International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 9 • Issue 2 pp. 186-195 June 2017-Dec 2017 www.csjournals.com

State-of-the-art Modeling Techniques in Speaker Recognition

Atul Sharma[1] and Sunil Kumar Singla[2]
[1] Research Scholar, Department of Electrical and Instrumentation Engineering, Thapar University, Patiala, Punjab
[2] Associate Professor, Department of Electrical and Instrumentation Engineering, Thapar University, Patiala, Punjab

Abstract: The voice based recognition system also referred as automatic speaker recognition system has attracted many researchers over the last few decades due to its tremendous applications in the field of biometric authentication. Although, in the field of speaker recognition research spans over about 2-3 decades and a lot of methods / techniques have been developed & implemented in different practical applications but, still there are many open issues that are required to be addressed so as to obtain the accurate / desired results. In this paper, a state of the art in speaker recognition has been presented with emphasis on speaker modeling techniques. Moreover, detailed insight has been presented for each categorization of modeling techniques. Important advancements in chronological order in the field of speaker recognition are also presented.

Keywords: biometrics, identification, soft computing, speaker modeling, speaker recognition, statistical.

1. INTRODUCTION

Recognition of an individual using voice can be put under the category of non-contact identification technology. Speaker recognition can be considered as performance biometric i.e. one has to perform a task to be recognized [1]. It is a dynamic biometric as compared to static biometrics like fingerprint, face, iris [2, 3]. Speaker recognition is an application of the broad field of pattern recognition. The goal of any speaker recognition system is to work like a human while recognizing a speaker. As this is a common and natural ability of human beings but for machines this task is not so simple. Although, in the field of speaker recognition research spans over about 2-3 decades and a lot of methods / techniques have been developed & implemented in different practical applications but, still there are many open issues that are required to be addressed so as to obtain the accurate / desired results. Basic components of a speaker recognition system are shown in (Figure 1). The basic idea is to extract speaker specific features and then apply a technique to identify the speaker.

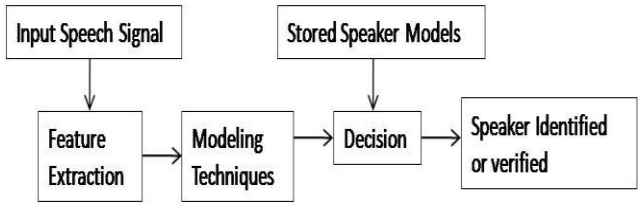

Figure 1: Components of speaker recognition System

The remainder of this paper is organized as follows. In Section 2 we discuss the various classifications of speaker recognition. Section 3 presents the state-of-the-art modeling techniques for speaker recognition. Section 4 presents the chronological advancements based on the modeling techniques provided in section 3. Finally the paper is concluded in section 5.

2. CLASSIFICATION OF SPEAKER RECOGNITION SYSTEMS

Broadly, speaker recognition is classified as verification and identification task [4]. In speaker verification (SV) it is decided whether or not a particular speaker produced the utterance while in speaker identification (SI) person's identity is chosen from a set of known speakers [5, 6]. SV comparatively results in faster computations and less complexity then SI as it involves binary comparison (either yes or no) while SI involves N+1 Decisions (where N are the registered users). Performance of SV is independent of population size while the performance of SI degrades with increase in number of speakers. Verification finds applications in security based transactions where firstly the user provides some pin or enrolment number and then the stored data of that particular number is matched with current data and acceptance or rejection is made. Identification finds applications in automated ID tagging or in forensics where random data is to be matched with the whole database and person is needed to be identified.

Other classification of Speaker recognition is made on

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 9 • Issue 2 pp. 186-195 June 2017-Dec 2017 www.csjournals.com

the basis of whether the identification is to be made from a group of N known speakers (closed set) or by adding the option "None of the above" to closed set recognition which leads to open set recognition. Generally open set recognition is implemented only in verification as identification among infinite users is not possible [7].

Another very important classification of speaker recognition is on the basis of data available for training and testing. If the system is modeled for fixed data then it comes under category of text- dependent system while if the data is different for training and testing then the system qualifies to be text-independent. So we can say the model is utterance specific and includes temporal dependencies between feature vectors in text-dependent systems while in text-independent systems feature distribution is modeled. For text- independent systems higher speech data is required for training and testing. Although text-dependent systems give better performance but sometimes complex speech recognition is required in such systems. In real life situations text-independent recognition systems are more in demand. Both systems can be defeated if recorded voice of registered speaker is played. Possible solution to this condition is to go for text-prompted systems in which a new text is prompted every time one operates a system. So using this technique recorded voice of an authorized user can also be rejected making the systems perform in real-time [8].

3. SPEAKER MODELING TECHNIQUES

Depending upon type of speech, ease of training, computational and storage requirements and expected performance the selection of modeling technique is made [9]. Modeling Techniques can be categorized as shown in (Figure 2).

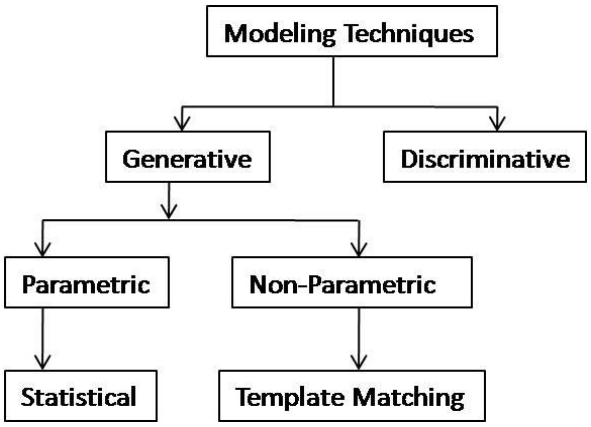

Figure 2: Classification of modeling techniques

Broadly the modeling techniques are classified as:
1. Generative Models
2. Discriminative Models

All the three sub-categories i.e. parametric, non-parametric and discriminative models can be further classified as statistical models, template matching models and soft-computing models respectively as discussed in further subsections.

3.1 Generative Models

Generative models estimate feature distribution within each speaker [10]. They only require training data samples from target speaker and form a statistical/non-statistical model that describes target speaker's feature distribution. It includes models like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), and Vector Quantization (VQ).

3.1.1 Parametric Models

Parametric models are those which assume structure characterized by specific parameters. It includes models like GMM, HMM. Form is limited and fewer data is needed to specify the density. Advantages of these models are that efficient use of data is there, it is possible to model and understand changes in data through changes in parameters and statistical summaries can be used rather complete data [7]. Disadvantage is that the structure is restrictive. GMM is the general parametric model which represents each speaker by a pdf (probability density function) governing the distribution of his/her feature vectors. It has the ability to form smooth densities of irregular shape [11]. In GMM for reliable density estimation the number of required training samples grows exponentially with number of features. This is also called as Curse of Dimensionality [12]. GMM can also use temporal information i.e. state transition probabilities; in that case it results in continuous HMM.

3.1.1.1 Probabilistic or Statistical Models

In statistical models, the pattern matching is probabilistic and results in a measure of likelihood or conditional probability of the observation given the specific model. Classification is based on probabilities or likelihoods rather than distances to average features. Each speaker is modeled as a probabilistic source with a fixed pdf. In training, parameters of pdf are estimated. There are two models: HMM (Hidden Markov Model) [13] and GMM (Gaussian Mixture Model) [14] . HMM are generally used for text-dependent tasks while GMM are used for Text-independent tasks. These can be trained using Baum-Welch or Forward-backward algorithm and Expectation-Maximization (EM) algorithm [15]. For text-independent speaker recognition

where there is no prior knowledge of what will be said the most successful model is GMM while when we have prior knowledge of what will be said as in text-dependent case additional temporal knowledge can be added using HMM [14].

Schwartz et al. [16] showed that using pdf estimation for text independent speaker identification under all conditions gave better results than distance metric methods like Mahalanobis distance method. They tested both parametric and non-parametric methods for pdf estimation and concluded that when less number of parameters are there non-parametric methods give good result but as number of parameters are increased performance of parametric methods is improved drastically.

Reynolds and Rose [14] introduced the GMM for Text Independent Speaker Identification. Some general speaker dependent spectral shapes which are effective for modeling speaker identity can be represented by individual Gaussian components of GMM. Comparisons of GMM for speakers with other modeling techniques like Unimodal Gaussian classifier, VQ codebook, Tied GMM and radial basis function showed that GMM outperforms other methods for Speaker Identification. But to achieve good identification, model order has to be kept high.

After that an important concept of Universal Background Model (UBM) to model a universal imposter was introduced in [17]. The GMM-UBM system is build around likelihood ratio test for speaker verification, using GMM as likelihood function and a UBM for representing alternative speakers. Effectiveness of this method was proved and as an alternative to individual speaker GMMs, an UBM can be trained and then speaker GMMs can be adapted using individual speaker data as adaptation data using Maximum A- Posteriori (MAP) adaptation. Although GMM has the advantage of being computationally inexpensive and insensitive to temporal aspects of speech but it is also a disadvantage as higher levels of information about speaker conveyed in temporal speech is not used.

GMM method with diagonal covariance matrix is widely used in speaker recognition field. But to enhance the speaker recognition performance, larger feature set is preferable which accounts for larger number of mixtures. These necessities require more storage and complexity of computation is increased. Seo, et al. [18] proposed a method to reduce dimension of feature vector using Principal Component Analysis (PCA). To obtain same performance number of mixtures required for conventional GMM is 4 times more than that required by their proposed method.

Zeinali, et al. [19] reported that by increasing the number of registered speakers in Speaker Identification (SI) systems, computation time for identifying an unknown speaker is significantly increased. Due to this limitation, we cannot use conventional Speaker Identification (SI) methods in real time applications. In this paper, a two-step method to overcome this limitation was proposed. They use different identification methods for each step. In the first step they reduce the search space using Nearest Neighbour method. In the second step they identify the target speaker using the conventional GMM based SI method. The experimental results show 3.4 times speed-ups without any accuracy loss using the proposed method.

In the case of speaker verification, the difference between two utterances can be due to inter-speaker variability or inter-session/channel variability. Although in state of the art methods for speaker verification inter-speaker variability is primarily important but inter-session variability also cannot be ignored. Over the last few years, to deal with speaker and channel/session variabilities in Gaussian Mixture Models (GMMs) the Joint Factor Analysis (JFA) approach has become the state of the art in the speaker verification field [20]. JFA is a model of speaker and channel variability in Gaussian mixture models (GMMs). As a model of inter-session variability, eigen channel maximum a posteriori (MAP) and as models of inter-speaker variability ,classical MAP and eigen voice MAP, are used to produce a joint model which is referred to as a joint factor analysis of speaker and channel effects . One major drawback of JFA is that it requires a training set in which a speaker is recorded under all possible channel conditions which are sufficient to cover most of the channel variation that is likely to be encountered at recognition time. This drawback led to another approach to the problem of session variability referred to as eigen channel modeling [21]. This model only attempt to deal with channel effects at recognition time and not at enrollment time but it is easier to implement and it puts less conditions on training data. But eigen channel model is weaker in that it only compensates for channel effects in test utterances whereas the joint factor analysis model handles channel effects in enrollment utterances as well.

Recently Dehak, et al. [22] proposed a new front-end factor analysis technique, termed i-vector extraction has evolved from JFA. The main difference between JFA and i-vectors is that i-vectors do not distinguish between speaker and channel space. Both work with a total variability space containing simultaneously speaker and channel variabilities, whereas JFA treats both spaces individually. This approach has the advantage that scoring uses a simple Cosine Similarity Scoring (CSS) kernel directly to perform verification, making the scoring process faster and less complex than other speaker verification methods, including

JFA or SVM super-vector approaches.

3.1.2 Non-Parametric Models

In non-parametric models minimal assumptions regarding pdf is made. The most popular methods are Nearest Neighbour (NN)[23], Vector Quantization (VQ) [24], and Dynamic Time Warping (DTW)[25]. In VQ centroid of each cluster of features is computed. Collection of centroids is called codebook. Codebook for each speaker is generated using Linde Buzo Gray (LBG) algorithm [26]. Identified speaker is the one with minimum distortion. VQ is identical to NN model except that distances to nearest data representations are measured. Therefore it reduces computation and memory demands as compared to NN. Advantages of VQ are reduced storage for spectral analysis information. Due to discrete representation of speech sounds we can associate a phonetic label with each codebook. Disadvantages of VQ include inherent spectral distortion since finite numbers of codebook vectors are there therefore there is certain level of quantization error. As the size of codebook increases quantization error decreases [27].

3.1.2.1 Template Matching

In template (time ordered set of features) models pattern matching is deterministic. The test data is assumed to be an imperfect replica of the training data and distance is measured and minimum distance template is chosen. Many different distance measures are there like Euclidean, Itakura, Mahalanobis [1]. This method does not require any model training. DTW is mainly used for text-dependent systems. Template models can be further categorized as time dependent (e.g. DTW) or time independent (e.g. VQ). In time independent models all temporal variations are ignored. A time dependent model is more complicated as it must accommodate variability of human speaking rate. In VQ larger codebook size is required for better efficiency and also VQ codebook has to be updated time to time to compensate intra speaker variations.

Soong et al. [24] used vector quantization codebook to identify the speakers based on minimum distance rule. LPC vectors were used as feature vectors. Memory requirements for storage and computational complexity become large when feature vectors are used directly, so training data is efficiently compressed using concept of codebook. LPC vectors were vector quantized using N codebooks corresponding to N speakers. Distortion for whole test sequence was accumulated with respect to each codebook. Codebook with minimum distortion was representative of speaker. But in this method larger codebook size was required for better efficiency and also VQ codebook had to be updated from time to time to compensate the intra-speaker variations.

3.2 Discriminative Models

Discriminative models model the boundary between speakers. These require training data for both target and non-target speakers and derive an optimal separation between the different speakers. It includes models like Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs). These modeling techniques can also be termed as soft computing modeling. Decision function between speakers is trained rather forming individual speaker models [28]. Advantages of these models include flexible architectures and discriminate training power but the optimal structure is selected by trial and error. SVM classifiers separate complex regions through non linear boundary. SVMs can achieve comparable or superior performance to GMM with much less training data.

Under the umbrella of soft computing models major methods include Artificial Neural Networks (ANNs) [29, 30] and Support Vector Machines (SVMs) [31, 32]. These models are explicitly trained to discriminate between speakers. Being a non-linear classifier, ANN has the ability to discriminate the characteristics of different speakers but the performance of ANN is inferior to that of GMM [33]. Drawback of any ANN technique is that the complete network must be retrained when a new speaker is added to the system. Also topology design of network and initial weight settings are not theoretically supported and often neural network get stuck in local minima.

Mafra and Simões [34] used Self Organizing Map (SOM) neural networks to implement speaker recognition task. The voice of each speaker is modeled by a SOM, trained to specialize in the quantization of feature vectors (MFCCs) extracted from his voice. When a test sample is presented, it is quantized by all SOMs that compete for the speaker and the SOM with smallest quantization error identifies the speaker.

Campbell, et al. [32] applied SVM to task of speaker recognition. SVMs have proven to be a powerful technique for pattern classification. SVMs map inputs into a high-dimensional space and then separate classes with a hyper plane. The sequence kernel was based upon generalized linear discriminants which proved to have low computational complexity and more accuracy. EER (Equal Error Rate) and minDCF (minimum Detection Cost Function) calculations showed that SVM performed comparably to GMM. Additionally, the SVM was shown to provide complementary scoring information resulting in substantially lower error rates when it was fused with a

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 9 • Issue 2 pp. 186-195 June 2017-Dec 2017 www.csjournals.com

GMM system.

Since speaker identification is a multiclass classification problem, SVM is not suitable directly as conventional SVM's usually solve two class problems. Fuzzy SVM can solve this problem but due to large training data lower identification rate occurs. So YuJuan, et al. [35] proposed a novel speaker identification method based on Fuzzy c-means and Fuzzy SVM where first FCM clustering technique is used to partition whole training data into several clusters and after that FSVM is trained by cluster centers to make final decision. Also identification rate of FCM+FSVM was more than that of FSVM alone.

Early work in text-dependent speaker recognition was dominated by template models. But currently statistical and soft computing models have offered more flexibility and better results.

4. Chronological Advancements

Research in the area of speaker recognition has been an active area for several decades and due to diversity in the field (because of different parameters like choice of features, modeling techniques and different databases used) direct comparisons are tough to make. Advantages and disadvantages of various modeling techniques along with relevant literature references have been discussed in section 3. Research in this field has focused on finding novel set of features, reducing computational loads, improving noise robustness, modeling through different techniques e.g. statistical and soft computing techniques to improve the efficiency of ASR systems. Table 1 presents the chronological advancements in the field of speaker recognition.

Experimental demonstration of the task of speaker identification for a basic 2 speaker classification has been carried out by using MFCC features and Neural Network (NN) classifier. 20 dimensional MFCC (Mel Frequency Cepstral Features) are extracted as shown in Figure. 3 and neural pattern recognition tool in MATLAB is used as a classifier. Plot of confusion matrix and Receiver operating characteristics (ROC) are shown in Figure. 4 and Figure. 5. As this paper aims at facilitating new researchers with present scenario of modeling techniques in the field of speaker identification, details about features, their extraction and presentation of all simulation results obtained are beyond the scope of this paper.

At present the research in field of speaker recognition is emphasizing more on novel features for speech signal representation and hybrid applications of above discussed different modeling techniques. Many efforts are being put into improving recognition under the conditions of session variabilities and noise.

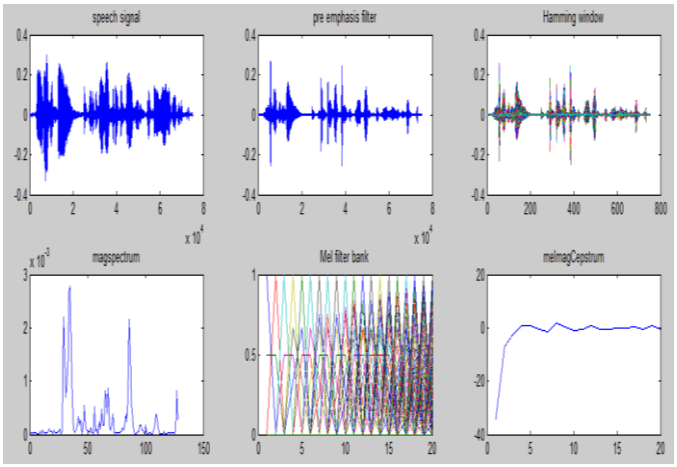

Figure 3: 20 dimensional MFCC Feature extraction in MATLAB

5. Conclusion

Traditionally access control was based on token-based and knowledge-based identification systems. Due to uniqueness of biometric identifiers to individuals, they are more reliable in authentication than token and knowledge-based methods. Using the voice as a biometric to verify or search for the identity of users is currently the most natural way. High security applications require speaker recognition to perform almost perfectly, which is quite difficult, especially when dealing with hundreds of speakers and thousands of possible imposters. Computationally powerful back-ends and high volumes of speech data are required. Challenge lies in applications where real time decision is required.

Speaker recognition has achieved fairly good performance under controlled conditions as reported in the NIST annual speaker recognition evaluations. But in real conditions mismatches exist between training and testing phases, such as wide band vs. narrow band, quite room environment vs. noisy street environment, and land-line channel vs. cell phone channel etc. These factors consequently induce performance degradation in automatic speaker recognition systems. Apart from these, factors like illness, aging, emotional variations also affect the accuracy of systems. Many techniques that have proven to be effective for countering these limitations are Universal Background Modeling, Score Normalizations, Feature compensation and Missing data approaches to name a few. One major challenge of Automatic Speaker Recognition (ASR) is its very high computational cost. Therefore research has been

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 9 • Issue 2 pp. 186-195 June 2017-Dec 2017 www.csjournals.com

focusing on decreasing the computational load of identification while attempting to keep the recognition accuracy reasonably high. Nevertheless, voice based biometrics is still a very natural and promising method for biometric authentication and hence speaker recognition is worthy of more efforts to be put in for its improvement.

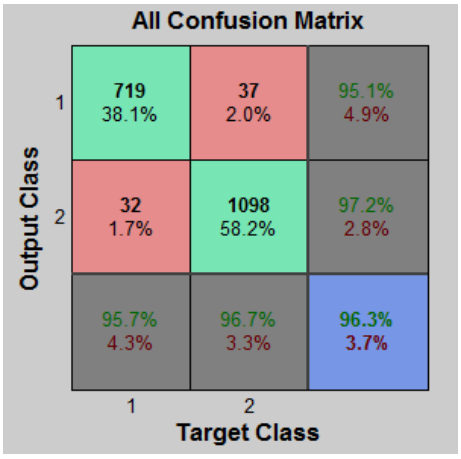

Figure 4: Confusion matrix from NPR tool in MATLAB (96.3 % identification rate is achieved)

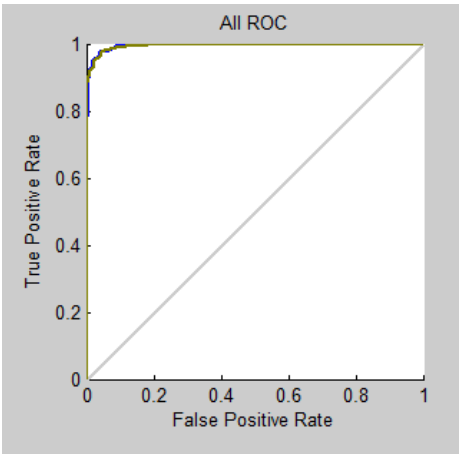

Figure 5: ROC from NPR tool in MATLAB

Table 1: Chronological advancements in text-Independent speaker recognition (M: Male, F: Female, MFCC: Mel Frequency Cepstral Coefficients, LPCC: Linear Prediction Cepstral Coefficients)

Year	Features (Coefficients)	Method Used	Input (Database)	Population (No. of speakers)	Results
1976 [36]	12 Linear Prediction orthogonal Parameters	Distance Generator (using mean values)	Quiet Environment + 6 Recordings per day for 6 different days (Self generated database)	21 M	93.7% Identification Rate (IR)
1990 [11]	20 MFCC coefficients	GMM	Studio conditions + Conversational Database	12 (8 M+4 F)	89% (IR)

Year	Features (Coefficients)	Method Used	Input (Database)	Population (No. of speakers)	Results
1993 [29]	Mean rate Response	Predictive NN + Min. Prediction error	TIMIT database	24 F	95.7 % (IR) for 1-state model
1994 [37]	12 LPCC coefficients	Modified Tree NN with pruning criteria	TIMIT database	38 (24 M+ 14 F)	7 pruning level (98% IR) 4 pruning level (88% IR)
1995 [38]	30 MFCC	GMM	TIMIT and NTIMIT (Training 24 sec) (Testing 3 sec)	630 (438 M + 192 F)	TIMIT 99.5% IR NTIMIT 60.7% IR
1999 [39]	16 MFCC features	VQ	YOHO database	40 (20 M +20 F)	Codebook size (64) LBG 92.7% IR GVQ 95.7% IR
2001 [18]	12 LPC cepstrum+ 13 delta cepstrum	GMM + PCA with VQ	Self generated database	50 (25 M + 25 F)	98-99 % IR
2003 [33]	19 MFCC + 19 delta MFCC	Structural GMM+ Multilayer feed forward NN	Switchboard II Telephony speech data	230 M + 309 F	Computational reduction of factor 17 with 5% reduction in Equal Error Rate (EER)
2004 [40]	Lower formants (f1,f2,f2-f1)	DTW	Self Gene rated database	26 speakers	92% IR
2006 [32]	36 LPCC and 38MFCC	SVM and GMM	NIST 2003 speaker and language recognition corpus	356 speakers	SVM-LPCC EER 7.72 SVM-MFCC EER 9.57 SVM-L +GMM EER 5.73
2006 [41]	LP residual and 19 LPCC	AutoAssociative Neural Network (AANN)	NIST 2001 and 2002 database	5 sets of 20 male speakers each	IITM2 EER 23.8 IITM1 EER 17.2 OGI1 + IITM1 + IITM2 EER 7.1
2007 [42]	Wavelet octave Coefficients of	GMM-UBM	NIST 2001 database	74 M	MFCC+WOCOR EER 7.67% MFCC EER 9.30%

Year	Features (Coefficients)	Method Used	Input (Database)	Population (No. of speakers)	Results
	residues (WOCOR) + MFCC				
2009 [43]	60 MFCC	SVM + Joint Factor Analysis (JFA)	NIST 2006 SRE Dataset	JFA (300 Speaker factors and 100 channel factors) SVM (1875 Imposters)	EER 4.23%
2010 [28]	DWT Coefficients (Daubechis, Symlets, Coiflets)	GMM + Multilayer perceptron (MLP)	Two Czech Speaker Corpora	I-10 (5 M + 5 F) II- 50 (25 M + 25 F)	98% IR
2011 [44]	18 MFCC	Fuzzy Min-Max Network	Self Generated database (Marathi Language)	50 speakers	99.9% IR
2012 [45]	MFCC + Phase Information	GMM	NTT database and JNAS database	35 (22 M + 13 F)	98.8% IR
2012 [35]	13 MFCC + 13 delta(Δ) + 13 (ΔΔ)	Fuzzy c-means + Fuzzy SVM	King Speech Database	51 speakers	FSVM 94.53% IR FCM+ FSVM 98.76% IR
2013 [46]	Temporal Teager Energy based Subband Cepstral Coefficients (TTESBCC)	GMM	Self Generated Database (Marathi Speech) (Neutral and Whisper)	25 speakers	For Neutral speech 98.62% IR For Whisper speech 55.8% IR
2013 [47]	MFCC	HMM+ GFM HMM+GMM	VoxForge Speech Corpus and NIST 2003 evaluation data set	100 speakers	HMM+GFM 92% IR GMM+GFM 92% IR
2015 [48]	i-vector d-vector	DNN, DTW, PLDA	Self generated 10 phrases 2~5 Chinese characters	100 speakers	EER ~ 2%
2017 [49]	Modified MFCC's	Neural Network (NN)	ELSDSR	22 speakers	91.9 % IR

REFERENCES

[1] J. P. Campbell Jr, "Speaker recognition: a tutorial," Proceedings of the IEEE, vol. 85, pp. 1437-1462, 1997.
[2] C. D. Shaver and J. M. Acken, "Effects of equipment variation on speaker recognition error rates," in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 1814-1817.
[3] S. K. Singla and A. S. Arora, "A Karnaugh-Map based fingerprint minutiae extraction method," Sonklanakarin Journal of Science and Technology, vol. 32, pp. 247-254, 2010.
[4] S. Furui, "An overview of speaker recognition technology," in Automatic speech and speaker recognition, ed: Springer, 1996, pp. 31-56.
[5] R. D. Peacocke and D. H. Graf, "An introduction to speech and speaker recognition," Computer, vol. 23, pp. 26-33, 1990.
[6] S. K. Singla and A. S. Arora, "Speaker Verification System using LabVIEW," IETE Technical Review, vol. 24, pp. 403-412, 2007.
[7] H. Gish and M. Schmidt, "Text-independent speaker identification," Signal Processing Magazine, IEEE, vol. 11, pp. 18-32, 1994.
[8] S. Furui, "50 years of progress in speech and speaker recognition," SPECOM 2005, Patras, pp. 1-9, 2005.
[9] D. Reynolds, "An overview of automatic speaker recognition," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075), 2002.
[10] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," Circuits and Systems Magazine, IEEE, vol. 11, pp. 23-61, 2011.
[11] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on, 1990, pp. 293-296.
[12] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, pp. 4-37, 2000.
[13] A. E. Rosenberg, C.-H. Lee, and S. Gokcen, "Connected word talker verification using whole word hidden Markov models," in Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on, 1991, pp. 381-384.
[14] D. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," Speech and Audio Processing, IEEE Transactions on, vol. 3, pp. 72-83, 1995.
[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), pp. 1-38, 1977.
[16] R. Schwartz, S. Roucos, and M. Berouti, "The application of probability density estimation to text-independent speaker identification," in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82., 1982, pp. 1649-1652.
[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital signal processing, vol. 10, pp. 19-41, 2000.
[18] C. Seo, K. Y. Lee, and J. Lee, "GMM based on local PCA for speaker identification," Electronics Letters, vol. 37, pp. 1486-1488, 2001.
[19] H. Zeinali, H. Sameti, and B. Babaali, "A fast Speaker Identification method using nearest neighbor distance," in Signal Processing (ICSP), 2012 IEEE 11th International Conference on, 2012, pp. 2159-2162.
[20] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, pp. 1448-1460, 2007.
[21] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, pp. 1435-1447, 2007.
[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, pp. 788-798, 2011.
[23] A. Higgins, L. Bahler, and J. Porter, "Voice identification using nearest-neighbor distance measure," in Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, 1993, pp. 375-378.
[24] F. K. Soong, A. E. Rosenberg, B.-H. Juang, and L. R. Rabiner, "Report: A vector quantization approach to speaker recognition," AT&T technical journal, vol. 66, pp. 14-26, 1987.
[25] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 26, pp. 43-49, 1978.
[26] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," Communications, IEEE Transactions on, vol. 28, pp. 84-95, 1980.
[27] L. R. Rabiner and B.-H. Juang, Fundamentals of speech recognition vol. 14: PTR Prentice Hall Englewood Cliffs, 1993.
[28] P. Král, "Discrete Wavelet Transform for automatic speaker recognition," in Image and Signal Processing (CISP), 2010 3rd International Congress on, 2010, pp. 3514-3518.
[29] H. Hattori, "Text-independent speaker recognition using neural networks," IEICE TRANSACTIONS on Information and Systems, vol. 76, pp. 345-351, 1993.
[30] M. A. Franzini, M. J. Witbrock, and K.-F. Lee, "Speaker-independent recognition of connected utterances using recurrent and non-recurrent neural networks," in Neural Networks, 1989. IJCNN., International Joint Conference on, 1989, pp. 1-6.
[31] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in NEURAL NETWORKS SIGNAL PROCESS PROC IEEE, 2000, pp. 775-784.
[32] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Computer Speech & Language, vol. 20, pp. 210-229, 2006.
[33] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural gaussian mixture models and

neural network," Speech and Audio Processing, IEEE Transactions on, vol. 11, pp. 447-456, 2003.

[34] A. T. Mafra and M. G. Simões, "Text independent automatic speaker recognition using selforganizing maps," in Industry Applications Conference, 2004. 39th IAS Annual Meeting. Conference Record of the 2004 IEEE, 2004, pp. 1503-1510.

[35] X. YuJuan, L. Hengjie, and T. Ping, "Hierarchical fuzzy speaker identification based on FCM and FSVM," in Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on, 2012, pp. 311-315.

[36] M. Sambur, "Speaker recognition using orthogonal linear prediction," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 24, pp. 283-289, 1976.

[37] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," Speech and Audio Processing, IEEE Transactions on, vol. 2, pp. 194-205, 1994.

[38] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," Signal Processing Letters, IEEE, vol. 2, pp. 46-48, 1995.

[39] J. He, L. Liu, and G. Palm, "A discriminative training algorithm for VQ-based speaker identification," Speech and Audio Processing, IEEE Transactions on, vol. 7, pp. 353-356, 1999.

[40] N. Fatima, S. Aftab, R. Sultan, S. A. H. Shah, B. M. Hashmi, A. Majid, et al., "Speaker recognition using lower formants," in Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International, 2004, pp. 125-130.

[41] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," Speech Communication, vol. 48, pp. 1243-1261, 2006.

[42] N. Zheng, T. Lee, and P.-C. Ching, "Integration of complementary acoustic features for speaker recognition," Signal Processing Letters, IEEE, vol. 14, pp. 181-184, 2007.

[43] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, et al., "Support vector machines and joint factor analysis for speaker verification," in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009, pp. 4237-4240.

[44] N. Jawarkar, R. Holambe, and T. Basu, "Use of fuzzy min-max neural network for speaker identification," in Recent Trends in Information Technology (ICRTIT), 2011 International Conference on, 2011, pp. 178-182.

[45] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, pp. 1085-1095, 2012.

[46] N. P. Jawarkar, R. S. Holambe, and T. K. Basu, "Speaker Identification Using Whispered Speech," in Communication Systems and Network Technologies (CSNT), 2013 International Conference on, 2013, pp. 778-781.

[47] S. Bhardwaj, S. Srivastava, M. Hanmandlu, and J. Gupta, "GFM-Based Methods for Speaker Identification," Cybernetics, IEEE Transactions on, vol. 43, pp. 1047-1058, 2013.

[48] Y. L. Lantian Li, Zhiyong Zhang, Dong Wang, "Improved Deep Speaker Feature Learning for Text-Dependent Speaker Recognition," in APSIPA Annual Summit and Conference, 2015, pp. 426-429.

[49] H. M. Mohammad Soleymanpour, "Text-independent speaker identification based on selection of the most similar feature vectors," International Journal of Speech Technology, vol. 20, pp. 99-108, 2017.