Deep Learning Approaches for Speech Analysis: A Critical Insight

Alisha Goyal[1], Advikaa Kapil[2], Sparsh Sharma[3], Garima Jaiswal[1],
and Arun Sharma[1](✉)

[1] Indira Gandhi Delhi Technical University for Women, Delhi, India
arunsharma@igdtuw.ac.in
[2] Sanskriti School, Chanakyapuri, New Delhi, India
[3] Delhi Technological University, Delhi, India

Abstract. The main objective of speaker recognition is to identify the voice of an authenticated and authorized individual by extracting features from their voices. The number of published techniques for speaker recognition algorithms is text-dependent. On the other hand, text-independent speech recognition appears to be more advantageous since the user can freely interact with the system. Several scholars have suggested a variety of strategies for detecting speakers, although these systems were difficult and inaccurate. Relying on WOA and Bi-LSTM, this research suggested a text-independent speaker identification algorithm. In presence of various degradation and voice effects, the sample signals were obtained from a available dataset. Following that, MFCC features are extracted from these signals, but only the most important characteristics are chosen from the available features by utilizing WOA to build a single feature set. The Bi-LSTM network receives this feature set and uses it for training and testing. In the MATLAB simulation software, the proposed model's performance is assessed and compared to that of the standard model. Various dependent factors, like accuracy, sensitivity, specificity, precision, recall, and Fscore, were used to calculate the simulated outputs. The findings showed that the suggested model is more efficient and precise at recognizing speaker voices.

Keywords: Speaker recognition system · Artificial intelligence · Whale optimization algorithm · Recurrent neural network

1 Introduction

Speaker recognition is considered a biometric technology that uses typical features collected from a user's speech sample to validate their claimed identification [1]. Recently, speaker recognition along with the technologies like depreciation and speech recognition is becoming a key component in speech analysis of audio and video content. Automatic subtitling, automatic metadata, and generation Query-by-voice for television and movies are considered instances of practical applications of these techniques [2]. Aside from physical distinctions, each speaker does have a distinct way of speaking, which includes the usage of a unique word, intonation style, pronunciation pattern,

A. Dev et al. (Eds.): AIST 2021, CCIS 1546, pp. 73–84, 2022.
https://doi.org/10.1007/978-3-030-95711-7_7

accent, choice, rhythm, etc. In speaker recognition system, speaker identification and verification is considered as the most essential parts [3] (Fig. 1).

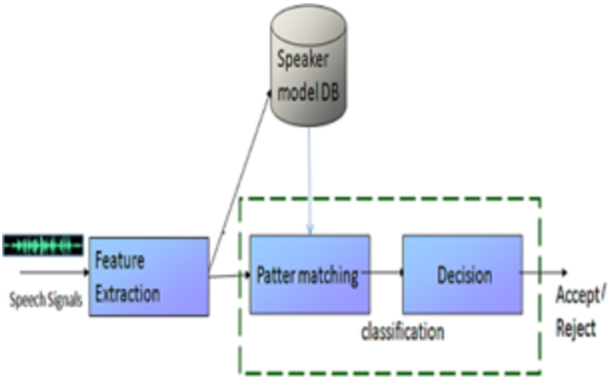

Fig. 1. Block diagram of standard speaker recognition system [8]

By utilizing module of feature extraction, speaker-specific elements of the speech signal are extracted. The speaker-specific elements of the speech signal are extracted using the feature extraction module. The characteristics are intended to offer the greatest possible separation amongst speakers in a group. Then after extracting features, the pattern recognition module matches the predicted characteristics of test speech samples to the system's speaker models [9]. This module matches the sample of a test speech to all of the stored models in a recognition task and returns a semantic similarity among the sample of test speech and all of the registered speaker models. The predicted features were calculated through an alternate model representing speakers besides the enrolled speakers in the task of the speaker authentication [10]. Furthermore, to make decisions, the decision module examines the semantic scores offered by the module of pattern matching. The decision module's outcome is determined by the mode and task type's for which the system is utilized. For instance, the decision module chooses the speaker model with the closest match to the test speech sample in closed-set mode [11]. A threshold value in the recognition task of open-set is needed to determine whether the match is good enough to recognize the speaker. Accuracy or the percentage of correct identifications is a performance criterion in an identification task. False Acceptance Rate (FAR) and False Rejection Rate (FRR) are the performance criteria in verification tasks. A high FAR gives the system unsafe to utilize, whereas a high FRR makes it inconvenient. In an ideal world, systems would be turned towards a low Total Error Rate (TER), which is equal to a low FAR + FRR. This state is obtained by altering the decision module's threshold value.

Forensics are considered a significant application of speaker recognition technology. In telephone conversations, a lot of data is transferred across two people, especially criminals, and there's been a growing interest in incorporating automatic speaker recognition to enhance semi-automatic and auditory analysis approaches [12–14]. The speaker recognition system can be modelled into two types Text-Dependent or Text-Independent system. Text-based recognition is used in situations where the user has a good command over the input [15]. Since the system is already familiar with the spoken content, this form of recognition provides improved system performance. Text-independent method is utilized to recognize any form of conversational speech or a phrase chosen by the user. The text-independent SRS does not have any previous knowledge of the text that the user has uttered. This is commonly utilized in applications when the user has less control over the input [16].

Recently, Artificial Intelligence (AI) has played a major role in speaker recognition, because AI provides various independent methods that utilize the unique characteristics to identify human voice. AI techniques like Support Vector Machine (SVM), Hidden Markov Modeling (HMM), Artificial Neural Networks (ANN), K-NEAREST NEIGHBOR (KNN), Convolutional Neural Network (CNN) etc. are useful in the In the areas of forensic voice verification, electronic voice eavesdropping, security and surveillance, mobile banking and shopping, etc. Other than this several researchers have proposed various techniques for speaker recognition that is described in the next section.

2 Related Work

For speaker identification, many researchers have developed a number of methods out of which some are discussed here: El-Moneim et al. [17] described the text-independent SRS in the presence of depletion factors like distortion and echoes. The suggested system by authors of this paper faced considerable difficulties in recognizing speakers in various recording situations. As a result, several speech improvement methods like wavelet denoising and spectrum subtraction were applied to increase recognition ability. **X. Zhao and Y. Wei** [18] investigated the SRS's performance as a function of input type and NN configuration, as well as the best feature parameters and NN architecture for an SRS. Additionally, traditional deep learning-based SR techniques like CNN and DNN, have been evaluated, and various enhanced deep learning-based models have been constructed and evaluated. The network model, when combined, has a greater recognition rate in speaker recognition than the standard network architecture.

Nammous et al. [19] integrated the prior research and implemented it to the issue of multi-speaker conversation speaker recognition. On Speakers in the Field, the authors tracked the performance and offered what they believe were the best described failure rates for this database. The suggested technique was also more resistant to domain shifts and produced results that were comparable to those acquired with a well-tuned threshold. Mobin et al. [20] developed a text-independent system relying on LSTM, with the goal of collecting spectral speaker-related data by working with standard speech attributes. For speaker verification, the LSTM framework was taught to build a discrimination space for verifying match and non-match pairings. When contrasted to other established approaches, the suggested design demonstrated its advantages in the text-independent domain. **S.** Shon et al. [21] presented an SRS is relying on CNNs for obtaining a reliable speaker embedding. With linear activation in the embedding layer, the embedding may be retrieved quickly. The frame level model permits authors to evaluate networks at the frame level, as well as perform additional analysis to enhance speaker recognition.

Jagiasi et al. [22],presented the development of an automatic speaker recognition system that incorporates classification and recognition of Sepedi home language speakers.

The performance of each model is evaluated in WEKA using 10-fold cross validation. MLP and RF yielded good accuracy surpassing the state-of-the-art with an accuracy of 97% and 99.9%, respectively; the RF model is then implemented in a graphical user interface for development testing. Mokgonyane et al. [23] in their work implemented a text-independent, language-independent speaker recognition system using dense & convolutional neural networks. The researcher of this paper explored a system that uses MFCC along with DNN and CNN as the model for building a speaker recognition system. Soufiane et al. [24] proposed a new technique to employ DNN (Deep neural network) in speaker recognition to understand the feature distribution. Experiments were carried out on the THUYG-20 SRE corpus, and excellent findings were obtained. Furthermore, in both noisy and clean situations, this innovative technique beated both i-vector/PLDA and their baseline approach. Mohammadi et al. [25] used the statistical properties of target training vectors to suggest a technique for balancing the framework and test. To analyze the system, the TIMIT database was employed. The experimental findings showed that using the suggested weighted vectors lowers the SVS's failure rate considerably. Duolin Huang et al. [26] presented a SRS technique depending on latent discriminative model learning. A reconstruction restriction was also added to develop a linear mapping matrix, making representation discriminative. Depending on the Apollo datasets utilized in the Fearless Steps Challenge at INTERSPEECH2019 and the TIMIT dataset, test findings showed that the presented approach surpassed common techniques.

After reviewing the literature it is concluded that the speaker recognition systems are playing important role in number of applications. Various researchers have proposed different approaches for recognizing the speaker from their voice. MFCC features are important feature model that are is recommended in most of the algorithms. Other than this there are few more features as frequency domain, time domain features that are also used in few of the studies. From the study it is concluded that the features processing is very crucial part of a recognition system, therefore selecting informative features can enhance the recognition rate. Very less studies are focus on feature selection models. Other than this artificial intelligence algorithms are the fundamental requirement of effective speaker recognition systems. Currently different machine learning algorithm such as SVM, ANN etc. are used for recognition applications, deep learning has reduced the system's complexity due to its high speed and better recognition capability. CNNs and RNNs are one of the examples of such models. But still the scope of modification and improvements are there, inspired from this the proposed scheme provides an improved speaker recognition system that is given in next section.

3 Proposed Work

To overcome the issues related to the traditional models, a novel and effective technique are proposed. The proposed work provides efficient algorithm to handle the complexity of the system and also gives a high recognition rate. Relying on WOA and Bi-LSTM, this research suggested a text-independent speaker identification algorithm. MFCC based features are extracted in the proposed model and using Whale optimization algorithm the informative feature are selected for all the speaker's audio samples. In addition to this, an upgraded variant of RNN that is BI-LSTM classification model is used for recognition.

The reason behind using the BI-LSTM network instead of conventional LSTM is the features of this model that are: It has two methods for dealing with inputs. It can not only process the inputs that have already passed through it, but it can also manage the inputs that have passed through it in the past. The simulation of the proposed model is done in MATLAB software by following the below written methodology.

Methodology

The proposed model's first step is to obtain input data from a dataset that is either available online or can be obtained in the real world. The dataset utilized in the suggested work was obtained from Kaggle.

Dataset Information

The audio sample of five famous political leaders can be found in the Kaggle dataset. The dataset comprises a collection of 500 audio samples, with 100 audio samples for each leader. Each audio file in the files is a PCM encoded one-second 16000 sample rate audio file. Figure 2 depicts a sample of audio signal collected from the dataset.

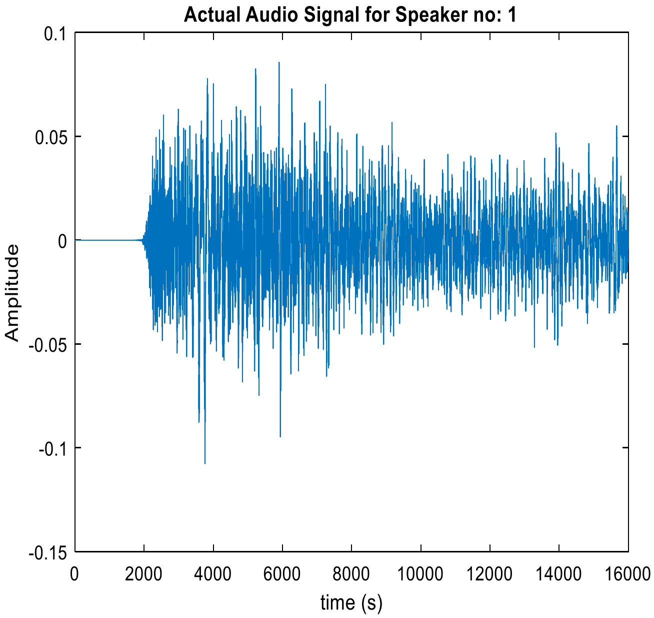

Fig. 2. Sample data taken from dataset

Following the collection of data, the signals must be processed and converted into a definite set of features. With 14 coefficients, the proposed model obtains MFCCs features. Figure 3 depicts each coefficient, which includes 100 characteristics.

Furthermore, in each audio signal sample, the average value of al 14 MFCC coefficients is determined and evaluated as represented in Fig. 4.

Moreover, other audio samples MFCC features are evaluated with their average MFCC coefficient of audio signals. At last, the WOA optimization method, as well as its

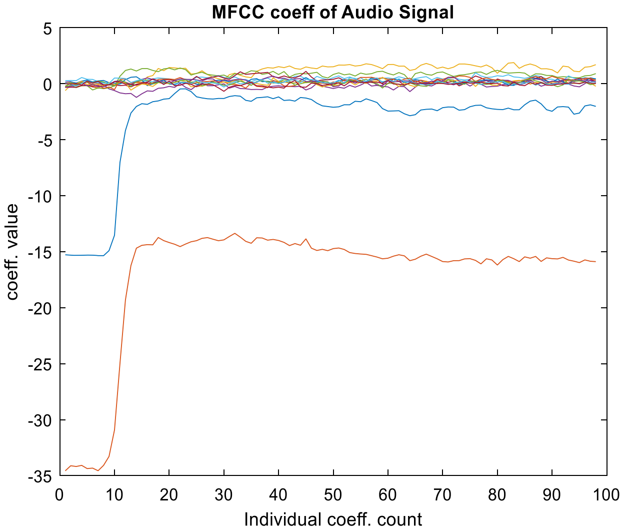

Fig. 3. MFCC features of 14 coefficients

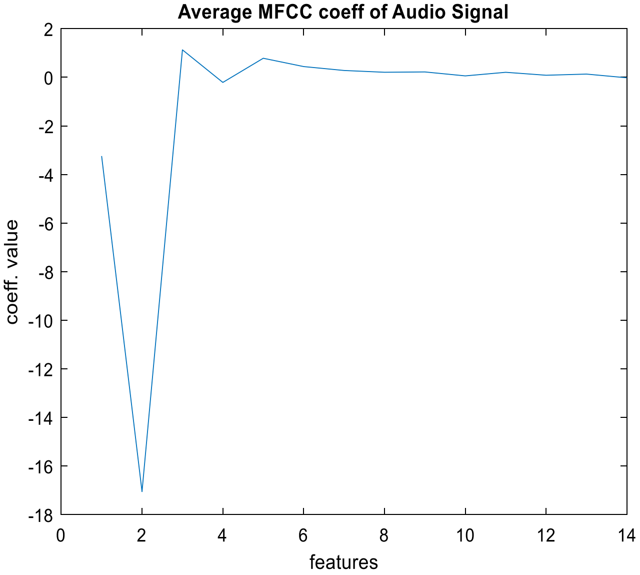

Fig. 4. Average MFCC coefficients of Audio signal

various parameters and their configurational values, must be initialized next to choose features from the available feature set. Table 1 shows the particular value of each and every parameter.

Then, the WOA algorithm only chooses the features that are essential for recognizing speaker from the available datasets. Only six of the 14 retrieved features are chosen as

Table 1. Configuration table for WOA feature selection model

WOA parameter	Value
Max Iteration	50
Population	10
Decision variable	6
Coefficient a	2 to 0
r	[0 1]
Population dimension	6

essential features with the highest fitness value. For each and every value, the optimal fitness value is evaluated by utilizing Eq. 1

$$Fitness = \frac{1}{mean(Std(SelectedFeatures))} \tag{1}$$

Following the creation of the final feature set, the proposed RNN-Bi-LSTM classification network is initialized by specifying different parameters like max epochs and threshold input size. Aside from that, the proposed network defines a number of other parameters, which are listed in Table 2.

Table 2. Configuration table for RNN BI-LSTM Classification model

Network's parameter	Value
Max epochs	150
Gradient threshold	1
Num of hidden Layer	120
Input size	1
No of layer	5
Min. batch size	10

For training and testing, the final feature set is send into the suggested RNN-BILSTM classification model. The model is trained by supplying trained data. At last, by supplying the testing data, the suggested model efficiency is tested. At last, the suggested model's efficiency is evaluated by supplying it with test data. The performance is assessed by using a number of performance parameters that are briefly described in the next section.

4 Results and Discussion

In the MATLAB simulation software, the proposed model's performance is simulated and examined. The simulation results were evaluated using a variety of performance metrics, including accuracy, sensitivity, specificity, precision, recall, and Fscore. Furthermore, the suggested model's performance is compared to that of the standard model, which is briefly addressed in this section.

- **Performance Evaluation**

The suggested model's efficiency is first assessed in terms of the convergence curve depicted in Fig. 5.

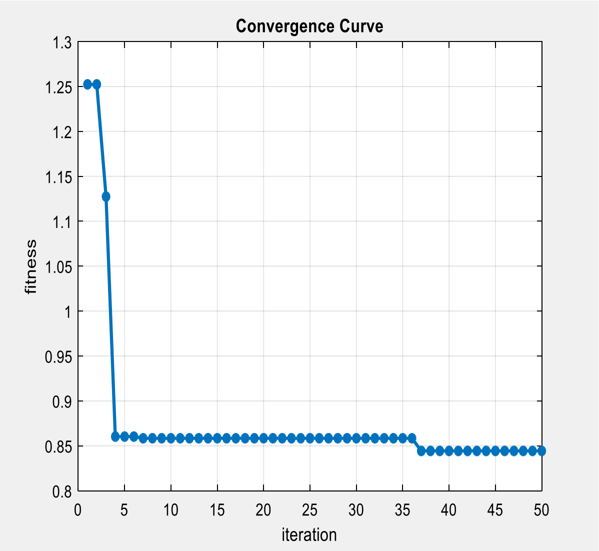

Fig. 5. Convergence curve of proposed model

The fitness graph obtained by the WOA algorithm during the feature selection process in the proposed model is shown in Fig. 5. The total number of iterations conducted is represented on the x-axis, while the fitness value is represented on the y-axis. The graph shows that the value of fitness is initially high, but as the number of iterations increases, the value of fitness drops dramatically, from 1.25 to 0.86 in just 4 iterations. After 50 iterations, the fitness value continues to decline and falls slightly below 0.85. This indicates that the suggested model is capable of extracting features efficiently and effectively.

The suggested model's performance is assessed using a variety of dependence variables such as accuracy, sensitivity, specificity, precision, recall, and Fscore. The suggested model's performance is first evaluated using several performance parameters, as illustrated in Fig. 6.

In Fig. 6 illustrates the several dependencies factors of the graphs that include the accuracy, sensitivity, specificity, precision, recall and Fscore of the proposed model. The blue, orange, and yellow colored bars represent the performance f-of accuracy, sensitivity, and specificity. The precision, recall, and Fscore performance are shown

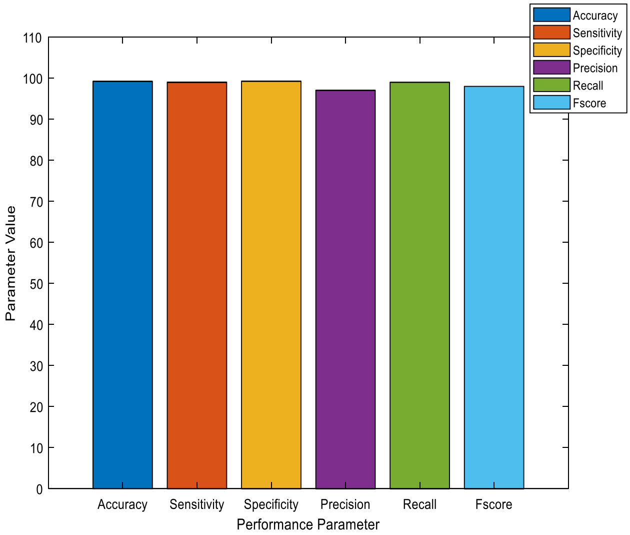

Fig. 6. Different performance parameters of proposed model

by purple, green, and sky-blue colored bars, respectively. The accuracy achieved in the suggested model is equivalent to 99% while the sensitivity and specificity achieved in the proposed model is equal to 98% and 99% respectively. Similarly, the suggested model examines the value of precision, recall, and Fscore, which come out to be 97%, 98%, and 98%, respectively. These results demonstrate that the suggested model is capable of reliably recognizing speakers. Table 3 shows the particular values of these parameters.

Table 3. Specific value of different parameters

Parameters	Values (% age)
Accuracy	99.2016
Sensitivity	98.9899
Specificity	99.2537
Precision	97.0297
Recall	98.9899
Fscore	98

In addition, the proposed model's accuracy is evaluated and compared to that of the standard model. Figure 7 represents the proposed model's comparison graph and proposed model in terms of accuracy.

Figure 7 illustrates the comparison of the suggested model and traditional MFCCs, Log-spectrum, spectrum technique, MFCCs-GMM-UBM models in terms of their accuracy value. Blue, orange, yellow, and purple colored bars represent the performance of the standard MFCCs, Log-spectrum, spectrum approach, and MFCCs-GMM-UBM models, respectively, while the green colored bar represents the suggested model's performance. According to the graph, the degree of accuracy reached in the standard MFCC and log-spectrum methods is 95% and 98%, respectively. The accuracy of traditional spectrum methods and MFCCs-GMM-UBM techniques was 98% and 94%, respectively. when the accuracy value for the proposed model was calculated, it came out to be 99%. This demonstrates that the suggested model is more precise and efficient at recognizing speaker signals. Table 4 shows the specific accuracy value in the traditional and proposed models.

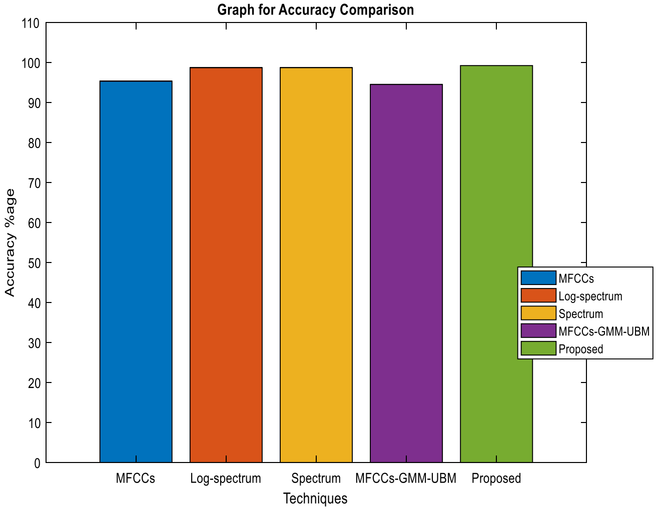

Fig. 7. Comparison graph of proposed and traditional model for accuracy

Table 4. Accuracy values in traditional and proposed model

Technique	Accuracy values (%age)
MFCCs	95.33
Log-spectrum	98.7
Spectrum	98.7
MFCCs-GMM-UBM	94.5
Proposed	99.2016

From graphs and tables, it is analyzed that the proposed model is more accurate and efficient in recognizing distinct speakers.

5 Conclusion

Nowadays, various researchers and experts have expressed their interest in automatic speech and speaker identification. Various scholars have developed a variety of methodologies, although the key problem for researchers is determine how precisely the system can identify and recognize speakers. The WOA and Bi-LSTM were used to suggest a model in this research. In terms of accuracy, sensitivity, specificity, precision, recall, and Fscore, the suggested model is compared to conventional MFCCs, Log-spectrum, spectrum method, and MFCCs-GMM-UBM models. The proposed model's sensitivity and specificity were calculated to be 98.9899% and 99.2537%, respectively. Furthermore, the precision, recall, and Fscore values were discovered to be 97.0297%, 98.9899%, and 98%, respectively. Furthermore, the traditional MFCCs and Log-spectrum accuracy values were 95.33% and 98.7%, respectively, but the traditional spectrum approach and MFCCs-GMM-UBM models accuracy values were 98.7% and 94.5%, respectively. The proposed model, on the other hand, achieved a value of accuracy of 99.2016%, demonstrating that it is more efficient and effective in recognizing distinct speakers.

References

1. Zilovic, M.S., Ramachandran, R.P., Mammone, R.J.: Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions. IEEE Trans. Speech Audio Process. **6**, 260–267 (1998)
2. Tranter, S., Reynolds, D.: An overview of automatic speaker diarization systems. IEEE Trans. Audio Speech Lang. Process. **14**, 1557–1565 (2006)
3. Alexander, A., Botti, F., Dessimoz, D., Drygajlo, A.: The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. Forensic Sci. Int. **146S**, 95–99 (2004)
4. Hansen, J., Hasan, T.: Speaker recognition by machines and humans: a tutorial review. Sign. Process. Mag. IEEE **32**, 74–99 (2015)
5. Jothilakshmi, S., Gudivada, V.N.: Large scale data enabled evolution of spoken language research and applications. Elsevier **35**, 301–340 (2016)
6. Kekre, H., Kulkarni, V.: Closed set and open set Speaker Identification using amplitude distribution of different transforms. In: 2013 International Conference on Advances in Technology and Engineering, pp. 1–8 (2013)
7. Mathu, S., et al.: Speaker recognition system and its forensic implications. Open Access Scientific Reports (2013)
8. Imdad, M.N., et al.: Speaker recognition in noisy environment. Int. J. Adv. Res. Comput. Sci. Electron. Eng. **1**, 52–57 (2012)
9. Imam, S.A., et al.: Review: speaker recognition using automated systems. AGU Int. J. Eng. Technol. **5**, 31–39 (2017)
10. Dhakal, P., Damacharla, P., Javaid, A.Y., Devabhaktuni, V.: A near real-time automatic speaker recognition architecture for voice-based user interface. Mach. Learn. Knowl. Extr. **1**, 504–520 (2019)
11. Varun, S., Bansal, P.K.: A review on speaker recognition approaches and challenges. Int. J. Eng. Res. Technol. (IJERT) **2**, 1581–1588 (2013)
12. Niemi-Laitinen, T., Saastamoinen, J., Kinnunen, T., Fränti, P.: Applying MFCC-based automatic speaker recognition to GSM and forensic data. In: Proceedings of the Second Baltic Conference on Human Language Technologies, pp. 317–322 (2005)

13. Pfister, B., Beutler, R.: Estimating the weight of evidence in forensic speaker verification. In: Proceedings of the 8th European Conference on Speech Communication and Technology, pp. 701–704 (2003)
14. Thiruvaran, T., Ambikairajah, E., Epps, J.: FM features for automatic forensic speaker recognition. In: Proceedings of the Interspeech 2008, pp. 1497–1500 (2008)
15. Hebert, M.: Text-dependent speaker recognition. Springer handbook of speech processing. Springer Verlag, pp. 743–762, 2008. https://doi.org/10.1007/978-3-540-49127-9_37
16. Nayana, P.K., Mathew, D., Thomas, A.: Comparison of text independent speaker identification systems using GMM and i-Vector methods. Procedia Comput. Sci. **115**, 47–54 (2017)
17. El-Moneim, S., Nassar, M., Dessouky, M.I., Ismail, N., El-Fishawy, A., Abd El-Samie, F.: Text-independent speaker recognition using LSTM-RNN and speech enhancement. Multimedia Tools Appl. (2020). https://doi.org/10.1007/s11042-019-08293-7
18. Zhao, X., Wei, Y.: Speaker recognition based on deep learning. In: 2019 IEEE International Conference on Real-time Computing and Robotics (RCAR), pp. 283–287 (2019)
19. Nammous, M.K., Saeed, K., Kobojek, P.: Using a small amount of text-independent speech data for a BiLSTM large-scale speaker identification approach. J. King Saud Univ.- Comput. Inf. Sci. (2020)
20. Mobin, A., Najarian, M.: Text-independent speaker verification using long short-term memory networks. arXiv:1805.00604 (2018)
21. Shon, S., Tang, H., Glass, J.: Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model. In: 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 1007–1013 (2018)
22. Jagiasi, R., Ghosalkar, S., Kulal, P., Bharambe, A.: CNN based speaker recognition in language and text-independent small scale system. In: 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 176–179 (2019)
23. Mokgonyane, T.B., Sefara, T.J., Modipa, T.I., Mogale, M.M., Manamela, M.J., Manamela, P.J.: Automatic speaker recognition system based on machine learning algorithms. In: 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA), pp. 141–146 (2019)
24. Hourri, S., Kharroubi, J.: A deep learning approach for speaker recognition. Int. J. Speech Technol. **23**(1), 123–131 (2019). https://doi.org/10.1007/s10772-019-09665-y
25. Mohammadi, M., Mohammadi, H.R.S.: Weighted I-vector based text-independent speaker verification system. In: 2019 27th Iranian Conference on Electrical Engineering (ICEE), pp. 1647–1653 (2019)
26. Huang, D., Mao, Q., Ma, Z., et al.: Latent discriminative representation learning for speaker recognition. Front Inform. Technol. Electron. Eng. **22**, 697–708 (2021)