
Speaker Recognition: Advancements and Challenges

Homayoon Beigi

Additional information is available at the end of the chapter

<http://dx.doi.org/10.5772/52023>

1. Introduction

Speaker Recognition is a multi-disciplinary 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*. Recently, a comprehensive book on all aspects of speaker recognition was published [1]. Therefore, here we are not concerned with details of the standard modeling which is and has been used for the recognition task. In contrast, we present a review of the most recent literature and briefly visit the latest techniques which are being deployed in the various branches of this technology.

Most of the works being reviewed here have been published in the last two years. Some of the topics, such as alternative features and modeling techniques, are general and apply to all branches of speaker recognition. Some of these general techniques, such as whispered speech, are related to the advanced treatment of special forms of audio which have not received ample attention in the past. Finally, we will follow by a look at advancements which apply to specific branches of speaker recognition [1], such as verification, identification, classification, and diarization.

This chapter is meant to complement the summary of speaker recognition, presented in [2], which provided an overview of the subject. It is also intended as an update on the methods described in [1]. In the next section, for the sake of completeness, a brief history of speaker recognition is presented, followed by sections on specific progress as stated above, for globally applicable treatment and methods, as well as techniques which are related to specific branches of speaker recognition.

2. A brief history

The topic of speaker recognition [1] has been under development since the mid-twentieth century. The earliest known papers on the subject, published in the 1950s [3, 4], were in search of finding personal traits of the speakers, by analyzing their speech, with some statistical underpinning. With the advent of early communication networks, Pollack, *et al.* [3] noted the need for speaker identification. Although, they employed human listeners to do the identification of individuals and studied the importance of the duration of speech and other facets that help in the recognition of a speaker. In most of the early

activities, a text-dependent analysis was made, in order to simplify the task of identification. In 1959, not long after Pollack's analysis, *Shearme, et al.* [4] started comparing the formants of speech, in order to facilitate the identification process. However, still a human expert would do the analysis. This first incarnation of speaker recognition, namely using human expertise, has been used to date, in order to handle forensic speaker identification [5, 6]. This class of approaches have been improved and used in a variety of criminal and forensic analyses by legal experts. [7, 8]

Although it is always important to have a human expert available for important cases, such as those in forensic applications, the need for an automatic approach to speaker recognition was soon established. *Prunzansky, et al.* [9, 10] started by looking at an automatic statistical comparison of speakers using a text-dependent approach. This was done by analyzing a population of 10 speakers uttering several unique words. However, it is well understood that, at least for speaker identification, having a text-dependent analysis is not practical in the least [1]. Nevertheless, there are cases where there is some merit to having a text-dependent analysis done for the speaker verification problem. This is usually when there is limited computation resource and/or obtaining speech samples for longer than a couple of seconds is not feasible.

To date, still the most prevalent modeling techniques are the Gaussian mixture model (GMM) and support vector machine (SVM) approaches. Neural networks and other types of classifiers have also been used, although not in significant numbers. In the next two sections, we will briefly recap GMM and SVM approaches. See *Beigi* [1] for a detailed treatment of these and other classifiers.

2.1. Gaussian Mixture Model (GMM) recognizers

In a GMM recognition engine, the models are the parameters for collections of multi-variate normal density functions which describe the distribution of the features [1] for speakers' enrollment data. The best results have been shown on many occasions, and by many research projects, to have come from the use of Mel-Frequency Cepstral Coefficient (MFCC) features [1]. Although, later we will review other features which may perform better for certain special cases.

The *Gaussian mixture model (GMM)* is a model that expresses the probability density function of a random variable in terms of a weighted sum of its components, each of which is described by a *Gaussian (normal)* density function. In other words,

$$p(\mathbf{x}|\boldsymbol{\varphi}) = \sum_{\gamma=1}^{\Gamma} p(\mathbf{x}|\boldsymbol{\theta}_{\gamma})P(\boldsymbol{\theta}_{\gamma}) \quad (1)$$

where the supervector of parameters, $\boldsymbol{\varphi}$, is defined as an augmented set of Γ vectors constituting the free parameters associated with the Γ mixture components, $\boldsymbol{\theta}_{\gamma}, \gamma \in \{1, 2, \dots, \Gamma\}$ and the $\Gamma - 1$ mixture weights, $P(\boldsymbol{\theta} = \boldsymbol{\theta}_{\gamma}), \gamma = \{1, 2, \dots, \Gamma - 1\}$, which are the prior probabilities of each of these mixture models known as the *mixing distribution* [11].

The parameter vectors associated with each mixture component, in the case of the Gaussian mixture model, are the parameters of the normal density function,

$$\boldsymbol{\theta}_{\gamma} = \begin{bmatrix} \boldsymbol{\mu}_{\gamma}^T & \mathbf{u}^T(\boldsymbol{\Sigma}_{\gamma}) \end{bmatrix}^T \quad (2)$$

where the *unique parameters* vector is an invertible transformation that stacks all the free parameters of a matrix into vector form. For example, if $\boldsymbol{\Sigma}_{\gamma}$ is a full covariance matrix, then $\mathbf{u}(\boldsymbol{\Sigma}_{\gamma})$ is the vector of

the elements in the upper triangle of $\mathbf{\Sigma}_\gamma$ including the diagonal elements. On the other hand, if $\mathbf{\Sigma}_\gamma$ is a diagonal matrix, then,

$$(\mathbf{u}(\mathbf{\Sigma}_\gamma))_d \triangleq (\mathbf{\Sigma}_\gamma)_{dd} \quad \forall d \in \{1, 2, \dots, D\} \quad (3)$$

Therefore, we may always reconstruct $\mathbf{\Sigma}_\gamma$ from \mathbf{u}_γ using the inverse transformation,

$$\mathbf{\Sigma}_\gamma = \mathbf{u}_\gamma^{-1} \quad (4)$$

The parameter vector for the mixture model may be constructed as follows,

$$\boldsymbol{\phi} \triangleq \left[\boldsymbol{\mu}_1^T \cdots \boldsymbol{\mu}_\Gamma^T \quad \mathbf{u}_1^T \cdots \mathbf{u}_\Gamma^T \quad p(\boldsymbol{\theta}_1) \cdots p(\boldsymbol{\theta}_{\Gamma-1}) \right]^T \quad (5)$$

where only $(\Gamma - 1)$ mixture coefficients (prior probabilities), $p(\boldsymbol{\theta}_\gamma)$, are included in $\boldsymbol{\phi}$, due to the constraint that

$$\sum_{\gamma=1}^{\Gamma} p(\boldsymbol{\phi}_\gamma) = 1 \quad (6)$$

Thus the number of free parameters in the prior probabilities is only $\Gamma - 1$.

For a sequence of *independent and identically distributed (i.i.d.)* observations, $\{\mathbf{x}\}_1^N$, the log of likelihood of the sequence may be written as follows,

$$\begin{aligned} \ell(\boldsymbol{\phi}|\{\mathbf{x}\}_1^N) &= \ln \left(\prod_{n=1}^N p(\mathbf{x}_n|\boldsymbol{\phi}) \right) \\ &= \sum_{n=1}^N \ln p(\mathbf{x}_n|\boldsymbol{\phi}) \end{aligned} \quad (7)$$

Assuming the mixture model, defined by Equation 1, the likelihood of the sequence, $\{\mathbf{x}\}_1^N$, may be written in terms of the mixture components,

$$\ell(\boldsymbol{\phi}|\{\mathbf{x}\}_1^N) = \sum_{n=1}^N \ln \left(\sum_{\gamma=1}^{\Gamma} p(\mathbf{x}_n|\boldsymbol{\theta}_\gamma) P(\boldsymbol{\theta}_\gamma) \right) \quad (8)$$

Since maximizing Equation 8 requires the maximization of the logarithm of a sum, we can utilize the incomplete data approach that is used in the development of the *EM algorithm* to simplify the solution. *Beigi* [1] shows the derivation of the incomplete data equivalent of the maximization of Equation 8 using the *EM algorithm*.

Each multivariate distribution is represented by Equation 9.

$$p(\mathbf{x}|\boldsymbol{\theta}_\gamma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{\Sigma}_\gamma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_\gamma)^T \mathbf{\Sigma}_\gamma^{-1} (\mathbf{x} - \boldsymbol{\mu}_\gamma) \right\} \quad (9)$$

where $\mathbf{x}, \boldsymbol{\mu}_\gamma \in \mathcal{R}^D$ and $\boldsymbol{\Sigma}_\gamma: \mathcal{R}^D \mapsto \mathcal{R}^D$.

In Equation 9, $\boldsymbol{\mu}_\gamma$ is the mean vector for cluster γ computed from the vectors in that cluster, where,

$$\boldsymbol{\mu}_\gamma \triangleq \mathcal{E}\{\mathbf{x}\} \triangleq \int_{-\infty}^{\infty} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \quad (10)$$

The *sample mean* approximation for Equation 10 is,

$$\boldsymbol{\mu}_\gamma \approx \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \quad (11)$$

where N is the number of samples and \mathbf{x}_i are the MFCC [1].

The *Covariance* matrix is defined as,

$$\boldsymbol{\Sigma}_\gamma \triangleq \mathcal{E}\left\{(\mathbf{x} - \mathcal{E}\{\mathbf{x}\})(\mathbf{x} - \mathcal{E}\{\mathbf{x}\})^T\right\} = \mathcal{E}\{\mathbf{x}\mathbf{x}^T\} - \boldsymbol{\mu}_\gamma \boldsymbol{\mu}_\gamma^T \quad (12)$$

The diagonal elements of $\boldsymbol{\Sigma}_\gamma$ are the variances of the individual dimensions of \mathbf{x} . The off-diagonal elements are the covariances across the different dimensions.

The *unbiased estimate* of $\boldsymbol{\Sigma}_\gamma$, $\tilde{\boldsymbol{\Sigma}}_\gamma$, is given by the following,

$$\tilde{\boldsymbol{\Sigma}}_\gamma = \frac{1}{N-1} \left[\mathbf{S}_{\gamma|N} - N(\boldsymbol{\mu}_\gamma \boldsymbol{\mu}_\gamma^T) \right] \quad (13)$$

where the *sample mean*, $\boldsymbol{\mu}_\gamma$, is given by Equation 11 and the *second order sum matrix (Scatter Matrix)*, $\mathbf{S}_{\gamma|N}$, is given by,

$$\mathbf{S}_{\gamma|N} \triangleq \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T \quad (14)$$

Therefore, in a general GMM model, the above statistical parameters are computed and stored for the set of Gaussians along with the corresponding mixture coefficients, to represent each speaker. The features used by the recognizer are *Mel-Frequency Cepstral Coefficients (MFCC)*. *Beigi* [1] describes details of such a **GMM**-based recognizer.

2.2. Support Vector Machine (SVM) recognizers

In general, *SVM* are formulated as *two-class* classifiers. Γ -class classification problems are usually reduced to Γ two-class problems [12], where the γ^h two-class problem compares the γ^h class with the rest of the classes combined. There are also other generalizations of the SVM formulation which are geared toward handling Γ -class problems directly. *Vapnik* has proposed such formulations in Section 10.10 of his book [12]. He also credits *M. Jaakkola* and *C. Watkins, et al.* for having proposed similar generalizations independently. For such generalizations, the constrained optimization problem becomes much more complex. For this reason, the approximation using a set of Γ two-class problems has been

preferred in the literature. It has the characteristic that if a data point is accepted by the decision function of more than one class, then it is deemed as *not classified*. Furthermore, it is not classified if no decision function claims that data point to be in its class. This characteristic has both positive and negative connotations. It allows for better rejection of outliers, but then it may also be viewed as giving up on handling outliers.

In application to speaker recognition, experimental results have shown that *SVM* implementations of speaker recognition may perform similarly or sometimes even be slightly inferior to the less complex and less resource intensive *GMM* approaches. However, it has also been noted that systems which combine *GMM* and *SVM* approaches often enjoy a higher accuracy, suggesting that part of the information revealed by the two approaches may be complementary [13].

The problem of *overtraining* (*overfitting*) plagues many learning techniques, and it has been one of the driving factors for the development of support vector machines [1]. In the process of developing the concept of *capacity* and eventually *SVM*, *Vapnik* considered the generalization capacity of learning machines, especially *neural networks*. The main goal of support vector machines is to maximize the generalization capability of the learning algorithm, while keeping good performance on the training patterns. This is the basis for the *Vapnik-Chervonenkis theory* (*VC theory*) [12], which computes bounds on the risk, $R(o)$, according to the definition of the *VC dimension* and the *empirical risk* – see *Beigi* [1].

The multiclass classification problem is also quite important, since it is the basis for the speaker identification problem. In Section 10.10 of his book, *Vapnik* [12] proposed a simple approach where one class was compared to all other classes and then this is done for each class. This approach converts a Γ -class problem to Γ two-class problems. This is the most popular approach for handling multi-class *SVM* and has been dubbed the *one-against-all*¹ approach [1]. There is also, the *one-against-one* approach which transforms the problem into $\Gamma(\Gamma + 1)/2$ two-class *SVM* problems. In Section 6.2.1 we will see more recent techniques for handling multi-class *SVM*.

3. Challenging audio

One of the most important challenges in speaker recognition stems from inconsistencies in the different types of audio and their quality. One such problem, which has been the focus of most research and publications in the field, is the problem of channel mismatch, in which the enrollment audio has been gathered using one apparatus and the test audio has been produced by a different channel. It is important to note that the sources of mismatch vary and are generally quite complicated. They could be any combination and usually are not limited to mismatch in the handset or recording apparatus, the network capacity and quality, noise conditions, illness related conditions, stress related conditions, transition between different media, etc. Some approaches involve normalization of some kind to either transform the data (raw or in the feature space) or to transform the model parameters. Chapter 18 of *Beigi* [1] discusses many different channel compensation techniques in order to resolve this issue. *Vogt, et al.* [14] provide a good coverage of methods for handling modeling mismatch.

One such problem is to obtain ample coverage for the different types of phonation in the training and enrollment phases, in order to have a better performance for situations when different phonation types are uttered. An example is the handling of whispered phonation which is, in general, very hard to collect and is not available under natural speech scenarios. Whisper is normally used by individuals who desire to have more privacy. This may happen under normal circumstances when the user is on a telephone and does not want others to either hear his/her conversation or does not wish to bother others in the

¹ Also known as one-against-rest.

vicinity, while interacting with the speaker recognition system. In Section 3.1, we will briefly review the different styles of phonation. Section 3.2 will then cover some work which has been done, in order to be able to handle whispered speech.

Another challenging issue with audio is to handle multiple speakers with possibly overlapping speech. The most difficult scenario would be the presence of multiple speakers on a single microphone, say a telephone handset, where each speaker is producing similar level of audio at the same time. This type of cross-talk is very hard to handle and indeed it is very difficult to identify the different speakers while they speak simultaneously. A somewhat simpler scenario is the one which generally happens in a conference setting, in a room, in which case, a far-field microphone (or microphone array) is capturing the audio. When multiple speakers speak in such a setting, there are some solutions which have worked out well in reducing the interference of other speakers, when focusing on the speech of a certain individual. In Section 3.4, we will review some work that has been done in this field.

3.1. Different styles of phonation

Phonation deals with the acoustic energy generated by the vocal folds at the larynx. The different kinds of phonation are *unvoiced*, *voiced*, and *whisper*.

Unvoiced phonation may be either in the form of *nil phonation* which corresponds to zero energy or *breath phonation* which is based on relaxed vocal folds passing a turbulent air stream.

Majority of voiced sounds are generated through *normal voiced phonation* which happens when the vocal folds are vibrating at a periodic rate and generate certain resonance in the upper chamber of the vocal tract. Another category of voiced phonation is called *laryngealization (creaky voice)*. It is when the arytenoid cartilages fix the posterior portion of the vocal folds, only allowing the anterior part of the vocal folds to vibrate. Yet another type voiced phonation is a falsetto which is basically the un-natural creation of a high pitched voice by tightening the basic shape of the vocal folds to achieve a false high pitch.

In another view, the emotional condition of the speaker may affect his/her phonation. For example, speech under stress may manifest different phonetic qualities than that of, so-called, *neutral speech* [15]. Whispered speech also changes the general condition of phonation. It is thought that this does not affect unvoiced consonants as much. In Sections 3.2 and 3.3 we will briefly look at whispered speech and speech under stressful conditions.

3.2. Treatment of whispered speech

Whispered phonation happens when the speaker acts like generating a voiced phonation with the exception that the vocal folds are made more relaxed so that a greater flow of air can pass through them, generating more of a turbulent airstream compared to a voiced resonance. However, the vocal folds are not relaxed enough to generate an unvoiced phonation.

As early as the first known paper on speaker identification [3], the challenges of whispered speech were apparent. The general text-independent analysis of speaker characteristics relies mainly on the *normal voiced phonation* as the primary source of speaker-dependent information.[1] This is due to the high-energy periodic signal which is generated with rich resonance information. Normally, very little natural whisper data is available for training. However, in some languages, such as *Amerindian*

languages² (e.g., *Comanche* [16] and *Tlingit* – spoken in Alaska) and some old languages, voiceless vocoids exist and carry independent meaning from their voiced counterparts [1].

An example of a whispered phone in English is the *egressive pulmonic whisper* [1] which is the sound that an [h] makes in the word, “home.” However, any utterance may be produced by relaxing the vocal folds and generating a whispered version of the utterance. This partial relaxation of the vocal folds can significantly change the vocal characteristics of the speaker. Without ample data in whisper mode, it would be hard to identify the speaker.

Pollack, *et al.* [3] say that we need about three times as much speech samples for whispered speech in order to obtain an equivalent accuracy to that of normal speech. This assessment was made according to a comparison, done using human listeners and identical speech content, as well as an attempted equivalence in the recording volume levels.

Jin, *et al.* [17] deal with the insufficient amount of whisper data by creating two GMM models for each individual, assuming that ample data is available for the normal-speech mode for any target speaker. Then, in the test phase, they use the *frame-based score competition (FSC)* method, comparing each frame of audio to the two models for every speaker (normal and whispered) and only using the result for that frame, from the model which produces the higher score. Otherwise, they continue with the standard process of recognition.

Jin, *et al.* [17] conducted experiments on whispered speech when almost no whisper data was available for the enrollment phase. The experiments showed that noise greatly impacts recognition with whispered speech. Also, they concentrate on using a throat microphone which happens to be more robust in terms of noise, but it also picks up more resonance for whispered speech. In general, using the two-model approach with FSC, [17] show significant reduction in the error rate.

Fan, *et al.* [18] have looked into the differences between whisper and neutral speech. By neutral speech, they mean normal speech which is recorded in a modal (voiced) speech setting in a quiet recording studio. They use the fact that the unvoiced consonants are quite similar in the two types of speech and that most of the differences stem from the remaining phones. Using this, they separate whispered speech into two parts. The first part includes all the unvoiced consonants, and the second part includes the rest of the phones. Furthermore, they show better performance for unvoiced consonants in the whispered speech, when using *linear frequency cepstral coefficients (LFCC)* and *exponential frequency cepstral coefficients (EFCC)* – see Section 4.3. In contrast, the rest of the phones show better performance with MFCC features. Therefore, they detect *unvoiced consonants* and treat them using LFCC/EFCC features. They send the rest of the phones (e.g., voiced consonants, vowels, diphthongs, triphthongs, glides, liquids) through an MFCC-based system. Then they combine the scores from the two segments to make a speaker recognition decision.

The unvoiced consonant detection which is proposed by [18], uses two measures for determining the frames stemming from unvoiced consonants. For each frame, l , the energy of the frame in the lower part of the spectrum, $E_l^{(l)}$, and that of the higher part of the band, $E_l^{(h)}$, (for $f \leq 4000\text{Hz}$ and $4000\text{Hz} < f \leq 8000\text{Hz}$ respectively) are computed, along with the total energy of the frame, E_l , to be used for normalization. The relative energy of the lower frequency is then computed for each frame by Equation 15.

$$R_l = \frac{E_l^{(l)}}{E_l} \quad (15)$$

² Languages spoken by native inhabitants of the Americas.

It is assumed that most of spectral energy of unvoiced consonants is concentrated in the higher half of the frequency spectrum, compared to the rest of the phones. In addition, the Jeffreys' divergence [1] of the higher portion of the spectrum relative to the previous frame is computed using Equation 16.

$$\mathcal{D}_J(l \leftrightarrow l-1) = -P_{l-1}^{(h)} \log_2(P_l^{(h)}) - P_l^{(h)} \log_2(P_{l-1}^{(h)}) \quad (16)$$

where

$$P_l^{(h)} \triangleq \frac{E_l^{(h)}}{E_l} \quad (17)$$

Two separate thresholds may be set for R_l and $\mathcal{D}_J(l \leftrightarrow l-1)$, in order to detect unvoiced consonants from the rest of the phones.

3.3. Speech under stress

As noted earlier, the phonation undergoes certain changes when the speaker is under stressful conditions. *Bou-Ghazale, et al.* [15] have shown that this may effect the significance of certain frequency bands, making MFCC features miss certain nuances in the speech of the individual under stress. They propose a new frequency scale which it calls the *exponential-logarithmic (expo-log)* scale. In Section 4.3 we will describe this scale in more detail since it is also used by *Bou-Ghazale, et al.* [18] to handle the unvoiced consonants. On another note, although research has generally shown that cepstral coefficients derived from FFT are more robust for the handling of neutral speech [19], *Bou-Ghazale, et al.* [15] suggest that for speech, recorded under stressful conditions, cepstral coefficients derived from the linear predictive model [1] perform better.

3.4. Multiple sources of speech and far-field audio capture

This problem has been addressed in the presence of microphone arrays, to handle cases when sources are semi-stationary in a room, say in a conference environment. The main goal would amount to extracting the source(s) of interest from a set of many sources of audio and to reduce the interference from other sources in the process [20]. For instance, *Kumatani, et al.* [21] address the problem using the, so called, beamforming technique[20, 22] for two speakers speaking simultaneously in a room. They construct a generalized sidelobe canceler (GSC) for each source and adjusts the active weight vectors of the two GSCs to extract two speech signals with *minimum mutual information* [1] between the two. Of course, this makes a few essential assumptions which may not be true in most situations. The first assumption is that the number of speakers is known. The second assumption is that they are semi-stationary and sitting in different angles from the microphone array. *Kumatani, et al.* [21] show performance results on the far-field PASCAL speech separation challenge, by performing speech recognition trials.

One important part of the above task is to localize the speakers. *Takashima, et al.* [23] use an HMM-based approach to separate the acoustic transfer function so that they can separate the sources, using a single microphone. It is done by using an HMM model of the speech of each speaker to estimate the acoustic transfer function from each position in the room. They have experimented with up to 9 different source positions and have shown that their accuracy of localization decreases with increasing number of positions.

3.5. Channel mismatch

Many publications deal with the problem of channel mismatch, since it is the most important challenge in speaker recognition. Early approaches to the treatment of this problem concentrated on normalization of the features or the score. *Vogt, et al.* [14] present a good coverage of different normalization techniques. *Barras, et al.* [24] compare cepstral mean subtraction (CMS) and variance normalization, Feature Warping, T-Norm, Z-Norm and the cohort methods. Later approaches started by using techniques from factor analysis or discriminant analysis to transform features such that they convey the most information about speaker differences and least about channel differences. Most GMM techniques use some variation of *joint factor analysis (JFA)* [25]. An offshoot of JFA is the i-vector technique which does away with the channel part of the model and falls back toward a PCA approach [26]. See Section 5.1 for more on the i-vector approach.

SVM systems use techniques such as *nuisance attribute projection (NAP)* [27]. NAP [13] modifies the original *kernel*, used for a *support vector machine (SVM)* formulation, to one with the ability of telling specific channel information apart. The premise behind this approach is that by doing so, in both training and recognition stages, the system will not have the ability to distinguish channel specific information. This channel specific information is what is dubbed nuisance by *Solomonoff, et al.* [13]. NAP is a projection technique which assumes that most of the information related to the channel is stored in specific low-dimensional subspaces of the higher dimensional space to which the original features are mapped. Furthermore, these regions are assumed to be somewhat distinct from the regions which carry speaker information. This is quite similar to the idea of *joint factor analysis*. *Seo, et al.* [28] use the statistics of the eigenvalues of background speakers to come up with discriminative weight for each background speaker and to decide on the between class scatter matrix and the within-class scatter matrix.

Shanmugapriya, et al. [29] propose a *fuzzy wavelet network (FWN)* which is a neural network with a wavelet activation function (known as a *Wavener*). A fuzzy neural network is used in this case, with the wavelet activation function. Unfortunately, [29] only provides results for the TIMIT database [1] which is a database acquired under a clean and controlled environment and is not very challenging.

Villalba, et al. [30] attempt to detect two types of low-tech spoofing attempts. The first one is the use of a far-field microphone to record the victim's speech and then to play it back into a telephone handset. The second type is the concatenation of segments of short recordings to build the input required for a text-dependent speaker verification system. The former is handled by using an SVM classifier for spoof and non-spoof segments trained based on some training data. The latter is detected by comparing the pitch and MFCC feature contours of the enrollment and test segments using dynamic time warping (DTW).

4. Alternative features

As seen in the past, most classic features used in speech and speaker recognition are based on LPC, LPCC, or MFCC. In Section 6.3 we see that *Dhanalakshmi, et al.* [19] report trying these three classic features and have shown that MFCC outperforms the other two. Also, *Beigi* [1] discusses many other features such as those generated by *wavelet filterbanks*, *instantaneous frequencies*, *EMD*, etc. In this section, we will discuss several new features, some of which are variations of cepstral coefficients with a different frequency scaling, such as *CFCC*, *LFCC*, *EFCC*, and *GFCC*. In Section 6.2 we will also see the *RMFCC* which was used to handle speaker identification for gaming applications. Other features

are also discussed, which are more fundamentally different, such as *missing feature theory (MFT)*, and *local binary features*.

4.1. Multitaper MFCC features

Standard MFCC features are usually computed using a periodogram estimate of the spectrum, with a window function, such as the Hamming window. [1] MFCC features computed by this method portray a large variance. To reduce the variance, multitaper spectrum estimation techniques [31] have been used. They show lower bias and variance for the multitaper estimate of the spectrum. Although bias terms are generally small with the windowed periodogram estimate, the reduction in the variance, using multitaper estimation, seems to be significant.

A multitaper estimate of a spectrum is made by using the mean value of periodogram estimates of the spectrum using a set of orthogonal windows (known as tapers). The multitaper approach has been around since early 1980s. Examples of such taper estimates are *Thomson* [32], *Tukey's split cosine taper* [33], *sinusoidal taper* [34], and *peak matched estimates* [35]. However, their use in computing MFCC features seems to be new. In Section 5.1, we will see that they have been recently used in accordance with the i-vector formulation and have also shown promising results.

4.2. Cochlear Filter Cepstral Coefficients (CFCC)

Li, et al. [36] present results for speaker identification using *cochlear filter cepstral coefficients (CFCC)* based on an auditory transform [37] while trying to emulate natural cochlear signal processing. They maintain that the CFCC features outperform MFCC, PLP, and RASTA-PLP features [1] under conditions with very low signal to noise ratios. Figure 1 shows the block diagram of the CFCC feature extraction proposed by *Li, et al.* [36]. The *auditory transform* is a *wavelet transform* which was proposed by *Li, et al.* [37]. It may be implemented in the form of a filter bank, as it is usually done for the extraction of MFCC features [1]. Equations 18 and 19 show a generic wavelet transform associated with one such filter.

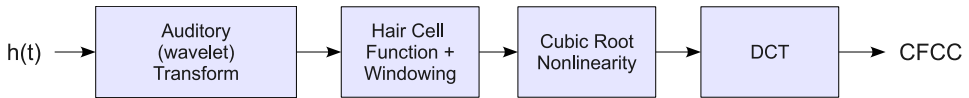


Figure 1. Block Diagram of Cochlear Filter Cepstral Coefficient (CFCC) Feature Extraction – proposed by *Li, et al.* [36]

$$T(a, b) = \int_{-\infty}^{\infty} h(t) \psi_{(a,b)}(t) dt \quad (18)$$

where

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) \quad (19)$$

The *wavelet basis functions* [1], $\{\psi_{(a,b)}(t)\}$, are defined by *Li, et al.* [37], based on the *mother wavelet*, $\psi(t)$ (Equation 20), which mimics the cochlear impulse response function.

$$\psi(t) \triangleq t^{\alpha} \exp[-2\pi h_L \beta t] \cos[2\pi h_L t + \theta] \quad (20)$$

Each wavelet basis function, according to the scaling and translation parameters $a > 0$ and $b > 0$ is, therefore, given by Equation 21.

$$\psi_{(a,b)}(t) = \frac{1}{\sqrt{|a|}} \left(\frac{t-b}{a} \right)^\alpha \exp \left[-2\pi h_L \beta \left(\frac{t-b}{a} \right) \right] \cos \left[2\pi h_L \left(\frac{t-b}{a} \right) + \theta \right] \quad (21)$$

In Equation 21, α and β are strictly positive parameters which define the shape and the bandwidth of the cochlear filter in the frequency domain. *Li, et al.* [36] determine them empirically for each filter in the filter bank. $u(t)$ is the units step (Heaviside) function defined by Equation 22.

$$u(t) \triangleq \begin{cases} 1 & \forall t \geq 0 \\ 0 & \forall t < 0 \end{cases} \quad (22)$$

4.3. Linear and Exponential Frequency Cepstral Coefficients (LFCC and EFCC)

Some experiments have shown that using *linear frequency cepstral coefficients* (LFCC) and *exponential frequency cepstral coefficients* (EFCC) for processing unvoiced consonants may produce better results for speaker recognition. For instance, *Fan, et al.* [18] use an unvoiced consonant detector to separate frames which contain such phones and to use LFCC and EFCC features for these frames (see Section 3.2). These features are then used to train up a GMM-based speaker recognition system. In turn, they send the remaining frames to a GMM-based recognizer using MFCC features. The two recognizers are treated as separate systems. At the recognition stage, the same segregation of frames is used and the scores of two recognition engines are combined to reach the final decision.

The EFCC scale was proposed by *Bou-Ghazale, et al.* [15] and later used by *Fan, et al.* [18]. This mapping is given by

$$E = (10^{\frac{f}{k}} - 1)c \quad \forall 0 \leq f \leq 8000Hz \quad (23)$$

where the two constants, c and k , are computed by solving Equations 24 and 25.

$$(10^{\frac{8000}{k}} - 1)c = 2595 \log \left(1 + \frac{8000}{700} \right) \quad (24)$$

$$\{c, k\} = \min \left\{ \left| \left(10^{\frac{4000}{k}} - 1 \right) - \frac{4000}{k^2} c \times 10^{\frac{4000}{k}} \ln(10) \right| \right\} \quad (25)$$

Equation 24 comes from the requirement that the exponential and Mel scale functions should be equal at the Nyquist frequency and Equation 24 is the result of minimizing the absolute values of the partial derivatives of E in Equation 23 with respect to c and k for $f = 4000Hz$ [18]. The resulting c and k which would satisfy Equations 24 and 25 are computed by *Fan, et al.* [18] to be $c = 6375$ and $k = 50000$. Therefore, the exponential scale function is given by Equation 26.

$$E = 6375 \times (10^{\frac{f}{50000}} - 1) \quad (26)$$

Fan et al. [18] show better accuracy for unvoiced consonants, when EFCC is used over MFCC. However, it shows even better accuracy when LFCC is used for these frames!

4.4. Gammatone Frequency Cepstral Coefficients (GFCC)

Shao, et al. [38] use *gammatone frequency cepstral coefficients (GFCC)* as features, which are the products of a cochlear filter bank, based on psychophysical observations of the total auditory system. The Gammatone filter bank proposed by *Shao, et al.* [38] has 128 filters, centered from 50Hz to 8kHz, at equal partitions on the *equivalent rectangular bandwidth (ERB)* [39, 40] scale (Equation 28)³.

$$E_c = \frac{1000}{(24.7 \times 4.37)} \ln(4.37 \times 10^3 f + 1) \quad (27)$$

$$= 21.4 \log(4.37 \times 10^3 f + 1) \quad (28)$$

where f is the frequency in Hertz and E is the number of ERBs, in a similar fashion as Barks or Mels are defined [1]. The bandwidth, E_b , associated with each center frequency, f , is then given by Equation 29. Both f and E_b are in *Hertz (Hz)* [40].

$$E_b = 24.7(4.37 \times 10^3 f + 1) \quad (29)$$

The impulse response of each filter is given by Equation 30.

$$g(f, t) \triangleq \begin{cases} t^{(a-1)} e^{-2\pi b t} \cos(2\pi f t) & t \geq 0 \\ 0 & \text{Otherwise} \end{cases} \quad (30)$$

where t denotes the time and f is the center frequency of the filter of interest. a is the order of the filter and is taken to be $a = 4$ [38], and b is the filter bandwidth.

In addition, as it is done with other models such as MFCC, LPCC, and PLP, the magnitude also needs to be warped. *Shao, et al.* [38] base their magnitude warping on the method of cubic root warping (magnitude to loudness conversion) used in PLP [1].

The same group that published [38], followed by using a *computational auditory scene analysis (CASA)* front-end [43] to estimate a binary spectrographical mask to determine the useful part of the signal (see Section 4.5), based on *auditory scene analysis (ASA)* [44]. They claim great improvements in noisy environments, over standard speaker recognition approaches.

4.5. Missing Feature Theory (MFT)

Missing feature theory (MFT) tries to deal with bandlimited speech in the presence of non-stationary background noise. Such missing data techniques have been used in the speech community, mostly to handle applications of noisy speech recognition. *Vizinho, et al.* [45] describe such techniques by

³ The ERB scale is similar to the Bark and Mel scales [1] and is computed by integrating an empirical differential equation proposed by *Moore and Glasberg* in 1983 [39] and then modified by them in 1990 [41]. It uses a set of rectangular filters to approximate human cochlear hearing and provides a more accurate approximation to the psychoacoustical scale (Bark scale) of *Zwicker* [42].

estimating the reliable regions of the spectrogram of speech and then using these reliable portions to perform speech recognition. They do this by estimating the noise spectrum and the SNR and by creating a mask that would remove the noisy part from the spectrogram. In a related approach, some feature selection methods use Bayesian estimation to estimate a spectrographic mask which would remove unwanted part of the spectrogram, therefore removing features which are attributed to the noisy part of the signal.

The goal of these techniques is to be able to handle non-stationary noise. *Seltzer, et al.* [46] propose one such Bayesian technique. This approach concentrates on extracting as much useful information from the noisy speech as it can, rather than trying to estimate the noise and to subtract it from the signal, as it is done by *Vizinho, et al.* [45]. However, there are many parameters which need to be optimized, making the process quite expensive, calling for suboptimal search. *Pullella, et al.* [47] have combined the two techniques of spectrographic mask estimation and dynamic feature selection to improve the accuracy of speaker recognition under noisy conditions. *Lim, et al.* [48] propose an optimal mask estimation and feature selection algorithm.

4.6. Local binary features (slice classifier)

The idea of statistical boosting is not new and was proposed by several researchers, starting with *Schapire* [49] in 1990. The *Adaboost algorithm* was introduced by *Freund, et al.* [50] in 1996 as one specific boosting algorithm. The idea behind statistical boosting is that a combination of weak classifiers may be combined to build a strong one.

Rodriguez [51] used the statistical boosting idea and several extensions of the Adaboost algorithm to introduce face detection and verification algorithms which would use features based on local differences between pixels in a 9×9 pixel grid, compared to the central pixel of the grid.

Inspired by [51], *Roy, et al.* [52] created local binary features according to the differences between the bands of the *discrete Fourier transform (DFT)* values to compare two models. One important claim of this classifier is that it is less prone to overfitting issues and that it performs better than conventional systems under low SNR values. The resulting features are binary because they are based on a threshold which categorizes the difference between different bands of the FFT to either 0 or 1. The classifier of [52] has a built-in discriminant nature, since it uses certain data as those coming from impostors, in contrast with the data which is generated by the target speaker. The labels of impostor versus target allow for this built-in discrimination. The authors of [52] call these features, *boosted binary features (BBF)*. In a more recent paper [53], *Roy, et al.* refined their approach and renamed the method a *slice classifier*. They show similar results with this classifier, compared to the state of the art, but they explain that the method is less computationally intensive and is more suitable for use in mobile devices with limited resources.

5. Alternative speaker modeling

Classic modeling techniques for speaker recognition have used *Gaussian mixture models (GMM)*, *support vector machines (SVM)*, and *neural networks* [1]. In Section 6 we will see some other modeling techniques such as non-negative matrix factorization. Also, in Section 4, new modeling implementations were used in applying the new features presented in the section. Generally, most new modeling techniques use some transformation of the features in order to handle mismatch conditions, such as joint factor analysis (JFA), Nuisance attribute projection (NAP), and principal component

analysis (PCA) techniques such as the i-vector implementation.^[1] In the next few sections, we will briefly look at some recent developments in these and other techniques.

5.1. The i-vector model (total variability space)

Dehak, *et al.* [54] recombined the *channel variability space* in the JFA formulation [25] with the *speaker variability space*, since they discovered that there was considerable leakage from the speaker space into the channel space. The combined space produces a new projection (Equation 31) which resembles a PCA, rather than a factor analysis process.

$$\mathbf{y}_n = \boldsymbol{\mu} + \mathbf{V}\boldsymbol{\theta}_n \quad (31)$$

They called the new space *total variability space* and in their later works [55–57], they referred to the projections of feature vectors into this space, *i-vectors*. *Speaker factor coefficients* are related to the speaker coordinates, in which each speaker is represented as a point. This space is defined by the *Eigenvoice matrix*. These speaker factor vectors are relatively short, having in the order of about 300 elements [58], which makes them desirable for use with *support vector machines*, as the observed vector in the observation space (\mathbf{x}).

Generally, in order to use an i-vector approach, several recording sessions are needed from the same speaker, to be able to compute the within class covariance matrix in order to do within class covariance normalization (WCCN). Also, methods using *linear discriminant analysis (LDA)* along with WCCN [57] and recently, *probabilistic LDA (PLDA)* with WCCN [59–62] have also shown promising results.

Alam, *et al.* [63] examined the use of *multitaper MFCC features* (see Section 4.1) in conjunction with the i-vector formulation. They show improved performance using multitaper MFCC features, compared to standard MFCC features which have been computed using a *Hamming window* [1].

Glembek, *et al.* [26] provide simplifications to the formulation of the *i-vectors* to reduce the memory usage and to increase the speed of computing the vectors. Glembek, *et al.* [26] also explore linear transformations using principal component analysis (PCA) and Heteroscedastic Linear Discriminant Analysis⁴ (HLDA) [64] to achieve orthogonality of the components of the Gaussian mixture.

5.2. Non-negative matrix factorization

In Section 6.3, we will see several implementations of extensions of non-negative matrix factorization [65, 66]. These techniques have been successfully applied to classification problems. More detail is given in Section 6.3.

5.3. Using multiple models

In Section 3.2 we briefly covered a few model combination and selection techniques that would use different specialized models to achieve better recognition rates. For example, Fan, *et al.* [18] used two different models to handle unvoiced consonants and the rest of the phones. Both models had similar form, but they used slightly different types of features (MFCC vs. EFCC/LFCC). Similar ideas will be discussed in this section.

⁴ Also known as Heteroscedastic Discriminant Analysis (HDA) [64]

5.3.1. Frame-based score competition (FSC):

In Section 3.2 we discussed the fact that *Jin, et al.* [17] used two separate models, one based on the normal speech (neutral speech) model and the second one based on whisper data. Then, at the recognition stage, each frame is evaluated against the two models and the higher score is used. [17] Therefore, it is called a frame-based score competition (FSC) method.

5.3.2. SNR-Matched Recognition:

After performing voice activity detection (VAD), *Bartos, et al.* [67] estimate the signal to noise ratio (SNR) of that part of the signal which contains speech. This value is used to load models which have been created with data recorded under similar SNR conditions. Generally, the SNR is computed in *decibels* given by Equations 32 and 33 – see [1] for more.

$$SNR = 10 \log_{10} \left(\frac{\mathcal{P}_s}{\mathcal{P}_n} \right) \quad (32)$$

$$= 20 \log_{10} \left(\frac{|H_s(\omega)|}{|H_n(\omega)|} \right) \quad (33)$$

Bartos, et al. [67] consider an SNR of 30dB or higher to be clean speech. An SNR of 30dB happens to be equivalent to the signal amplitude being about 30 times that of the noise. When the SNR is 0, the signal amplitude is roughly the same as the energy of the noise.

Of course, to evaluate the SNR from Equation 32 or 33, we would need to know the power or amplitude of the noise as well as the true signal. Since this is not possible, estimation techniques are used to come up with an instantaneous SNR and to average that value over the whole signal. *Bartos, et al.* [67] present such an algorithm.

Once the SNR of the speech signal is computed, it is categorized within a quantization of 4dB segments and then identification or verification is done using models which have been enrolled with similar SNR values. This, according to [67], allows for a lower equal error rate in case of speaker verification trials. In order to generate speaker models for different SNR levels (of 4dB steps), [67] degrades clean speech iteratively, using some additive noise, amplified by a constant gain associated with each 4db level of degradation.

6. Branch-specific progress

In this section, we will quickly review the latest developments for the main branches of speaker recognition as listed at the beginning of this chapter. Some of these have already been reviewed in the above sections. Most of the work on speaker recognition is performed on speaker verification. In the next section we will review some such systems.

6.1. Verification

As we mentioned in Section 4, *Roy, et al.* [52, 53] used the so-called boosted binary features (slice classifier) for speaker verification. Also, we reviewed several developments regarding the i-vector

formulation in Section 5.1. The i-vector has basically been used for speaker verification. Many recent papers have dealt with aspects such as LDA, PLDA, and other discriminative aspects of the training.

Salman, et al. [68] use a neural network architecture with very deep number of layers to perform a greedy discriminative learning for the speaker verification problem. The *deep neural architecture (DNA)*, proposed by [68], uses two identical subnets, to process two MFCC feature vectors respectively, for providing discrimination results between two speakers. They show promising results using this network.

Sarkar, et al. [69] use multiple background models associated with different *vocal tract length (VTL)* [1] estimates for the speakers, using MAP [1] to derive these background models from a root background model. Once the best VTL-based background model for the training or test audio is computed, the transformation to get from that universal background model (UBM) to the root UBM is used to transform the features of the segment to those associated with the VTL of the root UBM. *Sarkar, et al.* [69] show that the results of this single UBM system is comparable to a multiple background model system.

6.2. Identification

In Section 5.3.2 we discussed new developments on SNR-matched recognition. The work of *Bartos, et al.* [67] was applied to improving speaker identification based on a matched SNR condition.

Bharathi, et al. [70] try to identify phonetic content for which specific speakers may be efficiently recognized. Using these speaker-specific phonemes, a special text is created to enhance the discrimination capability for the target speaker. The results are presented for the TIMIT database [1] which is a clean and controlled database and not very challenging. However, the idea seems to have merit.

Cai, et al. [71] use some of the features described in Section 4, such as MFCC and GFCC in order to identify the voice of signers from a monophonic recording of songs in the presence of sounds of music from several instruments.

Do, et al. [72] examine the speaker identification problem for identifying the person playing a computer game. The specific challenges are the fact that the recording is done through a far-field microphone (see Section 3.4) and that the audio is generally short, apparently based on the commands used for gaming. To handle the reverberation and background noise, *Do, et al.* [72] argue for the use of the, so-called, *reverse Mel frequency cepstral coefficients (RMFCC)*. They propose this set of features by reversing the triangular filters [1] used for computing the MFCC, such that the lower frequency filters have larger bandwidths and the higher frequency filters have smaller bandwidths. This is exactly the opposite of the filters being used for MFCC. They also use LPC and F_0 (the fundamental frequency) as additional features.

In Section 3.2 we saw the treatment of speaker identification for whispered speech in some detail. Also, *Ghiurcau, et al.* [73] study the emotional state of speakers on the results of speaker identification. The study treats happiness, anger, fear, boredom, sadness, and neutral conditions; it shows that these emotions significantly affect identification results. Therefore, they [73] propose using emotion detection and having emotion-specific models. Once the emotion is identified, the proper model is used to identify the test speaker.

Liu, et al. [74] use the Hilbert Huang Transform to come up with new acoustic features. This is the use of intrinsic mode decomposition described in detail in [1].

In the next section, we will look at the multi-class SVM which is used to perform speaker identification.

6.2.1. Multi-Class SVM

In Section 2.2 we discussed the popular one-against-all technique for handling multi-class SVM. There have been other more recent techniques which have been proposed in the last few years. One such technique is due to *Platt, et al.* [75], who proposed the, so-called, *decision directed acyclic graph (DDAG)* which produces a classification node for each pair of classes, in a Γ -class problem. This leads to $\Gamma(\Gamma - 1)/2$ classifiers and results in the creation of the *DAGSVM* algorithm [75].

Wang [76] presents a tree-based multi-class SVM which reduces the number of matches to the order of $\log(\Gamma)$. Although at the training phase, the number of SVM are similar to that of DDAG, namely, $\Gamma(\Gamma - 1)/2$. This can significantly reduce the amount of computation for speaker identification.

6.3. Classification and diarization

Aside from the more prominent research on speaker verification and identification, audio source and gender classification are also quite important in most audio processing systems including speaker and speech recognition.

In many practical audio processing systems, it is important to determine the type of audio. For instance, consider a telephone-based system which includes a speech recognizer. Such recognition engines would produce spurious results if they were presented with non-speech, say music. These results may be detrimental to the operation of an automated process. This is also true for speaker identification and verification systems which expect to receive human speech. They may be confused if they are presented with music or other types of audio such as noise. For *text-independent speaker identification* systems, this may result in mis-identifying the audio as a viable choice in the database and resulting in dire consequences!

Similarly, some systems are only interested in processing music. An example is a music search system which would look for a specific music or one resembling the presented segment. These systems may be confused, if presented with human speech, uttered inadvertently, while only music is expected.

As an example, an important goal for audio source classification research is to develop filters which would tag a segment of audio as speech, music, noise, or silence [77]. Sometimes, we would also look into classifying the genre of audio or video such as movie, cartoon, news, advertisement, etc. [19].

The basic problem contains two separate parts. The first part is the segmentation of the audio stream into segments of similar content. This work has been under development for the past few decades with some good results [78–80].

The second part is the classification of each segment into relevant classes such as speech, music, or the rejection of the segment as silence or noise. Furthermore, when the audio type is *human speech*, it is desirable to do a further classification to determine the gender of the individual speaker. *Gender classification* [77] is helpful in choosing appropriate models for conducting better speech recognition, more accurate speaker verification, and reducing the computation load in large-scale speaker identification. For the speaker diarization problem, the identity of the speaker also needs to be recognized.

Dhanalakshmi, et al. [19] report developments in classifying the genre of audio, as stemming from different video sources, containing movies, cartoons, news, etc. *Beigi* [77] uses a *text* and *language*

independent speaker recognition engine to achieve these goals by performing audio classification. The classification problem is posed by *Beigi* [77] as an identification problem among a series of speech, music, and noise models.

6.3.1. Age and Gender Classification

Another goal for classification is to be able to classify age groups. *Bocklet, et al.* [81] categorized the age of the individuals, in relation to their voice quality, into 4 categories (classes). These classes are given by Table 1. With the natural exception of the child group (13 years or younger), each group is further split into the two male and female genders, leading to 7 total age-gender classes.

Class Name	Age
Child	Age \leq 13 years old
Young	14 years \leq Age \leq 19 years
Adult	20 years \leq Age \leq 64 years
Senior	65 years \geq Age

Table 1. Age Categories According to Vocal Similarities – From [81]

Class Name	Age
Young	18 years \leq Age \leq 35 years
Adult	36 years \leq Age \leq 45 years
Senior	46 years \leq Age \leq 81 years

Table 2. Age Categories According to Vocal Similarities – From [82]

Bahari, et al. [82] use a slightly different definition of age groups, compared to those used by [81]. They use 3 age groups for each gender, not considering individuals who are less than 18 years old. These age categories are given in Table 2.

They use *weighted supervised non-negative matrix factorization (WSNMF)* to classify the age and gender of the individual. This technique combines *weighted non-negative matrix factorization (WNMF)* [83] and *supervised non-negative matrix factorization (SNMF)* [84] which are themselves extensions of *non-negative matrix factorization (NMF)* [65, 66]. NMF techniques have also been successfully used in other classification implementations such as that of the identification of musical instruments [85].

NMF distinguishes itself as a method which only allows additive components that are considered to be parts of the information contained in an entity. Due to their additive and positive nature, the components are considered to, each, be part of the information that builds up a description. In contrast, methods such as principal component analysis and vector quantization techniques are considered to be learning holistic information and hence are not considered to be parts-based [66]. According to the image recognition example presented by *Lee, et al.* [66], a PCA method such as Eigenfaces [86, 87] provide a distorted version of the whole face, whereas the NMF provides localized features that are related to the parts of each face.

Subsequent to applying WSNMF, according to the age and gender, *Bahari, et al.* [82] use a *general regression neural network (GRNN)* to estimate the age of the individual. *Bahari, et al.* [82] show a

gender classification accuracy of about 96% and an average age classification accuracy of about 48%. Although it is dependent on the data being used, but an accuracy of 96% for the gender classification case is not necessarily a great result. It is hard to make a qualitative assessment without running the same algorithms under the same conditions and on exactly the same data. But *Beigi* [77] shows 98.1% accuracy for gender classification.

In [77], 700 male and 700 female speakers were selected, completely at random, from over 70,000 speakers. The speakers were non-native speakers of English, at a variety of proficiency levels, speaking freely. This introduced significantly higher number of pauses in each recording, as well as more than average number of humming sounds while the candidates would think about their speech. The segments were live responses of these non-native speakers to test questions in English, aimed at evaluating their linguistic proficiency.

Dhanalakshmi, et al. [19] also present a method based on an *auto-associative neural network* (AANN) for performing audio source classification. AANN is a special branch of feedforward neural networks which tries to learn the nonlinear principal components of a feature vector. The way this is accomplished is that the network consists of three layers, an input layer, an output layer of the same size, and a hidden layer with a smaller number of neurons. The input and output neurons generally have linear activation functions and the hidden (middle) layer has nonlinear functions.

In the training phase, the input and target output vectors are identical. This is done to allow for the system to learn the principal components that have built the patterns which most likely have built-in redundancies. Once such a network is trained, a feature vector undergoes a dimensional reduction and is then mapped back to the same dimensional space as the input space. If the training procedure is able to achieve a good reduction in the output error over the training samples and if the training samples are representative of the reality and span the operating conditions of the true system, the network can learn the essential information in the input signal. Autoassociative networks (AANN) have also been successfully used in speaker verification [88].

Class Name	Advertisement	Cartoon	Movie	News	Songs	Sports
------------	---------------	---------	-------	------	-------	--------

Table 3. Audio Classification Categories used by [19]

Dhanalakshmi, et al. [19] use the audio classes represented in Table 3. It considers three different front-end processors for extracting features, used with two different modeling techniques. The features are LPC, LPCC, and MFCC features [1]. The models are Gaussian mixture models (GMM) and autoassociative neural networks (AANN) [1]. According to these experiments, *Dhanalakshmi, et al.* [19] show consistently higher classification accuracies with MFCC features over LPC and LPCC features. The comparison between AANN and GMM is somewhat inconclusive and both systems seem to portray similar results. Although, the accuracy of AANN with LPC and LPCC seems to be higher than that of GMM modeling, for the case when MFCC features are used, the difference seems somewhat insignificant. Especially, given the fact that GMM are simpler to implement than AANN and are less prone to problems such as encountering local minima, it makes sense to conclude that the combination of MFCC and GMM still provides the best results in audio classification. A combination of GMM with MFCC and performing Maximum a-Posteriori (MAP) adaptation provides very simple and considerable results for gender classification, as seen in [77].

6.3.2. Music Modeling

Beigi [77] classifies musical instruments along with noise and gender of speakers. Much in the same spirit as described in Section 6.3.1, [77] has made an effort to choose a variety of different instruments or sets of instruments to be able to cover most types of music. Table 4 shows these choices. A total of 14 different music models were trained to represent all music, with an attempt to cover different types of timbre [89].

An equal amount of music was chosen by *Beigi* [77] to create a balance in the quantity of data, reducing any bias toward speech or music. The music was downsampled from its original quality to 8kHz, using 8-bit μ -Law amplitude encoding, in order to match the quality of speech. The 1400 segments of music were chosen at random from European style classical music, as well as jazz, Persian classical, Chinese classical, folk, and instructional performances. Most of the music samples were orchestral pieces, with some solos and duets present.

Although a very low quality audio, based on highly compressed telephony data (AAC compressed [1]), was used by *Beigi* [77], the system achieved a 1% error rate in discriminating between speech and music and a 1.9% error in determining the gender of individual speakers once the audio is tagged as speech.

Category	Model	Category	Model	Category	Model
Noise	Noise	Speech	Female	Speech	Male
Music	Accordion	Music	Bassoon	Music	Clarinet
Music	Clavier	Music	Gamelon	Music	Guzheng
Music	Guitar	Music	Oboe	Music	Orchestra
Music	Piano	Music	Pipa	Music	Tar
Music	Throat	Music	Violin		

Table 4. Audio Models used for Classification

Beigi [77] has shown that MAP adaptation techniques used with GMM models and MFCC features may be used successfully for the classification of audio into speech and music and to further classify the speech by the gender of the speaker and the music by the type of instrument being played.

7. Open problems

With all the new accomplishments in the last couple of years, covered here and many that did not make it to our list due to shortage of space, there is still a lot more work to be done. Although incremental improvements are made every day, in all branches of speaker recognition, still the channel and audio type mismatch seem to be the biggest hurdles in reaching perfect results in speaker recognition. It should be noted that perfect results are *asymptotes* and will probably never be reached. Inherently, as the size of the population in a speaker database grows, the intra-speaker variations exceed the inter-speaker variations. This is the main source of error for large-scale speaker identification, which is the holy grail of the different goals in speaker recognition. In fact, if large-scale speaker identification approaches acceptable results, most other branches of the field may be considered trivial. However, this is quite a complex problem and will definitely need a lot more time to be perfected, if it is indeed possible to do so. In the meanwhile, we seem to still be at infancy when it comes to large-scale identification.

Author details

Homayoon Beigi

President of Recognition Technologies, Inc. and an Adjunct Professor of Computer Science and Mechanical Engineering at Columbia University
Recognition Technologies, Inc., Yorktown Heights, New York, USA

References

- [1] Homayoon Beigi. *Fundamentals of Speaker Recognition*. Springer, New York, 2011. ISBN: 978-0-387-77591-3.
- [2] Homayoon Beigi. Speaker recognition. In Jucheng Yang, editor, *Biometrics*, pages 3–28. Intech Open Access Publisher, Croatia, 2011. ISBN: 978-953-307-618-8.
- [3] I. Pollack, J. M. Pickett, and W.H. Sumby. On the identification of speakers by voice. *Journal of the Acoustical Society of America*, 26(3):403–406, May 1954.
- [4] J. N. Shearme and J. N. Holmes. An experiment concerning the recognition of voices. *Language and Speech*, 2(3):123–131, 1959.
- [5] Francis Nolan. *The Phonetic Bases of Speaker Recognition*. Cambridge University Press, New York, 1983. ISBN: 0-521-24486-2.
- [6] Harry Hollien. *The Acoustics of Crime: The New Science of Forensic Phonetics (Applied Psycholinguistics and Communication Disorder)*. Springer, Heidelberg, 1990.
- [7] Harry Hollien. *Forensic Voice Identification*. Academic Press, San Diego, CA, USA, 2001.
- [8] Amy Neustein and Hemant A. Patil. *Forensic Speaker Recognition – Law Enforcement and Counter-Terrorism*. Springer, Heidelberg, 2012.
- [9] Sandra Pruzansky. Pattern matching procedure for automatic talker recognition. 35(3):354–358, Mar 1963.
- [10] Sandra Pruzansky, Max. V. Mathews, and P.B. Britner. Talker-recognition procedure based on analysis of variance. 35(11):1877–, Apr 1963.
- [11] Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 2nd edition, 2000. ISBN: 0-471-00626-2.
- [12] Vladimir Naumovich Vapnik. *Statistical learning theory*. John Wiley, New York, 1998. ISBN: 0-471-03003-1.
- [13] A. Solomonoff, W. Campbell, and C. Quillen. Channel compensation for svm speaker recognition. In *The Speaker and Language Recognition Workshop Odyssey 2004*, volume 1, pages 57–62, 2004.
- [14] Robbie Vogt and Sridha Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech and Language*, 22(1):17–38, Jan. 2008.

- [15] Sahar E. Bou-Ghazale and John H. L. Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4):429–442, Jul 2002.
- [16] Elliott D. Canonge. Voiceless vowels in comanche. *International Journal of American Linguistics*, 23(2):63–67, Apr 1957. Published by: The University of Chicago Press.
- [17] Qin Jin, Szu-Chen Stan Jou, and T. Schultz. Whispering speaker identification. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1027–1030, Jul 2007.
- [18] Xing Fan and J.H.L. Hansen. Speaker identification within whispered speech audio streams. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1408–1421, Jul 2011.
- [19] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam. Classification of audio signals using aann and gmm. *Applied Soft Computing*, 11(1):716 – 723, 2011.
- [20] Lucas C. Parra and Christopher V. Alvino. Geometric source separation: merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, Sep 2002.
- [21] K. Kumatani, U. Mayer, T. Gehrig, E. Stoimenov, and M. Wolfel. Minimum mutual information beamforming for simultaneous active speakers. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 71–76, Dec 2007.
- [22] M. Lincoln. The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 357–362, Nov 2005.
- [23] R. Takashima, T. Takiguchi, and Y. Ariki. Hmm-based separation of acoustic transfer function for single-channel sound source localization. pages 2830–2833, Mar 2010.
- [24] C. Barras and J.-L. Gauvain. Feature and score normalization for speaker verification of cellular data. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 2, pages II–49–52, Apr 2003.
- [25] P. Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. Technical report, CRIM, Jan 2006.
- [26] Ondrej Glembek, Lukas Burget, Pavel Matejka, Martin Karafiat, and Patrick Kenny. Simplification and optimization of i-vector extraction. pages 4516–4519, May 2011.
- [27] W.M. Campbell, D.E. Sturim, W. Shen, D.A. Reynolds, and J. Navratil. The mit-ill/ibm 2006 speaker recognition system: High-performance reduced-complexity recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–217–IV–220, Apr 2007.
- [28] Hyunson Seo, Chi-Sang Jung, and Hong-Goo Kang. Robust session variability compensation for svm speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6):1631–1641, Aug 2011.

- [29] P. Shanmugapriya and Y. Venkataramani. Implementation of speaker verification system using fuzzy wavelet network. In *Communications and Signal Processing (ICCSP), 2011 International Conference on*, pages 460–464, Feb 2011.
- [30] J. Villalba and E. Lleida. Preventing replay attacks on speaker verification systems. In *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on*, pages 1–8, Oct 2011.
- [31] Johan Sandberg, Maria Hansson-Sandsten, Tomi Kinnunen, Rahim Saeidi Patrick Flandrin, , and Pierre Borgnat. Multitaper estimation of frequency-warped cepstra with application to speaker verification. *IEEE Signal Processing Letters*, 17(4):343–346, Apr 2010.
- [32] David J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, Sep 1982.
- [33] Kurt S. Riedel, Alexander Sidorenko, and David J. Thomson. Spectral estimation of plasma fluctuations. i. comparison of methods. *Physics of Plasma*, 1(3):485–500, 1994.
- [34] Kurt S. Riedel. Minimum bias multiple taper spectral estimation. *IEEE Transactions on Signal Processing*, 43(1):188–195, Jan 1995.
- [35] Maria Hansson and Göran Salomonsson. A multiple window method for estimation of peaked spectra. *IEEE Transactions on Signal Processing*, 45(3):778–781, Mar 1997.
- [36] Qi Li and Yan Huang. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6):1791–1801, Aug 2011.
- [37] Qi Peter Li. An auditory-based transform for audio signal processing. In *IEEE Workshop on Applications of Signal Processing to audio and Acoustics*, pages 181–184, Oct 2009.
- [38] Yang Shao and DeLiang Wang. Robust speaker identification using auditory features and computational auditory scene analysis. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1589–1592, 2008.
- [39] Brian C. J. Moore and Brian R. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation. *Journal of Acoustical Society of America*, 74(3):750–753, 1983.
- [40] Brian C. J. Moore and Brian R. Glasberg. A revision of zwicker’s loudness model. *Acta Acustica*, 82(2):335–345, Mar/Apr 1996.
- [41] Brian R. Glasberg and Brian C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1–2):103–138, 1990.
- [42] E. Zwicker, G. Flottorp, and Stanley Smith Stevens. Critical band width in loudness summation. *Journal of the Acoustical Society of America*, 29(5):548–557, 1957.
- [43] Xiaojia Zhao, Yang Shao, and DeLiang Wang. Robust speaker identification using a casa front-end. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5468–5471, May 2011.

- [44] Albert S. Bergman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Bradford, 1994.
- [45] A. Vizinho, P. Green, M. Cooke, and L. Josifovski. Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: An integrated study. In *Eurospeech 1999*, pages 2407–2410, Sep 1999.
- [46] Michael L. Seltzer, Bhiksha Raj, and Richard M. Stern. A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43(4):379–393, 2004.
- [47] D. Pullella, M. Kuhne, and R. Togneri. Robust speaker identification using combined feature selection and missing data recognition. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4833–4836, 2008.
- [48] Shin-Cheol Lim, Sei-Jin Jang, Soek-Pil Lee, and Moo Young Kim. Hard-mask missing feature theory for robust speaker recognition. *Consumer Electronics, IEEE Transactions on*, 57(3):1245–1250, Aug 2011.
- [49] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.
- [50] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, pages 148–156, 1996.
- [51] Yann Rodriguez. *Face Detection and Verification Using Local Binary Patterns*. Ecole Polytechnique Fédérale de Lausanne, 2006. PhD Thesis.
- [52] Anindya Roy, Mathew Magimai-Doss, and Sébastien Marcel. Boosted binary features for noise-robust speaker verification. volume 6, pages 4442–4445, Mar 2010.
- [53] A. Roy, M. M. Doss, and S. Marcel. A fast parts-based approach to speaker verification using boosted slice classifiers. *IEEE Transactions on Information Forensic and Security*, 7(1):241–254, 2012.
- [54] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brummer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *InterSpeech*, pages 1559–1562, Sep 2009.
- [55] Najim Dehak, Reda Dehak, James Glass, Douglas Reynolds, and Patrick Kenny. Cosine similarity scoring without score normalization techniques. In *The Speaker and Language Recognition Workshop (Odyssey 2010)*, pages 15–19, Jun-Jul 2010.
- [56] Mohammed Senoussaoui, Patrick Kenny, Najim Dehak, and Pierre Dumouchel. An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. In *The Speaker and Language Recognition Workshop (Odyssey 2010)*, pages 28–33, June 2010.
- [57] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):788–798, May 2011.

- [58] Najim Dehak, Patrick Kenny, Réda Dehak, O. Glembek, Pierre Dumouchel, L. Burget, V. Hubeika, and F. Castaldo. Support vector machines and joint factor analysis for speaker verification. pages 4237–4240, Apr 2009.
- [59] M. Senoussaoui, P. Kenny, P. Dumouchel, and F. Castaldo. Well-calibrated heavy tailed bayesian speaker verification for microphone speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4824–4827, May 2011.
- [60] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Briimmer. Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4832–4835, May 2011.
- [61] S. Cumani, N. Brummer, L. Burget, and P. Laface. Fast discriminative speaker verification in the i-vector space. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4852–4855, May 2011.
- [62] P. Matejka, O. Glembek, F. Castaldo, M.J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky. Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4828–4831, May 2011.
- [63] M.J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O’Shaughnessy. Multi-taper mfcc features for speaker verification using i-vectors. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 547–552, Dec 2011.
- [64] Nagendra Kumar and Andreas G. Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26(4):283–297, 1998.
- [65] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [66] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.
- [67] A.L. Bartos and D.J. Nelson. Enabling improved speaker recognition by voice quality estimation. In *Signals, Systems and Computers (ASIOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 595–599, Nov 2011.
- [68] A. Salman and Ke Chen. Exploring speaker-specific characteristics with deep learning. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 103–110, 2011.
- [69] A.K. Sarkar and S. Umesh. Use of vtl-wise models in feature-mapping framework to achieve performance of multiple-background models in speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4552–4555, May 2011.
- [70] B. Bharathi, P. Vijayalakshmi, and T. Nagarajan. Speaker identification using utterances correspond to speaker-specific-text. In *Students’ Technology Symposium (TechSym), 2011 IEEE*, pages 171–174, Jan 2011.

- [71] Wei Cai, Qiang Li, and Xin Guan. Automatic singer identification based on auditory features. In *Natural Computation (ICNC), 2011 Seventh International Conference on*, volume 3, pages 1624–1628, Jul 2011.
- [72] Hoang Do, I. Tashev, and A. Acero. A new speaker identification algorithm for gaming scenarios. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5436–5439, May 2011.
- [73] M.V. Ghiurcau, C. Rusu, and J. Astola. A study of the effect of emotional state upon text-independent speaker identification. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4944–4947, May 2011.
- [74] Jia-Wei Liu, Jia-Ching Wang, and Chang-Hong Lin. Speaker identification using hht spectrum features. In *Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on*, pages 145–148, Nov 2011.
- [75] John C. Platt, Nello Cristianini, and John Shawe-Taylor. Large margin dags for multiclass classification. In S.A. Solla, T.K. Leen, and K.R. Müller, editors, *Advances in Neural Information processing Systems*. MIT Press, Boston, 2000.
- [76] Yuguo Wang. A tree-based multi-class svm classifier for digital library document. In *International Conference on MultiMedia and Information Technology (MMIT)*, pages 15–18, Dec 2008.
- [77] Homayoon Beigi. Audio source classification using speaker recognition techniques. World Wide Web, Feb 2011. Report No. RTI-20110201-01.
- [78] Stephane H. Maes Homayoon S. M. Beigi. Speaker, channel and environment change detection. Technical Report, 1997.
- [79] Homayoon S.M. Beigi and Stephane S. Maes. Speaker, channel and environment change detection. In *Proceedings of the World Congress on Automation (WAC1998)*, May 1998.
- [80] Scott Shaobing Chen and Ponani S Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian inromation criterion. In *IBM Techical Report, T.J. Watson Research Center*, 1998.
- [81] Tobia Bocklet, Andreas Maier, Josef G. Bauer, Felix Burkhardt, and Elmar Nöth. Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. pages 1605–1608, Apr 2008.
- [82] M.H. Bahari and H. Van Hamme. Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In *Biometric Measurements and Systems for Security and Medical Applications (BIOMS), 2011 IEEE Workshop on*, pages 1–6, Sep 2011.
- [83] N. Ho. *Nonnegative Martix Factorization Algorithms and Applications*. Université Catholique de Louvain, 2008. PhD Thesis.
- [84] H. Van-Hamme. Hac-models: A novel approach to continuous speech recognition. In *Interspeech*, pages 2554–2557, Sep 2008.

- [85] Emmanouil Benetos, Margarita Kotti, and Constantine Kotropoulos. Large scale musical instrument identification. In *Proceedings of the 4th Sound and Music Computing Conference*, pages 283–286, Jul 2007.
- [86] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, Jan. 1990.
- [87] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.
- [88] S.P. Kishore and B. Yegnanarayana. Speaker verification: minimizing the channel effects using autoassociative neural network models. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, volume 2, pages III1101–III1104, Jun 2000.
- [89] Keith Dana Martin. *Sound-Source Recognition: A Theory and Computational Model*. Massachusetts Institute of Technology, Cambridge, MA, 1999. PhD Thesis.

