Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 1, pp. 121-132
DOI : 10.47974/JIOS-1857

Review of speaker recognition : Concepts, challenges, architectures, and future directions

Neelam Nehra *
Geetanjali Sharma †
Department of Electronics & Communication Engineering
Maharaja Surajmal Institute of Technology
Janakpuri 110058
Delhi
India

Parveen Kumar §
Department of Information Technology
Maharaja Surajmal Institute of Technology
Janakpuri 110058
Delhi
India

Dinesh Sheoran ‡
Department of Electronics & Communication Engineering
Maharaja Surajmal Institute of Technology
Janakpuri 110058
Delhi
India

Amita Yadav @
Department of Computer Science and Engineering
Maharaja Surajmal Institute of Technology
Janakpuri 110058
Delhi
India

* *E-mail:* **neelamnehra@msit.in (Corresponding Author)**
† *E-mail:* **gsharma@msit.in**
§ *E-mail:* **parveen@msit.in**
‡ *E-mail:* **dineshsheoran@msit.in**
@ *E-mail:* **amita.yadav@msit.in**

Abstract

Speaker recognition refers to the task of recognizing a person from a spoken utterance with the aid of a machine. The distinctive features extracted from the speech characteristics of various individuals are utilized to create a unique speaker model for each person. These models are then employed to identify an enrolled speaker from a pool of available speakers by comparing the features of the test speaker with the stored models. This paper explores various feature extraction techniques, including MFCC, LPC, RASTA-PLP, ZCR, and PSMFCC, among others. The paper also aims to explore various techniques for speaker modeling like GMM, I-Vector, ANN, DNN, SVM with an emphasis on Deep Neural Network (DNN) techniques. Additionally, this paper also presents a detailed comparative performance analysis of the techniques used in the domain of speech processing.

Subject Classification: *68T07.*

Keywords: *Speaker recognition, Feature extraction, Classifier, Mel frequency cepstral coefficient, Deep neural network.*

1. Introduction

Speaker Recognition (SR) refers to identifying a speaker based on voice features in the input speech sample. Each person has their own distinct voice and the size of the vocal tract, the structure and specific movement of the articulatory organs, and variations in the speaking habits all contribute to the uniqueness in the voice features [1]. The most common biometric identification technique for authenticating and tracking humans is by using the speech signals [2]. Mel Frequency Cepstral Coefficients (MFCC), Relative Spectral –perceptual linear prediction (RASTA-PLP), Linear Prediction Cepstral Coefficients (LPCC), Linear Predictive coefficient (LPC), is features extraction techniques available in speaker recognition. The spectrograms of seven speakers' utterances were obtained by [3] and the frequency location of formants was recorded over a period of up to 29 years, as well as the pitch of voiced sounds, shift to lower frequencies. In [4] employed three speech coding methods were employed based on voice pitch and formant frequencies determined from time aligned, un-code. Employed Mel warping, this puts less focus on high frequencies. The log spectrum's spectrum can be used to represent the cepstrum. In [5], a methodology based on posteriori probability and likelihood ratio to reduce computation costs in order to identify normalization terms. [6] investigated the different feature extraction approaches, feature normalization techniques, and also feature compensation approaches those included and excluded noise compensation techniques. In [7] the comparison of supervised and

unsupervised classifiers for text-independent speaker recognition systems was discussed. Although numerous review papers have been published in the field of speaker recognition systems, most focus exclusively on either feature extraction techniques or classification methods [8]. This paper is structured as: Section 2 presents the various databases used in different fields of speech. Section 3 describes the pre-processing steps in a speech. Section 4 explains the different time domain and Frequency domain techniques. Section 5 explains the different classifiers

2. Database

The database selection is an important point in implementing any proposed approach by researchers. As the various database are available for speech processing and Language Identification Systems [9, 3, 4]. The table 1 below lists the many databases that are utilized in various applications.

Table 1

Summary of different data sets used in different domains

Reference	Application	Database used
[5]	Speaker Identification	Domain Adaptation challenge data DAC
[6]	Language identification	NIST 2011 Language Recognition (LRE)Data
[7]	Speaker Identification	IITG
[8]	Speaker Recognition	TIMIT
[9]	Speaker Identification	USM
[10]	Speaker Identification	ELDSR
[11]	Language identification	SOLID
[12]	Language identification	IIIT-H
[13]	Speech Recognition	IITG-Hing Cos
[14]	Speech Recognition	CSLU2002

3. Pre-Processing

Preprocessing is the most important stage in identifying a speaker. The speaker's voice is non-stationary and contains a number of features that may or may not be beneficial in identifying the speaker. Frame

segmentation, detection of active frames, and windowing methods are all part of the preprocessing as given in Fig.1.

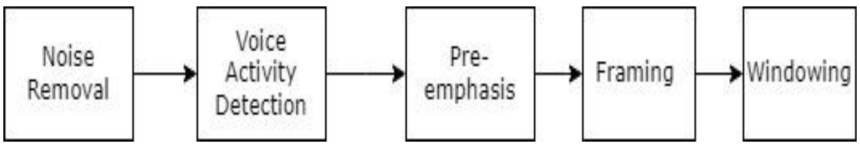

Figure 1

Pre-processing steps of a speech signal

1. **Noise Removal:** Signal-to-Noise Ratio (SNR) is given as follows and is generally expressed in decibels (dB) [15,16].
2. **Voice Activity Detection (VAD):** VADs (voice activity detectors) are devices that separate speech into voiced and unvoiced speech parts. VADs have to adjust to changes in noise characteristics while estimating noise characteristics during non-speech segments [17].
3. **Pre-emphasis:** A pre-emphasis filter is applied to flatten the voice spectrum. A high-pass FIR, or finite impulse response, filter having a single zero at the point of origin is the most often used pre-emphasis filter. To eliminate -6 dB slopes, pre-emphasis is used.
4. **Framing:** In order to handle audio signals segment-wise, continuous waveforms are divided into fixed-length segments. Speech, too, can be thought of as a quasi-stationary signal that only remains stationary for a limited time [18].
5. **Windowing** In the process in which window every individual frame in order to remove signal discontinuities at the start and finish of each frame. In windowing, the frame is multiplied with the window function point by point [4].

4. Feature Extraction Techniques

4.1 *Mel Frequency Cepstrum Coefficients (MFCC)*

Mel Frequency Cepstrum is a nonlinear frequency scale that represents the short-term power spectrum of sound, derived from the logarithmic power spectrum using a linear cosine transform. The MFCC [13] is made up of the coefficients that make up Mel frequency cepstral coefficients. Next step is framing in signal is segmented into N-sample frames, with M samples between each frame. Framing in followed by

windowing in which the signal discontinuity is minimized. Each frame of NNN samples is then converted from the time domain to the frequency domain by applying the Fast Fourier Transform. The frequency response magnitude of each frame is determined using FFT. By Discrete Cosine Transform (DCT) log Mel spectrum is converted to a time domain spectrum.

4.2 *Linear Predictive Coefficients (LPC)*

LPC has the drawback of not being consistent with human hearing because it conveys same detail to all frequencies, resulting in the presence of high frequency coefficients, which generally increase noise [19]. Because LPC provides similar detail to all frequencies, it is incompatible with human hearing, resulting in the existence of high frequency coefficients, which generate noise [20].

4.3 *Relative Spectral-perceptual linear prediction (RASTA-PLP)*

The primary goal of the RASTA-PLP method is to minimize additive and unwanted noise in speaker recognition while also enhancing the quality of the signal [21]. RASTA-PLP involves computing the critical power spectrum in the same way as PLP, then compressing the static nonlinear transform [22].

5. Speaker Modelling

The modeling technique is selected based on the type of speech, ease of training, computational and expected performance. Modeling techniques are classified as shown in Fig.2.

5.1 *Generative Models*

Generative models estimate the distribution of features within each speaker. They only need training data samples from the target speaker to create a statistical/non-statistical model that describes the feature distribution of the target speaker. It includes models like the Gaussian Mixture Model (GMM) [23], Hidden Markov Model (HMM) [24], and Vector Quantization (VQ) [25].

5.2 *Discriminative Models*

Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) models falls in this category. Flexible architectures and

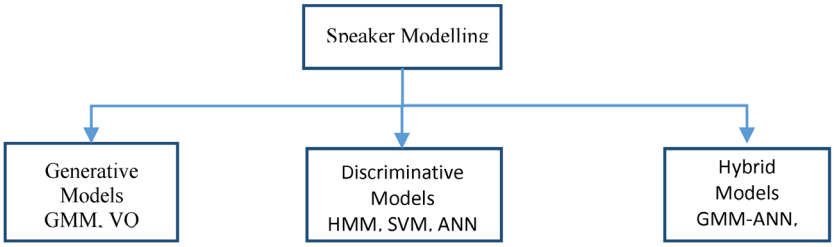

Figure 2

Classification Technique of Speaker Modelling

discriminating training power are two advantages of these models, SVM classifiers use nonlinear boundaries to separate complex regions [23, 24]. ANN can discriminate the features of different speakers because it is a non-linear classifier, but its performance is lower to that of GMM [26].

Artificial Neural Network (ANN): A data processing system called an artificial neural network (ANN) was developed with inspiration from how organic nerve systems, including the human brain, process information. It consists of the following basic steps:

- Neural network (NN) with multiple inputs are presented (each vector represents a pattern)
- Examine how closely the actual output produced for a given input matches to the desired output.

Deep Neural Network (DNN): DNN is a more advanced version of the Artificial Neural Network (ANN), having several hidden layers between the input and output layer [27]. The output layer converts the hidden layer activations to the scale you chose for your output. The neurons are structured into layers. The Back-Propagation algorithm is used to train ANNs with multiple layers [28, 29]. Due to the error propagation method and several hidden layers in between, ANN becomes very complex and time consuming. The different network architectures in DNN are given in Table 2.

5.3 *Hybrid Modelling*

A hybrid technique combines all of the above-mentioned models, such as GMM–HMM, ANN–HMM [29].

Table 2
Comparative Table of different type of DNN's

Ref.	Data Set	Input Features	DNN Type	Classifier	Score
[23]	US English Telephonic speech	20 dimension MFCC	4-layered, temporal pooling	UBM/i-vector	5.3
[24]	NIST SRE 12	60 dimension MFCC	5-layered	UBM/i-vector	1.58
[25]	NIST SRE 12, Noisy Narrowband	40 log Mel-filter bank coefficients	7-layered, fully connected	UBM/i-vector	1.39
[28]	Fisher, CSLT-CUDGT2014	Phone aware40 dimensional d-vectors	7-layered, time-delay	UBM/i-vector	8.37
[29]	NIST SRE 10	40 dimension MFCC	7-layered, multi spice time delay	GMM-UBM	7.2
[30]	SR 2015	39 dimension PLP	4-layered, fully connected	GMM-UBM,d-vector,j-vector	0.54
[31]	RSR 2015	20 dimension MFCC	4-layered, fully connected	UBM/i-vector GMM-DTW	0.2

Gaussian Mixture Model (GMM): GMM is among the successful techniques because it uses a Gaussian mixture probability density function. A Gaussian Mixture Model (GMM) is a weighted sum of MMM component densities, where MMM represents the number of Gaussian components. A GMM can be employed to model and represent each individual speaker. The speaker's identity is confirmed by selecting the model that yields the highest likelihood scores [28].

I-Vector: Two types of classifiers, CDS and PLDA, are commonly used in implementing I-Vector-based speaker recognition. In both methods, likelihood scores are computed in two different ways, and the speaker is identified by the speaker model with highest score. The simplified Gaussian PLDA approach is utilized here [22].

Table 3
A comparative table for various existing approaches

Reference	Domain	Feature Extraction Technique	Classifier	Accuracy
[19]	Speaker Identification	MFCC+LPC	ANN	91
[21]	Speaker Identification	RASTA PLP	GMM	79.7
[22]	Speaker Identification	RASTA PLP+Pitch+Formants	I-Vector	77
[25]	Speaker Identification	MFCC	RNN	96.9
[26]	Language Identification	MFCC	SVM	73.4
[31]	Language Identification	MFCC+SDC	BPNN	98.1
[32]	Speaker Identification	MFCC	VQ	78.94
[32]	Language Identification	MFCC+RASTA PLP	ANN	94.6
[33]	Language Identification	MFCC+PLP	GMM	88.75
[34]	Language Identification	MFCC+RASTA-PLP	FFBPNN	95.1

6. Discussion and Conclusion

In this paper several approaches to speaker recognition have been proposed, including generative and discriminative modelling such as GMM, HMM and VQ ANN, SVM and kernel based discriminative learning. Hybrid approaches such as ANN-GMM are also emphasized All of these approaches contributed to higher scores on tasks involving speaker recognition in comparison to other traditional approaches. The best models for getting encouraging outcomes in voice processing and computer vision are shown in Table 3. Better speaker attributes are being implicitly estimated in the output space using deep architectures in both supervised and unsupervised learning frameworks. This is done without making any assumptions about the explicit separation of information components. These are often referred to as shallow architectures, including

SVM, Multilayer Perceptron (MLP), Nearest Neighbor Classifiers, Principal Component Analysis (PCA), and Independent Component Analysis (ICA). Although similar techniques have been used in the past for speaker recognition tasks, they frequently don't produce good enough results on huge, extremely varied datasets. In contrast, many layers of hierarchical learning, where each layer translates nonlinear mathematical alterations from the preceding layer, can be used to train Deep Neural Networks (DNNs) to approximate task-specific knowledge given sufficient information. Majority of the work in the area of SR systems depends on different acoustic features. Performance of any speech processing system depends on the feature extraction block. The future direction of speaker recognition is being shaped by advancements in artificial intelligence, machine learning, and digital signal processing. In real-time voice authentication, businesses are enhancing customer service and fraud prevention by integrating speaker verification in call centers and mobile banking. This technology also supports secure transactions in e-commerce and remote healthcare applications

References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462 (1997).

[2] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4-20 (2004).

[3] D. Deshwal, P. Sangwan, and D. Kumar, "A structured approach towards robust database collection for language identification," in *2020 21st International Arab Conference on Information Technology (ACIT)*, pp. 1-6 (2020).

[4] D. Deshwal, P. Sangwan, N. Dahiya, N. Nehra, and A. Dahiya, "A comprehensive approach for performance evaluation of Indian language identification systems," *Journal of Intelligent & Fuzzy Systems*, (Preprint), pp. 1-17.

[5] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," arXiv preprint arXiv:1504.00923 (2015).

[6] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction,"

in *Twelfth Annual Conference of the International Speech Communication Association* (2011).

[7] K. J. Devi and K. Thongam, "Automatic speaker recognition with enhanced swallow swarm optimization and ensemble classification model from speech signals," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-14 (2019).

[8] R. Chakroun and M. Frikha, "Robust features for text-independent speaker recognition with short utterances," *Neural Computing and Applications*, pp. 1-21 (2020).

[9] H. Hassanzadeh, J. A. Qadir, S. M. Omer, M. H. Ahmed, and E. Khezri, "Deep learning for speaker recognition: A comparative analysis of 1D-CNN and LSTM models using diverse datasets," in *2024 4th Interdisciplinary Conference on Electrics and Computer (INTCEC)*, pp. 1-8 (2024).

[10] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using fusion of features with feedforward artificial neural network and support vector machine," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 170-176 (2020).

[11] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov, "A large-scale semi-supervised dataset for offensive language identification," arXiv preprint arXiv:2004.14454 (2020).

[12] H. Mukherjee, S. Ghosh, S. Sen, O. S. Md, K. C. Santosh, S. Phadikar, and K. Roy, "Deep learning for spoken language identification: Can we visualize speech signal patterns?," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8483-8501 (2019).

[13] S. Ganji, K. Dhawan, and R. Sinha, "IITG-HingCoS corpus: A Hinglish code-switching database for automatic speech recognition," *Speech Communication*, vol. 110, pp. 76-89 (2019).

[14] M. K. Mustafa, T. Allen, and K. Appiah, "A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition," *Neural Computing and Applications*, vol. 31, no. 2, pp. 891-899 (2019).

[15] B. Mak, J. C. Junqua, and B. Reaves, "A robust speech/non-speech detection algorithm using time and frequency-based features," in *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 269-272 (Mar. 1992).

[16] R. V. Pawar, R. M. Jalnekar, and J. S. Chitode, "Review of various stages in speaker recognition system, performance measures and recognition toolkits," *Analog Integrated Circuits and Signal Processing*, vol. 94, no. 2, pp. 247-257 (2018).

[17] B. Singh, V. Rani, and N. Mahajan, "Preprocessing in ASR for computer machine interaction with humans: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 3, pp. 396-399 (2012).

[18] B. Singh, V. Rani, and N. Mahajan, "Preprocessing in ASR for computer machine interaction with humans: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 3, pp. 396-399 (2012).

[19] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pp. 130-133 (2019).

[20] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616-1629 (2019).

[21] P. K. Nayana, D. Mathew, and A. Thomas, "Performance comparison of speaker recognition systems using GMM and i-vector methods with PNCC and RASTA PLP features," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pp. 438-443 (2017).

[22] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of text independent speaker identification systems using GMM and i-vector methods," *Procedia Computer Science*, vol. 115, pp. 47-54 (2017).

[23] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5115-5119 (2016).

[24] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association* (2015).

[25] S. Dey, S. Madikeri, M. Ferras, and P. Motlicek, "Deep neural network based posteriors for text-dependent speaker verification," in *2016*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5050-5054 (2016).

[26] A. H. T. Mahde, "Speech recognition by improving the performance of algorithms used in discrimination," *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 11 (2019).

[27] A. Kanagasundaram, D. Dean, S. Sridharan, and C. Fookes, "DNN based speaker recognition on short utterances," arXiv preprint arXiv:1610.03190 (2016).

[28] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 165-170 (2016).

[29] P. Kenny, T. Stafylakis, P. Ouellet, V. Gupta, and M. J. Alam, "Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition," in *Odyssey*, vol. 2014, pp. 293-298 (2014).

[30] D. Deshwal, P. Sangwan, and D. Kumar, "A language identification system using hybrid features and back-propagation neural network," *Applied Acoustics*, vol. 164, p. 107289 (2020).

[31] C. C. Bhanja, M. A. Laskar, and R. H. Laskar, "A pre-classification-based language identification for Northeast Indian languages using prosody and spectral features," *Circuits, Systems, and Signal Processing*, vol. 38, no. 5, pp. 2266-2296 (2019).

[32] P. Kumar, A. Biswas, A. N. Mishra, and M. Chandra, "Spoken language identification using hybrid feature extraction methods," *arXiv preprint arXiv:1003.5623*, 2010.

[33] A. Hajavi and A. Etemad, "A deep neural network for short-segment speaker recognition," arXiv preprint arXiv:1907.10420 (2019).

[34] P. Sangwan, D. Deshwal, and N. Dahiya, "Performance of a language identification system using hybrid features and ANN learning algorithms," *Applied Acoustics*, vol. 175, p. 107815 (2021).

Received August, 2024