Periodica Polytechnica Electrical Engineering and Computer Science, 67(3), pp. 300–336, 2023

Speaker Recognition through Deep Learning Techniques

A Comprehensive Review and Research Challenges

Nirupam Shome[1*], Anisha Sarkar[1], Arit Kumar Ghosh[1], Rabul Hussain Laskar[2], Richik Kashyap[1]

[1] Department of Electronics and Communication Engineering, Assam University, Dargakona, 788011 Silchar, Assam, India
[2] Department of Electronics and Communication Engineering, National Institute of Technology, NIT Road, Fakiratilla, 788010 Silchar, Assam, India
[*] Corresponding author, e-mail: nirupam.shome@aus.ac.in

Received: 09 August 2022, Accepted: 13 December 2022, Published online: 30 March 2023

Abstract

Deep learning has now become an integral part of today's world and advancement in the field of deep learning has gained a huge development. Due to the extensive use and fast growth of deep learning, it has captured the attention of researchers in the field of speaker recognition. A detailed investigation regarding the process becomes essential and helpful to the researchers for designing robust applications in the field of speaker recognition, both in speaker verification and identification. This paper reviews the field of speaker recognition taking into consideration of deep learning advancement in the present era that boosts up this technology. The paper continues with a systematic review by firstly giving a basic idea of deep learning and its architecture with its field of application, then entering into the high-lighted portion of our paper i.e., speaker recognition which is one of the important applications of deep learning. Here we have mentioned its types, different processing techniques, challenges that come across in this technology, performance evaluation criteria, deep learning implementation frameworks, and lastly various databases used in the field of speaker identification (SI) and Speaker Verification (SV).

Keywords

deep learning, speaker recognition, speaker identification, speaker verification

1 Introduction

Deep learning provides encouraging results especially when we are dealing with a large dataset that in turn gives accurate and satisfactory results. Also, it outshines where we don't have knowledge or information about the domain on which we are using it and thus in turn handles complex problems efficiently as compared to other existing techniques. It always provides an ease to the user to handle with great convenience in terms of handling and proves better in the existing scenario, when compared with other machine learning techniques. Deep learning has been extensively used to solve problems for different domain, such as translation of written text into another language, image detection by self-driving cars, image classification, fraud detection of credit card, prediction of next word in an online email editor, melody harmonization, speech recognition, speaker recognition, disease detection, networking, and communication etc.

Speech signal or voice print is an important aspect of speech technology. Speech signal varies from person to person based on some acoustic characteristics such as speaking rate, emotions, sex, gender, size of the vocal tract, the accent of speaking, and rate of vibration of the vocal fold, etc., and enables researchers and scientists to use these unique features to distinguish different speakers. A person authentication system with or without the physical presence of a speaker adds another dimensionality to its application. Remote person authentication (over the phone) has become more popular in this decade.

Speaker recognition is a broader category of speech technology that comprises speaker verification (SV) and speaker identification, which are much popular in speech technology for their various application in the field of biometrics, forensic science, security purpose, authentication technique [1–3]. This part of speech technology has versatile applications and researchers were started to work in this field, because of its versatility it becomes challenging to review. So, we tried to accomplish a systematic review out of it for having future benefits to the people work in this field. And due to the significant growth of

Cite this article as: Shome, N., Sarkar, A., Ghosh, A. K., Laskar, R. H., Kashyap, R. "Speaker Recognition through Deep Learning Techniques: A Comprehensive Review and Research Challenges", Periodica Polytechnica Electrical Engineering and Computer Science, 67(3), pp. 300–336, 2023. https://doi.org/10.3311/PPee.20971

deep learning and its easily accessible software and hardware tools, it gives a good boost to speaker recognition technology. Our systematic review aims to give a brief introduction of speaker recognition that comes into play in today's era and have a profound elaboration about speaker identification (SI) and speaker verification (SV), databases used in these respective areas, classification, and feature extraction techniques. Also, pre-processing and post-processing parts are covered that improves the deep learning system performance effectively. Speaker recognition with deep learning uses the voice prints of humans and is used to train the model by providing the raw input to the system and hence can serve in different fields where authentication of a person is needed. As voiceprints of different individuals attend uniqueness, so can be distinguished easily with the help of some acoustic features of the speech signal [4–8]. So, to execute this many databases are available in the field of speaker recognition, dedicated to speaker identification and speaker verification respectively to perform different tasks and processes.

1.1 Different aspects of the speaker recognition system

Speaker recognition system comprises of different aspects to execute a certain. It consists of the use of databases (discussed in this paper), various processes including pre-processing, feature engineering (that includes feature extraction and feature reduction), classification (discriminative, generative, and hybrid), and evaluation process (accuracy, precision, Re-call, F-measure, Equal Error Rate, Receiver Operating Characteristics etc.) are explained in this paper. Again, this process includes various steps that are needed to be done in every zone which is having separate importance of its own. In the pre-processing part which is the first step of speaker recognition where speech signal is filtered for silence removal, then pre-emphasis, framing, windowing, endpoint detection, and lastly normalization of the speech signal. All of them are also explained in detail in this paper. Fig. 1 shows the of classification of speaker recognition system. It gives a detailed description of requirement and possible techniques for speaker recognition. The comparative analysis of different survey papers with our paper is presented in Table 1 [9–15]. Table 2 gives the acronyms used in our paper with definitions.

The contributions of our paper are as follows:

1. To provide a brief idea about the deep learning mechanism and its widespread applications.
2. To offer an overview of the speaker recognition system, its variants, challenges in the development of speaker recognition system, and speaker identity information extraction challenges.
3. To review recent developments in the field of speaker recognition with help of a deep learning network and discuss the progress of deep neural networks in pre-processing, feature extraction, and classification.

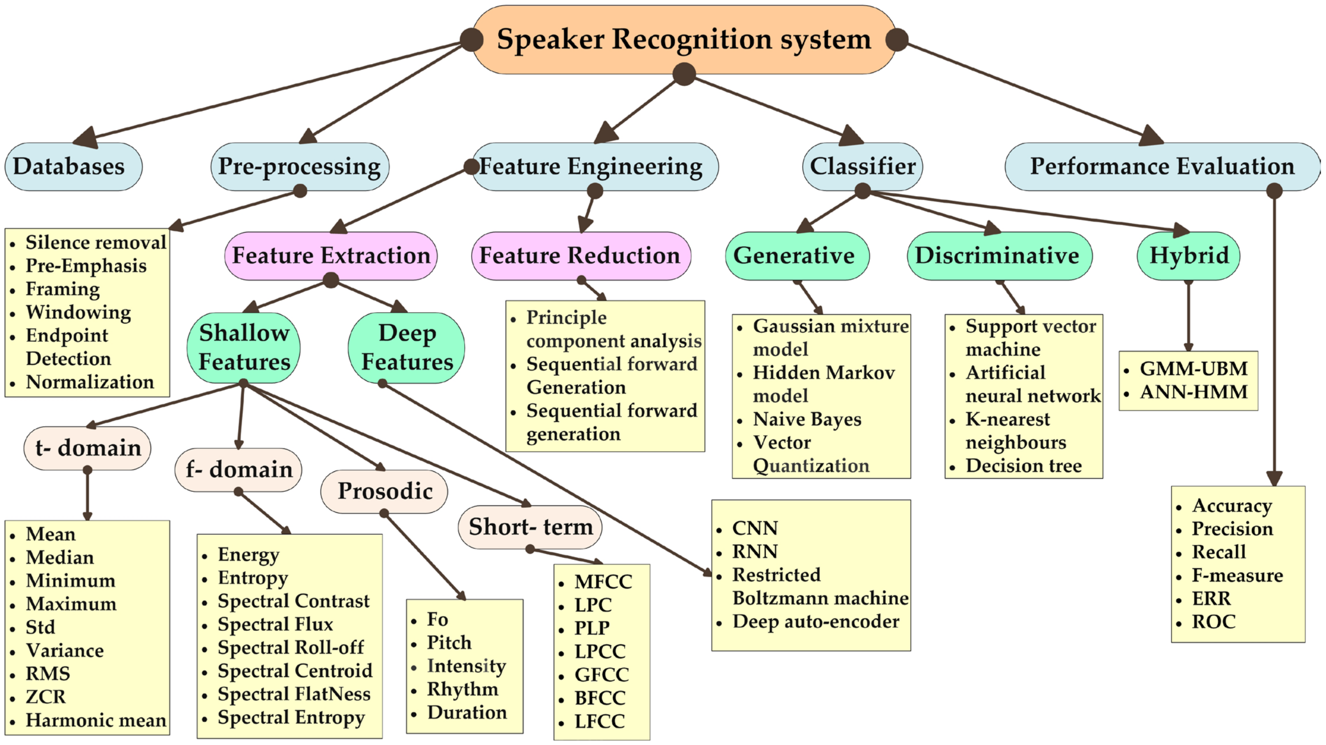

Fig. 1 Classification of speaker recognition system

Table 1 Comparison of our review with that of the other surveys. The comparison is based on preprocessing methods (Pre.), features selection (FeatSe.), features extractor (FeatEx.), Deep classifier (Dclass.), databases (Data), evaluation metrics (EvMet), machine and deep learning methods (ML/DL) and DL implementation frameworks (ImpFwk).

Paper	Year	Field of research			Review domain								
		ASR	SR		Pre.	FeatSe.	FeatEx.	Dclass.	Data.	EvMet.	ML	DL.	ImpFwk.
			SI	SV									
[9]	2014	×	✓	×	×	×	✓	×	✓	×	✓	×	×
[10]	2016	×	✓	×	×	×	×	✓	×	×	✓	✓	×
[11]	2017	×	✓	×	×	×	✓	✓	✓	×	✓	×	×
[12]	2017	×	✓	×	×	×	✓	✓	×	×	✓	×	×
[13]	2018	✓	×	×	×	×	✓	✓	✓	×	×	✓	×
[14]	2019	✓	×	×	×	×	×	✓	×	×	×	✓	×
[15]	2021	×	✓	×	✓	✓	✓	×	✓	✓	✓	✓	✓
Our paper	2022	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

4. To give a detailed investigation of database, deep learning framework, and performance evaluation parameters required for speaker recognition system analysis. To highlight research gaps and future challenges in the development of the speaker recognition system.

This paper is arranged as follows. In Section 2, we have discussed the methodology adopted to accomplish this review. Section 3 provides a basic idea of deep learning, about its architecture, working mechanism, and its application areas. A brief idea about speaker recognition, its challenges, and speaker uniqueness aspects are discussed in Section 4. Section 5 describes speaker recognition by deep learning with detailed discussion in pre-processing, feature extraction, classification techniques, and post-processing techniques. Then Section 6 shows Performance evaluation criteria, and Section 7 discusses the framework of deep learning implementation. Section 8 discusses the application of speaker recognition in different fields. In Section 9, databases used in the field of speaker recognition, both in speaker verification and identification are explained and lastly, future challenges and conclusions are discussed in Sections 10 and 11 respectively, which gives a systematic review of the domain of speaker recognition.

2 Methodology adopted

For writing this systematic review paper various information is fetched from different information sources. The list of digital libraries which were searched to have some valuable information for the publication of this paper is IEEE Explorer [16], Springer Link [17], ACM [18], Science Direct [19], Web of Science [20], MDPI [21], Research Gate [22], and help of internet is also taken to write this review paper. So, these are some of the sources from where information is gathered to review this paper, and different aspects were researched and based on this the collected information is provided in this paper.

The publication made by this review paper will be in English and this review paper contains the review of publications from the year 1991 to 2021 to form a systematic and detailed review on speaker recognition which is a topic under the field of attention and not much research and review are being carried out in this field. This review is done to have a basic idea or to gather information on what is going on in this dedicated field of speaker recognition to date and what portion of it needs the attention of research and development in the upcoming future. The publication made by this review paper will be in English and this review paper contains the review of publications from the year 1991 to 2021 to form a systematic and detailed review on speaker recognition which is a topic under the field of attention and not much research and review are being carried out in this field. This review is done to have a basic idea or to gather information on what is going on in this dedicated field of speaker recognition to date and what portion of it needs the attention of research and development in the upcoming future.

3 The idea about deep learning

Deep learning (DL) which is also known as Hierarchical learning or deep machine learning or deep structured learning is completely based on some sort of algorithms that trains the model and data by using multiple layers of the network which is of complex structures or composed of multiple non-linear transformations [23–28]. Deep learning is a machine learning (ML) technique that possesses

Table 2 List of abbreviations with definitions

Abbreviations	Definitions	Abbreviations	Definitions
SR	Speaker recognition	SRE	Speaker Recognition Evaluation
SI	Speaker identification	DL	Deep learning
SV	Speaker verification	PLDA	Probabilistic Linear Discriminant Analysis
ASR	Automatic speech recognition	TDNN	Time-Delay Neural Network
ML	Machine learning	F-BANK	Filter-bank
AI	Artificial intelligence	MKMFCC	Multiple Kernel Weighted Mel Frequency Cepstral Coefficient
GPU	Graphics processing unit	IMMFC	Incremental Multiple Medoids Based Fuzzy Clustering
AE	Autoencoder	PLPC	Perceptual Linear Prediction Coefficient
UDBN	Universal Deep Belief Network	IMFCC	Inverted Mel Frequency Cepstral Coefficients
DCT	Discrete Cosine Transform	MFCC	Mel Frequency Cepstral Coefficients
CLPC	Cepstral Linear Prediction Coefficients	MFSC	Mel Frequency Spectral Coefficients
GMM	Gaussian Mixture Model	LFCC	Linear Frequency Cepstral Coefficients
HMM	Hidden Markov Model	GFCC	Gammatone Frequency Cepstral Coefficients
CDBN	Convolutional Deep Belief Network	BFCC	Bark Frequency Cepstral Coefficients
K-NN	K-Nearest Neighbor	LPCC	Linear Predictor Cepstral Coefficients
SVM	Support Vector Machine	LPC	Linear Predictor Coefficients
ANN	Artificial neural network	PLP	Perceptual Linear Prediction
CNN	Convolutional Neural Network	CFCC	Cochlear Cepstrum Coefficients
RNN	Recurrent neural network	RASTA-PLP	Relative Spectra-Perceptual Linear Predictive
DNN	Deep neural network	DWT	Discrete Wavelet Transform
RF	Random Forest	WSBC	Wavelet Sub-Band Coding
NB	Naive Bayes	WPT	Wavelet Packet Transform
PCA	Principle Component Analysis	LBP	Local Binary Pattern
UBM	Universal Background Model	ZCR	Zero-Crossing Rate
FCNN	Fully Convolutional Neural Network	BP	Back Propagation
RBM	Restricted Boltzmann Machine	CAE	Convolutional Variational autoencoder
DBN	Deep Belief Network	ELU	Exponential Linear Rectification
LSTM	Long Short Term Memory	IBP	Invariant Backpropagation
DAE	Deep autoencoder	ReLu	Rectified Linear Unit
SAE	Stack autoencoder	Resnet	Residual Neural Network
DT	Decision Tree	VGG	Visual Geometry Group
RASTA	Representations Relative Spectra	DGS	Deep Gaussian Supervector
GAN	Generative Adversarial Network	DGCS	Deep Gaussian Correlation Super Vector
Bi-LSTM	Bidirectional Long Short-Term Memory	HT	Haman Transform
ANC	Adaptive Noise Canceller	FBLPCS	Frame-Based Linear Predictive Coding Spectrum
SG	Savitzky Golay	RMSE	Root Mean Square Error
T2IS	Type-2 Information Set	ROC	Receiver Operating Characteristics
VAE	Variational autoencoder	EER	Equal Error Rate
ADAM	Adaptive Moment Estimation	FAR	False Acceptance Rate
RMSProp	Root Mean Squared Propagation	FRR	False Rejection Rate
SGD	Stochastic Gradient Descent	DAN	Deep Attention Neural Network

networks that is capable of learning from unsupervised or unlabeled data and is also known as deep neural network (DNN) that broadly comes under the category of Artificial intelligence (AI). Deep learning function can mimic the working and functioning or behavior of the human brain, which is used in processing data and to create patterns for the task of decision making, detecting objects, translating languages, recognizing speech, etc. It can learn by itself from both unstructured and unlabeled data without the supervision of human beings. It can be used in the field of detecting frauds, money laundering and for solving many real-life problems. It requires high-end machines as compared to machine learning because its computational complexity is quite high. GPUs have become an integral part nowadays of deep learning to execute their algorithms. The time required to create a model mainly depends on the GPU specification.

3.1 Working mechanism of DL

All computer program uses deep learning techniques learned on its own to classify or do a particular task. The deep learning algorithm applies a non-linear transformation to its input and learns to create a statistical model that produces the output. The learning process continues with every iteration until the model achieves the highest acceptable level of accuracy. The deep learning technique requires a very large amount of training data to train the model and to produce output with high accuracy, accordingly a huge data set is to be fed into its input.

3.2 Comparison of DL and ML in technology development

In the existing technologies, deep learning (DL) is achieving much more popularity and attention as compared to other machine learning (ML) techniques. This is due to the ability of deep learning to deal with many datasets without compromising in its high accuracy rate. Also, it achieves great power and flexibility by learning from an uncontrolled environment and representing the world more merely. On the other hand, other machine learning needs lots of expertise, human supervision, or intervention for the implementation of the competent learning models. Due to its ability to handle a huge amount of data it provides opportunities for innovations in learning models.

3.2.1 When to use deep learning over other techniques

Here we have discussed several conditions under which deep learning can be a better option compared to other techniques:

- To deal with the large-size dataset: Considering all the existing techniques, deep learning has an excellent ability to deal with huge amounts of data and also gives out a high accuracy as compared to other techniques. For example, machine learning gives a good performance in terms of accuracy rate when it deals with small data sets. So, in this area by the implementation of DL, we can increase the capacity of the data dealing factor without compromising its accuracy rate.
- Lack of feature/characteristics knowledge in a particular field: Under such circumstances deep learning to give away all the worries, as in DL we don't have to worry about the features when dealing in a particular domain. So, without the proper understanding or knowledge of one particular domain, one can implement DL over other existing techniques.
- Solving complex problems: In dealing with complex scenarios, deep learning performs better as compared to other techniques. Some of the examples of complex scenarios are speech recognition, speaker recognition, language processing, image processing, image classifications, etc.

3.3 The architecture of deep learning

Now in this part, we will briefly discuss the understanding of the architecture of DL. Deep learning architectures like recurrent neural network (RNN) or Feedback neural network, deep belief neural network, Convolutional Neural Network, and many more are applied to different fields such as automatic speech recognition, speaker recognition, natural language processing, audio recognition, bioinformatics, etc., where they showed the great impact of DL on various applications that proofs the significance of deep learning in the modern world [29, 30]. Deep learning network consists of three main layers, namely input layer, an output layer, and hidden layer.

The depth or the deepness of a particular deep learning neural network is defined by the number of hidden layers. And depending on the type of hidden layer used, the system can learn different non-linear functions. The activation function in DL determines the characteristics of the non-linearity represented by the DNN model and specifically this non-linearity is that is observed within the data represented by these hidden layers.

The input layer accepts the input in various forms and the output layer translates these characteristics of non-linearity into the prediction or predicted output. The deep learning neural network consists of multiple layers of hidden layers that in turn are responsible for the accuracy of a particular deep learning model.

A deep neural network (DNN) achieves higher accuracy as compared to other networks. In simple words, DNN is a neural network incorporated with some level of complexity and consisting of multiple hidden layers that process data in complex ways by indulging sophisticated modeling.

Fig. 2 describes the architecture of the DNN where x1, x2, ... xn are the input layers that feds on data as input, and y1, y2, ... yn are the output layers that produces the output of a particular deep neural network after processing from the hidden layers or also known as a black box. Here, w1, w2, ... wn are the synaptic weights incorporated with each neuron or perceptron that reflects the strength or amplitude of the connecting nodes or neurons. Hidden layers are more than one in a DNN as mentioned here also in the above diagram, which is represented as layer 1, layer 2, ... layer n that forms a multi-layer DNN.

3.4 Types of deep learning networks

Deep learning networks can be classified based on their architecture and applications. A detailed discussion of them is given here.

3.4.1 Deep neural networks (DNNs)

A deep neural network (DNN) is a category of ANN where it consists of multiple hidden layers. When there are two or more two hidden layers, DNN comes into play. So, it adds complexity to a particular model and is used to process a complex set of data by applying mathematical modelling. It is much beneficial in the scenario whenever want to replace human labor with consistent work without compromising the efficiency rate. A deep neural network is a kind of human brain that keeps on learning things based on which it reacts in different scenarios. DNN whenever gets new information, it learns or trains its model to act accordingly. Whenever we must solve a difficult problem, the process of learning becomes very difficult and requires deeper information. It uses various nodes and various layers to deal with a complex task. DNN consists of various small nodes in the layers of a model, which are the smallest unit of the network that mimics the role of neurons of human brain. Whenever these nodes are triggered by some sort of stimuli, it starts to give a response or starts to react like the human brain. They are forming groups and are grouped into layers and are connected according to necessity. The interesting fact about DNN is that it provides accurate results without having marked data or labelled data and that is how it becomes unique in the field of machine learning.

3.4.2 Deep Belief Networks (DBNs)

In a Deep Belief Network, Restricted Boltzmann Machines (RBM) are connected sequentially. It can be used as a substitute for deep feedforward networks. In this particular network the Boltzmann machine train itself until convergence is achieved and again the output of each Boltzmann machine is fed into the next Boltzmann machine as input to train the system. So, this all trains itself until converges. These networks are used for various purposes like to recognize clusters, images, in sequencing video, and capture data in motion etc. Deep Belief Networks continuum decimal instead of binary data. They are a graphical representation that is usually generative in nature, means which produces all the possible values that are generated according to the required scenario. DBNs comprise multiple layers incorporated with some sort of values whose main objective is to help the system to classify the given feed data into various categories based on the characteristics on

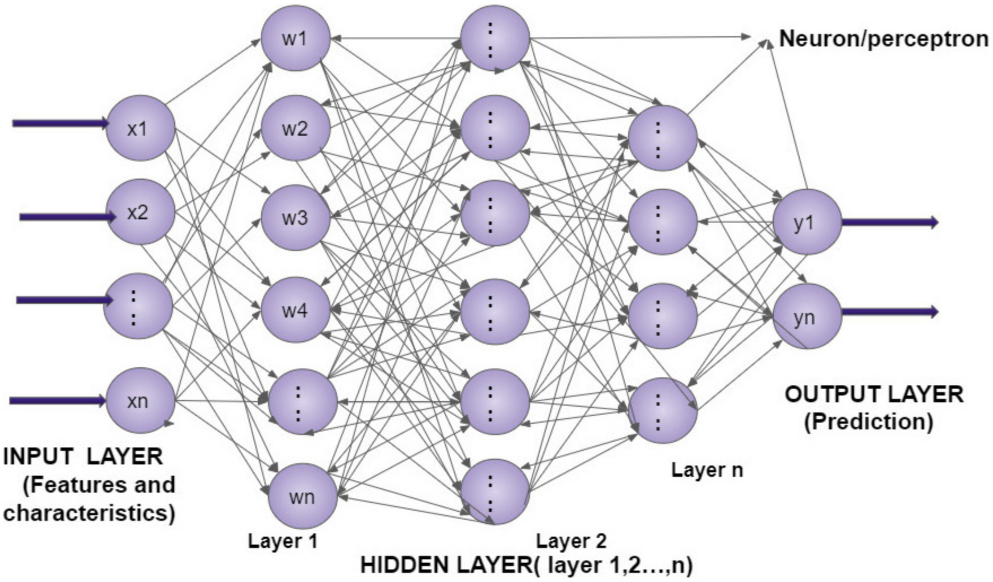

Fig. 2 Representation of DNN and its different layers

which the network model classifies those data. Training of Deep Belief Network comprises of few steps. Initially, we must train the layers of the network that has several properties and have the capability of obtaining the input signals from the pixels provided on the perceptron directly. Now immediately after this stage, we have to treat the pixel values of the layers and will learn the features that are derived or obtained in the previous hidden layers. And as subsequently the layers are added to the networks. There will be an improvement in the previous layers of the network as it is learning on its own according to the new information that it is encountering in terms of training data.

3.4.3 Convolutional Neural Network (CNN)

This particular neural network contains one or more than one convolutional layer that is the class of neural networks. All the convolutional layers perform a linear operation i.e., convolution that is why named the convolutional layer, where a set of multiplication of weights and inputs are done. A Convolutional Neural Network has an excellent capability of capturing the spatial and temporal dependencies present in the signal with the application of relevant filters. The major role of the Convolutional Neural Network is to reduce the data requirement, that helps the neural network to process data easily with great convenience and without compromising the crucial features that are used to get a good, predicted output. A Convolutional Neural Network comprises multiple layers of neurons that mimic the human brain. When input signals are feed into the input layer then activation functions are generated that subsequently passed to the next layers. First layer of the CNN usually extracts some basic features whose output are then passed to the next layers as input and now it detects more complex features as compared to the first layer. This process continues and with the increase in layers complexity increases, otherwards CNN become deeper and will extract complex features to process further. It uses pooling layers that help it to reduce the computational power required to process data.

3.4.4 Recurrent neural networks (RNNs)

Recurrent neural networks consists of internal memory. It is a generalized form of feedforward neural network where it does not form a loop and hence, recurrent where it performs the same functions for every input data of the network. In this network, the output of the recent input will depend on the previous output of the network. This network is used in speech recognition and natural language processing. With the help of this network, it recognizes the data characteristics or features sequentially and uses patterns that predicts different scenarios. These networks are used to solve complex problems with the help of feedback loops that process data sequentially to get the output. RNN uses the backpropagation technique to loop the information back into the input throughout the computational process. RNN converts the independent activation function into a dependent activation function by introducing the same weights and biases to all neurons present in the layers. This in turn reduces the complexity of the network with an increase in parameters. It then memorizes all its previous steps by providing the output of the previous step to the input of the current step. RNN is used along with convolutional layers that extend the signal points of the neighborhood neurons. And the main advantage behind using the RNN is that it remembers every step that is very useful in dealing with complex scenarios. But the training of the RNN model is very complex.

3.4.5 Convolutional Deep Belief Networks (CDBNs)

This network is one type of artificial neural network that contains multiple layers of convolutional Restricted Boltzmann Machine. Also, it is used as a generative model in deep learning neural networks. It uses max-pooling which in turn is used to reduce dimensions in the layer of the network. This network is translational invariant, and the main feature is the ability to scale under high dimensional signals. It is a type of Deep Belief Network that accepts the continuum decimals rather than accepting binary data, unlike other networks. It consists of multiple layers where connections are made between layers, not between the units of the layers. The first step behind training the CDBN network is to accomplish the layers in a systematic manner like that of a Deep Belief Network and then training of the model is done.

3.4.6 Autoencoders (AEs)

Autoencoders fall under the category of artificial neural networks that are used to encode the set of data by learning data coding. Here, input and output remain the same. The input is compressed into a latent space representation and by using this representation the output is reconstructed. It is used to learn the compressed form of raw input data. It is composed of decoder sub-models with an encoder incorporated with them. It is also used in the feature extraction of raw data that is eventually used to train the model. It is unsupervised learning method but uses supervised learning methods to train its model. It is a special type of neural network that is being trained to copy the input set of raw data into its output. These networks are also trained to remove noise from the model. Here, it learns important features present in the dataset by minimizing the reconstruction error present between the input and the output data. And in autoencoders, the number of neurons that is present in the output layer is the same as that of several neurons present in the input layers.

3.5 Difficulties with deep neural networks (DNN)

Here we have highlighted the short coming associated with the DNN. Researchers must come with some solutions to that in near future, so as to improve the overall performance of the DNN systems.

3.5.1 Deep learning requires a large amount of data

DL needs a large amount of data as it comprises of complex network having many hidden layers and to give accurate results it needs a lot of datasets. Deep neural network is very much dependent on amount of dataset, as the accuracy rate is fully achieved on the amount of data supplied to the model. So, this in turn creates a problem for the researchers and scientists in dealing with large number of datasets. Due to the highly complex structure, a DNN consists of many hidden layers that feed on a large set of databases to achieve desirable result with high accuracy rate. Deep learning neural network performs well if it provided large amount of data. Due to its high complexity in the layers model became very slow to train and consumes a huge amount of computational power. So, with all this, it creates difficulty for the researchers and the scientists to handle huge amounts of datasets that requires for DNN model in the practical scenario.

3.5.2 Deep learning neural networks don't give reasons, or the understanding based on which it is providing the output or the conclusion

The process that the DL network undergoes is not being explained by the model and that is why it is known as a black box. As we are not able to know what is going on in the processing part based on which we are getting the conclusion and how it mimics the behavior of the human brain. Like the human brain, the work we used to do, the voluntary movements that we used to execute, the reflex action that we produce is everything controlled by the nervous system of our body. But we are not aware of the production and working techniques behind the systematic behavior, which is a black box as we don't know what is happening between our neurons. And not clear about the procedure that the nervous system is opting to produce these behaviors. So, DNN is also playing the similar role in the field of artificial intelligence (AI) for which we have not enough information about the procedure of DNN. So, in turn, it doesn't have reasons based on which it is giving us a conclusion.

3.5.3 Deep learning neural network algorithms are very costly to build

The algorithm used in DL draws a large amount of computational power, which directly depends on the amount of data used, depth, complexity of the network. Hence, the implementation of this needs a lot of money and becomes expensive to implement. To execute DNN we need a lot of set up which in turn makes it more expensive to use as compared to other machine learning techniques. A deep learning neural network becomes expensive when it comes to training a DNN model, which is very much dependent upon some of the factors like dataset size, model size, the volume of the training, etc. Whenever these factors increase in training a DL model, the expense of a particular DNN model execution rises.

3.6 Deep learning application area

The way we look at technology or the perception towards the technology is changed by deep learning. According to some predictions, applications in the field of deep learning will largely affect our lifestyle shortly, and it is having an effect nowadays also. Within the next 10 years, development, or advancement in the field of deep learning in its tools, libraries, programming languages will become a standard component for every software toolkit. Table 3 [31–42] gives us an idea about different application areas of deep learning technology in different fields.

Fig. 3 shows fields of application of speech processing, which uses deep learning to model their system. Various fields such as speaker recognition, speech recognition, language identification, speech enhancement, etc., are mentioned in the following pie-chart that uses DL to model their system to get better classification results.

From the description presented here, we can see that there are many applications of DL in various fields and many more to evolve according to our requirements soon. Till now we have discussed about the DL methods and its various applications. Now in our further discussion, we will be going to have a detailed discussion on one of the application areas of DL i.e., speaker recognition (verification and identification).

Table 3 Different application areas of deep learning technology

Application area	Data	Target value/result value	The algorithm used/model used
Translation of written text into another language [31]	Raw text	The text is written will be translated into another language	Encoder-Decoder network that is built with RNN (recurrent neural network)
Image detection by self-driving cars [32]	Raw images	Labeling will be against the identified image	CNN (Convolutional Neural Network)
Image classification [33]	Raw images	Labeling against the classified image	CNN
Fraud detection of credit card [34]	Valuable details about the credit card including its past historical transaction, amount, date, place, etc.	Detecting whether the transaction made is fraud or not by producing binary value	Random Forest (RF) algorithm
Prediction of next word in an online email editor [35]	Large chunk of textual data	A word that fits with the sentence	RNN
Melody harmonization [36]	Melody	Harmonized melody	Generative Adversarial Network (GAN)
Speech recognition [37]	Raw speech signal	Unknown speech is recognized	Bi-LSTM (Bidirectional Long Short-Term Memory), LSTM (Long Short-Term Memory), CNN
Speaker recognition [38]	Raw speech signal	Unknown speaker's speech is recognized	CNN, DNN, Res-Nets
Speaker verification [39]	Raw speech signal	Unknown speaker's speech is verified	CNN, LSTM, Bi-LSTM
Disease detection [40]	Raw image of the suspected body part of the patient	The disease is detected or not in a patient	CNN, LSTM, Bi-LSTM
Network Traffic Classification [41]	ALL/L7 layers [256÷2304 B] 4–6 fields [4÷32 packets] Packet directions	Network Traffic is estimated	SAE, LSTM, 1D-CNN, 2D-CNN, Hybrid LSTM+2D-CNN
End-To-End Reconstruction Task [42]	Raw IQ samples or symbols to codewords or bits.	Signal Reconstruction in communication	Radio Transformer Networks (RTNs), CNNs

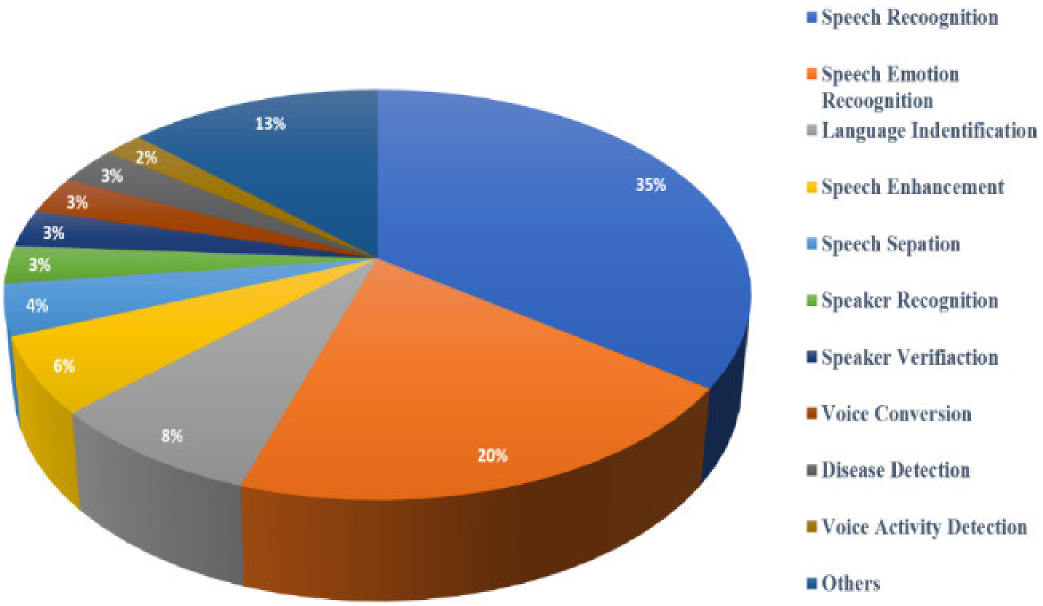

Fig. 3 Application of DL in speech processing

4 Speaker recognition system

Speaker recognition (SR) is the process of identifying or verifying a speaker's voiceprints from the collection of several speech samples. The main objective of speaker recognition is to create a system that can verify a person's voice and can also be proficient in identifying a person from his/her voice. Speech signal contains information about one's identity, gender, age, emotions, sex, health condition, etc., with some amount of ambient noise and redundant information. Speech signal varies from person to person based on some acoustic characteristics such as speaking rate, emotions, sex, gender, size of the vocal tract, the accent of speaking, and rate of vibration of the vocal fold, etc., and enables researchers and scientists to use these unique features to distinguish different speakers. A person authentication system with or without the physical presence of a speaker adds another dimensionality to its application. Remote person authentication (over the phone or the internet) has become more popular in this decade.

The speaker recognition system comprises different aspects to execute a certain task. It consists of various processes including pre-processing, feature engineering (that includes feature extraction and feature reduction), classification (discriminative, generative, and hybrid), and evaluation process (accuracy, precision, Re-call, F-measure, Equal Error Rate, Receiver Operating Characteristics, etc.) with the help of various datasets. Again, this process includes different steps that are needed to be executed in depth which are having separate importance of their own. In the pre-processing part which is the first step of speaker recognition where speech signal is filtered for silence removal, then pre-emphasis, framing, windowing, endpoint detection, and lastly normalization of the speech signal.

Speaker recognition is a very important aspect of the field of speech technology. It is a very popular topic where researchers are keen to research its application areas like various fields of forensic labs, biometric verification or authentication, security areas, and many more. Speaker recognition, especially identification and verification are grabbing great attention to the scientist and researchers as it is a very growing topic nowadays. An ample number of studies have been done in this field and thereby new methods have been discovered but still, it is a very versatile topic that requires in-depth investigation. The popularity of deep learning growing rapidly in the present technology due to its easily accessible or reachable software and inexpensive hardware equipment. Deep learning (DL) is adapted in every field to find out the solution to any task. Now, the next task is to get an idea about speaker identification and speaker verification as these are required in performing speaker recognition tasks [30]. For a clear understanding of speaker recognition, we are recommending the work mentioned in the paper [43]. It gives us an idea about the deep learning techniques that are applied specifically in the fields of speaker identification (SI) and speaker verification (SV) of speaker recognition tasks that are used earlier and nowadays.

Fig. 4 shows the different steps of a speaker recognition system, comprised of the pre-processing part, the feature extraction part, the classification part, and lastly the output part. Firstly, the raw input speech sample of the speakers is fed into the system. This speech signal comprises various additive noises, so in pre-processing these signals are filtered to remove the redundant parts and followed by this normalizing the signals to get them ready for the feature extraction process. Now, in the feature extraction process, the feature that is important or based on which the speech signal gets classified is being extracted. And finally, in the classification part, the model is trained with the help of extracted features and at the end, further normalization is done to produce the required desired output.

In speaker recognition system especially in the training phase, the model of the SR system is trained with a huge amount of data set usually known as training data which consists of many speech signals. Now, in the feature extraction process of the training phase, features or characteristics are being extracted from the large number of the speech signal which is used to train the SR model

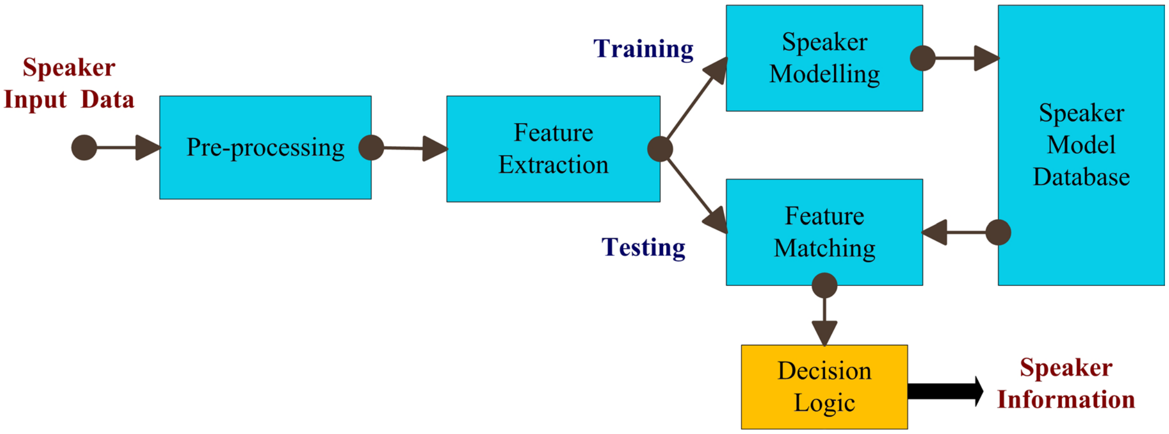

Fig. 4 Block diagram of speaker recognition system

and this trained dataset is stored in the database. In the testing phase of the SR, it is fed with a testing speech signal and from which features or characteristics are being extracted for further processing. Now, these extracted features of the speech signal are matched with features of the training data that is stored in the dataset of the SR system. The process which is common in both phases is the feature extraction phase which is an important step in the SR system. And depending on the result whether the speech data is matched with that of the trained dataset, the SR system recognize the speaker [44]. In Section 4.1 we will discuss the types of speaker recognition systems i.e., speaker identification (SI) and speaker verification (SV). Here we have listed the various work from the literature, that gives a clear idea about the progress and availability of the different techniques for various modules of speaker recognition.

4.1 Types of speaker recognition

Speaker recognition can be broadly categorized into two categories: speaker identification and speaker verification, which are discussed in Sections 4.1.1 to 4.3.5.

4.1.1 Speaker identification

Speaker identification is the task of identifying the speaker from the set of given samples of the speaker. It is done by matching the input voice sample of the speaker with the speech dataset of the different speakers and the speaker whose sound matches or closest with that of the input one will be get identified. So, when the speaker is identified within the set of the given data, then it is called closest-set or in-set speaker identification. On the other hand, if the speaker is identified as, it doesn't have the potential test subject and is not within the given set of speakers, then it is said to be the out-of-set speaker identification [45].

Fig. 5 shows the speaker identification system comprised of the training phase and testing phase. So, in the training phase of SI, firstly pre-processing of the speech signal is done and then features of the input speech are extracted by feature extraction technique, which is used to model or train the SI model and this trained dataset is being stored in the database of the speaker model. So, here is the end of the training phase of the SI system. In the testing phase, the SI model is fed with a test speech of a speaker, which is being matched with the dataset of the trained model that is stored in the database, and accordingly, the decision logic is implemented where a decision is taken whether the speaker is being identified or not.

4.1.2 Speaker verification

Speaker verification is the task of verifying the speaker present at the input of the model that claims to be an identity of the speaker or is the identity of the speaker. Another way means that the speech sample of the test utterance is compared with the existing model of the individual subject and depending on the closeness with the model, the claim of that speaker is accepted or rejected. Here, the system has to verify whether the subject is the person that is being identified by the speaker verification system or not. So, this is done by comparing the test sample of the speaker with the speech of the speaker present in the background model [45].

Fig. 6 shows the speaker verification system which consists of two phases, namely the training and testing phase. In the training phase, pre-processing of a raw speech signal

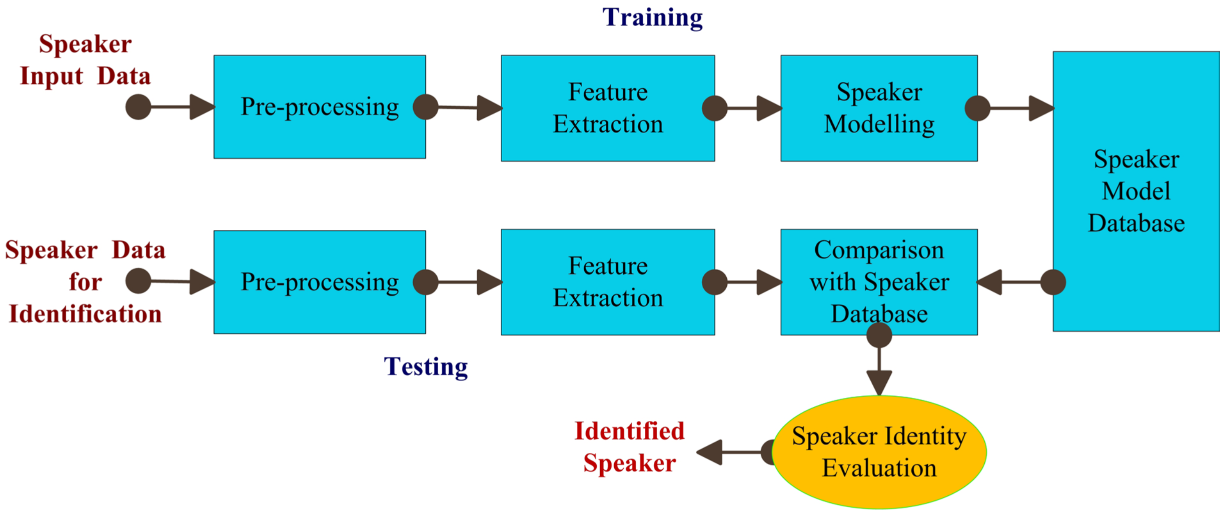

Fig. 5 Block diagram of speaker identification of SR system

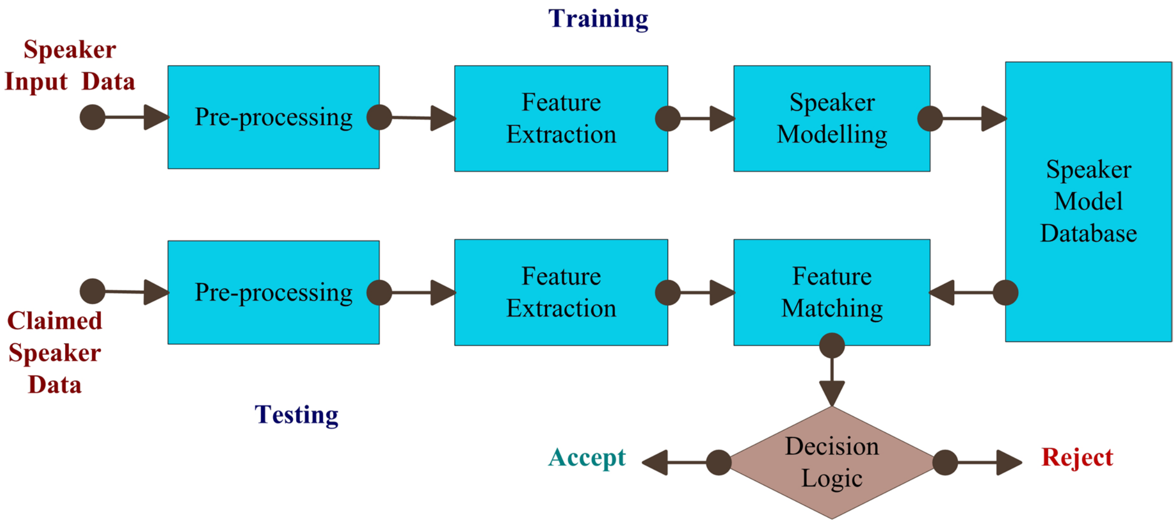

Fig. 6 Block diagram of the testing phase and training phase of speaker verification of SR system

is done and then features are extracted from the speech signal and finally with these extracted features SV model is trained and stored in the database. Next, in the testing phase of the SV process, the utterance of the speaker that claims to be the original speaker is verified by matching the features that were extracted in the feature extraction phase and the features are being matched with those that are stored in the database and afterward, if it matches, the claim is accepted and if not, it is being rejected.

4.2 Challenges in speaker recognition

Challenges in speaker recognition define the problem faced by the researcher and scientist in implementing speaker recognition technology or the obstacles that they face. Different factors cause a great effect in the execution of this particular technology, and some are discussed in Sections 4.2.1 to 4.2.6.

4.2.1 Environment

The surrounding environment consists of a lot of different types of background noises such as color noise, music, etc., which in turn affects the speaker recognition system in terms of performance. This disturbs the modelling, training, and testing phase of the system. So, whenever there is many background noises present in the environment, the speaker recognition system is very challenging. Techniques like normalization, robust modelling, speech enhancement, feature compensation are used to compensate for such problems.

4.2.2 Channel

In real-time application of speaker recognition, the utterance of the speakers is recorded utilizing sensors like microphones or headphones through some transmission channel. These channels are categorized into wired and wireless channels. Most of the unwanted counterparts of the speech are introduced in this phase and thus affect the quality of the speech signal. In comparison to wired channels, wireless channels introduce more error or unwanted signals to the desired speech. Since the channel is an integral part of any speaker recognition system; it is not possible to completely remove this problem. So, for the researchers, it is an open field to find the solution to this problem in an inefficient manner.

4.2.3 Speaker characteristics

With the help of speaker characteristics such as identity, gender, age, health status, the accent of language, emotional state, etc., causes a variation in the speech of a person that makes it different from another individual. Change in any of the speaker characteristics creates a unique voiceprint of a person that is considered in speaker recognition in case of identifying and authentication of speech signals in various scenarios. Due to these changes in speaker characteristics, a pre trained system faces enormous difficulties in speaker recognition.

4.2.4 Speech variability

It is the variability of speech due to amplitude, emotions, speaking rate, gender, accent, etc. A speech signal consists of various information regarding the health state of a person, emotional state, gender, age, social origin, regional origin, etc. So, such a huge variability of speech signal affects the speaker recognition system and makes it difficult to verify a particular voiceprint in a speaker recognition process. Also, session variability causes a great effect on the speech of an individual as when speech signals are recorded in different sessions it alters due to various factors such as amplitude, emotions, vocal tract size, speaking rate, gender, accent, etc. irrespective of recording voice prints in a continuous time frame in one go. This creates another challenge in the field of speaker recognition.

4.2.5 Language characteristics

A particular speech is represented in some languages and language varies from speaker to speaker. A particular segment of speech spoken in some languages has various characteristics such as grammar, lexicon, phonology, etc. So, the performance of the speaker recognition system differs in terms of language based on different levels of linguistic information. Such levels include word, sentence, unit and signal processing level, etc. Also, while dealing with multiple languages the performance of the speaker recognition system degrades.

4.2.6 Transducing characteristics

A transducing device such as a headphone or microphone converts mechanical waves into an electrical signal. The process of transduction may not always be linear for all transducing devices. So, due to this, there is a variation in the speech signal that results in distortion of the original signal. The spectral characteristics of the recorded speech signal depend on the handset or performance of speaker recognition device used to record the voice. And the frequency response of transducers is not the same for different recording devices. If the bandwidth of the transducer is small then the signal captured by this has less information, which deteriorates the. performance of speaker recognition.

4.3 Factors responsible for the uniqueness in the voice of a speaker

Some acoustic features are present in a person’s speech that makes it different from other individuals. There are particular factors or features that give uniqueness to the voice of a speaker and those factors are discussed in Sections 4.3.1 to 4.3.5.

4.3.1 Shape and size of the vocal tract

In terms of the structure of the vocal tract system, the shape and size matter in the distinction of speech of a speaker. It is different in terms of construction and dimensions that cause uniqueness in the voice and varies for different speakers. The time-varying speech or voiceprints from the vocal tract system is due to the unique profile of different individuals.

4.3.2 The dynamics of the articulators

Dynamics in speech is used to represent the loudness of the voice and articulation is used to represent the note of a particular voiceprint, which are different for different people. Articulators are broadly categorized as active and passive articulators. In most cases, active articulators are parts of the tongue whereas passive articulators are parts of the roof of the mouth. For speech production, the tongue is the most important articulator and plays a vital role in the dynamics of sound. Due to these large verities of articulators and their corresponding position, several sound units are produced, and this adds extra challenges to speaker recognition.

4.3.3 The rate of vibration of the vocal folds

The vibration of the vocal folds causes sounds and it is dissimilar in specific individuals. This is due to the different rate the vibration in the vocal cord which causes a significant effect on the voiceprint of a speaker. The rate of vibration of the vocal folds can be a very good discriminating criterion for speaker recognition.

4.3.4 The accent imposed by the speaker

Usually, different people impose different accents in their speech while speaking and it is depending on their community and geographical location. It is the observer that when a variety of language is spoken by a person then the original accent of that speaker gets affected. This unique accent of a speaker helps the system to identify a subject.

4.3.5 The speaking rates

Due to variability in the speaking rate of different people, the texture of the voice gets affected and it is reflected in the voice of a speaker. This property of speech signal varies from person to person and can be used to identify speakers. Speaking rates also depend on the language and ethnicity of the person.

5 Speaker recognition (SR) by deep learning

Speaker recognition is the process of identifying and verifying a speaker's voiceprints from the collection of various other voices. The main objective of introducing artificial intelligence in the field of speaker recognition is to create a system that can verify a person's voice and can also be proficient in identifying a person by his voice. Speech signal contains the information of one's identity, gender, age, emotions, sex, health condition, etc., with some amount of ambient noise and redundant information. So, as raw data, we must give speech signals to the

SR system which may contain distortion or additive noise. Hence, processing techniques are applied in the SR system to nullify the redundant portion and the additive noise incorporated with the speech signal. A speaker recognition system comprises of two measures, namely Feature extraction and classification. There are numerous feature extraction methods available that are being used in the field of SR such as LPC, CLPC, MFCC. MFCC is the most widely used in SR as its performance comparatively better in extracting features from speech signals [46].

The other measure of SR is a classification that again has two sub-divisions i.e., the training phase and the testing phase. In the training process, an SR model is trained with the extracted features by the feature extraction process and then saved in the database. And in the testing phase, an unknown speaker's speech is given to the input of the model, and based on the stored database, a decision is made (In case of identification and verification).

By the application of deep learning in the field of speaker recognition drastically changed the scenario. And this in turn grabs the attention of the researchers and reviewers to have a comprehensive review on this field that has great progress as compared to the older models. Deep learning has provided many advantages including the ability of representation to produce and extract acoustic features from the voice samples and many more when compared with other techniques. With the help of deep learning, the field of speaker recognition touches the sky of progress due to its advancement as compared to the other existing techniques.

Fig. 7 shows different steps of a speaker recognition system, comprised of the pre-processing part, the feature extraction part, the classification part, and lastly the output part. Firstly, the raw input speech sample of the speakers are fed into the system. This speech signal comprises various additive noises, so in pre-processing part, these signals are filtered by removing the redundant part that comprises unnecessary noises and followed by this normalization of the signals to get it ready for the feature extraction process. Now, in the feature extraction process, the feature that is important or based on which the speech signal gets classified are being extracted. And now finally in the classification part, the model is trained with the help of extracted feature and at the end, further normalization is done to produce the required desired output.

Here, we have discussed the different modules of speaker recognition. Here we have listed the various work from literature, that gives a clear idea about the progress and availability of the different techniques for various modules of speaker recognition.

5.1 Pre-processing techniques in speaker recognition

In the presence of additive noise, it becomes difficult to extract appropriate features for the SR system. In the pre-processing step, we are required to remove this redundant part to make it compatible with the feature extraction process. So, the first and foremost step is to lessen the distortion in the speech signal to enhance the performance of the system. Numerous speech enhancement methods are used to improve the quality of the speech, but mostly spectral subtraction is used for speech denoising [47]. Also, filtering methods used to filter the raw speech by Adaptive Noise Canceller (ANC) and Savitzky Golay (SG) filters are reported in [39, 40]. The ANC is a type of filter that has two kinds of inputs, 1. *A primary input*: contains distortion form of the speech signal, and 2. *A reference input*: contains noise correlated with that of the primary signal.

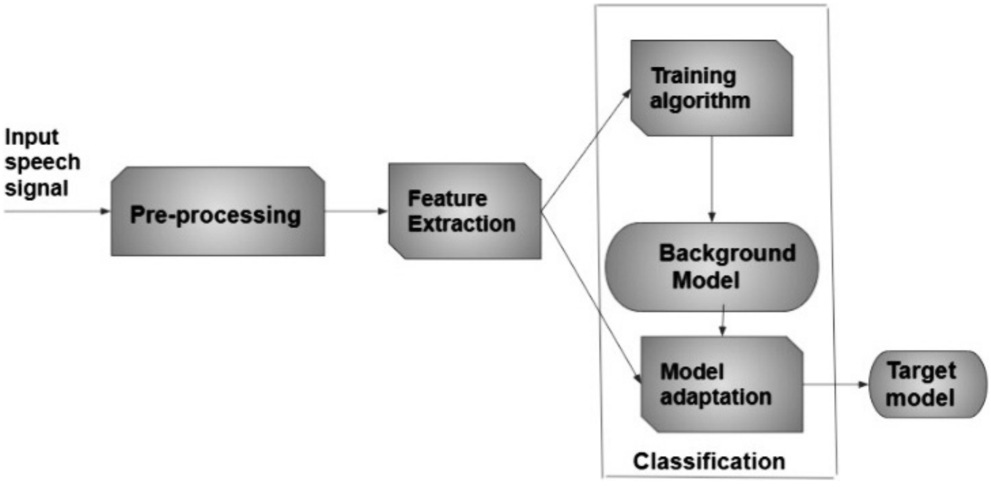

Fig. 7 Block diagram of speaker recognition illustrating pre-processing, feature extraction, and classification

Here, we have discussed the important pre-processing techniques that are used in speaker recognition and are illustrated in Table 4 [48–70]:

1. Method of silence removal: Technique to reduce the processing time and hence increasing the performance of the system. It also increases the performance of the particular system by eliminating the unwanted segments of the speech signal.
2. Method of pre-emphasis: It is defined as the process used to increase the magnitude of high frequency concerning a low frequency to improve signal quality. It boosts the energy of high-frequency signals to become stronger than that of the high-frequency noise components and directly helps to improve the signal-to-noise ratio.
3. Method of Hamming window: This process involves multiplying the ideal impulsive response with a window function which is used to generate a filter that will in turn tapers the ideal impulse response. These are widely used to design digital filters and conversion of an impulsive response having infinite duration to a finite duration impulse response. The whole process is called the window method.
4. End-point detection: Endpoint detection for speech is the process of detecting speech boundaries by digital processing methods. In a voice sample, the redundant pieces of information are mainly present before and after the actual speech region. This endpoint detection is broadly categorized into threshold-based and pattern-matching approaches. Acoustic features are extracted and compared with a predefined threshold to identify speech frames in the threshold-based approach. Whereas in pattern matching, speech and noise models are created and speech and non-speech classification are done based on these models.
5. Normalization method: It is the method of normalizing each feature on the same scale by performing a linear transformation on the original data. It is a very important step in speaker recognition technology where acoustic features, or the extracted features are normalized.
6. Spectrogram method: This process involves a visual representation of the strength and the characteristics of the signal over time. It gives a clear understanding in the form of a visual representation of the signal.
7. Method of recursive least squares: This process involves finding coefficients that will minimize a weighted linear least-squares cost function relating to the input signals.
8. Method of median filtering: It is a non-linear method that is used to remove noises and filter the unwanted signals from the speech. This method is implemented basically to remove the noise or unwanted speech signal or the redundancy from the informative speech signal.

5.2 Feature extraction process in speaker recognition

The feature extraction technique is usually used to extract valuable information from a raw signal which may have been incorporated with additive noise. This process is done mainly to discard the unwanted or redundant part from the signal and convert the raw acoustic speech signal into a desired informative signal that is needed for further processing in the case of speaker recognition [74, 75]. The quality of a speech and accuracy in speaker recognition degrades due to various factors such as background

Table 4 List of pre-processing techniques used in literature

Pre-processing methods	Pros	Cons	References
Silence removal	Reduction of processing time	Suits with white Gaussian noise	[48–57]
Pre-emphasis	Negative spectral slopes of voice parts are improved	Gain of low-frequency signals becomes high	[49, 54, 58–61]
Hamming window method	Reduces errors in case of distortion	Large variance in spectral estimation	[49, 59, 61–63]
End-point detection	Minimize the computational resources by incorporating spectral information	In low SNR it fails	[49, 64]
Normalization	Reduces errors that are due to change in the level of volume of speakers	Increment in the magnitude of jitter i.e., a precision measure for displacement estimation	[49, 50, 64–66]
Recursive least squares	Faster coverage can easily change into real-time systems, requires small memory	Intensive computation	[67]
Median filtering	Reduces noise	Not works well with multi-dimensional signals	[48]
Spectrogram	Speech is expressed in terms of time, frequency, energy in combined form	Resolution of time and frequency is very low	[53, 58, 61, 68–70]

noise, noise in the environment, variability in the utterance of speech by speaker due to illness or emotional state, noise in the transmission channel, etc. So, to overcome all this and extract the accurate features, it is needed to do a feature extraction process for speaker recognition. These features of a speech can be divided into various classes namely spectral-temporal, prosodic, source, and high-level features [76]. Various feature extraction techniques exist such as Linear Predictive Cepstral Coefficient (LPCC), Mel-frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction Coefficient (PLPC), etc. In Table 5 we have discussed the various aspects of different well-known feature extractors in detail [44, 71–73, 77–85].

In Section 5.2, we have discussed the applications of different features extraction techniques or approaches for speaker recognition. It is clear from the Table 6 [48, 51–54, 56–58, 60–63, 67–70, 73, 86–106], MFCC is the most widely used feature extraction technique that is being used mostly in the field of speaker recognition [107]. In Fig. 8 it is shown how these MFCC features extracted from the row input signal.

Fig. 8 shows the block diagram of the MFCC feature extraction technique. MFCC is a kind of feature extraction technique that aims to extract crucial features from a speech signal for further processing.

It involves the use of framing and windowing of the speech signal, then the executing fast Fourier transform followed by taking the logarithm of the magnitude of the speech signal. And after that, the speech signals are being wrapped with frequencies on a Mel scale and then applying the inverse of DCT (Discrete cos transform) to get the feature vector. Velocity and acceleration information can also be obtained by taking delta and double delta from the extracted features.

5.3 Classification process in speaker recognition

After the feature extraction classification is performed where the extracted features are applied to the classifier, the extracted features are being compared with the stored features. Based on the comparison, a particular classifier is used to detect the voiceprint or speech of a particular speaker. Broadly, the classifier is divided into two categories i.e., supervised when deal with training data set and unsupervised when doesn't deal with training data set.

Nowadays, classification technique in the field of speaker recognition is fully based on the statistical approach. The selection or the choice of a particular classifier depends on the feature used that are being extracted in the course of the feature extraction process. Table 7 [51,

Table 5 Different characteristics of feature extraction techniques

Feature extraction techniques	Performance with the noisy dataset	Implementation complexity	About	Extraction method used	Accuracy rate
PLP	Poor performance due to spectral balance	Moderately complex	Removes unwanted noise and increases the recognition rate	Combined method of spectral analysis and linear prediction [71]	Better than MFCC and LPCC [72]
MFCC	Poor performance [73]	Less complex	Widely used specifically in the bandwidth of the human ear	Dynamic method [73]	92% [44]
LPC	Poor performance [73]	Less complex [73]	Used in recognition of sounds and its extraction is done in lower rates	Static method [77]	Good accuracy rate and reliability [44]
LPCC	Poor performance [73]	Simple [73]	Used mainly in cepstral field or domain	Autocorrelation analysis [73]	88% [78]
PCA	Poor performance	Moderately complex	Don't remove noise completely. Based on Eigenvectors and it reduces the component features	Non-linear method [73]	54.66% [79]
RASTA	Good performance [73]	Moderately complex	Low-frequency modulations are captured and extract features in highly noisy data	Non-linear compression [80]	Robust technique with high accuracy [72]
Codebook quantized spectral entries	Moderate performance	Moderately complex	The approximate location of spectrum in acoustic space is taken as features	Quantization either by a VQ codebook or a GMM	Moderate accuracy [81, 82]
Pitch and energy	Poor performance	Simple	Pitch and energy of the signal is used as the feature	Learn gestures by modeling the joint slope dynamics of pitch and energy	Low accuracy [83, 84]
Prosodic statistics	Poor performance	Simple	For long speech segments various measurements like energy, duration and pitch are estimated	Prosodic idiosyncrasies of individual speakers is extracted	Low accuracy [83, 85]

Table 6 Various feature extraction processes used by different authors

Feature extraction methods used	References
MFCC, Spectrogram, Log-Mel, Filter Bank	[58]
MFCC, Spectrogram, MFSC	[69]
MFCC, Spectrogram	[61, 68, 86, 87]
MFCC, Delta MFCC, Delta-Delta MFCC	[88, 89]
MFCC, Spectral Roll-off, Roughness, Brightness, Irregularity, ZCR	[90]
MFCC, LFCC, LPC, ZCR, Spectral Roll-off, Roughness, Brightness, Irregularity	[91]
MFCC, Delta MFCC, Delta-Delta MFCC, GFCC	[63, 92]
MFCC, LPCC, DGS, DGCS	[60, 73]
MFCC, LPC, LBP	[93]
MFCC, Spectral Roll-off, Brightness, Roughness, Irregularity	[94]
MFCC, ZCR, Spectral Centroid, Spectral Entropy, Spectral Flux, Spectral Spread, Spectral Roll-off, Energy, Entropy of Energy, Chroma Deviation, Chroma Vector	[95]
MFCC, LPC	[96]
MFCC, DWT, WPT, WSBC	[97]
MFCC, LPCC, LPC residual, phase	[98]
MFCC	[63, 99]
MFCC, GFCC	[52, 57]
MFCC. LFCC	[51]
MFCC, Spectral and Cepstrum periodicities	[62]
MFCC, Delta MFCC, Delta-Delta MFCC, Spectrogram	[100]
MFCC, CFCC, GFCC, RASTA, RASTA-PLP	[101]
MFCC, Delta MFCC	[102]
MFCC, BFCC, PLP, RASTA-PLP	[103]
MFCC, LPCC	[104]
MFCC, IMMFC	[105]
MKMFCC	[59]
Haman Transform (HT), T21S	[54]
FBLPCS	[56]
Spectrogram	[53, 70]
Statistical features, Gabor filter, Spectrogram	[67]
WPT, DWT	[106]
LPC, DWT	[48]

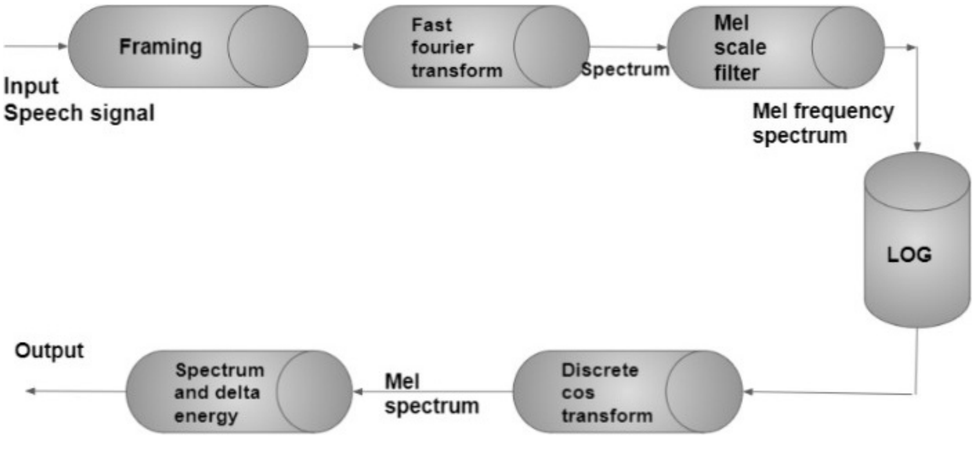

Fig. 8 MFCC feature extraction technique

Table 7 Different classification techniques for speaker classification

Method	Strengths	Weaknesses	References
DT	Very fast for forecasting unfamiliar records and easy to design disregards inappropriate features	Easy to overfit and huge Decision Trees can be challenging to understand, small dissimilarities in training data may produce huge variations in decision logic	[67, 108, 109]
NB	Requires a small set of training data to estimate the parameters essential for the forecast	It cannot learn the connection among the features	[108, 109]
SVM	It is memory effective and performs fine with non-overlapping target classes	Not appropriate for large databases and underperform once the number of features is greater than the number of training samples	[67, 93, 108]
K-NN	No training data is needed before making predictions. So, it is quite faster than other classifiers	It is not appropriate for high dimensional and categorical features due to difficulties in finding the distance in every dimension. And it also has a high classification cost	[93, 95, 108]
GMM	It requires a smaller number of parameters for training and can be precisely estimated by implementing the expectation maximization	It needs enough data to classify the speaker	[51, 66, 89]
ANN	It gives robust and efficient methods to learn feature representation automatically from complex data	It needs massive training data and has a possibility to stuck at the local optima and challenging to build an unambiguous model.	[66, 89, 110]

66, 67, 89, 93, 95, 108–110] illustrates the different classification techniques with their strengths and weaknesses used for speaker classification.

6 Deep learning in speaker recognition

In the last decade, deep learning techniques show immense improvement in various machine learning applications, now it replaces almost all state of art machine learning techniques in different fields. Deep learning has intensely transformed speech technology and then speaker recognition. It can produce highly abstract embedding features from sounds, which is the key benefit of deep learning over conventional methods. Due to these influential feature extraction capabilities, a lot of deep learning-based speaker recognition methods were developed [111–113]. In speech recognition, deep learning illustrates its great success in an uncontrolled environment, encouraged by the success researchers implemented deep learning in speaker recognition [114, 115]. With the easily available software and affordable hardware, it become very popular in the field where machine learning is applicable. The same trend can be observed also in speaker recognition (SR).

6.1 Feature extraction in deep learning

In speaker recognition, deep learning is helpful and extensively used for feature extraction and classification. In deep learning models, features can be extracted either from the conventional hand-crafted methods (like MFCC, F-BANK, etc.) or with the help of convolution operation from the raw speech signal. In classification, speaker models are built with the help of extracted features, and with a probabilistic approach decision are made. The deep feature extraction method [116] uses multiple subsets of autoencoder originally proposed in [117] to extract bottleneck features. Here a hybrid learning strategy has been adopted where the weights of the middle layer are shared across multiple adjacent frames by a cost function. Some of the popular deep feature extraction techniques are discussed here.

6.1.1 The *d*-vector

Many researchers had tried to extract the hidden layer character of DNN to use these as features. A method proposed in [112] consists of multiple fully connected layers and the average of the activations of the last hidden layer are taken as features, which is called ***d***-vector (shown in Fig. 9). Cosine distance comparison has been adopted for speaker verification. Supervised learning has been used to train the model using 13-dimensional perceptual linear predictive (PLP) features with Δ and ΔΔ coefficients. Then by removing the output layer the activations from the last hidden layers are taken as features.

6.1.2 The *j*-vector

In reality the individual speakers have their style of uttering syllables or words, so it becomes difficult to recognize a speaker directly from the speech sample. To address this issue, an extension of the ***d***-vector has been proposed in [118] which develops a multi-task learning setup using speaker ids and texts. Like the ***d***-vector, the output layer is removed after the supervised learning and the output of the last hidden layer has been taken as a feature vector, defined as a ***j***-vector (joint vector). The overview of the ***j***-vector is shown in Fig. 10.

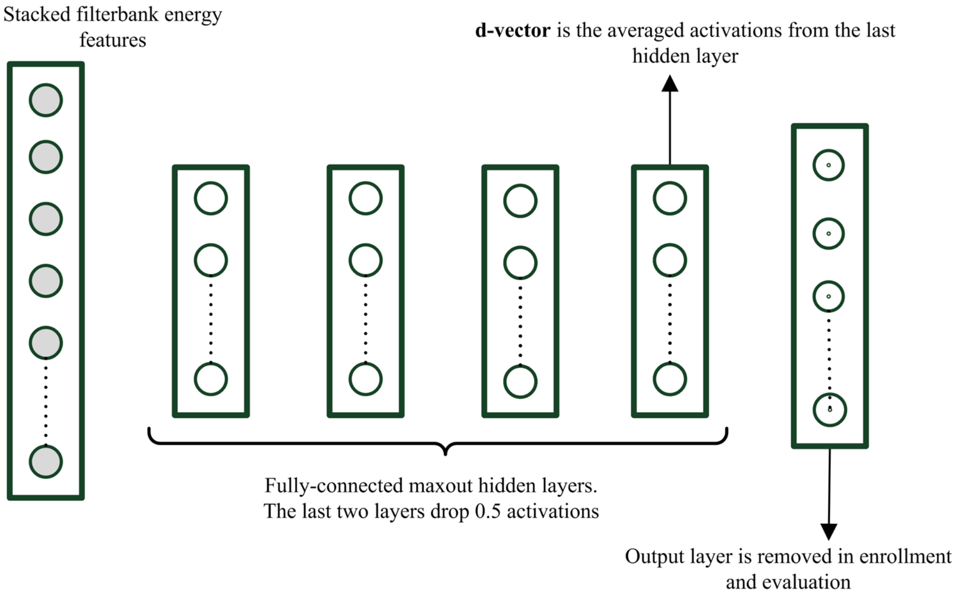
Stacked filterbank energy features
d-vector is the averaged activations from the last hidden layer
Fully-connected maxout hidden layers. The last two layers drop 0.5 activations
Output layer is removed in enrollment and evaluation

Fig. 9 The ***d***-vector model adapted from [112]

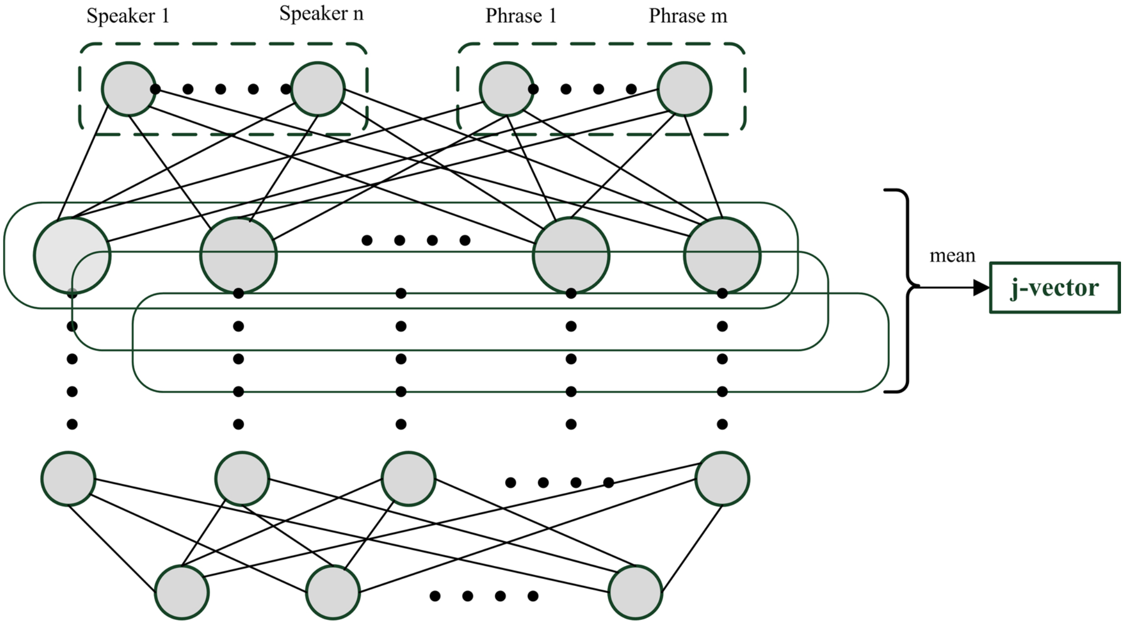

Fig. 10 The ***j***-vector model adapted from [118]

6.1.3 The *x*-vector

The ***x***-vector uses DNN embedding to learn the model characteristics. It employs multiple layered DNN architecture with fully connected layers to extract different frame-level temporal information. As it uses a wider temporal context, the architecture is referred to as a Time-Delay Neural Network (TDNN). The hidden layer extracted feature vector, ***x***-vector model [113] and its TDNN embedding architecture is shown in Fig. 11. To change the identity, speaker representations are used in [119]. Here ***x***-vectors are used for speaker anonymization. The extracted vector values are modified by re-synthetizing to generate anonymized speech to change the speaker characterization. In order to use in a multi-task learning scenario [120], after learning speaker identities, the model also learns higher-order statistics of the input vector and it gives better performance over the standard ***x***-vector. Due to the classification loss over training speakers, ***x***-vectors cannot yield maximum advantage from unlabeled utterances.

6.1.4 End-to-end systems

To obtain speaker representation vectors an end-to-end solution for speaker verification with deep networks has been proposed in [121]. Here, a speaker model is estimated by *N* enrolment utterances instead of using cosine distance or PLDA classification. DNN and LSTMs are applied for

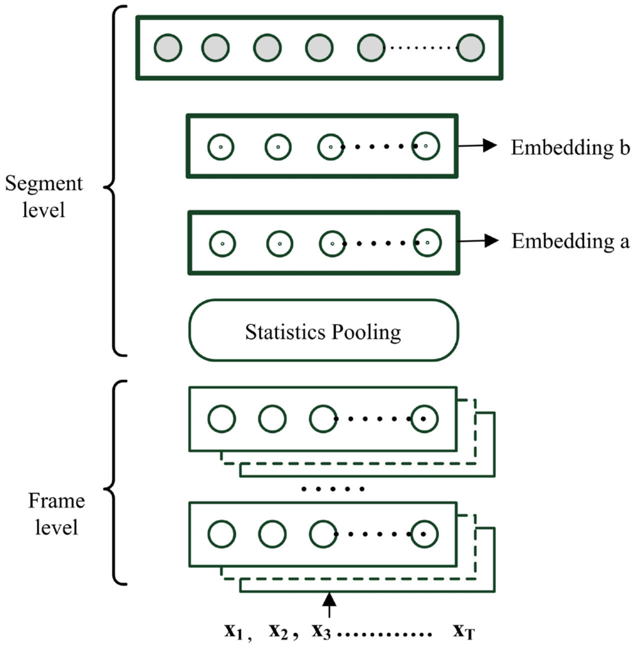

Fig. 11 The ***x***-vector DNN embedding architecture adapted from [113]

speaker representation and computation and the network is optimized by the end-to-end loss. The architecture is shown in Fig. 12. It is observed that the end-to-end architecture performs similarly to the ***d***-vector approach when the same feature extractor (DNN) is used. And LSTM performs better than DNN.

6.1.5 Deep Belief Networks

Another type of deep learning network is Deep Belief Networks (DBN) used in speaker recognition [86, 122], given in Fig. 13. These are generative models with several binary hidden layers. The neurons in the same layers are not linked with each other and adjacent layers are directionless. Training of DBNs is tough because the complexity of concluding the posterior distribution from the hidden layers. In [123] Stacked Restricted Boltzmann Machines (RBMs) is applied as a DBN architecture and shown in Fig. 13. Transfer Learning abstract hierarchical depictions of unlabeled speech input is the objective of DBN. In [122], to reduce dimensionality, spectrograms (25 ms frame size, 10 ms frameshift) have been applied as input after performing PCA transformation. With conventional MFCC features activations of the first and second layers of the RBM are appended to get the final feature set. After feature extraction, different classifiers can be used to perform speaker recognition.

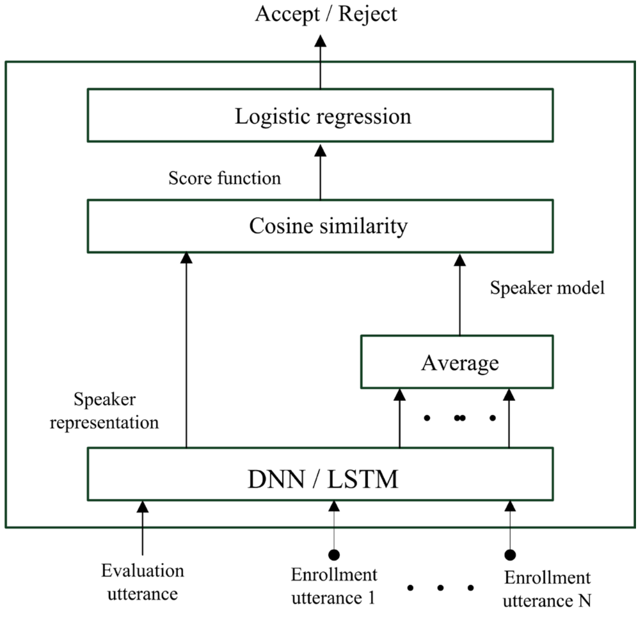

Fig. 12 End-to-end architecture adapted from [121]

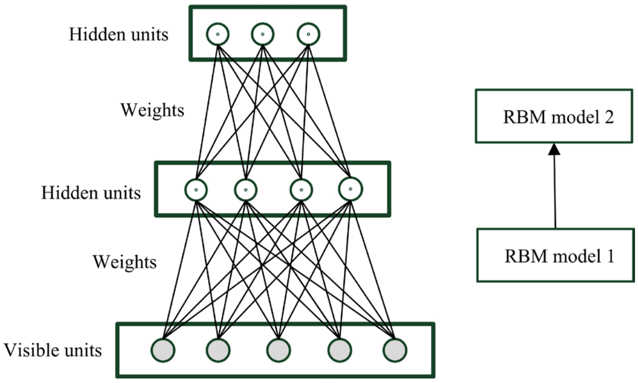

Fig. 13 Structure of the DBN used for extraction of short-term spectral features adapted from [122]

Table 8 [124–129] shows the few popular deep learning model implementations for speaker recognition with their performance.

6.2 Classification by deep learning

Apart from deep feature extraction to replace handcrafted, to further improve system performance, DNN is more robust and better than other scoring and comparison methods, like PLDA and cosine distance. Deep learning models can adopt several techniques. Here some of the important methods are discussed.

6.2.1 Variational autoencoder

Variational autoencoder (VAE) is a generative model for speech modeling and is proposed in [130, 131]. It belongs to the families of probabilistic graphical representations and variational Bayesian approaches. Statistical inference problems are rephrased as statistical optimization problems in Variational autoencoders to find the parameter values that minimize some objective function. VAE

Table 8 Performance of different classification techniques

References	DNN modelling technique used	Recognition rate (%)
[124]	DNN + HMM	13.57 WER (Word Error Rate)
[125]	DNN + HMM	17.53 WER
[126]	DNN	86.19% FDA (Feature Detection Accuracy)
[127]	DNN	59.90% WA (Weighted Accuracy)
[128]	DNN	82.00% WA
[129]	DNN	10.63% WER

maps input variables to a multivariate latent distribution. This model was proposed to use for unsupervised learning, but its usefulness has also been proven for semi-supervised learning and supervised learning.

6.2.2 Multi-task recurrent model

The conceptual design of the multi-task recurrent model is shown in Fig. 14. Here the output of one task at the present frame is used as a piece of supplementary information to supervise other tasks at the time of processing the next frame. After implementing it as a computational model, other many changes need to be carefully considered. In [132], the recurrent Long Short-Term Memory (LSTM) model has been realized to build automatic speech recognition (ASR) component and Speaker Recognition Evaluation (SRE) component. The two components implemented here are identical and accept the same input signal with an exception, one is trained for speaker identification discrimination and the other for phoneme identification. Here some inter-task recurrent links are present that interconnect the two components in a single network and the advantages of individual networks can be enchased in a combine network.

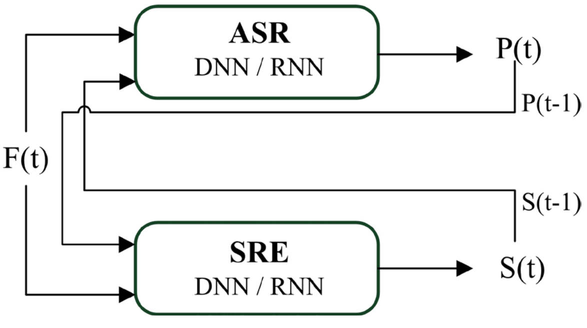

Fig. 14 Multi-task recurrent learning for ASR and SRE adapted from [132]

6.2.3 Deep learning backend with *i*-vectors

Inspired by the success of ***i***-vector and deep learning (DL) techniques in their respective field of applications, the [133] combined both techniques for robust speaker recognition. Here DL parameters have been used to create a background model and able to replace the UBM. A two-class hybrid DBN-DNN model has been trained for the individual target speaker to increase the discrimination between target speakers ***i***-vector/s and the other speakers ***i***-vectors. The Universal Deep Belief Network (UDBN) is initialized with speaker-specific parameters adapted from a global model. Then using the back-propagation algorithm the cross entropy between the class labels and the outputs is minimized. The train/test phases of this method are shown in Fig. 15. The impostor samples for target speakers are first identified by the impostor selection algorithm. The number of impostor samples are reduced, and the number of target ones are augmented reasonably and effectively by the balanced training block. For each target speaker model, the selected impostors are clustered, and the cluster centroids are taken as final impostor samples. To maintain impostor and target data in balance, impostor centroids and target samples are divided into equal mini batches. The DBN adaptation block is introduced to recompense the lack of input samples. The whole background ***i***-vectors are used to build a UDBN as unlabeled

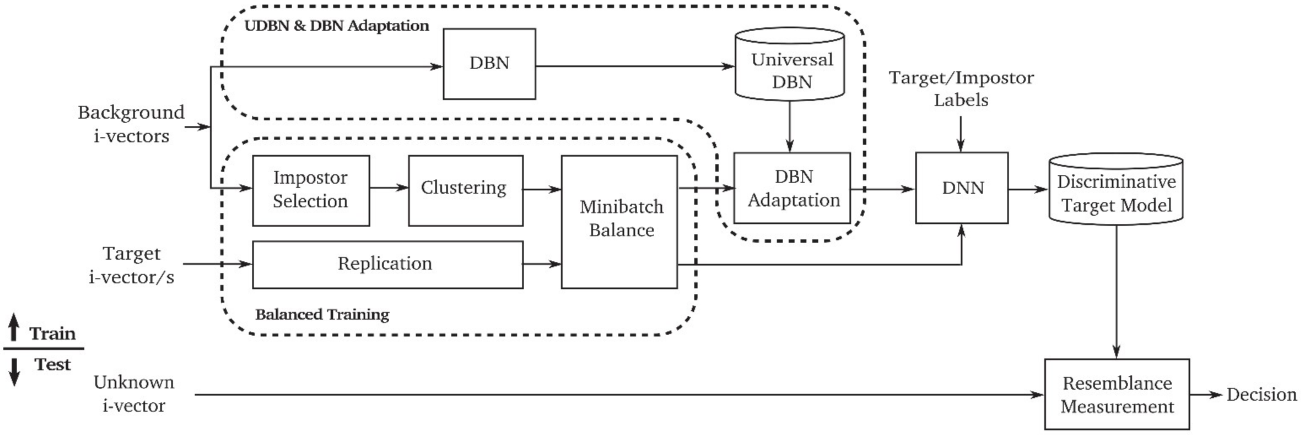

Fig. 15 Block diagram of the train/test phases of the deep learning backend for ***i***-vectors adapted from [133]

samples are sufficient for DBN training. For each target speaker, UDBN parameters are then adapted to the balanced data. The DNN is discriminatively trained for each target speaker by the adapted DBN and the balanced data to the target/impostor labels.

6.2.4 Using contrastive loss for vector comparison

In speaker identification, SoftMax layers are useful to generate a DNN backend system as it is a straightforward classification task. But for speaker verification, the comparison of two vectors is essential. To realize this, a general discriminative method for learning complex similarity metrics has been proposed in [134]. This method is useful where the number of classes are too large, and samples of all the classes are not available at the time of training. Here a loss function is defined and by minimizing this function the system approaches the anticipated behavior. The loss function is discriminative by driving the system to make the right decision without any probabilistic estimation. Here a convolutional network architecture has been chosen that exhibits robustness to variations of the input and reduces the need for an accurate representation of the data.

6.2.5 SincNet model

A special CNN architecture, called SincNet model [135] was introduced for speaker recognition. Instead of using any handcrafted features, it uses raw speech signals to train the model. The authors proposed a unique SincNet layer, which extracts more meaningful and effective features. In this approach first CNN layer was replaced by SincNet layer, which is a compact method of feature extraction. The SincNet layer learns the lower and higher frequencies of the band-pass input signal. These bandpass cut-off frequencies are learned by convolving Sinc function with speech waveform. Here only lower and higher cut-off frequencies are learned. This reduces the burden of learning nonessential (noisy and incongruous) features. The system learns only the meaningful parameters, which makes the model elegant and lightweight. By gradient-based optimization technique, cut-off frequencies of the filters can be optimized together with other CNN parameters as whole processes involved in SincNet are fully differentiable. After the first Sinc-based convolution a standard CNN pipeline (pooling, normalization, activations, dropout) has been employed as shown in Fig. 16. After that multiple standard convolutional and fully connected were introduced to accomplish speaker classification with a SoftMax classifier. This model has three distinct advantages over conventional models and is fast convergence, few learnable parameters, and high interpretability. Considering the mutual information as an objective function for embedding vector comparison, this model was extended for unsupervised speaker embedding learning as proposed in [136].

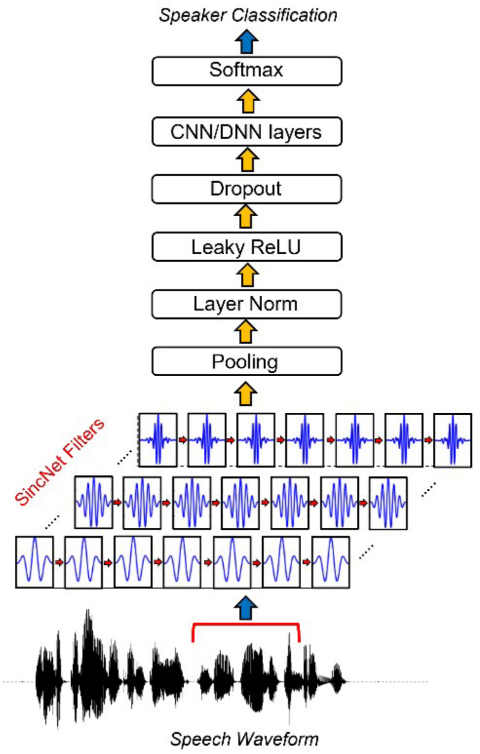

Fig. 16 Architecture of SincNet adapted from [135]

6.2.6 DNN/*i*-vector hybrid framework for speaker recognition

An automatic speech recognition (ASR) deep neural network (DNN) system for speaker recognition has been proposed in [111]. The flow diagram of this method is shown in Fig. 17. Here, a DNN trained for ASR was used to extract necessary statistics for the ***i***-vector model. This

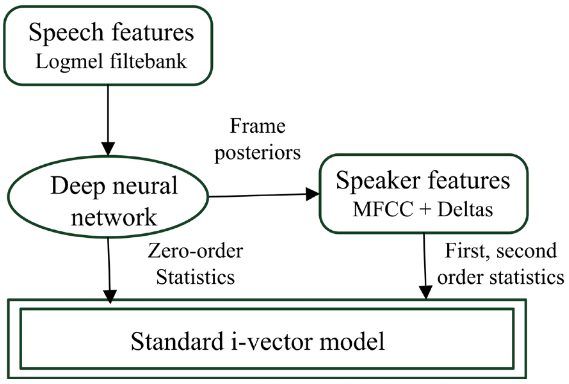

Fig. 17 Flow diagram of the DNN/***i***-vector hybrid framework adapted from [111]

model provides large enhancements over other state-of-the-art methods by substituting the UBM-GMM to yield frame alignments. To calculate the posterior of the frames for each of the classes in the model, DNN substitutes the GMM. The DNN classes are senones (tied triphone states) and are obtained using a standard Decision Tree for automatic speech recognition. After computing posteriors, the zeroth and first-order statistics are computed before they are fed into the ***i***-vector/PLDA. With the original backend, adequate statistics are then extracted using these posteriors to calculate ***i***-vector. The authors claimed that the DNN approach significantly improved the ***i***-vector speaker recognition system performance as related to the UBMGMM based model. As the ASR-DNN system integrates the information from speech content directly into the statistics keeping standard backends unchanged, it becomes a good option to implement in speaker recognition.

6.2.7 Speaker classification unlabeled data

To add extra dimensionality to speaker recognition from labelled to unlabeled data, many researchers implemented some models which can be able to learn through the samples are unlabeled. To increase the performance of the ML system, the DNN model must be trained with a large amount of data. these data may not be always properly labelled as the samples are taken from various sources under different conditions. Few methods use large-scale unlabeled data for training by curriculum learning [137, 138]. Here, the DNN models start learning from labelled data and sequentially unlabeled samples are fed to the model. TDNN and LSTM based models are supposed to give better results compared to baseline models.

7 Difficulties in speaker recognition

In real-life scenarios, speech processing systems must operate under varied data conditions. The estimation of speech quality under degraded data conditions is a very challenging task [139]. So, the system should be robust to different acoustic factors such as speaker differences, transmission channels, and background noise under diverse data conditions. Most speech applications perform digital filtering to enhance classification performance by passing noisy utterances through a linear filter to get a clean utterance approximation. The fundamental idea is to reduce noise by designing an optimum filter without significant loss of useful speech and speaker evidence. In this direction, several studies were conducted to minimize the effects of the environment so that speech signals can be classified correctly. In [140], a spectral subtraction method was proposed that overlaps a small portion of background noise over the speech signals. The components that are equivalent to noise were hidden in speech signals. However, spectral subtraction removes some valuable features of the original speech signal [141]. Support Vector Machine (SVM) [142] overcome this issue by classifying speech features into several classes. The difference among speech features of the same class is minimized to enhance classification accuracy. This method usually requires many training data and is not suitable for real-time applications.

7.1 Signal quality subject to practical applications

The efficiency and effectiveness of speaker recognition systems are depended on the quality of the data used to develop the system. Many experiential studies have shown that unwanted signals (like noise) intensely decrease prediction accuracy. In prediction applications, the problem of identifying and handling noise has drawn the large attention of researchers in the last decade. This noise can impact the performance of speaker recognition systems in terms of accuracy, complexity, and time requirement. For real-world data, sources of noise are from attributes and class labels which set a massive impact on system performance. The data collection process may be inaccurate or incompatible with the signal processing applications due to unreliable data collection instruments and human error. As a result, the collected data contains a significant amount of redundant portion/information which results in misinterpretation problems. This makes the whole system unreliable for unknown test data. When we deal with real-life data, it is observed that the samples are get affected by several components, noise is a key factor particularly [143]. And for speech signals, non-speech segments are the most important factor to be considered in designing a system because it is also the unwanted portion of the signal as noise. The source of this problem is like that of noise, like errors added by transducers and errors at the time of data collection, such as the time between on-off the transducers. The performance of any classifier not only depends on the training data conditions but also largely depends on the non-speech regions. To achieve maximum accuracy by a classifier, along with a learning algorithm, the data condition is an equally important parameter. Corruption in data in any form results in improper learning of parameters which leads to the degradation of system performance.

7.2 Identification of the desired segments in signal

The performance of a system mainly depends on input signal conditions. The input signal contains redundant parts like background noise, chirping noise, background speech, channel mismatch, sensor mismatch, and other environmental conditions, etc. along with valuable information. The redundant information (non-speech regions) is primarily present before and after the region of interest and it drastically degrades the performance of many speech-processing applications [144–146]. It is essential to remove these unwanted portions to improve the system performance of a speech-based application. The removal of these unwanted parts by identifying the speech boundary is called end-point detection. Precise end-point detection reduces the computational complexity by optimizing the dimensionality of speech signal [83], which is very much essential in applications like speech and speaker recognition.

In speaker recognition, classification is a complex problem in presence of such nonspeech regions, and finding an appropriate solution is often difficult but very essential for the proper learning of a model. When these non-speech data are present, it affects the intrinsic characteristics of a classification problem and introduces new properties in speaker recognition. The non-speech regions can be responsible for the creation of small clusters of examples of a particular speaker class in areas that corresponds to another speaker class. And it may also cause desertion of examples located in key areas within a specific speaker class. The overlapping between two speaker classes is caused by the non-speaker embeddings. Due to these various reasons, knowledge extraction from the data becomes difficult and pampers the models trained on non-speaker embeddings. Data collection in the real-world scenario can never be perfect and by the presence of non-speech regions, the performance of the system will be affected in terms of classification accuracy, training time, size, and interpretability of the classifier.

7.3 Complexities in model learning

In speaker recognition, many studies have been done on extracting more and more meaningful speaker embeddings to improve the system's performance. These models perform quite satisfactorily under clean and explicit data conditions. When the input speech samples have noise and non-speech segments, the functioning of the system drastically drops down. For real-world scenarios, class noise is more significant as compared to attribute noise for classification problems. The same instances in a sample with different class labels and different classes with similar instances are referred to as class noise. The non-speech regions in a speech sample carry the same properties when subjected to a classification problem. We have observed that these non-speaker embeddings play a crucial role in the operation of a system. This study emphasizes the non-speaker embeddings which are very essential for developing an effective model. But in this direction, very less importance given and sufficient research has not been done yet. In this study, we concentrate on this problem and analyses the effect of non-speaker embeddings on the speaker recognition process.

7.4 Non-speaker information in model learning

When machine learning models are trained with some speech samples, along with speech segments many non-speech segments are present. These non-speech segments may be silent, background/environmental noise, system-generated noise, etc. With the help of effective end-point detection techniques, the non-speech regions before and after the actual speech can be removed. In the presence of noise, proper endpoint detection is really a challenging task [147, 148]. Besides this, the non-speech segments are still present between the two speech segments. Eliminating these frames are near to impossible. When a machine learning model learns the parameters from the data, it gets confused with the speech and non-speech segments. The machine considers the non-speech segments as speech segments when it trains the model. This leads to errors in model hyperparameters, and the testing gives poor performance. This problem becomes more substantial in speaker recognition as ‘n’ numbers of speaker classes may be present in a complete system. The non-speech region characteristics are similar for all speaker utterances irrespective of the person’s identity. When a model learns a particular speaker pattern, the non-speech embeddings are the same for every utterance of all speakers. This creates an overlapping between all speakers and the extracted features are the same for each speaker in these regions. Because of this system gets perplexed as it cannot discriminate the speaker characteristics from these segments resulting misclassification problem. This problem becomes more significant when the speech samples are processed in a short frame. In the same sample, few frames are classified correctly but many frames are misclassified due to improper training. In the decision-making process, this creates a major issue in the final classification results. To overcome this problem the non-speech or non-speaker information should be handled carefully. To address this issue, we have introduced non-speaker information as embeddings to the model at the time of learning. The non-speaker embeddings are defined as some separate classes along with the speaker class. By doing this system can be able to learn this non-speaker information at the time of training. This will help the system to learn the speaker embedding more conveniently and the model can distinguish speaker and non-speaker effectively which in turn helps the system to recognize individual speakers more conveniently. This will increase the overall recognition ability of the model under variable data conditions.

8 Performance evaluation metric

In classification, system performance measures play an important role in critical study. So, several performance measures are used as evaluation criteria of classifiers. The confusion matrix estimation for evaluating classifier performance is the most widely used metrics in many applications. The confusion matrix representation is shows in Table 9 and is used to predict positive and negative instances. Speaker recognition system performance can be optimized by reducing "false positive" and "false negative".

In Table 9:

- T_{pos} is the true positive (instances where actual and predicted classes are produced correct).
- T_{neg} is the true negative (instances where actual and predicted classes are produced negative).
- F_{pos} is the false positive (instances, where the actual class comes out to be negative and predicted class, is positive).
- F_{neg} is the false negative (instances where actual class comes out to be positive and predicted class is negative).

The most popular evaluation criteria that are used in the field of speaker recognition for classification problem are accuracy, recall, *F*-measures, and precision, which are discussed below.

8.1 Accuracy

It is the ratio of correctly predicted instances to an entire number of instances present. It is widely used in the classification system and is the most usable evaluation criteria. The mathematical representation of accuracy is shown in Eq. (1):

Table 9 The confusion matrix

		Actual instances	
		Yes	No
Predicted instances	Yes	T_{pos}	F_{neg}
	No	F_{pos}	T_{neg}

$$\text{Accuracy} = \frac{\left(T_{pos} + T_{neg}\right)}{\left(T_{pos} + T_{neg} + F_{pos} + F_{neg}\right)} . \tag{1}$$

8.2 Precision

It is the ratio of the negative instances that are predicted to be negative. The precision rate and F_{pos} rate are inversely proportional to each other. And this is used especially when the scenario is to check for the exactness of the classification system as an evaluation criterion.

$$\text{Precision} = \frac{T_{pos}}{\left(T_{pos} + F_{pos}\right)} \tag{2}$$

8.3 Recall

It is the ratio of the positive instances when the accurate class and predicated class turns out to be positive or correct i.e., T_{pos} to the total number of positive instances. It is also said to be a True Positive Rate (TPR).

$$\text{Recall} = \frac{T_{pos}}{\left(T_{pos} + F_{neg}\right)} \tag{3}$$

8.4 *F*-measure

It is the weighted harmonic mean of Recall and Precision on the condition of extreme balance of F_{pos} and F_{neg}. The mathematical representation comes out to be as follows in Eq. (4):

$$F\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\left(\text{Precision} + \text{Recall}\right)} . \tag{4}$$

8.5 Equal Error Rate

Equal Error Rate (EER) represent the performance of a speaker recognition system by considering false acceptance Rate (FAR) and False Rejection Rate (FRR). The lower value of EER is always desirable as it indicates higher accuracy rate. The FAR, FRR and EER can be calculated by in Eqs. (5), (6) and (7):

$$\text{FAR} = \frac{F_{pos}}{\left(T_{neg} + F_{pos}\right)} , \tag{5}$$

$$\text{FRR} = \frac{F_{neg}}{\left(T_{pos} + F_{neg}\right)} , \tag{6}$$

$$\text{EER} = \frac{\left(\text{FAR} + \text{FRR}\right)}{2} . \tag{7}$$

8.6 Root Mean Square Error (RMSE)

The RMSE is a standard parameter, used to compute the error of a trained model for prediction of quantitative data. It is expressed by Eq. (8):

$$\text{RMSE} = \sqrt{\sum_{j=1}^{n} \frac{\left(\hat{x}_j - x_j\right)^2}{n}}\,, \tag{8}$$

where $\hat{x}_1, \hat{x}_2, \hat{x}_3, \ldots, \hat{x}_n$ are the predicted values and $x_1, x_2, x_3, \ldots, x_n$ are the observed values.

When the distribution of error is likely to be Gaussian, then RMSE is used to evaluate model performance.

8.7 Receiver Operating Characteristics (ROC)

The ROC is represented as the curve that expresses the relationship between FAR and FRR at different values. The ROC shows the performance of a classifier with respect to discriminative threshold. The FAR and FRR are data dependent, their values change with the size of dataset. The result of a classifier changes with the change in decision threshold.

9 Frameworks of DL implementation

The framework of deep learning implementation is very important in terms of training the deep learning neural network or DNN with the help of GPUs to gear up the training process in less time.

The intention of researchers of building various software frameworks in the field of deep learning to access different deep learning architectures, implement deep learning most easily, evaluate results efficiently. As deep learning techniques become popular in today's world, it gained a lot of attention from the researchers. The researchers and developers give much attention to develop an effective framework to properly utilize the advantages of deep learning with great ease. There is a list of the most widely used frameworks in the field of deep learning which are discussed below in Table 10 [149–165].

10 Application of speaker recognition system

Some of the applications of speaker recognition systems are discussed below:

1. In surveillance:
   Security agencies use speaker recognition systems in electronic telephone and radio conversions to collect information regarding various aspects and it requires good quality data, filter mechanisms [166].
2. In authentication:
   In this area, it is used to match or identify different voiceprints of different speakers based on different features present in respective voice prints like voice quality, utterance rate, etc. [166, 167].
3. In forensic:
   It is having great importance in this sector. A recorded speech sample can be matched with that of the suspect's voiceprint to recognize whether the person is a criminal or not [166].
4. Insecurity:
   This is used for security measures where it is incorporated into the whole authentication process such as in biometric authentication, in the transaction process, and many more [166, 167].
5. In multiple speaker tracking:
   In multi-speaker tasks, there are 3 processes, namely speaker detection, speaker tracking, and speaker segmentation. In detection presence of a particular speaker is checked, in tracking speakers speaking intervals are recorded, and finally in segmentation intervals of different speakers are located such as in conference calls. And this speaker segmentation is used in news broadcasts [167].
6. In personalized user interface:
   Here, it is used in the case of voice mails where the speaker on the other side can be accepted by the system according to his or her voice mails based on reference [166].

11 Databases used in SI and SV

As in many technologies, the choice of databases is very crucial. Numerous databases are used in the field of speaker recognition. It is very hard to say that whether an approach will perform better as it is provided with a different dataset. Due to these various data conditions are considered by selecting the training and evaluating or testing database. There are a huge number of databases that are used in this field with different properties that can be found publicly either free or paid. It enables us to use many datasets that are present in the database during the training phase. There are various kinds of databases available in this field, and a list of the databases that are used in speaker recognition specifically for speaker identification and speaker verification domain [114, 168–170] are given in Table 11 [58, 62, 89, 90, 103, 168–205].

12 Future scopes and challenges

In recent days with the increase in technological development, there is a great demand for the integration of biometrics into different applications. The methods of speaker verification and speaker identification are growing very

Table 10 Different frameworks of DL implementation

Software/ Toolbox	Supportive language	License/ Access	GPU	Platform that supports	Supporting DL techniques	Optimizer used	Activation function used
Deep learning toolbox	MATLAB	BDS [149]	Yes	Windows, LINUX	FCNN, SAE, DAE, CNN, DBN	BP	Sigmoid
Keras	Python	MIT [150]	Yes	Windows, LINUX, Apple	CNN, RNN	SGD ADAM RMSprop	ELU, ReLU SoftMax SeLU, Tanh Sigmoid
Theano	Python	Numpy [151]	Yes	Windows, LINUX, Apple	FCNN, DAE, CNN, DAN	SGD	ReLU, Tanh SoftMax
Tensorflow	Python, C++	Apache [152]	Yes	Windows, LINUX, Apple	CAE, CNN, RNN	SGD BP	ReLU, Sigmoid
Caffe	Python, MATLAB, C++	BSD [153]	Yes	Windows, LINUX, Apple	CNN	SGD	ReLU, Tanh Sigmoid, ELU
MatConvNet	MATLAB, C++	BSD [154]	Yes	Windows, LINUX, Apple	CNN	SGD	ReLU, Sigmoid
SimpleDNN	Kotlin	Mozilla [155]	Yes	Windows, LINUX, Apple	FCNN, RNN	-	ELU, SoftMax
RNNLIB	C++	MIT [156]	Yes	LINUX, Apple	RNN	-	-
CNTK	C++, C#, Python, Java	BSD [157]	Yes	Windows, LINUX	CNN, RNN	SGD	ReLU, Tanh ELU
PyBrain	Python	BSD [158]	Yes	Windows, LINUX, Apple	FCNN, RNN, RBM	BP	Sigmoid
Torch	Lua C	BSD [159]	Yes	Windows, LINUX, Apple	FCNN, AE, CNN	SGD ADAM	ReLU, Tanh SoftMax
Neural Networks	Java	MIT [160]	Yes	Windows, LINUX, Apple	FCNN, SAE, DAE, CNN, RBM DBN	-	ReLU, Tanh SoftMax, LRN Sigmoid
ConvNet	MATLAB	BSD [161]	Yes	Windows, LINUX	CNN	IBP	ReLU, Sigmoid SoftMax
OpenNN	Python, C++	GNU [162]	Yes	Windows, LINUX, Apple	FCNN	-	-
MDLTB	MATLAB	Information is not available	Yes	Windows, LINUX, Apple	FCNN, SAE, DAE, CNN, RNN	ADAM	ReLU, PReLU
DL4J	Scala Java	Apache [163]	Yes	Windows, LINUX, Apple	FCNN, AE, CNN, RNN	SGD ADAM RMSPrp	Tanh, ReLU SoftMax, ELU Sigmoid
Chainer	Python	MIT [164]	Yes	Windows, LINUX, Apple	FCNN, AE, CNN, RNN	SGD	ReLU
SincNet	Python	MIT [165]		Windows, LINUX, Apple	CNN	RMSprop	ReLU, SoftMax

rapidly for their excellent prospects in the fields of banking & technological security, disease detection, voice detection, biometric authentication, speaker recognition, virtual voice assistants, etc. Researchers and scientists have come up with numerous methods of speaker recognition. The modern state is very complex and mature hence hard to put it in simple run-throughs. Nowadays, the advancements of deep learning (DL) are steadily increasing because of its easy availability and low-cost software and hardware integrated environment. To improve performance and accuracy with more complex and challenging datasets, scientists started using deep learning methods in the field of speech processing and speaker recognition. Mobile phones and computers are the platforms of applications for digital assistants like Apple's Siri and Cortana, nowadays this has changed the whole scenario drastically. Now people have become more interested in voice-activated home speakers like Amazon Echo(Alexa), Google Home, and Samsung's Bixby, etc. The applications like Google Home or Amazon Echo can be easily controlled with a vast category of Internet of Things (IoT) devices. The IoT devices include smart TVs, smart fridges, headphones, and smoke alarms, along with a large and increasing list of third-party integrations and customization. Biometric security techniques like speaker recognition in smart home, and office devices play an important role in the safety and security of home and office appliances.

Table 11 List of databases used in SI and SV

Database	SV/SI	Year	Speaker (M/F)	Sessions	Language	Environment	Age info	Access	Sampling frequency (kHz)	Format	References
YOHO	SV/SI	1995	138 (106/32)	14	English	Quiet	No	Paid	8	WAV	[171, 172]
BT-DAVID	SV	1996	31(15/16)	5	English	Quiet	Yes	Paid	16	WAV	[173]
M2VTS	SV	1997	37(30/7)	5	English	Quiet	No	Paid	16	WAV	[174]
PolyVAR	SV	1997	143(85/58)	1–129	English	Quiet	Yes	Paid	16	WAV	[175]
OGI Speaker Recognition	SV	1998	91(43/48)	12	English	Quiet/Noisy	Yes	Paid	16	WAV	[176]
XM2VTS	SV	1999	295(158/137)	4	English	Quiet	No	Paid	8	WAV	[177]
Ahumada	SV	2000	104(104/0)	6	Spanish	Quiet	Yes	Paid	16	WAV	[178]
PolyCOST	SV/SI	2000	134(74/60)	5–14	English, European	Quiet	Yes	Paid	8	A-LAW	[179]
Verivox	SV	2000	50(50/0)	2	Swedish	Quiet	No	Paid	16	WAV	[180, 181]
Smart Kom	SV	2002	45(20/25)	2	German	Quiet	No	Paid	16	WAV	[182]
BANCA	SV	2003	208(104/104)	12	English, France, Italian, Spanish	Quiet/Noisy	No	Paid	16	WAV	[183]
BIOMET	SV	2003	91(45/46)	3	France	Quiet	Yes	Paid	16	WAV	[184]
STC	SV	2003	89(54/35)	1–15	Russian	Quiet	No	Paid	16	WAV	[185]
MyIdea	SV	2005	30(30/0)	3	English, France	Quiet/Noisy	No	Paid	16	WAV	[186]
Valid	SV	2005	106(79/30)	5	English	Quiet	Yes	Paid	16	WAV	[187]
CCC-VPR2 C200510000	SV	2006	10000(-/-)	2	Putonghua	Quiet	No	Paid	16	WAV	[188]
MIT-MDSVC	SV	2006	88(49/39)	2	English	Quiet/Noisy	No	Paid	16	WAV	[188]
M3	SV	2006	39(29/10)	3	Cantonese, English, Putonghua	Quiet/Noisy	Yes	Paid	8	WAV	[189]
BIOSEC	SV	2007	250(-/-)	4	English, Spanish	Quiet	Yes	Paid	8	WAV	[190, 191]
BiosecureID	SV	2007	400(-/-)	4	Spanish	Quiet	Yes	Paid	8	WAV	[192]
MbioID	SV	2007	120(-/-)	2	English, France	Quiet	Yes	Paid	16	WAV	[193]
Biosecure	SV	2010	400(-/-)	2	European	Quiet	Yes	Paid	16	WAV	[194]
UNMC-VIER	SV	2011	123(74/49)	2	English	Quiet	No	Paid	16	WAV	[195]
RSR2015	SV/SI	2012	300(157/143)	9	English	Quiet	Yes	Free	16	WAV	[169, 196]
TIMIT	SI	2011	630(438/192)	1	English	Quiet	No	Paid	16	WAV	[168, 197, 198]
VoxCeleb 2	SI	2018	6112(3761/2351)	2	English	Quiet/Noisy	No	Free	16	WAV	[58, 199, 200]
ELSDSR	SI	2012	22(12/10)	1	English	Quiet	Yes	Free	16	WAV	[62, 89, 103]
CHAIN	SI	2015	36(16/20)	1	English	Quiet	No	Free	44.1	WAV	[90, 201, 202]
LibriSpeech	SV/SI	2015	1166(602/564)	1	English	Quiet	No	Free	16	WAV	[203]
IITG-MV SR (Phase I II III & IV)	SV/SI	2012	544(-/-)	Multiple sessions	English (IND) + 13 regional languages	Quiet/Noisy	Yes	Paid	16	WAV	[204]
NITS database	SV	2018	300(247/51)	Multiple sessions	English (IND)	Noisy	Yes	Paid	16	WAV	[205]

Although deep learning techniques attains a new height in the field of speaker recognition. Reverberation effect, microphone mismatches, language mismatches, short utterances, background noise, channel distortion, and other issues in practical applications have a great impact on performance and these need to be addressed effectively to be used for practical deployable systems. The denoising algorithms still have great limitations and handling multiple noises at the same time is a great challenge for researchers. Feature extraction and classification are the major challenges under such degraded data conditions. In this direction of the smart future, the challenge lies in improving accuracy, noise reduction, range, etc. At present solutions to the problems are scenario-specific and finding a generalized solution is the future challenge in speaker recognition. For any speaker recognition system, security is the major issue in development and the system must be robust against spoofing attacks. These problems are caused by replay, speech synthesis, voice conversion, and adversarial samples. The spoofing-aware speaker recognition [206] becoming popular to address such issues. To improve overall system performance, model compression techniques like network pruning [207] and knowledge distillation [208] can be implemented. In deep model design Transformer layer, few-shot learning, and attention-based network are the few options that can be introduced for robust speaker recognition. Achieving good system performance and runtime efficiency at the same time is the major challenge in developing a speaker recognition system. To address these issues lightweight deep learning models [209] and multimodal deep learning models [210] are good options to explore in the future.

13 Conclusion

In this paper, we reviewed the speaker recognition techniques, their applications and gives an idea about deep learning for speaker recognition. We tried to describe the speaker recognition process including speaker verification and identification and also describes the factors that affect the speech signal which provides uniqueness in voiceprints and various challenges regarding this field. So, our main objective behind writing this paper is to review the speaker recognition technique in the field of speaker identification and speaker verification. As till date research and development or advancement in the field of the speaker, recognition doesn't grab so much attention, so this paper has been written intending to give light on the field of speaker recognition for the upcoming world.

References

[1] Sreenu, G., Girija, P. N., Prasad, M. N., Nagamani, M. "A Human Machine Speaker Dependent Speech Interactive System", In: Proceedings of the IEEE INDICON 2004. First India Annual Conference, 2004., Kharagpur, India, 2004, pp. 349–351. ISBN 0-7803-8909-3
https://doi.org/10.1109/INDICO.2004.1497769

[2] Zhu, W., Pelecanos, J. "A Bayesian Attention Neural Network Layer for Speaker Recognition", In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6241–6245. ISBN 978-1-4799-8132-8
https://doi.org/10.1109/ICASSP.2019.8682953

[3] Shuai, G., Renkai, C., Tuo, H., Guodong, W., Yong, W. "A Convenient and Extensible Offline Chinese Speech Recognition System Based on Convolutional CTC Networks", In: 2019 Chinese Control Conference (CCC), Guangzhou, China, 2019, pp. 7606–7611. ISBN 978-1-7281-2329-5
https://doi.org/10.23919/ChiCC.2019.8865580

[4] Yadav, S., Rai, A. "Frequency and Temporal Convolutional Attention for Text-Independent Speaker Recognition", In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6794–6798. ISBN 978-1-5090-6632-2
https://doi.org/10.1109/ICASSP40776.2020.9054440

[5] Zhang, Y., Yu, M., Li, N., Yu, C., Cui, J., Yu, D. "Seq2seq Attentional Siamese Neural Networks for Text-Dependent Speaker Verification", In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6131–6135. ISBN 978-1-4799-8132-8
https://doi.org/10.1109/ICASSP.2019.8682676

[6] Huang, L., Pun, C.-M. "Audio Replay Spoof Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, pp. 1813–1825, 2020.
https://doi.org/10.1109/TASLP.2020.2998870

[7] Cao, G., Tang, Y., Sheng, J., Cao, W. "Emotion Recognition from Children Speech Signals Using Attention Based Time Series Deep Learning", In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2020, pp. 1296–1300. ISBN 978-1-7281-1868-0
https://doi.org/10.1109/BIBM47256.2019.8982992

[8] Zhou, T., Zhao, Y., Li, J., Gong, Y., Wu, J. "CNN with Phonetic Attention for Text-Independent Speaker Verification", In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, Singapore, 2019, pp. 718–725. ISBN 978-1-7281-0307-5
https://doi.org/10.1109/ASRU46091.2019.9003826

[9] Larcher, A., Lee, K. A., Ma, B., Li, H. "Text-Dependent Speaker Verification: Classifiers, Databases and RSR2015", Speech Communication, 60, pp. 56–77, 2014. https://doi.org/10.1016/j.specom.2014.03.001

[10] Bai, Z., Zhang, X.-L. "Speaker Recognition Based on Deep Learning: An Overview", Neural Networks, 140, pp. 65–99, 2021. https://doi.org/10.1016/j.neunet.2021.03.004

[11] Dişken, G., Tüfekçi, Z., Saribulut, L., Çevik, U. "A Review on Feature Extraction for Speaker Recognition under Degraded Conditions", IETE Technical Review, 34(3), pp. 321–332, 2017. https://doi.org/10.1080/02564602.2016.1185976

[12] Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., Wang, R. "Speaker Identification Features Extraction Methods: A Systematic Review", Expert Systems with Applications, 90, pp. 250–271, 2017. https://doi.org/10.1016/j.eswa.2017.08.015

[13] Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E.-D., Jin, W., Schuller, B. "Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments", ACM Transactions on Intelligent Systems and Technology, 9(5), 49, 2018. https://doi.org/10.1145/3178115

[14] Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., Shaalan, K. "Speech Recognition Using Deep Neural Networks: A Systematic Review", IEEE Access, 7, pp. 19143–19165, 2019. https://doi.org/10.1109/ACCESS.2019.2896880

[15] Jahangir, R., Teh, Y. W., Nweke, H. F., Mujtaba, G., Al-Garadi, M. A., Ali, I. "Speaker Identification through Artificial Intelligence Techniques: A Comprehensive Review and Research Challenges", Expert Systems with Applications, 171, 114591, 2021. https://doi.org/10.1016/j.eswa.2021.114591

[16] IEEE Explorer "Speaker Recognition using Deep learning", [online] Available at: https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=SpeakerRecognitionDeepLearning [Accessed: 17 December 2020]

[17] Springer Link "Speaker Recognition using Deep learning", [online] Available at: https://link.springer.com/search?query=Speaker+Recognition+Deep+Learning+ [Accessed: 17 December 2020]

[18] ACM Digital Library "Speaker recognition through deep learning techniques'", [online] Available at: https://dl.acm.org/action/doSearch?AllField=Speaker+Recognition+through+Deep+Learning+Techniques [Accessed: 17 December 2020]

[19] Science Direct "Speaker recognition using Deep Learning", [online] Available at: https://www.sciencedirect.com/search?qs=Speaker%20Recognition%20Deep%20Learning [Accessed: 17 December 2020]

[20] Web of Science " Speaker recognition using Deep Learning", [online] Available at: https://mjl.clarivate.com/search-results?issn=1443-1394,1918-2902,1479-4403,2000-7426,2076-8184,1541-5015,1863-0383,2538-1032,1492-3831,1835-5196&hide_exact_match_fl=true&utm_source=mjl&utm_medium=share-by-l [Accessed: 17 December 2020] .

[21] MDPI "Speaker recognition using Deep Learning", [online] Available at: https://www.mdpi.com/search?q=Speaker+Recognition+Deep+Learning+ [Accessed: 17 December 2020]

[22] Research Gate "Speaker recognition through deep learning techniques", [online] Available at: https://www.researchgate.net/search/publication?q=Speaker+Recognition+Through+Deep+Learning+Techniques [Accessed: 17 December 2020]

[23] Deng, L., Yu, D. "Deep Learning: Methods and Applications", Foundations and Trends® in Signal Processing, 7(3–4), pp. 197–387, 2014. https://doi.org/10.1561/2000000039

[24] Bengio, Y. "Learning Deep Architectures for AI", Foundations and Trends® in Machine Learning, 2(1), pp. 1–127, 2009. https://doi.org/10.1561/2200000006

[25] Bengio, Y., Courville, A., Vincent, P. "Representation Learning: A Review and New Perspectives", IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), pp. 1798–1828, 2013. https://doi.org/10.1109/TPAMI.2013.50

[26] Schmidhuber, J. "Deep Learning in Neural Networks: An Overview", Neural Networks, 61, pp. 85–117, 2015. https://doi.org/10.1016/j.neunet.2014.09.003

[27] Komar, M., Yakobchuk, P., Golovko, V., Dorosh, V., Sachenko, A. "Deep Neural Network for Image Recognition Based on the Caffe Framework", In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 2018, pp. 102–106. ISBN 978-1-5386-2875-1 https://doi.org/10.1109/DSMP.2018.8478621

[28] Arel, I., Rose, D. C., Karnowski, T. P. "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]", IEEE Computational Intelligence Magazine, 5(4), pp. 13–18, 2010. https://doi.org/10.1109/MCI.2010.938364

[29] Collobert, R. "Deep Learning for Efficient Discriminative Parsing", In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 2011, pp. 224–232.

[30] Gomes, L. "Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts", IEEE Spectrum, 20 October 2014. [online] Available at: https://spectrum.ieee.org/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts [Accessed: 08 August 2022]

[31] Monroe, D. "Deep Learning Takes on Translation", Communications of the ACM, 60(6), pp. 12–14, 2017. https://doi.org/10.1145/3077229

[32] Simhambhatla, R., Okiah, K., Kuchkula, S., Slater, R. "Self-Driving Cars: Evaluation of Deep Learning Techniques for Object Detection in Different Driving Conditions", SMU Data Science Review, 2(1), 23, 2019.

[33] Vailaya, A., Figueiredo, M. A. T., Jain, A. K., Zhang, H.-J. "Image Classification for Content-Based Indexing", IEEE Transactions on Image Processing, 10(1), pp. 117–130, 2001. https://doi.org/10.1109/83.892448

[34] Dorronsoro, J. R., Ginel, F., Sgnchez, C., Cruz, C. S. "Neural Fraud Detection in Credit Card Operations", IEEE Transactions on Neural Networks, 8(4), pp. 827–834, 1997. https://doi.org/10.1109/72.595879

[35] Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., van den Bosch, A. "Prediction during Natural Language Comprehension", Cerebral Cortex, 26(6), pp. 2506–2516, 2016. https://doi.org/10.1093/cercor/bhv075

[36] Freitas, A. R. R., Guimarães, F. G. "Melody Harmonization in Evolutionary Music Using Multiobjective Genetic Algorithms", presented at 8th Sound and Music Computing Conference, Padova, Italy, Jul., 6-9, 2011.

[37] Juang, B. H., Rabiner, L. R. "Hidden Markov Models for Speech Recognition", Technometrics, 33(3), pp. 251–272, 1991. https://doi.org/10.1080/00401706.1991.10484833

[38] Chung, J. S., Nagrani, A., Zisserman, A. "VoxCeleb2: Deep Speaker Recognition", [preprint] arXiv, arXiv:1806.05622, 27 June 2018. https://doi.org/10.48550/arXiv.1806.05622

[39] Reynolds, D. A., Quatieri, T. F., Dunn, R. B. "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 10(1–3), pp. 19–41, 2000. https://doi.org/10.1006/dspr.1999.0361

[40] Durmuş, H., Güneş, E. O., Kırcı, M. "Disease Detection on the Leaves of the Tomato Plants by Using Deep Learning", In: 2017 6th International Conference on Agro-Geoinformatics, Fairfax, VA, USA, 2017, pp. 1–5. ISBN 978-1-5386-3885-9 https://doi.org/10.1109/Agro-Geoinformatics.2017.8047016

[41] Aceto, G., Ciuonzo, D., Montieri, A., Pescapé, A. "Mobile Encrypted Traffic Classification Using Deep Learning: Experimental Evaluation, Lessons Learned, and Challenges", IEEE Transactions on Network and Service Management, 16(2), pp. 445–458, 2019. https://doi.org/10.1109/TNSM.2019.2899085

[42] O'Shea, T., Hoydis, J. "An Introduction to Deep Learning for the Physical Layer", IEEE Transactions on Cognitive Communications and Networking, 3(4), pp. 563–575, 2017. https://doi.org/10.1109/TCCN.2017.2758370

[43] Hansen, J. H. L., Hasan, T. "Speaker Recognition by Machines and Humans: A Tutorial Review", IEEE Signal Processing Magazine, 32(6), pp. 74–99, 2015. https://doi.org/10.1109/MSP.2015.2462851

[44] Gupta, H., Gupta, D. "LPC and LPCC Method of Feature Extraction in Speech Recognition System", In: 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, India, 2016, pp. 498–502. ISBN 978-1-4673-8204-5 https://doi.org/10.1109/CONFLUENCE.2016.7508171

[45] Reynolds, D. A., Rose, R. C. "Robust Text-Independent Speaker Identification Using Gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing, 3(1), pp. 72–83, 1995. https://doi.org/10.1109/89.365379

[46] Das, A., Jena, M. R., Barik, K. K. "Mel-Frequency Cepstral Coefficient (MFCC) - A Novel Method for Speaker Recognition", Digital Technologies, 1(1), pp. 1–3, 2015. https://doi.org/10.12691/dt-1-1-1

[47] Kaladharan, N. "Speech Enhancement by Spectral Subtraction Method", International Journal of Computer Applications, 96(13), pp. 45–48, 2014. https://doi.org/10.5120/16858-6739

[48] Sarma, M., Sarma, K. K. "Vowel Phoneme Segmentation for Speaker Identification Using an ANN-Based Framework", Journal of Intelligent Systems, 22(2), pp. 111–130, 2013. https://doi.org/10.1515/jisys-2012-0050

[49] Avci, D. "An Expert System for Speaker Identification Using Adaptive Wavelet Sure Entropy", Expert Systems with Applications, 36(3), pp. 6295–6300, 2009. https://doi.org/10.1016/j.eswa.2008.07.012

[50] Daqrouq, K. "Wavelet Entropy and Neural Network for Text-Independent Speaker Identification", Engineering Applications of Artificial Intelligence, 24(5), pp. 796–802, 2011. https://doi.org/10.1016/j.engappai.2011.01.001

[51] Fan, X., Hansen, J. H. L. "Speaker Identification within Whispered Speech Audio Streams", IEEE Transactions on Audio, Speech, and Language Processing, 19(5), pp. 1408–1421, 2011. https://doi.org/10.1109/TASL.2010.2091631

[52] Jawarkar, N. P., Holambe, R. S., Basu, T. K. "Effect of Nonlinear Compression Function on the Performance of the Speaker Identification System under Noisy Conditions", In: PerMIn '15: Proceedings of the 2nd International Conference on Perception and Machine Intelligence, Kolkata, India, 2015, pp. 137–144. ISBN 978-1-4503-2002-3 https://doi.org/10.1145/2708463.2709049

[53] Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T. "Speaker Identification and Clustering Using Convolutional Neural Networks", In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), Vietri sul Mare, Italy, 2016, pp. 1–6. ISBN 978-1-5090-0747-9 https://doi.org/10.1109/MLSP.2016.7738816

[54] Jeevan, M., Hanmandlu, M. "Higher Order Information Set Based Features for Text-Independent Speaker Identification", International Journal of Speech Technology, 21(3), pp. 451–461, 2018. https://doi.org/10.1007/s10772-017-9472-7

[55] Novotný, O., Plchot, O., Glembek, O., Černocký, J. H., Burget, L. "Analysis of DNN Speech Signal Enhancement for Robust Speaker Recognition", Computer Speech & Language, 58, pp. 403–421, 2019. https://doi.org/10.1016/j.csl.2019.06.004

[56] Wu, J.-D., Tsai, Y.-J. "Speaker Identification System Using Empirical Mode Decomposition and an Artificial Neural Network", Expert Systems with Applications, 38(5), pp. 6112–6117, 2011. https://doi.org/10.1016/j.eswa.2010.11.013

[57] Zhao, X., Wang, Y., Wang, D. "Robust Speaker Identification in Noisy and Reverberant Conditions", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), pp. 836–845, 2014. https://doi.org/10.1109/TASLP.2014.2308398

[58] An, N. N., Thanh, N. Q., Liu, Y. "Deep CNNs with Self-Attention for Speaker Identification", IEEE Access, 7, pp. 85327–85337, 2019. https://doi.org/10.1109/ACCESS.2019.2917470

[59] Faragallah, O. S. "Robust Noise MKMFCC–SVM Automatic Speaker Identification", International Journal of Speech Technology, 21(2), pp. 185–192, 2018. https://doi.org/10.1007/s10772-018-9494-9

[60] Sun, L., Gu, T., Xie, K., Chen, J. "Text-Independent Speaker Identification Based on Deep Gaussian Correlation Supervector", International Journal of Speech Technology, 22(2), pp. 449–457, 2019. https://doi.org/10.1007/s10772-019-09618-5

[61] Liu, Z., Wu, Z., Li, T., Li, J., Shen, C. "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition", IEEE Transactions on Industrial Informatics, 14(7), pp. 3244–3252, 2018. https://doi.org/10.1109/TII.2018.2799928

[62] Al-Rawahy, S., Hossen, A., Heute, U. "Text-Independent Speaker Identification System Based on the Histogram of DCT-Cepstrum Coefficients", International Journal of Knowledge-based and Intelligent Engineering Systems, 16(3), pp. 141–161, 2012. https://doi.org/10.3233/KES-2012-0239

[63] Mporas, I., Safavi, S., Gan, H. C., Sotudeh, R. "Evaluation of Classification Algorithms for Text Dependent and Text Independent Speaker Identification", In: IEICE Information and Communication Technology Forum, Patras, Greece, 2016, B3-3. https://doi.org/10.34385/proc.24.B3-3

[64] Renisha, G., Jayasree, T. "Cascaded Feedforward Neural Networks for Speaker Identification Using Perceptual Wavelet Based Cepstral Coefficients", Journal of Intelligent & Fuzzy Systems, 37(1), pp. 1141–1153, 2019. https://doi.org/10.3233/JIFS-182599

[65] Daqrouq, K., Tutunji, T. A. "Speaker Identification Using Vowels Features through a Combined Method of Formants, Wavelets, and Neural Network Classifiers", Applied Soft Computing, 27, pp. 231–239, 2015. https://doi.org/10.1016/j.asoc.2014.11.016

[66] Wu, J.-D., Lin, B.-F. "Speaker Identification Based on the Frame Linear Predictive Coding Spectrum Technique", Expert Systems with Applications, 36(4), pp. 8056–8063, 2009. https://doi.org/10.1016/j.eswa.2008.10.051

[67] Dhakal, P., Damacharla, P., Javaid, A. Y., Devabhaktuni, V. "A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface", Machine Learning & Knowledge Extraction, 1(1), pp. 504–520, 2019. https://doi.org/10.3390/make1010031

[68] Bunrit, S., Inkian, T., Kerdprasop, N., Kerdprasop, K. "Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network", International Journal of Machine Learning and Computing, 9(2), pp. 143–148, 2019. https://doi.org/10.18178/ijmlc.2019.9.2.778

[69] Imran, A. S., Haflan, V., Shahrebabaki, A. S., Olfati, N., Svendsen, T. K. "Evaluating Acoustic Feature Maps in 2D-CNN for Speaker Identification", In: ICMLC '19: Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Zhuhai, China, 2019, pp. 211–216. ISBN 978-1-4503-6600-7 https://doi.org/10.1145/3318299.3318386

[70] Yadav, S., Rai, A. "Learning Discriminative Features for Speaker Identification and Verification", In: Interspeech 2018, Hyderabad, India, 2018, pp. 2237–2241. https://doi.org/10.21437/Interspeech.2018-1015

[71] Vimala, C., Radha, V. "Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words", International Journal of Computer Science and Information Technologies (IJCSIT), 5(1), pp. 378–383, 2014.

[72] Khara, S., Singh, S., Vir, D. "A Comparative Study of the Techniques for Feature Extraction and Classification in Stuttering", In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 887–893. ISBN 978-1-5386-1975-9 https://doi.org/10.1109/ICICCT.2018.8473099

[73] Singh, P. P., Rani, P. "An Approach to Extract Feature Using MFCC", IOSR Journal of Engineering (IOSRJEN), 4(8), pp. 21–25, 2014. https://doi.org/10.9790/3021-04812125

[74] Anusuya, M. A., Katti, S. K. "Speech Recognition by Machine, A Review", [preprint], arXiv, arXiv:1001.2267, 13 January 2010. https://doi.org/10.48550/arXiv.1001.2267

[75] Tiwari, V. "MFCC and Its Applications in Speaker Recognition", International Journal on Emerging Technologies, 1(1), pp. 19–22, 2010.

[76] Kinnunen, T., Li, H. "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors", Speech Communication, 52(1), pp. 12–40, 2010. https://doi.org/10.1016/j.specom.2009.08.009

[77] Dhingra, S. D., Nijhawan, G., Pandit, P. "Isolated Speech Recognition Using MFCC and DTW", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2(8), pp. 4085–4092, 2013.

[78] Kianisarkaleh, A., Ghassemian, H. "Spatial-Spectral Locality Preserving Projection for Hyperspectral Image Classification with Limited Training Samples", International Journal of Remote Sensing, 37(21), pp. 5045–5059, 2016. https://doi.org/10.1080/01431161.2016.1226523

[79] Shiva Prasad, K. M., Kodanda Ramaiah, G. N., Manjunatha, M. B. "Speech Features Extraction Techniques for Robust Emotional Speech Analysis/Recognition", Indian Journal of Science and Technology, 10(3), pp. 1–9, 2017. https://doi.org/10.17485/ijst/2017/v10i3/110571

[80] Kuncheva, L. I., Faithfull, W. J. "PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data", IEEE Transactions on Neural Networks and Learning Systems, 25(1), pp. 69–80, 2014. https://doi.org/10.1109/TNNLS.2013.2248094

[81] Soong, F. K., Rosenberg, A. E., Juang, B.-H., Rabiner, L. R. "Report: A Vector Quantization Approach to Speaker Recognition", AT&T Technical Journal, 66(2), pp. 14–26, 1987. https://doi.org/10.1002/j.1538-7305.1987.tb00198.x

[82] Ramachandran, R. P., Sondhi, M. M., Seshadri, N., Atal, B. S. "A Two Codebook Format for Robust Quantization of Line Spectral Frequencies", IEEE Transactions on Speech and Audio Processing, 3(3), pp. 157–168, 1995. https://doi.org/10.1109/89.388142

[83] Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A. "Modeling Prosodic Feature Sequences for Speaker Recognition", Speech Communication, 46(3–4), pp. 455–472, 2005. https://doi.org/10.1016/j.specom.2005.02.018

[84] Campbell, J. P., Reynolds, D. A., Dunn, R. B. "Fusing High- and Low-Level Features for Speaker Recognition", In: Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech 2003), Geneva, Switzerland, 2003, pp. 2665–2668. https://doi.org/10.21437/Eurospeech.2003-727

[85] Adami, A. G., Mihaescu, R., Reynolds, D. A., Godfrey, J. J. "Modeling Prosodic Dynamics for Speaker Recognition", In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03), Hong Kong, China, 2003, pp. IV-788–IV-791. ISBN 0-7803-7663-3

[86] Ali, H., Tran, S. N., Benetos, E., d'Avila Garcez, A. S. "Speaker Recognition with Hybrid Features from a Deep Belief Network", Neural Computing and Applications, 29(6), pp. 13–19, 2018. https://doi.org/10.1007/s00521-016-2501-7

[87] Bisio, I., Garibotto, C., Grattarola, A., Lavagetto, F., Sciarrone, A. "Smart and Robust Speaker Recognition for Context-Aware In-Vehicle Applications", IEEE Transactions on Vehicular Technology, 67(9), pp. 8808–8821, 2018. https://doi.org/10.1109/TVT.2018.2849577

[88] Dovydaitis, L., Rudžionis, V. "Building LSTM Neural Network Based Speaker Identification System", Computational Science and Techniques, 6(1), pp. 574–580, 2018. https://doi.org/10.15181/csat.v6i1.1579

[89] Soleymanpour, M., Marvi, H. "Text-Independent Speaker Identification Based on Selection of the Most Similar Feature Vectors", International Journal of Speech Technology, 20(1), pp. 99–108, 2017. https://doi.org/10.1007/s10772-016-9385-x

[90] Sardar, V. M., Shirbahadurkar, S. D. "Speaker Identification of Whispering Speech: An Investigation on Selected Timbrel Features and KNN Distance Measures", International Journal of Speech Technology, 21(3), pp. 545–553, 2018. https://doi.org/10.1007/s10772-018-9527-4

[91] Sardar, V. M., Shirbahadurkar, S. D. "Timbre Features for Speaker Identification of Whispering Speech: Selection of Optimal Audio Descriptors", International Journal of Computers and Applications, 43(10), pp. 1047–1053, 2021. https://doi.org/10.1080/1206212X.2019.1652788

[92] Zhang, X., Zou, X., Sun, M., Wu, P. "Robust Speaker Recognition Using Improved GFCC and Adaptive Feature Selection", In: International Conference on Security with Intelligent Computing and Big-Data Services, Guilin, China, 2018, pp. 159–169. ISBN 978-3-030-16945-9 https://doi.org/10.1007/978-3-030-16946-6_13

[93] Abdul, Z. K. "Kurdish Speaker Identification Based on a One Dimensional Convolutional Neural Network", Computational Methods for Differential Equations, 7(4), pp. 566–572, 2019.

[94] Sardar, V. M., Shirbahadurkar, S. D. "Speaker Identification of Whispering Sound: Effect of Different Features on the Identification Accuracy", International Journal of Pure and Applied Mathematics, 118, 24, 2018.

[95] Mokgonyane, T. B., Sefara, T. J., Manamela, M. J., Modipa, T. I. "The Effects of Data Size on Text-Independent Automatic Speaker Identification System", In: 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Winterton, South Africa, pp. 1–6. ISBN 978-1-5386-9237-0 https://doi.org/10.1109/ICABCD.2019.8851018

[96] Almaadeed, N., Aggoun, A., Amira, A. "Text-Independent Speaker Identification Using Vowel Formants", Journal of Signal Processing Systems, 82(3), pp. 345–356, 2016. https://doi.org/10.1007/s11265-015-1005-5

[97] Almaadeed, N., Aggoun, A., Amira, A. "Speaker Identification Using Multimodal Neural Networks and Wavelet Analysis", IET Biometrics, 4(1), pp. 18–28, 2015. https://doi.org/10.1049/iet-bmt.2014.0011

[98] Kawakami, Y., Wang, L.., Kai, A., Nakagawa, S. "Speaker Identification by Combining Various Vocal Tract and Vocal Source Features", In: International Conference on Text, Speech, and Dialogue, Brno, Czech Republic, 2014, pp. 382–389. ISBN 978-3-319-10815-5 https://doi.org/10.1007/978-3-319-10816-2_46

[99] Tirumala, S.S. "A Deep Autoencoder Approach for Speaker Identification", In: ICSPS 2017: Proceedings of the 9th International Conference on Signal Processing Systems, Auckland, New Zealand, 2017, pp. 175–179. ISBN 978-1-4503-5384-7 https://doi.org/10.1145/3163080.3163097

[100] Zhang, C., Koishida, K., Hansen, J. H. L. "Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(9), pp. 1633–1644, 2018. https://doi.org/10.1109/TASLP.2018.2831456

[101] Li, Z., Gao, Y. "Acoustic Feature Extraction Method for Robust Speaker Identification", Multimedia Tools and Applications, 75(12), pp. 7391–7406, 2016. https://doi.org/10.1007/s11042-015-2660-z

[102] Michalevsky, Y., Talmon, R., Cohen, I. "Speaker Identification Using Diffusion Maps", In: 19th European Signal Processing Conference (EUSIPCO 2011), Barcelona, Spain, 2011, pp. 1299–1302.

[103] Abdalmalak, K. A., Gallardo-Antolín, A. "Enhancement of a Text-Independent Speaker Verification System by Using Feature Combination and Parallel Structure Classifiers", Neural Computing and Applications, 29(3), pp. 637–651, 2018. https://doi.org/10.1007/s00521-016-2470-x

[104] Sadıç, S., Gülmezoğlu, M. B. "Common Vector Approach and Its Combination with GMM for Text-Independent Speaker Recognition", Expert Systems with Applications, 38(9), pp. 11394–11400, 2011. https://doi.org/10.1016/j.eswa.2011.03.009

[105] Chakroborty, S., Saha, G. "Improved Text-Independent Speaker Identification Using Fused MFCC & IMFCC Feature Sets Based on Gaussian Filter", International Journal of Electronics and Communication Engineering, 3(11), pp. 1968–1976, 2009. https://doi.org/doi.org/10.5281/zenodo.1073555

[106] Wu, J.-D., Lin, B.-F. "Speaker Identification Using Discrete Wavelet Packet Transform Technique with Irregular Decomposition", Expert Systems with Applications, 36(2), pp. 3136–3143, 2009. https://doi.org/10.1016/j.eswa.2008.01.038

[107] Fang, Z., Guoliang, Z., Zhanjiang, S. "Comparison of Different Implementations of MFCC", Journal of Computer science and Technology, 16(6), pp. 582–589, 2001. https://doi.org/10.1007/BF02943243

[108] Jahangir, R., Teh, Y. W., Ishtiaq, U., Mujtaba, G., Nweke, H. F. "Automatic Speaker Identification through Robust Time Domain Features and Hierarchical Classification Approach", In: ICDPA 2018: Proceedings of the International Conference on Data Processing and Applications, Guangdong, China, 2018, pp. 34–38. ISBN 978-1-4503-6418-8 https://doi.org/10.1145/3224207.3224213

[109] Indumathi, A., Chandra, E. "Speaker Identification Using Bagging Techniques", In: 2015 International Conference on Computers, Communications, and Systems (ICCCS), Kanyakumari, India, 2015, pp. 223–229. ISBN 978-1-4673-9757-5 https://doi.org/10.1109/CCOMS.2015.7562905

[110] Krothapalli, S. R., Yadav, J., Sarkar, S., Koolagudi, S. G., Vuppala, A. K. "Neural Network Based Feature Transformation for Emotion Independent Speaker Identification", International Journal of Speech Technology, 15(3), pp. 335–349, 2012. https://doi.org/10.1007/s10772-012-9148-2

[111] Lei, Y., Scheffer, N., Ferrer, L., McLaren, M. "A Novel Scheme for Speaker Recognition Using a Phonetically-Aware Deep Neural Network", In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 1695–1699. ISBN 978-1-4799-2893-4 https://doi.org/10.1109/ICASSP.2014.6853887

[112] Variani, E., Lei, X., McDermott, E., Moreno, I. L., Gonzalez-Dominguez, J. "Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification", In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014, pp. 4052–4056. ISBN 978-1-4799-2893-4 https://doi.org/10.1109/ICASSP.2014.6854363

[113] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. "X-Vectors: Robust DNN Embeddings for Speaker Recognition", In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329–5333. ISBN 978-1-5386-4659-5 https://doi.org/10.1109/ICASSP.2018.8461375

[114] Nagrani, A., Chung, J. S., Zisserman, A. "VoxCeleb: A Large-Scale Speaker Identification Dataset", [preprint], arXiv, arXiv:1706.08612, 30 May 2018. https://doi.org/10.48550/arXiv.1706.08612

[115] McLaren, M., Ferrer, L., Castan, D., Lawson, A. "The Speakers in the Wild (SITW) Speaker Recognition Database", In: Interspeech 2016, San Francisco, CA, USA, 2016, pp. 818–822. ISBN 9781510833135 https://doi.org/10.21437/Interspeech.2016-1129

[116] Chen, K., Salman, A. "Learning Speaker-Specific Characteristics with a Deep Neural Architecture", IEEE Transactions on Neural Networks, 22(11), pp. 1744–1756, 2011. https://doi.org/10.1109/TNN.2011.2167240

[117] Hinton, G. E., Salakhutdinov, R. R. "Reducing the Dimensionality of Data with Neural Networks", Science, 313(5786), pp. 504–507, 2006. https://doi.org/10.1126/science.1127647

[118] Chen, N., Qian, Y., Yu, K. "Multi-Task Learning for Text-Dependent Speaker Verification", In: Interspeech 2015, Dresden, Germany, 2015, pp. 185–189. https://doi.org/10.21437/Interspeech.2015-81

[119] Fang, F., Wang, X., Yamagishi, J., Echizen, I., Todisco, M., Evans, N., Bonastre, J.-F. "Speaker Anonymization Using X-Vector and Neural Waveform Models", In: 10th ISCA Workshop on Speech Synthesis (SSW 10), Vienna, Austria, 2019, pp. 155–160. https://doi.org/10.21437/SSW.2019-28

[120] You, L., Guo, W., Dai, L.-R., Du, J. "Multi-Task Learning with High-Order Statistics for X-Vector Based Text-Independent Speaker Verification", In: Interspeech 2019, Graz, Austria, 2019, pp. 1158–1162. ISBN 9781510896833 https://doi.org/10.21437/Interspeech.2019-2264

[121] Heigold, G., Moreno, I., Bengio, S., Shazeer, N. "End-to-End Text-Dependent Speaker Verification", In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5115–5119. ISBN 978-1-4799-9988-0 https://doi.org/10.1109/ICASSP.2016.7472652

[122] Banerjee, A., Dubey, A., Menon, A., Nanda, S., Nandi, G. C. "Speaker Recognition Using Deep Belief Networks", [preprint], arXiv, arXiv:1805.08865v1, 09 May 2018. https://doi.org/10.48550/arXiv.1805.08865

[123] Hinton, G. E., Osindero, S., Teh, Y.-W. "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, 18(7), pp. 1527–1554, 2006. https://doi.org/10.1162/neco.2006.18.7.1527

[124] Patel, T., Krishna, D. N., Fathima, N., Shah, N., Mahima, C., Kumar, D., Iyengar, A. "Development of Large Vocabulary Speech Recognition System with Keyword Search for Manipuri", In: Interspeech 2018, Hyderabad, India, 2018, pp. 1031–1035. https://doi.org/10.21437/Interspeech.2018-2133

[125] Kadyan, V., Mantri, A., Aggarwal, R. K., Singh, A. "A Comparative Study of Deep Neural Network Based Punjabi-ASR System", International Journal of Speech Technology, 22(1), pp. 111–119, 2019. https://doi.org/10.1007/s10772-018-09577-3

[126] Bhowmik, T., Chowdhury, A., Das Mandal, S. K. "Deep Neural Network based Place and Manner of Articulation Detection and Classification for Bengali Continuous Speech", Procedia Computer Science, 125, pp. 895–901, 2018. https://doi.org/10.1016/j.procs.2017.12.114

[127] Dey, A., Zhang, W., Fung, P. "Acoustic Modeling for Hindi Speech Recognition in Low-Resource Settings", In: 2014 International Conference on Audio, Language and Image Processing, Shanghai, China, 2014, pp. 891–894. ISBN 978-1-4799-3902-2 https://doi.org/10.1109/ICALIP.2014.7009923

[128] Mandal, P., Jain, S., Ojha, G., Shukla, A. "Development of Hindi Speech Recognition System of Agricultural Commodities Using Deep Neural Network", In: Interspeech 2015, Dresden, Germany, 2015, pp. 1241–1245. ISBN 978-1-5108-1790-6 https://doi.org/10.21437/Interspeech.2015-312

[129] Pandey, A., Srivastava, B. M. L., Gangashetty, S. V. "Adapting Monolingual Resources for Code-Mixed Hindi-English Speech Recognition", In: 2017 International Conference on Asian Language Processing (IALP), Singapore, 2017, pp. 218–221. ISBN 978-1-5386-1982-7 https://doi.org/10.1109/IALP.2017.8300583

[130] Kingma, D. P., Welling, M. "Auto-Encoding Variational Bayes", [preprint], arXiv, arXiv:1312.6114, 20 December 2013. https://doi.org/10.48550/arXiv.1312.6114

[131] Rezende, D. J., Mohamed, S., Wierstra, D. "Stochastic Backpropagation and Approximate Inference in Deep Generative Models", Proceedings of Machine Learning Research, 32(2), pp. 1278–1286, 2014.

[132] Tang, Z., Li, L., Wang, D. "Multi-Task Recurrent Model for Speech and Speaker Recognition", In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, South-Korea, 2016, pp. 1–4. ISBN 978-1-5090-2401-8 https://doi.org/10.1109/APSIPA.2016.7820893

[133] Ghahabi, O., Hernando, J. "Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4), pp. 807–817, 2017. https://doi.org/10.1109/TASLP.2017.2661705

[134] Chopra, S., Hadsell, R., LeCun, Y. "Learning a Similarity Metric Discriminatively, with Application to Face Verification", In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 539–546. ISBN 0-7695-2372-2

[135] Ravanelli, M., Bengio, Y. "Speaker Recognition from Raw Waveform with SincNet", In: 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 1021–1028. ISBN 978-1-5386-4335-8
https://doi.org/10.1109/SLT.2018.8639585

[136] Ravanelli, M., Bengio, Y. "Learning Speaker Representations with Mutual Information", In: Interspeech 2019, Graz, Austria, 2019, pp. 1153–1157. ISBN 9781510896833
https://doi.org/10.21437/Interspeech.2019-2380

[137] Ranjan, S., Hansen, J. H. L. "Curriculum Learning Based Approaches for Noise Robust Speaker Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(1), pp. 197–210, 2018.
https://doi.org/10.1109/TASLP.2017.2765832

[138] Zheng, S., Liu, G., Suo, H., Lei, Y. "Autoencoder-Based Semi-Supervised Curriculum Learning for Out-of-Domain Speaker Verification", In: Interspeech 2019, Graz, Austria, 2019, pp. 4360–4364. ISBN 9781510896833
https://doi.org/10.21437/Interspeech.2019-1440

[139] Shome, N., Laskar, R. H., Das, D. "Reference Free Speech Quality Estimation for Diverse Data Condition", International Journal of Speech Technology, 22(3), pp. 585–599, 2019.
https://doi.org/10.1007/s10772-018-9537-2

[140] Lim, J. S., Oppenheim, A. V. "Enhancement and Bandwidth Compression of Noisy Speech", In: Proceedings of the IEEE, 67(12), pp. 1586–1604, 1979.
https://doi.org/10.1109/PROC.1979.11540

[141] Wu, Z., Cao, Z. "Improved MFCC-Based Feature for Robust Speaker Identification", Tsinghua Science and Technology, 10(2), pp. 158–161, 2005.
https://doi.org/10.1016/S1007-0214(05)70048-1

[142] Cristianini, N., Shawe-Taylor, J. "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods", Cambridge University Press, 2013. ISBN 9780511801389
https://doi.org/10.1017/CBO9780511801389

[143] Wang, R. Y., Storey, V. C., Firth, C. P. "A Framework for Analysis of Data Quality Research", IEEE Transactions on Knowledge and Data Engineering, 7(4), pp. 623–640, 1995.
https://doi.org/10.1109/69.404034

[144] Campbell, J. P. "Speaker Recognition: A Tutorial", Proceedings of the IEEE, 85(9), pp. 1437–1462, 1997.
https://doi.org/10.1109/5.628714

[145] Karray, L., Martin, A. "Towards Improving Speech Detection Robustness for Speech Recognition in Adverse Conditions", Speech Communication, 40(3), pp. 261–276, 2003.
https://doi.org/10.1016/S0167-6393(02)00066-3

[146] Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R., Prasad, R. V., Gaurav, V. "VAD Techniques for Real-Time Speech Transmission on the Internet", In: 5th IEEE International Conference on High Speed Networks and Multimedia Communication (Cat. No. 02EX612), Jeju, South-Korea, 2002, pp. 46–50.
https://doi.org/10.1109/HSNMC.2002.1032545

[147] Shome, N., Laskar, R. H., Kashyap, R., Bandyopadhyay, S. "A Robust Technique for End Point Detection Under Practical Environment", In: International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, Silchar, India, 2020, pp. 131–144. ISBN 978-981-15-6317-1
https://doi.org/10.1007/978-981-15-6318-8_12

[148] Shome, N., Laskar, R. H., Kashyap, R. "Effect of End Point Detection on Fixed Phrase Speaker Verification", In: Emerging Electronics and Automation: Select Proceedings of E2A 2021, Assam, India, 2021, pp. 343–353. ISBN 978-981-19-4299-0
https://doi.org/10.1007/978-981-19-4300-3_30

[149] rasmusbergpalm "DeepLearnToolbox", [online] Available at: https://github.com/rasmusbergpalm/DeepLearnToolbox [Accessed: 21 May 2021]

[150] keras-team "Keras", [online] Available at: https://github.com/keras-team/keras [Accessed: 21 May 2021]

[151] Deep Learning "Theano", [online] Available at: http://deeplearning.net/software/theano/ [Accessed: 21 May 2021]

[152] TensorFlow "Webpage of TensorFlow", [online] Available at: www.tensorflow.org [Accessed: 21 May 2021]

[153] BVLC "Caffe", [online] Available at: https://github.com/BVLC/caffe/ [Accessed: 21 May 2021]

[154] MatConvNet "MatConvNet: CNNs for MATLAB", [online] Available at: http://www.vlfeat.org/matconvnet/ [Accessed: 21 May 2021]

[155] KotlinNLP "SimpleDNN", [online] Available at: https://github.com/KotlinNLP/SimpleDNN [Accessed: 21 May 2021]

[156] cudamat "CUDAMat", [online] Available at: https://github.com/cudamat/cudamat [Accessed: 21 May 2021]

[157] Microsoft "CNTK", [online] Available at: https://github.com/Microsoft/cntk [Accessed: 21 May 2021]

[158] pybrain "PyBrain", [online] Available at: https://github.com/pybrain/pybrain [Accessed: 21 May 2021]

[159] Torch "torch7", [online] Available at: https://github.com/torch/torch7 [Accessed: 21 May 2021]

[160] Vasilev, I. "Deep Neural Networks with GPU support", [online] Available at: https://github.com/ivan-vasilev/neuralnetworks [Accessed: 21 May 2021]

[161] Demyanov, S. "ConvNet", [online] Available at: https://github.com/sdemyanov/ConvNet [Accessed: 21 May 2021]

[162] Artificial Intelligence Techniques, Ltd. "OpenNN", [online] Available at: http://www.opennn.net/download/index.html [Accessed: 21 May 2021]

[163] Eclipse Deeplearning4j "DL4j", [online] Available at: https://github.com/eclipse/deeplearning4j [Accessed: 21 May 2021]

[164] Chainer "Chainer", [online] Available at: https://github.com/chainer/chainer [Accessed: 21 May 2021]

[165] Ravanelli, M. "SincNet", [online] Available at: https://github.com/mravanelli/SincNet [Accessed: 21 May 2021]

[166] Jin, Q. "Robust Speaker Recognition", PhD Thesis, Carnegie Mellon University, 2007.

[167] Kinnunen, T. "Spectral Features for Automatic Text-Independent Speaker Recognition", Licentiate's thesis, University of Joensuu, 2003.

[168] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. "DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1", U.S. Department of Commerce, Gaithersburg, MD, USA, Rep. NISTIR 4930, 1993.
https://doi.org/10.6028/NIST.IR.4930

[169] Larcher, A., Lee, K. A., Ma, B., Li, H. "The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases", In: Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 2012, pp. 1578–1581. ISBN 9781622767595

[170] Greenberg, C., Martin, A., Graff, D., Brandschain, L., Walker, K. "2010 NIST Speaker Recognition Evaluation Test Set LDC2017S06", National Institute of Standards and Technology, 2017. ISBN 1-58563-795-5
https://doi.org/10.35111/fjsq-a117

[171] Campbell, J., Higgins, A. "YOHO Speaker Verification", International Telephone & Telegraph, 1994. ISBN 1-58563-042-X
https://doi.org/10.35111/3wc3-n668

[172] Campbell, J. P. "Testing with the YOHO CD-ROM Voice Verification Corpus", In: 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 1995, pp. 341–344. ISBN 0-7803-2431-5
https://doi.org/10.1109/ICASSP.1995.479543

[173] Mason, J. S. "Project: DAVID (Digital Audio Visual Integrated Database)", Department of Electrical and Electronic Engineering, University of Wales Swansea, Swansea, UK, Technical Report, 1996.

[174] Pigeon, S., Vandendorpe, L. "The M2VTS Multimodal Face Database (Release 1.00)", In: International Conference on Audio- and Video-Based Biometric Person Authentication, Crans-Montana, Switzerland, 1997, pp. 403–409. ISBN 978-3-540-62660-2
https://doi.org/10.1007/BFb0016021

[175] Chollet, G., Cochard, J.-L., Constantinescu, A., Jaboulet, C., Langlais, P. "Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-Speaker Variability", IDIAP, Martigny, Switzerland, Rep. 96-01, 1996.

[176] Cole, R. A., Noel, M., Noel, V. "The CSLU Speaker Recognition Corpus", In: 5th International Conference on Spoken Language Processing (ICSLP 1998), Sydney, Australia, 1998, 0856.
https://doi.org/10.21437/ICSLP.1998-610

[177] Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G. "XM2VTSDB: The Extended M2VTS Database", In: Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99), Washington, DC, USA, 1999, pp. 965–966.

[178] Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V. "AHUMADA: A Large Speech Corpus in Spanish for Speaker Characterization and Identification", Speech Communication, 31(2–3), pp. 255–264, 2000.
https://doi.org/10.1016/S0167-6393(99)00081-3

[179] Hennebert, J., Melin, H., Petrovska, D., Genoud, D. "POLYCOST: A Telephone-Speech Database for Speaker Recognition", Speech Communication, 31(2–3), pp. 265–270, 2000.
https://doi.org/10.1016/S0167-6393(99)00082-5

[180] Karlsson, I. "Within Speaker Variability in the VeriVox Database", Gothenburg Papers in Theoretical Linguistics, 81(1), pp. 93–96, 1999.

[181] Karlsson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F., Scherer, K. "Speaker Verification with Elicited Speaking Styles in the VeriVox Project", Speech Communication, 31(2–3), pp. 121–129, 2000.
https://doi.org/10.1016/S0167-6393(99)00073-4

[182] Steininger, S., Rabold, S., Dioubina, O., Schiel, F. "Development of User-State Conventions for the Multimodal Corpus in Smartkom", In: Proceedings of Workshop on Multimodal Resources and Multimodal Systems Evaluation, Las Palmas, Spain, 2002, pp. 33–37.

[183] Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.-P. "The BANCA Database and Evaluation Protocol", In: 4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK, 2003, pp. 625–638. ISBN 978-3-540-40302-9
https://doi.org/10.1007/3-540-44887-X_74

[184] Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Leroux les Jardins, J., Lunter, J., Ni, Y., Petrovska-Delacrétaz, D. "BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", In: AVBPA 2003: Audio- and Video-Based Biometric Person Authentication, Guildford, UK, 2003, pp. 845–853. ISBN 978-3-540-40302-9
https://doi.org/10.1007/3-540-44887-X_98

[185] Raev, A., Koval, S., Smirnova, N., Khitrova, D., Stepanov, V. "Russian through Switched Telephone Network (RuSTeN)", Philadelphia: Linguistic Data Consortium, 2006. ISBN 1-58563-388-7
https://doi.org/10.35111/bw5g-8741

[186] Fox, N., O'Mullane, B., Reilly, R. B. "VALID: A New Practical Audio-Visual Database, and Comparative Results", In: 5th International Conference of Audio and Video-Based Person Authentication, Hilton Rye Town, NY, USA, 2005, pp. 777–786. ISBN 978-3-540-27887-0
https://doi.org/10.1007/11527923_81

[187] Zheng, T. F. "The Voiceprint Recognition Activities over China", In: Oriental COCOSDA 2005, Jakarta, Indonesia, 2005, pp. 54–58. ISBN 9781467382809

[188] Woo, R. H., Park, A., Hazen, T. J. "The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments", In: 2006 IEEE Odyssey - The Speaker and Language Recognition Workshop, San Juan, PR, USA, 2006, pp. 1–6.
https://doi.org/10.1109/ODYSSEY.2006.248083

[189] Meng, H., Ching, P. C., Lee, T., Mak, M. W., Mak, B., Moon, Y. S., Siu, M.-H., Tang, X., Hui, H. P. S., Lee, A., Lo, W.-K., Ma, B., Sio, E. K. T. "The Multi-Biometric, Multi-Device and Multilingual (M3) Corpus", presented at Second International Workshop on Multimodal User Authentication, Toulouse, France, May, 11-12, 2006.

[190] Fierrez, J., Ortega-Garcia, J., Toledano, D. T., Gonzalez-Rodriguez, J. "BioSec Baseline Corpus: A Multimodal Biometric Database", Pattern Recognition, 40(4), pp. 1389–1392, 2007.
https://doi.org/10.1016/j.patcog.2006.10.014

[191] Toledano, D., Hernandez-Lopez, D., Esteve-Elizalde, C., Ortega-Garcia, J., Ramos, D., Gonzalez-Rodriguez, J. "BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition", In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008, pp. 892–895. ISBN 2-9517408-4-0

[192] Fierrez, J., Galbally, J., Ortega-Garcia, J., Freire, M. R., Alonso-Fernandez, F., Ramos, D., Toledano, D. T., …, Gracia-Roche, J. J. "BiosecureID: A Multimodal Biometric Database", Pattern Analysis and Applications, 13(2), pp. 235–246, 2010.

[193] Dessimoz, D., Richiardi, J., Champod, C., Drygajlo, A. "Multimodal Biometrics for Identity Documents (MBIoD)", Forensic Science International, 167(2–3), pp. 154–159, 2007. https://doi.org/10.1016/j.forsciint.2006.06.037

[194] Ortega-Garcia, J., Fierrez, J., Alonso-Fernandez, F., Galbally, J., Freire, M. R., Gonzalez-Rodriguez, J., Garcia-Mateo, C., …, Savran, A. "The Multiscenario Multienvironment BioSecure Multimodal Database (BMDB)", IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(6), pp. 1097–1111, 2010. https://doi.org/10.1109/TPAMI.2009.76

[195] Wong, Y. W., Ch'ng, S. I., Seng, K. P., Ang, L.-M., Chin, S. W., Chew, W. J., Lim, K. H. "A New Multi-Purpose Audio-Visual UNMC-VIER Database with Multiple Variabilities", Pattern Recognition Letters, 32(13), pp. 1503–1510, 2011. https://doi.org/10.1016/j.patrec.2011.06.011

[196] Larcher, A., Bousquet, P.-M., Lee, K. A., Matrouf, D., Li, H., Bonastre, J.-F. "I-Vectors in the Context of Phonetically-Constrained Short Utterances for Speaker Verification", In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 2012, pp. 4773–4776. ISBN 978-1-4673-0045-2 https://doi.org/10.1109/ICASSP.2012.6288986

[197] Ajmera, P. K., Jadhav, D. V., Holambe, R. S. "Text-Independent Speaker Identification Using Radon and Discrete Cosine Transforms Based Features from Speech Spectrogram", Pattern Recognition, 44(10–11), pp. 2749–2759, 2011. https://doi.org/10.1016/j.patcog.2011.04.009

[198] Krobba, A., Debyeche, M., Selouani, S. A. "Maximum Entropy PLDA for Robust Speaker Recognition under Speech Coding Distortion", International Journal of Speech Technology, 22(4), pp. 1115–1122, 2019. https://doi.org/10.1007/s10772-019-09642-5

[199] Hajavi, A., Etemad, A. "A Deep Neural Network for Short-Segment Speaker Recognition", [preprint], arXiv, arXiv:1907.10420, 22 July 2019. https://doi.org/10.48550/arXiv.1907.10420

[200] Jung, J.-W., Heo, H.-S., Yang, I.-H., Shim, H.-J., Yu, H.-J. "Avoiding Speaker Overfitting in End-to-End DNNs Using Raw Waveform for Text-Independent Speaker Verification", In: Interspeech 2018, Hyderabad, India, 2018, pp. 3583–3587. https://doi.org/10.21437/Interspeech.2018-1608

[201] Manikandan, K., Chandra, E. "Speaker Identification Using a Novel Prosody with Fuzzy Based Hierarchical Decision Tree Approach", Indian Journal of Science and Technology, 9(44), pp. 1–6, 2016. https://doi.org/10.17485/ijst/2016/v9i44/90003

[202] Wang, J.-C., Chin, Y.-H., Hsieh, W.-C., Lin, C.-H., Chen, Y.-R., Siahaan, E. "Speaker Identification with Whispered Speech for the Access Control System", IEEE Transactions on Automation Science and Engineering, 12(4), pp. 1191–1199, 2015. https://doi.org/10.1109/TASE.2015.2467311

[203] Panayotov, V., Chen, G., Povey, D., Khudanpur, S. "Librispeech: An ASR Corpus Based on Public Domain Audio Books", In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206–5210. ISBN 978-1-4673-6997-8 https://doi.org/10.1109/ICASSP.2015.7178964

[204] Haris, B. C., Pradhan, G., Misra, A., Prasanna, S. R. M., Das, R. K., Sinha, R. "Multivariability Speaker Recognition Database in Indian Scenario", International Journal of Speech Technology, 15(4), pp. 441–453, 2012. https://doi.org/10.1007/s10772-012-9140-x

[205] Das, R. K., Jelil, S., Prasanna, S. R. M. "Multi-Style Speaker Recognition Database in Practical Conditions", International Journal of Speech Technology, 21(3), pp. 409–419, 2018. https://doi.org/10.1007/s10772-017-9475-4

[206] Jung, J.-W., Tak, H., Shim, H.-J., Heo, H.-S., Lee, B.-J., Chung, S.-W., Yu, H.-J., Evans, N., Kinnunen, T. "SASV 2022: The First Spoofing-Aware Speaker Verification Challenge", In: Interspeech 2022, Incheon, South-Korea, 2022, pp. 2893–2897. https://doi.org/10.21437/Interspeech.2022-11270

[207] Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H. P. "Pruning Filters for Efficient ConvNets", [preprint], arXiv, arXiv:1608.08710, 31 August 2016. https://doi.org/10.48550/arXiv.1608.08710

[208] Hinton, G., Vinyals, O., Dean, J. "Distilling the Knowledge in a Neural Network", [preprint], arXiv, arXiv:1503.02531, 09 March 2015. https://doi.org/10.48550/arXiv.1503.02531

[209] Zhao, Y., Yin, Y., Gui, G. "Lightweight Deep Learning Based Intelligent Edge Surveillance Techniques", IEEE Transactions on Cognitive Communications and Networking, 6(4), pp. 1146–1154, 2020. https://doi.org/10.1109/TCCN.2020.2999479

[210] Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., Pescapé, A. "XAI Meets Mobile Traffic Classification: Understanding and Improving Multimodal Deep Learning Architectures", IEEE Transactions on Network and Service Management, 18(4), pp. 4225–4246, 2021. https://doi.org/10.1109/TNSM.2021.3098157