International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Cluster-Based Speaker Diarization System Combined with Dimensionality Reduction Techniques

D. Indu*[1], Y. Srinivas[2]

Submitted: 07/12/2023 **Revised:** 18/01/2024 **Accepted:** 28/01/2024

Abstract: In this article, we introduced an unsupervised speaker diarization system for speaker detection in noisy environments, we introduced a statistical mixture-based model to model the input segment and cluster features obtained using MFCC for effective speaker identification of this segment. The concepts of KL- divergence is considered to effectively identify a speaker based on the a maximum Likelihood estimate of the speaker.

Keywords: *Speaker diarization, statistical model, segmentation clustering KL divergence.*

1. Introduction

A automatic speaker diarization is a process of knowing who said what when and why during a recorded speech that includes the understanding of automatic speech from the written and conference environment usually during diarization speech input is recorded in the form of audio sequences and these signals chopped into short signal segments and we try to embedded the segments of the speech a reason that characterizes the speaker characteristics these embedded segments are clustered and the process of general architecture for speech recognition process is given as follows

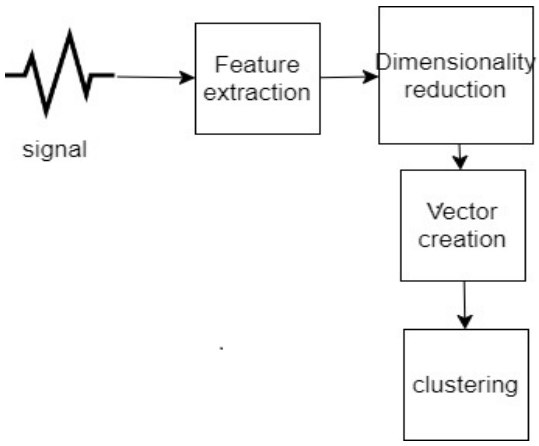


Fig. 1. General Architecture For Speech Recognition Process

In order to identify the similarity among these features generally the likelihood of audio streams are to be mapped. In general. this process in most of the cases is carried out using minimum likelihood function. In our case we have considered concepts of KL divergence for this process recently many articles have been presented in the architecture where the spectral clustering is taken into consideration also the concept of bayesian information criteria is considered for estimating the relevant speaker

In this process that is the BIC and spectral clustering the efficiency of the cluster totally lies on the values of k that is chosen to cluster the audio signals into streams in general this selection of K is studious process if the value is not choose and properly it may lead to over clustering or under clustering in order to overcome this challenge hierarchical clustering is generally considered

In hierarchical clustering, identifying the peak threshold on which the cluster is carried out is a matter of concern to overcome this disadvantage. Our article considered the Gaussian mixture model for clustering the data. The main advantage beyond this consideration is that Gaussian clustering helps cluster data into appropriate segments despite noise in general. The output signal will be associated with narrative noise whenever a speed signal is recorded. Every audio frame is connected as the signal converts to pulse additive noise.

$$S=S1+A \quad (1)$$

S=Output Speech

S1=Generated Audio Signal

A=Embedded Noise

In order to model the speech and cluster the relative segments' acoustic features based on MFCC, the main advantage of considering MFCC features is that it can be considered the low speech spectrum and high-speech spectrum. Then, we can model the data efficiently in order to cluster the data. The speed sample is divided into segments and uses the concepts of dimensionality

[1] *Research Scholar*
[2] *Professor*
[1,2]*Department of Computer Science and Engineering*
GITAM School of Technology GITAM (Deemed to be University)
Visakhapatnam, Andhra Pradesh, India
ORCID ID : 0009-0005-4525-544X
** Corresponding Author Email: idasari@gitam.in*

production using the eigenvector generated, and this vector is considered for the identification of relative features. The article's remaining content is explained as follows: The article's second section provides a succinct overview of the literature produced in this field of study. In Section 3, the data set considered is presented in Section 4 of the article deals with acoustic feature extracting based on FCC in Section 5, the process of clustering based on GMM is presented in thecorresponding Section 6, and the concept of dimensionality reduction is highlighted in Section 7, of the article together with methodology. In the concluding Section 8, the article is summarized with the conclusion.

2. Related Work

Much research has been undertaken in this area of research, with the latest advancements in the area of speech processing. The major contributions in this area of research are highlighted below.

Principal Methods Top-down and bottom-up speaker diarization systems comprise the bulk of today's state-of-the-art systems. In contrast to the bottom-up strategy, which starts with many clusters (often more than expected speakers), the top-down approach starts with very few clusters, generally just one. The goal is to iteratively converge towards the ideal number of clusters in both situations. An under-clustering of the system occurs when the final number exceeds the optimal. It is considered to be over cluster if it is lower. The hidden Markov models (HMMs), on which each state is a Gaussian mixture model (GMM) corresponding to a speaker, are typically the foundation of both top-down and bottom-up techniques.

Speaker turns coincide with state transitions. This section provides a brief overview of the conventional top-down and bottom-up methodologies, two recently suggested , one is on a nonparametric Bayesian approach and the other on information theory. These novel techniques have demonstrated great promise on NIST RT assessment datasets, despite the fact that they have not yet been documented in the context of official NIST RT evaluations. For this reason, they are included here. Furthermore, a few additional papers [5]–[7] provide sequential single-pass clustering and segmentation techniques, even though they frequently perform less efficiently than state-of-the-art techniques.

. This more straightforward method typically produces comparable results [8]. In every instance, the audio stream is first excessively divided into numerous fragments, surpassing the projected upper limit of speakers. After that, the bottom-up method repeatedly chooses clusters that are closely matched to merge, thereby lowering the total for every repetition, by one cluster.Typically, a GMM is used to model clusters. When two clusters mergethe data that was previously assigned to each of the initial clusters is used to train a new GMM..

The nearest clusters are determined using common distance measures, as those mentioned in Section III-c. After every cluster merging, for example, a reassignment of frames to clusters is commonly carried out by Viterbi realignment. This process is repeated iteratively until a stopping threshold is met, after which there should be only one cluster for each detected speaker. The Bayesian Information Criterion (BIC) [9], Kullback-Leibler (KL)-based metrics [10], the generalized likelihood ratio (GLR) [11], or the recently suggested measure [12] are examples of threshold techniques that could be used as halting criterion. Bottom-up systems have continuously fared well when they were submitted for the NIST RT evaluations [9], [13]..

2) Top-Down Approach: In contrast to the earlier technique, this one begins by modeling the whole audio stream using a single speaker model. It then progressively adds more models to the stream until it is believed that every speaker is present.. All of the accessible speech segments are used to train a single GMM model, and each segment is labeled as unlabeled. Viterbi realignment and adaptation are interleaved when new speaker models are iteratively introduced to the model one at a time, with a selection technique used to find appropriate training data from the unlabeled segments. Any segment linked to one of these novel models has a label attached to it.

The process can either continue until no more relevant unlabeled segments are available for training new speaker models, or it can be terminated using stopping criteria similar to those used in bottom-up systems. Compared to bottom-up approaches, top-down approaches are far less prevalent.. Here are a few instances: [14]–[16]. Top-down methods have regularly outperformed other bottom-up entrants in the field, even though the best bottom-up systems usually outperform them. Additionally very computationally efficient are top-down methods, which can be enhanced by cluster purification. [17].

3) Alternative Methods: A newer alternative method that is based on an information-theoretic framework and is also bottom-up in nature is motivated by rate-distortion theory [18]. It is entirely nonparametric, and despite substantial computational savings, its outcomes are on par with those of cutting-edge parametric systems Clustering is based on mutual information, which measures the mutual dependence of two variables [19]. There is just one global GMM tuned for the whole audio stream, and mutual information is computed in a new space of relevance variables specified by the GMM components. The method's objective is to minimize the loss of mutual information across succeeding groupings while retaining as much information as feasible from the original dataset.

The sequential information bottleneck (sIB) [19] and the agglomerative information bottleneck (aIB) [18] are two appropriate techniques that have been documented. Although the new system does not outperform parametric methods, it does provide results that are on par with the most advanced GMM systems while saving a significant amount of work.

However, speaker diarization has made use of Bayesian machine learning, which gained popularity by the end of the 1990s. A cornerstone of Bayesian inference is to concentrate on the parameters of an Anguera system rather than the system's real parameters (i.e., point estimations). Hyperparameters in Linked Distribution in Speaker Diarization: An Overview of Current Studies 359, written by Miro et al.. As a result, the system can automatically adjust to observations (e.g., the model's complexity depends on the data) and the diarization problem can avoid any hasty decisions.

However, intractable integrals are frequently needed for the construction of posterior distributions; hence, the statistics field has created approximation inference techniques. The initial use of Monte Carlo Markov chains (MCMCs) was made possible by the introduction of Bayesian techniques [20] by offering a methodical way to compute distributions through sampling. Nevertheless, when the volume of data is high, sampling techniques are typically prohibitively expensive and sluggish. They also need to be repeated since chains may become trapped and not converge after a reasonable number of rounds.

A further alternative method that attempts to produce a deterministic approximation of the distributions is called variational bayes, and it has gained popularity since 1993 [21], [22]. By approximating the intractable distribution with a tractable approximation produced by reducing the K L divergence between them, it allows an inference issue to be transformed into an optimization problem. In [23], the merging criterion, a change detection procedure, and a GMM speaker model are optimized by the application of a variational Bayes-EM algorithm. Variational Bayes is effectively coupled in [24] with eigen voice modeling—which is explained in [25]—for speaker diarization of phone calls. These systems are different from nonparametric Bayesian systems in that they still take into account traditional Viterbi decoding during the classification phase.

Lastly, speaker diarization in meetings has been effectively accomplished with the newly suggested speaker binary keys [26, 27]. This technology performs similarly to state-of-the-art systems while also conserving a significant amount of computational power (operating at 0.1 times the real-time rate). Small binary vectors known as speaker binary keys are derived from the audio data using a paradigm akin to the universal background model (UBM). All processing actions take place in the binary domain after calculation. Additional speaker diarization initiatives that prioritize speed include [28], [29], which use different processing techniques applied to a standard bottom-up approach ([28]) or parallelizing most of the processing on a GPU unit ([29]) to accomplish processing faster than real-time. managing incredibly big datasets or using diarization as a preliminary technique before using further speech algorithms.

Many gaps existed in the field of clustering and segmenting, despite important research in this domain. This article presents an elementary solution to this problem.

3. Dataset Considered

3.1. AVA:

AVA stands for "AVenue for ASR," and it is a dataset of speech data from meetings. It is an important resource for research in automated speech recognition (ASR) and related topics since it contains audio recordings and transcriptions from a variety of meetings. This dataset is used to evaluate and train models and algorithms for tasks like meeting transcription and speaker diarization.speech recognition, and natural language interpretation.

3.2. CHIME5:

The CHiME-5 dataset, which consists of over 50 hours of conversational audio recordings, was sourced from twenty real dinner parties in real households. The recordings were made with a variety of 4-channel microphone arrays, and they have been thoroughly transcribed. The features of the dataset include: real dialogue, i.e., talkers chatting casually and naturally; simultaneous recordings from multiple arrays of microphones; a range of room acoustics from twenty different houses, each with two or three distinct recording areas; and real domestic noise backgrounds, i.e., air conditioning, movement, kitchen appliances, etc.There is continuous audio available, complete transcriptions of all spoken words, ground truth speaker labels, and start/end time annotations for segmentation.

3.3. DIHARD:

Often referred to as DIHARD11, the The goal of the DIHARD (Directional Hearing in Noisy Environments) speaker diarization challenges is to increase the diarization systems' resilience to changes in the conversational domain, noise levels, and recording equipment. Eleven different domains and two speech activity conditions—diarization from a reference speech activity vs diarization from scratch—were used to assess speaker diarization. The domains cover a variety of recording scenarios and interaction kinds, such as web videos, clinical interviews, read-aloud audiobooks, meeting speech, and, for the first time, conversational telephone speech.

3.4. RADIO TALK:

The speech recognition transcripts in the RadioTalk corpus

were taken from talk radio shows that aired in the US between October 2018 and March 2019. Researchers in conversational analysis, natural language processing, and the social sciences are the target audience for this corpus. The corpus includes metadata about the speech, including speaker turn borders, gender, location, and radio program information, together with over 2.8 billion words of mechanically transcribed speech from 284,000 hours of radio.

4. Feature Extractions

In this article, we have considered MFCC features; MFCC features extract speech signals and, transform them into audio signals and capture the important frequency and temporal information during the MFCC feature extraction process. The audio signal passes through several steps, including pre-emphasis,, window, DFT, MEL filter BANK, log TFT, and inverse DFT feature transformation. During the pre-emphasis, each signal's energy will be boasted to high frequency. This process helps to understand the speech signal in a better way. The speech samples are sliced into a uniform during the windowing process, and each frame is subjected to noise elimination. During the DFT phase,

The speech information is transferred into the frequency domain to better enhance hearing perception to the individual mel filter bank. However, this Mel filter Bank generates a power spectrum, and since a human cannot perceive this energy spectrum, the log is applied of transformation is applied. The Mel spectrum coefficient transformed speech the spectrum into an audible range. There are 39 mFCc features and, 12 cepstral coefficients, and one energy coefficient. Using this feature parameter, we can model both the low-level frequency modulation as well and high level hence, are considered.

5. Gaussian Mixture Model

WE generally identify the speaker and the speed signal should be sampled into the appropriate signal. Many segmentation algorithms are presented in the literature base hierarchy clustering. Each of audio signal attributed from the speech sample will follow a distribution, and its range is general - infinity to + infinity. So, as mentioned in section 1 of the article, k means algorithm Hierarchical clustering suffers from imitations. Also, they cannot segment which is of infinite size; therefore, to model the data more efficiently, the equation for Gaussian distribution is given by

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

$\pi \approx 3.14159\ldots$

σ = Standard Deviation

$e \approx 2.71828\ldots$

μ = Mean

Once the speech signal is modeled using Gaussian distribution, we formulate a vector of probability density function against the peach sample. These speech samples are modeled using dimensionality reduction techniques to reduce the dimension in this article. PCA was considered for this analysis, and the modeled data is compared to that of the existing speakers using the concept of k l divergence. The KL divergence is used to identify the relativeness between the input speaker and the speaker in the database.

$$D_{KL}(p||q) = \int_x p(x) log \frac{p(x)}{q(x)} dx \quad (3)$$

P and Q are the speech segments under consideration. The limits are 0 to 1 if the speakers are mapped. The value is approximately 0 else to 1.

6. Methodology and Results

This article presents a methodology for speaker diarization using Gaussian clustering with MFCC features and K L divergence. The input speech is segmented into clusters, each using GMM based on MFCC coefficients. These clusters are mapped against relevance based on K L divergence for estimating the speaker. The results derived are presented below Speech Accent Data Set, Audio Files :2500, Speakers: Over 100 Countries.

Table 1. Speech Accent Data Set By Contury And Its Languages

Country	***Actual Languages and Regions***
US	'UnitedStates English'
India	'India and South Asia (India, Pakistan, Sri Lanka)'
New Zealand	'NewZealand English'
Singapore	'Singaporean English'
Philippines	'Filipino'
Hong Kong	'HongKong English'
Australia	Australian English'
Malaysia	'Malaysian English'
England	'England English'
Canada	'Canadian English'

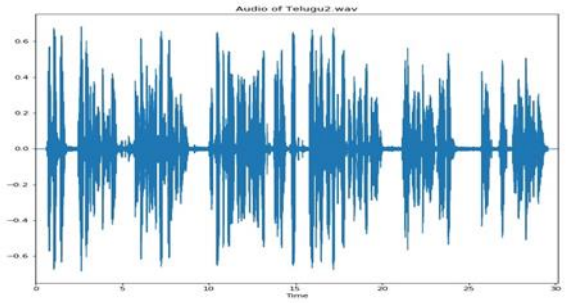


Fig. 2. Raw Audio Signal

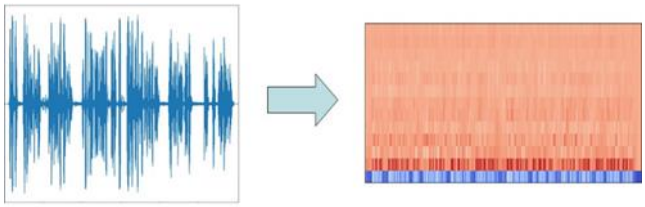


Fig. 3. Raw Audio File To MFCC

Periodogram power estimate formula

$$P_i(k) = \frac{1}{N} |S_i(k)|^2 \qquad (4)$$

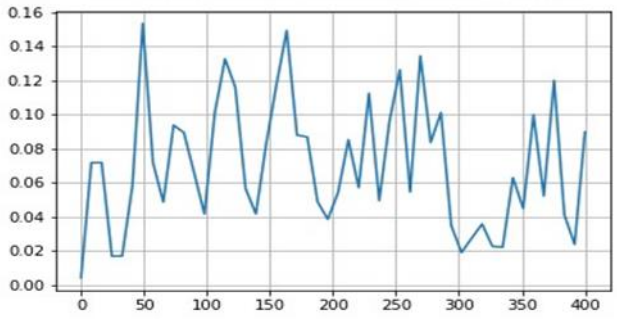


Fig. 4. Fourier Transform of Raw Audio

MEL Scale Conversion:

$$m = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \qquad (5)$$

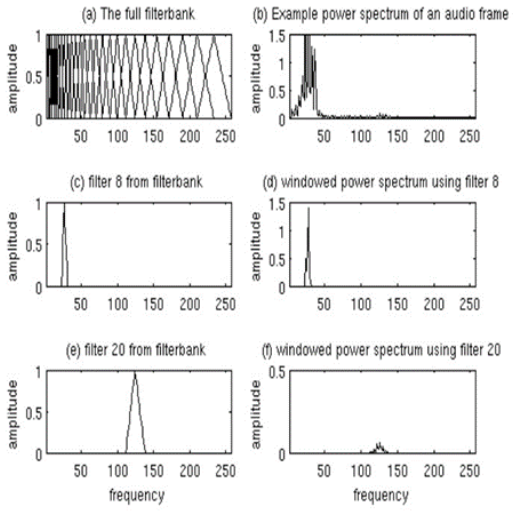


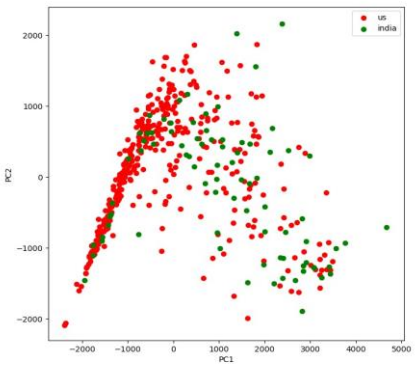
Fig. 5. Filter Bank

Log Of All Three Filters: So We Have 26 Co Efficients.

Dct Of Log Filterbank Energies:

$$\boldsymbol{X[n] = c\,(n) \sum_{m=0}^{N-1} x[m] cos\left(\frac{(2m+1)n\pi}{2N}\right) \quad (n = 0, \dots, N-1)} \qquad \boldsymbol{(5)}$$

PCA: PCA ON THREE CLASS (US,INDIA,UK):



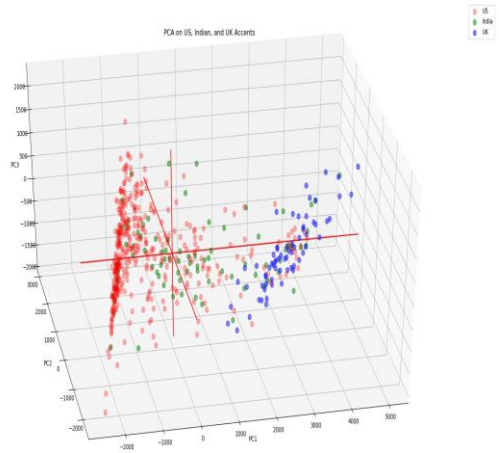


Fig. 6. PCA on three Class

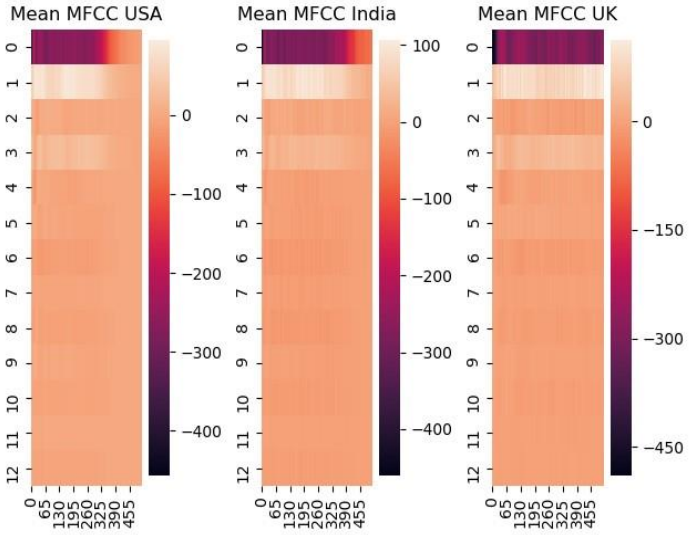


Fig. 7. Difference In Accents

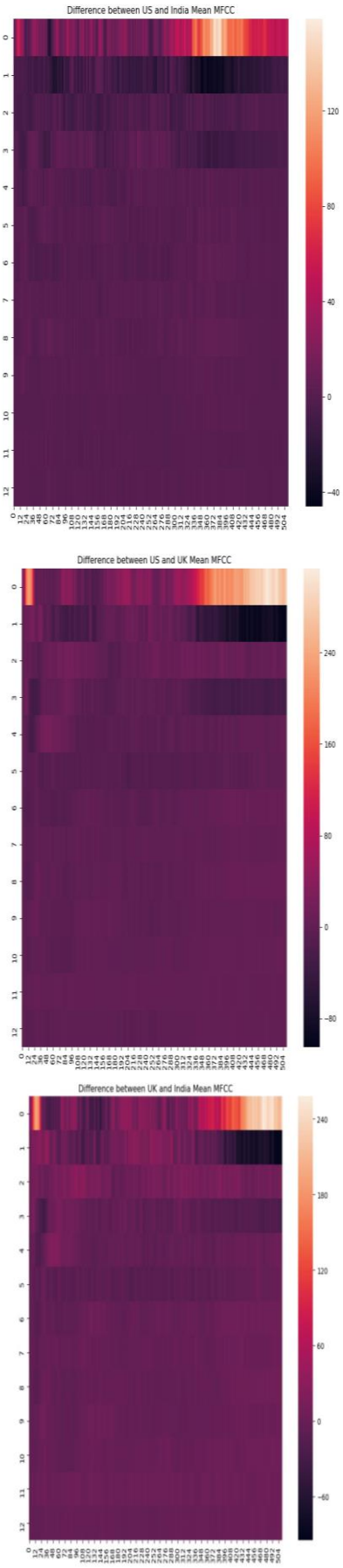


Fig. 8. Difference In Accents

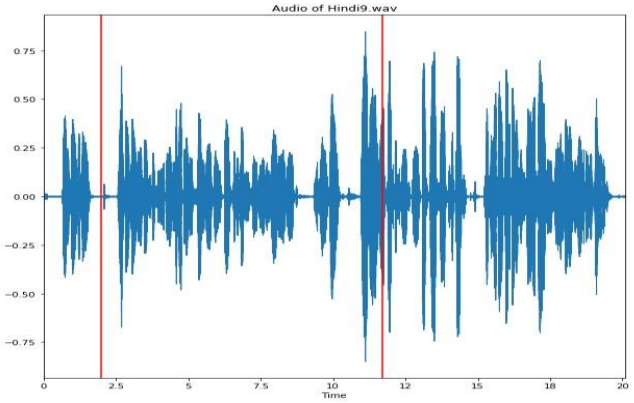


Fig. 9 Audio of Hindi

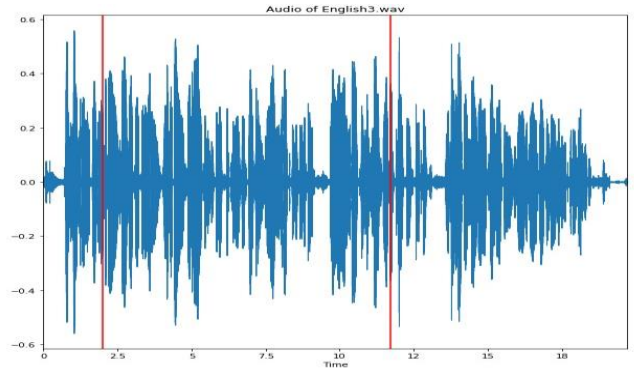


Fig. 9. Audio of English

TWO ACCENT CLASSIFICATION RESULTS:

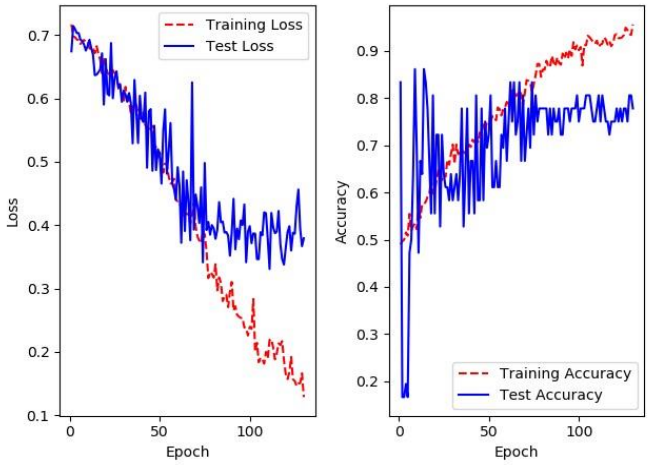


Fig. 10. Test Accuracy On Hold Data:79%

Table2 Performance Evaluation

Model	***MFCC Feature***	***Accuracy***
GMM	26	79%
LSTM	26	72%

When compared to the LSTM model, our proposed method's GMM findings provide the best accuracy.

7. Conclusion

The proposed research work includes a thorough evaluation of the methodology used for various applications in speech recognition, such as detecting the features, using MFCC, and comparing the quality of proposed results by considering the maximum likelihood using KL Divergence Method. In addition, the method of identifying the speaker was analyzed for the speaker diarization using the processes of segmentation and clustering. Derived results showed better recognition accuracy.

References

[1] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using autoassociative neural networks," Eng. Applicat. Artif. Intell., vol. 22, no. 4-5, pp. 667–675, 2009.

[2] X. Anguera, C. Wooters, and J. Hernando, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in Proc. ICSLP, Pittsburgh, PA, Sep. 2006.

[3] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8–11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

[4] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, "Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast," in Proc. ICASSP, May 2006, vol. 5, pp. 521–524.

[5] M. Kotti, E. Benetos, and C. Kotropoulos, "Computationally efficient and robust bic-based speaker segmentation," IEEE TASLP, vol. 16(5), 2008.

[6] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage Speaker Diarization for Conference and Lecture Meetings," in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

[7] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using autoassociative neural networks," Engineering Applications of Artificial Intelligence, vol. 22(4-5), 2009.

[8] X. Anguera, C. Wooters, and J. Hernando, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in Proc. ICSLP, Pittsburgh, USA, September 2006.

[9] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

[10] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, "Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast," in Proc. ICASSP, vol. 5, May 2006.

[11] W. Tsai, S. Cheng, and H. Wang, "Speaker clustering of speech utterances using a voice characteristic reference space," in Proc. ICSLP, 2004.

[12] T. H. Nguyen, E. S. Chng, and H. Li, "T-test distance and clustering criterion for speaker diarization," in Proc. Interspeech, Brisbane, Australia, 2008.

[13] T. Nguyen et al., "The IIR-NTU Speaker Diarization Systems for RT 2009," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[14] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in Proc. Odyssey Speaker and Language Recognition Workshop, Chania, Creete, June 2001, pp. 175–180.

[15] C. Fredouille and N. Evans, "The LIA RT'07 speaker diarization system," in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 520–532.

[16] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[17] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 Speaker Diarization System: enhancements in speaker modelling and cluster purification," in Proc. ICASSP, Dallas, Texas, USA, March 14-19 2010.

[18] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in Proc. ASRU, Dec. 2007, pp. 250–255.

[19] D. Vijayasenan, F. Valente and H. Bourlard, "An information theoretic approach to speaker diarization

of meeting data," IEEE TASLP, vol. 17, pp. 1382–1393, September 2009.

[20] S. McEachern, "Estimating normal means with a conjugate style dirichlet process prior," in Communications in Statistics: Simulation and Computation, vol. 23, 1994, pp. 727–741.

[21] G. E. Hinton and D. van Camp, "Keeping the neural networks simple by minimizing the description length of the weights," in Proceedings of the sixth annual conference on Computational learning theory, ser. COLT '93. New York, NY, USA: ACM, 1993, pp. 5–13. [Online]. Available: http://doi.acm.org/10.1145/168304.168306

[22] M. J. Wainwright and M. I. Jordan, "Variational inference in graphical models: The view from the marginal polytope," in Forty-first Annual Allerton Conference on Communication, Control, and Computing, Urbana-Champaign, IL, 2003.

[23] F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, Thesis, 09 2005.

[24] D. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in Proc. Interspeech. ISCA, 2009.

[25] P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," CRIM, Montreal, Technical Report, 2008.

[26] X. Anguera and J.-F. Bonastre, "A novel speaker binary key derived from anchor models," in Proc. Interspeech, 2010.

[27] X. Anguera and J. -F. Bonastre, "Fast speaker diarization based on binary keys," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp. 4428-4431.

[28] Y. Huang, O. Vinyals, G. Friedland, C. Muller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Kyoto, Japan, December 2007, pp. 693–698.