Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition

H. S. Jayanna and S. R. Mahadeva Prasanna

Department of Electronics and Communication Engineering, Indian Institute of Technology, Guwahati-781 039, India.

Abstract

Speaker recognition system may be viewed as working in four stages, namely, analysis, feature extraction, modeling and testing. This paper gives an overview of the major techniques developed in each of these stages. Such a review helps in understanding the developments that have taken place in each stage and also the available choices of techniques, along with their relative merits and demerits. A comparative study of different techniques is done at the end of each section to justify the choice of techniques available in the state-of-the-art speaker recognition systems. The paper is concluded with a discussion on the possible future direction for the development of techniques in each stage.

Keywords

Feature extraction, Speaker modeling, Speaker recognition, Speech analysis, Testing.

1. Introduction

The goal of automatic speaker recognition (termed more commonly as speaker recognition) is to analyze, extract, characterize and recognize information about the speaker identity [1,2]. Depending on the task objective, speaker recognition can be classified as speaker verification and speaker identification. The speaker verification involves accepting or rejecting the identity claim of a speaker. In speaker identification since there is no identity claim, the system identifies the most likely speaker of the test speech signal. Speaker identification can be further classified into closed-set identification and open-set identification. The task of identifying a speaker who is known *a priori* to be a member of the set of *N* enrolled speakers is known as closed-set speaker identification [3,4]. The limitation of this system is that the test speech signal from an unknown speaker will be identified to be one among the *N* enrolled speakers. Thus there is a risk of false identification. Therefore, closed-set mode should be employed in applications where it is surely to be used always by the set of enrolled speakers. On the other hand, speaker identification system which is able to identify the speaker who may be from outside the set of *N* enrolled speakers is known as open-set speaker identification [3,4]. In this case, first the closed-set speaker identification system identifies the speaker closest to the test speech data. The verification system is then used to compare the distance of this speaker with a threshold to come up with a decision. If the speaker is accepted, then the speaker is the identified speaker for the test data. Otherwise, the system has to generate an error message that the speaker is unknown. Depending on the mode of operation, speaker recognition can be further classified as text-dependent recognition and text- independent recognition. The former method requires the speaker to produce speech for the same text, both during training and testing; whereas the latter does not rely on a specific text being spoken [5].

The 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 segmenting the speech signal for further analysis and feature extraction. As will be discussed later, this is influenced by the type of speaker information to be extracted. The speech analysis is done using one of the following techniques: Segmental analysis, subsegmental analysis and supra-segmental analysis. The feature extraction stage deals with extracting the relevant speaker-specific information in terms of feature vectors at reduced data rate. The features may correspond to the vocal tract, excitation source and behavioral aspect of speaker information. The feature vectors of each speaker are further processed by a suitable modeling technique to enhance the speaker-specific information and also reduce the data rate. The modeling techniques may be either generative type or discriminative type. In this way, one model is built for each enrolled speaker. During testing, the speech signal is analyzed and features are extracted using the same techniques employed during training. The feature vectors are compared with the reference models using some distance measure like Euclidean distance; and based on the comparison result, the speaker in the test speech data will be recognized.

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

state-of- the-art speaker recognition systems mainly use segmental analysis, mel-frequency cepstral coefficients (MFCCs) quantifying vocal tract information, Gaussian mixture model (GMM) and likelihood ratio testing techniques in the analysis, feature extraction, modeling and testing stages, respectively. Even though this choice of techniques provides good performance, there are practical issues in the speaker recognition field for which other techniques may also have to be used for achieving good speaker recognition performance. Some of the practical issues are as follows:

- The state-of-the-art speaker recognition systems yield good performance when the speech data is clean. In practice, the speech data is noisy due to sensor, environment and channel conditions, and hence the degradation in performance. Improving the performance in such conditions is an important issue. The existing solution is the use of speech enhancement as a preprocessing stage [6-8].
- The use of speaker recognition system in the multilingual context is a requirement in India, where there is coexistence of a large number of languages. Efforts need to be made in this direction to build a reliable speaker recognition system. The existing solution is the use of a language identification system as a preprocessing stage [9].
- The use of speaker recognition system in an e-transaction needs a robust performance under stressed conditions. This is mostly due to the production of stressed speech during testing, like loud, fast and angry. Getting a reliable and satisfactory performance of a speaker recognition system under stressed conditions is also an important issue. The existing solution is the use of a stress compensation system [10].
- For remote biometric person authentication, the state-of-the-art systems employ large amounts of speech data. This is labor, resource and computation intensive. If we develop techniques that provide satisfactory performance using a relatively small amount of data, then acceptability of the speaker recognition technology increases. The existing solution is the use of Gaussian mixture model–universal background model (GMM-UBM) modeling technique [11,12].

Apart from the existing solutions that treat the speaker recognition system as a black box and add additional stages to take care of the practical issues, it may be possible to refine and improve the robustness of the existing techniques in each stage of the speaker recognition system. It may also be possible to develop new techniques in some of the stages. For this, first we should understand the various techniques that have been developed so far, mentioned in the literature – which is the scope of the present work. It can be found in the literature that for each stage, researchers in the field have developed many techniques. This paper gives an overview of these approaches. The rest of the paper is organized as follows. Speech analysis techniques developed for speaker recognition are discussed in section 2. Section 3 presents different feature extraction techniques for speaker recognition. Different modeling techniques are discussed in section 4. Section 5 presents different testing strategies and decision logics. Finally, summary of the review and scope for future research are mentioned in section 6.

2. 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 the vocal tract, as well as of the excitation source, are unique for each speaker. This uniqueness is embedded into the speech signal during speech production and can be used for speaker recognition. Further, the behavioral traits like how the vocal tract and excitation source are controlled during speech production are also unique for each speaker.

The information about the behavioral trait is also embedded into the speech signal and can be used for speaker recognition. These different types of speaker information are manifested best at different levels in the speech signal. 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 at extracting the relevant features in the feature extraction stage. State-of-the-art speaker recognition systems mainly employ the following analysis techniques:

2.1 Segmental Analysis

In this case, speech is analyzed using the frame size and shift in the range of 10-30 ms to extract the speaker information mainly due to the vocal tract. The speaker-specific vocal tract information may be assumed to be stationary for all practical analyses and processing when viewed in frames of size and shift in the range of 10-30 ms [13]. Studies made in [14-19] used segmental analysis to extract the vocal tract information for speaker recognition.

2.2 Sub-segmental Analysis

Speech analyzed using the frame size and shift in the range of 3-5 ms is known as sub-segmental analysis [20].

This technique is used mainly to analyze and extract the characteristics of the excitation source. Since the excitation source information is relatively fast varying compared to the vocal tract information, small frame size and shift are required to best capture the speaker-specific information, which is the reason for the choice of 3-5 ms for frame size and shift. Studies made in [21-26] demonstrated that speaker-specific excitation source information captured using the sub-segmental analysis, indeed, contains speaker information.

2.3 Supra-segmental Analysis

In this case, speech is analyzed using the frame size and shift in the range of 100-300 ms. This technique is used mainly to analyze and extract characteristics due to the behavioral traits of the speaker. These include word duration, intonation, speaking rate, accent, etc. The behavioral traits vary relatively slowly compared to the vocal tract information, which is the reason for the choice of large frame size and shift. Studies made in [17,22,27-29] demonstrated that some behavioral traits can be captured using supra-segmental analysis and can be used for speaker recognition.

State-of-the-art speaker recognition systems mainly use segmental analysis. Speaker-specific vocal tract information is therefore used for speaker recognition. As described above, speaker-specific vocal tract information is one component of rich speaker information 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.

3. Feature Extraction Techniques

The purpose of feature extraction stage is to extract the speaker-specific information in the form of feature vectors at reduced data rate. The feature vectors represent the speaker-specific information due to one or more of the following: Vocal tract, excitation source and behavioral traits. A good feature set should have representation due to all the components of speaker information. To develop such a good feature set, it is necessary to understand the different feature extraction techniques developed so far. This section describes the same. Spoken digit recognition conducted by P Denes *et al.* in 1960 suggested that inter-speaker differences exist in the spectral patterns of speakers [30]. 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 their usefulness for speaker recognition [31]. Further, he reported a study using the analysis of variance in 1964 [32]. In this work, a subset of features was selected from the analysis of variance using *F* ratio test defined as the ratio of variance of speaker means to average within speaker variance [32]. 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 *et al.* in 1966 using adaptive linear threshold elements [16]. This study used spectral representation of the input speech, obtained from a bank of 15 bandpass filters spanning the frequency range 300-4000 Hz. Two stages of adaptive linear threshold elements operate on the rectified and smoothed filter outputs. These elements are trained with fixed speech utterances. The training process results in a set of weights for the various frequency bands and time segments. The weights characterize the speaker. This study demonstrated that the spectral band energies as features contain speaker information. The study in [33] used pitch and formant information in addition to these band energies to improve the speaker verification performance. A study by Glenn *et al.* in 1967 suggested that acoustic parameters produced during the nasal phonation are highly effective for speaker recognition [34]. In this study, average power spectra of nasal phonation were used as the features for speaker recognition. In 1969, fast Fourier transform (FFT)–based cepstral coefficients were used in the speaker verification study [14]. 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 the speaker. A study made by G R Doddington in [15] reported an approach for speaker verification different from the approaches in [16] and [33]. He did not use a filter bank but converted the speech directly to pitch, intensity and formant frequency values, all sampled 100 times per second. These features were also demonstrated to provide good performance.

Most of the above studies used spectral patterns of speech as features for speaker recognition. Atal in 1972 demonstrated the use of variations in pitch as a feature for speaker recognition [17]. In addition to variations in pitch, other acoustic parameters such as glottal source spectrum slope, word duration and voice onset time were proposed as features for speaker recognition by Wolf in 1971 [18]. The concept of linear prediction for speaker recognition was introduced by Atal in 1974 [35]. 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. In general, the advantage of the cepstral coefficients is that they can be derived from a set of parameters which are invariant to any fixed frequency-response distortion introduced by the recording or transmission system [1].

Earlier studies neglected the features such as formant bandwidth, glottal source poles and higher formant frequencies, due to non-availability of measurement techniques. However, studies introduced after the linear prediction analysis, explored the speaker-specific potential of these features for speaker recognition[36]. A study carried out by Rosenberg and Sambur suggested that adjacent cepstral coefficients are highly correlated and hence all coefficients may not be necessary for speaker recognition [37]. In 1976, Sambur proposed to use orthogonal linear prediction coefficients as features in speaker identification [38]. 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 apparatus and contain little or no information about the linguistic content of the speech. In 1977, long-term parameter averaging, which includes pitch, gain and reflection coefficients for speaker recognition, was studied [39]. In this study, it was shown that the reflection coefficients are highly informative and effective for speaker recognition. In 1981 Furui introduced the concept of dynamic features, to track the temporal variability in the feature vector in order to improve the speaker recognition performance [40,41].

A study by Reynolds in 1994 compared the different features like Mel frequency cepstral coefficients (MFCCs), linear frequency cepstral coefficients (LFCCs), LPCCs and perceptual linear prediction cepstral coefficients (PLPCCs) for speaker recognition [19]. He reported that among these features, MFCCs and LPCCs gave better performance than the other features. Though the MFCCs and LPCCs are used to extract the same vocal tract information, in practice these features differ in their performance due to the different principle involved in extracting it [13], that is, the MFCC computation first applies discrete Fourier transform (DFT) on each frame and then weights the DFT spectrum by a Mel-scaled filter bank. The filter bank outputs are then converted to cepstral coefficients by applying the inverse discrete cosine transform (IDCT). In case of LPCCs, first, LPCs are obtained for each frame using Durbins-recursive method, and then these coefficients are converted to cepstral coefficients.

Most of the studies discussed above considered vocal tract information as speaker characteristics for speaker recognition. In [42], it is reported that linear prediction (LP) residual also contains speaker-specific source information that can be used for speaker recognition. Also, it has been reported that though the energy of the LP residual alone gives less performance, combining it with LPCC improves performance as compared to that of the LPCC alone. On similar lines, several studies demonstrated that though the information from the LP residual alone gives less performance compared to the MFCC, combining it with MFCC improves the performance as compared to that of MFCC alone [21,24-26]. Recently, it has been reported that LP residual phase also contains speaker-specific source information [23]. In this study, it was demonstrated that the LP residual phase combined with MFCC improved the performance as compared to that of MFCC alone [23]. Plumpe *et al.* developed a technique for estimating and modeling the glottal flow derivative waveform from speech for speaker recognition. 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 the combined coarse and fine structured parameters gave better performance than the individual parameter alone [43].

Most of the studies discussed so far have not considered features like word duration, intonation, speaking rate, speaking style, etc., representing the behavioral traits, for speaker recognition. A study carried out in [44] demonstrated the significance of long-term pitch and energy information for speaker recognition. In another study, pitch tracks and local dynamics in pitch were also used in speaker verification [45]. A study in [46] reported that the combination of prosodic features like long-term pitch with spectral features provided significant improvement as compared to only the pitch features. A study carried out in [22] demonstrated the use of features like long-term pitch and duration information obtained using dynamic time warping (DTW), along with source and spectral features, for text-dependent speaker recognition. In [27], supra-segmental features like duration and intonation captured using neural networks were used for speaker recognition. In [47], amplitude modulation (AM)–frequency modulation (FM)–based parameters of speech were proposed for speaker recognition. In this study, it was demonstrated that using different instantaneous frequencies due to the presence of formants and harmonics in the speech signal, it is possible to discriminate speakers.

The different feature extraction techniques described above may be summarized as follows:

- Spectral features like band energies, formants, spectrum and cepstral coefficients representing mainly the speaker-specific information due to the vocal tract.
- Excitation source features like pitch, variations in pitch, information from LP residual and glottal source parameters.

- Long-term features like duration, intonation, energy, AM and FM components representing mainly the speaker-specific information due to the behavioral traits.

Among these, the mostly used ones are the spectral features, in particular, MFCCs and LPCCs. The main reasons for the same may be the less intra-speaker variability and also availability of rich spectral analysis tools. However, the speaker-specific information due to excitation source and behavioral trait represents different aspects of speaker information. Thus the feature extraction stage will benefit by using feature extraction techniques for excitation source and behavioral traits; however, the main limitation for the same is the non-availability of suitable tools for extracting the features, but this is where the future lies for the feature extraction stage.

4. Speaker Modeling Techniques

The objective of modeling technique is to generate speaker models using speaker-specific feature vectors. Such models will have enhanced speaker-specific information at reduced data rate. This is achieved by exploiting the working principles of the modeling techniques. State-of-the-art speaker recognition systems employ different modeling techniques, which are briefly described in this section. Most of these techniques may be broadly grouped into generative and discriminative types. Earlier studies on speaker recognition used direct template matching between training and testing data [14,31,34-38]. In the direct template matching, training and testing feature vectors are directly compared using similarity measure. For the similarity measure, any of the techniques like spectral or Euclidean distance or Mahalanobis distance is used. Furui introduced the concept of dynamic time warping (DTW) for text-dependent speaker recognition [41]. However, it was originally developed for speech recognition [48]. In this approach, the sequence of feature vectors of the training-speech signal is the text-dependent template model. The DTW finds the match between the template model and the input sequence of feature vectors from the testing-speech signal. The disadvantage of template matching is that it is time consuming, as the number of feature vectors increases. For this reason, it is common to reduce the number of training feature vectors by some modeling technique like clustering. The cluster centers are known as *codevectors*, and the set of codevectors is known as *codebook*.

The most well-known codebook generation algorithm is the *K-means* algorithm [49,50]. In 1985, Soong *et al.* [51] used the LBG algorithm for generating speaker-based vector quantization (VQ) codebooks for speaker recognition. It is demonstrated that larger codebook and larger test data give good recognition performance. Also, the study suggested that VQ codebook can be updated from time to time to alleviate the performance degradation due to different recording conditions and intra-speaker variations [51]. The disadvantage of the VQ classification is, it ignores the possibility that a specific training vector may also belong to another cluster. As an alternative to this, fuzzy vector quantization (FVQ) using the well-known fuzzy *C-means* method was introduced by Dunn, and its final form was developed by Bezdek [52,53]. In [54] and [55], FVQ was used as a classifier for speaker recognition. It was demonstrated that FVQ gives better performance than the traditional *K-means* algorithm. This is because the working principle of FVQ is different from VQ, in the sense that the soft decision-making process is used while designing the codebooks in FVQ [52]; whereas in VQ, the hard decision process is used. Moreover, in VQ each feature vector 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 association decided by the membership function [52]. Since all the feature vectors are associated with all the clusters, there are relatively more number of feature vectors for each cluster; and hence the representative vectors, viz., *codevectors,* may be more reliable than VQ. Therefore, clustering may be better in FVQ and may lead to better performance compared to VQ.

In order to model the statistical variations, the hidden Markov model (HMM) for text-dependent speaker recognition was studied in [56-58]. In HMM, time-dependent parameters are observation symbols. Observation symbols are created by VQ codebook labels. Continuous probability measures are created using Gaussian mixtures models (GMMs). The main assumption of HMM is that the current state depends on the 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 observations for a given speaker model is calculated for speaker recognition. Kimbal *et al.* studied the use of HMM for text-independent speaker recognition under the constraint of limited data and mismatched channel conditions [59]. In this study, the MFCC feature was extracted for each speaker and then models were built using the broad phonetic category (BPC) and the HMM-based maximum likelihood linear regression (MLLR) adaptation technique. The BPC modeling is based on identification of phonetic categories in an utterance and modeling them separately. In HMM-MLLR, first, speaker-independent (SI) model is created using HMM, and then MLLR technique is used to adapt SI model to each speaker. It was shown that the speaker model built using the adaptation technique gave better performance than the BPC and GMM for cross-channel conditions.

The capability of neural networks to discriminate

between patterns of different classes is exploited for speaker recognition [60-62]. 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 associated with it. The concept of multi-layer perception (MLP) was used for speaker recognition in [63]. In this study, it was demonstrated that one–hidden layer network with 128 hidden nodes gave the same performance as that achieved with the 64 codebook VQ approach. The disadvantage of MLP is that it takes more time for training the network. This problem was alleviated using the radial basis function (RBF) in [64]. In this study, it was shown that the RBF network took lesser time than the MLP and outperformed both VQ and MLP.

Kohonen developed self-organization map (SOM) as an unsupervised learning classifier. SOM is a special class of neural network based on competitive learning [65]. Thus the performance of SOM depends on the parameters such as neighborhood, learning rate and number of iterations. These parameters are to be fine-tuned for good performance. The SOM and associative memory model were used together as a hybrid model for speaker identification in [66]. It was shown that the hybrid model gave better recognition performance than the MLP. A text-independent speaker recognition system based on SOM neural networks was also studied in [67]. 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 alleviated by using Kohonen learning vector quantization (LVQ) [65]. LVQ is a supervised learning technique that uses class information to optimize the positions of codevectors obtained by SOM, so as to improve the quality of the classifier-decision regions. An input vector is picked at random from the input space. If the class label of the input vector and the codevector agree, then the codevector is moved in the direction of the input vector. Otherwise, the codevector is moved away from the input vector. Due to this fine-tuning, there may be improved recognition rate compared to SOM. LVQ was proposed for speaker recognition in [68]. Speaker recognition using VQ, LVQ and GVQ (group vector quantization) was demonstrated for YOHO database in [69]. The experimental results show that LVQ gives better performance when the data is small, as compared to the traditional VQ and proposed GVQ; but GVQ yields better recognition performance when the data size is large.

In 1995, Reynolds proposed Gaussian mixture modeling (GMM) classifier for speaker recognition task [70]. This is the most widely used probabilistic modeling technique in speaker recognition. The GMM needs sufficient data to model the speaker, and hence good performance. In the GMM modeling technique, the distribution of feature 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 [71]. It was shown that GMM outperformed the other modeling techniques. Therefore, state-of-the-art speaker recognition systems use GMM as classifier due to the better performance, probabilistic framework and training methods scalable to large data sets [72].

The disadvantage of GMM is that it requires sufficient data to model the speaker well [70]. To overcome this problem, Reynolds *et al*. introduced GMM–universal background model (UBM) for the speaker recognition task [73]. 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-dependent 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 the UBM-based modeling technique is that it provides good performance even though the speaker-dependent data is small. The disadvantage is that a gender-balanced large speaker set is required for UBM training.

As an alternative to the GMM, an auto-associative neural network (AANN) has been developed for pattern recognition task [62,74,75]. AANN is a feed-forward neural network which tries to map an input vector onto itself. The number of units in the input and output layers is equal to the size of the input vectors. The number of nodes in the middle layer is less than the number of units in the input or output layers. The activation function of the units in the input and output layer is linear, whereas the activation function of the units in the hidden layer can be either linear or nonlinear. The advantage of AANN over GMM is that, it does not impose any distribution. The application of AANN has been extensively studied for speaker recognition in [21-23,76,77].

A learning method based on the statistical learning theory, a special theory on machine learning, is the support vector machine (SVM). The SVM has many desirable properties, including the 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 a multi-class problem by making it a one-versus-others two-class problem. SVM works by increasing the dimensionality of the input data space. The dimensionality is increased until it finds a maximum-margin linear hyperplane that can be used to separate the two classes. This is accomplished by using kernels and dot products. Moreover,

SVM is discriminative in nature, whereas other classifiers are generative in nature. Vincent Wan and Steve Renals studied SVM for speaker recognition [78,79]. In these studies, different kernels, like the polynomial, the Fisher, a likelihood ratio and the pair HMM, were studied. It was reported that using these kernels it is indeed possible to achieve state-of-the-art speaker recognition performance. Further, the same authors have used score space kernels for speaker verification study in [78]. The score space kernels generalize Fisher kernels and are based on underlying generative models such as GMM. In this study, it was demonstrated that SVM reduced the error rate compared to GMM likelihood ratio system.

W. M. Campbell *et al.* proposed generalized linear discriminant sequence (GLDS) kernel for speaker recognition and language identification tasks [72]. In this study, it was shown that though the SVM results were at par with GMM, combination of SVM with GMM yielded good recognition performance than the individual systems. The combination of SVM with GMM was also studied for speaker recognition in [80,81].

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

5. Speaker Testing and Decision Logic

Testing stage in the 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 close 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 (EER), 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 order to improve the visualization, the detection error trade-offs (DETs) plot is used where miss and false alarm probabilities are plotted according to their corresponding Gaussian deviations [82]. On the other hand, the computation of speaker identification performance is direct and simple. This is measured as a ratio of the number of correctly identified examples to the total number of examples considered for the testing.

In both speaker verification and identification, for matching test feature vectors to the reference model, we can use either the distance measurement techniques or the probabilistic scoring. Earlier studies employed spectral or Euclidean or Mahalanobis distance measurement techniques for comparison [14,31,34-38]. Reynolds used the concept of log likelihood ratio test for speaker recognition [70]. In 2001, H Jiang and L Deng studied the Bayesian approach for speaker recognition [83]. It was demonstrated that Bayesian approach moderately improved the performance compared to well-trained baseline system using the conventional likelihood ratio test.

In order to improve speaker recognition performance at the decision level, a combination of multiple classifiers has been proposed [84]. In this study, voting method was used for speaker identification based on the results of various resolution filter banks. A study conducted in [22] reported that by combining the evidences from source, supra-segmental and spectral features, it is indeed possible to improve the performance of the speaker recognition system. On similar lines, studies in [21,23] have also demonstrated the combination of evidences from system and source features to improve performance. In [72], it has been reported that the performance of the speaker recognition system can be improved by combining the evidences from SVM and GMM classifiers.

6. Conclusions and Future Works

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

As summarized at the end of each section, there are issues that may be taken up as directions for future research in the speaker recognition field. These include integrating the segmental, sub-segmental and supra-segmental techniques in a unified framework so that the 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 components of speech. Also, new methods may also be developed for extracting speaker-specific vocal tract information, which alleviate the problem of block processing present in the short-term processing of speech. New modeling techniques that maximize the speaker-specific information after modeling using only limited data need to be explored. Finally, different testing and combining methods may be explored for maximizing the speaker recognition performance under various practical conditions — like the amount of speech data being limited, stressed, obtained from different languages and uncontrolled environments.

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 speaker 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," *Proce. IEEE*, vol. 64(4), pp. 475-87, Apr. 1976.
4. H. Gish, and M. Schmidt, "Text-independent 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. S.R. Mahadeva Prasanna, "Event based analysis of speech," Ph.D. dissertation, Indian Institute of Technology Madras, Dept. of Computer Science, Chennai, India, Mar. 2004.
7. P. Krishnamoorthy, "Combined temporal and spectral processing methods for speech enhancement," Ph.D. dissertation, Indian Institute of Technology Guwahati, Dept. of Electronics and Communication Engg., Guwahati, India, Oct. 2008.
8. P. Krishnamoorthy, and S.R.M. Prasanna, "Reverberant Speech Enhancement by Temporal and Spectral Processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17(2), p. 253-66, Feb. 2009.
9. P.H. Arjun, "Speaker recognition in indian languages: A feature based approach," Ph.D. dissertation, Indian Institute of Technology Kharagpur, Dept. of Electrical Engg., Kharagpur, India, Jul. 2005.
10. G. Senthil Raja, "Feature analysis and compensation for speaker recognition under stressed condition," Ph.D. dissertation, Indian Institute of Technology Guwahati, Dept. of Electronics and Communication Engg., Guwahati, India, Jul. 2007.
11. P. Angkititrakul, and J.H.L. Hansen, "Discriminative In-Set/Out-of-Set speaker recognition," *IEEE Trans. Audio Speech Language Process.*, vol. 15(2), pp. 498-508, Feb. 2007.
12. V. Prakash, and J.H.L. Hansen, "In-Set/Out-of-Set speaker recognition under sparse enrollment," *IEEE Trans. Audio Speech Language Process.*, vol. 15(7), pp. 2044-51, Sep. 2007.
13. L. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition*. Singapore: Pearson Education, 1993.
14. James E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Soc. Amer.*, vol. 46(2), pp. 1026-32, Nov. 1969.
15. G. Doddington, "Speaker recognition -identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651-64, 1985.
16. 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.
17. B.S. Atal, "Automatic speaker recognition based on pitch contours," *J. Acoust. Soc. Amer.*, vol. 52, no. 6(part 2), pp. 1687-97, 1972.
18. J.J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, no. 6(part 2), pp. 2044-56, 1971.
19. D.A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2(4), pp. 639-43, Oct. 1994.
20. P. Satyanarayana, "Short segment analysis of speech for enhancement," Ph.D. dissertation, Indian Institute of Technology Madras, *Dept. of Computer Science and Engg., Chennai, India*, Feb. 1999.
21. S.R.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. 1243-61, 2006.
22. 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 speaker verification system," *IEEE Trans. Speech Audio Process.*, vol. 13(4), pp. 575-82, July 2005.
23. K.S.R. Murthy, and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13(1), pp. 52-6, Jan. 2006.
24. B. Yegnanarayana, K. Sharat Reddy, and S.P. Kishore, "Source and system features for speaker recognition using AANN models," in *proc. Int. Conf. Acoust., Speech, Signal Process.*, Utah, USA, Apr. 2001.
25. K. Sharat Reddy, "Source and system features for speaker recognition," Master's thesis, *Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India*, 2001.
26. C.S. Gupta, "Significance of source features for speaker recognition," Master's thesis, *Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India*, 2003.
27. L. Mary, K.S. Rao, S.V. Gangashetty, and B. Yegnanarayana, "Neural network models for capturing duration and intonation knowledge for language and speaker identification," in *Proc. Int. Conf. Cognitive Neural Systems, Boston, Massachusetts*, May 2004.
28. F. Farahani, P.G. Georgiou, and S.S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," in *proc. Int. Conf. Acoust., Speech, Signal Process.*, Montreal, Canada, May 2004, pp. 89-92.
29. F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, London, UK, April. 2002, pp. 141-4.
30. 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.
31. S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *J. Acoust. Soc. Amer.*, vol. 35(3), pp. 354-8, Mar. 1963.
32. S. Pruzansky, and M.V. Mathews, "Talker-recognition procedure based on analysis of variance," *J. Acoust. Soc. Amer.*, vol. 36(11), pp. 2041-7, Nov. 1964.
33. S.K. Das, W.S. Mohn, and S.L. Saleeby, "Speaker verification experiments," *J. Acoust. Soc. Amer.*, vol. 49, p. 138(A), 1971.
34. J.W. Glenn, and N. Kleiner, "Speaker identification based on nasal phonation," *J. Acoust. Soc. Amer.*, vol. 43(2), pp. 368-72, June 1967.
35. B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304-12, 1974.
36. M.R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23(2), pp. 176-82, Apr. 1975.
37. A.E. Rosenberg, and M.R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-23(2), pp. 169-76, Apr. 1975.
38. M.R. Sambur, "Speaker recognition using orthogonal linear

prediction," *IEEE Trans. Acoust., Speech, Signal Process..*, vol. ASSP-24(4), pp. 283-9, Aug. 1976.

39. J.D. Markel, B.T. Oshika, and A.H. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25(4), pp. 330-7, Aug. 1977.
40. 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. 1986.
41. Sasaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29(2), pp. 254-72, Apr. 1981.
42. P. Thevenaz, and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145-57, 1995.
43. 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.
44. M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," in *proc. Int. Spoken Language Process.*, Philadelphia, PA, USA, Oct. 1996.
45. M.K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *proc. Int. Spoken Language Process.*, Sydney, NSW, Australia, Nov-Dec. 1998.
46. B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D.A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition," in *Int. Conf. Acoust., Speech, Signal Process.*, vol. IV, Hong Kong, Apr. 2003, pp. 784-7.
47. M. Grimaldi, and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16(6), pp. 1097-111, Aug. 2008.
48. H. Sakoe, and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 43-9, Feb. 1978.
49. 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.
50. R. Gray, "Vector quantization," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 1, pp. 4-29, Apr. 1984.
51. 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, Apr. 1985, pp. 387-90.
52. 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.
53. H.J. Zimmermann, *Fuzzy set theory and its applications*, 1st ed. Kluwer academic, 1996.
54. L. Lin, and S. Wang, "A Kernel method for speaker recognition with little data," in *Int. Conf. signal Process.*, Budapest, Hungery, May, 2006.
55. 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.
56. 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, May 1996, pp. 81-4.
57. 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, May 1989, pp. 524-7.
58. 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.
59. 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, Sep. 1997, pp. 967-70.
60. R.P. Lipmann, "An introduction to computing with neural nets," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 4, pp. 4-22, Apr. 1989.
61. G. Bannani, and P Gallinari, "Neural networks for discrimination and modelization of speakers," *Speech Communication.*, vol. 17, pp. 159-75, 1995.
62. B. Yegnanarayana, *Artificial neural networks*. New Delhi: Prentice-Hall, 1999.
63. J. Oglesby, and J.S. Mason, "Optimization of neural models for speaker identification," in *proc. Int. Conf. Acoust., Speech, Signal Process.*, Albuquerque, NM, May 1990, pp. 261-4.
64. "Radial basis function for speaker recognition," in *proc. Int. Conf. Acoust., Speech, Signal Process.*, Toronto, Canada, May 1991, pp. 393-6.
65. T. Kohonen, "The self-organizing map," *Proce. IEEE*, vol. 78(9), pp. 1464-80, Sep. 1990.
66. 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, Sep. 2002, pp. 71-4.
67. A.T. Mafra, and M.G. Simoes, "Text independent automatic speaker recognition using self-organizing maps," in *proc. Ind. App. Society conf.*, vol. 3, Victoria, British Columbia, Oct. 2004, pp. 1503-10.
68. G. Bannani, F. Fogelman, and P. Gallinari, "A connectionist approach for speaker identification," in *proc. Int. Conf. Acoust., Speech, Signal Process.*, Albuquerque, NM, May 1990, pp. 265-8.
69. J. He, L. Liu, and G. Palm, "A discriminative training algorithm for VQ-based speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 353-6, May 1999.
70. D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, 1995.
71. 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.
72. W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and PA. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210-29, 2006.
73. 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.
74. B. Yegnanarayana, and S.P. Kishore, "AANN: An alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459-69, 2002.
75. M. Shajith Iqbal, Hemanth Misra, and B. Yegnanarayana, "Analysis of auto associative neural networks," in *Int. Joint Conf. Neural Networks*, Washington, USA, 1999.
76. B. Yegnanarayana, K.S. Redddy, and S.P. Kishore, "Source and system features for speaker recognition using AANN models," in *Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, Utah, USA, Apr. 2001, pp. 409-12.
77. N. Dhananjaya, and B. Yegnanarayana, "Correlation-based similarity between signals for speaker verification with limited amount of speech data," in *proc. International Workshop, MRCS 2006*, Istanbul, Turkey, Sep. 2006.
78. V. Wan, and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 203-10, 2005.
79. V. Wan, and S. Renals, "Evaluation of kernel methods for speaker verification and identification," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, pp. 1-669 - 1-672, 2002.
80. 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.

81. 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.

82. G.R. Doddington, M.A. Przybocki, A.F. Martin, and D.A. Reynolds, "The NIST speaker recognition evaluation overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225-54, 2000.

83. 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.

84. 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, May 2006, pp. 917-20.

AUTHORS

H. S. Jayanna was born in India in 1970. He received the B.E. degree in instrumentation and electronics engineering from Dr. Ambedkar Institute of Technology, Bangalore University, Bangalore, India, in 1992 and the M.E. degree in electronics from University Vishweshwaraya College of Engineering, Bangalore, India, in 1995. He is currently pursuing the Ph.D. degree in electronics and communication engineering at Indian Institute of Technology Guwahati, India.

His research interests are in speech and speaker recognition.

E-mail: h.jayanna@iitg.ernet.in

S. R. Mahadeva Prasanna was born in India in 1971. He received the B.E. degree in electronics engineering from Sri Siddartha Institute of Technology, Bangalore University, Bangalore, India, in 1994, the M.Tech. degree in industrial electronics from the National Institute of Technology, Surathkal, India, in 1997, and the Ph.D. degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, India, in 2004. He is currently an Associate Professor in the Department of Electronics and Communication Engineering, Indian Institute of Technology, Guwahati.

His research interests are in speech and signal processing, application of AI tools for pattern recognition tasks in speech, and signal processing.

E-mail: prasanna@iitg.ernet.in

DOI: 10.4103/0256-4602.50702; Paper No TR 25_09;