Speaker Recognition

Homayoon Beigi
Research, Recognition Technologies, Inc., South Salem, NY, USA

Synonyms

Speaker biometrics; Speaker identification and verification (SIV); Talker recognition; Voice biometrics; Voice recognition; Voiceprint recognition

Definition

Speaker Recognition is a multi-disciplinary technology which uses the vocal characteristics of speakers to deduce information about their identities. It is a branch of biometrics that may be used for identification, verification, and classification of individual speakers, with the capability of tracking, detection, and segmentation by extension.

Background

In addressing the act of *speaker recognition*, many different terms have been coined, some of which have caused great confusion. *Speech recognition* research has been around for a long time, and, naturally, there is some confusion in the public between *speech* and *speaker* recognition. One term that has added to this confusion is *voice recognition*.

The term *voice recognition* has been used in some circles to double for *speaker recognition*. Although it is conceptually a correct name for the subject, it is recommended that the use of this term is avoided. *Voice recognition*, in the past, has been mistakenly applied to *speech recognition*, and these terms have become synonymous for a long time. In a speech recognition application, it is not the voice of the individual which is being recognized, but the contents of his/her speech. Alas, the term has been around and has had the wrong association for too long.

Other than the aforementioned, there have been a myriad of different terminology used to refer to this subject. These include *voice biometrics*, *speech biometrics*, *biometric speaker identification*, *talker identification*, *talker clustering*, *voice identification*, *voiceprint identification*, and so on. With the exception of the term *speech biometrics* which also introduces the addition of a speech knowledge-base to speaker recognition, the rest do not present any additional information.

A human child develops an inherent ability to identify the voice of his/her parents before even learning to understand the content of their speech. In humans, speaker recognition is performed in the right (less dominant) hemisphere of the brain in conjunction with the functions for processing pitch, tempo, and other musical discourse. This is in contrast with most of the language functions

S. Jajodia et al. (eds.), *Encyclopedia of Cryptography, Security and Privacy*,
https://doi.org/10.1007/978-3-642-27739-9_747-2

(production and perception) in the brain which are processed by the *Broca* and *Wernicke* areas in the left (dominant) hemisphere of the *cerebral cortex* (Beigi 2011).

A speaker recognition system first tries to model the vocal tract characteristics of a person. This may be a mathematical model of the physiological system producing the human speech or simply a statistical model with similar output characteristics as the human vocal tract. Once a model is established and has been associated with an individual, new instances of speech may be assessed to determine the likelihood of them having been generated by the model of interest in contrast with other observed models. This is the underlying methodology for all speaker recognition applications. The earliest known papers on speaker recognition were published in the 1950s (Pollack et al. 1954; Shearme and Holmes 1959). Initial speaker recognition techniques relied on a human expert examining representations of the speech of an individual and making a decision on the person's identity by comparing the characteristics in this representation with others. The most popular representation was the *formant* representation. In the recent decades, fully automated speaker recognition systems have been developed and are in use (Beigi 2011).

As for the importance of speaker recognition, it is noteworthy that *speaker identity* is the only biometric which may be easily tested (identified or verified) remotely through the existing infrastructure, namely, the telephone network. This makes speaker recognition quite valuable and unrivaled in many real-world applications. It needs not be mentioned that with the growing number of cellular (mobile) telephones and their ever-growing complexity, speaker recognition will become more popular in the future.

Speaker Enrollment

The first step required in most manifestations of speaker recognition is to enroll the users of interest. This is usually done by building a mathematical model of a sample speech from the user and storing it in association with an identifier. This model is usually designed to capture statistical information about the nature of the audio sample and is mostly irreversible – namely, the enrollment sample may not be reconstructed from the model.

Speaker Verification (Authentication)

In a generic speaker verification application, the person being verified (known as the test speaker) identifies himself/herself, usually by non-speech methods (e.g., a username, an identification number, etc.). The provided ID is used to retrieve the enrolled model for that person which has been stored according to the enrollment process, described earlier, in a database. This enrolled model is called the *target speaker model* or the *reference model*. The speech signal of the test speaker is compared against the target speaker model to verify the test speaker.

Of course, comparison against the target speaker's model is not enough. There is always a need for contrast when making a comparison. Therefore, one or more competing models should also be evaluated to come to a verification decision. The competing model may be a so-called (universal) background model or one or more cohort models. The final decision is made by assessing whether the speech sample given at the time of verification is closer to the target model or to the competing model(s). If it is closer to the target model, then the user is verified and otherwise rejected.

The speaker verification problem is known as a one-to-one comparison since it does not necessarily need to match against every single person in the database. Therefore, the complexity of the matching does not increase as the number of enrolled subjects increases. Of course in reality, there is more than one comparison for speaker verification, as stated – comparison against the target model and the competing model(s).

Speaker Identification

There are two different types of speaker identification, *closed-set* and *open-set*. Closed-set identification is the simpler of the two problems. In close-set identification, the audio of the test speaker is compared against all the available speaker models, and the speaker ID of the model with the closest match is returned. In practice, usually, the top best matching candidates are returned in a ranked list, with corresponding confidence or likelihood scores. In closed-set identification, the ID of one of the speakers in the database will always be closest to the audio of the test speaker; there is no rejection scheme.

One may imagine a case where the test speaker is a 5-year-old child where all the speakers in the database are adult males. In closed-set identification, still, the child will match against one of the adult male speakers in the database. Therefore, closed-set identification is not very practical. Of course, like anything else, closed-set identification also has its own applications. An example would be a software program which would identify the audio of a speaker so that the interaction environment may be customized for that individual. In this case, there is no great loss by making a mistake. In fact, some match needs to be returned just to be able to pick a customization profile. If the speaker does not exist in the database, then there is generally no difference in what profile is used, unless profiles hold personal information, in which case rejection will become necessary.

Open-set identification may be seen as a combination of closed-set identification and speaker verification. For example, a closed-set identification may be conducted, and the resulting ID may be used to run a speaker verification session. If the test speaker matches the target speaker based on the ID, returned from the closed-set identification, then the ID is accepted and passed back as the true ID of the test speaker. On the other hand, if the verification fails, the speaker may be rejected all-together with no valid identification result. An open-set identification problem is therefore at least as complex as a speaker verification task (the limiting case being when there is only one speaker in the database), and most of the time it is more complex. In fact, another way of looking at verification is as a special case of open-set identification in which there is only one speaker in the list. Also, the complexity generally increases linearly with the number of speakers enrolled in the database since, theoretically, the test speaker should be compared against all speaker models in the database – in practice this may be avoided by tolerating some accuracy degradation (Beigi et al. 1999).

Speaker and Event Classification

The goal of classification is a bit more vague. It is the general label for any technique that pools similar audio signals into individual bins. Some examples of the many classification scenarios are gender classification, age classification, and event classification. Gender classification, as is apparent from its name, tries to separate male speakers and female speakers. More advanced versions also distinguish children and place them into a separate bin; classifying male and female is not so simple in children since their vocal characteristics are quite similar before the onset of puberty. Classification may use slightly different sets of features from those used in verification and identification, depending on the problem at hand. Also, either there may be no enrollment, or enrollment may be done differently (Beigi 2011).

Although these methods are called speaker classification, sometimes, the techniques are used for doing event classification such as classifying speech, music, blasts, gun shots, screams, whistles, horns, etc. The feature selection and processing methods for classification are mostly dependent on the scope and could be different from mainstream speaker recognition.

Speaker Segmentation, Diarization, Detection, and Tracking

Automatic segmentation of an audio stream into parts containing the speech of distinct speakers, music, noise, and different background conditions has many applications. This type of segmentation is elementary to the practical considerations of speaker recognition as well as speech and other audio-related recognition systems. Different specialized recognizers may be used for recognition of distinct categories of audio in a stream.

An example is the ever-growing teleconferencing application. In a tele-conference, usually, a host makes an appointment for a conference call and notifies attendees to call a telephone number and to join the conference using a special access code. There is an increasing interest from the involved parties to obtain transcripts (minutes) of these conversations. In order to fully transcribe the conversations, it is necessary to know the speaker of each statement. If an enrolled model exists for each speaker, then prior to identifying the active speaker (*speaker detection*), the audio of that speaker should be segmented and separated from adjoining speakers. When speaker segmentation is combined with speaker identification and the resulting index information is extracted, the process is called *speaker diarization*. In case one is only interested in a specific speaker and where that speaker has spoken within the conversation (the timestamps), the process is called *speaker tracking*.

Speaker Verification Modalities

There are two major ways in which speaker verification may be conducted. These two are called the *modalities* of speaker verification, and they are *text-dependent* and *text-independent*. There are also variations of these two modalities such as *text-prompted*, *language-independent text-independent*, and *language-dependent text-independent*.

In a purely *text-dependent* modality, the speaker is required to utter a predetermined text at enrollment and the same text again at the time of verification. Text-dependence does not really make sense in an identification scenario. It is only valid for verification. In practice, using such text-dependent modality will be open to *spoofing* attacks; namely, the audio may be intercepted and recorded to be used by an impostor at the time of the verification. Practical applications that use the text-dependent modality do so in the text-prompted flavor. This means that the enrollment may be done for several different textual contents, and at the time of verification, one of those texts is requested to be uttered by the test speaker. The chosen text is the prompt and the modality is called *text-prompted*.

A more flexible modality is the *text-independent* modality in which case the texts of the speech at the time of enrollment and verification are completely random. The difficulty with this method is that because the texts are presumably different, longer enrollment and test samples are needed. The long samples increase the probability of better coverage of the idiosyncrasies of the person's vocal characteristics.

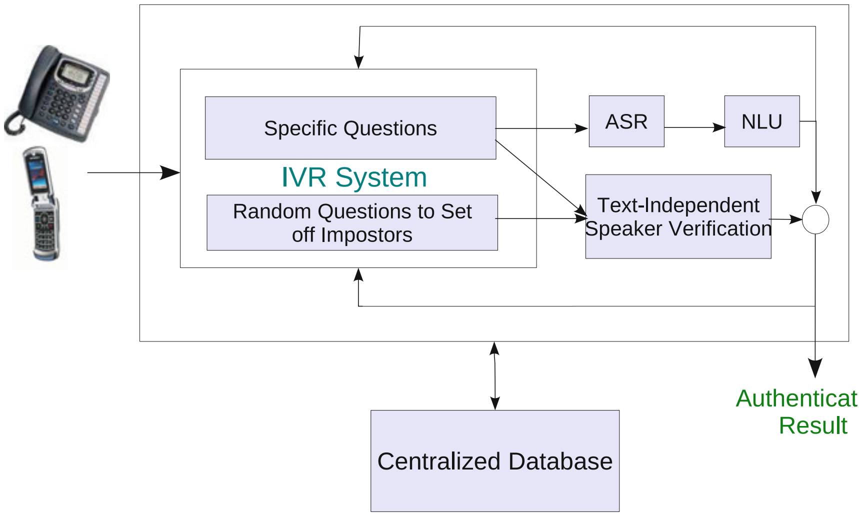
The general tendency is to believe that in the text-dependent and text-prompted cases, since the enrollment and verification texts are identical, they can be designed to be much shorter. One must be careful, since the shorter segments will only examine part of the dynamics of the vocal tract. Therefore, the text for text-prompted and text-dependent engines must still be designed to cover enough variation to allow for a meaningful comparison.

The problem of spoofing is still present with text-independent speaker verification. In fact, any recording of the person's voice should now get an impostor through. For this reason, text-independent systems would generally be used with another source of information in a multi-factor authentication scenario.

In most cases, *text-independent* speaker verification algorithms are also *language-independent*, since they are concerned with the vocal tract characteristics of the individual, mostly governed by the shape of the speaker's vocal tract. However, because of the coverage issue discussed earlier, some researchers have developed text-independent systems which have some internal models associated with phonemes in the language of their scope. These techniques produce a text-independent, but somewhat language-dependent speaker verification system. The language limitations reduce the space and, hence, may reduce the error rates.

Knowledge-Based Speaker Recognition (Speech Biometrics)

A knowledge-based speaker recognition system is usually a combination of a speaker recognition system and a speech recognizer and sometimes a natural language understanding engine or more. It is somewhat related to the *text-prompted* modality with the difference that there is another abstraction layer in the design. This layer uses knowledge from the speaker to test for liveness or act as an additional authentication factor. As



Speaker Recognition, Fig. 1 A practical speaker recognition system utilizing speech recognition and natural language understanding

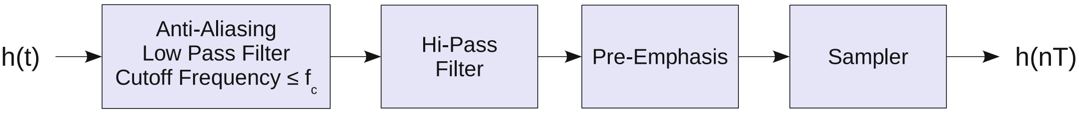
an example, at the enrollment time, specific information such as a Personal Identification Number (PIN) or other private data may be stored about the speakers. At the verification time, randomized questions may be used to capture the test speaker's audio and the content of interest. The content is parsed by doing a transcription of the audio and using a natural language understanding (Manning 1999) system to parse for the information of interest. This will increase the factors in the authentication and is usually a good idea for reducing the chance of successful impostor attacks – see Fig. 1.

Theory

Speaker recognition is a multi-disciplinary science. In its theory and implementation, it has a great deal in common with speech recognition (Rabiner and Juang 1990). It is impossible to cover the theory in this limited venue. The following disciplines are directly relevant: *signal processing*, *phonetics and phonology*, *information theory*, *Bayesian statistics and learning*, *optimization theory*, *parameter estimation*, *artificial intelligence and deep learning*, and *applied mathematics*. Reference Beigi (2011) provides a comprehensive coverage of the theory. An attempt is made here to list the different techniques which are used for speaker recognition.

As mentioned, the first step is to store the vocal characteristics of the speakers in the form of speaker models in a database for future reference. First, the features should be defined such that they would best represent the vocal characteristics of the speaker of interest. The most prevalent features used in the field happen to be identical to those used for speech recognition, namely, Mel Frequency Cepstral Coefficients (MFCCs).

Before extracting features, the audio signal should be sampled and made available with a fixed frequency which is determined based on the sampling theorem such that most of the information in the speech sample is preserved. There are many aspects to consider when the speech is sampled and stored in a standard format to be used by the speaker recognition engine. Figure 2 shows a typical sampling process which starts with an analog signal and produces a series of discrete samples at a fixed frequency, representing the speech signal. The discrete samples are usually stored using a Codec (Coder/Decoder) format such as linear PCM, MU-Law, A-Law, etc. Stan-



Speaker Recognition, Fig. 2 Block diagram of a typical sampling process

dardization is quite important for interoperability of different engines (Beigi 2011). The system of Fig. 2 should be designed so that it reduces *aliasing*, *truncation*, and *band-limitation* by choosing the right parameters such as the sampling rate and volume normalization. Note that in an ideal case, the filtering and pre-emphasis should be done in the analog domain, before sampling takes place, to avoid aliasing and loss of precision. The low-pass filter handles anti-aliasing, the high-pass filter removes unwanted DC components, and the pre-emphasis balances the power of low and high frequency components of the signal by increasing the naturally low energy levels of the higher frequency components so that they may provide useful information about high frequency phones, such as fricatives.

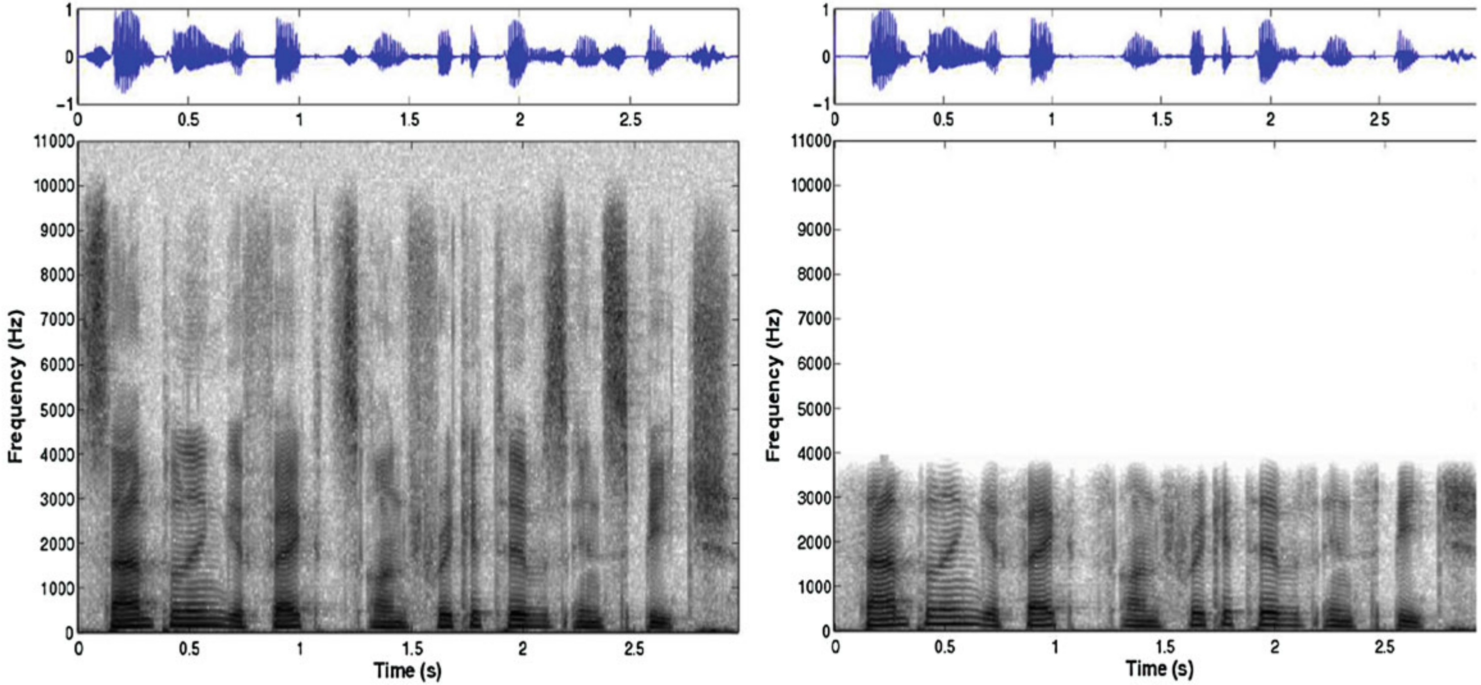
Figure 3 shows how most of the fricative information is lost going from a 22 kHz sampling rate to 8 kHz. Normal telephony sampling rates are at best 8 kHz. mostly everyone is familiar with having to qualify fricatives on the telephone by using statements such as “S” as in “Sam” and “F” as in “Frank.”

Cepstral Coefficients have stemmed from studies in exploring the arrival of echos in nature (Bogert et al. 1963). They are related to the spectrum of the log of the spectrum of a speech signal. The frequency domain of the signal in computing the MFCCs is warped to the Melody (Mel) scale. It is based on the premise that human perception of pitch is linear up to 1000 Hz and then becomes nonlinear for higher frequencies (somewhat logarithmic). There are models of the human perception based on other warped scales such as the Bark scale – which is also related to the Mel scale (Beigi 2011). There are several ways of computing Cepstral Coefficients. They may be computed using the Direct Method, also known as Moving Average (MA) which utilizes the Fast Fourier Transform (FFT) for the first pass and the inverse Discrete Cosine Transform (DCT) for the second pass to ensure real coefficients. Figure 4 shows the block-diagram for the main steps in the computation of the MFCCs.

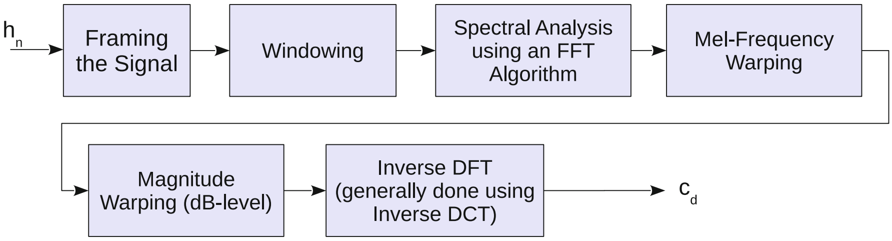
Some use the Linear Predictive, also known as AutoRegressive (AR) features by themselves: Linear Predictive Coefficients (LPC), Partial Correlation (PARCOR) – also known as reflection coefficients – or log area ratios (LAR). However, mostly the LPCs are converted to cepstral coefficients using autocorrelation techniques. These features are called Linear Predictive Cepstral Coefficients (LPCCs). There are also the Perceptual Linear Predictive (PLP) (Hermansky 1990) features, shown in Fig. 5. PLP works by warping the frequency and spectral magnitudes of the speech signal based on auditory perception tests. The domain is changed from magnitudes and frequencies to loudness and pitch (Beigi 2011).

There have been an array of other features used such as *wavelet filterbanks* (Burrus et al. 1997), for example, in the form of Mel-Frequency Discrete Wavelet Coefficients and Wavelet Octave Coefficients of Residues (WOCOR). There are also Instantaneous Amplitudes and Frequencies which are in the form of Amplitude Modulation (AM) and Frequency Modulation (FM). These features come in different flavors such as Empirical Mode Decomposition (EMD), FEPSTRUM, Mel Cepstrum Modulation Spectrum (MCMS), and so on (Beigi 2011).

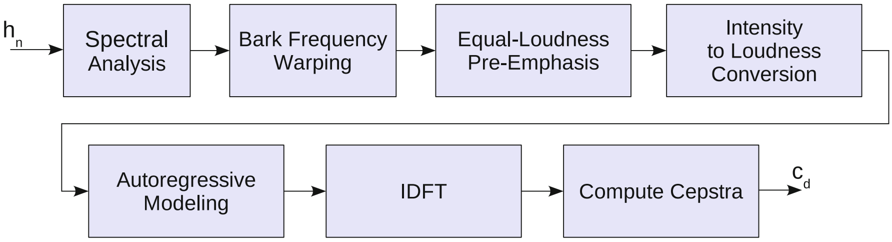
It is important to note that most audio segments include a good deal of silence. Addition of features extracted from silent areas in the speech will increase the similarity of models, since silence does not carry any information about the speaker’s vocal characteristics. Therefore, Silence Detection (SD) or Voice Activity Detection (VAD) (Beigi 2011) is quite important for better results. Only segments with vocal sig-



Speaker Recognition, Fig. 3 Utterance: “Sampling Effects on Fricatives in Speech,” sampled at 22 kHz (left) and 8 kHz (right)


h_n
Framing the Signal
Windowing
Spectral Analysis using an FFT Algorithm
Mel-Frequency Warping
Magnitude Warping (dB-level)
Inverse DFT (generally done using Inverse DCT)
c_d

Speaker Recognition, Fig. 4 Mel Frequency Cepstral Coefficient (MFCC) computation



Speaker Recognition, Fig. 5 A typical Perceptual Linear Predictive (PLP) system

nals should be considered for recognition. Other preprocessing such as Audio Volume Estimation and normalization and Echo Cancellation may also be necessary for obtaining desirable result (Beigi 2011).

Once the features of interest are chosen, models are built based on these features to represent the speakers' vocal characteristics. At this point, depending on whether the system is text-dependent (including text-prompted) or text-independent, different methods may be used.

Gaussian Mixture Model (GMM): Generic Case

The models are tied to the type of learning that is done. A historically popular technique is the use of a Gaussian Mixture Model (GMM) (Duda and Hart 1973) to represent the Speaker (Beigi 2011). This is mostly relevant to the text-independent case which encompasses speaker identification and text-independent verification. Even text-dependent techniques can use GMMs, but they usually use a GMM to initialize Hidden Markov Models (HMMs) (Poritz 1988) built to have an inherent model of the content of the speech as well. Of course a GMM is also considered to be a so-called degenerate single state HMM. Many speaker diarization (segmentation and ID) systems use GMMs. To build a Gaussian Mixture Model of a speaker's speech, one should make a few assumptions and decisions. The first assumption is the number of Gaussians to use. This is dependent on the amount of data that is available and the dimensionality of the feature vectors. Standard clustering techniques are usually used for the initial determination of the Gaussians. Once the number of Gaussians is determined, some large pool of features is used to train these Gaussians (learn the parameters). This step is called training.

After the training is done, generally, the basis for a speaker independent model is built. At this stage, depending on whether a Universal Background Model (UBM) (Reynolds et al. 2000) or Cohort Models are desired, different processing is done. For a UBM, a pool of speakers is used to optimize the parameters of the Gaussians as well as the mixture coefficients, using standard techniques such as Maximum Likelihood Estimation (MLE), Maximum a-Posteriori (MAP) adaptation, and Maximum Likelihood Linear Regression (MLLR). There may be one or more Background models. For example, some create a single background model called the UBM; others may build one for each gender, by using separate male and female databases for the training. Cohort models are built in a similar fashion. A cohort is a set of speakers that have similar vocal characteristics to the target speaker.

At this point, the system is ready for performing the enrollment. The enrollment may be done by taking a sample audio of the target speaker and adapting it to be optimal for fitting this sample. This ensures that the likelihoods returned by matching the same sample with the modified model would be maximal.

At the identification and verification stage, a new sample is obtained for the test speaker. In the identification process, the sample is used to compute the likelihood of this sample being generated by the different models in the database. The identity of the model that returns the highest likelihood is returned as the identity of the test speaker. In identification, the results are usually ranked by likelihood. To ensure a good dynamic range and better discrimination capability, log of the likelihood is computed.

At the verification stage, the process becomes very similar to the identification process described earlier, with the exception that instead of computing the log likelihood for all the models in the database, the sample is only compared to the model of the target speaker and the background or cohort models. If the target speaker model provides a better log likelihood, the test speaker is verified and otherwise rejected. The comparison is done using the Log Likelihood Ratio (LLR).

Of course, there are many other techniques used for the modeling of speakers, including speaker-space based models which may utilize GMMs on their own or in a *factor analytic* setting and of course the use of *neural network* techniques, as well as kernel-based methods such as *support vector machines* (Vapnik 1998) and other learning and classification approaches.

Speaker Space

The idea is to be able to represent each speaker as a point in a multidimensional speaker space. This would be some kind of speaker embedding. It was first introduced in the late 1990s (Beigi et al. 1999). Once a speaker may be formed as a point in such a space, it may be possible to define metrics and divergences in order to compare two speakers in that space. Once such a measure is established, as long as it is robust to variations in channel mismatch, conditions, and content, all modalities of speaker recognition can be used efficiently.

About a decade later, similar ideas for considering a speaker representation space stemmed from work on different variants of Factor Analysis (Campbell et al. 2006; Kenny 2006) which attempted to reduce speaker variability under different channel conditions. A simplification of the factor analysis approach started combining the channel/session and speaker variability in a so-called *total variability space* (Dehak et al. 2011), showing that most of the difference in the variability between sessions and channels may be resolved using other compensation techniques, once different data points have been established for speakers in the total variability space (speaker space) and they would be as effective as the more complex *joint factor analysis*. These so-called speaker identity vectors (*i-vectors*) may then undergo channel/session normalization and compensation such as *Within-Class Covariance Normalization*, *Linear Discriminant Analysis*, *Nuisance Attribute Projection*, or other techniques to handle mismatches in session and channel conditions. Once these normalizations, mostly arriving at a lower dimensional vector through projections, are compared to the original longer vectors, it would be easy to compare two speakers using simple distances in the speaker space, such as a simple cosine distance, etc.

Around the same time that researchers in the speaker recognition field were looking into the above, similar work was being done, based on the concept of *Linear Discriminant Analysis (LDA)* to provide an inference for the facial identity of individuals (Prince and Elder 2007). This technique is known as *Probabilistic Linear Discriminant Analysis* which is based on the premise that a parametric probability density may be defined for different conditions and classes, in such a way that it would allow scores for unseen labels (classes) to be estimated based on seen classes. This would allow the determination of a scoring system that would generally operate on the output of a projected set of vectors from an LDA, to provide optimal scores that would separate vectors in the speaker (facial) space, even for classes (speakers or faces) that have not been seen before, in a variety of channel conditions (Burget et al. 2011).

Speaker Embeddings (x-Vectors): Neural Network Kernels

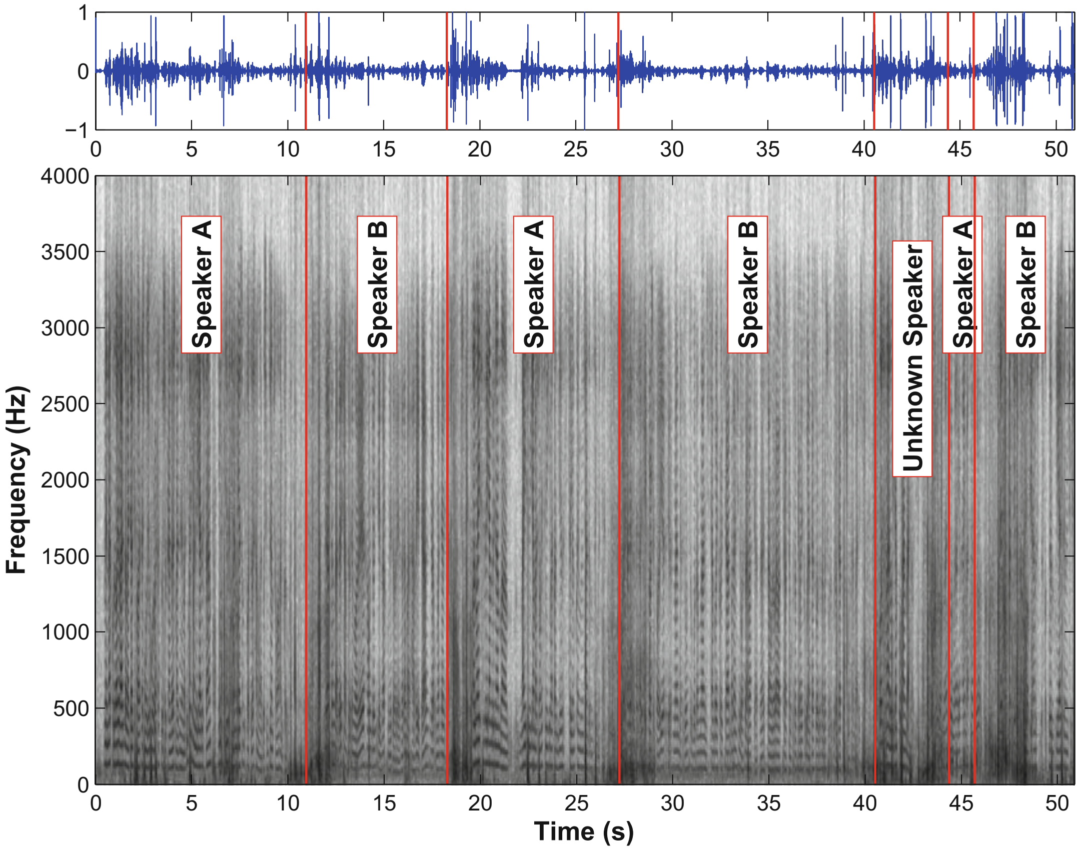
The latest and most successful techniques for doing speaker recognition use a neural network of some kind, such as a *Time Delay Neural Network (TDNN)* to create speaker embeddings by training the neural network on very large datasets with many speakers, recorded under a variety of conditions (Snyder et al. 2018). By training on a variety of speakers, a network is trained to tell the speakers apart. Generally a neural network kernel works in such a way that lower layers of the network learn the more fundamental features of speech and speaker identity. As the layers progress, more discourse and higher level reasoning is built into the outputs of the neurons in those layers. Utilizing this feature of neural networks, once training has been completed for the large training set, producing speaker identities at a standard output layer such as a Softmax layer, somewhat lower layers leading to the final layer would qualify as embeddings that would describe the speaker identity in this speaker space using a so-called *speaker embedding* – also known as *x-vectors*. These speaker embeddings may, for example, be the output of a layer of neurons just a few layers short of the final layer. Once such a vector, generally in the order of hundreds of dimensions, is established, a standard projection technique such as LDA, followed by a PLDA, can provide a reduction in dimensionality, as well as a metric for comparing the reduced dimensionality speaker coordinates. Given the property of the PLDA discussed earlier, features from an unseen speaker may be forwarded through the trained network to obtain the speaker embedding. This embedding may then be projected using an LDA. If another set of features undergoes the same process, the outputs of the two samples after the LDA may be compared using the PLDA parameters computed from the large training set, resulting in a score that would generally go in one direction for similar speakers and in the opposite direction for sufficiently different speakers. This completes the requirements set by Beigi et al. (1999) in order to be able to perform all modalities of speaker recognition. In order to improve the generalization of the speaker embeddings, many different data augmentation procedures have been used (Snyder et al. 2018).

Speaker Diarization

An extension of speaker recognition is diarization which includes segmentation followed by speaker identification and sometimes verification. The segmentation finds abrupt changes in the audio stream. Bayesian Information Criterion (BIC) (Chen and Gopalakrishnan 1998) and Generalized Likelihood Ratio (GLR) techniques and their combination (Ajmera and McCowan 2004) as well as other techniques (Beigi and Maes 1998) have been used for the initial segmentation of the audio. Once the initial segmentation is done, a limited speaker identification procedure allows for tagging the different parts with different labels. Figure 6 shows such results for a two-speaker segmentation.

Speaker identification results are usually presented in terms of the error rate. They may also be presented as the error rate based on the result being present in the top *N* matches. This case is usually more prevalent in the cases where identification is used to prune a large set of speakers to only a handful of possible matches so that another expert system (human or machine) would finalize the decision process.

In the case of speaker verification, the method of presenting the results is somewhat more controversial. In the early days in the field, a *Receiver Operating Characteristic (ROC)* curve was used (Beigi 2011). For the past two decades, the *Detection Error Trade-Off (DET)* curve (Martin et al. 1997; Martin and Mark 2000) has been more prevalent, with a measurement of the cost of producing the results, called the *Detection Cost Function (DCF)* (Martin



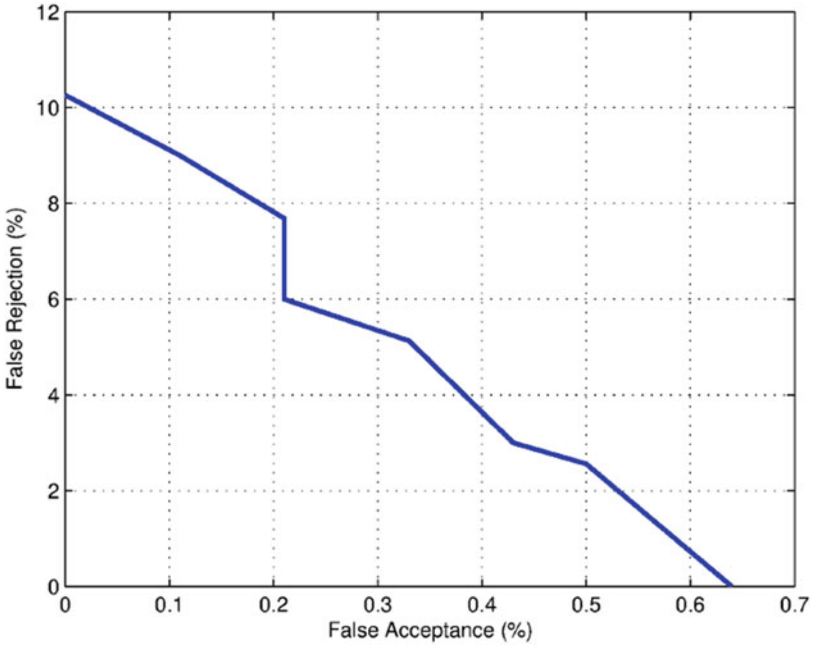
Speaker Recognition, Fig. 6 Segmentation and labeling of two speakers in a conversation using turn detection followed by identification

and Mark 2000), defined by in the process of providing National Institute of Standards in Technology (NIST) trials for speaker recognition, designed to advance the speaker recognition research. Figures 7 and 8 show sample ROC curves for two sets of data underscoring the difference in performances. Recognition results are usually quite data-dependent. The next section will speak about some open problems which degrade results. The DET curve uses logarithmic plots of the false rejection, which happens to be the *miss probability*, according to hypothesis testing definitions (Beigi 2011) against false acceptance which is referred to as the *False Alarm Probability*. Figures 9 and 10 show a sample ROC curve together with its corresponding DET curve side-by-side.

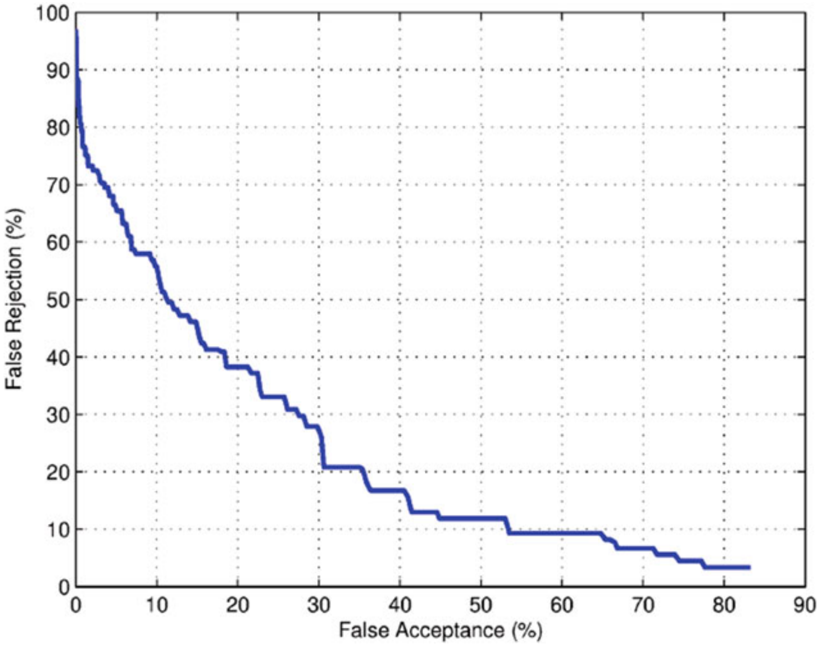
There is a controversial operating point on the DET curve which is usually marked as the point of comparison between different results. This point is called the *Equal Error Rate (EER)* and signifies the operating point where the false rejection rate and the false acceptance rate are equal. This point does not carry any real preferential information about the "correct" or "desired" operating point. It is mostly a point of convenience which is easy to denote on the curve.

Application

There are countless numbers of applications for the different branches of speaker recognition. These include, but are certainly not limited to, *financial*, *forensic and legal* (Nolan 1983; Tosi



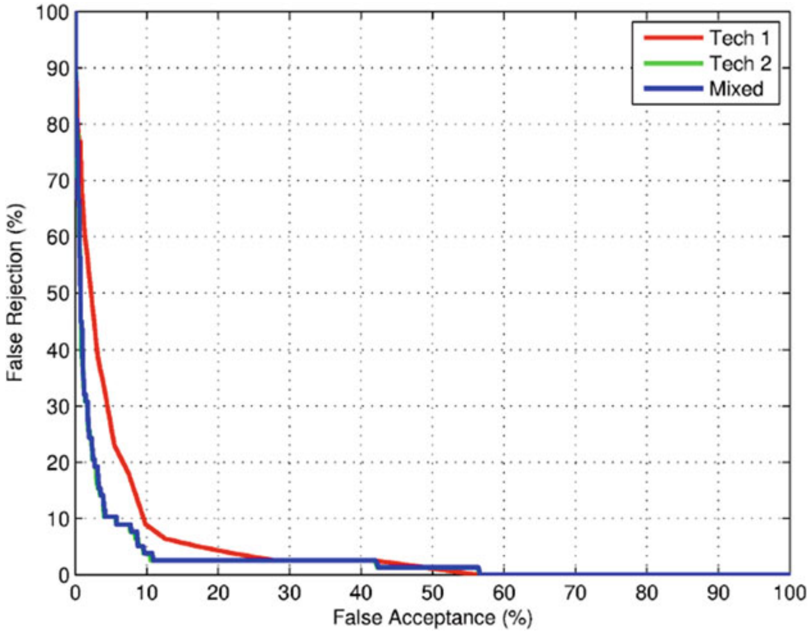
Speaker Recognition, Fig. 7 ROC curve for quality data



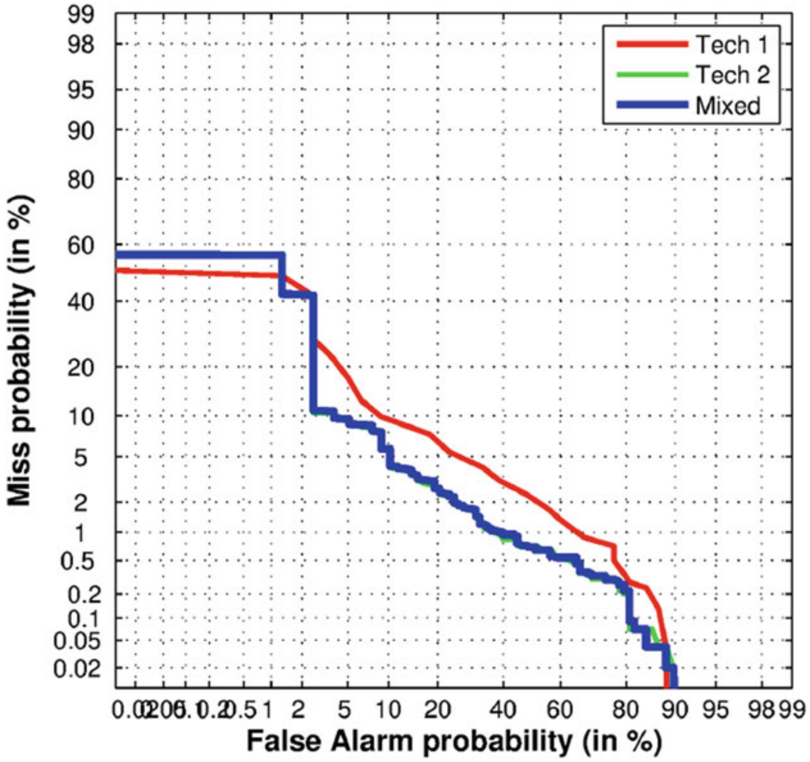
Speaker Recognition, Fig. 8 ROC curve – mismatched, noisy data

1979), *access control and security*, *audio/video indexing and diarization*, *surveillance*, *teleconferencing*, and *proctor-less distance learning*.

In designing a practical speaker recognition system, one should try to affect the interaction between the speaker and the engine to be able

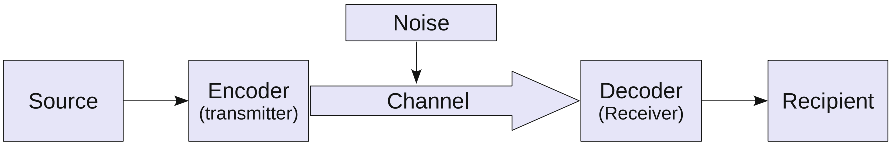


Speaker Recognition, Fig. 9 Sample ROC curve



Speaker Recognition, Fig. 10 Equivalent DET curve

to capture as many vowels as possible. Vowels are periodic signals which carry much more information about the resonance subtleties of the vocal tract. In the text-dependent and text-prompted cases, this may be done by actively designing prompts that include more vowels. For text-independent cases, the simplest way is to require more audio in hopes that many vowels



Speaker Recognition, Fig. 11 One-way communication

would be present. Also, when speech recognition and natural language understanding modules are included (Fig. 1), the conversation may be designed to allow for higher vowel production by the speaker.

Open Problems

The greatest challenge in speaker recognition is the so-called channel-mismatch problem. Considering the general communication system given by Fig. 11, it is apparent that the channel and noise characteristics at the time of communication are modulated with the original signal. Removing these channel effects is the most important problem in information theory. This is of course a problem when the goal is to recognize the message being sent. It is, however, a much bigger problem when the quest is the estimation of the model that generated the message – as it is with the speaker recognition problem. In that case, the channel characteristics have mixed in with the model characteristics and their separation is nearly impossible. Once the same source is transmitted over an entirely different channel with its own noise characteristics, the problem of learning the source model becomes even harder.

Many techniques are used for alleviating this problem, but it is still the most important source of errors in speaker recognition. It is the reason why most systems that have been trained on a predetermined set of channels such as landline telephone could fail miserably when cellular (mobile) telephones are used. Some of the techniques that have been traditionally used in the industry are listed here, but there are more techniques being introduced every day:

- ***Spectral Filtering and Cepstral Liftering***
  - Cepstral Mean Subtraction (CMS) or Cepstral Mean Normalization (CMN) (Benesty et al. 2008)
  - Cepstral Mean and Variance Normalization (CMVN) (Benesty et al. 2008)
  - Histogram Equalization (HEQ) (de la Torre et al. 2005) and Cepstral Histogram Normalization (CHN) (Benesty et al. 2008)
  - AutoRegressive Moving Average (ARMA) (Benesty et al. 2008)
  - RelAtive SpecTrAl (RASTA) Filtering (Hermansky 1991; van Vuuren 1996)
  - J-RASTA (Hardt and Fellbaum 1997)
  - Kalman Filtering (Kim 2002)
- ***Data Augmentation*** (Snyder et al. 2018)
  - Convolutional Noise such as Reverberation Random Room Characteristics
  - Additive Background Noise – Random signal to noise ratios
    Addition of Babble
    Addition of Music
    Addition of Noise
- ***Other Techniques***
  - Vocal Tract Length Normalization (VTLN) – first introduced for speech recognition: Chau et al. (2001) and later for speaker recognition (Grashey and Geibler 2006)
  - Feature Warping (Pelecanos and Sridharan 2001)
  - Feature Mapping (Reynolds 2003)
  - Speaker Model Synthesis (SMS) (Teunen et al. 2000)
  - Speaker Model Normalization (Beigi 2011)

– H-Norm (Handset Normalization) (Dunn et al. 2000)
– Z-Norm and T-Norm (Auckenthaler et al. 2000)

There are many challenges that have not been fully addressed in different branches of speaker recognition. For example, the large-scale speaker identification problem is one that is quite hard to handle. In most cases when researchers speak of large-scale in the identification arena, they speak of a few thousand enrolled speakers. As the number of speakers increases to millions or even billions, the problem becomes quite challenging. As the number of speakers increases, doing an exhaustive match through the whole population becomes almost computationally implausible. Hierarchical techniques (Beigi et al. 1999) would have to be utilized to handle such cases. In addition, the speaker space is really a continuum. This means that if one considers a space where speakers who are closer in their vocal characteristics would be placed near each other in that space, then as the number of enrolled speakers increases, there will always be a new person that would fill in the space between any two neighboring speakers. Since there are intra-speaker variabilities (differences between different samples taken from the same speaker), the intra-speaker variability will be at some point more than inter-speaker variabilities, causing confusion and eventually identification errors. Since there are presently no large databases (in the order of millions and higher), there is no indication of the results, both in terms of the speed or processing and accuracy.

Another challenge is the fact that over time, the voice of speakers may change due to many different reasons such as illness, stress, aging, etc. One way to handle this problem is to have models which constantly adapt to changes (Beigi 2009).

Yet another problem plagues speaker verification. Neither background models nor cohort models are error-free. Background models generally smooth out many models, and unless the speaker is considerably different from the norm, they may score better than the speaker's own model. This is especially true if one considers the fact that nature is usually Gaussian and that there is a high chance that the speaker's characteristics are close to the smooth background model. If one were to only test the target sample on the target model, this would not be a problem. But since a test sample which is different from the target sample (used for creating the model) is used, the intra-speaker variability might be larger than the inter-speaker variability between the test speech and the smooth background model.

There are, of course, many other open problems. Some of these problems have to do with acceptable noise levels until break-down occurs. Using a cellular telephone with its inherently band-limited characteristics in a very noisy venue such as a subway (metro) station is one such challenging problem.

Given the number of different operating conditions in invoking speaker recognition, it is quite difficult for technology vendors to provide objective performance results. Results are usually quite data-dependent, and different data sets may pronounce particular merits and downfalls of each provider's algorithms and implementation. A good speaker verification system may easily achieve an 0% EER for clean data with good inter-speaker variability in contrast with intra-speaker variability. It is quite normal for the same "good" system to show very high equal error rates under severe conditions such as high noise levels, bandwidth limitation, and small relative inter-speaker variability compared to intra-speaker variability. However, under most controlled conditions, equal error rates below 5% are readily achieved. Similar variability in performance exists in other branches of speaker recognition, such as identification, etc.

Cross-References

▶ Speaker Adaptation
▶ Speaker Classification
▶ Speaker Detection
▶ Speaker Diarization
▶ Speaker Enrollment
▶ Speaker Identification

S

- ▶ Speaker Segmentation
- ▶ Speaker Tracking
- ▶ Speaker Verification
- ▶ Speech Biometrics
- ▶ Speech Recognition

References

Ajmera J, McCowan H, Bourlard I (2004) Robust speaker change detection. IEEE Signal Process Lett 11(8):649–651

Auckenthaler R, Carey M, Lloyd-Thomas H (2000) Score normalization for text-independent speaker verification systems. Digit Signal Process 10(1–3):42–54

Beigi H (2009) Effects of time lapse on speaker recognition results. In: 16th Internation Conference on Digital Signal Processing, July, pp 1–6

Beigi H (2011) Fundamentals of speaker recognition. Springer, New York. ISBN:978-0-387-77591-3

Beigi H, Maes SS (1998) Speaker, channel and environment change detection. In: Proceedings of the World Congress on Automation (WAC'98), May

Beigi H, Maes SH, Chaudhari UV, Sorensen JS (1999) A hierarchical approach to large-scale speaker recognition. In: EuroSpeech 1999, Sept, vol 5, pp 2203–2206

Benesty J, Sondhi MM, Huang Y (2008) Handbook of speech processing. Springer, New York. ISBN:978-3-540-49125-5

Bogert BP, Healy MJR, Tukey JW (1963) The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In: Rosenblatt M (ed) Time series analysis, chap 15, pp 209–243

Burget L, Pichot O, Cumani S, Glembek O, Matejka P, Brummer N (2011) Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May, pp 4832–4835

Burrus CS, Gopinath RA, Guo H (1997) Introduction to wavelets and wavelet transforms: a primer. Prentice Hall, New York. ISBN:0-134-89600-9

Campbell WM, Sturim DE, Reynolds DA, Solomonoff A (2006) SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06), June, vol 1, pp 14–19

Chau CK, Lai CS, Shi BE (2001) Feature vs. model based vocal tract length normalization for a speech recognition-based interactive toy. In: Active media technology. Lecture notes in computer science. Springer, Berlin/Heidelberg, pp 134–143. ISBN:978-3-540-43035-3

Chen SS, Gopalakrishnan PS (1998) Speaker, environemnt and channel change detection and clustering via the bayesian inromation criterion. In: IBM techical report, T.J. Watson Research Center

de la Torre A, Peinado AM, Segura JC, Perez-Cordoba JL, Benitez MC, Rubio AJ (2005) Histogram equalization of speech representation for robust speech recognition. IEEE Trans Speech Audio Process 13(3):355–366

Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798

Duda RO, Hart PE (1973) Pattern classification and scene analysis. Wiley, New York. ISBN:0-471-22361-1

Dunn RB, Reynolds DA, Quatieri TF (2000) Approaches to speaker detection and tracking in conversational speech. Digit Signal Process 10:92–112

Grashey S, Geibler C (2006) Using a vocal tract length related parameter for speaker recognition. In: Speaker and language recognition workshop. IEEE Odyssey 2006: The, June 2006, pp 1–5

Hardt D, Fellbaum K (1997) Spectral subtraction and rasta-filtering in text-dependent HMM-based speaker verification. In: Acoustics, speech, and signal processing. ICASSP-97. 1997 IEEE International Conference on, vol 2, Apr 1997, pp 867–870

Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. J Acoust Soc Am (JASA) 87(4):1738–1752

Hermansky H (1991) Compensation for the effect of the communication channel in the auditory-like analysis of speech (RASTA-PLP). In: Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH-91), pp 1367–1370

Kenny P (2006) Joint factor analysis of speaker and session varaiability: theory and algorithms. Technical report, CRIM, Jan 2006

Kim NS (2002) Feature domain compensation of nonstationary noise for robust speech recognition. Speech Commun 37(3–4):59–73

Manning CD (1999) Foundations of statistical natural language processing. The MIT Press, Boston. ISBN:0-26-213360-1

Martin A, Przybocki M (2000) The nist 1999 speaker recognition evaluation – an overview. Digit Signal Process 10:1–18

Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M (1997) The det curve in assessment of detection task performance. In: Eurospeech 1997, pp 1–8

Nolan F (1983) The phonetic bases of speaker recognition. Cambridge University Press, New York. ISBN:0-521-24486-2

Pelecanos J, Sridharan S (2001) Feature warping for robust speaker verification. In: A speaker odyssey – the speaker recognition workshop, June, pp 213–218

Pollack I, Pickett JM, Sumby WH (1954) On the identification of speakers by voice. J Acoust Soc Am 26:403–406

Poritz AB (1988) Hidden Markov models: a guided tour. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP'88), vol 1, pp 7–13

Prince SJD, Elder JH (2007) Probabilistic linear discriminant analysis for inference about identity. In: IEEE International Conference on Computer Vision (ICCV), Oct, pp 1–8

Rabiner L, Juang B-H (1990) Fundamentals of speech recognition. Prentice Hall signal processing series. PTR Prentice Hall, Englewood Cliffs. ISBN:0-13-015157-2

Reynolds DA (2003) Channel robust speaker verification via feature mapping. In: Acoustics, speech, and signal processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on, Apr, vol 2, pp II–53–6

Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian miscture models. Digit Signal Process 10:19–41

Shearme JN, Holmes JN (1959) An experiment concerning the recognition of voices. Lang Speech 2:123–131

Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors: robust DNN embeddings for speaker recognition. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 15–20 Apr

Teunen R, Shahshahani B, Heck L (2000) A model-based transformational approach to robust speaker recognition. In: International Conference on Spoken Language Processing, vol 2, pp 495–498

Tosi OI (1979) Voice identification: theory and legal applications. University Park Press, Baltimore. ISBN:978-0-839-11294-5

van Vuuren S (1996) Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch. In: International Conference on Spoken Language Processing (ICSLP), Oct, pp 784–787

Vapnik VN (1998) Statistical learning theory. Wiley, New York. ISBN:0-47-103003-1