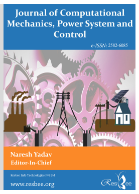
Vol.4 No.4 Oct 2021

Journal of Computational Mechanics, Power System and Control

Received 20 August, Revised 13 September, Accepted 17 October


Journal of Computational Mechanics, Power System and Control
Naresh Yadav
Editor-In-Chief
www.resbee.org

Hybrid Moth Flame Optimization and Teachers Learning based Optimization Algorithm based Artificial Neural Network Model for Speaker Diarization

Heyan Zhang

Xidian University, Xian, China

Abstract: Speaker diarization or indexing is the procedure of automatically dividing conversation including numerous speakers into homogeneous segments and clustering together all segments, which symbolize a similar speaker. Till now, numerous researches are carried out in this field, but the requirement of precise partitioning procedure gates lagged in particular conditions. However, this work aspires to develop a novel speaker diarization or indexing technique (Telugu language) which at first includes the Mel Frequency Cepstral coefficient (MFCC) based feature extraction. Then, for the process of clustering, a novel optimized Artificial Neural Network (ANN) is developed. The main aim of the process of clustering is to train the ANN and it is performed using the optimization algorithm which updates the ANN weight by exploiting a hybrid Moth Flame Optimization (MFO) and Teaching learning-based Optimization Algorithm (TLBO). Thus, the adopted technique is called Hybrid MFO-TLBO and the overall performance is evaluated with the conventional models regarding diverse performance metrics.

Keywords: Mfcc, Neural Network, Optimization Model, Speaker Diarization, Speakers.

Nomenclature

Abbreviations	Descriptions
GMM	Gaussian mixture models
MFO	Moth Flame Optimization
HMI	Human Machine Interaction
VBEM	Variational Bayes Expectation Maximization
DFT	Discrete Fourier Transform
Whale-IpCA	Whale-Imperialist Optimization algorithm
MFCC	Mel Frequency Cepstral coefficient
CSPHMM2s	Circular Suprasegmental Hidden Markov Models
TLBO	Teaching learning based Optimization Algorithm
HDP-HMM	Hierarchical Dirichlet Process Hidden Markov Modeling
FASR	False Alarm Speech Rate
ANN	Artificial Neural Network
MSR	Missed Speech Rate
UBM	Universal Background Model
Multi-SVNN	Multiple Support Vector Neural Network
SAD	Speech Activity Detection
AI	Artificial Intelligence

1. Introduction

The advancement of voice interactive systems is led by the development of HMI technologies. Numerous up-to-date devices like computers/laptops, smart televisions, smart phones home assistants, and so on possess these interactive technologies. Aforesaid systems can comprehend the human speech data and produce human-like speech. Most importantly two units are used as the interaction system to carry out this task such as the speech synthesis unit as well as the speech recognition unit. Current researchers have shown interest in emotion recognition from the speech signal as it creates a manner to build the artificial intelligence system. Numerous fields namely artificial intelligence as well as pattern recognition

https://doi.org/10.46253/jcmps.v4i4.a5

have exploited the recognition of emotion models to build the HMI. In the speech signal, the emotions have deviated based on the speaker's style. In addition, the attendance of higher duration speech is equivalent to the single emotion reducing the other emotions. Therefore, in the speech signal, it is vital to examine the emotions by deviating the frame length [1].

In the speech signals, a few techniques concentrate which involve the speech quotes used by the speakers. Specifically, the spoken signals are single-channel inputs that comprise numerous audio sources. Moreover, these are attained from several noises, speakers, music, etc as well as the information and format regarding the audio input is the application-specific regarding the sources. Additionally, the phrase speaker diarization is indicated as the audio segmentation, indexing, clustering so on following the researching people trend. Particularly, speaker diarization recognizes the speech signals like nonspeech signals, and input from speakers, and divides the audio into the same segments [2].

Not like English, numerous Indian languages are "Free-word-order" and also they are morphologically prosperous. Therefore, it has been recommended that "Free-word-order" is handled and enhanced by the dependency-based model than the constituency. There are numerous current efforts to build the dependency parses because of the dependency treebanks' availability. To build the current dependency parsers two CoNLL shared tasks were aimed to exploit diverse languages. Currently, in two ICON tools contests as well as rule-based, constraint-based, Hindi parsing shared task, statistical as well as hybrid techniques were searched towards building dependency parsers for three Indian languages such as Hindi, Telugu as well as Bangla. Using the renowned data-driven parsers, current accuracies are attained [3].

Both the speaker clustering as well as speaker segmentation is encompassed by the Taged speaker diarization. After the primary non-speech/speech recognition phase, the segmentation phase divides the incessant speech segments of the audio stream into homogenous segments with one active speaker [8]. Subsequently, the clustering phase gathers produced segments into clusters indicating the single speakers. In speaker diarization, numerous techniques are available like K-means clustering, HDP-HMM, spectral clustering, iterative mean shift clustering, VBEM-GMM, and so on [4].

The major objective of this work is to address a novel speaker diarization or indexing technique that is achieved using speaker clustering. In general, in speaker diarization, speaker clustering plays an important role. Therefore, by exploiting the optimized ANN, the speaker clustering process is optimally performed. In addition, the adopted ANN is trained optimally using the effectiveness of the meta-heuristic optimization approach called ANN-Hybrid MFO-TLBO. At last, the developed model performance is evaluated with the conventional techniques

2. Literature Review

In 2017, Brecht Desplanques et al [1], worked on the iVector-based diarization system that uses the adaptation on all levels as well as factor analysis. The diarization segments appropriate for the current research were: the speaker segmentation which explores speaker clustering as well as speaker boundaries that aspires at grouping speech segments of a similar speaker. By exploiting the eigenvoices, speaker segmentation lies on speaker factors that were extracted on a frame-by-frame base. Also, in this extraction process, they had integrated soft voice activity recognition as the speaker change recognition must be based on the speaker information merely to ignore the non-speech frames by using speech posteriors.

In 2018, B. Venkata Seshu Kumari and Ramisetty Rajeshwara Rao [2] explored diverse statistical dependency parsers to parse the Telugu. Here, five well-known dependencies parsers were considered such as "MaltParser, MSTParser, TurboParser, ZPar, and Easy-First Parser". Finally, the experimentation was performed with the diverse parser as well as the feature settings and exhibits the impact of the diverse settings. Additionally, a complete analysis was provided for all the parser's performances on the main dependency labels.

In 2018, Kasiprasad Mannepalli et al [3], developed a new emotion recognition model, named Whale-IpCA on the basis of the Multi-SVNN classifier, to identify the emotions in the speech signal. The recently adopted Whale-IpCA approach integrates WOA as well as IpCA, and it trains the Multi-SVNN classifier to identify emotions. In addition, from the input signal, the spectral feature set was extracted and presented to the adopted WhaleIpCA-based Multi-SVNN for detection purposes. Experimentation of the adopted model was performed with aid of the conventional emotion databases, like Telugu as well as Berlin.

In 2018, Varun Tiwari et al [4], worked on the speech recognition system. One of the main issues was communication including the short duration of speech utterances. The conventional GMM-based systems had attained suitable outcomes for recognition of speakers only while the speech lengths were adequately high. The present technique uses ivector-based technique by exploiting a GMM-UBM. It develops an i-vector speaker model from a speaker's enrollment data as well as exploits it to distinguish any new test speech. Here, a multi-model i-vector system for short speech lengths was presented.

In 2014, Ismail Shahin [5], worked on the recognition of speakers and it carry out the approximately perfect in neutral talking environments; though, these systems carry out badly in emotional talking environments. This work was proposed to enhance minimum text-independent as well as identification of emotion-dependent speaker performance in emotional talking environments on the basis of using the Second-Order CSPHMM2s as classifiers. Here, experimentation was performed on the speech database that was constituted of 50 speakers talking in six diverse emotional conditions.

3. Proposed Speaker Diarization Approach

Fig 1 illustrates the block diagram of adopted speaker diarization model. Here, the speaker diarization process is stated they are feature extraction, Speaker Segmentation, SAD, as well as the process of Speech Clustering. At first, the input audio signal is subjected to the system. The aforesaid are performed to obtain the diarization outcome. In this work, to set up a new speaker diarization or indexing technique by exploiting the Telugu language through optimized trained ANN for the process of clustering wherein the training is performed by exploiting a hybrid algorithm (hybrid Hybrid MFO-TLBO).

For the adopted speaker diarization model, let S represents the input audio signal with multiple speakers, that is $S = \{S_1, S_2, \ldots S_M\}$, wherein M represents the number of speakers. MFCC features get extracted that is "$F_x(N) = \{F_x(1), F_x(2), \ldots\}$ from the given input signal S, wherein $N = 1,2,\ldots Z$".Then, recognition of speech activity comes in 2 means namely removal of silence as well as the removal of music. At last, by exploiting the optimally trained ANN, the clustering process is obtained into n clusters on basis of speaker identity that is indicated as $I = I_1, I_2, \ldots I_c$ wherein $1 \le i \le N$.

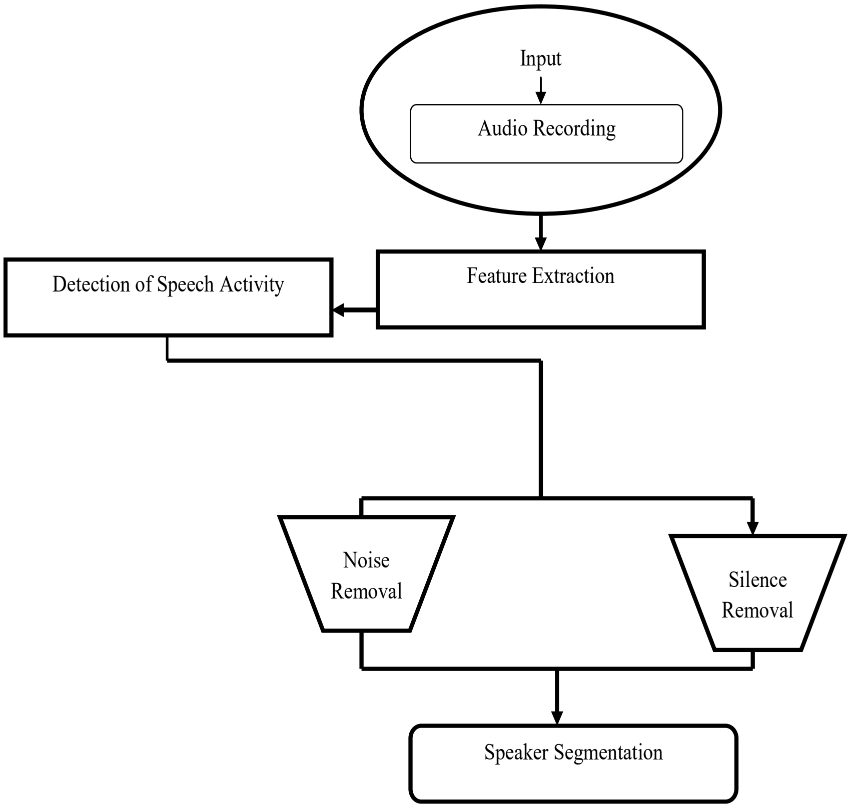


Fig. 1. *Block diagram of adopted Speaker Diarization approach*

3.1 Feature Extraction

In the speaker diarization technique, the MFCC features are extracted. Moreover, the audio signal is exploited for speaker indexing which is transferred as a frame signal with the reference of the frame i. Let $y(u)$ represents framed as well as input signal $y_i(u)$ in that i indicates the frame number as well as u indicates the sample number. As stated in eq. (1), the power spectrum is a computer in that $Y_i(l)$ represents the DFT for the framed signal as well as it is stated in Eq. (2). Let H_f and L_f the high as well as low frequency as well as it is transferred as Mel scale as represented in Eq. (3). This formulation is reformulated by exploiting Eq. (4). As stated in eq. (5), aforesaid values are exploited to generate a filter bank, where $k = 1 to S$ represents the filter number and $\mathrm{e}(\ )$ indicates $\mathrm{k}+2$ Mel spaced frequencies.

$$\mathrm{p_i}(\mathrm{l}) = \frac{1}{\mathrm{V}}\left|\mathrm{Y_i}(\mathrm{l})\right|^2 \tag{1}$$

$$\mathrm{Y_i}(\mathrm{l}) = \sum_{\mathrm{u}-1}^{\mathrm{V}} \mathrm{y_i}(\mathrm{u})\mathrm{h}(\mathrm{u})\mathrm{e}^{-\mathrm{j}2\Pi\,\mathrm{ln}};\ 1 \le \mathrm{l} \le \mathrm{Z} \tag{2}$$

$$\mathrm{S}(\mathrm{F}) = 1125 + \mathrm{zu}\left(1 + \frac{\mathrm{F}}{700}\right) \tag{3}$$

$$e(i) = (nfft) * h(i) \tag{4}$$

$$\mathrm{G_k}(\mathrm{l}) = \begin{cases} 0 & \mathrm{l} < \mathrm{d}(\mathrm{k}-1) \\ \dfrac{\mathrm{z}-\mathrm{d}(\mathrm{k}-1)}{\mathrm{d}(\mathrm{k})-\mathrm{d}(\mathrm{k}-1)} & \mathrm{d}(\mathrm{k}-1) \le \mathrm{l} \le \mathrm{d}(\mathrm{k}) \\ \dfrac{\mathrm{d}(\mathrm{k}-1)-1}{\mathrm{d}(\mathrm{k}+1)-\mathrm{d}(\mathrm{k})} & \mathrm{d}(\mathrm{k}) \le \mathrm{l} \le \mathrm{d}(\mathrm{k}+1) \\ 0 & \mathrm{l} > \mathrm{d}(\mathrm{k}-1) \end{cases} \tag{5}$$

The extraction of MFCC features is performed; furthermore fed to the speaker activity recognition procedure.

3.2. Speech Activity Detection

In order to partition the speech from the non-speech signal, a technique is developed in an audio recording process. Here, two most important confronts are involved namely least missed speech and the least false alarm speech. The speech signal misclassification percentage is represented as the non-speech signals through the SAD named as MSR when the percentage of the non-speech signal is recognized as speech signals that are indicated as FASR. For SAD, these two rates are represented as the analytical metrics. The technique-based classifier is exploited in the developed SAD subsystem. Moreover, from training speech and non-speech data, it is free. Moreover, the SAD subsystem is developed through two decoupled phases. At first, the silence attending in the whole recording is eradicated using the energy-based bootstrapping repeated with recurring classification. Then, non-speech signals are identified from the recording, music, and additional perceptible. In this, the process of music eradication takes silence eradicated audio against "speech bootstrap discriminator" for music. To train the music approach, a music audio signal which possesses higher confidence is used and is continuously refined.

Silence Removal: By exploiting 19 MFCC features integrated to STE and its 1st as well as 2nd order derivatives it is developed. For each frame, a confidence value is allocated using the segmentation of bootstrap to both the silences and speech classes. Additionally, to train the bootstrap silence technique, Gaussian mixtures within the size of 60-dimensional feature spaces are exploited. Likewise, speech technique is trained to utilize equivalent size. Here, all frames are categorized as two classes namely silence as well as speech in this iterative step. In reality, to train the silence and speech signals for consecutive iterations, the silence with higher confidence and speech frame is used. To model the silence and speech GMM is augmented with the augmentation in the iteration number, the number of 60 dimensional Gaussian used. When the speech GMM size is 32 and the nonspeech is 16, the optimal results were obtained. At present, the silences and the pauses are eradicated yet the jingles as well as music; the audible non speech is present.

Music Removal: In addition, audible nonspeech represents the utmost energy nonspeech, which is recognized as speech signals because of MFCCs and music frame energy is the same as speech signals.

By exploiting the ZCR as well as the STE of the windowed signal, the model fitting basic music versus speech classifier was developed. Concurrently, when the music and the speech signals are present the music speech discriminator suffers. This tends to the missed speech and also important speech data will also get lost. Here, the bootstrap segmentation is used as the classifier outcome to overcome the aforesaid disadvantages. To train the initial estimate technique, maximum confidence frames, as well as speech classes, are used. As same as the iterative classification, the silence eradication process is performed in this phase from music segments the speech is refined.

3.3 Segmentation of Speaker

In this system, the speaker segmentation approach is exploited; w represents a raising window size by means of Bayesian Inference criteria ΔBIC distance. At first, for a single speaker, an examination is performed and it changes from preliminary audio as well as for every encounter alters; an examination is restarted on the subsequent frame. In addition, the search window is announced and ΔBIC distance is calculated for every frame positioned with the window. While maxima go beyond a threshold value ψ , a maxima note is a alter. While no maxima are identified "on the window, window size" is increased as well as this procedure is continued until an alter meets. By exploiting the SAD, after the removal of the non-speech frame, speech signals are processed as well as after change points detection from audio signals its corresponding locations from original audio are recognized as well as mentioned as a changed point. Formerly, by exploiting two phases the segmentation process is developed. At first, the ΔBIC based change detection is performed based on the aforesaid threshold value. Then, the integration of consecutive segments happens only for those possessing positive ΔBIC scores. Aforesaid two phases are developed due to the over-segmentation issue in"0" threshold ΔBIC based segmentation. The maxima value which goes beyond the threshold ψ is represented for additional processes to eradicate these two-phase processes, as well as efficiently reduce the over-segmentation as represented in Eq. (6).

$$\Delta \mathrm{BIC}(y_i) = N \log|E| - N_1 \log|E_1| - N_2 \log|E_2| - \frac{\gamma}{2}\left(b + \frac{1}{2} b(b+1)\right) \log N \tag{6}$$

4. Adopted Clustering Process

4.1 Clustering Process using Adopted Optimized ANN

The process of clustering includes gathering as well as integrating the segments of the same speakers. Therefore, ANN [6] is developed to cluster segments in that training is performed using the adopted technique. In addition, it is a renowned approach used in numerous applications because of its flexibility when compared with various other classifiers. In this work, chosen features are used within the ANN for important feature classification. To calculate the AI issues, the ANN is modeled as a biological neuron network. Generally, in ANN the weights are represented as associations with the neurons in that the positive weight is indicated as the excitatory association as well as negative one is the inhibitory association. In addition, each input is transferred as weights as well as each one is augmented. Ultimately, the output amplitude is regulated during an activation function. Basically, ANN consists of 3 layers such as a hidden layer, an input layer, as well an output layer exploited to train the outcome. in_i represents the input neuron and h_i represents the output neuron. As mentioned in eq. (7), the hidden layer output H is represented, in that, af_1 denotes the nonlinear activation function, wi_{bo_i} indicates bias weight to o_i^{th} output neuron F denotes input to ANN, and wi_{bh_i} signifies bias weight to h_i^{th} hidden layer. Moreover, in eq. (8), the network output is mentioned, in that $wi_{in_i h_i}$ signifies weight from in_i^{th} input neuron to h_i^{th} hidden neuron Ei_2^* signifies outcome of o_i^{th} output neuron, af_2 denotes an activation function, and $wi_{h_i o_i}$ signifies the weight to o_i^{th} output neuron from h_i^{th} hidden neuron. At last, the whole network outcome is exhibited in eq. (9). In addition, eq. (9) states the error between the actual as well as predicted values Ei_2^* in that $\hat{s}i$ referred to as derived output as well as si referred as the actual output.

$$H = af_1\left[wi_{bh_i} + \sum_{in_i=1}^{N_{in_i}} \left(F \times wi_{in_i h_i} \right) \right] \tag{7}$$

$$si = af_2\left[wi_{bo_i} + \sum_{h_i=1}^{N_{h_i}} \left(H \times wi_{h_i o_i} \right) \right] \tag{8}$$

$$Ei_2^* = \underset{\{wi_{bh_i}, wi_{bo_i}, wi_{in_i h_i}, wi_{h_i o_i}\}}{\arg\min} \sum_{o_i=1}^{N_{o_i}} \left| si - \hat{s}i \right| \tag{9}$$

The major objective of this paper is the training process i.e., in the offline process wherein the weights $wi_{bh_i}, wi_{bo_i}, wi_{in_i h_i}, wi_{h_i o_i}$ jointly called W are updated using the optimization ideas. Specifically, a hybrid based updating process occurs with the initiation of specific principle.

4.2 Proposed Model

At present, numerous studies are presenting various hybrid optimization techniques to develop the exploration as well as exploitation abilities of the present approaches. In this paper, two approaches are integrated such as TLBO and MFO approaches by exploiting the co-evolutionary low level. In the MFO approach, the effectiveness of the exploration is enhanced and the efficiency of the TLBO exploitation is enhanced to exhibit both the techniques of the power point [7].

The position of the moth is updated is using as follows:

$$Y_i = S(M_i, F_j) \tag{10}$$

$$S(M_i, F_j) = D_i \cdot e^{bt} \cdot \cos(2\pi t) + F_j \tag{11}$$

$$D_i = | F_j - M_i | \tag{12}$$

X_i represent the i^th^ Moth, S signifies spiral function, b signifies logarithmic spiral shape, D_i signifies distance of i^th^ Moth for j^th^ Moth flame.

In order to cover the widespread extents in the imprecise search spaces using the logarithmic issues, the MFO is used for the exploration phase. For the duration of the computation, the anonymous phrase search space over the iteration sequence from primary to great repetition bounds as jointly of the deviations are randomization techniques.

$$Y_{new,i} = Y_{old,i} + r.(Y_{teacher} - (T_F \cdot Y_{mean})) \tag{13}$$

T_F indicates teacher factor, $Y_{teacher}$ indicates the optimal individual in the population, Y_{mean} indicates the current mean or average value of the individual.

The exploration phase represents the variation ability to run via the large numbers of probable outcomes. The teacher position which is accountable for the verdict of the globally optimal outcome of the issue is substituted with the moth position which is equivalent to the teacher position however exceedingly efficient to the change an outcome in the manner to an optimum one MFO approach reduces the computational time as well as the points in the accurate direction to the learners in the optimum value direction. The TLBO is a swarm intelligence optimization model which is represented as an effectual resolution approach for various complications of optimizations. Therefore, an integration of aforesaid optimal individual exploration using MFO and exploitation using TLBO assures the attainment of the optimal possible global optimum outcome standard.

5. Result and Discussion

In this section, experimentation analysis of developed speaker diarization or indexing was demonstrated. Here, the "input audio data was divided into 5 test cases like Test cases 1, 2, 3, 4, and 5". Moreover, the proposed method was evaluated with the existing models. Here, the proposed method was evaluated for the positive as well as negative metrics. Here, the proposed model was compared with the LM-NN model.

Fig 2 explains the accurateness performance evaluation of the adopted speaker diarization approach with the conventional techniques for 5 test cases. Therefore, it is evident that the adopted technique obtained enhanced outcomes and exhibited its effectiveness from the performance analysis.

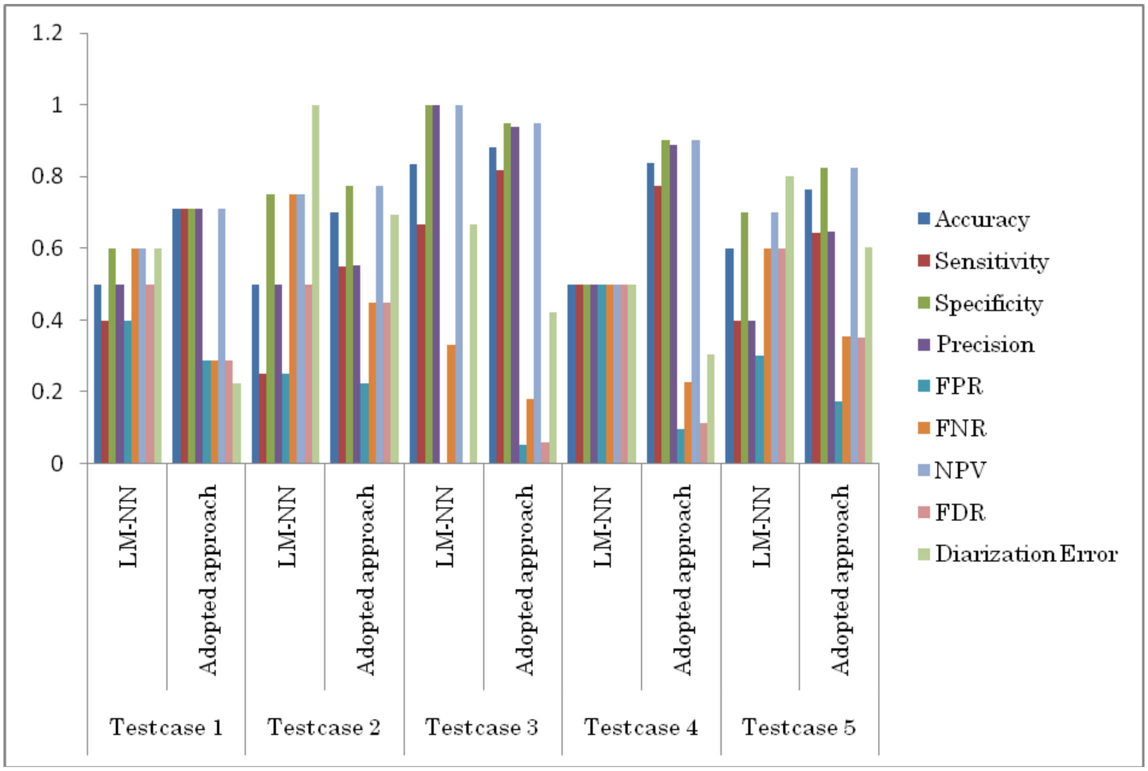


Fig. 2. *Performance evaluation of adopted as well as existing techniques for 5 test cases*

7. Conclusion

A new speaker diarization or indexing technique was developed in this work using an optimized ANN model. Hence, the adopted speaker diarization technique attained the stages such as speaker segmentation as well as clustering for that the process of clustering was developed through the adopted optimized ANN and the training was performed using a novel optimization approach by selecting the optimal weight by exploiting a hybrid approach called adopted ANN-ABC-LA. Finally, the performance evaluation was performed between developed as well as existing techniques to check the efficiency of the developed technique. Here, the overall analysis exhibits the minimum errors were attained by the proposed model than the conventional models.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Brecht DesplanquesKris DemuynckJean-Pierre Martens,"Adaptive speaker diarization of broadcast news based on factor analysis", Computer Speech & Language, 17 May 2017.

[2] B. Venkata Seshu KumariRamisetty Rajeshwara Rao,"Telugu dependency parsing using different statistical parsers", Journal of King Saud University - Computer and Information Sciences, January 2017.

[3] Varun TiwariMohammad Farukh HashmiN. C. Shivaprakash,"Speaker identification using multi-modal i-vector approach for varying length speech in voice interactive systems", Cognitive Systems Research, 31 October 2018.

[4] Ismail Shahin,"Speaker identification in emotional talking environments based on CSPHMM2s", Engineering Applications of Artificial Intelligence, August 2013.

[5] Kasiprasad MannepalliPanyam Narahari SastryMaloji Suman,"Emotion recognition in speech signals using optimization based multi-SVNN classifier", Journal of King Saud University - Computer and Information SciencesAvailable online, 23 November 2018.

[6] Y. Mohan, S. S. Chee, D. K. P. Xin and L. P. Foong, "Artificial neural network for classification of depressive and normal in EEG," 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), Kuala Lumpur, pp. 286-290, 2016.

[7] Reddy, K.N., Bojja, P. A new hybrid optimization method combining moth–flame optimization and teaching–learning-based optimization algorithms for visual tracking, Soft Comput 24, pp.18321–18347, 2020.

[8] Meghna Sangtani,"Improved Butterfly Optimization Algorithm: PI Controller for 7-Level Inverter", Journal of Computational Mechanics, Power System and Control, vol. 4, no. 3, July 2021.