A Review on Forensic Speaker Recognition System

Pardeep Sangwan
Department of Electronics & Communication Engineering
Maharaja Surajmal Institute of Technology, New Delhi, India

Abstract— The voice features like speaking style pitch, fundamental frequency of each speaker is unique. Speech recognition is the capability to recognize said words whereas speaker recognition recognizes the speaker on the specific characteristics obtained from speaker voice. Speaker recognition is a significant area in the field of speech processing. The standard databases existing for speaker recognition and forensic speaker recognition also discussed .The criterion for selection of these database also covered in this study. In speaker recognition various sources of variability has been discussed. Various factor responsible for variability has been explained. This paper describes the I-Vector and deep learning techniques.

Keywords: Variability, Database, I-Vector, Neural Network, DNN, Deep learning

I. INTRODUCTION

Communication in speech is one of natural form. A person's voice comprises several parameters that carries information such as health, sentiment, gender, attitude, and individuality. Speaker recognition could be considered as speaker verification and speaker identification. If the speaker claims to be of a specific identity and the voice is used to check this claim, this is called verification or authentication. Instead, identification is the task of describing an unidentified speaker's identity. Verification or authentication represents the specific uniqueness and voice claimed by the speaker .On the other hand speaker verification is a process in which the voice of one speaker is matched to an individual voice template in comparison to speaker identification where the voice of one speaker is matched to N templates. Thus verification is to be a 1:1 match and speaker identification is 1: N match [1]. There are two different for speaker identification task, namely, open set and closed set. Speaker .Closed set Speaker Identification represents highest level of resemblance of an unidentified speaker input speech signal with the template of reference speaker, from N number of speaker. This unknown speaker speech template is supposed to be taken from a certain set of speakers. Thus an enforced decision by a closed set SI is taken by selecting highest matched speaker from the speaker database. On the other hand in an open set SI system the matching of unidentified speaker speech reference template may not occur [2]. Another classification for SR systems is text dependent and text independent. The text dependent SR system are those in which a random sequence of text is already defined .whereas text independent SR system there are no restriction on the text which are to be used by the speakers therefor the training and testing speech signal may represent entirely dissimilar content thereby making the text independent Speaker recognition task very challenging. Automatic SR is an important application in the field of pattern recognition utilizing training and testing phases. In training, a speaker enrolls the speech samples in the SR system. The system thereby builds a speech models of the enrolled speaker by extracting specific speaker information. Whereas in testing phase a unknown voice sample is utilized by the SR system to find the similarity between the unknown sample and the sample of already enrolled speaker and subsequently make a decision the expected output of the system can be name of one of the training speakers or a rejection of voice [3]. Speaker recognition that is used in the field of Forensic Science. When a criminal leaves his/her voice as evidence it may be telephone recording or voice recorded in camera. Many criminals may try to change their voice to prevent them from being recognized. So, the speaker recognition in Forensic is more challenging. When there is appropriate speech material is available from the suspect and offender to extract speaker parameter from speech data and comparison is done among samples. Automatic Speaker Recognition systems, feature is extracted and signifies feature extraction for performing a significant comparison.

II. SPEAKER RECOGNITION SYSTEM

The SR system basically consist of three stages namely preprocessing, feature extraction and classification. Preprocessing step is done before feature extraction in order to remove redundant information from the speech signal. Feature extraction performs dimensionality reduction thereby simplifying the processing of speech signals. Audio signal likewise comprises of non-significant data. Consequently, significant features are separated and are utilized for advance processing of speech signal. Preprocessing includes segmentation of frames, identification of active frames and windowing method. In the event that speech data is improved during preprocessing, at that point it builds the identification rate of the framework .Spectral subtraction and adaptive noise cancellation techniques are utilized for speech improvement in single channel and multichannel audio data separately. In this manner, significant feature of speech signal are extracted for further processing. The generally utilized methods for feature extraction are Mel frequency cepstral coefficients (MFCC), Delta and double delta features, linear predictive coefficients (LPC), perceptual linear predictive (PLP), RASTA PLP, Shifted Delta Cepstrum (SDC) [4]. For various users classifiers are utilized to model the feature data after extraction the reliable feature.

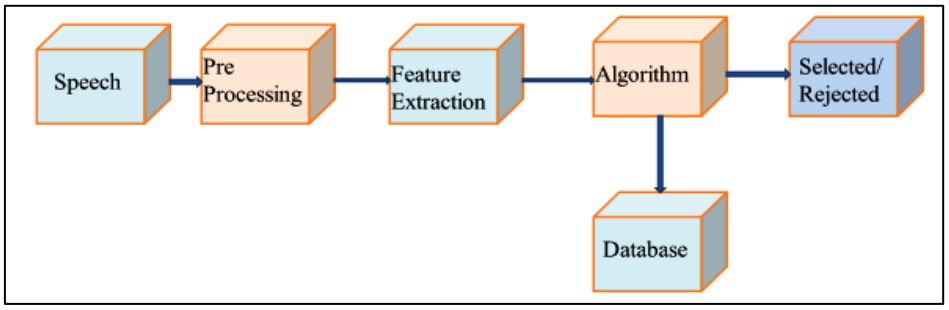

Fig. 1: Phases of Speaker Recognition System

Features are matched with the speaker model database whenever an individual speaks and accepted or rejected based on matching of speaker model. The classifiers modelling can be done by using techniques: Stochastic model and deterministic model. Pattern matching is probabilistic in nature and result is a degree of likelihood or conditional probability of given model e.g. Support Vector Machines (SVM), Vector Quantization (VQ) ,Artificial Neural Networks (ANN),Deep Neural Network (DNN).Further these classification techniques also categorized into supervised and unsupervised. In the first technique labeled data is used whereas the latter depends on unlabeled data.

III. Challenges in Speaker Recognition

The overall performance of the speaker recognitions is affected by means of many elements simultaneously. Same individual does not say the identical phrases in precisely the same way each and every time known as intraspeaker variability. Also, a variety of recording units and transmission techniques generally used worsen the problem. This makes speech signals susceptible to a great degree of variability. Variability can be interring (different or across) speaker or intra speaker (same or within) [5]. Factors responsible for variability are as under:

A. Time:

Quality of the speech is more likely to be degraded with time. In any biometric system, the accuracy of the system degrades due to ageing effect. The acoustic changes to the voice due to physiological changes in the vocal mechanism have been extensively studied in. Most of the noticeable changes in voice take place in childhood and the old age but voice keep on changing progressively throughout adulthood. Out of these, the most important changes are (i) a downward shift in fundamental frequency, and (ii) a change in resonance. Generally, in the speakers of age more than 60 years, the typical changes are instability of pitch and intensity of voice, and slowed rate of delivery. The problem of ageing and variation of speech quality go hand in hand. The ageing effect increases with time. In speaker recognition system, this problem of ageing can be overcome by updating the database at regular intervals of time but for large-scale system, this solution degrades the security of the system and hence not a feasible solution at all [6].

B. Disguise:

Speakers, who don't want to get identified i.e. un-cooperative speakers, may lower their voice or try to change their speaking behavior to avoid recognition and fool the system intentionally. The effect of intentional speaking behavior modification on the speaker recognition is investigated by Kajarekar et al. and presented vulnerability in speaker recognition systems. The performance of the speaker recognition system degrades greatly if mismatch appears between training and test.

C. Different Stress level and mental state of the speaker:

Speech samples could result in error if the speaker is in a different mental state at time of two utterances to be compared. For example, if one sample is recorded when mental state of the speaker was normal and another speech sample is taken when speaker is mentally ill or under the impression of some sedative drug. Both samples can have large intra-speaker variability [6].

D. Intrinsic Speaker variability:

Speaker variability can be affected when person is performing some physical task like driving a vehicle or there is hands free input. It is also affected by the speaking style of target speaker. Another factor is background noise while recording. Speaker variability also depends speaker physiological and emotional state. By utilizing a voice modification system the speaker may copy or change another person voice.

Situational talk Stress: the speaker is doing some activity while at the same time talking, for example, driving a vehicle, plane, truck, and so forth.), without hands voice input (industrial facility setting, crisis responders/firemen, and so on.), which can incorporate psychological just as physical assignment stress.

E. Intra-speaker variability:

Inter-session variability may be from same speaker speech samples because the acoustic features of the speech sample largely depend on the age, physiological and psychological health, emotional state etc. of the speaker. Thus, if there is a large time span between training and testing sample recording there may be huge change in the features of the voice of the same speaker. Furthermore, if target speaker is ill or in a different emotional state, then also features may change and cause inter-session variability.

F. Situational Mismatch and Different Speaking Style:

It include whether the speaker is interacting with other person, telephonic conversation, group discussion or addressing some audience. It may be speech by reading or someone disguising purposely to create confusion. It may also be the part of police suspect interviewer may be some information exchanged over phone [7].

G. Technology based Variability:

This may include where and how recording is done. This type of variability occurs due to different characteristics of transmission channel, environmental noise and data quality. It includes the following issues:

H. Electromechanical:

This occurs due to transmission channel, handset or microphone different characteristics [5] For example, audio signal captured from wired channel (landline) have different characteristics as compared to signal from wireless channel (mobile phone).

I. Environmental Noise:

It includes the background noise which may be due to traffic, some industry, and people talking near to the recording locality. Spectral characteristics of speech samples will be changed if recording of speech samples done at different location like library, sound proof room, auditorium, class room, market etc.

IV. Databases used for Speaker Recognition

The different databases and speech corpus used in recent studies are described in Table 1 and Table 2. There are various database available which are listed as follows:

S. No.	Name of Database	No. of Speakers		
		Male	Female	Total
1	NIST 2001 SRE Speech Corpus	74	100	174
2	NIST 2002 SRE Speech Corpus	139	191	330
3	NIST 2004 SRE Speech Corpus	248	368	616
4	TIMIT speech corpus	438	192	630
5	TSID Speech Corpus	31	4	35
6	POLYVAR Speech Corpus	85	58	143
7	POLYCOST Speech Corpus	74	59	133
8	YOHO Speech Corpus	106	32	138
9	ANDOSL	67	62	129
10	KING Speech Corpus	51	-	51
11	SWITCHBOARD-I Speech Corpus	-	-	543
12	SWITCHBOARD-II Speech Corpus	-	-	657
13	SIVA speech corpus	-	-	840
14	Speaker Recognition Corpus (OGI)	47	53	100

Table 1: Speech Corpus of Speaker Recognition

S. No.	Dataset	Language	Year	Speaking Styles used	Duration Size	Type of data/text used	No. of speakers		EER
							M	F	
1	ESTER-I,II	French	2006,2009	Reading	100 Hrs.	News, Debate ,TV Shows	-	-	1.1
2	AHUMADA	Spanish	2000	Reading/Extempore	-	Sentences/Digits/Word	150	250	.5
3	ETAPE	French	2011	Reading	30Hrs	News, Debate ,TV Shows	-	-	-
4	AUS-TALK	English	2012	Reading/Interview/Story telling	3000Hrs	Story/Sentence/Digit/ Word	1000		-
5	REPERE	French	2013	Reading	60 Hrs.	News, Debate ,TV Shows	-	-	-
6	WHI-SPE	Serbian	2013	Normal & whispered	5000 word	Words	5	5	-
7	CIVIL CORPUS	Spanish	2013	Disguise/Conversations/Reading	20 Hrs.	Sentences/Digits/Word	28	32	-
8	NFI-FRITS	Dutch	2014	Actual Conversations intercepted by Police	4188 Conversations	Conversational	604		12.1
9	FABIOLE	French	2016	Reading	3100 Utterances	News, Debate ,TV Shows	130		2.5

Table 2: Databases for Forensic Speaker Recognition

Speaker recognition systems are categorized according to their database, feature extraction method and classification techniques. Forensic database consists of speech samples recorded in a noisy environment by a non-cooperative speaker while in speaker recognition speech samples are recoded in a clean environment [6].

V. I-vectors

A. I-Vector in Speaker Recognition:

I vector can be used as a simple factor analysis to get a low channel and speaker dependent space, this space models both channel and speaker variability's hence named as total variability's. I-vector can be used to model both inter domain and intra domain variability's into similar small dimensional space. I-Vectors are mostly used as a feature in SVM classifier, but Cosine kernel classifier with cosine similarity score can also give better performance in terms of efficiency in comparison to SVM classifier. The I -vector provides is a technique for diminishing the dimensionality of input speech data and converts it into a feature vector of fixed length thereby maintaining the relevant information. I vector structure comprises of front end and back end. Front end part comprises of MFCC at the feature extraction stage and UBM for speaker modelling. For Feature extraction, voice activity detection based on energy is used to remove silenced part. In this technique computation of energy levels of the first frame

are done followed by normalization and finally classification is performed. Segments having greater mean value are supposed to be of speaker. After that log energy feature and MFCC along with first and second order derivative are calculated. This step is performed over different window frames. Back end part consists of I-Vector extraction, training of T-matrix, necessary statistics computation, cosine scoring and dimensionality reduction. I-Vector extractor built on different UBM size components is used to convert feature sequence of each utterance to its corresponding I-Vector. I-Vector is followed by Post Processing, in order to remove nuisance effect from total factor space channel compensation technique is required. Channel variability has modelled explicitly for channel compensation. Channel compensation is performed using Whitening, Linear Discriminant Analysis (LDA) projections and Class Covariance Normalization (WCCN) before calculating the verification score. Then it is followed by dimensionality reduction which is used to project I vector on a space in which intra-accent variability is minimized and between accents variability is maximized and, as I-Vectors contain both between accent and intra variation [8].Finally, by calculating the similarity score from the system identification result is given. The cosine distance is then calculated between I-Vector from the test segment and I-Vector from speaker, which is simplest and fast scoring model.

B. *I Vector in Forensic Speaker Recognition:*

Channel mismatch, voice disguise and voice over IP is the main problem in Forensic Automatic Speaker Recognition. Using I vector a more accurate and computationally less complex framework can be developed. I vector may be characterized as lower dimensional representation of GMM super vectors.

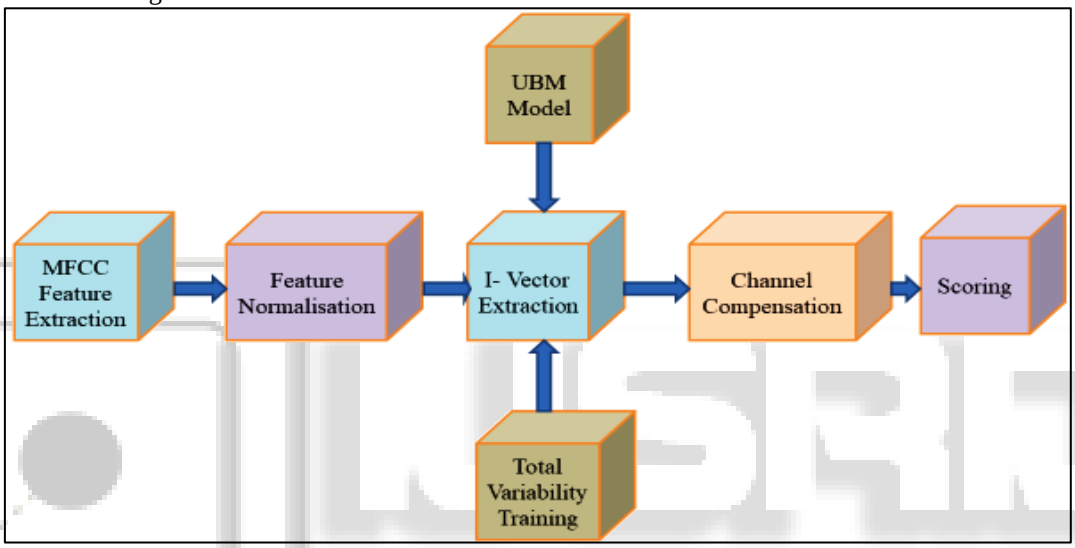

Fig. 2: Block Diagram of I Vector Based FASR system

ALZIZE is a toolkit that may be used to implementing I-vector based FASR System. Feature extraction is the first stage of speaker recognition system which converts raw speech signal into series of feature vectors. Second stage is feature normalization that is used to reduce channel variability. Gaussian normalization and Mean-variance normalization can be used for normalization. UBM is designed to capture a common form of speaker model. Higher order of GMM is UBM.A total variability space is utilized to design I-vector system. The session and speaker dependent Gaussian Super vector is represented by following equation,

$$M=m+Tw$$

m represents session and speaker independent super vector. T represents total variability space and column vector w is the I-vector. Channel compensation may be provided by PLDA for I-vectors [9]. To compare I-vectors PLDA scoring is used in verification phase to compare I-vectors.

VI. CLASSIFIERS

A. *Neural Network:*

Neural network contains various highly interconnected processing elements. It actually uses self-learning process to solve the pattern recognition problems. Such networks have mainly two methods one is probabilistic neural network (PNN) and another is feed-forward neural network (FFNN). PNN is an unsupervised feed-forward network and has four different layers as: input layer, output layer, pattern layer and lastly summation layers [10]. PNN is widely used to implement various Statistical algorithms. A Gaussian function is used for each pattern node as a probabilistic function and updating in the network weights is done as per the input patterns. The nearest neighborhood function can be utilized used to classify such patterns. Feed-forward neural networks are very easy to implement and some early type of neural networks. Such FFNN has 3 layers as input layer, hidden layer and output layer. In such networks, there is no close loop and the flow of data is possible only from input to the output layer in a feed forward manner.

B. *Deep Neural Network:*

In 2012, Deep Neural Networks (DNN) is used for image recognition that gained the Image net Large Scale Visual Recognition Challenge (ILSVCR). After that DNN were effectively used in various fields to resolve an extensive range of applications: translation, autonomous cars, speech recognition, speech understanding etc. DNN can be considered as universal approximates. In Deep neural network, neurons are interconnected. The neurons are

structured into layers. First layer is input layer, to which data features are provided. The final layer is the output layer by which output probabilities of labels or classes are provided. The output of the neuron is computed as the non-linear weighted sum of its input. DNN can be used for both feature extraction and classification. Deep learning is successful approach of machine learning than other machine learning approach like SVM and typical artificial neural network approach. Training mechanism of ANNs multiple layer is done with Back Propagation algorithm. ANN becomes very complex and time consuming due to error propagation mechanism and multiple hidden layers in between. To resolve this problem and to train multiple hidden layers network, a greedy-wise training mechanism is used [11]. The different network architectures in DNN are given in Table 3. In recent years, several Deep networks have been implemented and main networks are:

1) Convolutional Neural Network (CNN)
2) Deep Belief Networks (DBN)
3) Stacked auto-encoders (SAE)

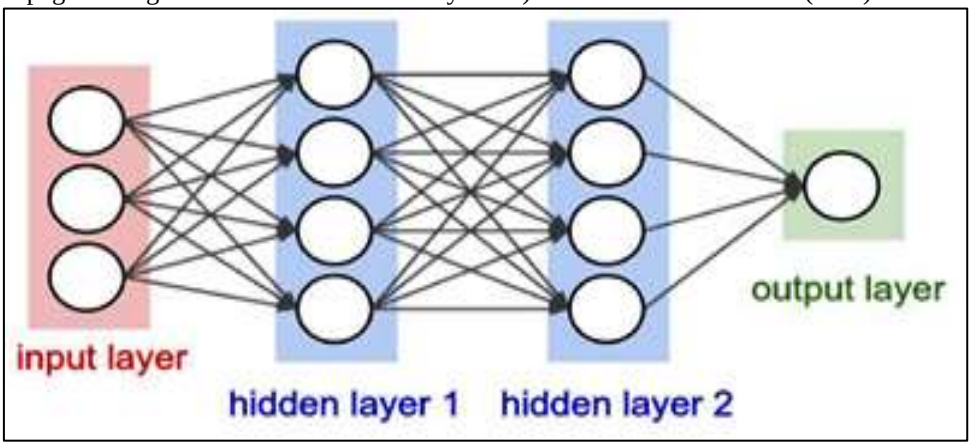

Fig. 3: Deep Neural Network

1	CNN	It is different type of feed forward ANN which can be used for feature extraction by applying sub sampling and convolution to each layer. In CNN, pooling of layers is applied after convolution layer. Pooling reduces the output dimensionality but keeps the most salient feature. Pooling provides a fixed size output matrix which is required for classification.
2	DBN	It has multiple interrelated hidden layer where each layer is input to next layer and visible only to next layer. DBN nodes are probabilistic nodes that can use activation function. Initially input is given to first layer and activation function is applied to output then this is treated as input to second layer.
3	SAE	It consists of decoder and encoder layer with various hidden layers. Firstly, random weights are applied to both networks and then trained by studying the difference between original and output data from decoding and encoding. After that error is back propagated first through decoder and then by encoder network.

1) DNN as Feature extractor:

In 2014 DNN was proposed to extract feature from speech signal instead of usual models for representing voice frames. Instead of convolutional network supervised training is used in this approach. DNN is used to extract speaker information frame by frame and then calculate utterance level information. DNN output is converted into i-vectors. Verification score is generated by probabilistic linear discriminant analysis (PLDA) on back end. In the training phase this approach for extracting known feature does not use any adaptation technique instead use DNN model for feature extraction in both enrollment and matching phase. DNN can be used alone or can be combined with conventional features.

2) DNN as Classifier:

A DNN classifier use feature such as MFCC as input to DNN classifier. Initial DNN use short frame of 20ms with a context of 10 frames for each segment of input. Each DNN is expected to predict probability of speaker for input frame that are fed to DNN. Each prediction of DNN of speaker class may be averaged to find out overall decision. Two different DNNs can be used one for frame level prediction and another one for classification. To extract feature vector each hidden layer is used to which weight is assigned after applying activation function. To reduce high dimensionality of hidden layers, dimensionality reduction technique like Principal Component Analysis can be applied before passing weights to next layer. Matrix factorization technique can be used to reduce dimensionality of output layer. DNN with multiple hidden layer can be used as stacked bottleneck Features [11].

VII. Conclusion

Speaker recognition can be widely used in different type of security systems as an identification marker. Various studies, methods have been proposed and implemented in the field of speaker recognition. This paper has discussed several features and characteristics of such high performance machine learning and deep learning techniques. It concludes that the deep neural networks presents much better results in speaker recognition than artificial neural networks as it automatically learn from its neuron or inputs without using any feature extraction methods. It gives better results in terms of accuracy and error if we have larger dataset. This paper has also discussed different datasets used in recent past in forensic and their feasibility with several. I vector can be used to model any type of inter and intra domain variability and I-Vector based FASR system has been discussed. Different type of neural network and their characteristics has been explained.

REFERENCES

[1] Singh, Nilu, R. A. Khan, and Raj Shree. "Applications of speaker recognition." Procedia engineering 38 (2012): 3122-3126.
[2] Kekre, H. B., and Vaishali Kulkarni. "Closed set and open set Speaker Identification using amplitude distribution of different Transforms." In 2013 International Conference on Advances in Technology and Engineering (ICATE), pp. 1-8. IEEE, 2013.
[3] Reynolds, Douglas A., and Richard C. Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models." IEEE transactions on speech and audio processing 3, no. 1 (1995): 72-83.
[4] Deshwal, Deepti, Pardeep Sangwan, and Divya Kumar. "Feature Extraction Methods in Language Identification: A Survey." Wireless Personal Communications (2019): 1-33.
[5] Hansen, John HL, and Taufiq Hasan. "Speaker recognition by machines and humans: A tutorial review." IEEE Signal processing magazine 32, no. 6 (2015): 74-99.
[6] Sangwan, Pardeep, and Saurabh Bhardwaj. "A Structured Approach towards Robust Database Collection for Speaker Recognition." Global Journal of Enterprise Information System 9, no. 3 (2017).
[7] Benzeghiba, Mohamed, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore et al. "Automatic speech recognition and speech variability: A review." Speech communication 49, no. 10-11 (2007): 763-786.
[8] Ibrahim, Noor Salwani, and Dzati Athiar Ramli. "I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction." Procedia Computer Science 126 (2018): 1534-1540.
[9] Ajit, Arathy P., Anu George, and Leena Mary. "I-Vectors for Forensic Automatic Speaker Recognition." In 2018 International CET Conference on Control, Communication, and Computing (IC4), pp. 284-287. IEEE, 2018.
[10] Richardson, Fred, Douglas Reynolds, and Najim Dehak. "Deep neural network approaches to speaker and language recognition." IEEE signal processing letters 22, no. 10 (2015): 1671-1675..
[11] Tirumala, Sreenivas Sremath, and Seyed Reza Shahamiri. "A review on Deep Learning approaches in Speaker Identification." In Proceedings of the 8th international conference on signal processing systems, pp. 142-147. ACM, 2016.

