Received: March 5, 2024. Revised: April 19, 2024.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Hybrid DMFCC-LPC Based Feature Extraction with DCNN Clustering for Speaker Diarization

Jhansi Rani Kaka[1]* **Vijay Kumar Kangala[2]**

[1]*Department of Electronics and Communication Engineering,*
Jawaharlal Nehru Technological University, Kakinada, India
[2]*Department of Computer Science and Engineering,*
Srinivsa Institute of Engineering and Technology, Amalapuram, India
* Corresponding author's Email: jhansikaka@jntucek.ac.in

Abstract: Speaker Diarization (SD) or speaker indexing is a procedure for automatically partitioning a conversation by the number of speakers into homogeneous segments. The trustworthy diarization method accurately estimates the variable length assertion, and it involves major steps such as speech detection, speaker merges, and speaker change. The major problem with the SD method is enhancing the readability of speech transcription. In this study, a hybrid method of feature extraction based on the Dynamic Mel Frequency Cepstral Coefficient (DMFCC) and Linear Prediction Coding (LPC) is proposed for SD. The Voice Activity Detection (VAD) method is utilized to detect the presence or absence of a speaker in the audio lecture, which is followed by speaker segmentation utilizing the extracted features. The Deep Convolutional Neural Network (DCNN) is utilized to determine the feature vector and cluster the speaker from the audio lecture. The results show that the proposed DMFCC-LPC delivers a robust performance on metrics such as accuracy, Diarization Error Rate (DER) and False Positive Rate (FPR) of 0.967, 0.31 and 0.119 on CALLHOME dataset, in contrast to the Speaker Siarization System using HXLP-DCNN with Sailfish Optimization Algorithm (SDS-HXLP-DCNN-SOA), Feature-Level fusion, and Self-supervised clustering with Path Integral Clustering (SSC-PIC).

Keywords: Dynamic Mel frequency cepstral coefficient, Linear prediction coding, Speaker diarization, Speaker segmentation, Voice activity detection.

1. Introduction

With the development of advanced technologies in the engineering field, a wide range of intelligent and efficient methodologies have appeared to improve the quality of the community [1]. When these technologies arrived in the interaction (HCI), there was a forever enhancing appeal to building an automated human language recognition model that accesses in conceptual ways for communication [2]. Efficient communication methods among humans and computer are challenging in advanced technologies. The easiest way for the community to enter information is by signals of speech [3]. Hence, the speech signal processing method and its tools are the most important things in information society. Diagnosis is an essential thing in the speech recognition process as it separates input audio reporting into various speech reports, each of which belongs to the individual speaker [4]. Anciently, recognition integrates an audio reporting segmentation into a single expression and the outcoming segment's clustering [5]. The speaker's diarization is a challenging task because of the development of reported speech, which contains audio broadcasts, voice messages, meetings, and television [6].

The diarization method majorly utilizes unsupervised Machine Learning (ML) algorithms when expressions are transferred among speakers [7]. However, it does not know which diarization labels are applied to a specific speaker, this method is called unsupervised diarization. A number of supervised

and unsupervised learning approaches are developed for the speaker diarization (SD) process, which is trained for segmentation and clustering [8]. The SD contains various components such as segmentation, Voice Activity Detection (VAD), re-segmentation and steering [9]. Segmentation recognizes where the speaker alters in an audio report and clustering group's speech segments according to the major components of speaker [10]. The major component of this differentiation is clustering, in which the segmentation involves the bottom-up, global optimization, neural network clustering, as well as up-down approach. The existing approaches for clustering contain Deep Neural Network (DNN), spectral clustering, and bottleneck-based methods [11]. The existing methods of the SD have traversed by clustering free methods or end-to-end approaches by multiple speakers' discussion [12]. The foremost contribution of this research are as follows:

- The FE using Hybrid DMFCC and LPC techniques-based speaker diarization is proposed and the noise removal approach is utilized for the pre-processing step to enhance the training data.
- The VAD approach is utilized for the identification of speech and non-speech signals, and segmented by using extracted features by the DMFCC and LPC.
- The Deep CNN algorithm is used for the speaker clustering process, while the performance of the proposed method is evaluated by utilizing the existing methods.

The rest of this paper is arranged as follows: Literature survey is explained in Section 2. The proposed methodology is described in Section 3, while the experimental results are explained in Section 4, and conclusion is described in Section 5.

2. Literature survey

Sailaja [13] developed a segmentation and classification of the diarization of the speaker. This method utilized Tangent weighted Mel-Frequency Cepstral Coefficient (TMFCC) and HXLPS as well as Linear Prediction Coding (LPC) by autocorrelation snapshot for the process of the Feature Extraction (FE). This method implemented the Deep Convolutional Neural Network (DCNN) for the clustering and Sailfish algorithm for the problem of optimization. The VAD approach was utilized to identify signals of a speech and non-speech. The TMFCC provided the most efficient results, as well as enhanced the effectives of this method by utilizing maximum energy. However, the developed approach had required the large labeled dataset and it had time-consumed to collect the large data.

Sethuram [14] implemented a novel MFCC-based FE method of the speaker diarization (Telugu language) model. Later, an Optimized Artificial Neural Network (ANN) was introduced for the process of clustering. ANN training played a vital role in optimization logic that updated ANN weight by integrating the Artificial Bee Colony (ABC) and Lion Algorithm (LA) called ANN-ABC-LA. The LA approach achieved minimized errors and achieved the most efficient results and flexibility. But, the MFCC FE approach was difficult to clustering the diarization and needed a greater computational cost and low scale ability.

Chauhan [15] introduced a Speaker Recognition (SR) approach-based on various types of features of the speech. The novel kind of feature aggregation approach utilized eighteen features namely, MFCC, LPC, PLP, RMS, centroid, delta, and delta-delta values. This approach was evaluated on various speech size datasets and a total of 315 feature fusion approaches were evaluated on the NIST-2008 dataset. The most efficient 35 approaches were chosen and evaluated on balancing datasets for quick computation to acquire better fusion approach. The introduced approach had provided the better results only by utilize the smaller fusion models, however, performance and training testing duration was affected due to utilize large datasets.

Rajeyyagari [16] presented a speaker diarization approach based on Deep Long Short-Term Memory (LSTM). This method utilized a noise removal approach to remove unwanted noise in the collected input from E-Khool users. The LPC FE extraction approach was utilized for extracting features from an audio discourse of E-Khool users. A VAD approach was deployed to identify the speaker's presence and absence in audio discourse, which was followed by speaker segmentation. Eventually, feature vectors were determined, and the audio discourse was clustered using deepLSTM. This approach effectively enhanced the tracking of the distance of the speech signal. However, the presented approach was sensitive to various random weight initializations.

Khoma [17] developed a multi-architecture of the speaker recognition systems based on the integration of identification and diarization approach. This approach was segmented according to segment-level or group-level classification, aside from an open-source PyAnnote module utilized to build the system. A SD approach's performance was estimated by an AMI Corpus audio dataset. This dataset contained the 100 *hz* of annotated and transcribed data of both the audio and video. This method had simplicity for utilization and was used in many real-time applications. However, there was no possibility to

adjust the system parameters and was dependent on the service provider infrastructure.

Niu [18] developed an Iterative based SSD with QDM, named QDM-SSD. The QDM-SSD enhanced the simulated data utilized for the conversion of model by QDM to reduce an error effect in the priority of speaker. Due to the distillation of quality of data, the QDM-SSD made the reduction data sparse by modifying speaker overlap ratios based on the quality of data. Additionally, by utilizing a sliding window over conversion data, cleaning regions in segmenting the speech could be localized. The QDM-SSM decreased both speaker false alarm rate and misclassification compared to the ISSD, but it was not directly utilized on speech segments as it contained few assured interfering speeches.

Singh and Ganapathy [19] developed a depiction learning and clustering approach that was iteratively performed for an enhanced SD. A depiction learning was established on concepts of self-supervised learning, simultaneously, a clustering approach was a graph structural model according to the Path Integral Clustering (PIC). The depiction learning procedure utilized cluster targets from PIC, while the cluster was employed on embedded learning from self-supervised deep model and was based on the Self-Supervised Clustering (SSC). The SSC provided the enhanced cluster separability with the enhanced association with ground-truth speakers. This method had a computational complexity due to trained on the DNN and PIC.

He [20] implemented a Neural Speaker Diarization (NSD) network framework for speaker recognition. This method contained the following three characteristics: Initially, Memory-aware Multi-Speaker Embedding (MA-MSE) approach was implemented to simplify the dynamic refinement among NSD and extraction of embedding. Then, the selection process of speaker was presented to the control condition when the detection of speaker was varied from NSD. Eventually, an adaptive process was implemented to enhance the needed information for nonoverlap speech segments in a provided expression during every iteration. This method reduced the computational complexity, but this method caused an overfitting problem by utilizing the neural network for collecting the information.

VijayKumar and Rao [21] developed a clustering with hybrid optimization technique was utilized to perform the SD. The Speech Activity Prediction (SAP) was deployed to process the extract features from audio signal according to the recognized segments of the speaker. The Deep Embedded Clustering (DEC) was employed to complete the process of diarization, in which the constants were trained through enhanced Fractional Anticorona Whale Optimization Algorithm (FrACWOA). The DEC continuously identified the most efficient solution for clustering objective by using the mapping function. Nonetheless, the developed method failed to compute with the number of speakers.

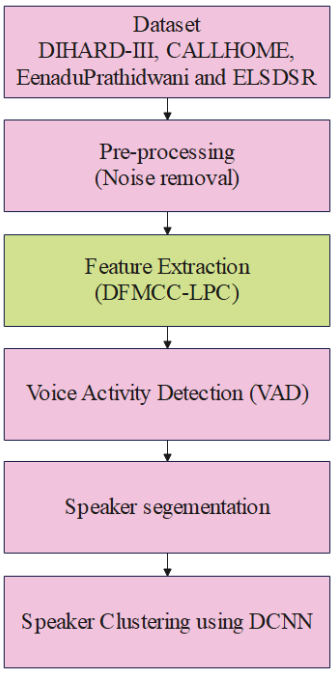

Figure. 1 Block diagram of the Speaker Diarization

3. Proposed methodology

This section provides an overall summary of the suggested method of SD. This proposed method consists of various processes such as input audio signal, pre-processing, feature extraction, Voice Activity Detection, segmentation, clustering and optimization. Fig. 1 depicts the block diagram of the proposed method.

3.1 Dataset

The proposed method utilizes the various datasets an input for the audio signal dataset named DIHARD-III [22], CALLHOME [23], EenaduPrathidwani [21] and English Language Speech Database for Speaker Recognition (ELSDSR) [24], which is extensively utilized for the speaker diarization. The DIHARD-III dataset involves the single channel input level improvement and its estimation set involves 5-10minutes long and it obtained from the 11 conversational domains, each consists of approximately 2h of audio. The CALLHOME is the group of various lingual telephone data of 500 recordings, the range of each recording is 2-5 min per

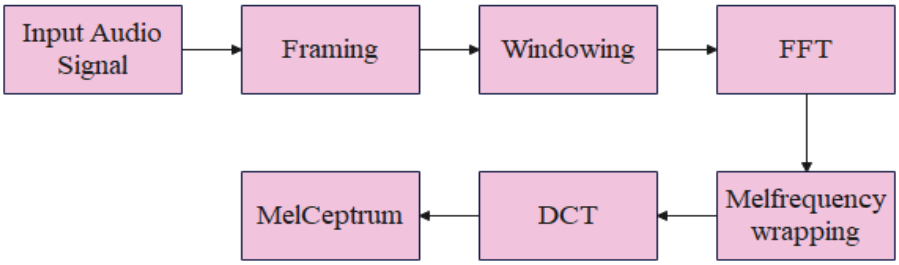

Figure. 2 MFCC process

file. The count of the speakers in each recording differs as 2-7. EenaduPrathidwani dataset involves an echo signal of human emotions with present problems such as economic, cultural dimensions as well as political.

The ELSDSR is a small corpus of read speech developed to give data for the evolution and automatic identification method. The ELSDSR design was recorded by the faculty, Doctor of Philosophy, and Master students from Informatics and Mathematical modelling department, Technical University of Denmark (DTU). In this dataset, 20 Danes, 1 Icelander and 1 Canadian spoke in English language. This dataset consists of speech data from 22 speakers such as 12 males and 10 females. The time taken the duration the training is 78 seconds and 88.3 seconds for males and females. For training, 154 speeches were recorded, each with seven phrases. For testing, 44 voices were given with two phrases were spoken by each person. The test duration takes 16.1 seconds for males and 19.6 seconds for female. The collected input speaker data are provided to the pre-processing step to improve the performance of the proposed method.

The collected four datasets have been divided into two parts such as 80% of the data is utilized for training and 20% of the data is utilized for the testing.

3.2 Pre-processing

The pre-processing step utilizes input from the collected speaker dataset, and at the initial stage after the collection of data, attempts are made to enhance the data quality for generating high-performance outcomes. The process of transforming the data into a form that is easily parsable by a machine is called data pre-processing. Initially, an input speaker signal is defeated for pre-processing at noisy regions which are removed by noise removal approach to acquire the maximum quality in audio signal. The reduction of noise is achieved by means of decomposition of the wavelet sub-band [25]. The procedures employed to remove the noise are:

- The decomposition of wavelet is evaluated by selecting a wavelet at N level.
- A threshold is chosen for each level 1 to N, where soft thresholds are put for estimation and attribute coefficients.
- The wavelet redevelopment is evaluated utilizing updated estimation and attributes.

This pre-processed input signal is further provided to the extraction process by utilizing the DMFCC and LPC.

3.3 Feature extraction

After the audio signal pre-processing, one of the most significant processes is the extraction of feature. The feature extraction takes place to parse the speech signals features, thus, it is usable as the differentiator among speech signals from one another. In this study, the feature extraction method is proposed by utilizing the hybrid method of Dynamic Mel Frequency Cepstral Coefficient (DFMCC) and Linear Predictive Coding (LPC).

3.3.1. Dynamic MFCC

In MFCC FE, an audial feature is extracted from speech signal to depict the features of the speech signal. The denoising process of the speech is utilized for wavelet transform, performed before MFCC FE [26]. The Fast Fourier Transform (FFT) feature data is determined in MFCC and is taken out to convert speech signal data into frequency domain. Fig. 2 shows the process of the MFCC.

3.1.1.1 Pre-emphasis

The pre-emphasis is a filtering process aiming to acquire a smoother spectral form of the speech signals frequency, alongside minimizing noise during the retrieval of speech. The pre-emphasis is formulated in Eq. (1) as follows:

$$y[n] = x[n] - a\,x[n-1] \tag{1}$$

Where, $x[n]$ and $y[n]$ are the input and output speech signals, and a is the filter constant with the characteristics $0.9 < a < 1.0$.

3.1.1.2 Framing and windowing

The framing is taken out to segment the signal into smaller parts in an overlapping manner. In this method, the speech signal is segmented at 25ms with overlapping along 5ms. The process of windowing is carried out to minimize the spectral leakage and to reduce the signal discontinuity in each process. The output of the hamming windowing process is formulated in Eq. (2) as follows:

$$y1(n) = x1(n)w(n), \quad 0 \le n \le N-1 \tag{2}$$

Where, $x1(n)$ and $y1(n)$ are the input and output signals in framing and windowing, $w(n)$ is the hamming window, N is the number of signal samples in each process, wherein 0.54 and 0.56 are the fixed coefficients.

3.1.1.3 Fast fourier transform

The FFT is commonly used in digital signal processing that serves to modify the signal from time into a frequency domain and is provided in Eq. (3) as follows.

$$X_n = \sum_{k=0}^{N-1} x_{k e^{-2\pi jkn/N}} \tag{3}$$

Where, X_n is the outcome of FFT, X_k is the input signal while the value of $N = 0, 1, 2, \dots, N-1$. The FFT outcomes are examined to identify the r threshold separating the noise-unaffected data (x) and noise-affected data (y). The (y) is then denoised utilizing the wavelet transform. Further, (x) does not experience the denoising process. The y is initially converted back to the time domain before the process of denoising by using an Inverse Fast Fourier Transform (IFFT), which is given in Eq. (4).

$$y_n = \sum_{k=0}^{N-1} y_{n e^{2\pi jkn/N}} \tag{4}$$

Where, y_n is the input of IFFT, y_k is the outcome which is transformed into the time domain. The outcome of y_k data transformation that becomes y_n, later integrates with the noise-unaffected data x. If the data x was equal to r, then, data y as the transformation result and it was equal to $256 - r$ (from data number r to 256). The integration of data x and y is represented in Eq. (5).

$$X_{total} = \left[x_{(1\dots r)}, y_{(r\dots 256)}\right] \tag{5}$$

Where, X_{total} is the integration outcome of noise-unaffected data x and the outcome of the data from y.

3.1.1.4 Mel frequency wrapping

Mel frequency wrapping is processed with the bandpass filter. The filter is designed in the triangle form which has a linear characteristic below the frequency of $1000Hz$ and a logarithmic above $1000Hz$. This filter is utilized to evaluate the number of spectral filter components, where the outcome is the total of filtered spectral components. It is later applied to the data of X_{total} after being designed. The number of triangle filters designed in this method is 26, thus the total MFCC outcome features are 26 per frame.

3.1.1.5 Discrete cosine transform

The final procedure of the MFCC is the DCT which is provided in Eq. (6).

$$\sum_{k=1}^{N} log(Y(i)cos[mx(k-0.5)x\pi \div N \tag{6}$$

Where, N is the number of bandpass filter triangle, the value of m is among 1 to L, while L is the total of coefficients outcome.

3.3.2. Linear predictive coding

The LPC [27] is employed for an encryption approach for data privacy, which evaluates the present sampling by utilizing a linear integration of previous samples. The IFFT is performed and the harmonic approach is eliminated from speech signals and balancing signal that is abandoned after IFFT is known as residue. The features of the LPC are evaluated utilizing a VQ-LBG algorithm. To minimize the bit rate, VQ is request to transform the features of LPC to Linear Spectral Frequency (LSF). This is a significant step for recognizing an autoregressive model of speech with respect to LPC. The audio signal is modelled in the p-th order. In this process, each sample is provided in Eq. (7).

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + u(n) \tag{7}$$

Each sample at n th instant depends on 'p' previous sample, extended with Gaussian noise $u(n)$, while a is the LPC coefficient. The Yule-Walker equations are utilized to evaluate coefficients and the final equation is provided in Eq. (8).

$$\sum_{k=1}^{p} a_k R(l-k) = R(l) \quad (8)$$

This method uses only 13 LPC coefficients, in order to minimize the complexity of the model.

3.3.3. Hybrid DMFCC-LPC

As described in the previous sections, both DMFCC and LPC models have various attributes of the speaker's voice which are further utilized separately for the recognition of the speaker. Even MFCC has the representation of high-fidelity and stable recognition process, but with lesser robustness to noise signals as noise signals change all MFCCs if at least one frequency band is oblique. To solve this problem, the DMFCC and LPC is combined to improve the performance of the proposed method. The LPC provides a robust and accurate method for evaluating the parameters that depict the vocal tract system. The final process of DMFCC is provided as input to the LPC method. The affirmation algorithms of DMFCC and LPC coefficients communicating the required speech features are developed. The hybrid utilization of cepstral of DMFCC and LPC in speech recognition system is proposed to enhance oblivious quality of the speech recognition system. The recognition system is separated into MFCC and LPC-based recognition systems. The training and recognition processes are accepted in both subsystems individually, while the recognition method obtains the choice with similar effects of every subsystem. The DMFCC features describe perceptual speech features, whereas the LPC features describe vocal tract model for speaker in speech audio. A hybrid method of DMFCC and LPC features are utilized for representing the irregular features of the speaker's message. In this method, a neural network-based feature level fusion approach is designed to combine and estimate voice features in DMFCC and LPC, into d-dimensional combined feature space. Now, value of d depends on architectures of the neural network. The LPC provides better results for speaker recognition rather than speech recognition. Now the 10th order LPC is carried, which means that the 11 coefficients of each frame are acquired. The 12 coefficients of DMFCC and 11 coefficients of LPC are integrated on frame-by-frame basis to earn 23 coefficients of each frame. The combined feature space is studied in such a way that the combined feature depiction selects effective speaker dependent voice features for enhancing the performance of speaker recognition. The speech features are extracted by using the hybrid DMFCC and LPC. Then, these extracted features are provided for voice activity detection.

3.4 Voice activity detection

The VAD [28] is the speaker signal detection approach where the speaker is evaluated by the goodness or badness of each signal of speech. The audio signals described by recognizing the speaker are utilized for speaker classification and to remove the non-speech area. The audio with MFCC features is applied in a GMM mode, whereas for the number of listeners, an audio signal depends on the vocal tract [29]. The GMM have binary characteristics: speaker identification through combining the Gaussian component as well as for achieving the smooth approximation. The Gaussian features provide a preferable solution. The Bayesian Inference Criterion (BIC) is utilized to determine the variability evaluations, as well as for the recognition of speaker. The GMM is identified for the audio signal, as expressed in Eq. (9).

$$ac(x) = \sum_{k=1}^{q} wt_k P(\mu_k, \sigma_k, x) \quad (9)$$

Where, $P(\mu_k, \sigma_k, x)$ is the Gaussian with mean and covariance, q is the Gaussian component, and wt is the weight. Then, BIC is provided significantly in enhanced probability measure for the activity diagnosis. A BIC supports to minimize false alarm rates in VAD. Now, the basic form BIC is identified by using Eq. (10).

$$R(s) = log\ Lg(V|S) - \xi \frac{1}{2} X\ log(M_a) \quad (10)$$

Where, $log\ Lg(V|S)$ is the likelihood log measure, M_a is the frame size V, ξ is the penalty weight, X is the perplexity model, and S is the desire model. The threshold value for speaker from input of audio signal is predetermined. The voice activity is recognized by the characteristic's speaker signal, as extracted by utilizing the Eq. (11).

$$S = \{S_1, S_2, \ldots S_p\} \quad (11)$$

Moreover, the signal score value from BIC is evaluated and the recognition of speech activity is completed if the BIC value is maximum, when compared to the value of threshold. The score value of BIC is utilized for speech activity. Then, the detected voice is further provided to the segmentation process.

3.5 Speaker segmentation

A speaker segmentation approach which develops the window size w by ΔBIC distance is utilized. Initially, an analysis is taken out for a single speaker modification from the beginning of the audio, as well as for each endure modification; an analysis is reworked on the following frame. Then, a search window is reported and ΔBIC distance is determined for all frames placed inside a window. Once the maxima surpass the threshold value Ψ , the modification approach will result maximum record in the segmentation. There are no maxima identified on window, its size is enhanced and this procedure is iterated until the modification is endured. After the non-speech removal by VAD, the signals of the speech are performed and the modification points from audio signal is detected, while its corresponding locations from traditional audio are recognized and recorded as modification points. Earlier, the procedure of segmentation is computed using two steps; initially, the ΔBIC based modification identification is taken out according to above produced threshold value. Next, an integration of achieved segments are carried only for those having successful ΔBIC score. These two steps are evaluated because of the problem of over-segmentation in zero threshold ΔBIC based segmentation. Through the purpose of removing the process of these two steps, the maxima value that surpasses the threshold Ψ is examined for the following process, and is effectively reduced for over-segmentation by using the Eq. (12).

$$\Delta BIC(y_i) = N\ log\ |E| - N_1 log|E_1| - N_2 log|E_2| - \frac{\gamma}{2}\left(b + \frac{1}{2}b(b+1)\right) log\ N \quad (12)$$

The features of the speaker are segmented by using the outcome of extracted features. Then, the similar speaker signals collect and cluster the signal of an individual speaker by using the DCNN algorithm.

3.6 Speaker clustering using DCNN

Convolutional Neural Network (CNN or deep CNN) is utilized in this section to classify clustering of a speaker. A supporting training procedure enhances an efficiently detected data when the speaker diarization method extends the additional features. The physiological theory based mathematical model is appeal to the neurons or training space correlation. Each function emerges to plays a significant part on the properties of the feedback for the framework. The CNN contains three layers such as input, hidden as well as output layer. An input layer is utilized to provide an input in network, hidden layers are utilized to determine suitable features of input data. Then, output layer is utilized to provide output from network. A convolutional, pooling and Fully Connected (FC) layers are building blocks, which are utilized to learn mapping features. The convolution is utilized to evaluate the stride value, kernel size, paddings as well as biases. The convolutional layers perform the input signal transformations through convolving it by a kernel. These transformations generate the number of feature maps correlated to an input signal. The outcome of this evaluation produces the number of feature maps for an input data. To find a difficult features, a group of number of convolutional layers by definite sizes of kernel is utilized. The evaluation of convolutional layer is formulated in Eq. (13) as follows;

$$O_n^i = \left(\sum m \epsilon X_n\ O_m^{i-1} \times K_{mn}^i + B_n^i\right) \quad (13)$$

Where, O_n^i and O_m^{i-1} is the outcome of ith and $(i-1)$ layer, K_{mn}^i is the convolutional kernel of ith layer, B_n^i is the bias of ith convolution layer, and X_n is the selection of input map. The ReLU layer is utilized after a convolutional layer to exchange all refuse activities to zero. The ReLU is selected for an activation function because of the access to a faster convergence. The ReLU activation function estimation is formulated in Eq. (14).

$$ReLU\ (a) = maximum(0, a) \quad (14)$$

Later, the ReLU layer's outcome is provided to a maxpooling layer to estimate the maximum value of signals. A max-pooling layer minimizes feature maps according to the sampling mask; hence, pooling layer minimizes various features for learning. Later, each convolution is followed by the batch normalization to normalize the input channel over mini-batch size, further accompanied by a ReLU layer. The clustered features are then provided in the results section to evaluate the performance.

4. Experimental results

The proposed DMFCC-LPC approach for SD is estimated by using the ELSDSR dataset and implemented in the MATLAB tool. The proposed method uses various performances metrics to evaluate the performance based on accuracy, tracking distance, False Alarm Rate (FAR), Diarization Error Rate (DER), False Discovery Rate (FDR), False

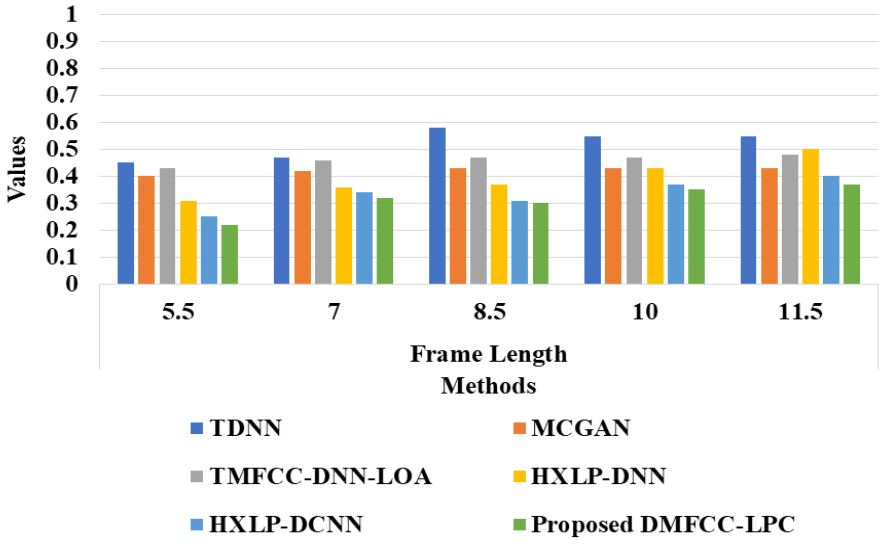

Figure. 3 Graphical representation of FAR

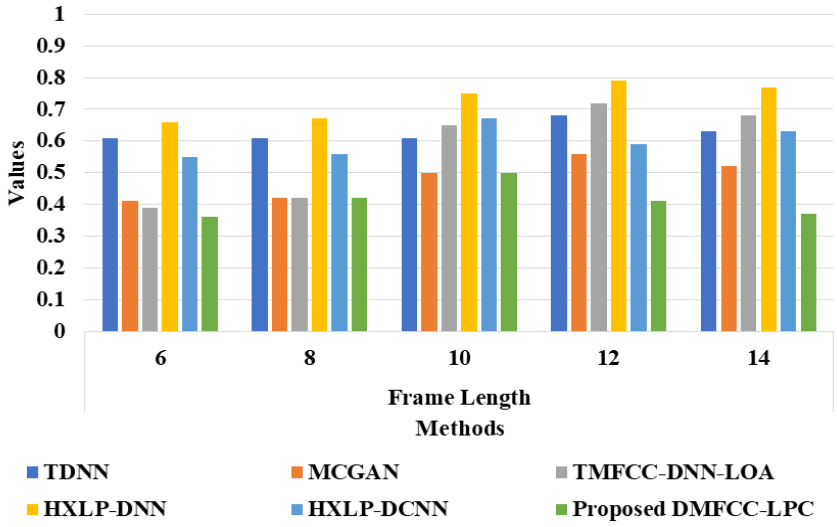

Figure. 4 Graphical representation of DER

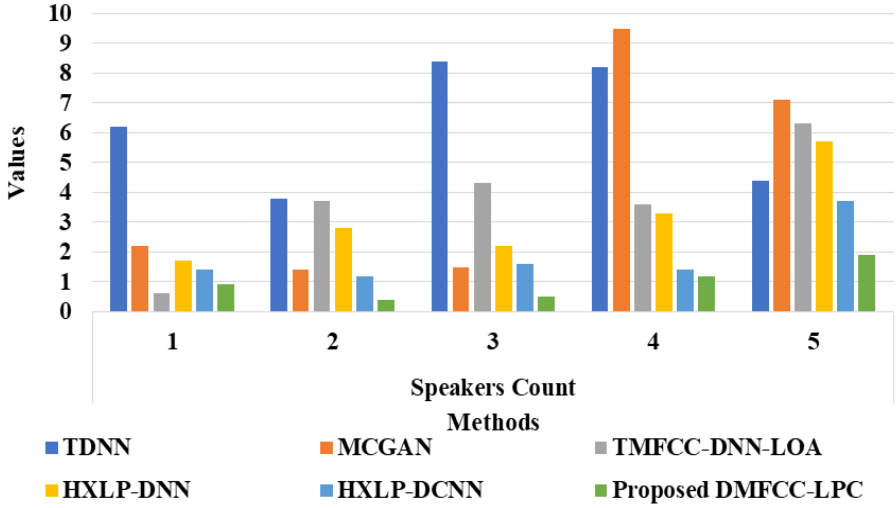

Figure. 5 Graphical representation of speakers counts with DER

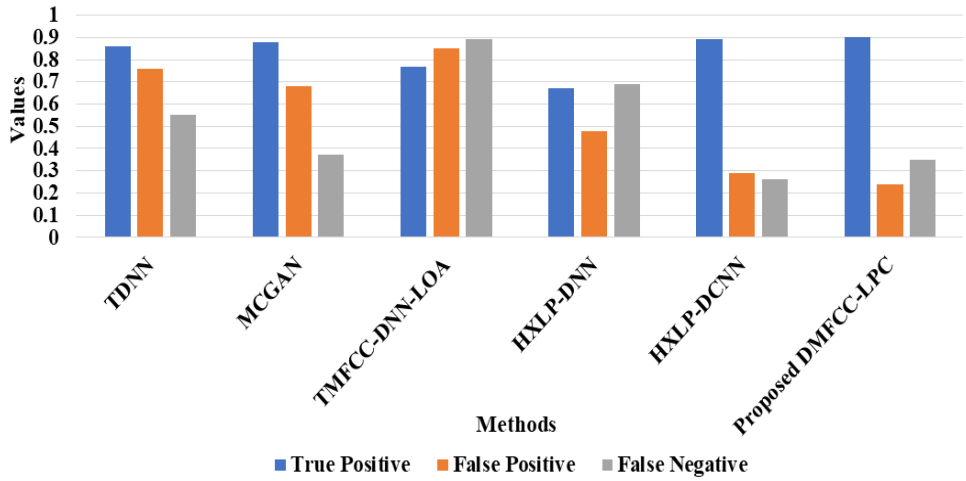

Figure. 6 Graphical representation of classifiers with FAR

Table 1. Performance analysis of FAR

Methods	Frame Length				
	5.5	7	8.5	10	11.5
TDNN	0.45	0.47	0.58	0.55	0.55
MCGAN	0.4	0.42	0.43	0.43	0.43
TMFCC-DNN-LOA	0.43	0.46	0.47	0.47	0.48
HXLP-DNN	0.31	0.36	0.37	0.43	0.52
HXLP-DCNN	0.25	0.34	0.31	0.37	0.49
Proposed DMFCC-LPC	0.22	0.32	0.35	0.40	0.43

Table 2. Performance analysis of DER

Methods	Frame Length				
	6	8	10	12	14
TDNN	0.61	0.60	0.57	0.52	0.49
MCGAN	0.57	0.52	0.50	0.46	0.42
TMFCC-DNN-LOA	0.55	0.51	0.49	0.44	0.41
HXLP-DNN	0.66	0.63	0.60	0.59	0.53
HXLP-DCNN	0.55	0.52	0.49	0.45	0.43
Proposed DMFCC-LPC	0.39	0.37	0.33	0.30	0.28

Positive Rate (FPR) and False Negative Rate (FNR). The mathematical expressions of these metrics are provided in Eqs. (15)-(21).

$$Accuracy = \frac{Number\ of\ speechs\ correctly\ identified}{Total\ number\ of\ audio\ files} \quad (15)$$

$$Tracking\ distance = \sqrt{\left(W_m^F - W_m^g\right)^2} \quad (16)$$

$$FAR = 1 - specificity \quad (17)$$

$$DER = \frac{Confusion\ error + Miss\ error + False\ alarmrate}{Total\ reference\ speech\ time} \quad (18)$$

$$FDR = \frac{P_f}{t} \quad (19)$$

$$FPR = \frac{FP}{FP+TN} \quad (20)$$

$$FNR = \frac{FN}{FN+TP} \quad (21)$$

Where, W_m^F is the original signal; W_m^g is the output signal. P_f – false discovery; t – discovery count; FP – False Positive; TN – True Negative; FN – False Negative; TP – True Positive.

4.1 Performance analysis

In this section, performance analysis of the proposed DMFCC-LPC approach for SD are discussed. Table 1 displays the performance analysis of FAR for the presented method with the existing methods.

Table 1 and Fig. 3 display the FAR analysis of the proposed method with existing methods. The existing methods such as TDNN, MCGAN, TMFCC-DNN-LOA, HXLP-DNN and HXLP-DCNN are analyzed with the proposed DMFCC-LPC method. The acquired outcomes prove that the DMFCC-LPC model attains good results through utilizing various frame lengths of 5.5, 7, 8.5, 10 and 11.5, and values of about 0.22. 0.32, 0.3, 0.35 and 0.37, respectively. Table 2 and Fig. 4 represent the analysis of DER for the proposed method with existing methods. The existing methods such as TDNN, MCGAN, TMFCC-DNN-LOA, HXLP-DNN and HXLP-DCNN are

Table 3. Performance analysis of speakers count with DER

Methods	Speakers Count				
	1	2	3	4	5
TDNN	4.2	3.8	3.4	3.2	2.9
MCGAN	3.2	3.0	2.9	2.7	2.6
TMFCC-DNN-LOA	3.1	3.0	2.8	2.6	2.3
HXLP-DNN	2.8	2.6	2.3	2.1	2.0
HXLP-DCNN	2.0	1.8	1.6	1.4	1.3
Proposed DMFCC-LPC	1.9	1.4	1.2	1.0	0.8

Table 4. Performance analysis of classifiers with FAR

Methods	True Positive	False Positive	False Negative
TDNN	0.86	0.76	0.55
MCGAN	0.88	0.68	0.37
TMFCC-DNN-LOA	0.77	0.85	0.89
HXLP-DNN	0.67	0.48	0.69
HXLP-DCNN	0.89	0.29	0.26
Proposed DMFCC-LPC	0.90	0.24	0.35

analysed with the proposed DMFCC-LPC method. The acquired outcomes prove that the DMFCC-LPC model attains better results through utilizing various frame lengths of 6, 8, 10, 12 and 14 and values of about 0.39. 0.37, 0.33, 0.30 and 0.28, respectively.

Table 3 and Fig. 5 represent the performance analysis of speaker counts with DER for proposed method with existing methods. The existing methods such as TDNN, MCGAN, TMFCC-DNN-LOA, HXLP-DNN and HXLP-DCNN are analysed with the presented DMFCC-LPC method. The acquired results display that the DMFCC-LPC model attains superior results through utilizing speakers count: 1, 2, 3, 4 and 5 with values of about 1.9. 1.4, 1.2, 1.0 and 0.9, respectively.

Table 4 and Fig. 6 display the performance analysis of classifiers with FAR for the proposed method with the existing methods. The existing methods such as TDNN, MCGAN, TMFCC-DNN-LOA, HXLP-DNN and HXLP-DCNN are analysed with the introduced DMFCC-LPC. The acquired results show that the DMFCC-LPC model attains good results and achieves True Positive, False Positive and False Negative values of about 0.90, 0.24, 0.35, respectively.

4.2 Comparative analysis

Table 5 shows the comparative analysis of proposed DMFCC-LPC using Accuracy, DER, FPR, FNR and FDR on various standard datasets. The proposed DMFCC-LPC method achieves the preferable outcomes when compared to other existing methods.

4.3 Discussion

The obtained results of this research are presented in Tables 1-4, and the graphical representations of these results are illustrated in Figs. 3-6. Table 5 denotes the comparative results of proposed with existing methods based on the performance metrics of Accuracy, DER, FPR, FNR and FDR on various standard datasets.

Table 5. Comparison Results

Method	Dataset	Accuracy	DER	FPR	FNR	FDR
DIHARD-III	SDS-HXLP-DCNN-SOA [13]	N/A	0.3	N/A	N/A	N/A
	QDM-SSD [18]	N/A	18.56	N/A	N/A	N/A
	ANSD-MA-MSE [20]	N/A	11.12	N/A	N/A	N/A
	Proposed DMFCC-LPC	0.957	0.29	0.118	0.213	0.245
CALLHOME	SDS-HXLP-DCNN-SOA [13]	N/A	0.33	N/A	N/A	N/A
	SSC-PIC [19]	N/A	7.0	N/A	N/A	N/A
	Proposed DMFCC-LPC	0.967	0.31	0.119	0.215	0.246
EenaduPrathidwani	ANN-ABC-LA [14]	0.764	N/A	0.175	0.354	0.351
	FrACWOA+DEC [21]	0.902	0.627	0.118	0.117	0.276
	Proposed DMFCC-LPC	0.953	0.612	0.117	0.116	0.254
ELSDSR	Feature-Level fusion [15]	N/A	4.31	N/A	N/A	N/A
	Proposed DMFCC-LPC	0.993	0.56	0.115	0.211	0.245

The SDS-HXLP-DCNN-SOA [13] obtained the DER of 0.3 and 0.33 on DIHARD-III and CALLHOME dataset. ANN-ABC-LA [14] had achieved the accuracy of 0.764, FPR of 0.175, FNR of 0.354 and FDR of 351 on EenaduPrathidwani dataset. The Feature-Level fusion [15] estimated only the DER and it achieved the 4.31 on ELSDSR dataset. The QDM-SSD [18] had achieved the DER of 18.56 on DIHARD dataset. SSC-PIC [19] had achieved the DER of 7.0 on CALLHOME dataset. ANSD-MA-MSE [20] had achieved the DER of 11.12 on DIHARD-III dataset and FrACWOA+DEC [21] achieved the accuracy of 0.902, DER of 0.627, FPR of 0.118, FNR of 0.117 and FDR of 0.276 on EenaduPrathidwani dataset. On the other hand, the proposed DMFCC-LPC achieves accuracy of 0.957, 0.967, 0.953 and 0.993 as well as DER of 0.29, 0.31, 0.627 and 0.56 on DIHARD-III, CALLHOME, EenaduPrathidwani and ELSDSR dataset.The introduced method achieves commendable results when compared to the existing methods. The proposed method utilizes various frame lengths and number of speakers to evaluate the performance of DER.

5. Conclusion

In this paper, a new speaker diarization approach based on the feature extraction using DFMCC and LPC is proposed. In this pre-processing step, a noise removal approach is utilized for removing an unwanted noise from the collected English dataset. The Voice Activity Detection (VAD) and speaker segmentation techniques are used after feature extraction for the detection of similar noises. The Deep Convolutional Neural Network (DCNN) is employed for the classification of speaker clustering. The outcomes evidence that the introduced DMFCC-LPC delivers commendable outcomes and it achieves accuracy of 0.957, 0.967, 0.953 and 0.993 as well as DER of 0.29, 0.31, 0.627 and 0.56 on DIHARD-III, CALLHOME, EenaduPrathidwani and ELSDSR dataset. In the future, the proposed method will be expanded to analyze the integration of various tasks, alongside traversing the different number of speakers and other complex frameworks.

Notation

Variables	Description
$x[n]$ and $y[n]$	Input and output speech signals
a	Filter constant with the characteristics $0.9 < a < 1.0$
$x1(n)$ and $y1(n)$	Input and output signals in framing and windowing
$w(n)$	Hamming window
N	Number of signal samples in each process
X_n	Outcome of FFT
X_k	Input signal while the value of $N = 0, 1, 2, \dots, N-1$
y_n	Input of IFFT
y_k	Outcome which is transformed into the time domain
r	Sample count
x	Noise-unaffected speech data
y	Noise-affected speech data
X_{total}	Integration outcome of noise-unaffected data x and the outcome of the data from y
N	Number of bandpass filter triangle
L	Total of coefficients outcome
p	Previous sample
$u(n)$	Gaussian noise
$P(\mu_k, \sigma_k, x)$	Gaussian with mean and covariance
q	Gaussian component
wt	Weight at time t
$g\ Lg(V\|S)$	Likelihood log measure
M_a	Frame size
ξ	Penalty weight
X	Perplexity model
S	Desire model
Ψ	Threshold value
O_n^i and O_m^{i-1}	Outcome of ith and $(i-1)$ layer
K_{mn}^i	Convolutional kernel of ith layer
B_n^i	Bias of ith convolution layer
X_n	Selection of input map
W_m^F	Original signal
W_m^g	Output signal
P_f	False discovery
t	Discovery count
FP	False Positive
TN	True Negative
FN	False Negative
TP	True Positive

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, have been done by 1st author. The supervision and project administration, have been done by 2nd author.

References

[1] M. India, J. Hernando, and J.A.R. Fonollosa, "Language modelling for speaker diarization in telephonic interviews", *Computer Speech & Language*, Vol. 78, p. 101441, 2023.

[2] A.B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, and T.K. Whangbo, "Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm", *Sensors*, Vol. 22, No. 21, p. 8122, 2022.

[3] A.I. Ahmed, J.P. Chiverton, D.L. Ndzi, and M.M. Al-Faris, "Channel and channel subband selection for speaker diarization", *Computer Speech & Language*, Vol. 75, p. 101367, 2022.

[4] G. Jaffino, R. Raman, and J.P. Jose, "Improved Speaker Identification System Based on MFCC and DMFCC Feature Extraction Technique", In: *Proc. of 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Erode, India, pp. 1-5, 2021, doi: 10.1109/ICECCT52121.2021.9616805

[5] J.M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation", In: *Proc. of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Cartagena, Colombia, IEEE, pp. 1139-1146, 2021, doi: 10.1109/ASRU51503.2021.9688044

[6] R. Ahmad, S. Zubair, and H. Alquhayz, "Speech enhancement for multimodal speaker diarization system", *IEEE Access*, Vol. 8, pp. 126671-126680, 2020.

[7] T.M. Al-Hadithy, and M. Frikha, "A Real-Time Speaker Diarization System Based On Convolutional Neural Networks Architectures", In: *Proc. of 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Istanbul, Turkiye, pp. 1-9, 2023, doi: 10.1109/HORA58378.2023.10156741

[8] Y. Takashima, Y. Fujita, S. Watanabe, S. Horiguchi, P. García, and K. Nagamatsu, "End-to-end speaker diarization conditioned on speech activity and overlap detection", In: *Proc. of 2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, IEEE, pp. 849-856, 2021, doi: 10.1109/SLT48900.2021.9383555.

[9] W. Wang, Q. Lin, D. Cai, and M. Li, "Similarity measurement of segment-level speaker embeddings in speaker diarization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 2645-2658, 2022.

[10] K.J. Naik, and A. Mishra, "Filter selection for speaker diarization using homomorphism: speaker diarization", In: *Artificial Neural Network Applications in Business and Engineering*, IGI Global, pp. 108-125, 2021, doi: 10.4018/978-1-7998-3238-6.ch005

[11] Y.X. Wang, J. Du, M. He, S.T. Niu, L. Sun, and C.H. Lee, "Scenario-Dependent Speaker Diarization for DIHARD-III Challenge", In: *Proc. of Interspeech*, Vol. 2021, pp. 3106-3110, 2021.

[12] A.A. Alnuaim, M. Zakariah, C. Shashidhar, W.A. Hatamleh, H. Tarazi, P.K. Shukla, and R. Ratna, "Speaker gender recognition based on deep neural networks and ResNet50", *Wireless Communications and Mobile Computing*, Vol. 2022, p. 4444388, 2022.

[13] C. Sailaja, S. Maloji, and K. Mannepalli, "A hybrid HXPLS-TMFCC parameterization and DCNN-SFO clustering based speaker diarization system", *Concurrency and Computation: Practice and Experience*, Vol. 34, No. 15, p. e6954, 2022.

[14] V. Sethuram, A. Prasad, and R.R. Rao, "Optimal trained artificial neural network for Telugu speaker diarization", *Evolutionary Intelligence*, Vol. 13, No. 4, pp. 631-648, 2020.

[15] N. Chauhan, T. Isshiki, and D. Li, "Text-Independent Speaker Recognition System Using Feature-Level Fusion for Audio Databases of Various Sizes", *SN Computer Science*, Vol. 4, No. 5, p. 531, 2023.

[16] S. Rajeyyagari, "Automatic speaker diarization using deep LSTM in audio lecturing of e-Khool Platform", *Journal of Networking and Communication Systems*, Vol. 3, No. 4, pp. 17-25, 2020.

[17] V. Khoma, Y. Khoma, V. Brydinskyi, and A. Konovalov, "Development of Supervised Speaker Diarization System Based on the PyAnnote Audio Processing Library", *Sensors*, Vol. 23, p. 2082, 2023.

[18] S.T. Niu, J. Du, L. Sun, Y. Hu, and C.H. Lee, "QDM-SSD: Quality-Aware Dynamic Masking for Separation-Based Speaker Diarization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 1037-1049, 2023.

[19] P. Singh, and S. Ganapathy, "Self-supervised representation learning with path integral clustering for speaker diarization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 1639-1649, 2021.

[20] M.K. He, J. Du, Q.F. Liu, and C.H. Lee, "ANSD-MA-MSE: Adaptive Neural Speaker Diarization Using Memory-Aware Multi-speaker Embedding", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 1561-1573, 2023.

[21] K. VijayKumar, and R.R. Rao, "Optimized speaker changes detection approach for speaker segmentation towards speaker diarization based on deep learning", *Data & Knowledge Engineering*, Vol. 144, p. 102121, 2023.

[22] DIHARD-III dataset link: https://dihardchallenge.github.io/dihard1/data.html.

[23] CALLHOME dataset link: https://catalog.ldc.upenn.edu/LDC97S42.

[24] ELSDSR dataset link: https://www.kaggle.com/datasets/kongaevans/speaker-recognition-dataset.

[25] K. Park, M. Chae, and J.H. Cho, "Image Pre-Processing Method of Machine Learning for Edge Detection with Image Signal Processor Enhancement", *Micromachines*, Vol. 12, p. 73, 2021.

[26] R. Hidayat, and A. Winursito, "A Modified MFCC for Improved Wavelet-Based Denoising on Robust Speech Recognition", *International Journal of Intelligent Engineering & Systems*, Vol. 14, No. 1, pp. 12-21, 2021, doi: 10.22266/ijies2021.0228.02.

[27] Y.G. Thimmaraja, B.G. Nagaraja, and H.S. Jayanna, "Speech enhancement and encoding by combining SS-VAD and LPC", *International Journal of Speech Technology*, Vol. 24, No. 1, pp. 165-172, 2021.

[28] H. Dinkel, S. Wang, X. Xu, M. Wu, and K. Yu, "Voice activity detection in the wild: A data-driven approach using teacher-student training", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 1542-1555, 2021.

[29] V.K. Kangala, and R.R. Ramisetty, "A Fractional Ebola Optimization Search Algorithm Approach for Enhanced Speaker Diarization", *Ingénierie des Systèmes d'Information*, Vol. 28, No. 4, 2023.