction system, it is necessary to make the system can understand the user's speech and provide an appropriate response.

The speaker diarization is a process of recognizing human thinking by using speech processing for speech, image processing for the handwriting, video processing for facial expression and body gesture. But the speech is the most popular research topic in the speaker diarization area. As a technology tool, the speaker diarization will make the interaction system easier, more functional and more natural. The following are examples of the past researches that involve speaker diarization.

The problem of speaker diarization involves answering “Who spoke when”. It is generally broken down into answering “is anyone speaking?” and then answering “which speaker in the audio is speaking?” The first step is called speech activity detection, which is a pre-processing step common in speaker recognition, speech recognition, speech coding and speech enhancement (Miró & Bozonnet, 2012). The latter problem can be approached as finding the change in speaker which is called speaker segmentation, and then combining the contiguous segments belonging to the same speaker under a unique label ,which is called speaker clustering.

Initially in the late 1990’s, when research in diarization was still in its nascent stages, few systems attempted to perform speech activity detection as a by-product of the segmentation and clustering (Miró et al., 2012). Nonspeech was thought to be just another speaker. But owing to the acoustic variability of nonspeech, systems with explicit speech activity detectors performed much better. Often, the speaker segmentation and speaker clustering are performed iteratively (Friedland et al., 2012).

2.1 Evaluation for speaker diarization

The evaluation of a diarization system is done using a metric called the Diarization Error Rate (DER) (Miró, Wooters & Pardo, 2006), which is the percentage of the time of the audio for which the speaker was wrongly labelled. The output of the system is compared with a segment level manually annotated temporal transcription indicating the speaker labels.

2.1.1 Meeting diarization

Equation (2.1) was provided for the NIST RT speaker diarization evaluations (Miró, Wooters and Pardo, 2006). In these evaluations, competitor systems had to perform speaker diarization on meeting conversation recordings. The meeting diarization is given by the following equation:

$$DER = \frac{\sum_{s=1}^{S} dur(s) \times (\max(N_{ref}, N_{hyp}) - N_{correct})}{\sum_{s=1}^{S} dur(s) \times N_{ref}} \quad (2.1)$$

Where N_{ref} and N_{hyp} are the number of speakers indicated by the annotations, hypothesized by the system respectively, and $N_{correct}$ is the number of speakers in segment s that were a correct match between the annotation and hypothesis.

2.1.2 Broadcast news diarization

In broadcast news, there is very little overlap. For the case where overlap is absent, the formula for DER can be simplified. This error calculation is used in the broadcast news diarization systems. Pyannote (Bredin, 2017) is a python based toolkit that facilitates calculation specifically for such systems.

$$S = C_1 + T_1 + T_3 \quad (2.2)$$

Where S is total speech time, C_1 is correctly labelled speech time, T_1 is Missed Speech time and T_3 is incorrectly labelled speech time.

$$NS = C_2 + T_2 \quad (2.3)$$

Where NS is total non-speech time, C_2 is correctly labelled non-speech time and T_2 is false alarm speech time.

The DER can be broken down systematically into 2 types. Consider an audio with speech and non-speech as indicated by the annotation. Non-speech includes silences, speaker pauses, music, jingles, noise etc. The two categories can be further classified exhaustively as shown in Equation (2.2) and Equation (2.3). Missed speech time is the time when the algorithm erroneously indicated a segment as non-speech. False alarm speech time, on the other hand, is the time when the algorithm erroneously indicated a segment as speech.

$$E_1 = \frac{T_1 \times 100}{S} \tag{2.4}$$

Where E_1 is missed speech rate (MSR), T_1 is missed speech time and S is total speech time.

$$E_2 = \frac{T_2 \times 100}{S} \tag{2.5}$$

Where E_2 is false alarm speech rate (FASR) and T_2 is false alarm speech time and S is total speech time.

$$E_3 = \frac{T_3 \times 100}{S} \tag{2.6}$$

Where E_3 is speaker error, T_3 is incorrectly labelled speech time and S is total speech time.

$$DER = E_1 + E_2 + E_3 \tag{2.6}$$

Where DER is diarization error rate, E_1 is missed speech rate, E_2 is false alarm speech rate and E_3 is speaker error.

Two errors in Equation (2.4) and Equation (2.5) occur during the speech activity detection, which is a pre-processing step in almost all diarization systems. E_3 in Equation (2.5) is partly contributed by errors in both the speaker segmentation and the speaker clustering. A speaker changes if missed during segmentation causes misclassification of the shorter of the two segments. If the system segments the audio into a greater number of segments than indicated by the annotation, often called over segmentation, there is a chance to make up for the over segmentation error during the

clustering step by merging neighboring segments together. However, if the segment happens to be small, the possibility to error increases due to less data to capture the speaker information in the segment. Erroneously clustered speaker segments are the main reason for E_3. Finally, DER is the total error in the systems hypothesis.

Calculating speaker error makes use of the Hungarian algorithm to perform a matching between algorithm labels and annotation labels. It iteratively finds the best matching cluster pair at each step based on maximum overlap between the 2 sets of labels. Finding the optimal mapping is needed because the system does not need to identify speakers by name and therefore its speaker labels will differ from the labels in the reference transcript.

2.2 Dataset for speaker diarization

In this section, we will present the details of dataset, which usually used in the speaker diarization research. Diarization experiments for the system developed in many research was trained and tested using various corpus.

The NDTV dataset which has been manually annotated. It consists of 22 episodes of the Hindu news Headlines Now show from the NDTV news channel. It consists manual annotations of 4h15m of English news reading with Indian accent. The anchor is the dominant speaker (maximum active time) in the episodes. The anchors are different across episodes. The dataset also has silence segments with lengths varying from 1s to 5s. No advertisement jingles are present in the dataset although the headlines of the show are announced with music in the background that is common across episodes. In cases where active speaker is a field correspondent background noise is also present. In the manually generated annotation the speakers in a single episode are labeled with the following information – gender of the speaker, background environment (clean /noisy /music), speaker ID in the episode (indicating anchor separately) in that order. The nonspeech labeled as silence, noise, speaker pause, or music consists of 7% of the total recording. Speaker overlap has been annotated with most dominant speaker in the overlap.

the REPERE dataset which was used in the REPERE video diarization campaigns. The REPERE corpus was annotated during the competition during 2012-2014. The REPRE dataset was used in the French Video Diarization campaign in 2012-

2014. The French REPERE challenge was a research evaluation competition aimed at systems performing multimodal person discovery in video recordings of broadcast news. The systems participating had to answer the question “Who is speaking”, “Who is present in the video”, “What names are cited” and “What names are displayed?” The original REPERE corpus is composed of various TV shows from two French TV channels that were diverse in their content - news, debates, celebrity interviews etc. It has been distributed by ELDA (Evaluation and Language resources Distribution Agency). 60 hours of data has been manually annotated. These annotations provide identity-labeled speech turns. The nonspeech segments which consist of lots of advertisements and show specific music events have not been annotated. The nonspeech duration out of the total 60 hours is 5%. Speaker overlap is annotated with every speaker present in the overlap. The annotation is in the standard RTTM format.

Subtitle-Aligned Movie (SAM) Corpus is a public SAD corpus consisting of 95 movies with 117 hours of movie audio, which 23 hours of speech and 1,595 different speakers.

CSTR VCTK Corpus is the English multi-speaker corpus, which consists of 109 native speakers of English with various accents. Each speaker reads out about 400 sentences, most of which were selected from a newspaper.

Hollywood movie dataset consists of around 2 hrs of speech and 6 hrs of nonspeech in four Hollywood movies. The dataset consists of speech with various acoustic backgrounds, which is a good benchmark for the entertainment media domain.

AVA speech dataset consists of around 40 hours with 40,000 annotated speech and non-speech segments. It contains around 350 different speakers in 192 videos which is available on YouTube.

CALLHOME American English is 30-minutes of 120 unscripted telephone conversation between 2 English native speakers.

2003 NIST Rich Transcription is developed for broadcast news speech and conversational telephone speech task in three languages, which is English, Mandarin Chinese and Arabic.

The 2000 NIST Speaker Recognition Evaluation has around 150 hours of telephone conversation.

2.3 Feature Extraction in speaker diarization

For the task of speaker diarization, acoustic features that discriminate speaker information in the spectrogram but are invariant to the phone sequence being uttered are desired. Mel-frequency cepstral coefficients (MFCC) or Perceptual Linear Prediction (PLP) coefficients, although not designed to distinguish between speakers, have been used widely in the areas of speaker verification and speaker recognition. Since a similar task of modelling speaker information is tackled in speaker diarization, MFCC and other cepstral features are the most commonly used features. During speaker segmentation 12-19 MFCC have been used along with the short time energy, while during clustering usage of higher order derivatives of these MFCC has been reported. LFCCs extracted using a linear filter bank instead of the Mel scale filter bank (Bozonnet, Evans & Fredouille, 2010) and Linear Prediction Cepstral Coefficients (LPCCs) (Narváez, Vera, Bedoya & Percybrooks, 2017) have also been tested but no conclusion has been reached regarding the better performance of either. Typical sizes of analyses windows are 25-30ms with frame hops of 10ms.

For speech activity detection, acoustic features that discriminate between speech and non-speech are sought after. Features such as energy (Narváez, Vera, Bedoya & Percybrooks, 2017), zero-crossing rate, spectral centroid, spectral roll-off and spectral flux (Bellagha, Labidi & Maraoui, 2017) have been used previously in speech activity detection. However, the use of these feature vectors has always been seen in concatenation with cepstral features.

Other than the above-mentioned short time analysis features, 4Hz modulation frequency features that convey long term characteristics of the acoustic signal have also been investigated (Maganti, Motlícek & Gatica-Perez, 2007), and have been applied in the speaker overlap detection and speech activity detection. A major challenge faced in these features though is the high dimensionality of the features and the computational cost associated with it. Long term cumulative features drawn over texture windows of 500ms such as median of pitch, long time average spectrum, deviation of the 4th and 5th formants, harmonics to noise ratio, formant dispersion etc. have shown to be of use for fast cluster initialization (Friedland et al., 2012), while features providing vocal source and vocal tract information (Chan, Zheng & Ouyang, 2006) have shown better speaker discrimination when used along with MFCC.

Recently Slaney et al. used features derived as activations of the bottleneck layer of a neural network. The artificial neural network was trained to discriminate 500ms segments as belonging to same or different speaker (Dawalatabad, Madikeri, Sekhar & Murthy, 2016). Ryant, Liberman and Yuan (2013) proposed a 50% relative improvement was reported for speech activity detection on a large Youtube corpus when a two dimensional softmax activation of a deep neural network was concatenated with 13 MFCC. Another interesting feature space explored in 2011 have sacrificed diarization error only slightly to obtain a 10x speed up using binary valued features for performing clustering (Delgado, Miró, Fredouille & Serrano, 2015). In this work, acoustic MFCC features of segments are transformed into a binary feature space using likelihoods obtained from GMMs.

2.4 Speech activity detection in speaker diarization

The task of finding contiguous segments of speech in an audio and segregating them from other types of sounds is called speech activity detection (SAD). It is beneficial for speech processing systems since it is practical to process only speech segments rather than entire recordings. It makes a design more efficient by saving computation time and resources. Apart from the computational advantages, the absence of an SAD often causes insertion errors in ASR systems. Hence speech activity detection is a fundamental task in almost all fields of speech processing coding, enhancement and recognition (Miró et al., 2012).

In speaker diarization, the error metric itself highlights the need for a speech activity detector since missed speech and false alarm speech are included in the diarization error rate metric. Moreover, with limited speaker data from small speech segments, presence of non-speech contaminates the estimated speaker models thereby affecting the performance of the diarization system. Initial approaches to diarization tried to let SAD be a by-product of the diarization system (Miró et al., 2012) by letting nonspeech be a single cluster which would be discarded at the end. However, it was soon noticed that systems having an explicit SAD gave better results.

SAD is often performed using frame-wise classification. Statistical models are trained and estimated on a feature space most suitable for discriminating the speech and nonspeech classes. In most cases, Gaussian mixture models are the statistical models

used and the feature space is in most cases cepstral features. Some works have reported use of acoustic features such as energy (Friedland et al., 2012), zero-crossing rate (Panagiotakis & Tziritas, 2005), spectral flux (Bellagha, Labidi & Maraoui, 2017).

2.4.1 Systems participating in NIST-RT evaluations

The NIST organised rich transcription evaluations which are now the current benchmark in meeting diarization. These benchmarks consist of results obtained by four participant systems (Miró et al., 2012). Typically, 1-3% missed speech error and 2-4% false alarm speech rates are the state-of-the-art in speech activity detection.

The SHoUT diarization toolkit for SAD (Zajíc et al., 2018) uses a bootstrap segmentation performed using speech and nonspeech models pre-trained on a dataset for Dutch broadcast news. It is followed by an iterative classification using a Viterbi decoder on 1 HMM with 2 states representing speech and nonspeech. The use of an HMM allows to control the minimum duration of the speech and nonspeech thereby preventing sporadic transitions from one class to another. The system uses 12 MFCC concatenated with zero crossing rate and their first and second derivatives in a 39-dimensional feature vector. The system was used by ICSI (Friedland et al., 2012) and LIA-Eurecom (Bozonnet, Evans & Fredouille, 2010), although the feature vectors used by the latter team for the iterative classification consisted of linear frequency cepstral coefficients (LFCC).

The UPC system (Luque, Miró, Temko & Hernando, 2007) made use of modified support vector machines (Proximal SVMs) with Gaussian kernels to segregate the speech and nonspeech in the audio. The modification allowed for faster retraining of SVMs as suited for an iterative classification.

The IIR-NTU (Friedland et al., 2012) system performed a bootstrap segmentation based on an energy derived confidence score. An iterative classification with GMMs trained for speech and nonspeech, using high confidence frames from the bootstrap segmentation, followed the initial segmentation to refine the speech and nonspeech classes. The authors reported use of Linear Prediction Cepstrum Coefficients (Friedland et al., 2012) for both the bootstrap segmentation and the iterative classification. This approach was completely independent of external training data for the speech and nonspeech models.

2.4.2 Broadcast news systems

In the LIUM diarization toolkit (Rouvier et al., 2013), the authors developed a model based segmentation system for speech activity detection using an 8 state HMM with 2 states of silence (wide and narrow band), 3 states of wide band speech (clean, over noise or over music), 1 state of narrow band speech, 1 state of jingles, and 1 state of music. Each state is modeled with a GMM of size 64 of MFCC, their deltas and delta-deltas. All the models were trained using the extensive data for each model from the ESTER1 dataset. This system resulted in a 1.1% false alarm speech and 3.9% missed speech on the dev0 subset of the REPERE corpus. Their results on other databases ESTER2 and ETAPE are reported by Rouvier et al. (2013). Besides speech activity detection the LIUM toolkit also performs a gender and bandwidth detection. This also uses a model-based segmentation with 128 sized diagonal GMMs for each of the 4 classes (2 genders x 2 bandwidths) and a feature warping.

The Albayzin 2010 campaign saw five competing systems. The best results for SAD were reported by Bellagha, Labidi and Maraoui (2017). Although the DER was much worse than others (55% DER), their SAD error stood best at 3.4% (1.1% missed and 2.3% false alarm). The authors reported using multi-layer perceptrons instead of GMMs to model emission probabilities of a five-states hybrid NN-HMM system. The feature space was also expanded. 16 MFCC concatenated with 8 other audio features - energy, zero-crossing rate, spectral centroid, spectral roll-off, maximum normalized correlation coefficient and its frequency, harmonicity measure and spectral flux. Information regarding other participating systems in the Albayzin campaign is mentioned by Zelenák, Schulz and Pericás (2010).

In the REPERE 2012-2014 evaluations, three consortia took part - SODA, QCompere and PERCOL (Bernard, Galibert & Kahn, 2014). The SODA consortia used the LIUM toolkit above. The QCompere system had a 4 state HMM similar to the LIUM toolkit one state each for speech, silence, noise and music (Zhu, Barras, Meignier & Gauvain, 2005) modelled by GMMs of size 64. The PERCOL system (Favre et al., 2013) performed a 3 class GMM based SAD. Interestingly their 3 classes were non-speech, overlapping speech and non-overlapping speech, each modeled by 256 sized GMMs trained from the ETAPE corpus. The overlap detection reportedly also improved the DER than the baseline clustering system.

2.5 Speaker segmentation in speaker diarization

In audio segmentation, the task is to create homogeneous and contiguous chunks of audio that show dissimilarity from its neighboring segments. It is also called acoustic change detection. We will look at two approaches to audio segmentation with more focus on methods used in speaker segmentation applied to speaker diarization.

2.5.1 Metric based segmentation

One of the most common audio segmentation methods to date is metric-based segmentation. These methods are very popular in music segmentation tasks as well. In metric-based segmentation a distance metric is first defined between two audio segments, that indicated their similarity. Then a change detection strategy is implemented using this metric. Compared to model-based methods, these methods have great advantage since they do not need any information about the data a priori.

For music segmentation, distances are calculated between the feature directly. In speech processing however, the features (generally cepstral features) used are not suitable for framewise distance computation for comparing speaker similarity, due to their variability with the phones uttered. To aggregate speaker information from longer segments, it is assumed that the features of every segment come from a probability distribution. Distance comparison is done between these probability distributions using statistical similarity measures such as the KL divergence, Cross Likelihood Ratio, Bayesian Inference Criterion etc. The most commonly used probability distribution for modeling chunks of feature vectors during speaker segmentation is the full covariance multivariate Gaussian distribution.

Bayesian Inference Criterion (BIC) is a model selection criterion i.e., is a statistical criterion that compares available models for representing the data. The aim during this selection is to calculate if there is any over-fitting. For a set of vectors X, the BIC of the model M is one such criterion defined as:

$$BIC(X, M) = \log P(P|M) - \lambda \cdot \#(M) \cdot \log N \tag{2.7}$$

The first term calculates the likelihood of the data given the model, whereas the latter term penalizes it proportional to the number of parameters $\#(M)$ that the model

uses and the size of available data on which the model was trained. The second term is called the complexity of the model.

The BIC can be applied to indicate whether the two sets of feature vectors being compared for similarity are drawn from the same distribution or from different distributions. To measure similarity between blocks $X1$ and $X2$ the following hypotheses need to be compared:

H0 : The feature vectors from $X1$ and $X2$ are from same distribution

H1 : The feature vectors from $X1$ and $X2$ are from separate distributions.

Let the model for H0 be M which would be trained on X i.e., $X1$ concatenated with $X2$ and let the models for H1 be $M1$ and $M2$ for $X1$ and $X2$ respectively. We define

$$\Delta BIC = BIC(M) - BIC(M1) - BIC(M2) \tag{2.8}$$

A positive value for ΔBIC suggests dissimilarity between the two blocks $X1$ and $X2$ and hence indicates that there is a change between segments $X1$ and $X2$.

Chen (1998) built a completely unsupervised system using ΔBIC and their method has been replicated a number of times for speaker/ environment change detection (Chen,1998) in the speaker diarization domain. Improvements were made by Labrunie et.al (2018) to make faster implementations of change detection using the BIC approach by reducing the number of computations with some compromise on accuracy.

Metric based segmentation methods are implemented in two strategies, a fixed sliding window strategy and a growing window search strategy. In the former, there is a window of fixed size, the center of which is being inspected for a change (Trancoso & Redol,2009). If the feature vectors on either side of the midpoint are better modeled by separate distributions, resulting in a higher distance between distributions, the midpoint is declared as a change point. The size of the sliding window is typically five second and the two two-second segments are compared for a similarity. With a larger window size, the two segments would be modeled better. However, it would have higher chances of missing a change if more change points enter the window under consideration, since the probability distribution estimated on the segment may get contaminated.

When implementing the growing window search strategy, a single change is pursued from the start of the recording in a window of certain size generally about five second. If no change is detected in this window, the size of the window is increased and a change is searched in the new window. After the first change is detected, the search is reset from the last change detected. The growing window method has been reported with the BIC metric (Chen,1998). In recent years, the Rouvier et al., 2013 shown the growing window BIC segmentation followed by a BIC clustering that merges only consecutive segments to reduce false alarm speaker changes.

2.5.2 Model based segmentation

Model based segmentation methods train a GMM for every segmentation class. These GMMs are used as a PDF in a hidden Markov model (HMM) where each state is connected to every other state with equal transition probability. The Viterbi decoding using this HMM gives a segmentation of the audio recording. A major disadvantage of the model-based segmentation is that the GMMs needs to be known before hand, hence need some external training data.

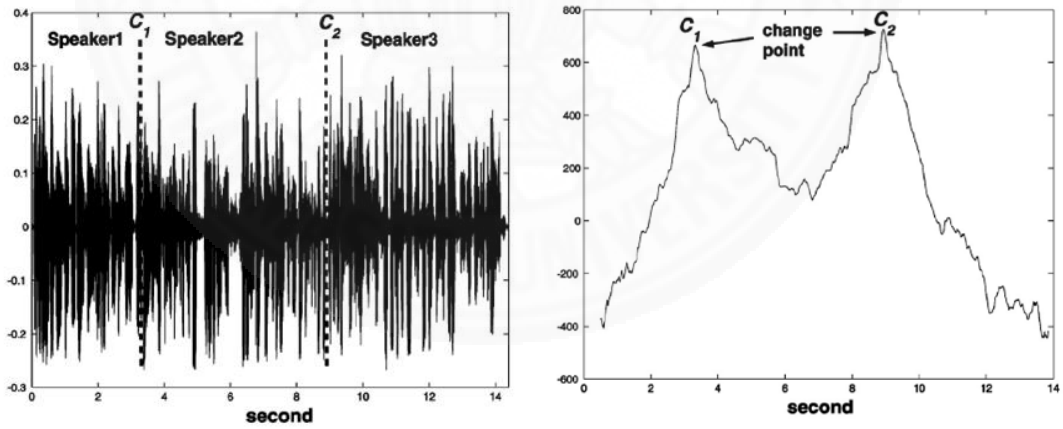

Figure 2.1: Sliding window search for speaker change detection.

In Figure 2.1, distance is computed between two halves of the sliding window and plotted with time. Peaks in the distance indicate a change (Lopez-Otero,2015).

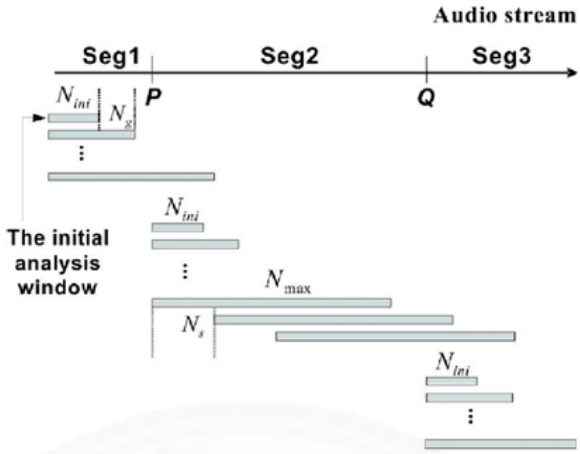

Figure 2.2: Growing window search for speaker change detection.

Figure 2.2 demonstrate that every search is for a single change point. Search is reset after the change is found by Lopez-Otero (2015). However, when segmentation and clustering are performed iteratively, the output produced by the speaker clustering algorithm gives a set of speaker segments as training data for GMMs of the next iteration to refine speaker segmentation. A pre-clustering is often performed to get an initial grouping of audio segments (Friedland et al., 2012) with each grouping showing resemblance of speaker information.

Often model based segmentation methods have been used only as a post processing step to achieve a refined segmentation (Moraru et al., 2004). Such model-based techniques are famous in segmentation during speech activity detection (Rouvier et al., 2013) where the acoustic change being looked at is between speech and nonspeech. In some telephone diarization systems (Reynolds & Torres-Carrasquillo, 2004), there has been a pre-segmentation of the audio recording based on bandwidth and gender using GMMs trained for each of the 4 classes (2 bandwidths x 2 genders).

The model-based segmentation methods are more famous in meeting diarization systems. With the advent of better clustering algorithms with i-vector speaker models, the focus has shifted to performing speaker segmentation and speaker clustering separately for diarization. However, use of the GMM-HMM framework for refinement is still popular (Rouvier et al., 2013).

2.6 Speaker clustering in speaker diarization

Clustering is a common problem in statistical data analysis. It has been addressed in many scientific fields right from exploratory data mining to community detection in social networks. It is the process of grouping a set of objects such that objects in each group, called cluster, are more similar to each other than they are to objects in other groups or clusters. The objects could be points in a vector space or even statistical models. The similarity mentioned above is a distance like measure defined between the objects by the user. The word similarity is used because the measure defined need not satisfy all the properties of a norm viz. non-negativity, triangle inequality and symmetry. The words similarity and distance have been used interchangeably here, with less distance meaning more similarity and vice versa.

The process of clustering is generally translation invariant and hence the relative position of the objects in their space is more relevant rather than the objects themselves. Indeed, this relative position of the objects is indicative of pairwise similarity. For the problem of speaker diarization, the aim is to perform clustering of segments of audio based on the active speaker in each segment. Each cluster should ideally represent a single speaker.

The dimensionality of the spectrogram of a single segment is large and comparison between these segments based on their spectrogram is not computationally viable. Hence the segment needs to be quantified in a low dimensional space to compare the similarity of their speaker information. A few speaker models that have been utilized in the past in the fields of speaker verification and speaker recognition have been reviewed. Every speech segment would have a representative vector or a statistical model which is characterizes the speaker information. Clustering is performed on these speaker models.

2.6.1 Speaker models for clustering

Since speaker diarization needs to capture speaker information from audio segments, speaker models commonly used in speaker verification and speaker recognition are adopted. The two main speaker models GMMs and i-vectors are explained below. Off these the i-vectors have recently become state of the art in speaker verification tasks.

2.6.1.1 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) of cepstral features are often used to model speakers. A Gaussian mixture model is a popular tool for modeling multi-modal data and possess the following form.

$$p(x) = \sum_{i=1}^{N} w_i N_{(\mu_i, \sum_i x)} \quad s.t. \ \sum_{i=1}^{N} w_i - 1 \tag{2.9}$$

Since the segment durations can be small, the number of feature vectors available from a segment is sometimes insufficient to estimate a full Gaussian mixture model. To overcome this problem pre-trained Universal Background Model (UBM) is adapted for the segment to obtain its speaker model (Reynolds, Quatieri & Dunn, 2000). The UBM is a comprehensive model for data from multiple speakers combined together that captures variability of speech. For GMMs of cepstral features, different statistical similarity measures have been investigated earlier such as the symmetric KL divergence, normalized cross likelihood ratio (NCLR) etc. (Beecks, Zimmer, Kirchhoff & Seidl, 2011). The Kullback–Leibler divergence is an information theoretic measure of how different the two probability distributions are from each other, while the cross likelihood ratio compare P (X1 |M2) and P (X2 |M1) (equation 3.2 & 3.3).

$$CLR(X_1, X_2) = \log \frac{P(X_1|M_1)}{P(X_1|M_2)} + \log \frac{P(X_2|M_2)}{P(X_2|M_1)} \tag{2.10}$$

$$NCLR(X_1, X_2) = \frac{1}{|X_1|} \log \frac{P(X_1|M_1)}{P(X_1|M_2)} + \frac{1}{|X_2|} \log \frac{P(X_2|M_2)}{P(X_2|M_1)} \tag{2.11}$$

where Mi is the model estimated on Xi . As we can see, if feature vectors of segment X1 and X2 come from the same speaker, X1 it fits the model of segment X2 well, so the cross likelihood increases, decreasing the distance.

Recent experiments in the SV and SR fields noted that only the means of gaussians in a GMM contain most speaker related information. Due to the high variability of the covariance matrices and mixture weights with respect to utterances (Reynolds, Quatieri & Dunn, 2000), they are not reliable indicators of speaker information. Hence, instead of calculating the above likelihood scores, the means of a GMM are concatenated to get a single vector (called the GMM supervector) in a high dimensional vector space. Distance measures such as the cosine distance and Mahalanobis distance (Curelaru, 2018) have been investigated on this space. To make

a comparison between two GMM supervectors they need to be adapted from the same UBM to make sure that the corresponding mean vectors of the GMM are being compared between segments. The adaptation algorithm is detailed by Reynolds, Quatieri and Dunn (2000).

2.6.1.2 I-vector representation

The concept of i-vectors was first introduced in speaker verification as a feature extraction from GMMs to reduce the dimensionality of the GMM hyper parameters. With the UBM sizes being of the order of 512, 1024 or even 2048 Gaussians in some GMM systems, the size of the supervector becomes very large to do further computation on the supervector. Instead, using factor analysis for reducing the dimensionality of the supervector led to a new representative vector with a few hundred dimensions. This subspace, called the total variation subspace, is hypothesized to contain spectral information of the speaker and background.

$$m = M + Tx \tag{2.12}$$

where m is the mean adapted supervector of the utterance for which the i-vector x is sought. M is the mean supervector of the UBM. The matrix T is a tall low rank matrix representing the total variability subspace which needs to be learned on a training dataset. Although supervectors typically have tens of thousands of dimensions, this representation constrains all supervectors to lie in an affine subspace of the supervector space. The dimension of the affine subspace is at most a few hundred.

i-vector extraction requires speaker labeled training data with multiple utterances of the same speaker with possible variations in utterances in terms of their phonetic balance and background noise. The training algorithm for the total variability subspace (Curelaru, 2018) and the i-vector extraction from the Baum-Welch statistics of the utterance have been implemented in the MSR Identity toolkit (Sadjadi, Slaney & Heck, 2013).

2.6.2 Clustering algorithms

Given the similarity matrix between the speaker GMMs or i-vectors, a clustering algorithm aims at reaching the best set of clusters with minimum intra-cluster

variance and maximum intercluster variance We will look at two clustering algorithms used previously in diarization – (i) hierarchical agglomerative clustering (HAC) and (ii) integer linear program (ILP) clustering, which use different solving techniques and also have different criteria for arriving at the best set of speaker clusters.

2.6.2.1 Hierarchical Agglomerative Clustering

HAC is a greedy algorithm i.e. it makes a locally optimal choice at each stage with the hope of finding the global optimum.In an iterative process, the 2 most similar clusters are merged into a single cluster. The number of clusters reduces by 1 at each step. This iterative process continues until only one cluster remains. While merging 2 clusters, the data from segments corresponding to the 2 clusters is concatenated and a single speaker model is re-calculated on it. The distances of every other cluster with this newly formed cluster are re-calculated to update the similarity matrix for the next step.

The optimal set of clusters is chosen based on an optimality criterion. One optimality criterion is to choose the set of clusters where the minimum inter cluster distance is greater than a threshold. Another criterion was proposed by Nguyen, Siong and Li (2008) in which from among the clusters from every iteration, the set of clusters where the histograms of intra-cluster distances and inter-cluster distances are farthest from each other is chosen.

$$\Delta BIC = BIC(M) - BIC(M1) - BIC(M2) \tag{2.13}$$

where (X (k) i , X(k) j) is the similarity matrix in the k th iteration

$$\Delta BIC = BIC(M) - BIC(M1) - BIC(M2) \tag{2.14}$$

where m, σ and n are the mean, standard deviation and number of elements in the inter-cluster distances and similarly for the intra-cluster distances.

The ICSI system (Friedland et al., 2012) performed an initial clustering of segments of 1s duration using prosodic long-term features. This segmentation was then refined iteratively in a GMM-HMM framework over MFCC feature vectors derived from segments within each cluster. Each state of the HMM represented a speaker and was modelled by a GMM. The use of the HMM allowed adding a constraint for

obtaining contiguous speech turns of 2s. In each iteration, the number of clusters was reduced by merging state GMMs in a HAC using the BIC distance. The IIR-NTU system (Friedland et al., 2012) used LPCCs to form 30 initial clusters from uniformly dividing the audio, and then iteratively used the same GMM-HMM framework to perform HAC with CLR distance. The LIA-Eurecom system (Bozonnet, Evans & Fredouille, 2010) uses the same framework, although in their method, they took a top-down approach (also called divisive clustering) instead of HAC and performed a splitting of states starting from a single state. Many other systems have been implemented that use the HAC approach to clustering with different distances (Vijayasenan, Valente & Bourlard, 2007).

i-vector speaker models adopted from speaker verification were first used in speaker diarization of telephone conversations where the number of speakers in the recording was known a priori and hence k-means clustering of i-vectors was performed to arrive at 2 clusters of i-vectors (Shum et al., 2011). Later, an HAC like clustering of i-vectors for broadcast news were reported by Silovský and Prazak (2012) which demonstrated better performance over traditional BIC based GMM HAC architecture.

2.6.2.2 ILP based clustering

In an effort by Rouvier and Meigner in 2012, a global optimization approach was proposed to perform speaker clustering (Rouvier & Meignier, 2012) using an integer linear program (ILP). Clustering is posed as a combinatorial optimization problem on a complete graph (each node connected to every other node). The speaker segments are considered as nodes of the graph and the incidence matrix is the similarity matrix. The integer linear program to find the optimal clustering is a variation of the k-centres problem. In simple words, the k-centres problem is to choose K cities out of N for building warehouses so that the worst-case distance between a city and its closest warehouse is minimized. The ILP is adapted for unsupervised speaker diarization since the number of speakers (K) in not known a priori.

With the introduction of ILP clustering in the broadcast news domain, diarization systems now typically perform segmentation and clustering separately (Zelenák, Schulz & Pericás, 2010), and perform a post processing step of Viterbi decoding using the GMM-HMM framework. Recently in 2014, Dupuy, Meignier,

Deléglise and Estève (2014) improved upon the above mentioned ILP framework to reduce the redundancies in the constraints, making clustering extremely fast. Hence, in the recent years, the ILP based methods are in the limelight over the traditional HAC based clustering due to their better performance in terms of speed and accuracy. The LIUM toolkit reported a 17.19% DER with GMM based CLR clustering and 15.46% DER with i-vector based ILP clustering on the REPERE corpus (Dupuy, Meignier, Deléglise & Estève, 2014).

2.7 Summary

In this chapter a few background concepts used in segmentation, clustering and speech activity detection have been presented, and previously implemented techniques in speaker diarization literature are reviewed. Other than the speaker diarization task, where the number of speakers and the presence of specific speakers is not known a priori, there have been many specialized audio indexing tasks that have been investigated in the past. For example explicitly detecting the presence of music (Bouafif & Ellouze, 2019), helping find the structure of a broadcast program (Liu, Wang & Chen, 1998) or locating commercials to eliminate unwanted audio (Johnson & Woodland, 2000). In another work, making use of speech transcriptions, fast speaker change detection was applied by Liu & Kubala (1999). More generic systems specialized for different domains include the Alize speaker recognition toolkit which has a diarization sub-block, the SHoUT toolkit (Zajíc et al., 2018) that was designed for meeting diarization. INRIA, IDIAP, and DiarTK are some other toolkits that are still under development for diarization.

CHAPTER 3
SYSTEM DESCRIPTION

In this chapter the detail of proposed speaker diarization system was described. The proposed speaker diarization system consists of three processes, which is speech activity detection, speaker change detection and speaker clustering. At first, speech activity detection was implemented to find the speech segment. The speaker change detection was then implemented to find the specific time that speaker change in the output of the speech activity detection, which is speech segment. The speaker clustering was finally implemented to group the speech segment according to speaker id.

3.1 Speech activity detection

In this section, the proposed speech activity was described. The speech activity detection (SAD) is the technique for detecting speech under various noise conditions in the audio recording. We use the combination technique and fused model with Dempster-Shafer theory (DS theory).

3.1.1 Feature extraction

we chose five acoustic features, which is Mel Frequency Cepstral Coefficients (MFCC) , log mel-Spectrogram (Log-mel) ,chroma, spectral contrast, and tonnetz to extract the sound characteristic for identifying the speech segment in the audio recording.

3.1.1.1 Mel-Frequency Ceptrum Coefficient

The most widely used acoustic feature extraction technique is the Mel Frequency Cepstral Coefficient (MFCC). MFCC are short-term spectral based and dominant features and are widely used in the area of audio and speech processing. The mel frequency cepstrum has proven to be highly effective in recognizing the structure of music and speech signals and in modeling the subjective pitch and frequency content of audio signals. The MFCC have been applied in a range of audio mining tasks, and have shown good performance compared to other features. It computes the cepstral coefficients along with delta cepstral energy and power spectrum deviation which

results in various dimensional features. The low order MFCC contains information of the slowly changing spectral envelope while the higher order MFCC explains the fast variations of the envelope to capture the difference characteristic between speech and non-speech.

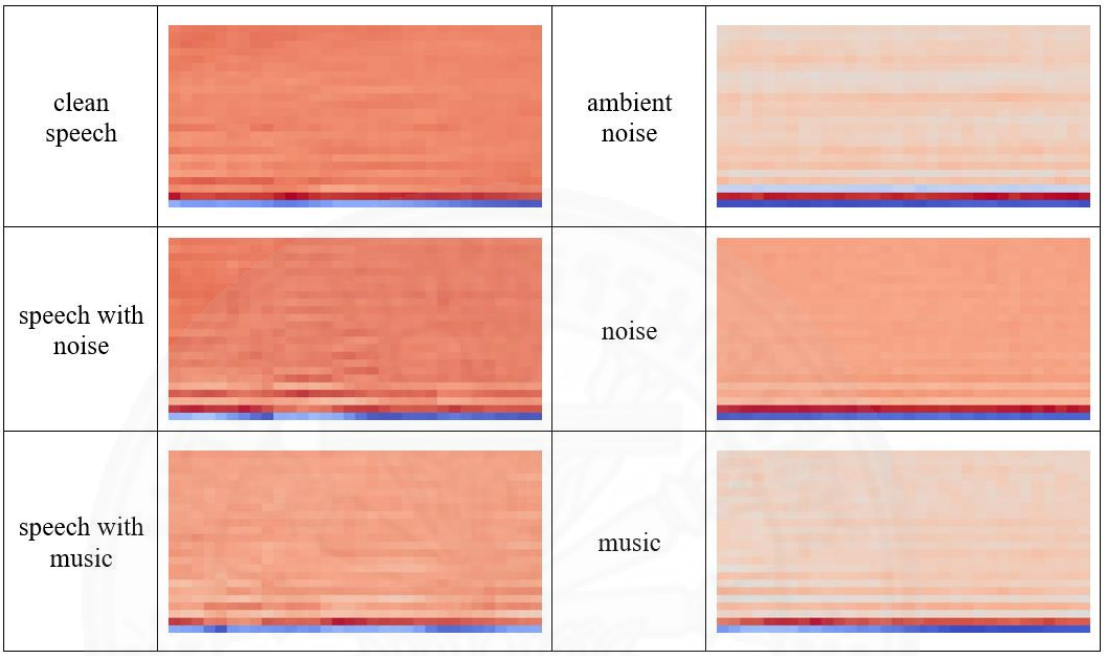

Figure 3.1 MFCC feature of different acoustic conditions.

3.1.1.2 Log-mel spectrogram

Log mel-Spectrogram (Log-mel) is a spectrogram with the mel scale, which is displayed in log-scaled or decibels. To plot the spectrogram, we break the audio signal into millisecond chunks and compute Short-Time Fourier Transform (STFT) for each chunk. We then plot this time chunk as a colored vertical line in the spectrogram. Spectrograms represent the frequency content in the audio as colors in an image. Frequency content of milliseconds chunks is stringed together as colored vertical bars. Spectrograms are basically two-dimensional graphs, with a third dimension represented by colors.

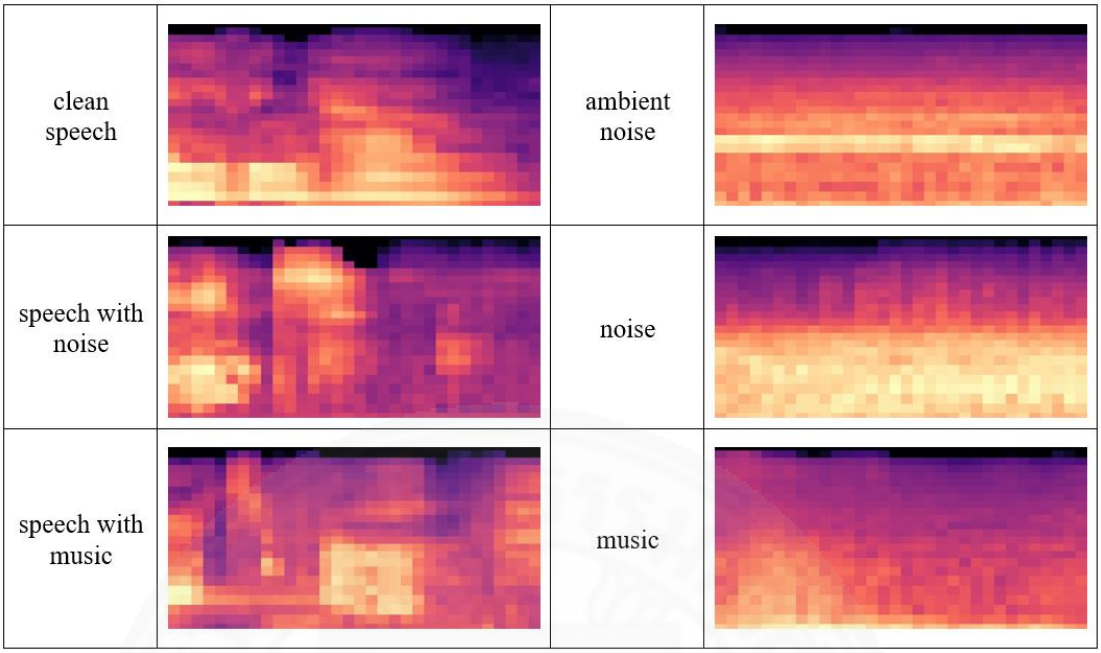

Figure 3.2 Log-mel spectrogram feature of different acoustic conditions.

3.1.1.3 Spectral Contrast

Spectral contrast takes only the spectral valley and spectral peak to calculate difference in each frequency sub-band separately, so that it could differentiate the spectral characteristics. spectral contrast feature in sound classification task demonstrates that the octave-based spectral contrast feature has a better discrimination among different sound types. This kind of features averages the spectra in each sub-band and reflects the average spectral characteristics, but it could not represent the relative spectral characteristics in each sub-band, which seem more important to discriminate different types of sound.Thus, we use octave-based spectral contrast to represent the spectral characteristics of music, noise and speech.

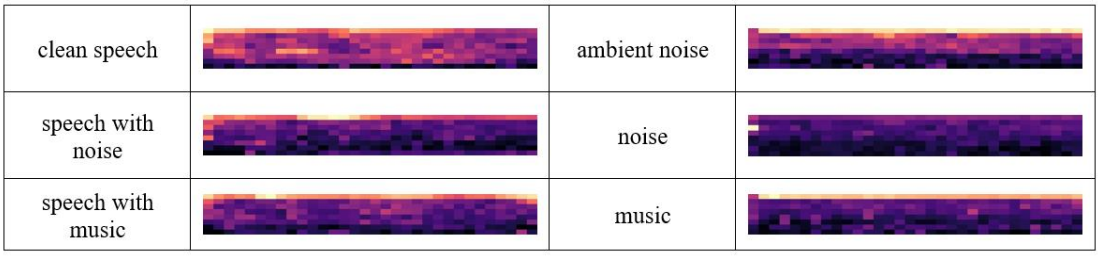

Figure 3.3 Spectral contrast feature of different acoustic conditions.

3.1.1.4 Tonnetz

Tonnetz are used to show harmonic relationships in European classical track. In simple term Tonnetz is a conceptual lattice layout representing tonal space in musical tuning and harmony.The tonnetz is a visual method for representing chords and voice leading relationships in the plane. Tonnetz was defined as a spatial metaphor for music theory, which was described intervals as displacements in the space of pitches (and pitch classes), before generalizing the concept of interval itself. So, we apply the displacements in the space of pitches to classify the speech and non-speech in the recording.

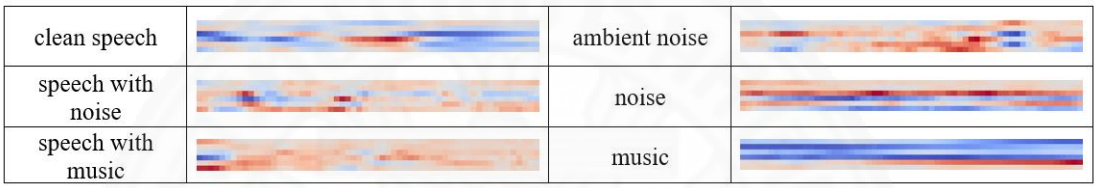

Figure 3.4 Tonnetz feature of different acoustic conditions.

3.1.1.5 Chroma

In music, the term chroma feature or chromagram closely relates to the twelve different pitch classes. Chroma-based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing music whose pitches can be meaningfully categorized, often into twelve categories, and whose tuning approximates to the equal tempered scale. One main property of chroma features is that they capture harmonic and melodic characteristics of sound, especially music, while being robust to changes in timbre and instrumentation. Chroma features aim at representing the harmonic content (eg:keys,chords) of a short-time window of audio. Thus, we believe that the harmonic and melodic characteristics can determine the speech in the recording.

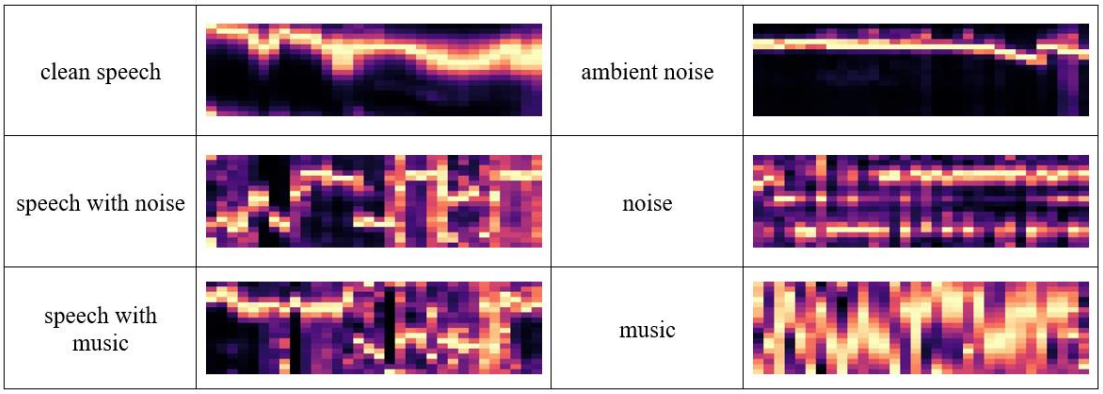

Figure 3.5 Chroma feature of different acoustic conditions.

In the feature extraction for speech activity detection, we extract the 32-dimensional log-mel spectrogram features and 32-dimensional mel frequency cepstral coefficient features. Then, 18-dimensional chroma features, 6-dimensional tonnetz features, and 8-dimensional spectral contrast features. After feature extraction, we first combined Log-mel with chroma, tonnetz and spectral contrast vertically and denoted the first feature set as LMCST. We then extract MFCC, chroma, spectral contrast, and tonnetz and denoted as MFCST. All feature sets were created by concatenating four acoustic features vertically and used as the input of neural network.

3.1.2 Classification

The dense Convolutional Network (DenseNet) was used as the classifier. All architecture in our experiment has the same Dense Convolutional Network (DenseNet) with the different acoustic feature sets.

For our first architecture (MFCST-DenseNet), I use the MFCST feature set as an input feature. Instead of MFCST feature set, I use the LMCST feature set for LMCST-DenseNet architecture. After the softmax value was obtained from each neural network, the two softmax values are fused through Dempster-Shafer theory (DS theory) to form the sound classification results in the final architecture. I denote the model as DS-DenseNet.

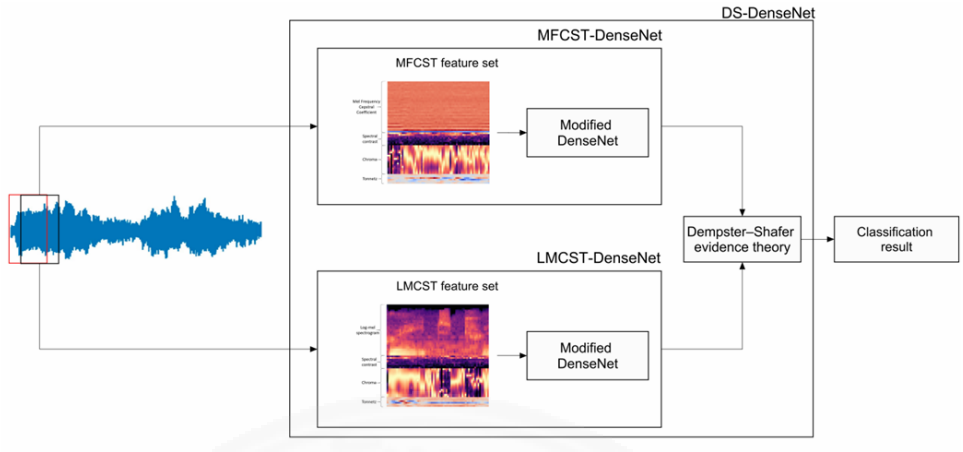

Figure 3.6 The framework of the speech activity.

3.2 Speaker change detection

In order to find the speaker changing in each speech segment, we propose a method way to integrate the pitch and spectrum continuity with pitch and spectrum related features such as chroma feature and spectral contrast. We first extract the chroma feature and spectral contrast from the speech segment. We then combined them with by concatenating two acoustic features vertically.

Chroma-based features, which are also referred as "pitch class profiles". Chroma closely relates to the twelve distinct pitch classes. One of the essential assets of chroma features is that it can capture harmonic and melodic characteristics of tune, while being robust to adjustments in timbre and instrumentation. The chroma feature is a mainstream audio melody feature in the field of music information retrieval. It is designed according to Twelve Tone Equal Temperament. Since, in music, notes exactly one octave apart are perceived as particularly similar, knowing the distribution of chroma even without the absolute frequency (the original octave) can give useful musical information about the audio, and may even reveal perceived musical similarity that is not apparent in the original spectra. Chroma feature is usually represented as a 12-dimensional vector, and each element of a vector is associated with one element of the set of {C,C#, D, D#, E, F, F#, G, G#, A, A#, B , reflecting the local energy distribution of the audio signal at semitones represented by the 12 pitch names.

Octave-based Spectral Contrast feature is proposed to represent the relative spectral characteristics of music. Octave-based Spectral Contrast feature considers the

strength of spectral peaks and spectral valleys in each sub-band separately, so that it could represent the relative spectral characteristics, and then roughly reflect the distribution of harmonic and nonharmonic components. Thus, Octave-based spectral contrast can be a measure of the relative distribution of harmonic and non-harmonic components in a spectrum. The Octave-based spectral Contrast represents the strength of spectral peaks and spectral valleys in each sub-band separately, so that it could represent the relative spectral characteristics. Spectral peaks and spectral valleys are obtained from the amplitude spectrum, thus discrete Fourier transform (DFT) is first applied on the signal, and the spectrum is divided into octave based subband. Spectral peaks and spectral valleys are estimated by averaging the values in the small neighborhood around maximum and minimum values of the amplitude spectrum respectively. The spectral contrast is defined as difference between spectral peak and spectral valley. The statistics of spectral peaks and spectral contrasts are used for the feature vector of texture window.

After we have the feature combination for the speaker change detection, we consider the continuity of the feature set by focusing on the difference of the column of the pixel value in RGB space.

3.3 Speaker clustering

The speaker clustering is applied after the speaker change detection was applied to find the accurate time that speaker change in the speech segment. The speaker clustering is the technique in which every speech segment was grouped according based on the speaker id.

3.3.1 Feature extraction

I apply Mel Frequency Cepstral Coefficients and log-mel spectrogram to find the similarities between each speech segments. cosine similarity is used as our similarity metric.

3.3.1.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) are mostly used in automatic speech and voice recognition. The sounds are generated from vocal tract of human. If

it can determine the vocal tract shape correctly, this will give us an representation of a short-time power spectrum of a speech signal which is called “MFCC”. The motivating idea of MFCC is to compress information about the vocal tract (smoothed spectrum) into a small number of coefficients based on an understanding of the cochlea. MFCC is a technique based on human hearing behavior that cannot recognize frequencies over 1Khz. MFCC are based on the difference of frequencies that the human ear can distinguish. The signal is expressed in the MEL scale, this scale is based on the perception of the pitches in an equally spaced intervals judged by observers. This scale uses a filter that is spaced linearly at frequencies below 1000 Hz and logarithmic spacing above 1000Hz. The filters are spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech and audio.

3.3.1.2 Log-mel spectrogram

In the source-filter model of speech, MFCC are understood to represent the filter (vocal tract). The frequency response of the vocal tract is relatively smooth, whereas the source of voiced speech can be modeled as an impulse train. The result is that the vocal tract can be estimated by the spectral envelope of a speech segment.

3.3.2 Clustering algorithm

In this clustering algorithm, each cluster is represented by the centroid of all its corresponding matrix. After that, I compute its similarities to centroids of all existing clusters. If they are all smaller than the threshold, then create a new cluster containing only this matrix. Otherwise, add this matrix to the most similar cluster and update the centroid. And we find the threshold by applying the 2-D CNN architecture.

Speech data in the entertainment media domain often has the case that one speaker will speak often, while other speakers will speak rarely. In order to due with cluster imbalance, we add the CSTR VCTK Corpus as the training set in the speaker clustering process. We use Mel frequency Cepstral Coefficient (MFCC) to extract the features from voice and Vector quantization technique to identify the speaker, this technique is usually used in data compression.

3.4 Evaluation the system

In broadcast news, there is very little overlap. For the case where overlap is absent, the formula for DER can be simplified. This error calculation is used in the broadcast news diarization systems. Pyannote (Bredin, 2017) is a python based toolkit that facilitates calculation specifically for such systems.

$$S = C_1 + T_1 + T_3 + T_4 \tag{3.1}$$

Where S is total speech time, C_1 is correctly labelled speech time, T_1 is Missed Speech time, T_3 is incorrectly labelled speech time from speaker change detection process and T_4 is incorrectly labelled speech time from speaker clustering process.

$$NS = C_2 + T_2 \tag{3.2}$$

Where NS is total non-speech time, C_2 is correctly labelled non-speech time and T_2 is false alarm speech time.

The DER can be broken down systematically into 2 types. Consider an audio with speech and non-speech as indicated by the annotation. Non-speech includes silences, speaker pauses, music, jingles, noise etc. The two categories can be further classified exhaustively as shown in Equation (2.2) and Equation (2.3). Missed speech time is the time when the algorithm erroneously indicated a segment as non-speech. False alarm speech time, on the other hand, is the time when the algorithm erroneously indicated a segment as speech.

$$E_1 = \frac{T_1 \times 100}{S} \tag{3.3}$$

Where E_1 is missed speech rate (MSR), T_1 is missed speech time and S is total speech time.

$$E_2 = \frac{T_2 \times 100}{S} \tag{3.4}$$

Where E_2 is false alarm speech rate (FASR) and T_2 is false alarm speech time and S is total speech time.

$$E_3 = \frac{T_3 \times 100}{S} \tag{3.5}$$

Where E_3 is speaker error from speaker change detection process, T_3 is incorrectly labelled speech time from speaker change detection process and S is total speech time.

$$E_4 = \frac{T_4 \times 100}{S} \tag{3.6}$$

Where E_4 is speaker error from speaker clustering process, T_4 is incorrectly labelled speech time from speaker clustering process and S is total speech time.

$$DER = E_1 + E_2 + E_3 + E_4 \tag{3.7}$$

Where DER is diarization error rate, E_1 is missed speech rate, E_2 is false alarm speech rate and E_3, E_4 is speaker error.

Calculating speaker error makes use of the Hungarian algorithm to perform a matching between algorithm labels and annotation labels. It iteratively finds the best matching cluster pair at each step based on maximum overlap between two sets of labels. Finding the optimal mapping is needed because the system does not need to identify speakers by name and therefore its speaker labels will differ from the labels in the reference transcript.

CHAPTER 4
EXPERIMENTS AND RESULTS

4.1 Evaluation of speech activity detection

Subtitle-Aligned Movie (SAM) Corpus was used as a training set to build the speech activity detection model for indexing the speech segment in the audio recording. The speech signal was sampled at 16 kHz with 16-bit resolution. From the statistic of the SAM corpus (Hebbar, Somandepalli & Narayanan, 2019), We chose a segment of length 1.28 second with a duration of pause 0.9 seconds. Thus, we received around 78,000 speech segments. The non-speech segment was separated into segments of the length of 1.28 seconds. We used segments of length 640 ms with 64 frames to extract 64-dimensional combined features in order to use the square input feature as the input of our models. We used the overlap percentage of 87. 5%, as suggested by corpus Hebbar, Somandepalli and Narayanan (2019). All models were trained on the SAM corpus at a segment level using a batch-size of 2048. Focal loss with Adam optimizer was used as a loss criterion with a focusing parameter of 2. The value of the focusing parameter is selected based on the results by Lin et al. (2020). Then,Adam optimizer by Kingma and Ba (2015) is applied as optimizer. Models are trained in the network for ten epochs. We then tuned filter shape and number of fully connected layers (FC). Models are evaluated on the Hollywood movie dataset, AVA-speech dataset, CALLHOME American English, 2003 NIST Rich Transcription and 2000 NIST Speaker Recognition Evaluation at segment-label. We use Missed speech rate (MSR) and False alarm speech rate (FASR) as the metric for speech activity detection evaluation.

In this section, we demonstrate the result of three independent experiments, which is the studies of the effect of acoustic feature extraction and neural network on speech activity detection, the studies of the effect of the dimension of feature in the combined feature on the model training and the studies of the effect of varying dimension of feature in the combined feature with Dempster-Shafer.

4.1.1 The studies of the effect of acoustic feature extraction and neural network on speech activity detection.

We first study the effect of acoustic feature extraction and neural network to compare the performance of each feature extractions and classifiers in the speech activity detection task. We evaluate each method on two different domains, which is the entertainment media domain and telephone conversation domain.

4.1.1.1 Evaluation on the entertainment media corpus

In order to evaluate the speech activity detection on the entertainment media corpus, the Hollywood movie dataset and AVA-speech dataset were used. We first implemented the speech activity detection with the two most widely used acoustic feature extraction in speech recognition task, which are Mel-Frequency Ceptrum Coefficient (Mfcc) and log mel spectrogram (log-mel). We use residual neural network (resnet) and Dense Convolutional Network (densenet), which are the convolution neural network based, as classifier. The results shown that the log-mel feature outperform the mfcc feature on the False Alarm Speech Rate (FASR). On the other hand, the mfcc feature can achieve the better result than the log-mel in term of Missed Speech Rate (MSR).

Table 4.1 MSR and FASR with general acoustic feature and classifier on the entertainment media datasets.

Speech activity detection		Hollywood movies dataset		AVA speech	
model	no of parameter	MSR	FASR	MSR	FASR
log-mel + resnet96	30M	13.23	8.98	10.55	10.76
log-mel + densenet121	7.2M	13.94	9.18	11.28	8.79
log-mel +modified densenet	0.6M	14.07	9.63	11.63	9.59
mfcc +resnet96	30M	8.74	9.96	9.70	11.91
mfcc + modified densenet	0.6M	10.91	12.02	9.46	11.85
mfcc + densenet121	7.2M	7.52	10.56	9.91	13.65

4.1.1.2 Evaluation on telephone conversation corpus.

We use CALLHOME American English, 2003 NIST Rich Transcription and 2000 NIST Speaker Recognition Evaluation to evaluate the speech activity detection on the telephone conversation domain. The results similar to the experiment on the entertainment media corpus.

Table 4.2 MSR and FASR with general acoustic feature and classifier on telephone conversation datasets.

Speech activity detection		CALLHOME American English		2003 NIST Rich Transcription		2000 NIST Speaker Recognition Evaluation	
model	no of parameter	MSR	FASR	MSR	FASR	MSR	FASR
log-mel + resnet96	30M	9.01	8.31	10.36	9.56	9.1	8.19
log-mel + densenet121	7.2M	7.87	7.87	10.95	10.33	9.88	11.38
log-mel +modified densenet	0.6M	9.41	8.68	8.78	8.88	10.99	10.11
mfcc +resnet96	30M	9.36	8.0	8.81	10.54	8.89	10.52
mfcc + modified densenet	0.6M	9.03	8.54	10.86	8.70	9.98	10.5
mfcc + densenet121	7.2M	9.39	9.29	9.77	11.44	10.94	10.08

4.1.2 The studies of the effect of the dimension of feature in the combined feature on the model training.

Because of the result from the of the effect of acoustic feature extraction and neural network on speech activity detection, we assume that the feature combination collects the advantages of each feature extraction and improve the performance of indicating the speech in audio recording. We first combined 24-dimensional log-mel with 24-dimensional mfcc. We second combined 24-dimensional log-mel with 8 dimentional spectral contrast, 12 dimensional chroma and 6 dimensional tonnetz. And we then combined 24-dimensional mfcc with 8 dimentional spectral contrast, 12 dimentional dimensional chroma and 6 dimensional tonnetz.

4.1.2.1 Evaluation on the entertainment media corpus

In the entertainment media corpus, the combination of the Log-mel, chroma, spectral contrast and tonnetz lead to the lowest FASR. The combination of the MFCC and the other acoustic feature can achieve the lowest MSR in the both datasets.

Table 4.3 the effect of using combined features in speech activity detection process on the entertainment media datasets.

method	Hollywood movies dataset		AVA speech	
	MSR	FASR	MSR	FASR
Log-mel,Mfcc + modified densenet	12.49	11.94	10.79	11.87
Log-mel,chroma, spectral contrast, tonnetz + modified densenet	8.75	7.86	8.96	8.62
mfcc,chroma, spectral contrast, tonnetz + modified densenet	7.98	8.21	7.94	9.34

4.1.2.2 Evaluation on telephone conversation corpus.

The result in the evaluation on telephone is similar to the previous experiment in which the combination of the Log-mel and other features can get the better result in FASR than using only Log-mel or MFCC as input feature. When we combined the MFCC with other features, it is also can achieve the MSR. The combination of MFCC and Log-mel lead us to the worse MSR and FASR.

Table 4.4 the effect of using combined features in speech activity detection process on telephone conversation datasets.

Method	CallHome American English Eval (%)		NIST RT-03 English (%)		2000 NIST Speaker Recognition (%)	
	MSR	FASR	MSR	FASR	MSR	FASR
Log-mel,Mfcc + modified densenet	9.44	8.58	9.25	9.1	9.02	9.45
Log-mel,chroma, spectral contrast, tonnetz + modified densenet	6.83	5.89	6.83	6.5	7.17	6.52
mfcc,chroma, spectral contrast, tonnetz + modified densenet	7.38	6.09	7.37	7.42	6.42	7.24

4.1.3 The studies of the effect of varying dimension of feature in the combined feature with Dempster-Shafer.

After the combination of Log-mel and MFCC get the worse result, we decided to apply the Dempster-Shafer theory for the model's result fusion. We first combined Log-mel with spectral contrast, chroma and dimensional tonnetz. We then combined MFCC with spectral contrast, chroma and dimensional tonnetz. We use the two-feature set as the input of the classifier separately. After we get results from two model, we fused two result with the Dempster-Shafer theory.

In this experiment, we studied the effect of the feature dimension in the combined feature with Dempster-Shafer. The dimension of features was varied and categorized into three feature sets, which were denoted as FS1, FS2, FS3, respectively. The first feature set was the combination of 24-dimensional log-mel, 24-dimensional mfcc, 16-dimensional chroma, 16-dimensional tonnetz, and 8-dimensional spectral contrast. The second feature set consists of 32-dimensional log-mel, 32-dimensional mfcc, 8-dimensional chroma, 8-dimensional tonnetz, and 16- dimensional spectral contrast. The final feature set only had 64-dimensional log-mel and 64-dimensional mfcc.

4.1.3.1 Evaluation on the entertainment media corpus

This experiment demonstrates the model fusion by using DS theory with the appropriate dimension feature can lead to the lowest MSR and FASR in both two datasets.

Table 4.5 the effect of varying dimension of feature in the combined feature with Dempster-Shafer on the entertainment media corpuses.

Method	Hollywood movies dataset (%)		AVA speech (%)	
	MSR	FASR	MSR	FASR
FS1 + modified densenet	8.78	3.4	5.93	6.64
FS2 + modified densenet	4.95	5.42	3.58	4.55
FS3 + modified densenet	5.78	10.4	6.56	9.4

4.1.3.2 Evaluation on telephone conversation corpus.

The result on the telephone conversation corpus is similar to the result on entertainment media corpus. With DS theory, SAD performance was improved in terms of MSR and FASR.

Table 4.6 the effect of varying dimension of feature in the combined feature with Dempster-Shafer on telephone conversation corpuses.

Method	CallHome American English Eval (%)		NIST RT-03 English (%)		2000 NIST Speaker Recognition (%)	
	MSR	FASR	MSR	FASR	MSR	FASR
FS1 + modified densenet	3.68	4.16	2.41	4.03	4.02	2.31
FS2 + modified densenet	2.51	3.20	1.38	2.95	2.52	4.12
FS3 + modified densenet	5.63	6.84	7.66	8.14	5.44	6.14

4.1.4 The studies of the effect of varying the number of multiple segments for segment level prediction on different testing sets.

In this experiment, we study the effect of varying the number of segments for multiple segment prediction. We varied the number of multiple segment prediction and use majority prediction after we get the single segment prediction.

4.1.4.1 Evaluation on the entertainment media corpus

The result of this evaluation show that the using multiple segment on the prediction can improve the performance of SAD. In the Holly wood movies dataset, we can achieve the best MSR and FASR with 800 millisecond segment length. And we can get the greatest MSR and FASR when 640 milliseconds segment length was chosen for the prediction.

Table 4.7 the effect of varying the number of multiple segments for segment level prediction on entertainment media testing sets.

Method	No. of segment	Segment	Hollywood movies dataset (%)		AVA speech (%)	
		length (ms)	MSR	FASR	MSR	FASR
FS2 + modified densenet	N = 1	640	5.95	6.42	3.58	4.55
	N = 2	720	4.95	6.42	2.58	8.45
	N = 3	800	5.27	5.18	3.43	7.85
	N = 4	880	7.14	5.46	4.47	5.46
	N = 5	960	6.49	4.87	4.55	3.59

4.1.4.2 Evaluation on telephone conversation corpus.

On the telephone conversation corpus, the result also shows that the segment length can affect to the prediction result. Because of using the 640 milliseconds prediction length on NIST RT-03 English and 2000 NIST Speaker Recognition, we can get the better SAD result than the other segment length. We can get the better result on the CallHome American English Eval dataset, when the prediction segment length is 880 milliseconds.

Table 4.8 effect of varying the number of multiple segments for segment level prediction on telephone conversation testing sets.

Method	No of segment	CallHome American English Eval (%)		NIST RT-03 English (%)		2000 NIST Speaker Recognition (%)	
		MSR	FASR	MSR	FASR	MSR	FASR
FS2 + modified densenet	N = 1	2.51	3.20	1.38	2.95	2.52	4.12
	N = 2	2.79	2.83	2.61	3.37	3.68	2.26
	N = 3	2.53	2.96	2.73	3.25	3.61	2.13
	N = 4	2.45	2.22	2.88	3.16	3.81	2.33
	N = 5	2.88	2.71	3.06	3.55	3.82	2.83

4.2 Evaluation of speaker change detection

Subtitle-Aligned Movie (SAM) Corpus was used as a training set to build the speaker change model for finding the actual time that the speaker change. The speech signal was sampled at 16 kHz with 16-bit resolution. From the statistic of the SAM corpus (Hebbar, Somandepalli & Narayanan, 2019), We chose a segment of length 1. 28 second with a duration of pause 0. 9 seconds. Thus, we received around 18,000 the speech segment, which contain more than one speaker. We used segments of length 640 ms with 64 frames to extract 64-dimensional combined features in order to extract the cosine similarity between the previous window and the window. We then use the array of the similarity as the input of our network for determining the continuity of the speech signal. We used the overlap percentage of 87. 5%, as suggested by Hebbar, Somandepalli and Narayanan (2019). Models are evaluated on the Hollywood movie dataset, AVA-speech dataset, CALLHOME American English, 2003 NIST Rich Transcription and 2000 NIST Speaker Recognition Evaluation at segment-label.

In this section, we study the effect of acoustic feature extraction and neural network on speech activity detection.

4.2.1 Evaluation on the entertainment media corpus

The combination of chroma and spectral contrast with Densenet architecture can improve the speaker change performance on the entertainment media corpus. And the acoustic feature combination with Densenet 121 layers can achieve a slightly better result than the combined feature with modified Densenet.

Table 4.9 the effect of using combined features in speaker change detection process on the entertainment media datasets.

Speaker Change detection			Speaker error rate	
Feature extraction	Classifier		Hollywood movie dataset	AVA-speech dataset
	model	No. of parameter		
Chroma	Densenet121	7.2M	2.23	3.91
Spectral contrast	Densenet121	7.2M	2.67	2.47
Chroma + spectral contrast	Densenet121	7.2M	1.22	1.87
Chroma	modified densenet	0.6M	2.56	3.13
Spectral contrast	modified densenet	0.6M	3.15	2.35
Chroma + spectral contrast	modified densenet	0.6M	1.88	1.95

4.2.2 Evaluation on telephone conversation corpus.

On the telephone conversation domain, the feature combination did not take much effect on the speaker change performance. The result of chroma with Densenet 121 layers is slightly better than the combined feature on the 2000 NIST Speaker Recognition Evaluation. And the combination of chroma and spectral contrast get the slightly better result than other methods CALLHOME American English and 2003 NIST Rich Transcription.

Table 4.10 the effect of using combined features in speaker change detection process on telephone conversation datasets.

Speaker Change detection			Speaker error rate		
Feature extraction	Classifier		CALLHOME American English	2003 NIST Rich Transcription	2000 NIST Speaker Recognition Evaluation
	model	No.of parameter			
Chroma	Densenet121	7.2M	1.36	1.73	1.04
Spectral contrast	Densenet121	7.2M	1.99	1.4	1.25
Chroma + spectral contrast	Densenet121	7.2M	1.06	1.04	1.61
Chroma	modified densenet	0.6M	1.18	1.3	1.57
Spectral contrast	modified densenet	0.6M	1.49	1.5	1.46
Chroma + spectral contrast	modified densenet	0.6M	1.17	1.07	1.48

4.3 Evaluation of Speaker clustering

Subtitle-Aligned Movie (SAM) Corpus and the CSTR VCTK Corpus were used as a training set to build the speech activity detection model for indexing the speech segment in the audio recording. The speech signal was sampled at 16 kHz with 16-bit resolution. From the statistic of the SAM corpus corpus (Hebbar, Somandepalli & Narayanan, 2019) and the CSTR VCTK Corpus, we received around 3,500 different speakers. Models are evaluated on the Hollywood movie dataset, AVA-speech dataset, CALLHOME American English, 2003 NIST Rich Transcription and 2000 NIST Speaker Recognition Evaluation at segment-label.

In this section, results for two independent experiments has been shown. the effect of acoustic feature extraction and neural network on speech activity detection.

4.3.1 Evaluation on the entertainment media corpus

On the media corpus, we noticed that the combination of Log-mel and MFCC can achieve a better speaker error rate on the speaker clustering process than using other acoustic features. We can get the lowest speaker error rate on the Hollywood movies dataset and AVA speech dataset with Densenet 121 layers.

Table 4.11 the effect of using combined features in speaker clustering on entertainment media datasets.

Speaker Clustering			Speaker error rate	
Feature extraction	Classifier		Hollywood movies dataset	AVA-speech dataset
	model	no of parameter		
Log-mel	Densenet121	7.2M	7.12	5.52
mfcc	Densenet121	7.2M	5.92	6.82
Log-mel + mfcc	Densenet121	7.2M	3.24	4.16
Log-mel	modified densenet	0.6M	6.05	5.95
mfcc	modified densenet	0.6M	7.15	6.37
Log-mel + mfcc	modified densenet	0.6M	4.39	4.36

4.3.2 Evaluation on telephone conversation corpus.

On the telephone corpus, we can get the lowest speaker error rate on CALLHOME American English by using the combination of Log-mel and MFCC with modified Densenet. We can achieve the lowest speaker error rate on 2003 NIST Rich Transcription and 2000 NIST Speaker Recognition Evaluation by using the combination feature with Densenet 121 layers.

Table 4.12 the effect of using combined features in speaker clustering on telephone conversation datasets.

Speaker Clustering			Speaker error rate		
Feature extraction	Classifier		CALLHOME American English	2003 NIST Rich Transcription	2000 NIST Speaker Recognition Evaluation
	model	no of parameter			
Log-mel	Densenet121	7.2M	4.39	4.83	4.21
mfcc	Densenet121	7.2M	5.34	4.53	4.09

Log-mel + mfcc	Densenet121	7.2M	3.76	3.55	3.38
Log-mel	modified densenet	0.6M	4.91	4.84	4.82
mfcc	modified densenet	0.6M	5.1	4.46	5.05
Log-mel + mfcc	modified densenet	0.6M	3.44	3.56	3.78

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

The aim of this thesis was to study the state-of-the-art techniques in speaker diarization for specific application to broadcast news audio recordings and develop a python based system. The proposed system has been evaluated using the diarization error rate metric and presented with new additions in combination feature technique. The system has three main precesses which is speech activity detection, speaker change detection and speaker clustering. The system has been evaluated for five speech databases, which is Hollywood movie dataset, AVA speech dataset, CALLHOME American English, 2003 NIST Rich Transcription and the 2000 NIST Speaker Recognition Evaluation.

The general purpose speech activity detection is capable of removing silences as well as audible nonspeech such as music from a recording. The speaker clustering process allows for representing segments with using the combination of Log-mel and MFCC as input of the modified Densenet, which can facilitate further work in fast diarization.

The system can perform speech activity detection with using external training data. From the studies, we can conclude that multiple segment prediction affects to the prediction performance in the SAD process. The difference of segment length depends on the speech length in each dataset. The four-feature combination has been used as the input of classifiers to construct silence, music and other non-speech sounds from the audio recording. A competitive speech activity detection has been achieved with using the DS theory in the model fusion. The results are comparable to a state-of-the-art based speech activity detection which only uses MFCC as the input of the neural network.

The lowest MSR and FASR in the speech activity detection can lead to lower diarization error rate. The improvement of the speech activity detection also led to the optimum of speaker change detection classifier and speaker clustering classifiers.

The combination of Log-mel and MFCC speaker models provide a low dimensional representation of the speaker information compared to traditional GMM

speaker models. They also offer a computational advantage since distance computation between the acoustic feature combination is much faster compared to cross-likelihood based similarity computation on the speaker models.

It has been verified in this thesis that speaker clustering is achieved better using a modified Densenet approach to reach the optimum set of speaker clusters rather than the traditional greedy optimization approach of the hierarchical agglomerative clustering (HAC) algorithm. HAC is computationally very expensive, and an erroneous merge step during the clustering significantly affects the later iterations i.e., error gets propagated. To verify the better performance of proposed system compared to the other was implemented for the diarization error rate (DER).

5.2 Future work

Future work on the system development should focus on the following aspects of speaker diarization:

Refinement of the diarization output by passing it through a Viterbi decoder should be attempted.

Cross-show diarization is the task of performing speaker clustering across different recordings to identify segments of the same speakers in different shows. Current momentum of diarization research is along solving this problem for large databases. Cross diarization should be attempted using the proposed PYTHON-BASED system

Improvements in ILP have shown sufficiently faster implementations by reducing the redundancies in the original ILP, although PYTHON-BASED does not support solving these optimization problems. Solvers such as GUROBI provide support to solving advanced integer linear programs.

It was observed that during the speaker clustering, the segments having background music were unable to show similarity with segments having clean background using the MFCC-GMM speaker models owing to the low SNR. Even after using background variability compensation techniques on i-vector speaker models, the problem persists. Speech enhancement and singing voice separation prior to parameterizing the audio recording should be attempted so that music in the background of a speaker is suppressed.

REFERENCES

Beecks, C., Zimmer, A.M., Kirchhoff, S., & Seidl, T. (2011). Modeling image similarity by Gaussian mixture models and the Signature Quadratic Form Distance. 2011 International Conference on Computer Vision, 1754-1761.

Bellagha, M.L., Labidi, M., & Maraoui, M. (2017). Speaker segmentation using adapted GMMs. 2017 International Conference on Engineering & MIS (ICEMIS), 1-6.

Bernard, G., Galibert, O., & Kahn, J. (2014). The second official REPERE evaluation. SLAM@INTERSPEECH.

Bouafif, L., & Ellouze, N. (2019). Speech-Music-Noise Discrimination in Sound Indexing of Multimedia Documents. Cmc-computers Materials & Continua, 61, 2-10.

Bozonnet, S., Evans, N., & Fredouille, C. (2010). The lia-eurecom RT'09 speaker diarization system: Enhancements in speaker modelling and cluster purification. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 4958-4961.

Bredin, H. (2017). pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. INTERSPEECH.

Chan, W., Lee, T., Zheng, N., & Ouyang, H. (2006). Use of Vocal Source Features in Speaker Segmentation. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 1, I-I.

Chen, S. (1998). Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion.

Curelaru, F. (2018). Front-End Factor Analysis For Speaker Verification. 2018 International Conference on Communications (COMM), 101-106.

Dawalatabad, N., Madikeri, S., Sekhar, C., & Murthy, H. (2016). Two-Pass IB Based Speaker Diarization System Using Meeting-Specific ANN Based Features. INTERSPEECH.

Delgado, H., Miró, X., Fredouille, C., & Serrano, J. (2015). Improved binary key

speaker diarization system. 2015 23rd European Signal Processing Conference (EUSIPCO), 2087-2091.

Dupuy, G., Meignier, S., Deléglise, P., & Estève, Y. (2014). Recent Improvements on ILP-based Clustering for Broadcast News Speaker Diarization. Odyssey.

Favre, B., Damnati, G., Béchet, F., Bendris, M., Charlet, D., Auguste, R., Ayache, S., Bigot, B., Delteil, A., Dufour, R., Fredouille, C., Linarès, G., Martinet, J., Senay, G., & Tirilly, P. (2013). PERCOLI: A Person Identification System for the 2013 REPERE Challenge. SLAM@INTERSPEECH.

Friedland, G., Janin, A., Imseng, D., Miró, X., Gottlieb, L., Huijbregts, M., Knox, M., & Vinyals, O. (2012). The ICSI RT-09 Speaker Diarization System. IEEE Transactions on Audio, Speech, and Language Processing, 20, 371-381.

Friedland, G., Vinyals, O., Huang, Y., & Mueller, C. (2009). Prosodic and other Long-Term Features for Speaker Diarization. IEEE Transactions on Audio, Speech, and Language Processing, 17, 985-993.

Hebbar, R., Somandepalli, K., & Narayanan, S.S. (2019). Robust Speech Activity Detection in Movie Audio: Data Resources and Experimental Evaluation. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4105-4109.

Johnson, S.E., & Woodland, P. (2000). A method for direct audio search with applications to indexing and retrieval. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), 3, 1427-1430 vol.3.

Kingma, D.P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. CoRR, abs/1412.6980.

Labrunie, M., Badin, P., Voit, D., Joseph, A., Frahm, J., Lamalle, L., Vilain, C., & Boë, L. (2018). Automatic segmentation of speech articulators from real-time midsagittal MRI based on supervised learning. Speech Commun., 99, 27-46.

Lin, T., Goyal, P., Girshick, R.B., He, K., & Dollár, P. (2020). Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42, 318-327.

Liu, D., & Kubala, F. (1999). Fast speaker change detection for broadcast news

transcription and indexing. EUROSPEECH.

Liu, Z., Wang, Y., & Chen, T. (1998). Audio Feature Extraction and Analysis for Scene Segmentation and Classification. Journal of VLSI signal processing systems for signal, image and video technology, 20, 61-79.

Lopez-Otero, P. (2015). Improved strategies for speaker segmentation and emotional state detection.

Luque, J., Miró, X., Temko, A., & Hernando, J. (2007). Speaker Diarization for Conference Room: The UPC RT07s Evaluation System. CLEAR.

Maganti, H.K., Motlícek, P., & Gatica-Perez, D. (2007). Unsupervised Speech/Non-Speech Detection for Automatic Speech Recognition in Meeting Rooms. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 4, IV-1037-IV-1040.

Miró, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. IEEE Transactions on Audio, Speech, and Language Processing, 20, 356-370.

Miró, X., Wooters, C., & Pardo, J. (2006). Robust speaker diarization for meetings: ICSI RT06s evaluation system. INTERSPEECH.

Moraru, D., Meignier, S., Fredouille, C., Besacier, L., & Bonastre, J. (2004). The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, I-373.

Narváez, P., Vera, K., Bedoya, N., & Percybrooks, W. (2017). Classification of heart sounds using linear prediction coefficients and mel-frequency cepstral coefficients as acoustic features. 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 1-6.

Nguyen, T., Siong, C.E., & Li, H. (2008). T-test distance and clustering criterion for speaker diarization. INTERSPEECH.

Panagiotakis, C., & Tziritas, G. (2005). A speech/music discriminator based on RMS and zero-crossings. IEEE Transactions on Multimedia, 7, 155-166.

Reynolds, D., & Torres-Carrasquillo, P. (2004). The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone

Conversations.

Reynolds, D., Quatieri, T., & Dunn, R. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Process., 10, 19-41.

Rouvier, M., & Meignier, S. (2012). A global optimization framework for speaker diarization. Odyssey.

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., & Meignier, S. (2013). An open-source state-of-the-art toolbox for broadcast news diarization. INTERSPEECH.

Ryant, N., Liberman, M., & Yuan, J. (2013). Speech activity detection on youtube using deep neural networks. INTERSPEECH.

Sadjadi, S.O., Slaney, M., & Heck, L. (2013). MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research.

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D., & Glass, J.R. (2011). Exploiting Intra-Conversation Variability for Speaker Diarization. INTERSPEECH.

Silovský, J., & Prazak, J. (2012). Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4193-4196.

Trancoso, I., & Redol, R.A. (2009). A BROADCAST NEWS PROCESSING CHAIN FOR SEVERAL VARIETIES OF PORTUGUESE.

Vijayasenan, D., Valente, F., & Bourlard, H. (2007). Agglomerative information bottleneck for speaker diarization of meetings data. 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), 250-255.

Zajíc, Z., Soutner, D., Hrúz, M., Müller, L., & Radová, V. (2018). Recurrent Neural Network Based Speaker Change Detection from Text Transcription Applied in Telephone Speaker Diarization System. TSD.

Zelenák, M., Schulz, H., & Pericás, F.J. (2010). Albayzin 2010 Evaluation Campaign: Speaker Diarization.

Zhu, X., Barras, C., Meignier, S., & Gauvain, J. (2005). Combining speaker identification and BIC for speaker diarization. INTERSPEECH.

BIOGRAPHY

Name	Ms. Pantid Chantangphol
Date of Birth	May 16, 1995
Education	2018: Bachelor of Engineering (Robotics and Automation Engineering) Institute of Field Robotics King Mongkut's University of Technology Thonburi 2020: Master of Engineering (Information and Communication Technology for Embedded System) Sirindhorn International Institute of Technology Thammasat University

Publications

Chantangphol, P., Usanavasin, S., Karnjana, J., Boonkla, S., Keerativittayanun, S., Rugchatjaroen, A., & Shinozaki, T. (2020). Speech Activity Detection Using a Fusion of Dense Convolutional Network in the Movie Audio. 2020 35th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), 9-14.