International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-3, February 2019

Multiple Speaker Recognition

Deepanshi Bansal, Pooja Gupta

Abstract*: *Multiple Speaker Recognition is a powerful tool determining the number of speakers in a random speech along with determination of time span and segregation of voice signal of each speaker according to his/her MFCC and delta MFCC , chroma factors etc. which are different for different people because of variation in frequency of vocal chord which in turn effect the places of stress and syllables used.The methodology used in the research paper enable user to extract and isolate the individual voice streams at the receiver end. It can be used in various situations such as office meetings, the country parliament sessions to identify the speaker and his/her content to draw conclusions. The system may help the user to highlight a particular speaker's voice amongst other speakers. Even if the number of speakers is not known, the algorithm used in the paper will be able to determine the number of speakers based on the concept of clustering and then extracting common features of voice , it will be able to distinguish the speakers and separate their voice with each other thus giving the output as the duration of the discussion done by each speaker along with separated voice saved in wav format. If the voice sample of the speaker is stored in the speaker recognition model then, the model will also be able to show the name of the speaker else the representation of different speakers will be in the form speaker 1 , speaker 2 and so on. The methodology discussed in the paper could be used in present day interview process during group discussions and in admission process for higher studies and this could ease the work load on the recruiters giving them a idea about contribution of different speakers in the discussion.

Index Terms*: *clustering, diarization, voice activity detection

I. INTRODUCTION

Group conversations have been common these days and a lot of important decisions take place through group discussions. Speaker Recognition methods as known which could identify which speaker is speaking but when multiple speaker are speaking with each other, it becomes difficult to identify the speaker. Issues like overlapping, time span, similarity in voice, similar age, same gender etc. could effect the results. Multiple Speaker Recognition could solve this problem without even knowing the number of speakers in the room or while in a conversation. This is all together an unsupervised learning model to identify the different speakers and the time span along with the speaker details. Having an efficient information of contribution of different speakers in a discussion could be fruitful. Also it help segregate the speech of different speakers. This could help listen to particular speaker again. Also it could help to produce an automatic machine to determine who has performed the best in any group discussion. This technology provide a platform for academic institute and companies to select valid candidates.

This paper talks about how a conversation can be broken down for different speakers in order know their contribution in the conversation.

II. RELATED RESEARCH

Segregation of voice is a kind of cocktail party problem discussed by Josh H. McDermott[5] in his research paper but it can be solved if voice is transmitted through different channels. He discusses about the problems faced during solving cocktail party problem like the distance between the microphone and the speaker may vary speaker to speaker. Also it discusses about the ICA (independent component analysis) that can produce effective results for multi- source. For single channel if there is a need to recognize multiple speaker, then the technique Fourier transform is needed.

Xavier Anguera[2] and his team says overlapping of voice is the main problem during speaker separation. It performed supervised learning algorithms for speaker separation and thus dependent on the amount of dataset present. It discusses about the process to be followed for speaker separation and the difficulties to be faced . S. Govinda Rao[3] discusses about how to determine the number of speakers in an audio because without knowing the number of speakers, it is difficult to distinguish the voice. Hence the writer implements silhouette method of determining number of clusters in k means algorithm. In his methodology he modified k means algorithm but it is efficient for large dataset. As this research paper is focused on unsupervised learning, it is better to use silhouette method with K means clustering. GMM model is studied using the research paper published by Douglas A Reynolds[6] and team .It gives knowledge about how to identify the speakers based on their features. Tomi Kinnunen[11] in his paper discusses about voice activity detection and using SVM to produce the output. It uses MFCC to get the most important features. R. Thiruvengatanadhan[9] implemented SVM and MFCC for speech recognition having 95% speech recognized rate. Students fromTufts University [10] researched about Gaussian Mixture Model (GMM) to supervised learning based on the Maximum Likelihood (ML) estimation using Expectation Maximization (EM). Stephan Herzog[7] describes the efficiency of using median filtering on audio signals. All the above research done by different authors are combined to work on multiple speaker recognition. None of the author has studied multiple speaker recognition through determining the number of speakers and applying clustering and voice activity detection for recognition.

Manuscript published on 28 February 2019.
***** Correspondence Author (s)
Deepanshi Bansal*, Department of CSE, Maharaja Agrasen Institute of Technology(MAIT), GGSIPU, New Delhi, India.
Pooja Gupta, Assistant Prof. Department of CSE, Maharaja Agrasen Institute of Technology (MAIT), GGSIPU, New Delhi, India.

III. PROPOSED METHODOLOGY

In this work multiple speaker recognition methodology has been discussed which enables to extract and isolate the individual voice stream from a group discussion recording. The process starts by taking mixed audio as input. The complete procedure consist of various process including feature extraction, voice activity detection, silhouette algorithm, clustering, vector quantization and median filtering. The detailed flow between the processes has been shown in figure 1.

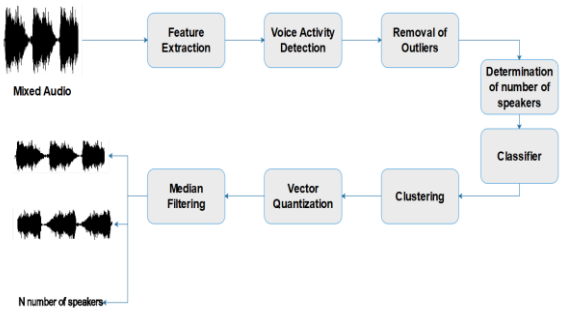

Fig.1 Process followed for multiple speaker segmentation

Mixed Audio is feeded as input. Multidimensional reduction is done as feature extraction using MFCC and chroma factors which is the first and most essential part of audio processing. The features are studied and normalized that is also called whitening used to remove channel effects. After the process of normalization, the audio is ready for us to perform further actions. Voice Action Detection is performed to check the presence or absence of a signal and extract particular speech rejecting noise. K-Means Silhouette method is used to determine the number of speakers in the audio clip. Then GMM classifier is used to get the normal distribution model and then clustering is performed using GMM model by putting the number of clusters as obtained by silhouette method. Then Linear Prediction Coefficient algorithm is used to create the segments of voice according to the clusters formed and finally at last median filtering is performed to get a clear start and end point of different clusters. Details of each process is discussed as below-

A. Feature Extraction

For processing any audio file; the signal need to be analyzed. Preprocessing needs to be done which includes loading data, pre-emphasis, framing, window, Fourier-transform, power spectrum, filter banks, MFCCs and mean normalization.
Hence MFCC(Mel Frequency Cepstral Coefficients) and Delta MFCC are used to produce vector of features of any audio signal. It includes various process like pre-emphasis, frame blocking, hamming window, Fourier transform etc. MFCC has proved that it can determine the major acoustic features like frequency, pitch, energy, entropy etc. which is responsible for variation in voice of different speakers. Feature extraction thus is the first step to perform any action on audio signal. Spectral features taken are zero crossing rate calculation, energy, energy entropy, spectral centroid spread, spectral entropy, spectral flux, and spectral roll off. With num_ceps=13 that usually is between 2-13, cepstral coefficients are retained. Other coefficients are discarded because they depict fast change in filter bank coefficients which does not effect the result of speech recognition.

Delta MFCC is used to have better features. Delta MFCC determines the change in mel frequency coefficients during change in time. Hence both MFCC features and Delta MFCC features together build up an efficient process for feature extraction. To obtain MFCC, analysis of spectrum is done for a small window obtained from speech and then the mel-scaled filterbank is applied. Formula for mel scale is

$$f_{mel} = 2595 \cdot log_{10}\left(1 + \frac{f_{linear}}{700}\right) \quad \text{-(1)}$$

Where,
f_{mel} - mel scale resulting frequency in mels
f_{linear} -normal frequency in Hz.
If (frequency is low)
mel filterbank is linear
Thus this contains more information and
if (frequency is high)
mel filterbank is logarithmic.
Filters in filter bank are distributed in mel scale evenly. Filter bank produces a feature vector that contains amount of energy captured in each frequency band.
Other than this , chroma features are also taken into consideration Some chroma names are 'A', 'A#', 'B', 'cen', 'cen#', 'D', 'D#', 'E', 'st_flux', 'st_flux#', 'G', 'G#' .

With this short term features vectors are extracted which in turn is used to get the mid term features. By the term 'Short-term features', it refers to formants, formant bandwidths, pitch and log energy, reflect local speech characteristics in a short time window. Mean of pitch, standard deviations of pitch, time envelopes of pitch and energy, reflect voice characteristics over a whole utterance refer to as long term features. Then come the normalization of features. Normalization is used to remove the channel effects in a speech signal. Mean of features is removed and normalization is done by variance. Through this, features come out to be in same space. The reason for performing normalization is to get the maximum volume and also the matching volume. There are various methods to normalize audio signal but the best ones are peak volume detection and RMS volume detection. Detecting peak volume considers how loud the waveform peaks are in the audio signal. RMS volume detection considers the overall loudness in the signal. It takes the average of peaks as there may be softer sections as well as larger peaks.

B. Voice Activity Detection(VAD)

VAD is known as voice activity detection or speech detection. Its main use is in speech recognition problem. It is based on spectral entropy. It processes the portion of audio where there is no speech and thus the output comes out to be the audio portion where there is speech. It discriminates the speech from unwanted noise and disturbance. Temporal energy variations, periodicity, and spectrum are few of the properties that are combined for building up the algorithm of voice activity detection.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Threshold is compared with the measured features that are extracted from the input signal. When the features measured exceeds the threshold, it is said to be a voice activity that is VAD=1. And where there is no speech activity, VAD = 0. Thus through this binary output we can make the decision whether that portion of frame is to be accepted or ignored.
SVM is used as binary classifier which acts as a hyperplane and models the decision boundary between the two classes. The kernel function used is linear. Implicit mapping into a high-dimensional feature space is the work that the kernel function does. Other kernel function could be the RDF(radial basis function). For classification, the two classes taken are short term features with short term energy greater than and less
than the mean value of the same.
The assumption taken is that the frame should be greater than input signal and the frame duration taken in this paper is 20ms.

C. Before determining the number of speakers, pairwise distance function is applied to the mid term normalized feature vector and then converting this distance vector to square form vector, the mean is calculated to check for outliers and focusing on the major section of input data.

D. Unsupervised Learning

Identification of pattern in data input as unlabeled data is what unsupervised learning is all about. It automatically splits the data into group without having trained on known outcomes. Clustering is majorly used for unsupervised learning. It can segregate the audio signal based on the different features extracted and form clusters of voice with similar features. K – means clustering is iterative clustering algorithm which is based on random initialization of centroid point and determining WCSS(within cluster sum of square). Less the WCSS , better are the clusters formed. As there are different clusters, input goes into the cluster based on the centroid distance. The centroids in each cluster define one or the other feature that distinguish that cluster from others and define that particular cluster.

E. Determining Number of Speakers

Next challenge is to determine the no of speakers .For the same , clustering techniques is required to apply on speech to determine various similar voice signals. Out of many algorithms to determine number of clusters to perform clustering , K means Silhouette algorithm has proved to be good for audio processing when used with K means clustering algorithm. The plot of silhouette displays the closeness of each point in one cluster with the data in neighboring clusters. The Silhouette coefficient has to be calculated and it ranges between -1 to 1. 0 value defines the overlapping clusters. 1 is considered the best value .It indicate larger distance between the neighboring clusters.

Consider ‘a’ as the mean distance of point with the other points in the same cluster and ‘b’ as the mean distance of point with the points present in neighboring clusters.
The formula for calculating silhouette coefficient ‘s’ is

$$s = \frac{b-a}{\max(a,b)} \qquad -(2)$$

Number of labels should be greater than or equal to 2 and less than or equal to 1 less than the number of samples.

Mean Silhouette score is calculated for the entire cluster as the mean of silhouette score for each data point is taken.

F. GMM Classifier

The role of classifier is to represent input data by its features while creating different categories for different types of objects to focus on input. GMM classifier is supervised as probability density function(pdf) is estimated Linear combination of multivariate Gaussian pdf is used as model for different classes. GMM uses maximum estimation algorithm with calculating maximum likelihood. Though GMM is a supervised learning algorithm, it is commonly used in unsupervised learning as it is able to produce cluster and data patterns that shows similar behavior together. Statistical models like GMM are able to provide complete description of the real problem unlike KNN because it uses probability distribution density and probabilities are manipulated to model entire hypothesis space and data distribution. GMM model has good computational properties which make it better than deterministic models K- nearest neighbor (KNN) , perceptron etc. Many different types of noise in the physical system could be approximated with GMM according to the central limit theorem. It is important to work on hidden variables as well along with maximum likelihood. It may or may not indicate indicator variables. These types of variables whose values are never observed can be analyzed by the estimation maximization algorithm if the probability distribution of variables is there with us. Estimation of parameters of gaussians and their weights can also be done with GMM.

G. Vector Quantization

Vector Quantization is the most preferred method for segregation of audio based on different speakers and Linear Prediction Coefficient algorithm is used.
It is been observed that vector quantization method has comparatively much less error rate than the gaussian mixture model in [1] . Here the vector quantization and linear prediction coefficient algorithm is used. Frames of test segment is represented as feature vector as
$S = \{s_1, \ldots, s_t\}$, and the reference features vectors of speaker i as $R_i = \{r, \ldots, R_k\}$.
The test is classified as the speaker that minimizes,

$$D_Q(S, R_i) = \frac{1}{T}\sum_{t=1}^{T} \min_{1 \le k \le K} d(s_t, r_k) \qquad -(3)$$

for some distance measure, d.
Reduction of feature factors is done through one of the clustering method with each reference set R. This creates a summary for each cluster. In this paper, k means clustering is done and hence , in k means, r_k is the cluster centroid.
As the audio signal contains random speech of multiple people, it is difficult to identify when one particular speaker will start and end. Due to this reason, it is preferred to classify each test frame individually rather than averaging the segment of speech entirely. The frame is classified as the speaker using equation 4,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$D_Q(s_j, R_i) = \frac{1}{T} \sum_{t=j-\frac{F_s}{2}}^{j+F_s} \min_{1 \le k \le K} d(s_t, r_k) \quad \text{- (4)}$$

where F_s is the number of frames per second, r_k is the cluster centroid, and S_t is feature vector of segment.

H. Median Filtering

Further after analyzing each frame, next step is to remove any noise that is present or denoising of data is required.
Median filtering is used for this purpose. It is a type of feature used for removal of noise or denoising.It is applied to final classified vector to have good results. There are various applications for median filters like audio and image processing. In audio processing, it is used as sound analysis, voice separation and audio noise reduction
Median filters can be used with doubly linked list for operations to be handled effectively. In this paper median filter is used for getting a better version of separated voice and determining accurate start and end points for each speaker according to the clusters formed with the help of vector quantisation. The input to the median filter is the n dimensional input array of size as mid term features having the class to which the features belong to. Kernel size taken is 11 .It depicts the size of the median filter window in each dimension. The output with the median filter process is the same size array having the different classes in which the features are segregated into. Entry by entry , the signal is run and for each entry, replacement of entry with median of neighbouring entries takes place. Window is the term used for pattern of neighbours. The window keep on sliding through the different entries so as to cover the complete signal. Because there will be more than one possible median if the entries are even, thus the kernel size should be odd so as to ease the process of finding the median. When all the entries are sorted , the median is the middle value if entries are odd.

I. Determining if it is a known speaker or unknown

Determining if speaker is known or unknown is the last process done after recognizing the starting and ending point for each speaker. Figure 2 explains the flow of steps that are followed.

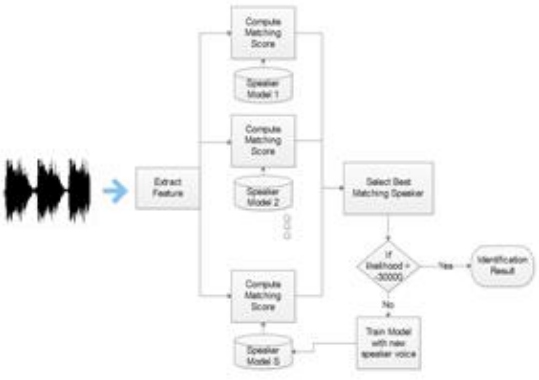

Fig.2 Determination of known speaker

Features of audio signal is extracted and GMM(gaussian mixture model) is created for each cluster points and compared with the trained dataset of known speakers whose GMM model is already created with us with prior training of model. GMM model uses weights and build up a normal distribution. The mixture of distribution is made from the sum of mixture model densities. Equation 5 is the formula for log likelihood that is used for comparing the existing GMM model with the GMM model formed from each cluster. If the likelihood is more than a particular preferred log likelihood that is set through observation then the speaker name is displayed according to the matched data and if the condition is false then that voice sample is used to train the model and create its GMM model to save it for future comparison.

$$lnp(\mathbf{x}|\mathbf{w}, \mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} ln\left\{ \sum_{i=1}^{K} w_i N|(\mathbf{x}_n|\mu_i, \mathbf{\Sigma}_i) \right\}. \quad \text{-(5)}$$

The mixture distribution is a sum of K Gaussian densities. Each mixture has its own mean vector μ_i, covariance matrix $\sum i$ and mixing weights w_i. The weights satisfy the constraint
$\sum^K{}_{i=1} w_i = 1$ and $\lambda = (w_i, \mu_i, \sum_i)$ where i = 1,... , K

IV. ANALYSIS AND RESULT

In this work segmentation of audio file into time constraints is done according to speakers who speak at that particular interval of time. For this detailed methodology used has been discussed in section III .To implement or test the same, real time voice discussion data has been searched on various free available sites. But due to unavailability of Indian voice data, in this work data has been generated by recording real time discussions. The data set such produced consist of 120 audio files. The audio file format that is supported is wav format and thus wav files were taken. Python language was used to implement the methodology discussed in this paper Accuracy of clustering could be done using commonly used indices. Because silhouette coefficient is used to determine the number of speakers, dunn index in used to check the accuracy of the model.

$$\text{Dunn Index(D)} = \frac{\text{minimum separation}}{\text{Maximum diameter}} \quad (6)$$

During the testing, 78 audio were predicted correctly and rest had issues due to overlapping and in few due to wrong prediction of number of clusters. The mechanism thus discussed in this paper is able to produce different audio files separating the different speaker by determining the number of clusters using silhouette algorithm. Moreover it helps to get the start and end point of each speaker to segregate the audio efficiently with an accuracy of 65%. The internal metrics validation technique is used for checking the accuracy. Compactness, connectedness and separation of cluster partitions are the common factors on which Internal validation measures vary. The mechanism in this paper is able to produce a working application for multiple speaker recognition that can be used during group discussions or in recordings and lectures.

V. CONCLUSION

Multiple Speaker Recognition through a single channel without knowing the number of speakers is implemented through unsupervised learning and later used dataset to determine the name of the speaker if his/her voice sample is taken beforehand.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The result was good and the model was able to distinguish large number of speakers in different voice samples. Building of dataset was a lengthy process but finally the result was fruitful and dataset of one twenty voice samples with Indian voice was collected and used in the implementation of this paper. Speech Recognition along with determination of number of speakers in a random conversation is implemented successfully and can be used in its various applications. Multiple Speaker Recognition has many applications such as a front-end for speaker and speech recognition, as a metadata extraction tool to aid navigation in broadcast TV, lecture recordings, meetings, and video conferences and even for applications such as media similarity estimation for copyright detection. Other than this , the implementation of this paper could be used in meeting , political sessions, group discussions and interviews.

REFERENCES

1. *Speaker Recognition for Multi-Source Single- Channel Recordings* by Jose Krause Perin, Maria Frank, and Neil Gallagher,2015
2. *Speaker Diarization: A Review of Recent Research* Xavier Anguera, Member, IEEE, Simon Bozonnet, Student Member, IEEE, Nicholas Evans, Member, IEEE, Corinne Fredouille, Gerald Friedland, Member, IEEE, Oriol Vinyals , 2011
3. *Performance Validation of the Modified KMeans Clustering Algorithm Clusters Data* by S. Govinda Rao Associate Professor and Dr.A. Govardhan
4. *The Clustering Validity with Silhouette and Sum of Squared Errors* published in Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 by Tippaya Thinsungnoena*, Nuntawut Kaoungkub , Pongsakorn Durongdumronchaib , Kittisak Kerdprasopb , Nittaya Kerdprasop
5. *The cocktail party problem* by Josh H. McDermott. http://mcdermottlab.mit.edu/papers/McDermott_2010_cocktail_party_problem.pdf 10.1016,2009
6. Douglas A Reynolds, Thomas F. Quatieri, and Robert B. Dunn. *Speaker verification using adapted Gaussian mixture models*. Digital signal processing, 10:19–41, 2000
7. E*fficient DSP implementation of median filtering for real-time audio Noise reduction.* Stephan Herzog. Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland, September 2-5, 2013.
8. Christopher M. Bishop. *Pattern Recognition and Machine learning*. Springer Science, 2006.
9. R. Thiruvengatanadhan, *Speech Recognition using SVM*, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 ,Sep 2018
10. *Practice on Classification using Gaussian Mixture Model* Dept. Computer Science, Tufts University, Medford, USA , 2010
11. *Voice Activity Detection Using MFCC Features and Support Vector Machine* Tomi Kinnunen1 , Evgenia Chernenko2 , Marko Tuononen2 , Pasi Fränti2 , Haizhou Li1 1 Speech and Dialogue Processing Lab, Institute for Infocomm Research (I2 R), Singapore 2 Speech and Image Processing Unit, Department of Computer Science, University of Joensuu, Finland ,2007

AUTHORS PROFILE

Deepanshi Bansal is a student studying BTech CSE from Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India graduating in 2019. Deepanshi Bansal has been Microsoft Student Partner for 2017-18 and 2018-19. Also she has been treasurer of society International Organization of Software Developers in 2017-18. Now Deepanshi is Vice President of the same society for 2018-19. She has won Deloitte Innovation Award In Smart India Hackathon 2017.

Pooja Gupta is Assistant Professor at Maharaja Agrasen Institute of Technology(MAIT) , GGSIPU, New Delhi India. She has completed Ph.D. Computer Science & Engineering from JIIT (April-2014) and M.Tech (CE) from YMCA, Faridabad (MDU, Rohtak) (2006).Her area of interest are Information Retrieval and Data Mining, Machine Learning, Computer Networks. She has written various research papers , few of them are “An Improved approach to rank web documents”, Journal of Information Processing Systems (JIPS), Korea, Vol. 9, No.2, pp-217-236 (indexed at DBLP and Scopus.) , “A Novel Technique for Back-Link Extraction and Relevance Evaluation”, International Journal of Computer Science & Information Technologies, Vol 3, No. 3, June 2011 pp-227-238. [indexed in docstoc, pubzone, DOAJ, Inspec, EBSCOS etc.] , “A Novel Framework for Context Based Distributed Focused Crawler (CBDFC)”, Int. J. Computer and Communication Technology, Vol. 1, No. 1, 2009 pp-14-26 (Copyright Inderscience Enterprise Ltd.) etc.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)