Speaker Features And Recognition Techniques: A Review

1. Dr. Mahesh S. Chavan
Professor and Head of Electronics Engg. Dept.
KIT's Collegege of Engineering, Kolhapur,
Maharashtra, India

2. Mrs. Sharada V. Chougule.
Assistant Professor,
Electronics and Telecommunication Engg. Dept.
Finolex Academy of Managrment and Technology,Ratnagiri,
Maharashtra, India

Abstract
This paper gives an overview of various methods and techniques used for feature extraction and modeling in speaker recognition. The research in speaker recognition have been evolved starting from short time features reflecting spectral properties of speech (low-level or physical traits) to the high level features (behavioural traits) such as prosody, phonetic information, conversational patterns etc. Low level acoustic information such as cepstral features has been dominated as these features gives very low error rates (especially in quiet conditions). But they are more prone to error in noisy conditions. In this paper various features along with modeling techniques used in speaker recognition are discussed.

Introduction
Speech is the product of a complex behaviour conveying different speaker-specific traits that are potential sources of complementary information. Human speech production can be modeled by so-called *source-filter model* featured. As the name suggests, the model considers the voice production mechanism as a combination of two components: the *voice source* and the *acoustic filter*. The "source" refers to the airstream generated by the larynx and the "filter" refers to the vocal tract. Both of the components are inherently time-varying and assumed to be independent of each other [1], [2]. In this model we separate the source G(w) from the articulation and radiation H(w) (Furui S. 2001). The speech signal is then represented by the cascade connection of G(w) and H(w) giving: $S(w) = G(w) H(w)$. For voiced speech, the source is modeled as an impulse train with period T. For unvoiced speech, the source is modeled as Gaussian white noise. In each case, the source is amplified by the gain factor G in proportion to the volume of speech. The signal is then passed through a time-varying digital filter to represent the articulation and radiation applied by the vocal track.

Historically, all speaker recognition systems have been mainly based on acoustic cues that is nothing but physical traits extracted from spectral characteristics of speech signals. So far the features derived from the speech spectrum have proven to be the most effective in automatic systems, because the spectrum reflects the geometry of system that generate the signal. Therefore the variability in the dimensions of the vocal track is reflected in the variability of the spectra between the speakers. However, studies [3] have proved that there is a large amount of information suitable for speaker recognition being the top part related to learned traits and the bottom part to physical traits.

Speaker Recognition
One objective in automatic speaker recognition is to decide which voice model from a known set of voice models best characterizes a speaker; this task is referred to as speaker identification. In the different task of speaker verification, the goal is to decide whether a speaker corresponds to a particular known voice or to some other unknown voice.

There are two modes of operation for speaker identification: in the closed-set mode, the system assumes that the unknown voice must come from the set of known voices; in open-set mode, the speakers that do not belong to the set of known voices are referred to as impostors. An important application of speaker identification technology is forensics, identifying the suspects among a set of known criminals. Automatic speaker recognition systems can be further classified according to the speech modality: text-dependent or text-independent. In text-dependent recognition, the user must speak a phrase known to the system, which can be fixed or prompted. The knowledge of a spoken phrase can provide better recognition results. In text-independent recognition, the system does not know the phrase spoken by the user. Despite the unconstrained phrase selection, this makes the system to be more complex. However, text-independent speaker recognition systems have more applications than text-dependent ones in real life.

There are generally two phases [1] in building or using a speaker recognition system. The first phase is called enrolment or training phase, in which a user enrols by providing voice samples to the system. The system extracts speaker-specific information from the voice samples to build a voice model of the enrolled speaker. The second phase is called the classification or recognition phase, in which a test voice sample is used by the system to measure the similarity of the user's voice to the previously enrolled speaker models to make a decision. In a speaker identification task, the system measures the similarity of the test sample to all stored voice models. In speaker verification task, the similarity is measured only to the model of the claimed identity. The decision also differs across systems. For example, a closed-set identification task outputs the identity of the recognized user; besides the identity, an open-set identification task can also choose to reject the user in case the test sample do not belong to any of the stored voice models; a verification task chooses to accept or reject the identity claim.

The effectiveness of speaker recognition system is in measures differently for different tasks. Since the output of a closed-set speaker identification system is a speaker identity from a set of known speakers, the identification accuracy is used to measure the performance. For open-set systems there are two types of errors: false acceptance of an impostor and false rejection of a known speaker. The performance measure can also incorporate the cost associated with each error.

Like most pattern recognition problems, a speaker recognition system can be partitioned into four modules: feature analysis and extraction, speaker modelling, pattern matching and decision logic.

Feature Analysis and Extraction

From human speech production mechanism, it is possible to identify individual using the speech data. Speech contains speaker specific information due to vocal track and excitation source. Larynx is the major excitation source, whereas vocal track is the major resonant structure. Speaker information is due to particular shape, size and dynamics of vocal track and also the excitation source. These features related to physiological nature of human speech production are called physical traits, which are used in state-of-art systems. However human speaker recognition relies on other sources of information like speaking style, pronunciation etc. Such features are referred to as behavioral traits. 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. Thus the information present in speech signal carries the identity of speaker at different levels. To properly represent speech data, it is necessary to analyse it using suitable analysis techniques. The analysis techniques aims at selecting proper frame size and shift for analysis and also at extracting the relevant features in the feature extraction stage [4].

The information about the audio category is contained in the excitation source (sub-segmental), system/physiological (segmental) and behavioral (suprasegmental) characteristics of the speech data.

Sub-segmental analysis is one in speech signal is analysed using the frame size and shift of very small duration (3-5ms). This technique is used mainly to analyse 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 [5].

In *segmental analysis* (used for extraction of short-term spectral features), features are computed from short frames of about 20-30 milliseconds in duration. They are usually descriptors of the short-term spectral envelope which is an acoustic correlate of timbre, i.e. the "colour" of sound, as well as the resonance properties of the supra-laryngeal vocal tract.

High-level features are generally related to a speaker's learned habits and style, such as particular word usage or idiolect. For humans, the information about the audio category is perceived by listening to a longer segment of audio signal. This other level of information contained in the audio signal is the *suprasegmental information* that is the variation of the signal over long duration. In this case, speech is analysed using the frame size and shift in the range of 50-200 ms. Studies made in [5],[6],[7] shows the significance of suprasegmental features in speaker recognition systems. These features are useful as their structure is not affected by the frequency characteristics of the transmission systems. Each of the four basic acoustic features of speech signal, i.e. pitch, intensity, duration and speech quality, is a carrier of a variety of types of linguistic, paralinguistic and non-linguistic information [8][9].

Feature Extraction Techniques

Feature extraction is said to be the heart of speaker recognition system.. Feature extraction is one where the input speech is processed to obtain those features of the input speech which are useful in speaker identification. The function of feature extraction is to convert speech waveform to some type of parametric representation called *feature vectors* for further analysis and processing by the classifiers. This stage 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 (Physical traits) and behavioral traits. For speaker recognition, features that exhibit high speaker discrimination power, high inter-speaker variability, and low intra-speaker variability are desired. Although there are no exclusive features conveying speaker identity in the speech signal, from the source-filter theory of speech production it is known that the speech spectrum shape encodes information about the speaker's vocal tract shape via resonances (formants) and glottal source via pitch harmonics.

Feature extraction is necessary for several reasons. First, speech is a highly complex signal which carries several features mixed together. In speaker recognition we are interested in the features that correlate with the physiological and behavioral characteristics of the speaker. Other information sources are considered as undesirable noise whose effect must be minimized. The second reason is a mathematical one, and relates to the phenomenon known as *curse of dimensionality* [10], which implies that the number of needed training vectors increases exponentially with the dimensionality. Furthermore, low-dimensional representations lead to computational and storage savings.

The ideal feature should have [10]:

- large between-speaker and small within-speaker variability
- be difficult to impersonate/mimic
- not be affected by the speaker's health or long-term variations in voice
- occur frequently and naturally in speech
- be robust against noises and distortions

It is unlikely that a single feature would fulfill all the listed requirements. Fortunately, due to the complexity of speech signals, a large number of complementary features can be extracted and combined to improve accuracy. The selection of features depends largely on the application (co-operative/non co-operative speakers, desired security/convenience balance, database size, amount of environmental noise).

Types of Features

A vast number of features have been proposed for speaker recognition. We divide them into the following classes:

- Spectral features
- Dynamic features
- Source features
- Suprasegmental features
- High-level features

Spectral features are descriptors of the short-term speech spectrum, and they reflect more or less the physical characteristics of the vocal tract. *Dynamic features* relate to time evolution of spectral (and other) features. *Source features* refer to the features of the glottal voice source. *Suprasegmental* features span over several segments. Finally, *high-level features* refer to symbolic type of information, such as characteristic word usage.

The anatomical structure of vocal apparatus is easy to extract in automatic fashion. For same sound the location and magnitude of peaks observed in spectra is different. Early text-dependent speaker recognition systems utilized information from short-time spectrum to provide unique features for speaker [8]. These features consisted of energy measurements from the outputs of a bank of a filter. LPC is one of the feature extraction method based on the source-filter model of speech production.The basic problem in LPC analysis is to determine prediction coefficients from the speech frame. B.S. Atal in 1976 [9] uses linear prediction model for parametric representation of speech derived features. The predictor coefficients and other speech parameters derived from them, such as the impulse response function, the auto-correlation function, the area function, and the cepstrum function were used as input to an automatic speaker recognition system ,and found the cepstrum to provide the best results for speaker recognition. Joseph P. Campbell in [1], uses all-pole LP (linear prediction) to model a signal by a linear combination of its past values and a scaled present input. Reynolds in 1994 [12] compared different - features useful for speaker recognition, such as Mel frequency cepstral coefficients (MFCCs), linear frequency cepstral

coefficients (LFCCs), LPCC (linear predictive cepstral coefficients) and perceptual linear prediction cepstral coefficients (PLPCCs). From the experiments conducted, he had concluded that , of these features, MFCCs and LPCCs give better performance than the other features. Both MFCC and LPCC coefficients are used to extract vocal track information, but uses different technique to extract the features . MFCC extraction is similar to the cepstrum calculation except that one special step is inserted, namely the frequency axis is warped according to the Mel-scale using mel 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. The predictor coefficients themselves are rarely used as features, but they can be transformed into robust and less correlated features such as LPCC, line spectral frequencies (LSFs) and *perceptual linear prediction cepstrum coefficients (PLPCC)* [12,13] or *eigen-MLLR coefficients* [14]. Experimental evaluation of recognition accuracy of the MFCC, LPCC and PLPCC was made in [13] and result of this report is that all features perform poorly without some form of channel compensation, however, with channel compensation MFCC slightly outperform other types. Cepstrum representation of the speech signal has shown to be useful in practice. The features discussed above are called short-term (spectral) or low-level features. These features are used in most state-of speaker recognition systems as these easy to compute and yield good performance (Reynolds et al.,2003). However, it is not without drawbacks. The main disadvantage of the cepstrum is that it is quite sensitive to the environment and noise [15].

Of all the various spectral features, MFCC, LPCC,LSF and PLP are the most recommended features which carry information about the resonance properties of vocal track [17]. Most of the current implementations use some kind of spectral envelope features to parameterize the voice (MFCC, LPCC...), achieving a great performance. But recent researches are trying to include long term information into the system, in order to reduce error rates.

Unlike short-term spectral information, long-term information is being used which convey supra-segmental information, such as prosodic and speaking style. Andre G. Adami and Douglas A. Reynolds in2003 [16], presented two new approaches that demonstrated effective ways to model and apply prosodic contours for text independent speaker verification tasks. In first approach the relation between dynamics of fundamental frequency (fo) and energy trajectories were used to characterize the speaker's identity. In this, global distribution of energy and fo features were such as log f0, log energy and their first order derivatives were created. In second accent and intonation information from a known set of frequently and naturally occurring words found in conversational speech. They had used n-grams to model the sequence.

From source-filter model, it was shown that speech signal can be decomposed into two parts: the source part and the system part. The system part consists of the smooth envelope of the power spectrum and is represented in the form of cepstrum coefficients, which can be computed by using either the linear prediction analysis or the mel-flter-bank analysis. Most of the automatic speaker recognition systems reported in the literature utilise the system information in the form of cepstral coefficients. These systems perform reasonably well. The source contains information about pitch and voicing. This information is also important for humans to identify a person from his/her voice. Hassan Euaidi and Jean Rouaf in 2004 [18] had proposed an approach which jointly exploits the information of the vocal tract and the glottis source. The approach synchronously takes into account the correlation between the two sources of information. The fundamental frequency and the MFCC coefficients were used to represent the information of the source and the vocal tract, respectively. Experiments that integrate the a-posteriori probability of observing a MFCC vector given the knowledge of the pitch frequency have been reported. It was also shown that systems based on voiced segments yield good scores. However, when the dependence of the source and vocal tract is taken into account, the best results were observed for durations T lower than 500 ms. Speech prosody refers to the intonation, energy and rate of the speech. It is well known that these features are characteristics of each person, so that they carry information about the speaker. Furthermore, prosody is uncorrelated with the spectral envelope shape. Therefore, supposedly adding prosodic features to the already used spectral features may lead to an improvement in the system's performance. Najim Dehak, Pierre Dumouchel in their work [19] , introduced the use of continuous prosodic features for speaker recognition and showed how they can be modeled using joint factor analysis. Similar features have been successfully used in language identification. These prosodic features were pitch and energy contours spanning a syllable-like unit. They were extracted using a basis consisting of Legendre polynomials. Tharmarajah Thiruvaran [20] used frequency modulation (FM) features for improving accuracy of speaker identification . Due to the similarity between amplitude modulation (AM) feature and the conventional Mel frequency cepstrum coefficients (MFCC), FM features were used. It was shown that, the correlation between FM feature components was observed to be very small compared with that of Mel filter bank log energies, thus reducing the need for decorrelation. FM feature components were shown to be very nearly Gaussian distributed, Digital Energy Separation Algorithm (DESA) was used as a front-end in speaker identification system.

E. Shriberg, L. Ferrer , S. Kajarekar in [21] described a new approach to modeling idiosyncratic prosodic behaviour for automatic speaker recognition. The approach computes various duration, pitch, and energy features for each estimated syllable in speech recognition output, quantizes the features, forms N-grams of the quantized values, and models normalized

counts for each feature N-gram using support vector machines (SVMs) referred to as SNERF-grams (N-grams of Syllable based Nonuniform Extraction Region Features).S.R. Mahadeva Prasanna, Jinu Mariam Zachariah and B. Yegnanarayana (2004) used the features from spectral, duration and pitch . The substantiation from the different sources were combined using a multilayer perceptron neural network. It was shown that not only that the performance of verification improved, but also the non- spectral features such as duration and pitch were found to be robust for variations due to channel [22]. In order to improve the speaker recognition accuracy, in [23], pitch was applied to GMM-based speaker recognition (SR). The circular average magnitude difference function (CAMDF) method was used to extract the pitch. An endpoint detection method based on the pitch was proposed. In this work, mel-frequency cepstral coefficient (MFCC) based on the pitch, the pitch contour, the pitch first-order difference and the pitch changed rate features were selected as the features of the SR. Experimental results showed improvement in recognition rate using proposed endpoint detection method, than that of the speaker recognition system using the MFCC parameters only.

The last decade has seen increased interest in exploring such higher-level features in automatic speaker recognition. High level features are based on voice timbre and accent/ pronunciation of speaker and also on lexicon - the kind of words the speakers tend to use in their conversations. The work on such "high-level" conversational features was initiated in [24] where a speaker's characteristic vocabulary, the so-called *idiolect*, was used to characterize speakers. The idea in "high-level" modeling is to convert each utterance into a sequence of *tokens* where the co-occurrence patterns of tokens characterize speaker differences. Elizabeth Shriberg in her article [25] demonstrated how higher-level features can contribute to performance in a state-of-the-art system. Various features such as cepstral and cepstral-derived, phonetic (acoustic tokenization), prosodic, lexical features along with their performance was discussed. It was shown that, systems based on frame-level cepstral or cepstral derived features show higher accuracy than longer-range systems. Within the set of cepstral-based systems, the MLLR system had best performance, because it takes advantage of linguistic information from ASR. Of the longer-range systems, the conditioned syllable-based prosody sequence system was shown to be the most successful. Recently a project titled *SuperSid* is undertaken to explore the effectiveness of high level information for speaker recognition [47]. Specifically, methods to extract and model speaker specific patterns in acoustic cues (the sound of the person's voice), speech prosody, word and phone pronunciations (idiosyncratic or dialectical distinctions), word usage (characteristic phrases or word selection), and interactions with conversational partners (taciturn or dominating in conversations) were examined. The fusion of features and classifiers to improve the recognition performance is proposed in this project.

In above section, we have discussed various features and features extraction techniques which is the front end of any speaker recognition system. The features discussed are from low-level spectral features such as MFCC, LPCC, PLP etc., representing the vocal track dynamics as well as features such as LP residues, pitch, pitch contours representing vocal fold or excitation source features. These features are related to physical traits of speaker. We have also discussed the prosodic and high level features which represent behavioural characteristics of speaker. The choice of feature is based on applications, accuracy demand and robustness in various operating conditions and environments, channel parameters, complexity of computation etc.

Feature Modeling

In the previous section, we have discussed so called *measurement* step in the speaker identification where a set of speaker discriminative characteristics is extracted from the speech signal. In this section, we go through the next step called *classification*, which is a decision making process of determining the author of a given speech signal based on the previously stored or learned information [1]. This step is usually divided into two parts, namely *matching* and *modeling*. The modeling is a process of enrolling speaker to the identification system by constructing a model of his/her voice, based on the features extracted from his/her speech sample. The matching is a process of computing a *matching score*, which is a measure of the similarity of the features extracted from the unknown speech sample and speaker model [26]. Once the feature vectors corresponding to the "speech" frames have been extracted , the associated speech data also known as training/enrolment data is used to build a speaker specific model. During the verification phase, the trained model is used to authenticate a sequence of feature vectors extracted from utterances of unknown speakers.

The statistical approaches for constructing the relevant models can be divided into two distinct categories: generative and discriminative. Training of generative models typically involves data specific to the target speakers where the objective is that the models can faithfully capture the statistical properties of the speaker specific speech signal. Training of discriminative models involves data corresponding to the target and imposter speakers and the objective is to faithfully estimate the parameters of the manifold which distinguishes the features for the target speakers from the features for the imposter speakers. An example of a popular generative model used in speaker verification is *Gaussian Mixture Models* (GMMs) and an example of a popular discriminative model is *Support Vector Machines* (SVMs).

Classical speaker models can be also categorized into nonparametric and parametric models. They are also called template models and stochastic models, respectively. Vector quantization (VQ) [30] and dynamic time warping (DTW) [31] are representative examples of template models for text-independent and text-dependent recognition, respectively. In stochastic

models, each speaker is modeled as a probabilistic source with an unknown but fixed probability density function. The training phase is to estimate the parameters of the probability density function from the training data. The likelihood of the test utterance with respect to the model is used for pattern matching. The Gaussian mixture model (GMM) [32,33] and the hidden Markov model (HMM) [34, 35] are the most popular stochastic models for text-independent and text-dependent speaker recognition, respectively. Speaker models can also be classified into generative and discriminative models. The generative models such as GMM and VQ estimate the feature distribution within each speaker independently. While the discriminative models such as artificial neural networks (ANNs) [36, 37] and support vector machines (SVMs) [38, 39] model the boundary between speakers.

Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm. In the earlier days, the design of a vector quantizer (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. In 1980, Linde, Buzo, and Gray (LBG) proposed a VQ design algorithm based on a training sequence. The use of a training sequence bypasses the need for multi-dimensional integration. A VQ that is designed using this algorithm are referred to in the literature as an LBG-VQ [40]. In 1985, Soong *et al.* used the LBG algorithm for generating speaker-based vector quantization (VQ) codebooks for speaker recognition [41].VQ is often used for computational speed-up techniques and lightweight practical implementations. It also provides competitive accuracy when combined with background model adaption(Kinnuen et al.2009). Zhong-Xuan [42] Yuan presented a new approach to vector quantization in which feature vector is represented by a binary vector called binary quantization (BQ). The performance criterion of vector quantization, distortion measure, was employed for investigating the effectiveness of BQ. The results shown very good performance in terms of memory space and computation required. Also the identification system had shown strong robustness in additive White Gaussian noise.

In 1995, Reynolds proposed Gaussian mixture modeling (GMM) classifier for speaker recognition task. GMM [33], is a stochastic model which has become the *de-facto* reference method in speaker recognition. The GMM needs sufficient data to model the speaker, and hence good performance. It can be considered as an extension of the VQ model, in which the clusters are overlapping. That is , a feature vector is not assigned to the nearest cluster as in VQ, but it has a nonzero probability of originating from each cluster. GMM is composed of multivariate Gaussian components [17]. A GMM super-vector characterizes a speaker's voice by the GMM parameters such as the mean vectors, covariance matrices and mixture weights.The parameters of the model are typically estimated by maximum likelihood estimation, using the *Expected-Maximization* algorithm. The matching function in GMM is defined in terms of *likelihood* [32,33]. 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 [43]. As GMM needs sufficient data to model the speaker, Reynolds in [44], introduced GMM-UBM (universal background model), in which UBM is trained from speech data collected from large number of speakers, which acts as a speaker independent model. In the GMM approach, speaker models are obtained from the adaption of a universal background model (UBM) through the maximum *a posteriori(MAP)* criterion. The UBM is usually trained by means of expectation-maximization (EM) algorithm from a background dataset, which includes a wide range of speakers, languages, communication channels, recording devices and environments. The GMM-UBM becomes a standard technique for text-independent speaker recognition due to its reliable performance. The discriminant classifier based on support vector machine (SVM) were of great interest in speech field. In speaker recognition, an important revolution was proposed , mainly by [45]. SVMs are typically trained in binary mode to discriminate between the speaker's data and impostor data. The impostor data consists of several speakers and can coincide with the data used to train the SI GMM. The resulting SVM is a hyperplane separating the two classes in the predefined kernel space. During testing, the same kernel is used to compute a signed distance between the test sample and the hyperplane. This distance is used as a similarity measure or score, with positive values indicating that the sample is on the target speaker side of the hyperplane (but note that the decision threshold may be set to a nonzero value to bias the outcome in accordance with a given decision cost model). One disadvantage of the GMM-based approach is that it models the features as a bag of frames ignoring sequence information. Researchers have explored other modeling techniques, such as hidden Markov models (HMMs), to model sequence information (Newman et al. 1996). HMM-based approaches have been shown to outperform the GMM-based approach given enough training data. Another approach has been to model blocks of features, preserving the temporal information (Gillick et al. 1995). It uses a mixed approach , associating the robustness of the statistical modeling provided by the GMM-UBM paradigm with the discriminating power of SVMs. This approach, denoted GMM super-vector SVM with linear kernel (GSL) , uses the GMM-UBM to model the training or testing data. A super-vector is extracted from the corresponding GMM (obtained from UBM by MAP procedure), composed by concatenation of the mean coefficients of all the GMM components. The super-vectors are then used as inputs of the SVM classifier [46].

Pattern matching and decision logic

The next step after computing of matching scores for every speaker model enrolled in the system is the process of assigning the exact classification mark for the input speech. Matching gives a score which represents how well the test feature vectors are close to the reference models This process depends on the selected matching and modeling algorithms. The feature extraction and pattern matching are same for different speaker recognition tasks, but the decision depends on the nature of task. In closed-set identification task, the decision is simply the speaker index that yields the maximum score. In template matching, decision is based on the computed distances, whereas in stochastic matching it is based on the computed probabilities. In template matching, the speaker model with smallest matching score is selected, whereas in stochastic matching, the model with highest probability is selected. Practically, decision process is not so simple and for example for so called open-set identification problem the answer might be that input speech signal does not belong to any of the enrolled speaker models. It is quite difficult to characterize the performance of speaker verification systems in all applications due to the complexities and differences in the enrolment/testing scenarios . Having computed a match score between the input speech-feature vector and a model of the claimed speaker's voice, a verification decision is made whether to accept or reject the speaker or request another utterance (or, without a claimed identity, an identification decision is made). If a verification system accepts an impostor, it makes a false acceptance (FA) error. If the system rejects a valid user, it makes a false rejection (FR) error. The FA and FR errors can be traded off by adjusting the decision threshold, as shown by a Receiver Operating Characteristic (ROC) curve. The operating point where the FA and FR are equal corresponds to the equal error rate. The accept or reject decision process can be an accept, continue, time-out, or reject hypothesis-testing problem. In this case, the decision making, or classification, procedure is a sequential hypothesis-testing problem. On the other hand, the computation of speaker identification is measured as a ratio of the number of correctly identified examples to the total number of examples considered for the testing.

Conclusion

In this paper, we have presented an overview of the various features, the extraction methods and modeling techniques of speaker recognition. The low level features such as cepstral features work well in ideal conditions, but their performance is degraded in real time situations. Use of high level information can add complementary knowledge to improve the performance of recognition system. In practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users. The techniques of robust feature extraction, feature normalization, model-domain compensation and score normalization methods are necessary. There are number of research problems that can be taken up, such as human-related error sources, real-time implementation, and forensic interpretation of speaker recognition scores. For this it is important to explore stable features that remain insensitive to variation of speakers voice over time and are robust against variation in voice quality due to physical states or disguises. The problem of distortion in the channels and background noise also requires being resolved with better techniques.

References

1. J.P. Campbell, Jr., "Speaker Recognition: A Tutorial," Proceedings of IEEE, vol. 85, no. 9 , Sept. 1997 .
2. Tomas F. Quatieri ,"Discrete Time Speech Signal Processing, Principles and Practice", Pearson Education 2006.
3. Marcos Faundez-Zanuy and Enric Monte-Moreno, "State-of-the-art in Speaker Recognition ", IEEE A&E Systems Magazine, May 2005
4. Jayanna, H.S. and S.R.M. Prasanna, 2009. Analysis,feature extraction, modeling and testing techniques for speaker recognition. IETE Technical Review, 2009, Volume 26, issue 3
5. 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.
6. Grazyna Demenko, "Analysis of suprasegmental features for speaker verification", 8th Australian International Conference on Speech Science and Technology, 2000
7. 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.
8. P.D. Bricker, " Statistical techniques for talker identification", Bell Svst. Techn. Jour. Vol.50 April 1971B. S. Atal "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification "J. Acoust. Soc. Am. Volume 55, Issue 6, pp. 1304-1312 (1974)
9. Tommy Kinnuen, "Spectral Features for Automatic Text-Independent Speaker Recognition " thesis *University of Joensuu* ,Dec.2003
10. Douglas A. Reynolds and Richard Rose , 'Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models ", IEEE transaction on Speech and Audio Processing, Vol.3, No.1, January 1995

11. D.A. Reynolds, "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech Audio Process. , vol. 2(4), pp. 639-43, Oct. 1994
12. X. Huang, A. Acero and H.-W. Hon, Spoken language processing, Upper Saddle River, New Jersey, Prentice Hall PTR, 2001.
13. N. J.-C. Wang, W.-H. Tsai, L.-S. Lee, "Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification", Eurospeech 2001 – Scandinavia
14. J. R. Deller, J. H. L. Hansen, J. G. Proakis, "Discrete-Time Processing of Speech Signals", Piscataway (N.J.), IEEE Press, 2000.
15. Andre G. Adami and Douglas A. Reynolds, "Modeling prosodic dynamics for Speaker Recognition ", IEEE ,ICASSP 2003
16. Tomi Kinnuen and Haizhou Li, "An overview of text independent speaker recognition: From features to supervectors ", ScienceDirect, Speech Communication 2010
17. Hassan Euaidi and Jean Rouaf, "Pitch and MFCC dependent GMM models for speaker identification systems " , CCECE IEEE, 2004
18. Najim Dehak, Pierre Dumouchel, and Patrick Kenny, "Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification" , IEEE Trans. Speech and Language Proces 2004.
19. Tharmarajah Thiruvaran, Eliathamby Ambikairajah, "Speaker Identification using FM Features ", ATP Research Laboratory, National ICT Australia
20. E. Shriberg, L. Ferrer , S. Kajarekar, A. Venkataraman, "Modeling prosodic feature equences for speaker recognition ", Science Direct, Speech Communication 2005
21. S.R. Mahadeva Prasanna, Jinu Mariam Zachariah and B. Yegnanarayana. " Neural Network Models for Combining Evidence from Spectral and Suprasegmental Features for Text-Dependent Speaker Verification" IEEE 2004
22. ZHU Jian-wei, SUN Shui-fa," Pitch in Speaker Recognition", IEEE 2009
23. Doddington, G. "Speaker recognition based on idiolectal differences between speakers", Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)
24. Elizabeth Shriberg, " Higher-Level Features in Speaker Recognition " SRI International, International Computer Science Institute, Berke
25. B. S. Atal, "Automatic Recognition of Speakers from their Voices", Proceedings of the IEEE, vol 64, 1976
26. H. Gish and M. Schmidt, "Text Independent Speaker Identification", IEEESignal Processing Magazine, Vol. 11, No. 4, 1994
27. J. M. Naik, "Speaker Verification: A Tutorial", IEEE Communications Magazine, January 1990
28. D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", ICASSP 2002
29. L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," Prentice-Hall, NJ, 1993
30. S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustics, Speech and Signal Processing ,April 1981
31. D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing , January 2000
32. D. Reynolds, and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. on Speech and Audio Processing 3, January 1995
33. J. Naik, L. Netsch, and G. Doddington, "Speaker verification over long distance telephone lines," In Proc. ICASSP, Glasgow, May 1989
34. M. BenZeghiba, and H. Bourland, "User-customized password speaker verification using multiple reference and background models," Speech Communication, September 2006
35. L. Heck, Y. Konig, M. Sonmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," Speech Communication , June 2000.
36. B. Yegnanarayana, and S. Kishore, "AANN: an alternative to GMM for pattern recognition," Neural Networks , April 2002
37. W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Computer Speech and Language, April 2006
39 V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," Proceedings of the 2000 IEEE Signal Processing Society Workshop, vol.2, 2000
40 Y. Linde, A. Buzo, and R. M. Gray, ``An Algorithm for Vector Quantizer Design," IEEE Transactions on Communications, January 1980.

41 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, Apr. 1985

42 Zhong-Xuan Yuan and Chong-Yu, "Binary quantization of feature vectors for robust text independent speaker identification", IEEE Trans. Speech Audio Process. , vol. 7(1), January 1999

43 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, 2006

44 D.A. Reynolds, T.F. Quateri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing , vol. 10, 2000.

45 W. Campbell, D. Sturim, D. Reynolds and A. Solomonoff, "SVM based speaker verification using GMM supervector kernel and NAP variability compensation", Proc. Int. Conf. Acoustics, Speech and Signal Processing, 2006

46 Joseph P. Campbell, W.M. Campbell, Wade Shen, " Forensic speaker recognition-A need for caution ", IEEE Signal Processing Magazine, March 2009

47 Douglas Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andre Adami, Qin Jin, Daviekc Klusacek, Joy Abramson, Radu Mihaescu, John Godfrey, Douglas Jones, Bing Xiang, "Exploiting High-Level Information for High-Performance Speaker Recognition" , 2002