Speaker Recognition: A Research Direction

Rita Yadav
Student, M.Tech (ECE)
IET Bhaddal
Ropar, India
rita.yadavs@gmail.com

Danvir Mandal
Assistant Professor (ECE Depts.)
IET Bhaddal
Ropar, India
danvir_mandal@rediffmail.com

***Abstract*— Speaker recognition system can be divided in four stages, namely, analysis, feature extraction, modeling and testing. This paper gives an overview of major techniques developed in each stage so far. Such a review helps in understanding the developments that have been taken place in each stage and also available choices of techniques.**

***Keywords*— speech analysis, feature extraction, speaker modeling,testing, speaker recognition, Particle Swarm Optimization (PSO).**

I. INTRODUCTION

Human speech is the foundation of Self-expression and communication with others. The goal of speaker recognition system is to analyze, extract, characterize and recognize information about the speaker identity [1,2]. Depending on the task, speaker recognition can be classified as speaker verification and speaker identification. The speaker verification involves a yes/ no decision to identify whether the speaker is who he/she claims to be. In speaker identification, the system identifies the speaker from some known group of speakers. Speaker recognition can be further categorized into two parts: a closed-set problem and an open-set problem. The closed set problem is to identify who makes a specific input speech among N speakers. Therefore, a difficulty in identification increases as N increases. The open set problem is to determine whether a claimed speaker makes a specific speech or not. It becomes a binary decision of determining whether the input speech is from a claimed speaker or not. The open-set problem is usually called the speaker identification and the closed-set problem is usually called the speaker verification [3,4]. Depending upon the mode of operation, speaker recognition can be classified into text-independent speaker recognition and text-dependent speaker recognition. In the text-independent speaker recognition, the input speech is unconstrained so that a user may make any speech he/she will. But in the text-dependent speaker recognition, a user makes a pre-defined keyword such as password [5].

Speaker recognition system may be viewed as working in four stages, namely, analysis, feature extraction, modeling and testing. The speech analysis stage deals with the selection of suitable frame size and frame shift for segmentation of speech for further analysis and feature extraction. The speech analysis is done using one of the following techniques: Segmental analysis, sub-segmental and supra-segmental analysis. The feature extraction stage deals with extracting the relevant speaker-specific information in terms of feature vectors. The modeling techniques may be either generative type or discriminative type. One model is built for each enrolled speaker. During testing, the speech signal is analyzed and features are extracted using same techniques employed during training. The feature vectors are compared with reference models using distance measure techniques and based on the comparison results, the speaker in the test will be recognized.

The performance of speaker recognition system depends on the techniques employed in the various stages of the speaker recognition system.

This paper is structured as follows. Section 2 describes speech analysis techniques. Section 3 describes, feature extraction techniques. Section 4 describes, speaker modeling techniques. Section 5 describes speaker testing and decision logic techniques. Finally, we have some conclusions in section 6.

II. SPEECH ANALYSIS TECHNIQUES

Speech data contains different types of information that convey speaker identity. These include speaker-specific information due to the vocal tract, excitation source and behavioral traits. The speech signal is produced from the vocal tract system by varying its dimension with the help of articulators and exciting with a time varying source of excitation. The physical structure and dimension of vocal tract as well as excitation source are unique for each speaker. This uniqueness is embedded in the speech signal during speech production and can be used for speaker used for speaker recognition. The behavioral tracts like how the vocal tract and excitation source are controlled during speech production are also unique for each user. The information about behavioral tracts is also embedded in the speech signal and can be used for speaker recognition.

In order to obtain good representation of these speaker characteristics, speech data needs to be analyzed using a suitable analysis technique. The analysis technique aims at selecting proper frame size and shift for analysis and also for extracting the relevant features in the feature extraction stage. Speaker recognition systems mainly employ the following analysis techniques.

A. Segmantal analysis

In this case, speech is analyzed using the frame size and shift of 10-30 ms to extract speaker information mainly due to vocal tract. The speaker –specific vocal tract information may be assumed to be stationary for all practical analyses and

processing when viewed in frames and shift in the range of 10-30 ms [6, 7, 8].

B. Sub-segmental analysis

In this case, speech is analyzed using the frame size and shift of 3-5 ms to extract speaker information mainly due to excitation source [9]. The excitation source information is relatively fast varying compared to vocal tract information, so small frame size and shift are required to best capture the speaker –specific information [10-15].

C. Supra-segmental analysis

In this case, speech is analyzed using the frame size and shift of 100-300 ms to extract speaker information mainly due behavioral tract. These include word duration, intonation, speaker rate, accent etc. the behavioral tracts vary restively slowly compares to the vocal tract information, which is the reason for the choice of large frame size and shift [11, 16-18].

Most of the speaker recognition system mainly uses the segmental analysis technique. Therefore, Speaker-specific vocal tract information is mainly used for speaker recognition. The, speaker-specific vocal tract information is one of the rich speaker information source present in the speech signal. We can also use speaker-specific excitation source information extracted using sub-segmental analysis and speaker-specific information representing behavioral trait extracted using supra-segmental analysis. Such a process will provide improved representation and modeling of the speaker and hence improved performance. Thus apart from the existing segmental analysis, we can also use sub-segmental and supra-segmental analysis techniques in the analysis stage.

III. FEATURE EXTRACTION TECHNIQUES

The purpose of feature extraction stage is to extract the speaker-specific information in the foam of feature vectors. The feature vectors represent the speaker-specific information due to one or more of the following: Vocal tract, excitation source and behavioral tracts. A good feature set should have representation due to all of the components of speaker information. To develop such a speaker a good feature set, it is necessary to understand the different feature extraction techniques. This section describes the same.

Spoken digit recognition conducted by P Denes in 1960, suggested that inter-speaker differences exists in the spectral patterns of speakers [19]. S Pruzansky, motivated from this study, conducted the first speaker identification study in 1963. In his study, spectral energy patterns were used as the features. It was shown that the spectral energy patterns yielded good performance, confirming the usefulness for the speaker recognition [20]. Further he reported a study using the analysis of variance in 1964 [21]. In this work, a subset of features was selected from the analysis of variance using F ratio test defined as the ratio of the variance of the speaker means to average within speaker variance [21]. It was reported that the subset of features provided equal performance, thus significantly reducing the number of computations. Speaker verification study was first conducted by Li in 1966 using adaptive linear threshold elements [22]. This study used spectral representation of the input speech, obtained from the bank of 15 band pass filters spanning the frequency range 300-4000Hz. Two stages of adaptive linear threshold elements operate on the rectified and smoothed filter outputs. These elements are trained with speech utterances. The training process results in a set of weights that characterize the speaker. This study demonstrated that the spectral band energies as feature contain speaker information. A study by Glenn in 1967 suggested that acoustic parameters produced during nasal phonation are highly effective for speaker recognition [23]. In this study, average power spectral of nasal phonation was used as the features for the speaker recognition. In 1969, Fast Fourier Transform (FFT) based cepstral coefficients were used in speaker verification study [7]. In this work, a 34- dimensional vector was extracted from speech data. The first 16 components were from FFT spectrum, the next 16 were from log magnitude FFT spectrum and the last two components were related to pitch and duration. Such a 34-dimensional vector seems to provide a good representation of speaker.

In 1972 Atal demonstrated the use of variations in pitch as a feature for speaker recognition [16]. In addition to the variation in pitch, other acoustic parameters such as glottal source spectrum slope, word duration and voice onset were proposed as features for speaker recognition by Wolf in 1971 [24]. The concept of linear prediction for speaker recognition was introduced by Atal in 1974 [25]. In this work, it was demonstrated that Linear Prediction Cepstral Coefficients (LPCCs) were better than the Linear Prediction Coefficients (LPCs) and other features such as pitch and intensity.

Earlier studies neglected the features such as formant bandwidth, glottal source poles and higher formant frequencies, due to non-availability of measurement techniques. The studies introduced after the linear prediction analysis, explored the speaker specific potential of these features for speaker recognition [26]. A study carried by Rosenberg and Sambur suggested that adjacent cepstral coefficients are highly correlated and hence all coefficients may not be necessary for speaker recognition [27]. In 1976, Smbur proposed to use orthogonal linear prediction coefficients as feature in speaker identification [28]. In this work, he pointed out that for a speech feature to be effective, it should reflect the unique properties of the speaker's vocal tract and contain little or no information about linguistic content of the speech.. In 1977, long term parameter averaging, which includes pitch, gain and reflection coefficients for speaker recognition was studied [29]. In this study, it was shown that reflection coefficients are informative and effective for speaker recognition. In 1981 Furui introduced the concept of dynamic features, to track the temporal variability in feature vector in order to improve the speaker recognition performance [30, 31]. A study made by G R Doddington in 1985 [8], converts the speech directly in to pitch, intensity and formant frequency, all sampled 100 times per second. These features were also demonstrated to provide good performance.

A study by Reynolds in 1994 compared the different features like Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs), LPCCs and Perceptual Linear Prediction Cepstral Coefficients (PLPCCs)

for speaker recognition [32]. He reported that among these features, MFCCs and LPCCs gave better performance than other features. In 1995 P. Thevenaz and H Hugli [33] reported that Linear Prediction (LP) residual also contains speaker-specific information that can be used for speaker recognition. Also, it has been reported that though the energy of LP residual alone gives less performance, combining it with LPCC improves the performance as compared to that of LPCC alone. Similarly, several studies reported that though the energy of LP residual alone gives less performance, combining it with MFCC improves the performance as compared to that of MFCC alone [10,13-14]. In 1996 Plumpe developed a technique for estimating and modeling the glottal flow derivative waveform from speech for speaker recognition [34]. In this study, the glottal flow estimate was modeled as coarse and fine glottal features, which were captured using different techniques. Also it was shown that combined coarse and fine structure parameters gave better performance than the individual parameter alone. In 1996, M J Carey, E S Paris carried out a study on the significance of long term pitch and energy information for speaker recognition [35]. In 1998, M K Sonmez, E Sriberg carried out a study on pitch tracks and local dynamics for speaker verification [36].

In 2003, B Peskin, J Navratil reported that combination of prosodic features like long-term pitch with spectral features provided significant improvement as comapared to only pitch features [37]. A study by L Mary, K S Rao, B Yegnanarayana in 2004 were carried out on supra-segmental features like duration and intonation capyurd using using neural network for speaker recognition [17]. In 2005, B Yegnanarayana, S R M Prasanna demonstrated the use of features such as long term pitch and duration information obtained using Dynamic Time Warping (DTW), along with source and spectral features for text –dependent speaker recognition [11].In 2008, M Girmaldi, F Cummins carried a study on Amplitude Modulation (AM)-Frequency Modulation (FM)-based parameter of speech for speaker recognition. In this study it was demonstrated that using different instantaneous frequencies due to the presence of formants and harmonics in speech signal, it is possible to discriminate speakers [38].

In 2007, Min-Seok Kim and Ha-Jin Yu introduced a new feature transformation method based on rotation for speaker identification [39]. In this study, they have proposed a new feature transformation method that is optimized for diagonal covariance Gaussian mixture models [58] which is used for a speaker identification system. They first have defined an object function as the distances between the Gaussian mixture components and rotate each plane in the feature space to maximize the object function. The optimal degrees of the rotations are found using the Particle Swarm Optimization [40] algorithm. In 2008, Min-Seok Kim, IL-Ho Yung and Ha-Jin Yu have proposed a feature transformation method to maximize the distance between the Gaussian mixture models for speaker verification using PSO [41].

Among these the most commonly used cepstral coefficients are MFCCs and LPCCs, because of less intra-speaker variability and also availability of spectral analysis tools. However, the speaker-specific information due excitation source and behavioral tract represents different aspects of speaker information. The main limitation for the use of excitation source and behavioral tract is non – availability of suitable feature extraction tools.

IV. SPEAKER MODELING TECHNIQUES

The objective of the modeling techniques is to generate speaker-specific feature vectors. Such models have enhanced speaker-specific information. This is achieved by exploiting the working principles of the modeling techniques. Various modeling techniques are briefly described in this section. Earlier studies on speaker recognition uses direct template matching between training and testing data [7, 20, 23-28]. In direct template matching, training and testing features vectors are directly compares using similarity measures. For similarity measures either of spectral or Euclidean distance or Mahalanobis distance is used.

In 1981 Furui introduced the concept of Dynamic Time Warping (DTW) for text- speaker recognition [31]. In this approach the sequence of feature vectors of the training speech signal is text -dependent template model. The DTW finds the match between the template model and the input sequence of feature vectors from the testing speech signals. The disadvantage of DTW is that it is time consuming, as the number of feature increases. For this reason, it is common to reduce the number of training feature vectors by some modeling techniques like clustering. The cluster centers are known as code vectors and the set of code vectors is known as codebook. The well known codebook generation algorithm is K-means algorithm [42, 43].

In 1985, Soong used the LBG algorithm for generating speaker-based Vector Quantization (VQ) codebooks for speaker recognition [44]. It is demonstrated that larger codebook and larger test data gives good recognition performance. Also study suggested that VQ codebook can be updated from time to time to alleviate the performance degradation due to different recoding and intra speaker variations. The disadvantage of VQ classification is, it ignores the possibility that a specific training vector may also belong to another cluster. As an alternate to this, Fuzzy Vector Quantization (FVQ) using the well-known fuzzy C-means method was introduced by Dunn and its final foam was developed by Bezdek in 1978 [45]. In 1999 and 2006, FVQ was used as classifier for speaker recognition [46,47]. It was demonstrated that FVQ gives better performance than traditional K-means algorithm because of working principal of FVQ is different from VQ in the sense that the soft decision making process is used while designing the codebooks in FVQ [45]; whereas in VQ, the hard decision process is used. In VQ each feature has an association with only one of the clusters; whereas in FVQ, each feature vector has an association with all the clusters, with varying degrees of associations decided by membership function [45]. Since all the feature vectors are associated with all the clusters, there are relatively more numbers of feature vectors for each cluster and hence the representative vectors i.e. code vectors may be more reliable than VQ. Therefore, clustering may be better in FVQ and may lead to better performance than VQ.

In order to model the statistical variations, the Hidden Markov Model (HMM) for text-dependent speaker recognition is studied in [48-50]. In HMM, time-dependent parameters are observation symbols. Observation symbols are created by VQ codebook labels. The main assumption of HMM is that the current state depends on previous state. In training phase, state transition probability distribution, observation symbol probability distribution and initial state probabilities are estimated for each speaker as a speaker model. The probability of observation for a given model is calculated for speaker recognition. Kimbel studied the use of HMM for text-independent speaker recognition under constrained of limited data and mismatched channel condition [51]. In this study, the MFCC features was extracted for each speaker and then models were built using Broad Phonetic Category (BPC) and the HMM- based Maximum Likelihood Linear Regression (MLLR) adaptation techniques. The BPC modeling is based on identification of phonetic categories in an utterance and modeling them separately. In HMM-MLLR, first speaker identification model is created using HMM and MLLR technique is used to adapt SI model to each speaker. It was shown that speaker model built using the adaptation techniques gave better performance than BPC and GMM for cross- channel conditions.

The capability of neural networks to discriminate between patterns of different classes is exploited for speaker recognition [52-54]. Neural network has an input layer, one or more hidden layers and an output layer. Each layer consists of processing units, where each unit represents model of an artificial neuron, and the interconnection between the two units as a weight associate with it. The concept of Multi-Layered Perception (MLP) was used for speaker recognition [55]. In this study, it was demonstrated that one hidden layer network with 128 hidden node gave same performance as that achieved with 64 codebook VQ approach. The disadvantage of MLP is that it takes more time for training the network. This problem was removed using the Radial Basis Function (RBF) [56]. In this study, it was shown that RBF network took lesser time than the MLP and outperformed both VQ and MLP.

Kohonen developed Self-Organization Map (SOM) as an unsupervised learning classier. SOM is a special class of neural network based on competitive learning [57]. Thus, the performance of SOM depends on the parameters such as neighborhood, learning rate and number of iterations. These parameters are to be fined tuned for good performance. The SOM and associative memory model were used together for speaker identification [58]. This was shown that hybrid model gave better recognition performance than MLP. The disadvantage of SOM is that it does not use class information while modeling speakers, resulting in a poor speaker model that leads to degradation in the performance. This can be removed by using Kohonen Learning Vector Quantization (LVQ). LVQ is a supervised learning technique that uses class information to optimize the positions of code vectors obtained by SOM, so as to improve the quality of decision classifier.

In 1995, Reynolds proposed Gaussian Mixture Model (GMM) for speaker recognition [59]. This is the most widely used probabilistic modeling technique for speaker recognition. The GMM needs sufficient data to model the speaker and hence good performance. In GMM modeling technique, the distribution of features vectors is modeled by the parameters mean, covariance and weight. In another study, Reynolds compared GMM performance with regard to speaker identification with that of other classifiers like unimodal Gaussian, VQ, tied Gaussian mixture and radial basis functions [60]. It was shown that GMM outperformed the other modeling techniques. Therefore most of the speaker recognition systems use GMM as classifier due to better performance, probabilistic framework and training methods scalable to large data sets [61].

The disadvantage of GMM is that it requires sufficient data to model the speaker [59]. To overcome this problem, Reynolds introduced GMM-Universal Background Model (UBM) for speaker recognition [62]. In this system speech data collected from a large number of speakers is pooled and the UBM is trained, which acts as a speaker-independent model. The speaker-independent model is then created from the UBM by performing Maximum A Posteriori (MAP) adaptation technique using speaker-specific training speech. As a result, the GMM-UBM gives better results than the GMM. The advantage of UBM-based modeling technique is that it provides good performance even though the speaker-dependent data is small. The disadvantage is that gender-balanced speaker set is required for UBM training.

As an alternate to the GMM, an Auto Associative Neural Network (AANN) has been developed for pattern recognition [54, 63, 64]. AANN is a feed-forward neural network which tries to map an input vector on to itself. The number of units in the input and the output layers is equal to the size of input vectors. The number of nodes in middle layer is less than the number of units in the input or output layers. The activation function of the units in the hidden layer can be either linear or non-linear. The advantage of AANN over GMM is that, it does not impose any distribution.

A learning method based on the statistical learning theory, a special theory on machine learning, is the Supervised Vector Machine (SVM). The SVM has many desirable properties, including ability to classify sparse data without over training. It is basically a solution to a two class problem, but it can be extended to solve multi-class problem. SVM works by increasing the dimensionality of the input data space. The dimensionality is increased until it finds a maximum- margin linear hyper plane that can be used to separate the two classes. This accomplished by using kernels and dot products. The SVM is discriminative in nature, whereas other classifiers are generative in nature.

W M Campbell proposed Generalized Linear Discriminate Sequence (GLDS) kernel for speaker recognition and language identification [61]. In this study, was shown that though the SVM results are much better than GMM, combination of SVM with GMM yielded good recognition performance than the individual systems. Combination of SVM with GMM was also studied for speaker recognition [65, 66].

H. R. S. Mohammadi and R. Saeidi, in 2006, introduced an efficient implementation of GMM based speaker verification using Sorted Gaussian Mixture Model (SGMMs) algorithm

providing the means to tradeoff performance for operational speed and thus permitting the speed-up of GMM-based classification schemes. The performance of the SGMM algorithm depends on the proper choice of the sorting function and the proper adjustment of its parameters [67]. In 2009 they, employed Particle Swarm Optimization (PSO) and an appropriate fitness function to find the most advantageous parameters of the sorting function. They evaluate the practical significance of this approach on the text-independent speaker verification. The experimental results demonstrate a superior performance of the SGMM algorithm using PSO when compared to the original SGMM [68]**.** In 2010, they have employed joint frame and Gaussian selection for text-independent speaker verification [69].In this study, they extend the SGMM method by using 2-dimensional indexing, which leads to simultaneous frame and Gaussian selection.

The various modeling techniques discussed so far may be summarizes as follows. In case of text-dependent speaker recognition, DTW technique is most commonly used. In case of text-independent speaker recognition, we have VQ and its variants like FVQ, SOM AND LVQ. Among these, from simplicity point of view, VQ is mostly used and performance point of view, LVQ is preferred one. The GMM technique is mostly used modeling technique from among the Gaussian classifiers. Among the neural networks, the MLP, RBF and AANN are mostly used. SVM has also been demonstrated to be a potential discriminatory-type classifier for speaker modeling, especially under condition of limited data. Also the GMM-SVM combination has been demonstrated to provide better modeling compared to either GMM or SVM alone. As a final comment, it should be stated that PSO-GMM combination has been demonstrated to provide better modeling techniques these days.

V. Speaker Testing And Decision Logic

Testing stage in speaker recognition system includes matching and decision logic. During testing, usually the test feature vectors are compared with the reference models. Hence matching gives a score which represents how well the test feature vectors are closed to the reference models. Decision will be taken on the basis of matching score, which depends on the threshold value. In the speaker verification system the performance is measured in terms of Equal Error Rate (ERR), which is defined as the error rate at which False Acceptance (FA) rate is equal to the False Rejection (FR) rate. Moreover, the detection probability as a function of false alarm probability, known as Receiver Operating Characteristics (ROC) plot, is also used for the assessment of speaker verification performance. In speaker identification system, performance measurement is simple and direct. This is measured as a ratio of number correctly identified examples to the total number of examples considered for testing.

In both speaker verification and identification, for matching test feature vectors to the reference model, either distance measurement or probabilistic score approaches are used. Earlier studies employed spectral or Eucliden or Mahalonabis distance measurement techniques for comparison [7, 20, 23-28]. Reynolds used the concept of log likehood ratio test for speaker recognition [56]. In 2001, H Jiang and L Deng studied the Bayesian approach for speaker recognition [70]. It was demonstrated that Bayesian approach moderately improved the performance compared to well -trained baseline system using of the conventional likelihood ratio test.

In order to improve the speaker recognition performance at decision level, a combination of multiple classifiers has been proposed [71]. In this study, voting method was used for speaker identification based on the results of various resolution filter banks. A study conducted [11] reported that by combining the evidence from source, supra-segmental and spectra features, it is indeed possible to improve the performance of speaker recognition system. On similar lines, studies in [10, 12] have also demonstrated the combination of evidences from system and source features to improve performance. In [61], it has been reported that the performance of the speaker recognition system can be improved by combining the evidence from SVM and GMM classifiers.

S.M. Mirrezaie & S.M. Ahadi, in 2008, introduced speaker diarization in multi-speaker environment using PSO and mutation information. In this study they present an approach comprising of PSO algorithm, which encodes possible segmentations of an audio record by measuring mutual information between these segments and the audio data. This measure is used as the fitness function for the PSO. This algorithm has been tested on two actual sets of data with up to 8 speakers for speaker diarization, and has led to very good results in all test problems [72]. Md. Tariquzzaman, Jin Young Kim, Seung You Na, in 2009 [73], introduces a technique for correction of missing reliability for robust bimodal speaker identification. In this study, they proposed a fuzzy membership function for adaptive threshold in different modalities reliability measure for robust bimodal speaker identification. In the bimodal speaker identification system, they also proposed an extension of a modified convection reliability function applied to both the audio and lip information to account optimal reliability simultaneously for audio and visual information integration. Petru-Marian Briciu in 2010 introduced the speaker identification using partially connected locally recurrent probabilistic neural networks [74].

VI. Conclutions And Future Work

In this paper, we have discussed the techniques developed for each stage of speaker recognition system. This includes different analysis, feature extraction, modeling and testing techniques. Among the developed techniques, segmental analysis for speech analysis, MFCC and its derivatives as features, PSO-GMM as a modeling technique and log likelihood ratio for testing.

There are the issues that may be taken up as directions for future research in the speaker recognition field. These includes integrating the segmental, sub- segmental and supra- segmental techniques in a unified framework so that speech signal is analyzed with all of them and relevant features are extracted.

Methods may be developed to extract feature vectors representing the speaker-specific information from the excitation source and the behavioral component of speech. New modeling techniques like soft computing that maximize the speaker specific information after modeling using only

limited data need to be explored. Finally, different testing and combining methods nay be explored for maximizing the speaker recognition performance under various practical conditions- like amount of speech data being limited, uncontrolled environment and stressed speaker conditions.

ACKNOWLEDGMENT

Rita Yadav received her Degree from Institute of Electronics & Telecommunication Engineers (IETE), New Delhi, India in 2007. She is currently pursuing M.Tech. Degree in Electronics & Communication Engineering from Punjab Technical University, Jalandhar, India. She worked as Junior Engineer (R &D) in Recorders and Medicare System Pvt. Ltd., Chandigarh, India in 2006. She worked in C-DAC, Mohali, India as a part of PCB Designing Lab. in 2006-2010. She is presently working as Junior Technical officer (JTO) in SQAE (L), Ministry of Defence (DGQA), New Delhi, India. She wants to thanks the college faculty members for their supervision and guidance.

Danvir Mandal received his B.Tech. Degree in Electronics & Communication Engineering from Punjab Technical University, Jalandhar, India in 2001, the M.Tech. Degree in Electronics & Communication Engineering from Punjab Technical University, Jalandhar, India in 2006. He is currently pursuing Ph.D degree from NITTTR, Chandigarh, India. He was a Lecturer with Department of Electronics & Communication Engineering, Institute of Engineering & Technology, Bhaddal, Punjab, India in 2006. Presently, he is an Assistant Professor with Department of Electronics & Communication Engineering, Institute of Engineering & Technology, Bhaddal, Punjab, India. His research interests include digital signal processing, image processing, antenna design and analysis, FDTD methods.

REFERENCES

[1] B.S. Atal, "Automatic recognition of speakers from their voices," Proc. IEEE, vol. 64(4), pp. 460-75, Apr. 1976.

[2] R.J. Mammone,X.Zhang,and R.P.Ramachandran, "Robust speakker recognition a feature-based approach," IEEE Signal Process. Mag., vol.13 (5), pp.58-71, Sep. 1996.

[3] A.E. Rosenberg, "Automatic speaker verification: A review," Proc IEEE, vol. 64(4), pp. 475-87, Apr. 1976.

[4] H. Gish, and M.Schmidt , " Text-indepent speaker identification," IEEE Signal Process. Mag. , vol. 18, pp.18-32, Oct. 2002.

[5] J. P. Campbell, Jr.., "Speaker recognition: A tutorial," Proc. IEEE, vol. 85 (9), pp. 1437-62, Sep. 1997.

[6] L. Rabiner, and B.H. Jung ,Fundamntals of speech recognition. Singapore: Pearson Education, 1993.

[7] James E.Luck, "Automatic speaker verification using cepstral measurements," J. Acoust. Soc. Amer., vol. 46(2), pp. 1026-32, Nov. 1969.

[8] G. Doddington, "Speaker recognition: identifying people by their voices," Proc. IEEE, vol. 73, pp. 1651-64, 1985.

[9] P. Satyanarayana, "Short segmant analysis of speech for enhancement," Ph. D. disseratation, Indian Institute of Technology Madras, Dept. of computer Science anf Engg., Chennai, India, Feb. 1999.

[10] S.R.M. Prasanna, C.S. Gupta , and B. Yegnanarayana "Extraction of speaker-specific excitation information frpm linear prediction residual of speech," Speech Communication, vol. 48, pp. 1243-61, 2006.

[11] B. Yegnanarayana, S.R.M. Prasanna, J. M. Zachariah, and C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed- text specker verification system ," IEEE Trans. Speech Audio Process. , vol. 13(4), pp. 575-82, July 2005.

[12] K.S.R. Murthy, and B. Yegnanarayana, "Combining evidence from residue phase and MFCC features for specker recognition ," IEEE Trans. Signal Process. Lett. , vol. 13(1), pp. 52-6, Jan. 2006.

[13] B. Yegnanarayana, K. Sharat Reddy, and S.P. Kishore, "Source and system features for specker recognition using AANN models ," in proc. Int. Conf. Acoust., Speech, Signal Process. , Utah, USA, Apr. 2001.

[14] K. Sharat Reddy, "Source and system features for specker recognition ," Master's thesis, Indian Institute of Technology Madras, Dept. of computer Science anf Engg., Chennai, India, 2003.

[15] C.S. Gupta, " Significance of source features for specker recognition ," Master's thesis, Indian Institute of Technology Madras, Dept. of computer Science anf Engg., Chennai, India, 2003.

[16] B.S. Atal, "Automatic specker recognition based on pitch contours ," J. Acoust. Soc. Amer., vol. 52, no. 6(part 2), pp. 1687-97, 1972.

[17] L. Mary, K.S. Rao ,S.V. Gangashetty, and B. Yegnanarayana, "Neural networls model for capturing duration anf intination knowlwdge for language and speaker identification," in Proc. Int.Conf. Cognitive Neural Syatem,Boston,Massachusetts, May 2004.

[18] F.Farahani, P.G. Georgiou, and S.S. Narayanan, "Speaker identification using suprasegmental pitch patterns dynamics," in Proc. Int.Conf. Acoust. , Speech Signal Process. , Montreal, Canada, May 2004, pp.89-92.

[19] P. Denes, and M.V. Mathews, "Spoken digit recognition using time-frequency pattern matching," J. Acoust. Soc. Amer., vol. 32(11), pp. 1450-5, Nov. 1960.

[20] S.Pruzansky, "Pattern-matching procedure for automatic talker recognition," J. Acoust. Soc. Amer., vol. 35(3), pp. 354-8, Mar. 1963.

[21] S.Pruzansky and M.V. Mathews, "Talker- recognition procedure based on analysis of varuance," J. Acoust. Soc. Amer., vol. 36(11), pp. 2041-7-8, Nov. 1964.

[22] K.P Li, J. E. Dammann, and W.D. Chapman, "Experimental studies in speaker verification using an adaptive system," J. Acoust. Soc. Amer., vol. 40(5), pp. 966-78, Nov. 1966.

[23] J.W. Glenn, and N. Kleiner, "Speaker identification based on nasal phonation," J. Acoust. Soc. Amer., vol. 43(2), pp. 368-72, June 1967.

[24] J.J. Wolf, "Efficient acoustic parameters for speaker recognition," J. Acoust. Soc. Amer., vol. 51, no.6(part2) , pp. 2044-56, 1971.

[25] B.S. Atal, "Effectness of linear prediction characteristics of the speech wave for Automatic specker identification and verification ," J. Acoust. Soc. Amer., vol. 55, pp. 1304-12, 1974.

[26] M.R.Sambur, "Selection of acoustic features for specker identification ," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-23 (2), pp. 176-82, Apr. 1975.

[27] A.E. Rosenberg , and M.R.Sambur, "New techniques for automatic specker verification ," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-23 (2), pp. 169-76, Apr. 1975.

[28] M.R.Sambur, "Specker recognition using orthogonal linear prediction ," IEEE Trans. Acoust., Speech Signal Process., vol. ASSP-24 (4), pp. 283-9, Aug. 1976.

[29] J.D. Markel, B.T. Oshika, and A.H. Grey, Jr., "Long-term feature avarging for specker recognition," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-25 (4), pp. 330-7, Aug. 1977.

[30] S. Furui., "Speaker-independent isolated word recognition using dynamic features of speech spectrum ," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, pp. 52-9, Feb. 1983.

[31] Sasaoki Furui., "Spectral analysis technique for automatic speaker verification ," IEEE Trans. Acoust., Speech, Signal Process., vol. 29 (2), pp. 254-72, Apr. 1981.

[32] D.A. Reynolds, "Experimaental evalution of features for robust speaker identification ," IEEE Trans. Speech Audio Process., vol. 2 (4), pp. 639-43, Oct. 1994.

[33] P. Thevenaz and H Hugli, "Usefullness of LPC-residue in text - independent speaker verification ," Speech communication., vol. 17, pp. 145-57, 1995.

[34] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification ," IEEE Trans. Speech Audio Process., vol. 7 (5), pp. 569-85, 1999.

[35] M J Carey, E S Paris,H. Lloyd-Thomas, and S.Bennett, "Robust prosotic features for speaker identification ," in proc. Int. Spoken Language Process., Philadelphis,PA,USA,Oct 1996.

[36] M K Sonmez, E Sriberg,L.Heck, amd M. Weintraub, "Modeling dynamic prosodic variation for speaker verification ," in proc. Int. Spoken Language Process., Sydney ,NSW, Australia Nov –Dec. 1998.

[37] B Peskin, J Navratil , J. Abramson, D.Jones, D. Klusacek , D.A. Reynolds,and B.Xiang, "Using prosodic and conservational feature for high- performance speaker recognition ," in Int. Conf Acoust., Speech ,Signal Process., vol. IV,Hong Kong pp. 784-7, Apr. 2003.

[38] M Girmaldi, F Cummins, "Speaker identification using instataneous frequencies ," IEEE Trans. Audio, Speech , Language Process., vol. 16(6), pp. 1097-111, Aug. 2008.

[39] Min-Seok Kim and Ha-Jin Yu, "A new feature transformation method based on rotation for speaker identification," 19th IEEE Int. Conf. on Tools with Artificial Intelligence , pp. 68-73, 2007.

[40] J. Kennedy and R. Eberhart, "Particle Swarm Optimization", Proceedings of IEEE International Conference on Neural Networks (ICNN'95), Vol. IV, pp.1942-1948, Perth, Australia, 1995.

[41] Min-Seok Kim, IL-Ho Yung and Ha-Jin Yu, "Maximize the distance between the Gaussian mixture models for speaker verification using PSO," 4th IEEE Int. Conf. on Natural Computation , pp. 175-78, 2008.

[42] Y. Linde, A.Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Communications, vol. COM 28(1), pp. 84-96, Jan. 1980.

[43] R. Gray, "Vector quantization," IEEE Acoust., Speech, Signal Process., Mag. , vol. 1, pp. 4-29, Apr. 1984.

[44] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.H. Juang, "A Vector quantization approach to speaker recognition," in Proc. IEEE Int Conf. Acoust., Speech, Signal Process., vol. 10, Detroit, Michingon, pp. 387-90, Apr. 1985.

[45] J.C. Bezdek, and J.D. Harris, "Fuzzy portions and relations;an axiomatic basis for clustering," Fuzzy Sets and Systems, vol. 1, pp. 111-27, 1978.

[46] L. Lin, and S. Wang, "A Kernel method for speaker recognition with little data," in Int. Conf. signal Process., Budapest, Hungery, May 2006.

[47] V. Chatzis, A.G. Bors, and I. Pitas , "Multimodal decision-level fusion for person authentication," IEEE Trans. Man Cybernetics Part A: Systems and Humans, vol. 29, pp. 674-81, Nov. 1999.

[48] A.E. Rosenberg, and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in proc.Int.Conf. Acoust. , Speech, Signal Process., Atlanta Georgia, pp. 81-4, May 1996.

[49] J.M. Naik, L.P. Nestch, and G.R. Doddington, "Speaker verification using long distance telephone lines," in proc. Int. Conf. Acoust., Speech, Signal Process., Glasgow, UK, pp. 524-7, May 1989.

[50] T. Matsui, and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and Discrete/continuous HMMs," IEEE Trans. Speech Audio Process., vol. 2(3), pp. 456-9, July 1994.

[51] O. Kimball, M. Schmidt, H. Gish, and J. Waterman, "Speaker verification with limited enrollment data," in proc. European Conf. Speech Commun. and Tech. (EUROSPEECH'97), Rhodes, Greecee, pp. 967-70, Sep.1997.

[52] R.P. Lipmann, "An introduction to computing with neural nets," IEEE Trans. Acoust., Speech, Signal Process., vol. 4, pp. 4-22, Apr 1989.

[53] G. Bannani, and P Gallinari, "Neural networks for discrimination and modelization of speakers," Speech Communication., vol. 17, pp. 159-75, 1995.

[54] B. Yegnanarayana, Artificial neural networks. New Delhi: Prentice-Hall, 1999.

[55] J. Oglesby, and J.S. Mason, "Optimization of neural models for speaker identification," in proc. Int. Conf. Acoust., Speech, signal Process., Albuquerque, NM, pp. 261-4, May 1990.

[56] "Radial basis function for speaker recognition," in proc. Int. Conf. Acoust., Speech, Signal Process., Toronto, Canada, pp. 393-6, May 1991.

[57] T. Kohonen, "The self-organizing map," Proce. IEEE, vol. 78(9), pp. 1464-80, Sep. 1990.

[58] M. Inal, and Y.S. Fatihoglu, "Self organizing map and associative memory model hybrid classifier for speaker recognition," in proc. Neu., Net., App., Elec., Engg. (NEUREL'02), Belgrade, Yugoslavia., pp. 71-4, Sep. 2002.

[59] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models,"Speech Communication, vol. 17, pp. 91-108, 1995.

[60] D.A. Reynolds, and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models,"IEEE Trans. Speech Audio Process., vol. 3, pp. 72-83, Jan. 1995.

[61] 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 and Language, vol. 20, pp. 210-29, 2006.

[62] D.A. Reynolds, T.F. Quateri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[63] B. Yegnanarayana, and S.P. Kishore, "AANN: An alternative to GMM for pattern recognition," Neural Networks, vol. 15, pp. 459-69, 2002.

[64] M. Shajith Iqbal, Hemanth Misra, and B. Yegnanarayana, "Analysis of auto associative neural networks," in Int. Joint Conf. Neural Networks, Washington, USA, 1999.

[65] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Lett., vol. 13(5), pp. 308-11, May 2006.

[66] C.H. You, K.A. Lee, and H. Li, "An SVM kernel with GMM- supervector based on the Bhattacharyya distance for speaker recognition,"IEEE Signal Process. Lett., vol. 16(1), pp. 49-52, Jan. 2009.

[67] H. R. S. Mohammadi and R. Saeidi, "Efficient implementation of GMM based speaker verification using sorted Gaussian mixture model," in Proc. EUSIPCO'06, Florence, Italy, Sep. 4–8, 2006.

[68] Rahim Saeidi, Hamid Reza Sadegh Mohammadi, " Particle swarm optimization for sprted adapted gaussian mixture models." IEEE Trans. Audio, Speech , Language Process., vol. 1672), pp. 344-353, Feb. 2009.

[69] Rahim Saeidi, Tomi Kinnunen, Hamid Reza Sadegh Mohammadi, Robert Rodman, Pasi Fränti, " Joint frame and gaussian selection for text independent speaker verification." IEEE Trans. ICASSP2010 , pp. 4530-4533, 2010.

[70] H. Jiang, and L. Deng, "A Bayesian approach to the verification problem: Applications to speaker verification," IEEE Trans. Speech, Audio Process., vol. 9(8), pp. 874-975, 2001.

[71] B-J Lee, S-W Yoon, H-G Kang, and D.H. Youn, "On the use of voting methods for speaker identification based on various resolution filterbanks," in proc. Int. Conf. Acoust., Speech, Signal Process., vol. I, Toulouse, France, , pp. 917-20, May 2006.

[72] S.M. Mirrezaie & S.M. Ahadi, "Speaker diarization in multi-speaker environment using PSO and mutation information," IEEE Trans. ICME 2008 , pp. 1533-1536, 2008.

[73] Md. Tariquzzaman, Jin Young Kim, Seung You Na, "A correction of missing reliability for robust bimodal speaker identification," IEEE Trans. Int.Conf. on information and Multimedia tech., pp. 239-243, 2009.

[74] Petru-Marian BRICIU, "Speaker identification using partially connected locally recurrent probabilistic neural networks," Proce. IEEE, pp. 87-90, 2010.