[1]Indu D.

[2]Y. Srinivas

# A Methodology for Speaker Diazaration System Based on LSTM and MFCC Coefficients

***Abstract:*** - Research on Speaker Identification is always difficult. A speaker may be automatically identified using by comparing their voice sample with their previously recorded voice, the machine learning strategy has grown in favor in recent years. Convolutional neural networks (CNN) , deep neural networks (DNN) are some of the machine learning techniques that has employed recently. The article will discuss a successful speaker verification system based on the d-vector to construct a new approach based on speaker diarization. In particular, in this article, we use the concept of LSTM to cluster the speech segments using MFCC coefficients and identify the speakers in the diarization system. The proposed system will be evaluated using benchmark performance metrics, and a comparative study will be made with other models. The need to consider the LSTM neural network using acoustic data and linguistic dialect is considered. LSTM networks could produce reliable speaker segmentation outputs.

***Keywords:*** Speaker Diarization, Deep Learning, Speech Sample, Segmentation, MFCC Coefficients, LSTM.

## I. INTRODUCTION

The method of identifying a speaker and determining their identity from voice samples is known as speaker recognition. Specifically, chunking the speech signal according to a speaker's identity is known as speaker diarization. This chunking or clustering technique is a crucial subtask in some specific speaker recognition scenarios. Speaker tracking is also a part of the recognition process. Splitting the verbal utterance into samples is important to monitor the speaker, and boundaries are typically considered. Something that divides two speech signals is called a border.

A two-phase method is used in speaker diarization; in the first phase, the obtained speech samples are grouped into segments, and in the second phase, preprocessing is done. The general first segmentation technique detects variations in voice samples within the input speech signal. We consider the speech samples fake and discard them if there are any relative changes. Most of the time, parametric models are considered in the literature. In the statistical modeling process, each speech sample is considered, and the acoustic or phase spectrum is identified generally to identify the speech signal and to interact with the signals for processing. These models are mostly statistical, including GMM and HMM [1][2].

The procedures commonly carried out are based on MFCC [3], employing digital amplitude and frequencies that are thought to model the speech sample when we extract speech signals. A vector with 13 columns will be produced. Every column can be considered a feature vector; prosodic or auditory features correspond to this column. The statistical modeling techniques take these features as input. The resultant output will take the shape of a peaks-containing signal or histogram. These peaks are kurtosis and skewness in statistics [4], [5]. Kurtosis is the term used to describe peak readings.

The majority of the voice samples may exhibit positive or negative skewness and left or right truncation; in other words, the data may not be uniformly distributed, with some signals exhibiting high and low peaks. Kurtosis is the term for the measurements of the peaks. Generally speaking, the signal can be lepto, meso, or platykurtic. This Methodology defines a normal distribution (beta = 3) if the data is uniformly skewed. It also generates hmm algorithms and identifies the corresponding class labels.

The homogenous signal that has been categorized and grouped must be identified to identify a specific speaker, i.e., during the speaker diarization procedure that asks who spoke when in a multilingual manner. The class label is

[1,2]Department of Computer Science and Engineering, GITAM School of Technology

GITAM (Deemed to be University), Andhra Pradesh, Visakhapatnam, India,

E-mail: idasri@gitam.in.

Copyright © JES 2024 on-line: journal.esrgroups.org

recognized when a border divides the speaker. Calculating the change within voice samples for which the histogram or frequency curve is considered is necessary for differentiating the speaker or speaker identification system.

In speaker diarization, a corpus or dataset is maintained with all such corpora. The speech samples are grouped into short sequence samples in this manner, and non-speech samples—non-trained samples—are filtered out as outliers in the next step, feature extraction, for which mfcc is considered. In the next step, the factors about the speaker are generated using a vector, and using this vector, the speakers are segmented. In the final phase, i.e., segmentation, these clusters are combined to form a refined output for identification or speaker diarization.

Neural networks have been used in speaker diarization systems rapidly in recent technological inventories [7], [8], [9], and [10]. In most literature, speaker diarization systems use text-based speaker verification and detection to identify the same speaker. In practice, text-independent recognition systems are necessary when working with realistic processes. Therefore, a system that can identify the relative speaker based on dynamic voice not labeled in the corpus must be built.

Thus, this article is an attempt to address this. An LSTM-based text-independent diarization system is used to develop the new model. This proposed article addresses the integration of Long Short-Term Memory (LSTM) networks with Mel-Frequency Cepstral Coefficients (MFCC) features for successful speaker diarization.

This is how the rest of the article is structured. An overview of recent research conducted in this field is provided in section 2 of the article.

Section 3 discusses the text-independent speaker verification system or LSTM. The algorithmic perspective of the suggested model is given in Section 4 of the paper. Section 5 of the article highlights the dataset under consideration. The clustering technique proposed is in section 6; the Methodology, together with experiments and the deduced outcomes, are described in section 7 of the article. The article is summarized in section 7.

## II. Related Work

Previous studies have divided the verification issue into more manageable but loosely related subproblems. As an illustration, the i-vector and probabilistic linear.

The most popular method for text-dependent speaker verification [9, 10, 11] and text-independent speaker verification [7, 8, 5, 6] is discriminant analysis (PLDA) [5, 6]. For text-independent speaker detection, hybrid techniques with deep learning-based components have also shown promise [12, 13, 14]. However, a more direct deep learning modeling might be a desirable substitute for tiny footprint systems [15, 4]. Recurrent neural networks have, to the best of our knowledge, not been used for the speaker verification job, but they have been used for related issues like speaker identification [16] and language identification [17].

Machine learning algorithms are considered in this article because of their capability to adapt to a diversified and dynamic dataset that suits different styles of speaker dialog pronunciation and dissimilar acoustic conditions. Another advantage associated with the machine learning algorithm is that it can automatically learn about relevant features from input data, thereby reducing the need for manual intervention. This capability is most beneficial in speaker diarization because it facilitates the extraction of discriminant features directly from the audio stream.

Among these machine learning techniques, lstm is preferred, and the advantages of choosing this technique are underlined in the following section of the article.

## III. Long Short- Term Memory (LSTM)

Leading-edge deep learning networks are designed in LSTM networks, a rapidly emerging subject. These techniques aid in determining the temporal correlations between subsequent data points. The basic ideas and importance of lengthy short-term memory are examined in this introduction within the framework of neural networks.

### *A. Speech Recognition Process*

By using long short-term memory banks (LSTMs) in speech recognition and transcription, we can improve accuracy by detecting recordings in audio. There is a separate application, which is Speaker discrimination, that can used for Long Short-Term Memory (LSTM) networks where it contains superior sequential data processing capabilities. By using Long Short-Term Memory (LSTM) networks, we can extract strong data by using suitable feature extraction techniques, including MFCCs

### *B. Mel-Frequency Cepstral Coefficients (MFCC):*

Recording an audio stream's spectral characteristics with a comprehensive feature is known as MFCC. By using discrete cosine transform, logarithmic compression, and Mel-filter bank analysis, we can produce accurate representations of the audio spectrum that can be useful in identifying speakers. Using Long Short-Term Memory (LSTM) with Recurrent neural networks (RNNs) will provide exceptionally good insight into identifying sample correlations in a series of data. LSTMs imitate the temporal patterns in MFCC sequences in speaker diary tasks, eventually enabling the network to identify complicated speech features. In brief, the characteristics of MFCC are combined with LSTM networks, which specifically identify speaker recognition solutions that improve performance and accuracy across different audio settings. The improvement of speaker recognition can be possible with deep learning and signal processing developments.

### *C.Clustering:*

The clustering techniques that are incorporated into our diarization system are described in this section. Our emphasis lies on the spectral offline clustering strategy, which demonstrated superior performance compared to alternative methods during testing, boasting a considerable margin of improvement.

Based on the run-time latency, clustering algorithms fall into two categories:

•Online clustering: When a segment becomes available, a speaker label is released immediately, without regard to subsequent segments.

•Offline clustering: Speaker labels are generated once all segment embeddings are available.

Because more contextual information is accessible in the offline scenario, offline clustering algorithms usually perform better than online clustering algorithms. Moreover, only in the offline configuration can the last re-segmentation step be applied. However, the application's nature ultimately determines whether to use it online or offline.

## IV. Process Model LSTM With Mfcc Features For Effective Speaker Diarization System

- Data preprocessing: Consider Librosa to extract MFCC features from audio samples.
- Segmentation: To facilitate analysis, split the audio stream into overlapping sections.
- LSTM Modeling: To examine temporal patterns in the MFCC sequences, build an LSTM network.
- Clustering: To classify segments with comparable speaker features, use clustering methods like k-means.
- Post-processing: Make sure that segment transitions are seamless and adjust speaker labels in light of contextual data.
- Obstacles: Using LSTM and MFCC characteristics to implement speaker diarization presents some obstacles, including.

## V. Dataset

Speech Accent Archive: The speech accent archive was created to consistently display a wide range of accents from different languages. The identical English text is read aloud by native and non-native speakers, and their responses are meticulously documented. The archive is designed to be both a research and teaching tool. Linguists and other individuals who want to listen to and contrast the accents of various English speakers are the intended users.

2140 voice samples from several talkers reading the same reading passage are included in this dataset. Talkers have 214 different native languages and are from 177 different countries. Every talker is using the English language.

*A. Speaker Diarization Procedure:*

Segmentation and clustering are the two primary phases in the diarization process. Initially, the audio stream is Speaker Diarization Procedure:Diarization procedure for speakers involves two phases: segmentation and grouping. Initially, the audio stream is separated into brief frames that represent temporal samples. Speaker diarization is the technique of categorizing audio portions in a recording based on individual speakers. Combining MFCC with LSTM networks is the most effective approach.

*B.* Mel-Frequency (MFCC):

Measures a sound stream's short-term power spectrum. These techniques are widely utilised in speech and audio processing due to their ability to accurately identify speakers.

*C. Diarizing Speakers:*

Audio streams are divided into portions with uniform speakers and assigned unique speaker identities. The diarization process involves two primary processes: segmentation and clustering.

*D.* Long Short-Term Memory with MFCC LSTM

networks may represent temporal dependencies inside MFCC sequences for each audio segment in a speaker diarization setting.

*E.* Workflow:

- Preprocess data to extract MFCC features from audio segments using programmes like Librosa.
- Segment the audio stream into brief, overlapping chunks.
- To extract temporal patterns from MFCC sequences, use an LSTM network.
- Clustering: Use k-means to classify segments with similar speaker attributes.
- Refinement: Adjust speaker labels for smooth transitions and contextual consistency.

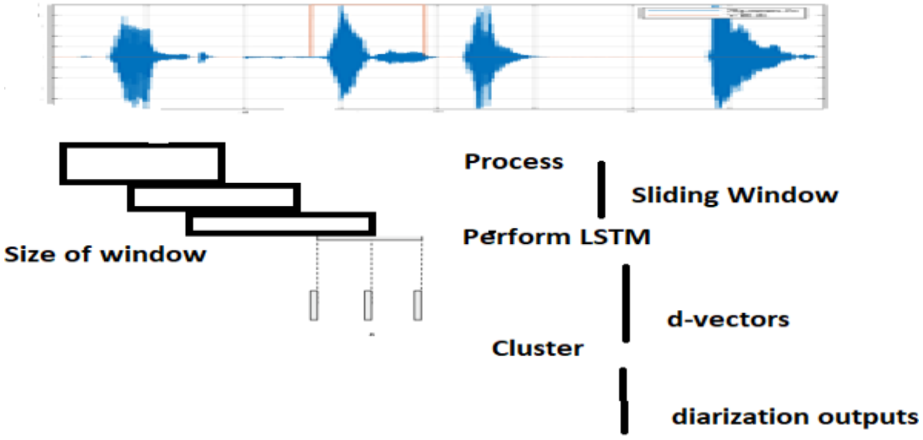

**Figure 1: Architecture**

### F. *Step-by-step process of Speaker Diarization Algorithm Employing LSTM and MFCC Features.*

- Each audio file contains multiple speakers. Before processing, separate the audio into short chunks, such as one or two seconds.
- Extract the MFCC features from each segment using Librosa or a comparable library.
- Building an LSTM Model: Define an LSTM network architecture for sequence modeling.
- Input layer: MFCC sequences for each segment.
- To capture temporal dependencies, use LSTM layers.
- A labeled dataset containing speaker annotations will be used to train the LSTM network.
- Choose an appropriate loss function for jobs involving sequence labeling.
- Modify hyperparameters, including batch size and learning rate.
- Partition-wise Deduction: Apply the learned LSTM model to each segmented audio fragment to perform inference.
- Get each segment's speaker embedding.
- Clustering: Apply a clustering technique to group speaker embedding (k- means, for example). The number of speakers and the number of clusters match.
- Enhancement: Adjust the speaker clusters in light of the surrounding situation.
- Following processing, seamless changes in speaker segments to improve the diarization outcomes. Also, address instances of background noise and overlapping voice.
- Results: final divided output, with portions identified to correspond to various speakers.
- Compute the keyword. A one mask value corresponds to a segment where the keyword was spotted. mask = classify (KWSNet,features.');
- Plot the test signal and the mask.

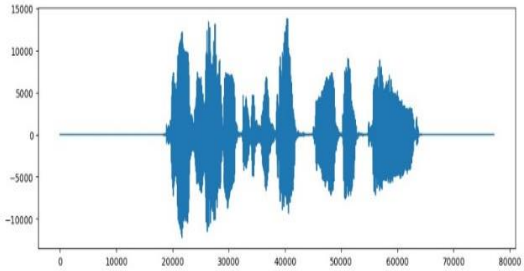

**Figure 2: Exploratory data analysis of audio**

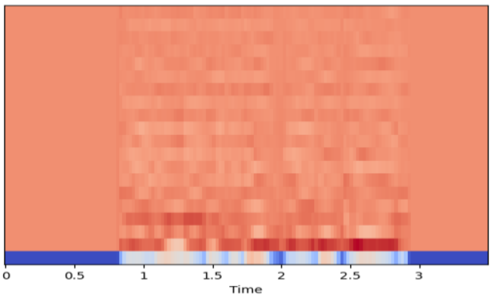

**Figure 3: Raw Audio to MFCC Feature Extraction**

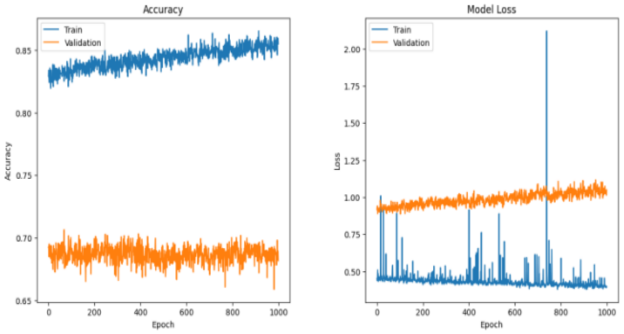

**Figure 4: Test Accuracy and Loss for The Hold Lstm Model**

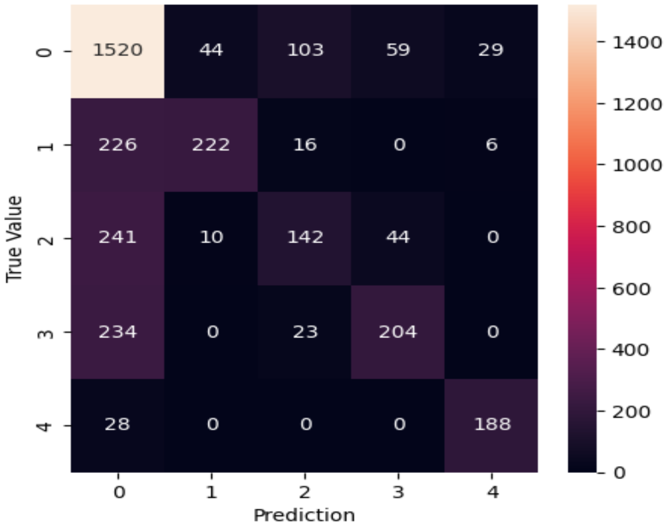

**Figure 5: Confusion Matrix for The Model**

LSTM Model Predictions: Relative speaker identification in given corpus, it is able to identify different accents with accuracy of78%.”.

**Table 1. LSTM Model prediction**

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.775 | 0.866 | 0.758 |
| 1 | 0.804 | 0.478 | 0.599 |
| 2 | 0.542 | 0.324 | 0.393 |
| 3 | 0.664 | 0.442 | 0.53 |
| 4 | 0.866 | 0.84 | 0.852 |

LSTM Model Predictions for Dynamic voice: prediction of voice that is not labeled in corpus it is able to identify different accents with accuracy of 77%

## VI. CONCLUSION

To create a novel d-vector-based method for speaker diarization, we expanded on the achievements of previous d-vector-based speaker verification systems in this study. To be more precise, we used recent advancements in non-parametric clustering with LSTM-based d-vector audio embeddings to produce an advanced system for speaker diarization.

## REFERENCES

[1] Betser, Michael. “Speaker Diarization Using Bottom-up Clustering Based on a Parameter-Derived Distance between Adapted GMMs.” Proc. ICSLP (2004):

[2] httS. Madikeri and H. Bourlard, "KL-HMM based speaker diarization system for meetings," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 4435-4439, doi: 10.1109/ICASSP.2015.7178809.

[3] Barhoush, M., Hallawa, A. & Schmeink, A. Speaker identification and localization using shuffled MFCC features and deep learning. Int J Speech Technol 26, 185–196 (2023). https://doi.org/10.1007/s10772-023-10023-2

[4] Nemer, Elias & Goubran, Rafik & Mahmoud, Samy. (2001). Robust voice activity detection using higher-order statistics in the LPC residual domain. Speech and Audio Processing, IEEE Transactions on. 9. 217 - 231. 10.1109/89.905996.

[5] Indu, D. ., & Srinivas, Y. . (2024). A Cluster-Based Speaker Diarization System Combined with Dimensionality Reduction Techniques . International Journal of Intelligent Systems and Applications in Engineering, 12(14s), 125–132.

[6] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, “Speaker diarization using deep neural network embeddings,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 4930–4934.

[7] Sepp Hochreiter and Jurgen Schmidhuber, “Long short-term ¨ memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] Ulrike Von Luxburg, “A tutorial on spectral clustering,” Statistics and computing, vol. 17, no. 4, pp. 395–416, 2007.

[9] Huazhong Ning, Ming Liu, Hao Tang, and Thomas S Huang, “A spectral clustering approach to speaker diarization.,” in INTERSPEECH, 2006.

[10] Philip Andrew Mansfield, Quan Wang, Carlton Downey, Li Wan, and Ignacio Lopez Moreno, “Links: A highdimensional online clustering method,” arXiv preprint arXiv:1801.10123, 2018.

[11] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in Proc. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, Jul. 2010.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[13] D. Reynolds, T. Quoter, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, no. 1, pp. 19–41, 2000.

[14] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1435–1447, 2007.

[15] H. Aronowitz, “Text-dependent speaker verification using a small development set,” in Proc. Odyssey Speaker and Language Recognition Workshop, Singapore, Jun. 2012.

[16] T. Stafylakis, P. Kenny, P. Ouellet, P. Perez, J. Kockmann, and P. Dumouchel, “Text-dependent speaker recognition using PLDA with uncertainty propagation,” in Interspeech, Lyon, France, Aug. 2013.

[17] A. Larcher, K.-A. Lee, B. Ma, and H. Li, “Phoneticallyconstrained PLDA modeling for text-dependent speaker verification with multiple short utterances,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, May 2013.

[18] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, “Improving speaker recognition performance in the domain adaptation challenge using deep neural networks,” in IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoie, NV, USA, Dec. 2014, pp. 378–383.

[19] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phoneticallyaware deep neural network,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, May 2014, pp. 1695–1699.

[20] F. Richardson, D. Reynolds, and N. Dehak, “Deep neural network approaches to speaker and language recognition,” IEEE Signal Processing Letters, 2005.

[21] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, May 2014.

[22] S. Parveen, A. Qadeer, and P. Green, “Speaker recognition with recurrent neural networks,” in Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, Oct 2000, pp. 16– 20.

[23] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. Moreno, “Automatic language identification using long short-term memory recurrent neural networks,” in Interspeech, Singapore, Sep. 2014, pp. 2155–2159.