SPEAKER DIARIZATION

By

ANSHUL PATEL 16BIT016
ANUJ SHAH 16BIT017

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
Ahmedabad 382481

SPEAKER DIARIZATION

Minor Project

Submitted in fulfilment of the requirements

For the degree of

Bachelor of Technology in Information Technology

By

ANSHUL PATEL 16BIT016
ANUJ SHAH 16BIT017

Guided By
PROF. SAPAN MANKAD
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
Ahmedabad 382481

CERTIFICATE

This is to certify that the project entitled "Playback Attack Detection On Speaker Verification Systems" submitted by ANSHUL PATEL (16BIT016) and ANUJ SHAH (16BIT017) towards the partial fulfilment of the requirements for the degree of Bachelor of Technology in Information Technology of Nirma University is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination.

Prof. Sapan Mankad
Assistant Professor
Department of Computer Science & Engg.,
Institute of Technology,
Nirma University,
Ahmedabad

Dr. Madhuri Bhavsar
Dept. of Computer Science & Engg.,
Institute of Technology,
Nirma University,
Ahmedabad

ACKNOWLEDGEMENT

This Minor Project was bolstered by the Department of Computer Science and Engineering, Nirma University. I thank my guide Prof. Sapan Mankad who gave knowledge and ability that extraordinarily helped the minor undertaking by giving help and assigning tasks on "Speaker Diarization".

I also want to thank my task accomplice for unravelling any of my questions and executing the minor project.

I also likewise want to show gratitude to all on the web and hypothetical sources that we had utilized so as to see any idea to its profundity and effectively actualize in our venture.

ABSTRACT

Speaker Diarization is an important task in audio retrieval and processing. Speaker Diarization is the process of determining the activity between different segments of an audio signal. The simplest of the activities can be determining the regions of speech and non speech. Non Speech regions includes background music, silence, laughter, etc. A more advanced version can be to classify the speech regions into speaker labels, that is identifying the total speakers through unsupervised learning and their corresponding time frames when each speaker spoke during the entire speech signal.

In this paper, we will focus on the state of the art methods and techniques required to undertake Speaker Diarization, while discussing their merits as well as disadvantages. In the final section, we will focus on the approach used by us along with the results.

CONTENTS

Certificate

Acknowledgement

Abstract

Table of Contents

List of figures

Chapter 1 Speaker Diarization

1.1 What is Speaker Diarization 7

1.2 Flow of Discussion 7

Chapter 2 Dataset

2.1 Dataset 8

Chapter 3 Methodology(Literature Survey)

3.1 Main Approaches 8

3.1.1 Top-Down Approach

3.1.2 Bottom-Up Approach

3.1.3 Alternative Approach

3.2 Main Algorithms 10

3.2.1 Acoustic Beamforming

3.2.2 Speech Activity Detection

3.2.3 Segmentation

3.2.4 Clustering

3.2.5 One step Segmentation and Clustering

Chapter 4 Feature Extraction

4.1 Features Extracted 11
4.2 Mid term Feature extraction 11
4.3 Linear Discriminant Analysis 12

Chapter 5 Signal Segmentation
5.1 Discussion 14
5.1.1 Model Based Segmentation
5.1.2 Distance Based Methods
5.2 Hybrid Based Segmentation 14

Chapter 6 Clustering
6.1 Discussion 16
6.2 kMeans Clustering
6.2.1 Elbow Method
6.2.2 kMeans Algorithm
6.3 Hierarchical Agglomerative Clustering 16
6.3.1 Dendrograms
6.3.2 Hierarchical Agglomerative Algorithm
Chapter 7 Conclusion 19

REFERENCES 20
Appendix – A List of Useful Websites 20

LIST OF FIGURES

Fig. 3.1: Approaches to Speaker Diarization
Fig. 3.2: Flow of Work
Fig. 4.1: MFCC Representation
Fig. 4.2: Feature Representation
Fig. 6.1: Elbow Method Visual representation
Fig. 6.2: Output after applying kMeans
Fig. 6.3: Dendrograms Visual Representation
Fig. 6.4: Output after applying hierarchical agglomerative

CHAPTER 1: Speaker Diarization

1.1 What is Speaker Diarization?

Speech Processing can be separated into two general classifications: first is recognition of speech, where the substances of sound elements are distinguished and second is the speaker identification which is an assignment of recognizing speakers in a discussion. The Speaker Diarization falls in the second class of speech processing technique where it is required to distinguish the speaker alongside recognizable proof of timespan of the speech expressed by a specific speaker.

Speaker Diarization is the job of distinguishing the beginning and end time of a speaker in a speech-signal file, together with the speaker's identification for example who talked when. It can upgrade the decipherable of an automatic speech transcription by organizing the sound stream into the speaker turns and by giving the speaker's actual identity.

Speaker diarization is a mix of speaker segmentation and speaker clustering. The primary targets discovering speaker change timespan in a sound stream. The latter targets gathering segments of speech based on speakers unique features.

Some applications of Speaker Diarization are:

- Speech-to-text Transcription/ Rich Transcription(RT)
- Broadcast News
- Conference Meetings
- Youtube video automatic caption generation

1.2 Flow of Discussion

We will first go through some of the initial research work on speaker diarization and elaborate on the approach used by other researchers. Next, we will attempt to explain how our framework is unique in relation to the previously existing frameworks.

We will discuss the approach used while cleaning the audio source and later, how to divide the audio into appropriate sized segments. We will focus on calculating the optimal cluster number for clustering and provide an extended discussion on comparison of results using different unsupervised clustering techniques.

CHAPTER 2: The Dataset

2.1 Dataset

The dataset for the purpose of Speaker Diarization can be any audio file consisting of multiple speakers including background music, silence and laughter. Typically, the duration of the audio file is approximately 4-5 minutes. For meeting recordings, however, the duration can be as long as 30 minutes.

The freely available dataset are 'The ICSI Meeting Corpus' and 'The AMI Meeting Corpus'. Both of them contains audio files of several meetings or conferences recorded from different microphones and language of speech being English.

CHAPTER 3: Methodology(Literature Survey)

3.1 Main Approaches

Among all the state-of-the-art systems implementing Speaker Diarization present in the industry, most of them are using either a Bottom-Up Approach or a Top-Down Approach. Both of these methods focuses on reaching an optimum cluster number either by shrinking or expanding the number of clusters they started with.

The bottom-up approach starts with multiple clusters usually greater than estimated speakers number and then merging them one by one based on a metric, we reach the optimum cluster number. While in the case of top-down approach we start with a single cluster and break it down into multiple clusters again on some metric until we reach an optimum cluster number. If number of clusters exceeds the optimum cluster number then it is said to be under-clustering while if it falls short of optimum number the the system is over-clustering.

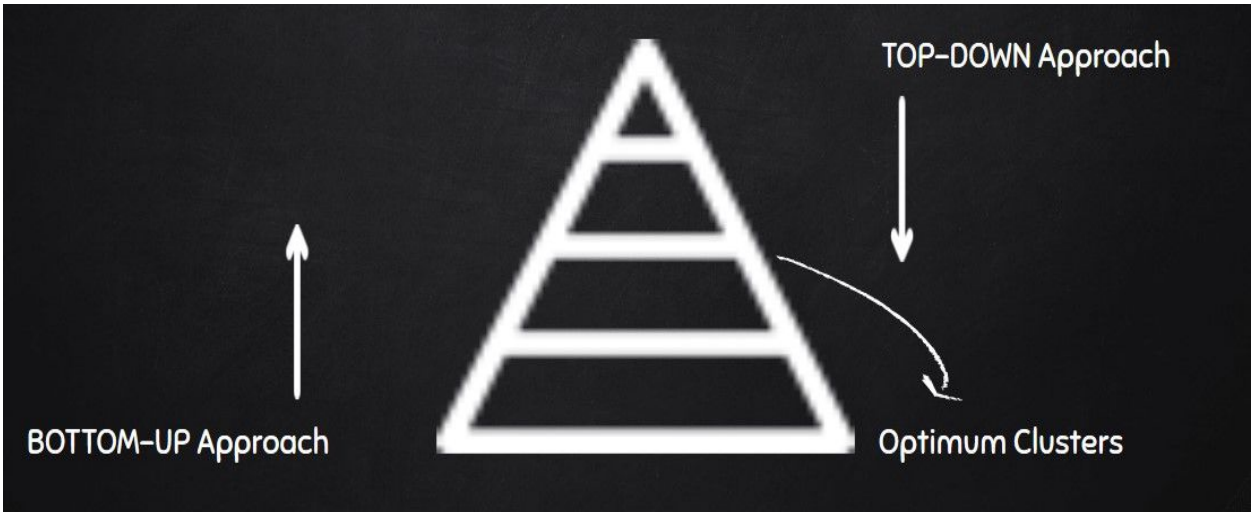

Fig. 3.1: Approaches to Speaker Diarization

Both of these approaches that are Bottom-Up and Top-Down are based on Hidden Markov Model in which each stage is represented by a Gaussian Mixture Model and it corresponds to a particular speaker. Speaker turns are defined as transition between states in a Hidden Markov Model.

Below in this chapter we will provide a description of the above two mentioned major approaches along with newly introduced alternatives which have shown great potential on the benchmark NIST RT evaluations.

3.1.1 Top-Down Approach:

The approach firstly fits the whole speech audio on a single speaker GMM model and then iteratively introduces new models by segmenting the original single model until it reaches the optimum cluster number. The segmentation for generating clusters is done with interleaved Viterbi Realignment and Adaptation, however it is a difficult task since we have to deal with unlabeled segments and then classify them into new segments.

Top-Down Approach is less popular than Bottom-Up Approach, but they are more computationally efficient and the results can be further enhanced by using cluster purification techniques.

3.1.2 Bottom-Up Approach:

The approach initially trains a huge number of clusters and the iteratively merges them until an optimum cluster number is reached which is one cluster assigned to a unique speaker in the speech signal. Clusters are modeled using the Gaussian Mixture model and such multiple models are merged into a single new model. The reassignment of segments to clusters is done using Viterbi Realignment Algorithm. The stopping criteria involves a metric that uses a threshold value like Bayesian Information Criterion, generalized likelihood ratio, and the Kullback–Leibler based metrics.

3.1.3 Alternative Approach:

The approach that has been recently gained acknowledgement is inspired from theory of rate-distortion which is based upon framework of information-theoretic. It is basically a bottom-up approach in nature. It is totally non parametric and its outcomes have been demonstrated to be equivalent to those of best in class parametric frameworks, with noteworthy savings in calculation. Clustering depends on shared information, which gauges the common reliance on two factors.

Initially, a single Gaussian Mixture Model is tuned for the full sound stream, and common information is processed in another space of applicable factors characterized by the GMM components. The methodology targets limiting the loss of shared information between progressive clusterings while saving however much

information as could be expected from the first dataset. Examples of this approach is Sequential Information Bottleneck and Agglomerative Information Bottleneck.

3.2 Main Algorithms:

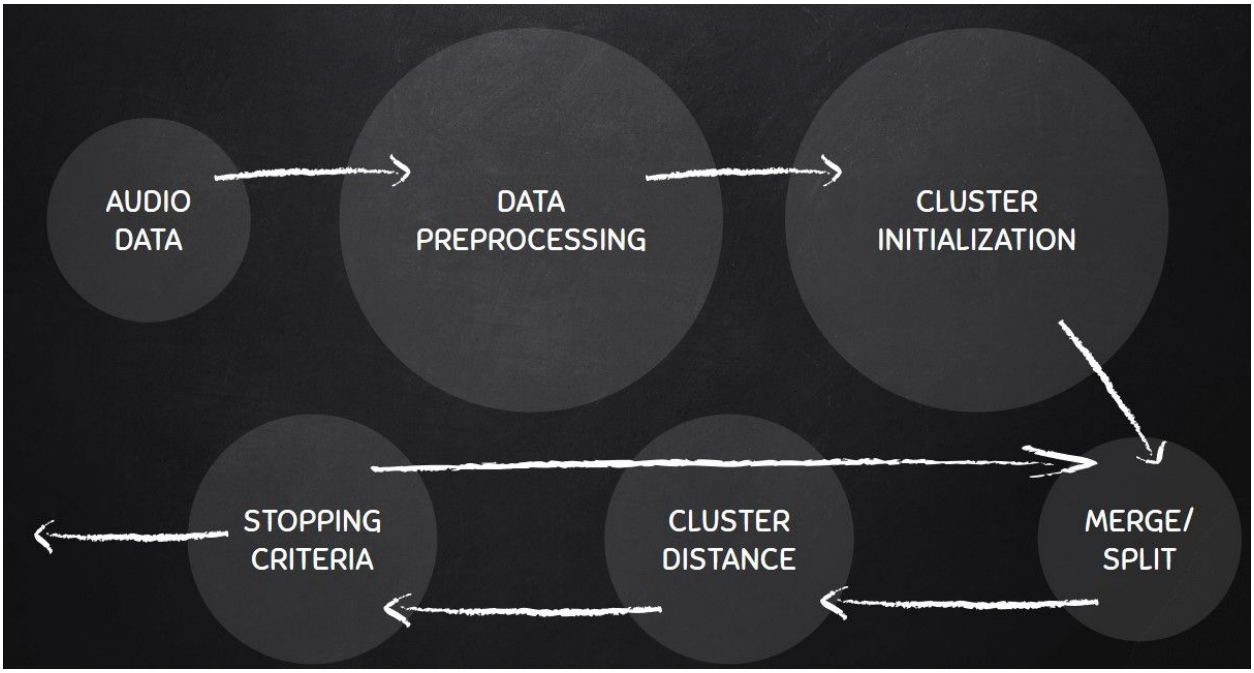

Fig. 3.2: Flow of work

The above figure represents the general architecture of a Speaker Diarization System. Audio Signal that may contain the speech recording of a meeting or a conference is passed to the preprocessing stage where techniques such as noise reduction, extraction of acoustic features such as Mel Frequency Cepstral Coefficients(MFCC) or Perceptual Linear Prediction (PLP), acoustic beamforming, and silence removal using speech activity detection. The next stage initializes clusters based on the approach taken mainly either Bottom-Up or Top-Down. Then an iterative processing occurs where the merge/split and cluster distance calculation is done until it satisfies the stopping criteria. Data purification can also be applied to increase the discriminative power of the system forming the clusters.

3.2.1 Acoustic Beamforming:

Speaker Diarization finds its main use case in conference meeting domain where it has to deal with multiple microphones that record the speech. NIST RT'04 was the first time the Multiple Distant Microphone(MDM) was introduced. To address the problem, we independently treat each channel and then merge the output into a single one. For merging, a two-axis merging technique is used which takes the longest speaker segments detected and iterates over the output of segmentation.

Recently, a late-stage fusion is a technique is used which segments speaker from only channels having the best ratio of signal-to-noise. In the next step, we can merge all the channels into one channel.

3.2.2 Speech Activity Detection:

The algorithm is used in separating speech segments from silence segments in the audio signal that allows to preprocess it required for further processing. Instead of using a feature extractor model for separating the segments, we rather use a classifier such as LDA or Support-Vector Machine. However due to their reliability on training data it causes a setback in achieving better results. In general, a system is said to have bad results if they have a higher Diarization Error Rate(DER) which is caused partly due to poor Speech Activity Detection.

3.2.3 Segmentation:

The critical part to any Diarization system is the segmentation stage which tries to attempt a split on training audio dataset in order to detect speaker transitions also called speaker turns. The old-style way to deal with segmentation plays out a theory-testing utilizing the acoustic portions in two sliding and conceivably covering, successive windows. For each considered change point there are two potential speculations: first that the two sections originate from a similar speaker (H0), and in this way that they can be very much spoken to by a solitary model; and second that there are two unique speakers (H1), and along these lines that two distinct models are progressively suitable.

Usage of Bayesian Information Criteria with its BIC Metric has gained popularity due to its methodology requiring the setting of a penalty term which controls the tradeoff between missed turns and those erroneously distinguished. However, to identify a correct penalty term is a difficult process. In contrast to BIC-metric an algorithm that uses Generalized Likelihood ratio as the metric which is a ratio between H0 and H1 provides concrete results. Recently, a newly developed algorithm known as Information Change Rate which is based on the principle of entropy that may be used to compare the similarity between a pair of neighboring segments. It has proven to be robust and an efficient algorithm when dealing with asymmetric data.

3.2.4 Clustering:

The next step after generating segments is to identify and merge the same speaker segments into their respective clusters. The general methodology includes Viterbi realignment where the speech signal is resegmented dependent on the present bunching speculation before the models are retrained on the new divisions.

An elective way to deal with clustering includes larger part voting a ballot whereby short windows of frames are completely relegated to the nearest bunch, i.e., that

which pulls in the most segments while decoding. This method results in saving resources in calculation yet is increasingly fit to on the web or live speaker diarization frameworks.

3.2.5 One Step Segmentation and Clustering:

Several advanced methodologies have now started combining the steps of segmentation and clustering into single step. It provides an early advantage in handling early errors while training the system and successively correcting them in resegmentation steps later on. However, it's performance is slower than two-stage performed segmentation and clustering.

CHAPTER 4: Feature Extraction

4.1 Features Extracted

For the purpose of diarization, we have extracted a total of 34 features. The entire audio signal is divided into frames each having a duration of 0.05*Sampling frequency(fs). All the 34 features are extracted for every frame present as part of the audio signal. The 34 features extracted are given below:

- **Zero Crossing Rate:** Signifies the change in signals from either +ve to -ve and vice-versa.
- **Energy:** The fourier transform of the signal in time domain.
- **Energy Entropy:** Entropy of the energy feature.
- **Spectral Centroid:** It is a measure of spectral position and acts as a centre of gravity of the spectrum.
- **Spectral Spread:** It is a measure of spectral shape and acts as the central moment of the spectrum.
- **Spectral Entropy:** Power Spectral density of your signal.
- **Spectral Flux:** How fast are the changes in the power spectrum.
- **Spectral Rolloff:** Total amount of frequencies below a threshold spectral energy, normally, 85%.
- **13 Mel Frequency Cepstral Coefficients:** Imitate the features responsible for generating the voice in humans. The frequency are scaled on the mel scale.

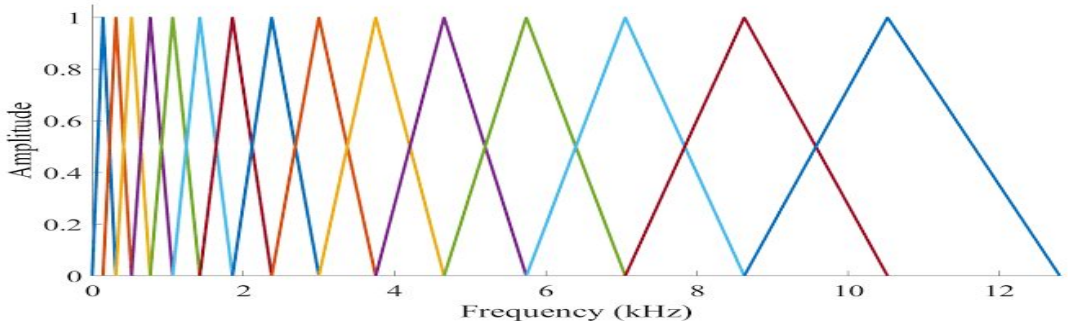

Fig.4.1 : MFCC representation

- **12 Chroma Features:** The entire spectrum is projected into 12 bins representing the 12 distinct pitch classes.

- **Chroma Deviation:** Standard deviation of the 12 chroma features.

4.2 Mid Term feature Extraction

The features mentioned above are the short term features. Now , we will calculate the mid term features. The number of the mid term features is twice the number of short term features. This is due to the fact that the mid term features are actually two statistics of the short term features, namely, average value and the standard deviation. What this means, that for every feature we will calculate the mean and standard deviation over the entire frame.

Fig. 4.2: Total features

The first half of the features will represent the mean and the later half represents the median.

4.3 Linear Discriminant Analysis

The feature extraction step will give us a total of 64 features, 34 mean and 34 standard deviation. Linear Discriminant Analysis is a dimensionality reduction technique used to map the high dimensional space to a lower dimensional space. The objective is to maximise the variance, only considering axes that capture more information.

For our problem, our 68 dimensional features are converted into lower dimensions space of 35 features. The number of features in the lower dimensions have been selected as a result of extensive in [4].

CHAPTER 5: Signal Segmentation

5.1 Discussion:

Speaker Segmentation refers to the process of dividing our signal into multiple frames and associate each segment with a particular speaker. However, before doing so. We have to divide our audio signals into segments. Segmentation, in case of speaker diarization can be performed using several methods. The simplest form of diarization can be to identify the change in gender or bandwidth . The different methods are described below

5.1.1 Model Based Segmentation

In this type of segmentation, set of models is trained for different speakers using a classification model. This approach can be effective if we have information about the total speakers number, there is no other speaker present other than that present in the dataset. In such a case, the unknown audio segments are classified and the change points are identified as the points where there is a change of the speaker class.

5.1.2 Distance Based Methods

These type of method is the most common and most widely used for the purpose of indexing the segments. In this, a distance metric is used between two consecutive segments is analysed to determine the change points. The change points points are detected by comparing the distance with a particular threshold which is empirically tuned for changes in audio type and features.

5.2 Hybrid Speaker Segmentation

Hybrid speaker segmentation is a combination of distance based and model based techniques. Initially, pre-segmentation of the input audio signal is done using distance based segmentation. The results are used to create speaker models. Then, to achieve refined segmentation, model based resegmentation is performed.

CHAPTER 6: Clustering

6.1 Discussion

The goal of the clustering is to identify segments belonging to the same speaker together. Speaker clustering produces a result of one cluster for each speaker. However, before associating each segment with a particular speaker, our aim is to identify the speaker's present or the number of clusters representing each of the speakers involved.

6.2 kMeans Clustering

6.2.1 Elbow Method

It is the most commonly used technique for determining the optimal cluster number for kMeans clustering. In this method, the sum of squares at each number of clusters in calculated and plotted. We focus on the change in the slope from steep to shallow in order to determine the optimal cluster number.

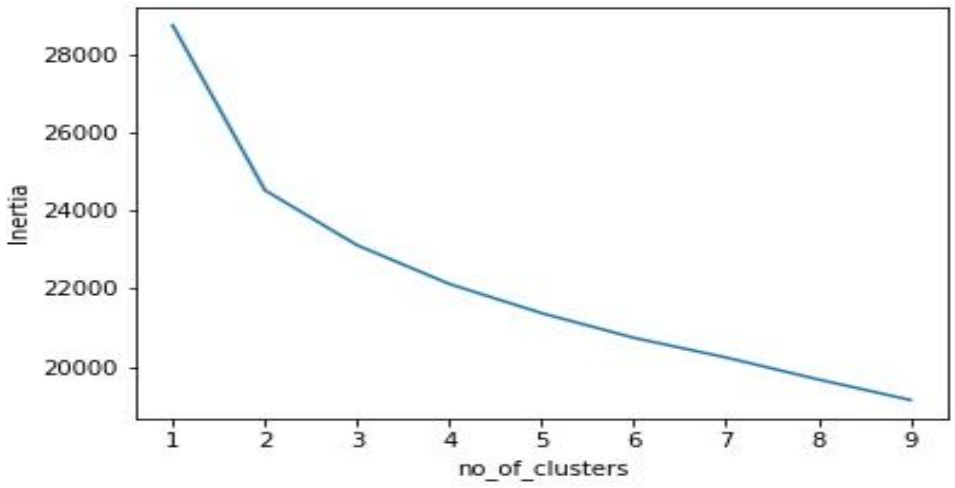

Fig.6.1 : Elbow Method Visual Plot

6.2.2 kMeans Algorithm

In kMeans Algorithm, we will randomly select unique points equal to the number of clusters. This points will be the centroids of the clusters. Then, for each point in the dataset, we will calculate the euclidean distance from each of the centroids of the clusters. The point will be assigned to the cluster from which it’s distance is minimum. Once we have assigned all the points to respective clusters, the centroids will be updated as the mean of the points in respective clusters. Again the same

procedure is repeated until there is no change in the elements of respective clusters.

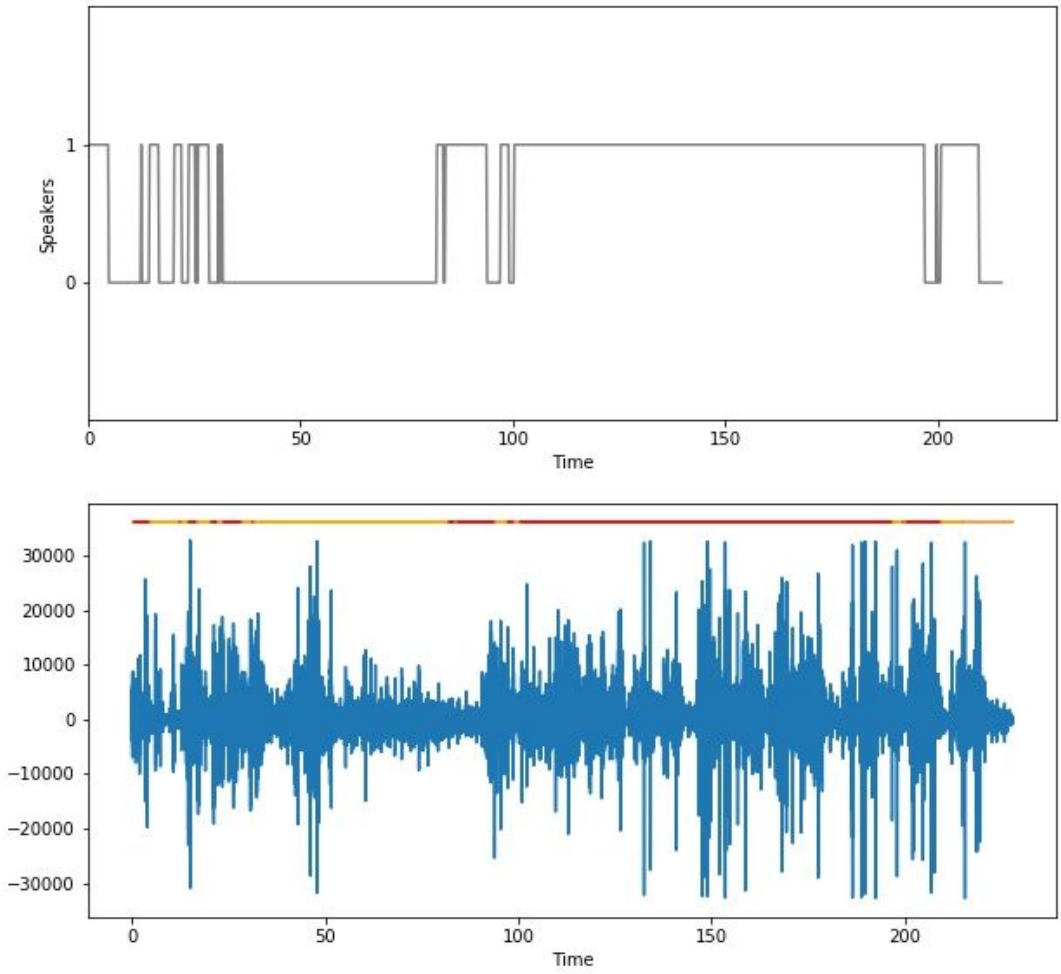

Fig.6.2 : Output after applying kMeans

6.3 Hierarchical Agglomerative Clustering

6.3.1 Dendrograms

It is most normally made as a yield from hierarchical clustering. The fundamental utilization of a dendrogram is to work out the most ideal approach to assign elements to clusters. A dendrogram is an outline that shows the progressive connection between elements.

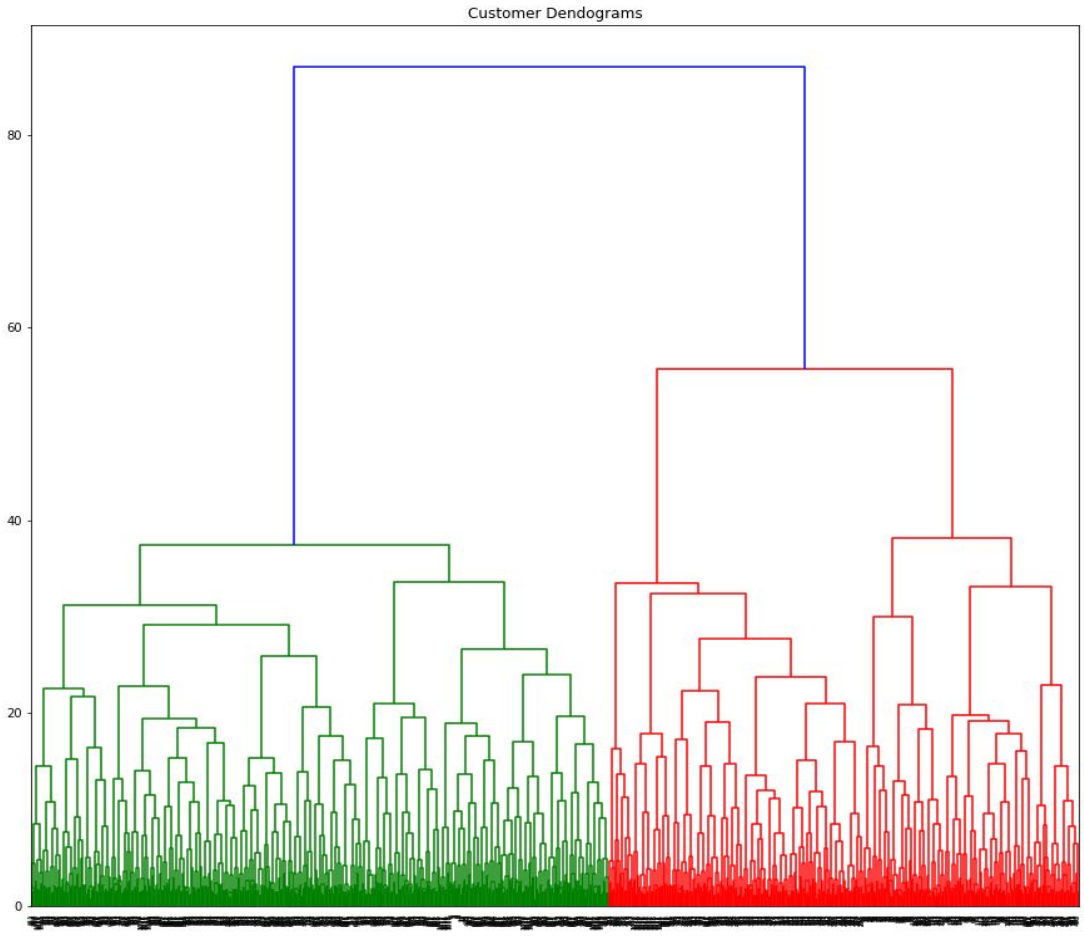

Fig.6.3: Dendrograms representation.

6.3.2 Hierarchical Agglomerative Algorithm

Initially, each point in the dataset is considered as an individual cluster. After each iteration, the closest clusters merge with each other to form the resultant k clusters.

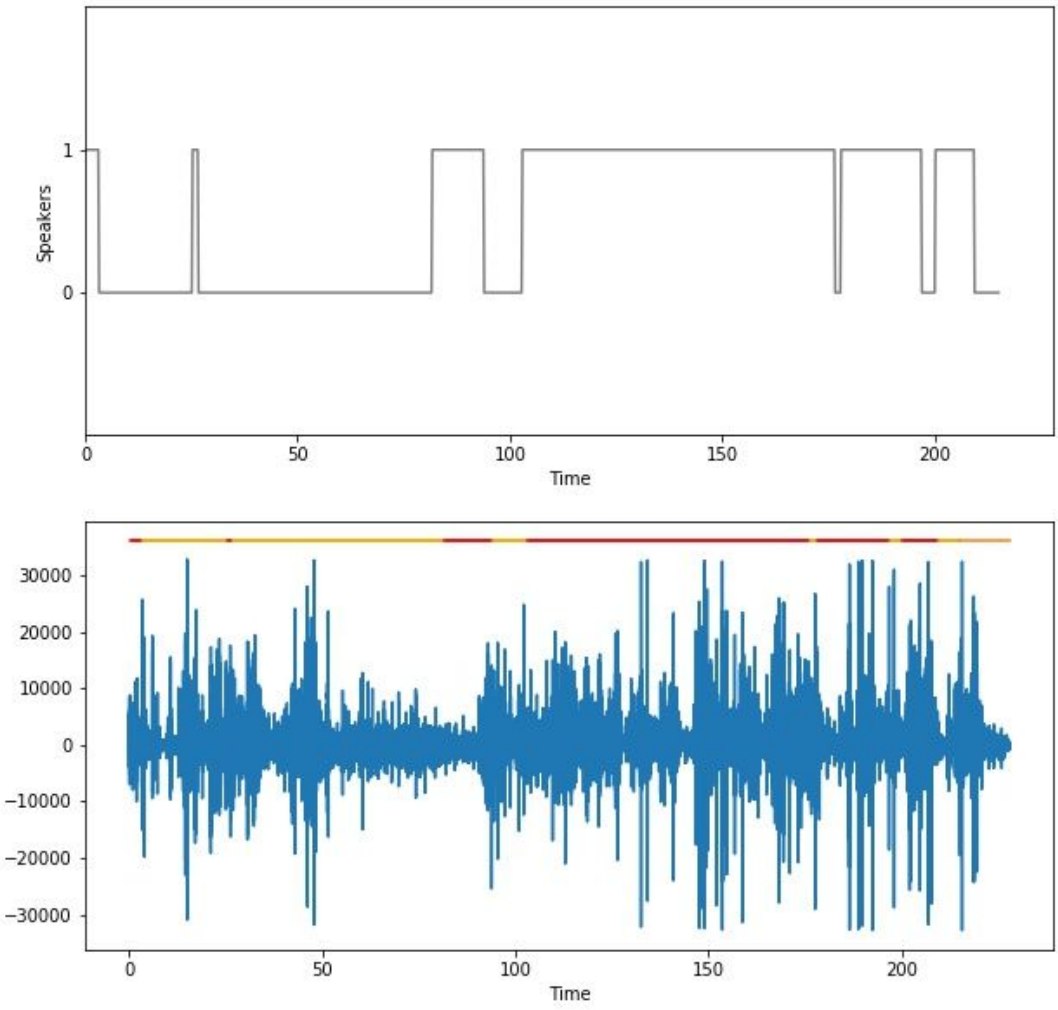

Fig.6.4 : Output After Applying Hierarchical Agglomerative Clustering

CONCLUSION

There has been tremendous progress in Speaker Diarization over the recent years. It can be applied to phone call conversations, broadcast news, and meetings recordings. Moreover, it has led to several by-products. The diarization techniques can further be applied for the betterment of automatic rich text in videos. Furthermore, it can also be identified the speakers in videos and index them helping the user to identify who is speaking at the particular moment.

Overall, the future of Speaker Diarization is brighter and broader than what is currently utilised and there is a scope for large improvements in the area, especially handling of overlapping speech, which needs to be attributed to multiple speakers.

References

[1] *"A review on speaker diarization systems and approaches"* by M.H. Moattar, M.M. Homayounpour at SciVerse ScienceDirect by Elsevier.

[2] *"Speaker Diarization: A Review of Recent Research"* by Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland and Oriol Vinyals.

[3] *"An Overview of Automatic Speaker Diarization Systems"* by Sue E. Tranter and Douglas A. Reynolds.

[4] *"pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis"* by Theodoros Giannakopoulos.

Appendix – A

- https://github.com/tyiannak/pyAudioAnalysis
- https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
- https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6
- https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/