EURASIP Journal on Applied Signal Processing 2004:4, 430–451

A Tutorial on Text-Independent Speaker Verification

Frédéric Bimbot,[1] Jean-François Bonastre,[2] Corinne Fredouille,[2] Guillaume Gravier,[1] Ivan Magrin-Chagnolleau,[3] Sylvain Meignier,[2] Teva Merlin,[2] Javier Ortega-García,[4] Dijana Petrovska-Delacrétaz,[5] and Douglas A. Reynolds[6]

[1] *IRISA, INRIA & CNRS, 35042 Rennes Cedex, France*
Emails: bimbot@irisa.fr; ggravier@irisa.fr

[2] *LIA, University of Avignon, 84911 Avignon Cedex 9, France*
Emails: jean-francois.bonastre@lia.univ-avignon.fr; corinne.fredouille@lia.univ-avignon.fr; sylvain.meignier@lia.univ-avignon.fr; teva.merlin@lia.univ-avignon.fr

[3] *Laboratoire Dynamique du Langage, CNRS, 69369 Lyon Cedex 07, France*
Email: ivan@ieee.org

[4] *ATVS, Universidad Politécnica de Madrid, 28040 Madrid, Spain*
Email: jortega@diac.upm.es

[5] *DIVA Laboratory, Informatics Department, Fribourg University, CH-1700 Fribourg, Switzerland*
Email: dijana.petrovski@unifr.ch

[6] *Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02420-9108, USA*
Email: dar@ll.mit.edu

Received 2 December 2002; Revised 8 August 2003

This paper presents an overview of a state-of-the-art text-independent speaker verification system. First, an introduction proposes a modular scheme of the training and test phases of a speaker verification system. Then, the most commonly speech parameterization used in speaker verification, namely, cepstral analysis, is detailed. Gaussian mixture modeling, which is the speaker modeling technique used in most systems, is then explained. A few speaker modeling alternatives, namely, neural networks and support vector machines, are mentioned. Normalization of scores is then explained, as this is a very important step to deal with real-world data. The evaluation of a speaker verification system is then detailed, and the detection error trade-off (DET) curve is explained. Several extensions of speaker verification are then enumerated, including speaker tracking and segmentation by speakers. Then, some applications of speaker verification are proposed, including on-site applications, remote applications, applications relative to structuring audio information, and games. Issues concerning the forensic area are then recalled, as we believe it is very important to inform people about the actual performance and limitations of speaker verification systems. This paper concludes by giving a few research trends in speaker verification for the next couple of years.

Keywords and phrases: speaker verification, text-independent, cepstral analysis, Gaussian mixture modeling.

1. INTRODUCTION

Numerous measurements and signals have been proposed and investigated for use in biometric recognition systems. Among the most popular measurements are fingerprint, face, and voice. While each has pros and cons relative to accuracy and deployment, there are two main factors that have made voice a compelling biometric. First, speech is a natural signal to produce that is not considered threatening by users to provide. In many applications, speech may be the main (or only, e.g., telephone transactions) modality, so users do not consider providing a speech sample for authentication as a separate or intrusive step. Second, the telephone system provides a ubiquitous, familiar network of sensors for obtaining and delivering the speech signal. For telephone-based applications, there is no need for special signal transducers or networks to be installed at application access points since a cell phone gives one access almost anywhere. Even for non-telephone applications, sound cards and microphones are low-cost and readily available. Additionally, the speaker recognition area has a long and rich scientific basis with over 30 years of research, development, and evaluations.

Over the last decade, speaker recognition technology has made its debut in several commercial products. The specific

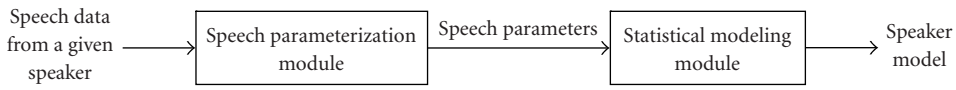


Figure 1: Modular representation of the training phase of a speaker verification system.

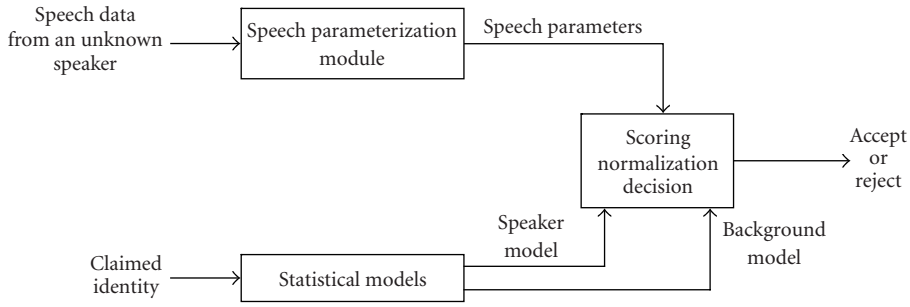


Figure 2: Modular representation of the test phase of a speaker verification system.

recognition task addressed in commercial systems is that of verification or detection (determining whether an unknown voice is from a particular enrolled speaker) rather than identification (associating an unknown voice with one from a set of enrolled speakers). Most deployed applications are based on scenarios with cooperative users speaking fixed digit string passwords or repeating prompted phrases from a small vocabulary. These generally employ what is known as text-dependent or text-constrained systems. Such constraints are quite reasonable and can greatly improve the accuracy of a system; however, there are cases when such constraints can be cumbersome or impossible to enforce. An example of this is background verification where a speaker is verified behind the scene as he/she conducts some other speech interactions. For cases like this, a more flexible recognition system able to operate without explicit user cooperation and independent of the spoken utterance (called text-independent mode) is needed. This paper focuses on the technologies behind these text-independent speaker verification systems.

A speaker verification system is composed of two distinct phases, a training phase and a test phase. Each of them can be seen as a succession of independent modules. Figure 1 shows a modular representation of the training phase of a speaker verification system. The first step consists in extracting parameters from the speech signal to obtain a representation suitable for statistical modeling as such models are extensively used in most state-of-the-art speaker verification systems. This step is described in Section 2. The second step consists in obtaining a statistical model from the parameters. This step is described in Section 3. This training scheme is also applied to the training of a background model (see Section 3).

Figure 2 shows a modular representation of the test phase of a speaker verification system. The entries of the system are a claimed identity and the speech samples pronounced by an unknown speaker. The purpose of a speaker verification system is to verify if the speech samples correspond to the claimed identity. First, speech parameters are extracted from the speech signal using exactly the same module as for the training phase (see Section 2). Then, the speaker model corresponding to the claimed identity and a background model are extracted from the set of statistical models calculated during the training phase. Finally, using the speech parameters extracted and the two statistical models, the last module computes some scores, normalizes them, and makes an acceptance or a rejection decision (see Section 4). The normalization step requires some score distributions to be estimated during the training phase or/and the test phase (see the details in Section 4).

Finally, a speaker verification system can be text-dependent or text-independent. In the former case, there is some constraint on the type of utterance that users of the system can pronounce (for instance, a fixed password or certain words in any order, etc.). In the latter case, users can say whatever they want. This paper describes state-of-the-art text-independent speaker verification systems.

The outline of the paper is the following. Section 2 presents the most commonly used speech parameterization techniques in speaker verification systems, namely, cepstral analysis. Statistical modeling is detailed in Section 3, including an extensive presentation of Gaussian mixture modeling (GMM) and the mention of several speaker modeling alternatives like neural networks and support vector machines (SVMs). Section 4 explains how normalization is used. Section 5 shows how to evaluate a speaker verification system. In Section 6, several extensions of speaker verification are presented, namely, speaker tracking and speaker segmentation. Section 7 gives a few applications of speaker verification. Section 8 details specific problems relative to the use of speaker verification in the forensic area. Finally, Section 9 concludes this work and gives some future research directions.

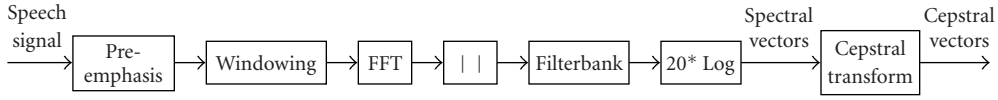


Figure 3: Modular representation of a filterbank-based cepstral parameterization.

2. SPEECH PARAMETERIZATION

Speech parameterization consists in transforming the speech signal to a set of feature vectors. The aim of this transformation is to obtain a new representation which is more compact, less redundant, and more suitable for statistical modeling and the calculation of a distance or any other kind of score. Most of the speech parameterizations used in speaker verification systems relies on a cepstral representation of speech.

2.1. *Filterbank-based cepstral parameters*

Figure 3 shows a modular representation of a filterbank-based cepstral representation.

The speech signal is first preemphasized, that is, a filter is applied to it. The goal of this filter is to enhance the high frequencies of the spectrum, which are generally reduced by the speech production process. The preemphasized signal is obtained by applying the following filter:

$$x_p(t) = x(t) - a \cdot x(t-1). \tag{1}$$

Values of a are generally taken in the interval $[0.95,\ 0.98]$. This filter is not always applied, and some people prefer not to preemphasize the signal before processing it. There is no definitive answer to this topic but empirical experimentation.

The analysis of the speech signal is done locally by the application of a window whose duration in time is shorter than the whole signal. This window is first applied to the beginning of the signal, then moved further and so on until the end of the signal is reached. Each application of the window to a portion of the speech signal provides a spectral vector (after the application of an FFT—see below). Two quantities have to be set: the length of the window and the shift between two consecutive windows. For the length of the window, two values are most often used: 20 milliseconds and 30 milliseconds. These values correspond to the average duration which allows the stationary assumption to be true. For the delay, the value is chosen in order to have an overlap between two consecutive windows; 10 milliseconds is very often used. Once these two quantities have been chosen, one can decide which window to use. The Hamming and the Hanning windows are the most used in speaker recognition. One usually uses a Hamming window or a Hanning window rather than a rectangular window to taper the original signal on the sides and thus reduce the side effects. In the Fourier domain, there is a convolution between the Fourier transform of the portion of the signal under consideration and the Fourier transform of the window. The Hamming window and the Hanning window are much more selective than the rectangular window.

Once the speech signal has been windowed, and possibly preemphasized, its fast Fourier transform (FFT) is calculated. There are numerous algorithms of FFT (see, for instance, [1, 2]).

Once an FFT algorithm has been chosen, the only parameter to fix for the FFT calculation is the number of points for the calculation itself. This number N is usually a power of 2 which is greater than the number of points in the window, classically 512.

Finally, the modulus of the FFT is extracted and a power spectrum is obtained, sampled over 512 points. The spectrum is symmetric and only half of these points are really useful. Therefore, only the first half of it is kept, resulting in a spectrum composed of 256 points.

The spectrum presents a lot of fluctuations, and we are usually not interested in all the details of them. Only the envelope of the spectrum is of interest. Another reason for the smoothing of the spectrum is the reduction of the size of the spectral vectors. To realize this smoothing and get the envelope of the spectrum, we multiply the spectrum previously obtained by a filterbank. A filterbank is a series of bandpass frequency filters which are multiplied one by one with the spectrum in order to get an average value in a particular frequency band. The filterbank is defined by the shape of the filters and by their frequency localization (left frequency, central frequency, and right frequency). Filters can be triangular, or have other shapes, and they can be differently located on the frequency scale. In particular, some authors use the Bark/Mel scale for the frequency localization of the filters. This scale is an auditory scale which is similar to the frequency scale of the human ear. The localization of the central frequencies of the filters is given by

$$f_{\text{MEL}} = 1000 \cdot \frac{\log\left(1 + f_{\text{LIN}}/1000\right)}{\log 2}. \tag{2}$$

Finally, we take the log of this spectral envelope and multiply each coefficient by 20 in order to obtain the spectral envelope in dB. At the stage of the processing, we obtain spectral vectors.

An additional transform, called the cosine discrete transform, is usually applied to the spectral vectors in speech processing and yields cepstral coefficients [2, 3, 4]:

$$c_n = \sum_{k=1}^{K} S_k \cdot \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1, 2, \ldots, L, \tag{3}$$

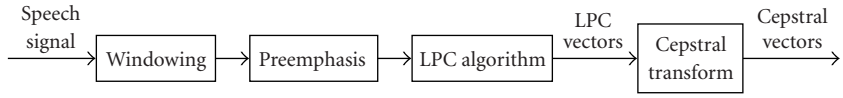


Figure 4: Modular representation of an LPC-based cepstral parameterization.

where K is the number of log-spectral coefficients calculated previously, S_k are the log-spectral coefficients, and L is the number of cepstral coefficients that we want to calculate ($L \leq K$). We finally obtain cepstral vectors for each analysis window.

2.2. LPC-based cepstral parameters

Figure 4 shows a modular representation of an LPC-based cepstral representation.

The LPC analysis is based on a linear model of speech production. The model usually used is an auto regressive moving average (ARMA) model, simplified in an auto regressive (AR) model. This modeling is detailed in particular in [5].

The speech production apparatus is usually described as a combination of four modules: (1) the glottal source, which can be seen as a train of impulses (for voiced sounds) or a white noise (for unvoiced sounds); (2) the vocal tract; (3) the nasal tract; and (4) the lips. Each of them can be represented by a filter: a lowpass filter for the glottal source, an AR filter for the vocal tract, an ARMA filter for the nasal tract, and an MA filter for the lips. Globally, the speech production apparatus can therefore be represented by an ARMA filter. Characterizing the speech signal (usually a windowed portion of it) is equivalent to determining the coefficients of the global filter. To simplify the resolution of this problem, the ARMA filter is often simplified in an AR filter.

The principle of LPC analysis is to estimate the parameters of an AR filter on a windowed (preemphasized or not) portion of a speech signal. Then, the window is moved and a new estimation is calculated. For each window, a set of coefficients (called predictive coefficients or LPC coefficients) is estimated (see [2, 6] for the details of the various algorithms that can be used to estimate the LPC coefficients) and can be used as a parameter vector. Finally, a spectrum envelope can be estimated for the current window from the predictive coefficients. But it is also possible to calculate cepstral coefficients directly from the LPC coefficients (see [6]):

$$\begin{aligned} c_0 &= \ln \sigma^2, \\ c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leq m \leq p, \\ c_m &= \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad p < m, \end{aligned} \tag{4}$$

where σ^2 is the gain term in the LPC model, a_m are the LPC coefficients, and p is the number of LPC coefficients calculated.

2.3. Centered and reduced vectors

Once the cepstral coefficients have been calculated, they can be centered, that is, the cepstral mean vector is subtracted from each cepstral vector. This operation is called cepstral mean subtraction (CMS) and is often used in speaker verification. The motivation for CMS is to remove from the cepstrum the contribution of slowly varying convolutive noises.

The cepstral vectors can also be reduced, that is, the variance is normalized to one component by component.

2.4. Dynamic information

After the cepstral coefficients have been calculated, and possibly centered and reduced, we also incorporate in the vectors some dynamic information, that is, some information about the way these vectors vary in time. This is classically done by using the Δ and $\Delta\Delta$ parameters, which are polynomial approximations of the first and second derivatives [7]:

$$\begin{aligned} \Delta c_m &= \frac{\sum_{k=-l}^{l} k \cdot c_{m+k}}{\sum_{k=-l}^{l} |k|}, \\ \Delta\Delta c_m &= \frac{\sum_{k=-l}^{l} k^2 \cdot c_{m+k}}{\sum_{k=-l}^{l} k^2}. \end{aligned} \tag{5}$$

2.5. Log energy and Δ log energy

At this step, one can choose whether to incorporate the log energy and the Δ log energy in the feature vectors or not. In practice, the former one is often discarded and the latter one is kept.

2.6. Discarding useless information

Once all the feature vectors have been calculated, a very important last step is to decide which vectors are useful and which are not. One way of looking at the problem is to determine vectors corresponding to speech portions of the signal versus those corresponding to silence or background noise. A way of doing it is to compute a bi-Gaussian model of the feature vector distribution. In that case, the Gaussian with the "lowest" mean corresponds to silence and background noise, and the Gaussian with the "highest" mean corresponds to speech portions. Then vectors having a higher likelihood with the silence and background noise Gaussian are discarded. A similar approach is to compute a bi-Gaussian model of the log energy distribution of each speech segment and to apply the same principle.

3. STATISTICAL MODELING

3.1. Speaker verification via likelihood ratio detection

Given a segment of speech Y and a hypothesized speaker S, the task of speaker verification, also referred to as detection, is to determine if Y was spoken by S. An implicit assumption often used is that Y contains speech from only one speaker. Thus, the task is better termed singlespeaker verification. If there is no prior information that Y contains speech from a single speaker, the task becomes multispeaker detection. This paper is primarily concerned with the single-speaker verification task. Discussion of systems that handle the multispeaker detection task is presented in other papers [8].

The single-speaker detection task can be stated as a basic hypothesis test between two hypotheses:

H0: Y is from the hypothesized speaker S,
H1: Y is *not* from the hypothesized speaker S.

The optimum test to decide between these two hypotheses is a likelihood ratio (LR) test[1] given by

$$\frac{p(Y|\mathrm{H0})}{p(Y|\mathrm{H1})} \begin{cases} > \theta, & \text{accept H0,} \\ < \theta, & \text{accept H1,} \end{cases} \tag{6}$$

where $p(Y|\mathrm{H0})$ is the probability density function for the hypothesis H0 evaluated for the observed speech segment Y, also referred to as the "likelihood" of the hypothesis H0 given the speech segment.[2] The likelihood function for H1 is likewise $p(Y|\mathrm{H1})$. The decision threshold for accepting or rejecting H0 is θ. One main goal in designing a speaker detection system is to determine techniques to compute values for the two likelihoods $p(Y|\mathrm{H0})$ and $p(Y|\mathrm{H1})$.

Figure 5 shows the basic components found in speaker detection systems based on LRs. As discussed in Section 2, the role of the front-end processing is to extract from the speech signal features that convey speaker-dependent information. In addition, techniques to minimize confounding effects from these features, such as linear filtering or noise, may be employed in the front-end processing. The output of this stage is typically a sequence of feature vectors representing the test segment $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, where $\vec{x}_t$ is a feature vector indexed at discrete time $t \in [1, 2, \ldots, T]$. There is no inherent constraint that features extracted at synchronous time instants be used; as an example, the overall speaking rate of an utterance could be used as a feature. These feature vectors are then used to compute the likelihoods of H0 and H1. Mathematically, a model denoted by λ_{hyp} represents H0, which characterizes the hypothesized speaker S in the feature space of $\vec{x}$. For example, one could assume that a Gaussian distribution best represents the distribution of feature vectors for H0 so that λ_{hyp} would contain the mean vector and covariance matrix parameters of the Gaussian distribution. The model

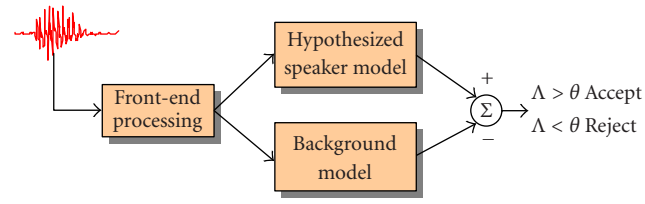


Figure 5: Likelihood-ratio-based speaker verification system.

$\lambda_{\overline{\mathrm{hyp}}}$ represents the alternative hypothesis, H1. The likelihood ratio statistic is then $p(X|\lambda_{\mathrm{hyp}})/p(X|\lambda_{\overline{\mathrm{hyp}}})$. Often, the logarithm of this statistic is used giving the log LR

$$\Lambda(X) = \log p(X|\lambda_{\mathrm{hyp}}) - \log p\left(X|\lambda_{\overline{\mathrm{hyp}}}\right). \tag{7}$$

While the model for H0 is well defined and can be estimated using training speech from S, the model for $\lambda_{\overline{\mathrm{hyp}}}$ is less well defined since it potentially must represent the entire space of possible alternatives to the hypothesized speaker. Two main approaches have been taken for this alternative hypothesis modeling. The first approach is to use a set of other speaker models to cover the space of the alternative hypothesis. In various contexts, this set of other speakers has been called likelihood ratio sets [9], cohorts [9, 10], and background speakers [9, 11]. Given a set of N background speaker models $\{\lambda_1, \ldots, \lambda_N\}$, the alternative hypothesis model is represented by

$$p\left(X|\lambda_{\overline{\mathrm{hyp}}}\right) = f(p(X|\lambda_1), \ldots, p(X|\lambda_N)), \tag{8}$$

where $f(\cdot)$ is some function, such as average or maximum, of the likelihood values from the background speaker set. The selection, size, and combination of the background speakers have been the subject of much research [9, 10, 11, 12]. In general, it has been found that to obtain the best performance with this approach requires the use of speaker-specific background speaker sets. This can be a drawback in applications using a large number of hypothesized speakers, each requiring their own background speaker set.

The second major approach to the alternative hypothesis modeling is to pool speech from several speakers and train a single model. Various terms for this single model are a general model [13], a world model, and a universal background model (UBM) [14]. Given a collection of speech samples from a large number of speakers representative of the population of speakers expected during verification, a single model λ_{bkg}, is trained to represent the alternative hypothesis. Research on this approach has focused on selection and composition of the speakers and speech used to train the single model [15, 16]. The main advantage of this approach is that a single speaker-independent model can be trained once for a particular task and then used for all hypothesized speakers in that task. It is also possible to use multiple background models tailored to specific sets of speakers [16, 17]. The use of a single background model has become the predominate approach used in speaker verification systems.

[1] Strictly speaking, the likelihood ratio test is only optimal when the likelihood functions are known exactly. In practice, this is rarely the case.

[2] $p(A|B)$ is referred to as a likelihood when B is considered the independent variable in the function.

3.2. Gaussian mixture models

An important step in the implementation of the above likelihood ratio detector is the selection of the actual likelihood function $p(X|\lambda)$. The choice of this function is largely dependent on the features being used as well as specifics of the application. For text-independent speaker recognition, where there is no prior knowledge of what the speaker will say, the most successful likelihood function has been GMMs. In text-dependent applications, where there is a strong prior knowledge of the spoken text, additional temporal knowledge can be incorporated by using hidden Markov models (HMMs) for the likelihood functions. To date, however, the use of more complicated likelihood functions, such as those based on HMMs, have shown no advantage over GMMs for text-independent speaker detection tasks like in the NIST speaker recognition evaluations (SREs).

For a D-dimensional feature vector $\vec{x}$, the mixture density used for the likelihood function is defined as follows:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\vec{x}). \tag{9}$$

The density is a weighted linear combination of M unimodal Gaussian densities $p_i(\vec{x})$, each parameterized by a $D\times 1$ mean vector $\vec{\mu}_i$ and a $D \times D$ covariance matrix Σ_i:

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)}. \tag{10}$$

The mixture weights w_i further satisfy the constraint $\sum_{i=1}^{M} w_i = 1$. Collectively, the parameters of the density model are denoted as $\lambda = (w_i, \vec{\mu}_i, \Sigma_i)$, $i = (1, \ldots, M)$.

While the general model form supports full covariance matrices, that is, a covariance matrix with all its elements, typically only diagonal covariance matrices are used. This is done for three reasons. First, the density modeling of an Mth-order full covariance GMM can equally well be achieved using a larger-order diagonal covariance GMM.[3] Second, diagonal-matrix GMMs are more computationally efficient than full covariance GMMs for training since repeated inversions of a $D\times D$ matrix are not required. Third, empirically, it has been observed that diagonal-matrix GMMs outperform full-matrix GMMs.

Given a collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation-maximization (EM) algorithm [18]. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors, that is, for iterations k and $k+1$, $p(X|\lambda^{(k+1)}) \geq p(X|\lambda^{(k)})$. Generally, five–ten iterations are sufficient for parameter convergence. The EM equations for training a GMM can be found in the literature [18, 19, 20].

[3] GMMs with $M > 1$ using diagonal covariance matrices can model distributions of feature vectors with correlated elements. Only in the degenerate case of $M = 1$ is the use of a diagonal covariance matrix incorrect for feature vectors with correlated elements.

Under the assumption of independent feature vectors, the log-likelihood of a model λ for a sequence of feature vectors $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$ is computed as follows:

$$\log p(X|\lambda) = \frac{1}{T}\sum_{t} \log p(\vec{x}_t|\lambda), \tag{11}$$

where $p(\vec{x}_t|\lambda)$ is computed as in equation (9). Note that the average log-likelihood value is used so as to normalize out duration effects from the log-likelihood value. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by T can be considered a rough compensation factor.

The GMM can be viewed as a hybrid between parametric and nonparametric density models. Like a parametric model, it has structure and parameters that control the behavior of the density in known ways, but without constraints that the data must be of a specific distribution type, such as Gaussian or Laplacian. Like a nonparametric model, the GMM has many degrees of freedom to allow arbitrary density modeling, without undue computation and storage demands. It can also be thought of as a single-state HMM with a Gaussian mixture observation density, or an ergodic Gaussian observation HMM with fixed, equal transition probabilities. Here, the Gaussian components can be considered to be modeling the underlying broad phonetic sounds that characterize a person's voice. A more detailed discussion of how GMMs apply to speaker modeling can be found elsewhere [21].

The advantages of using a GMM as the likelihood function are that it is computationally inexpensive, is based on a well-understood statistical model, and, for text-independent tasks, is insensitive to the temporal aspects of the speech, modeling only the underlying distribution of acoustic observations from a speaker. The latter is also a disadvantage in that higher-levels of information about the speaker conveyed in the temporal speech signal are not used. The modeling and exploitation of these higher-levels of information may be where approaches based on speech recognition [22] produce benefits in the future. To date, however, these approaches (e.g., large vocabulary or phoneme recognizers) have basically been used only as means to compute likelihood values, without explicit use of any higher-level information, such as speaker-dependent word usage or speaking style. Some recent work, however, has shown that high-level information can be successfully extracted and combined with acoustic scores from a GMM system for improved speaker verification performance [23, 24].

3.3. Adapted GMM system

As discussed earlier, the dominant approach to background modeling is to use a single, speaker-independent background model to represent $p(X|\lambda_{\overline{\text{hyp}}})$. Using a GMM as the likelihood function, the background model is typically a large GMM trained to represent the speaker-independent distribution of features. Specifically, speech should be selected that reflects the expected alternative speech to be encountered during recognition. This applies to the type and quality of speech as well as the composition of speakers. For

example, in the NIST SRE single-speaker detection tests, it is known a priori that the speech comes from local and long-distance telephone calls, and that male hypothesized speakers will only be tested against male speech. In this case, we would train the UBM used for male tests using only male telephone speech. In the case where there is no prior knowledge of the gender composition of the alternative speakers, we would train using gender-independent speech. The GMM order for the background model is usually set between 512–2048 mixtures depending on the data. Lower-order mixtures are often used when working with constrained speech (such as digits or fixed vocabulary), while 2048 mixtures are used when dealing with unconstrained speech (such as conversational speech).

Other than these general guidelines and experimentation, there is no objective measure to determine the right number of speakers or amount of speech to use in training a background model. Empirically, from the NIST SRE, we have observed no performance loss using a background model trained with one hour of speech compared to a one trained using six hours of speech. In both cases, the training speech was extracted from the same speaker population.

For the speaker model, a single GMM can be trained using the EM algorithm on the speaker's enrollment data. The order of the speaker's GMM will be highly dependent on the amount of enrollment speech, typically 64–256 mixtures. In another more successful approach, the speaker model is derived by adapting the parameters of the background model using the speaker's training speech and a form of Bayesian adaptation or maximum a posteriori (MAP) estimation [25]. Unlike the standard approach of maximum likelihood training of a model for the speaker, independently of the background model, the basic idea in the adaptation approach is to derive the speaker's model by updating the well-trained parameters in the background model via adaptation. This provides a tighter coupling between the speaker's model and background model that not only produces better performance than decoupled models, but, as discussed later in this section, also allows for a fast-scoring technique. Like the EM algorithm, the adaptation is a two-step estimation process. The first step is identical to the "expectation" step of the EM algorithm, where estimates of the sufficient statistics[4] of the speaker's training data are computed for each mixture in the UBM. Unlike the second step of the EM algorithm, for adaptation, these "new" sufficient statistic estimates are then combined with the "old" sufficient statistics from the background model mixture parameters using a data-dependent mixing coefficient. The data-dependent mixing coefficient is designed so that mixtures with high counts of data from the speaker rely more on the new sufficient statistics for final parameter estimation, and mixtures with low counts of data from the speaker rely more on the old sufficient statistics for final parameter estimation.

[4]These are the basic statistics required to compute the desired parameters. For a GMM mixture, these are the count, and the first and second moments required to compute the mixture weight, mean and variance.

The specifics of the adaptation are as follows. Given a background model and training vectors from the hypothesized speaker, we first determine the probabilistic alignment of the training vectors into the background model mixture components. That is, for mixture i in the background model, we compute

$$\Pr\left(i|\vec{x}_t\right) = \frac{w_i p_i(\vec{x}_t)}{\sum_{j=1}^{M} w_j p_j(\vec{x}_t)}. \tag{12}$$

We then use $\Pr(i|\vec{x}_t)$ and $\vec{x}_t$ to compute the sufficient statistics for the weight, mean, and variance parameters:[5]

$$\begin{aligned} n_i &= \sum_{t=1}^{T} \Pr\left(i|\vec{x}_t\right), \\ E_i(\vec{x}) &= \frac{1}{n_i} \sum_{t=1}^{T} \Pr\left(i|\vec{x}_t\right)\vec{x}_t, \\ E_i(\vec{x}^2) &= \frac{1}{n_i} \sum_{t=1}^{T} \Pr\left(i|\vec{x}_t\right)\vec{x}_t^2. \end{aligned} \tag{13}$$

This is the same as the expectation step in the EM algorithm.

Lastly, these new sufficient statistics from the training data are used to update the old background model sufficient statistics for mixture i to create the adapted parameters for mixture i with the equations

$$\begin{aligned} \hat{w}_i &= \left[\alpha_i n_i/T + (1-\alpha_i)w_i\right]\gamma, \\ \hat{\vec{\mu}}_i &= \alpha_i E_i(\vec{x}) + (1-\alpha_i)\vec{\mu}_i, \\ \hat{\vec{\sigma}}_i^2 &= \alpha_i E_i(\vec{x}^2) + (1-\alpha_i)\left(\vec{\sigma}_i^2 + \vec{\mu}_i^2\right) - \hat{\vec{\mu}}_i^2. \end{aligned} \tag{14}$$

The scale factor γ is computed over all adapted mixture weights to ensure they sum to unity. The adaptation coefficient controlling the balance between old and new estimates is α_i and is defined as follows:

$$\alpha_i = \frac{n_i}{n_i + r}, \tag{15}$$

where r is a fixed "relevance" factor.

The parameter updating can be derived from the general MAP estimation equations for a GMM using constraints on the prior distribution described in Gauvain and Lee's paper [25, Section V, equations (47) and (48)]. The parameter updating equation for the weight parameter, however, does not follow from the general MAP estimation equations.

Using a data-dependent adaptation coefficient allows mixture-dependent adaptation of parameters. If a mixture component has a low probabilistic count n_i of new data, then $\alpha_i \to 0$ causing the deemphasis of the new (potentially under-trained) parameters and the emphasis of the old (better trained) parameters. For mixture components with high probabilistic counts, $\alpha_i \to 1$ causing the use of the new speaker-dependent parameters. The relevance factor is a way

[5]$\vec{x}^2$ is shorthand for $\mathrm{diag}(\vec{x}\vec{x}')$.

of controlling how much new data should be observed in a mixture before the new parameters begin replacing the old parameters. This approach should thus be robust to limited training data. This factor can also be made parameter dependent, but experiments have found that this provides little benefit. Empirically, it has been found that only adapting the mean vectors provides the best performance.

Published results [14] and NIST evaluation results from several sites strongly indicate that the GMM adaptation approach provides superior performance over a decoupled system, where the speaker model is trained independently of the background model. One possible explanation for the improved performance is that the use of adapted models in the likelihood ratio is not affected by "unseen" acoustic events in recognition speech. Loosely speaking, if one considers the background model as covering the space of speaker-independent, broad acoustic classes of speech sounds, then adaptation is the speaker-dependent "tuning" of those acoustic classes observed in the speaker's training speech. Mixture parameters for those acoustic classes not observed in the training speech are merely copied from the background model. This means that during recognition, data from acoustic classes unseen in the speaker's training speech produce approximately zero log LR values that contribute evidence neither towards nor against the hypothesized speaker. Speaker models trained using only the speaker's training speech will have low likelihood values for data from classes not observed in the training data thus producing low likelihood ratio values. While this is appropriate for speech not for the speaker, it clearly can cause incorrect values when the unseen data occurs in test speech from the speaker.

The adapted GMM approach also leads to a fast-scoring technique. Computing the log LR requires computing the likelihood for the speaker and background model for each feature vector, which can be computationally expensive for large mixture orders. However, the fact that the hypothesized speaker model was adapted from the background model allows a faster scoring method. This fast-scoring approach is based on two observed effects. The first is that when a large GMM is evaluated for a feature vector, only a few of the mixtures contribute significantly to the likelihood value. This is because the GMM represents a distribution over a large space but a single vector will be near only a few components of the GMM. Thus likelihood values can be approximated very well using only the top C best scoring mixture components. The second observed effect is that the components of the adapted GMM retain a correspondence with the mixtures of the background model so that vectors close to a particular mixture in the background model will also be close to the corresponding mixture in the speaker model.

Using these two effects, a fast-scoring procedure operates as follows. For each feature vector, determine the top C scoring mixtures in the background model and compute background model likelihood using only these top C mixtures. Next, score the vector against only the corresponding C components in the adapted speaker model to evaluate the speaker's likelihood.

For a background model with M mixtures, this requires only $M + C$ Gaussian computations per feature vector compared to $2M$ Gaussian computations for normal likelihood ratio evaluation. When there are multiple hypothesized speaker models for each test segment, the savings become even greater. Typically, a value of $C = 5$ is used.

3.4. Alternative speaker modeling techniques

Another way to solve the classification problem for speaker verification systems is to use discrimination-based learning procedures such as artificial neural networks (ANN) [26, 27] or SVMs [28]. As explained in [29, 30], the main advantages of ANN include their discriminant-training power, a flexible architecture that permits easy use of contextual information, and weaker hypothesis about the statistical distributions. The main disadvantages are that their optimal structure has to be selected by trial-and-error procedures, the need to split the available train data in training and cross-validation sets, and the fact that the temporal structure of speech signals remains difficult to handle. They can be used as binary classifiers for speaker verification systems to separate the speaker and the nonspeaker classes as well as multicategory classifiers for speaker identification purposes. ANN have been used for speaker verification [31, 32, 33]. Among the different ANN architectures, multilayer perceptrons (MLP) are often used [6, 34].

SVMs are an increasingly popular method used in speaker verifications systems. SVM classifiers are well suited to separate rather complex regions between two classes through an optimal, nonlinear decision boundary. The main problems are the search for the appropriate kernel function for a particular application and their inappropriateness to handle the temporal structure of the speech signals. There are also some recent studies [35] in order to adapt the SVM to the multicategory classification problem. The SVM were already applied for speaker verification. In [23, 36], the widely used speech feature vectors were used as the input training material for the SVM.

Generally speaking, the performance of speaker verification systems based on discrimination-based learning techniques can be tuned to obtain comparable performance to the state-of-the-art GMM, and in some special experimental conditions, they could be tuned to outperform the GMM. It should be noted that, as explained earlier in this section, the tuning of a GMM baseline systems is not straightforward, and different parameters such as the training method, the number of mixtures, and the amount of speech to use in training a background model have to be adjusted to the experimental conditions. Therefore, when comparing a new system to the classical GMM system, it is difficult to be sure that the baseline GMM used are comparable to the best performing ones.

Another recent alternative to solve the speaker verification problem is to combine GMM with SVMs. We are not going to give here an extensive study of all the experiments done [37, 38, 39], but we are rather going to illustrate the problem with one example meant to exploit together the GMM and SVM for speaker verification purposes. One of the

problems with the speaker verification is the score normalization (see Section 4). Because SVM are well suited to determine an optimal hyperplan separating data belonging to two classes, one way to use them for speaker verification is to separate the likelihood client and nonclient values with an SVM. That was the idea implemented in [37], and an SVM was constructed to separate two classes, the clients from the impostors. The GMM technique was used to construct the input feature representation for the SVM classifier. The speaker GMM models were built by adaptation of the background model. The GMM likelihood values for each frame and each Gaussian mixture were used as the input feature vector for the SVM. This combined GMM-SVM method gave slightly better results than the GMM method alone. Several points should be emphasized: the results were obtained on a subset of NIST'1999 speaker verification data, only the Znorm was tested, and neither the GMM nor the SVM parameters were thoroughly adjusted. The conclusion is that the results demonstrate the feasibility of this technique, but in order to fully exploit these two techniques, more work should be done.

4. NORMALIZATION

4.1. Aims of score normalization

The last step in speaker verification is the decision making. This process consists in comparing the likelihood resulting from the comparison between the claimed speaker model and the incoming speech signal with a decision threshold. If the likelihood is higher than the threshold, the claimed speaker will be accepted, else rejected.

The tuning of decision thresholds is very troublesome in speaker verification. If the choice of its numerical value remains an open issue in the domain (usually fixed empirically), its reliability cannot be ensured while the system is running. This uncertainty is mainly due to the score variability between trials, a fact well known in the domain.

This score variability comes from different sources. First, the nature of the enrollment material can vary between the speakers. The differences can also come from the phonetic content, the duration, the environment noise, as well as the quality of the speaker model training. Secondly, the possible mismatch between enrollment data (used for speaker modeling) and test data is the main remaining problem in speaker recognition. Two main factors may contribute to this mismatch: the speaker him-/herself through the intraspeaker variability (variation in speaker voice due to emotion, health state, and age) and some environment condition changes in transmission channel, recording material, or acoustical environment. On the other hand, the interspeaker variability (variation in voices between speakers), which is a particular issue in the case of speaker-independent threshold-based system, has to be also considered as a potential factor affecting the reliability of decision boundaries. Indeed, as this interspeaker variability is not directly measurable, it is not straightforward to protect the speaker verification system (through the decision making process) against all potential impostor attacks. Lastly, as for the training material, the nature and the quality of test segments influence the value of the scores for client and impostor trials.

Score normalization has been introduced explicitly to cope with score variability and to make speaker-independent decision threshold tuning easier.

4.2. Expected behavior of score normalization

Score normalization techniques have been mainly derived from the study of Li and Porter [40]. In this paper, large variances had been observed from both distributions of client scores (intraspeaker scores) and impostor scores (interspeaker scores) during speaker verification tests. Based on these observations, the authors proposed solutions based on impostor score distribution normalization in order to reduce the overall score distribution variance (both client and impostor distributions) of the speaker verification system. The basic of the normalization technique is to center the impostor score distribution by applying on each score generated by the speaker verification system the following normalization. Let $L_\lambda(X)$ denote the score for speech signal X and speaker model λ. The normalized score $\widetilde{L}_\lambda(X)$ is then given as follows:

$$\widetilde{L}_\lambda(X) = \frac{L_\lambda(X) - \mu_\lambda}{\sigma_\lambda}, \tag{16}$$

where μ_λ and σ_λ are the normalization parameters for speaker λ. Those parameters need to be estimated.

The choice of normalizing the impostor score distribution (as opposed to the client score distribution) was initially guided by two facts. First, in real applications and for text-independent systems, it is easy to compute impostor score distributions using pseudo-impostors, but client distributions are rarely available. Secondly, impostor distribution represents the largest part of the score distribution variance. However, it would be interesting to study client score distribution (and normalization), for example, in order to determine theoretically the decision threshold. Nevertheless, as seen previously, it is difficult to obtain the necessary data for real systems and only few current databases contain enough data to allow an accurate estimate of client score distribution.

4.3. Normalization techniques

Since the study of Li and Porter [40], various kinds of score normalization techniques have been proposed in the literature. Some of them are briefly described in the following section.

World-model and cohort-based normalizations

This class of normalization techniques is a particular case: it relies more on the estimation of antispeaker hypothesis ("the target speaker does not pronounce the record") in the Bayesian hypothesis test than on a normalization scheme. However, the effects of this kind of techniques on the different score distributions are so close to the normalization method ones that we have to present here.

The first proposal came from Higgins et al. in 1991 [9], followed by Matsui and Furui in 1993 [41], for which the normalized scores take the form of a ratio of likelihoods as follows:

$$\widetilde{L}_\lambda(X) = \frac{L_\lambda(X)}{L_{\overline{\lambda}}(X)}. \quad (17)$$

For both approaches, the likelihood $L_{\overline{\lambda}}(y)$ was estimated from a cohort of speaker models. In [9], the cohort of speakers (also denoted as a cohort of impostors) was chosen to be close to speaker λ. Conversely, in [41], the cohort of speakers included speaker λ. Nevertheless, both normalization schemes equally improve speaker verification performance.

In order to reduce the amount of computation, the cohort of impostor models was replaced later with a unique model learned using the same data as the first ones. This idea is the basic of world-model normalization (the world model is also named "background model") firstly introduced by Carey et al. [13]. Several works showed the interest in world-model-based normalization [14, 17, 42].

All the other normalizations discussed in this paper are applied on world-model normalized scores (commonly named likelihood ratio in the way of statistical approaches), that is, $\widetilde{L}_\lambda(X) = \Lambda_\lambda(X)$.

Centered/reduced impostor distribution

This family of normalization techniques is the most used. It is directly derived from (16), where the scores are normalized by subtracting the mean and then dividing by the standard deviation, both estimated from the (pseudo)impostor score distribution. Different possibilities are available to compute the impostor score distribution.

Znorm

The zero normalization (Znorm) technique is directly derived from the work done in [40]. It has been massively used in speaker verification in the middle of the nineties. In practice, a speaker model is tested against a set of speech signals produced by some impostor, resulting in an impostor similarity score distribution. Speaker-dependent mean and variance—normalization parameters—are estimated from this distribution and applied (see (16) on similarity scores yielded by the speaker verification system when running. One of the advantages of Znorm is that the estimation of the normalization parameters can be performed offline during speaker model training.

Hnorm

By observing that, for telephone speech, most of the client speaker models respond differently according to the handset type used during testing data recording, Reynolds [43] had proposed a variant of Znorm technique, named handset normalization (Hnorm), to deal with handset mismatch between training and testing.

Here, handset-dependent normalization parameters are estimated by testing each speaker model against handset-dependent speech signals produced by impostors. During testing, the type of handset relating to the incoming speech signal determines the set of parameters to use for score normalization.

Tnorm

Still based on the estimate of mean and variance parameters to normalize impostor score distribution, test-normalization (Tnorm), proposed in [44], differs from Znorm by the use of impostor models instead of test speech signals. During testing, the incoming speech signal is classically compared with claimed speaker model as well as with a set of impostor models to estimate impostor score distribution and normalization parameters consecutively. If Znorm is considered as a speaker-dependent normalization technique, Tnorm is a test-dependent one. As the same test utterance is used during both testing and normalization parameter estimate, Tnorm avoids a possible issue of Znorm based on a possible mismatch between test and normalization utterances. Conversely, Tnorm has to be performed online during testing.

HTnorm

Based on the same observation as Hnorm, a variant of Tnorm has been proposed, named HTnorm, to deal with handset-type information. Here, handset-dependent normalization parameters are estimated by testing each incoming speech signal against handset-dependent impostor models. During testing, the type of handset relating to the claimed speaker model determines the set of parameters to use for score normalization.

Cnorm

Cnorm was introduced by Reynolds during NIST 2002 speaker verification evaluation campaigns in order to deal with cellular data. Indeed, the new corpus (Switchboard cellular phase 2) is composed of recordings obtained using different cellular phones corresponding to several unidentified handsets. To cope with this issue, Reynolds proposed a blind clustering of the normalization data followed by an Hnorm-like process using each cluster as a different handset.

This class of normalization methods offers some advantages particularly in the framework of NIST evaluations (text independent speaker verification using long segments of speech—30 seconds in average for tests and 2 minutes for enrollment). First, both the method and the impostor distribution model are simple, only based on mean and standard deviation computation for a given speaker (even if Tnorm complicates the principle by the need of online processing). Secondly, the approach is well adapted to a text-independent task, with a large amount of data for enrollment. These two points allow to find easily pseudo-impostor data. It seems more difficult to find these data in the case of a user-password-based system, where the speaker chooses his password and repeats it three or four times during the enrollment phase only. Lastly, modeling only the impostor distribution is a good way to set a threshold according to the global false acceptance error and reflects the NIST scoring strategy.

For a commercial system, the level of false rejection is critical and the quality of the system is driven by the quality reached for the "worse" speakers (and not for the average).

Dnorm

Dnorm was proposed by Ben et al. in 2002 [45]. Dnorm deals with the problem of pseudo-impostor data availability by generating the data using the world model. A Monte Carlo-based method is applied to obtain a set of client and impostor data, using, respectively, client and world models. The normalized score is given by

$$\widetilde{L}_\lambda(X) = \frac{L_\lambda(X)}{\mathrm{KL2}(\lambda, \overline{\lambda})}, \tag{18}$$

where $\mathrm{KL2}(\lambda, \overline{\lambda})$ is the estimate of the symmetrized Kullback-Leibler distance between the client and world models. The estimation of the distance is done using Monte-Carlo generated data. As for the previous normalizations, Dnorm is applied on likelihood ratio, computed using a world model.

Dnorm presents the advantage not to need any normalization data in addition to the world model. As Dnorm is a recent proposition, future developments will show if the method could be applied in different applications like password-based systems.

WMAP

WMAP is designed for multirecognizer systems. The technique focuses on the meaning of the score and not only on normalization. WMAP, proposed by Fredouille et al. in 1999 [46], is based on the Bayesian decision framework. The originality is to consider the two classical speaker recognitionhypotheses in the score space and not in the acoustic space. The final score is the a posteriori probability to obtain the score given the target hypothesis:

$$\mathrm{WMAP}\,(L_\lambda(X)) = \frac{P_{\mathrm{Target}} \cdot p(L_\lambda(X)|\mathrm{Target})}{P_{\mathrm{Target}} \cdot p(L_\lambda(X)|\mathrm{Target}) + P_{\mathrm{Imp}} \cdot p(L_\lambda(X)|\mathrm{Imp})}, \tag{19}$$

where P_{Target} (resp., P_{Imp}) is the a priori probability of a target test (resp., an impostor test) and $p(L_\lambda(X)|\mathrm{Target})$ (resp., $p(L_\lambda(X)|\mathrm{Imp})$) is the probability of score $L_\lambda(X)$ given the hypothesis of a target test (resp., an impostor test).

The main advantage of the WMAP[6] normalization is to produce meaningful normalized score in the probability space. The scores take the quality of the recognizer directly into account, helping the system design in the case of multiple recognizer decision fusion.

The implementation proposed by Fredouille in 1999 used an empirically approach and nonparametric models for estimating the target and impostor score probabilities.

[6]The method is called WMAP as it is a maximum a posteriori approach applied on likelihood ratio where the denominator is computed using a world model.

4.4. Discussion

Through the various experiments achieved on the use of normalization in speaker verification, different points may be highlighted. First of all, the use of prior information like the handset type or gender information during normalization parameter computation is relevant to improve performance (see [43] for experiments on Hnorm and [44] for experiment on HTnorm).

Secondly, HTnorm seems better than the other kind of normalization as shown during the 2001 and 2002 NIST evaluation campaigns. Unfortunately, HTnorm is also the most expensive in computational time and requires estimating normalization parameters during testing. The solution proposed in [45], Dnorm normalization, may be a promising alternative since the computational time is significantly reduced and no impostor population is required to estimate normalization parameters. Currently, Dnorm performs as well as Znorm technique [45]. Further work is expected in order to integrate prior information like handset type to Dnorm and to make it comparable with Hnorm and HTnorm. WMAP technique exhibited interesting performance (same level as Znorm but without any knowledge about the real target speaker—normalization parameters are learned a priori using a separate set of speakers/tests). However, the technique seemed difficult to apply in a target speaker-dependent mode, since few speaker data are not sufficient to learn the normalization models. A solution could be to generate data, as done in the Dnorm approach, to estimate the score models Target and Imp (impostor) directly from the models.

Finally, as shown during the 2001 and 2002 NIST evaluation campaigns, the combination of different kinds of normalization (e.g., HTnorm & Hnorm, Tnorm & Dnorm) may lead to improved speaker verification performance. It is interesting to note that each winning normalization combination relies on the association between a "learning condition" normalization (Znorm, Hnorm, and Dnorm) and a "test-based" normalization (HTnorm and Tnorm).

However, this behavior of current speaker verification systems which require score normalization to perform better may lead to question the relevancy of techniques used to obtain these scores. The state-of-the-art text-independent speaker recognition techniques associate one or several parameterization level normalizations (CMS, feature variance normalization, feature warping, etc.) with a world model normalization and one or several score normalizations. Moreover, the speaker models are mainly computed by adapting a world/background model to the client enrollment data which could be considered as a "model" normalization.

Observing that at least four different levels of normalization are used, the question remains: is the front-end processing, the statistical techniques (like GMM) the best way of modeling speaker characteristics and speech signal variability, including mismatch between training and testing data? After many years of research, speaker verification still remains an open domain.

5. EVALUATION

5.1. Types of errors

Two types of errors can occur in a speaker verification system, namely, false rejection and false acceptance. A false rejection (or nondetection) error happens when a valid identity claim is rejected. A false acceptance (or false alarm) error consists in accepting an identity claim from an impostor. Both types of error depend on the threshold θ used in the decision making process. With a low threshold, the system tends to accept every identity claim thus making few false rejections and lots of false acceptances. On the contrary, if the threshold is set to some high value, the system will reject every claim and make very few false acceptances but a lot of false rejections. The couple (false alarm error rate, false rejection error rate) is defined as the *operating point* of the system. Defining the operating point of a system, or, equivalently, setting the decision threshold, is a trade-off between the two types of errors.

In practice, the false alarm and nondetection error rates, denoted by P_{fa} and P_{fr}, respectively, are measured experimentally on a test corpus by counting the number of errors of each type. This means that large test sets are required to be able to measure accurately the error rates. For clear methodological reasons, it is crucial that none of the test speakers, whether true speakers or impostors, be in the training and development sets. This excludes, in particular, using the same speakers for the background model and for the tests. However, it may be possible to use speakers referenced in the test database as impostors. This should be avoided whenever discriminative training techniques are used or if across speaker normalization is done since, in this case, using referenced speakers as impostors would introduce a bias in the results.

5.2. DET curves and evaluation functions

As mentioned previously, the two error rates are functions of the decision threshold. It is therefore possible to represent the performance of a system by plotting P_{fa} as a function of P_{fr}. This curve, known as the system operating characteristic, is monotonous and decreasing. Furthermore, it has become a standard to plot the error curve on a normal deviate scale [47] in which case the curve is known as the detection error trade-offs (DETs) curve. With the normal deviate scale, a speaker recognition system whose true speaker and impostor scores are Gaussians with the same variance will result in a linear curve with a slope equal to -1. The better the system is, the closer to the origin the curve will be. In practice, the score distributions are not exactly Gaussians but are quite close to it. The DET curve representation is therefore more easily readable and allows for a comparison of the system's performances on a large range of operating conditions. Figure 6 shows a typical example of a DET curves.

Plotting the error rates as a function of the threshold is a good way to compare the potential of different methods in laboratory applications. However, this is not suited for the evaluation of operating systems for which the threshold has been set to operate at a given point. In such a case, systems are evaluated according to a cost function which takes into account the two error rates weighted by their respective costs, that is $C = C_{\text{fa}}P_{\text{fa}} + C_{\text{fr}}P_{\text{fr}}$. In this equation, C_{fa} and C_{fr} are the costs given to false acceptances and false rejections, respectively. The cost function is minimal if the threshold is correctly set to the desired operating point. Moreover, it is possible to directly compare the costs of two operating systems. If normalized by the sum of the error costs, the cost C can be interpreted as the mean of the error rates, weighted by the cost of each error.

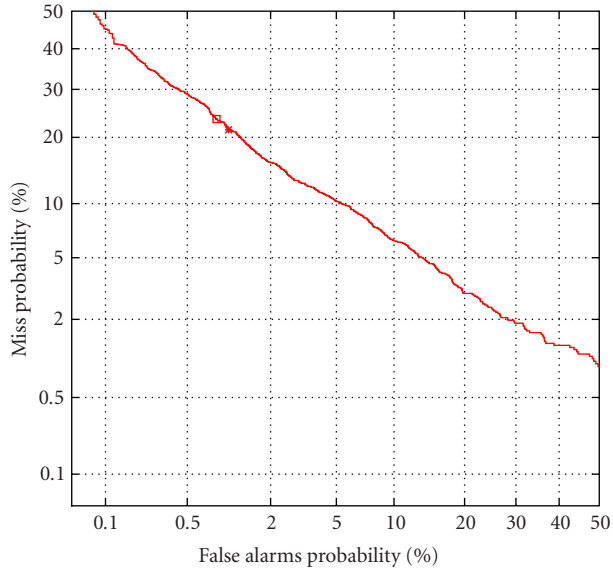


Figure 6: Example of a DET curve.

Other measures are sometimes used to summarize the performance of a system in a single figure. A popular one is the equal error rate (EER) which corresponds to the operating point where $P_{\text{fa}} = P_{\text{fr}}$. Graphically, it corresponds to the intersection of the DET curve with the first bisector curve. The EER performance measure rarely corresponds to a realistic operating point. However, it is a quite popular measure of the ability of a system to separate impostors from true speakers. Another popular measure is the half total error rate (HTER) which is the average of the two error rates P_{fa} and P_{fr}. It can also be seen as the normalized cost function assuming equal costs for both errors.

Finally, we make the distinction between a cost obtained with a system whose operating point has been set up on development data and a cost obtained with a posterior minimization of the cost function. The latter is always to the advantage of the system but does not correspond to a realistic evaluation since it makes use of the test data. However, the difference between those two costs can be used to evaluate the quality of the decision making module (in particular, it evaluates how well the decision threshold has been set).

5.3. Factors affecting the performance and evaluation paradigm design

There are several factors affecting the performance of a speaker verification system. First, several factors have an

impact on the quality of the speech material recorded. Among others, these factors are the environmental conditions at the time of the recording (background noise or not), the type of microphone used, and the transmission channel bandwidth and compression if any (high bandwidth speech, landline and cell phone speech, etc.). Second are factors concerning the speakers themselves and the amount of training data available. These factors are the number of training sessions and the time interval between those sessions (several training sessions over a long period of time help coping with the long-term variability of speech), the physical and emotional state of the speaker (under stress or ill), the speaker cooperativeness (does the speaker want to be recognized or does the impostor really want to cheat, is the speaker familiar with the system, and so forth). Finally, the system performance measure highly depends on the test set complexity: cross gender trials or not, test utterance duration, linguistic coverage of those utterances, and so forth. Ideally, all those factors should be taken into account when designing evaluation paradigms or when comparing the performance of two systems on different databases. The excellent performance obtained in artificial good conditions (quiet environment, high-quality microphone, consecutive recordings of the training and test material, and speaker willing to be identified) rapidly degrades in real-life applications.

Another factor affecting the performance worth noting is the test speakers themselves. Indeed, it has been observed several times that the distribution of errors varies greatly between speakers [48]. A small number of speakers (goats) are responsible for most of the nondetection errors, while another small group of speakers (lambs) are responsible for the false acceptance errors. The performance computed by leaving out these two small groups are clearly much better. Evaluating the performance of a system after removing a small percentage of the speakers whose individual error rates are the higher may be interesting in commercial applications where it is better to have a few unhappy customers (for which an alternate solution to speaker verification can be envisaged) than many ones.

5.4. Typical performance

It is quite impossible to give a complete overview of the speaker verification systems because of the great diversity of applications and experimental conditions. However, we conclude this section by giving the performance of some systems trained and tested with an amount of data reasonable in the context of an application (one or two training sessions and test utterances between 10 and 30 seconds).

For good recording conditions and for text-dependent applications, the EER can be as low 0.5% (YOHO database), while text-dependent applications usually have EERs above 2%. In the case of telephone speech, the degradation of the speech quality directly impacts the error rates which then range from 2% EER for speaker verification on 10 digit strings (SESP database) to about 10% on conversational speech (Switchboard).

6. EXTENSIONS OF SPEAKER VERIFICATION

Speaker verification supposes that training and test are composed of monospeaker records. However, it is necessary for some applications to detect the presence of a given speaker within multispeaker audio streams. In this case, it may also be relevant to determine who is speaking when. To handle this kind of tasks, several extensions of speaker verification to multispeaker case have been defined. The three most common ones are briefly described below.

(i) The n-speaker detection is similar to speaker verification [49]. It consists in determining whether a target speaker speaks in a conversation involving two speakers or more. The difference from speaker verification is that the test recording contains the whole conversation with utterances from various speakers [50, 51].

(ii) Speaker tracking [49] consists in determining if and when a target speaker speaks in a multispeaker record. The additional work as compared to the n-speaker detection is to specify the target speaker speech segments (begin and end times of each speaker utterance) [51, 52].

(iii) Segmentation is close to speaker tracking except that no information is provided on speakers. Neither speaker training data nor speaker ID is available. The number of speakers is also unknown. Only test data is available. The aim of the segmentation task is to determine the number of speakers and when they speak [53, 54, 55, 56, 57, 58, 59]. This problem corresponds to a blind classification of the data. The result of the segmentation is a partition in which every class is composed of segments of one speaker.

In the n-speaker detection and speaker tracking tasks as described above, the multispeaker aspect concerned only the test records. Training records were supposed to be monospeaker. An extension of those tasks consists in having multispeaker records for training too, with the target speaker speaking in all these records. The training phase then gets more complex, requiring speaker segmentation of the training records to extract information relevant to the target speaker.

Most of those tasks, including speaker verification, were proposed in the NIST SRE campaigns to evaluate and compare performance of speaker recognition methods for mono- and multispeaker records (test and/or training). While the set of proposed tasks was initially limited to speaker verification task in monospeaker records, it has been enlarged over the years to cover common problems found in real-world applications.

7. APPLICATIONS OF SPEAKER VERIFICATION

There are many applications to speaker verification. The applications cover almost all the areas where it is desirable to secure actions, transactions, or any type of interactions by identifying or authenticating the person making the transaction. Currently, most applications are in the banking

and telecommunication areas. Since the speaker recognition technology is currently not absolutely reliable, such technology is often used in applications where it is interesting to diminish frauds but for which a certain level of fraud is acceptable. The main advantages of voice-based authentication are its low implementation cost and its acceptability by the end users, especially when associated with other vocal technologies.

Regardless of forensic applications, there are four areas where speaker recognition can be used: access control to facilities, secured transactions, over a network (in particular, over the telephone), structuring audio information, and games. We briefly review those various families of applications.

7.1. On-site applications

On-site applications regroup all the applications where the user needs to be in front of the system to be authenticated. Typical examples are access control to some facilities (car, home, warehouse), to some objects (locksmith), or to a computer terminal. Currently, ID verification in such context is done by mean of a key, a badge or a password, or personal identification number (PIN).

For such applications, the environmental conditions in which the system is used can be easily controlled and the sound recording system can be calibrated. The authentication can be done either locally or remotely but, in the last case, the transmission conditions can be controlled. The voice characteristics are supplied by the user (e.g., stored on a chip card). This type of application can be quite dissuasive since it is always possible to trigger another authentication mean in case of doubt. Note that many other techniques can be used to perform access control, some of them being more reliable than speaker recognition but often more expensive to implement. There are currently very few access control applications developed, none on a large scale, but it is quite probable that voice authentication will increase in the future and will find its way among the other verification techniques.

7.2. Remote applications

Remote applications regroup all the applications where the access to the system is made through a remote terminal, typically a telephone or a computer. The aim is to secure the access to reserved services (telecom network, databases, web sites, etc.) or to authenticate the user making a particular transaction (e-trade, banking transaction, etc.). In this context, authentication currently relies on the use of a PIN, sometimes accompanied by the identification of the remote terminal (e.g., caller's phone number).

For such applications, the signal quality is extremely variable due to the different types of terminals and transmission channels, and can sometimes be very poor. The vocal characteristics are usually stored on a server. This type of applications is not very dissuasive since it is nearly impossible to trace the impostor. However, in case of doubt, a human interaction is always possible. Nevertheless, speaker verification remains the most natural user verification modality in this case and the easiest one to implement, along with PIN codes, since it does not require any additional sensors. Some commercial applications in the banking and telecommunication areas are now relying on speaker recognition technology to increase the level of security in a way transparent to the user. The application profile is usually designed to reduce the number of frauds. Moreover, speaker recognition over the phone complements nicely voice-driven applications from the technological and ergonomic point of views.

7.3. Information structuring

Organizing the information in audio documents is a third type of applications where speaker recognition technology is involved. Typical examples of the applications are the automatic annotation of audio archives, speaker indexing of sound tracks, and speaker change detection for automatic subtitling. The need for such applications comes from the movie industry and from the media related industry. However, in a near future, the information structuring applications should expand to other areas, such as automatic meeting recording abstracting.

The specificities of those types of applications are worth mentioning and, in particular, the huge amount of training data for some speakers and the fact that the processing time is not an issue, thus making possible the use of multipass systems. Moreover, the speaker variability within a document is reduced. However, since speaker changes are not known, the verification task goes along with a segmentation task eventually complicated by the fact that the number of speakers is not known and several persons may speak simultaneously. This application area is rapidly growing, and in the future, browsing an audio document for a given program, a given topic, or a given speaker will probably be as natural as browsing textual documents is today. Along with speech/music separation, automatic speech transcription, and keyword and keysound spotting, speaker recognition is a key technology for audio indexing.

7.4. Games

Finally, another application area, rarely explored so far, is games: child toys, video games, and so forth. Indeed, games evolve toward a better interactivity and the use of player profiles to make the game more personal. With the evolution of computing power, the use of the vocal modality in games is probably only a matter of time. Among the vocal technologies available, speaker recognition certainly has a part to play, for example, to recognize the owner of a toy, to identify the various speakers, or even to detect the characteristics or the variations of a voice (e.g., imitation contest). One interesting point with such applications is that the level of performance can be a secondary issue since an error has no real impact. However, the use of speaker recognition technology in games is still a prospective area.

8. ISSUES SPECIFIC TO THE FORENSIC AREA

8.1. Introduction

The term "forensic acoustics" has been widely used regarding police, judicial, and legal use of acoustics samples. This

wide area includes many different tasks, some of them being recording authentication, voice transcription, specific sound characterization, speaker profiling, or signal enhancement. Among all these tasks, forensic speaker recognition [60, 61, 62, 63, 64] stands out as far as it constitutes one of the more complex problems in this domain: the fact of determining whether a given speech utterance has been produced by a particular person. In this section, we will focus on this item, dealing with forensic conditions and speaker variability, forensic recognition in the past (speaker recognition by listening (SRL), and "voiceprint analysis"), and semi- and fully-automatic forensic recognition systems, discussing also the role of the expert in the whole process.

8.2. Forensic conditions and speaker variability

In forensic speaker recognition, the disputed utterance, which constitutes the evidence, is produced in crime perpetration under realistic conditions. In most of the cases, this speech utterance is acquired by obtaining access to a telephone line, mainly in two different modalities: (i) an anonymous call or, when known or expected, (ii) a wiretapping process by police agents.

"Realistic conditions" is used here as an opposite term to "laboratory conditions" in the sense that no control, assumption, or forecast can be made with respect to acquisition conditions. Furthermore, the perpetrator is not a collaborative partner, but rather someone trying to impede that any finding derived from these recordings could help to convict him.

Consequently, these "realistic conditions" impose on speech signals a high degree of variability. All these sources of variability can be classified [65] as follows:

(i) peculiar intraspeaker variability: type of speech, gender, time separation, aging, dialect, sociolect, jargon, emotional state, use of narcotics, and so forth;
(ii) forced intraspeaker variability: Lombard effect, external-influenced stress, and cocktail-party effect;
(iii) channel-dependent external variability: type of handset and/or microphone, landline/mobile phone, communication channel, bandwidth, dynamic range, electrical and acoustical noise, reverberation, distortion, and so forth.

Forensic conditions will be reached when these variability factors that constitute the so-called "realistic conditions" emerge without any kind of principle, rule, or norm. So they might be present constantly on a call, or else arise and/or disappear suddenly, so affecting in a completely unforeseeable manner the whole process.

The problem will worsen if we consider the effect of these variability factors in the comparative analysis between the disputed utterances and the undisputed speech controls. Factors like time separation, type of speech, emotional state, speech duration, transmission channel, or recording equipment employed acquire—under these circumstances—a preeminent role.

8.3. Forensic recognition in the past decades

Speaker recognition by listening

Regarding SRL [63, 66], the first distinctive issue to consider makes reference to the condition of familiar or unfamiliar voices. Human beings show high recognition abilities with respect to well-known familiar voices, in which a long-term training process has been unconsciously accomplished. In this case, even linguistic variability (at prosodic, lexical, grammatical, or idiolectal levels) can be comprised within these abilities. The problem here arises when approaching the forensic recognition area in which experts always deal with unfamiliar voices. Since this long-term training cannot be easily reached even if enough speech material and time are available, expert recognition abilities in the forensic field will be affected by this lack.

Nevertheless, several conventional procedures have been traditionally established in order to perform forensic SRL-based procedures, depending upon the condition (expert/nonexpert) of the listener, namely,

(1) by nonexperts: regarding nonexperts, which in the case of forensic cases include victims and witnesses, SRL refer to voice lineups. Many problems arise with these procedures, for both speakers and listeners, like size, auditory homogeneity, age, and sex; quantity of speech heard; and time delay between disputed and lineup utterances. Consequently, SRL by nonexperts is given just an indicative value, and related factors, like concordance with eyewitness, become key issues;
(2) by experts: SRL by experts is a combination of two different approaches, namely,
  (i) aural-perceptual approach which constitutes a detailed auditory analysis. This approach is organized in levels of speaker characterization, and within each level, several parameters are analyzed:
    (a) voice characterization: pitch, timbre, fullness, and so forth;
    (b) speech characterization: articulation, diction, speech rate, intonation, defects, and so forth;
    (c) language characterization: dynamics, prosody, style, sociolect, idiolect, and so forth;
  (ii) phonetic-acoustic approach which establishes a more precise and systematic computer-assisted analysis of auditory factors:
    (a) formants: position, bandwidth, and trajectories;
    (b) spectral energy, pitch, and pitch contour;
    (c) time domain: duration of segments, rhythm, and jitter (interperiod short-term variability).

"Voiceprint" analysis and its controversy

Spectrographic analysis was firstly applied to speaker recognition by Kersta, in 1962 [67], giving rise to the term "voiceprint." Although he gave no details about his research tests and no documentation for his claim ("My claim to voice pattern uniqueness then rests on the improbability that two speakers would have vocal cavity dimensions and articulator use-patterns identical enough to confound voiceprint

identification methods"), he ensured that the decision about the uniqueness of the "voiceprint" of a given individual could be compared, in terms of confidence, to fingerprint analysis.

Nevertheless, in 1970, Bolt et al. [68] denied that voiceprint analysis in forensic cases could be assimilated to fingerprint analysis, adducing that the physiological nature of fingerprints is clearly differentiated from the behavioral nature of speech (in the sense that speech is just a product of an underlying anatomical source, namely, the vocal tract); so speech analysis, with its inherent variability, cannot be reduced to a static pattern matching problem. These dissimilarities introduce a misleading comparison between fingerprint and speech, so the term voiceprint should be avoided. Based in this, Bolt et al. [69] declared that voiceprint comparison was closer to aural discrimination of unfamiliar voices than to fingerprint discrimination.

In 1972, Tosi et al. [70] tried to demonstrate the reliability of voiceprint technique by means of a large-scale study in which they claimed that the scientific community had accepted the method by concluding that "if trained voiceprint examiners use listening and spectrogram they would achieve lower error rates in real forensic conditions than the experimental subjects did on laboratory conditions."

Later on, in 1973, Bolt et al. [69] invalidated the preceding claim, as the method showed lack of scientific basis, specifically in practical conditions, and, in any case, real forensic conditions would decrease results with respect to those obtained in the study.

At the request of the FBI, and in order to solve this controversy, the National Academy of Sciences (NAS) authorized in 1976 the realization of a study. The conclusion of the committee was clear—the technical uncertainties were significant and forensic applications should be allowed with the utmost caution. Although forensic practice based on voiceprint analysis has been carried out since then [71]; from a scientific point of view, the validity and usability of the method in the forensic speaker recognition has been clearly set under suspect, as the technique is, as stated in [72], "subjective and not conclusive .... Consistent error rates cannot be obtained across different spectrographic studies." And, due to lack of quality, about 65% of the cases in a survey of 2,000 [71] remain inadequate to conduct voice comparisons.

8.4. Automatic speaker recognition in forensics

Semiautomatic systems

Semiautomatic systems refer to systems in which a supervised selection of acoustic phonetic events, on the complete speech utterance, has to be accomplished prior to the computer-based analysis of the selected segment.

Several systems can be found in the literature [66], the most outstanding are the following: (i) SASIS [73], semiautomatic speaker identification system, developed by Rockwell International in the USA; (ii) AUROS [74], automatic recognition of speaker by computer, developed jointly by Philips GmbH and BundesKriminalAmt (BKA) in Germany; (iii) SAUSI [75], semiautomatic speaker identification system, developed by the University of Florida; (iv) CAVIS [76], computer assisted voice identification system, developed by Los Angeles County Sheriff's Department, from 1985; or (v) IDEM [77], developed by Fundazione Ugo Bordoni in Rome, Italy.

Most of these systems require specific use by expert phoneticians (in order to select and segment the required acoustic phonetic events) and, therefore, suffer a lack of generalization in their operability; moreover, many of them have been involved in projects already abandoned by scarceness of results in forensics.

Automatic speaker recognition technology

As it is stated in [72], "automatic speaker recognition technology appears to have reached a sufficient level of maturity for realistic application in the field of forensic science." State-of-the-art speaker recognition systems, widely described in this contribution, provide a fully automated approach, handling huge quantities of speech information at a low-level acoustic signal processing [78, 79, 80]. Modern speaker recognition systems include features as mel frequency cepstral coefficients (MFCC) parameterization in the cepstral domain, cepstral mean normalization (CMN) or RASTA channel compensation, GMM modeling, MAP adaptation, UBM normalization, or score distribution normalization.

Regarding speaker verification (the authentication problem), the system is producing binary decisions as outputs (accepted versus rejected), and the global performance of the system can be evaluated in terms of false acceptance rates (FARs) versus miss or false rejection rates (FRRs), shown in terms of DET plots. This methodology perfectly suits the requirements of commercial applications of speaker recognition technology, and has led to multiple implementations of it.

Forensic methodology

Nevertheless, regarding forensic applicability of speaker recognition technology and, specially, when compared with commercial applications, some crucial questions arise concerning the role of the expert.

(i) Provided that the state-of-the-art recognition systems under forensic conditions produce nonzero errors, what is the real usability of them in the judicial process?
(ii) Is acceptance/rejection (making a decision) the goal of forensic expertise? If so, what is the role of judge/jury in a voice comparison case?
(iii) How can the expert take into account the prior probabilities (circumstances of the case) in his/her report?
(iv) How can we quantify the human cost related with FAR (innocent convicted) and with FRR (guilty freed)?

These and other related questions have led to diverse interpretation of the forensic evidence [81, 82, 83, 84]. In the field of forensic speaker recognition, some alternatives to the direct commercial interpretation of scores have been recently proposed.

(i) Confidence measure of binary decisions: this implies that for every verification decision, a measure of confidence of that decision is addressed. A practical implementation of this approach is the forensic automatic speaker recognition (FASR) system [72], developed at the FBI, based on standard speaker verification processing, and producing as an output, together with the normalized log LR score of the test utterance with respect to a given model, a confidence measurement associated with each recognition decision (accepted/rejected). This confidence measure is based on an estimate of the posterior probability for a given set of conditional testing conditions, and normalizes the score to a range from 0 to 100.

(ii) Bayesian approach through LR of opposite hypothesis: Bayesian approach posterior odds (a posteriori probability ratio)—assessments pertaining only to the court— are computed from prior odds (a priori probability ratio)—circumstances related with evidence—and LR (ratio between likelihood of evidence compared with H0 and likelihood of evidence compared with H1)—computed by expert [62]. In this approach, H0 stands for positive hypothesis (the suspected speaker is the source of the questioned recording), while H1 stands for the opposite hypothesis (the suspected speaker is not the source of the questioned recording). The application of this generic forensic approach to the specific field of forensic speaker recognition can be found in [85, 86] in terms of Tippet plots [87] (derived from standard forensic interpretation of DNA analysis); and a practical implementation as a complete system of the LR approach, denoted as IdentiVox [64], (developed in Spain by Universidad Politécnica de Madrid and Dirección General de la Guardia Civil) has shown to have encouraging results in real forensic approaches.

8.5. Conclusion

Forensic speaker recognition is a multidisciplinary field in which diverse methodologies coexist, and subjective heterogeneous approaches are usually found between forensic practitioners; although technical invalidity of some of these methods has been clearly stated, they are still used by several gurus in unscientific traditional practices. In this context, the emergence of automatic speaker recognition systems, producing robust objective scoring of disputed utterances, constitutes the milestone of forensic speaker recognition. This does not imply that all problems in the field are positively solved, as issues like availability of real forensic speech databases, forensic-specific evaluation methodology, or role of the expert are still open; but definitively, they have made possible a common-framework unified technical approach to the problem.

9. CONCLUSION AND FUTURE RESEARCH TRENDS

In this paper, we have proposed a tutorial on text-independent speaker verification. After describing the training and test phases of a general speaker verification system, we detailed the cepstral analysis, which is the most commonly used approach for speech parameterization. Then, we explained how to build a speaker model based on a GMM approach. A few speaker modeling alternatives have been mentioned, including neural network and SVMs. The score normalization step has then been described in details. This is a very important step to deal with real-world data. The evaluation of a speaker verification system has then been exposed, including how to plot a DET curve. Several extensions of speaker verification have then been enumerated, including speaker tracking and segmentation by speakers. A few applications have been listed, including on-site applications, remote applications, applications relative to structuring audio documents, and games. Issues specific to the forensic area have then been explored and discussed.

While it is clear that speaker recognition technology has made tremendous strides forward since the initial work in the field over 30 years ago, future directions in speaker recognition technology are not totally clear, but some general observations can be made. From numerous published experiments and studies, the largest impediment to widespread deployment of speaker recognition technology and a fundamental research challenge is the lack of robustness to channel variability and mismatched conditions, especially microphone mismatches. Since most systems rely primarily on acoustic features, such as spectra, they are too dependent on channel information and it is unlikely that new features derived from the spectrum will provided large gains since the spectrum is obviously highly affected by channel/noise conditions. Perhaps a better understanding of specific channel effects on the speech signal will lead to a decoupling of the speaker and channel thus allowing for better features and compensation techniques. In addition, there are several other levels of information beyond raw acoustics in the speech signal that convey speaker information. Human listeners have a relatively keen ability to recognize familiar voices which points to exploiting long-term speaking habits in automatic systems. While this seems a rather daunting task, the incredible and sustained increase in computer power and the emergence of better speech processing techniques to extract words, pitch, and prosody measures make these high-level information sources ripe for exploitation. The real breakthrough is likely to be in using features from the speech signal to learn about higher-level information not currently found in and complimentary to the acoustic information. Exploitation of such high-level information may require some form of event-based scoring techniques, since higher-levels of information, such as indicative word usage, will not likely occur regularly as acoustic information does. Further, fusion of systems will also be required to build on a solid baseline approach and provide the best attributes of different systems. Successful fusion will require ways to adjudicate between conflicting signals and to combine systems producing continuous scores with systems producing event-based scores.

Below are some of the emerging trends in speaker recognition research and development.

Exploitation of higher levels of information

In addition to the low-level spectrum features used by current systems, there are many other sources of speaker information in the speech signal that can be used. These include idiolect (word usage), prosodic measures, and other long-term signal measures. This work will be aided by the increasing use of reliable speech recognition systems for speaker recognition R&D. High-level features not only offer the potential to improve accuracy, they may also help improve robustness since they should be less affected by channel effects. Recent work at the JHU SuperSID workshop has shown that such levels of information can indeed be exploited and used profitably in automatic speaker recognition systems [24].

Focus on real-world robustness

Speaker recognition continues to be data driven, setting the lead among other biometrics in conducting benchmark evaluations and research on realistic data. The continued ease of collecting and making available speech from real applications means that researchers can focus on more real-world robustness issues that appear. Obtaining speech from a wide variety of handsets, channels, and acoustic environments will allow examination of problem cases and development and application of new or improved compensation techniques. Making such data widely available and used in evaluations of systems, like the NIST evaluations, will be a major driver in propelling the technology forward.

Emphasis on unconstrained tasks

With text-dependent systems making commercial headway, R&D effort will shift to more difficult issues in unconstrained situations. This includes variable channels and noise conditions, text-independent speech, and the tasks of speaker segmentation and indexing of multispeaker speech. Increasingly, speaker segmentation and clustering techniques are being used to aid in adapting speech recognizers and for supplying metadata for audio indexing and searching. This data is very often unconstrained and may come from various sources (e.g., broadcast news audio with correspondents in the field).

REFERENCES

[1] R. N. Bracewell, *The Fourier Transform and Its Applications*, McGraw-Hill, New York, NY, USA, 1965.

[2] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[3] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proc. of the Symposium on Time Series Analysis*, M. Rosenblatt, Ed., pp. 209–243, John Wiley & Sons, New York, NY, USA, 1963.

[4] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, 1968.

[5] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, The Netherlands, 1970.

[6] D. Petrovska-Delacrétaz, J. Cernocky, J. Hennebert, and G. Chollet, "Segmental approaches for automatic speaker verification," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 198–212, 2000.

[7] S. Furui, "Comparison of speaker recognition methods using static features and dynamic features," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 342–350, 1981.

[8] R. B. Dunn, D. A. Reynolds, and T. F. Quatieri, "Approaches to speaker detection and tracking in conversational speech," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 93–112, 2000.

[9] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, no. 2, pp. 89–106, 1991.

[10] A. E. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," in *Proc. International Conf. on Spoken Language Processing (ICSLP '92)*, vol. 1, pp. 599–602, Banff, Canada, October 1992.

[11] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[12] T. Matsui and S. Furui, "Similarity normalization methods for speaker verification based on a posteriori probability," in *Proc. 1st ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 59–62, Martigny, Switzerland, April 1994.

[13] M. Carey, E. Parris, and J. Bridle, "A speaker verification system using alpha-nets," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '91)*, vol. 1, pp. 397–400, Toronto, Canada, May 1991.

[14] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, vol. 2, pp. 963–966, Rhodes, Greece, September 1997.

[15] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Communication*, vol. 17, no. 1-2, pp. 109–116, 1995.

[16] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '96)*, vol. 1, pp. 81–84, Atlanta, Ga, USA, May 1996.

[17] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, vol. 2, pp. 1071–1074, Munich, Germany, April 1997.

[18] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[19] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, USA, 1973.

[20] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech, and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[21] D. A. Reynolds, *A Gaussian mixture modeling approach to text-independent speaker identification*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, Ga, USA, September 1992.

[22] M. Newman, L. Gillick, Y. Ito, D. McAllaster, and B. Peskin, "Speaker verification through large vocabulary continuous speech recognition," in *Proc. International Conf. on Spoken Language Processing (ICSLP '96)*, vol. 4, pp. 2419–2422, Philadelphia, Pa, USA, October 1996.

[23] M. Schmidt and H. Gish, "Speaker identification via support vector classifiers," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '96)*, vol. 1, pp. 105–108, Atlanta, Ga, USA, May 1996.

[24] *SuperSID Project at the JHU Summer Workshop, July-August 2002*, http://www.clsp.jhu.edu/ws2002/groups/supersid.

[25] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech, and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[26] J. Hertz, A. Krogh, and R. J. Palmer, *Introduction to the Theory of Neural Computation*, Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, Reading, Mass, USA. 1991.

[27] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan, New York, NY, USA, 1994.

[28] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[29] H. Bourlard and C. J. Wellekens, "Links between Markov models and multilayer perceptrons," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167–1178, 1990.

[30] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.

[31] J. Oglesby and J. S. Mason, "Optimization of neural models for speaker identification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '90)*, vol. 1, pp. 261–264, Albuquerque, NM, USA, April 1990.

[32] Y. Bennani and P. Gallinari, "Connectionist approaches for automatic speaker recognition," in *Proc. 1st ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 95–102, Martigny, Switzerland, April 1994.

[33] K. R. Farrell, R. Mammone, and K. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. Speech, and Audio Processing*, vol. 2, no. 1, pp. 194–205, 1994.

[34] J. M. Naik and D. Lubenskt, "A hybrid HMM-MLP speaker verification algorithm for telephone speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '94)*, vol. 1, pp. 153–156, Adelaide, Australia, April 1994.

[35] D. J. Sebald and J. A. Bucklew, "Support vector machines and the multiple hypothesis test problem," *IEEE Trans. Signal Processing*, vol. 49, no. 11, pp. 2865–2872, 2001.

[36] Y. Gu and T. Thomas, "A text-independent speaker verification system using support vector machines classifier," in *Proc. European Conference on Speech Communication and Technology (Eurospeech '01)*, pp. 1765–1769, Aalborg, Denmark, September 2001.

[37] J. Kharroubi, D. Petrovska-Delacrétaz, and G. Chollet, "Combining GMM's with support vector machines for text-independent speaker verification," in *Proc. European Conference on Speech Communication and Technology (Eurospeech '01)*, pp. 1757–1760, Aalborg, Denmark, September 2001.

[38] S. Fine, J. Navratil, and R. A. Gopinath, "Enhancing GMM scores using SVM "hints"," in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech '01)*, Aalborg, Denmark, September 2001.

[39] X. Dong, W. Zhaohui, and Y. Yingchun, "Exploiting support vector machines in hidden Markov models for speaker verification," in *Proc. 7th International Conf. on Spoken Language Processing (ICSLP '02)*, pp. 1329–1332, Denver, Colo, USA, September 2002.

[40] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '88)*, vol. 1, pp. 595–598, New York, NY, USA, April 1988.

[41] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '93)*, vol. 1, pp. 391–394, Minneapolis, Minn, USA, April 1993.

[42] G. Gravier and G. Chollet, "Comparison of normalization techniques for speaker recognition," in *Proc. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C '98)*, pp. 97–100, Avignon, France, April 1998.

[43] D. A. Reynolds, "The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '96)*, vol. 1, pp. 113–116, Atlanta, Ga, USA, May 1996.

[44] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification system," *Digital Signal Processing*, vol. 10, no. 1, 2000.

[45] M. Ben, R. Blouet, and F. Bimbot, "A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 1, pp. 689–692, Orlando, Fla, USA, May 2002.

[46] C. Fredouille, J.-F. Bonastre, and T. Merlin, "Similarity normalization method based on world model and a posteriori probability for speaker verification," in *Proc. European Conference on Speech Communication and Technology (Eurospeech '99)*, pp. 983–986, Budpest, Hungary, September 1999.

[47] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. European Conference on Speech Communication and Technology (Eurospeech '97)*, vol. 4, pp. 1895–1898, Rhodes, Greece, September 1997.

[48] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves, an analysis of individual differences in speaker recognition performances in the nist 1998 speaker recognition evaluation," in *Proc. International Conf. on Spoken Language Processing (ICSLP '98)*, Sydney, Australia, December 1998.

[49] M. Przybocki and A. Martin, "The 1999 NIST speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking," in *Proc. European Conference on Speech Communication and Technology (Eurospeech '99)*, vol. 5, pp. 2215–2218, Budpest, Hungary, September 1999.

[50] J. Koolwaaij and L. Boves, "Local normalization and delayed decision making in speaker detection and tracking," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 113–132, 2000.

[51] K. Sönmez, L. Heck, and M. Weintraub, "Speaker tracking and detection with multiple speakers," in *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech '99)*, vol. 5, pp. 2219–2222, Budpest, Hungary, September 1999.

[52] A. E. Rosenberg, I. Magrin-Chagnolleau, S. Parthasarathy, and Q. Huang, "Speaker detection in broadcast speech databases," in *Proc. International Conf. on Spoken Language Processing (ICSLP '98)*, Sydney, Australia, December 1998.

[53] A. Adami, S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 4, pp. 3908–3911, Orlando, Fla, USA, May 2002.

[54] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker based segmentation for audio data indexing," *Speech Communication*, vol. 32, no. 1-2, pp. 111–126, 2000.

[55] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '00)*, vol. 3, pp. 1423–1426, Istanbul, Turkey, June 2000.

[56] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. 2001: A Speaker Odyssey—The Speaker Recognition Workshop*, pp. 175–180, Crete, Greece, June 2001.

[57] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, vol. 2, pp. 89–92, Hong Kong, China, April 2003.

[58] D. A. Reynolds, R. B. Dunn, and J. J. McLaughlin, "The Lincoln speaker recognition system: NIST EVAL2000," in *Proc. International Conf. on Spoken Language Processing (ICSLP '00)*, vol. 2, pp. 470–473, Beijing, China, October 2000.

[59] L. Wilcox, D. Kimber, and F. Chen, "Audio indexing using speaker identification," in *Proc. SPIE Conference on Automatic Systems for the Inspection and Identification of Humans*, pp. 149–157, San Diego, Calif, USA, July 1994.

[60] H. J. Kunzel, "Current approaches to forensic speaker recognition," in *Proc. 1st ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 135–141, Martigny, Switzerland, April 1994.

[61] A. P. A. Broeders, "Forensic speech and audio analysis: the state of the art in 2000 AD," in *Actas del I Congreso Nacional de la Sociedad Española de Acústica Forense*, J. Ortega-Garcia, Ed., pp. 13–24, Madrid, Spain, 2000.

[62] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, no. 2-3, pp. 193–203, 2000.

[63] D. Meuwly, "Voice analysis," in *Encyclopaedia of Forensic Sciences*, J. A. Siegel, P. J. Saukko, and G. C. Knupfer, Eds., vol. 3, pp. 1413–1421, Academic Press, NY, USA, 2000.

[64] J. Gonzalez-Rodriguez, J. Ortega-Garcia, and J.-L. Sanchez-Bote, "Forensic identification reporting using automatic biometric systems," in *Biometrics Solutions for Authentication in an E-World*, D. Zhang, Ed., pp. 169–185, Kluwer Academic Publishers, Boston, Mass, USA, July 2002.

[65] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and S. Cruz-Llanas, "Speech variability in automatic speaker recognition systems for commercial and forensic purposes," *IEEE Trans. on Aerospace and Electronics Systems*, vol. 15, no. 11, pp. 27–32, 2000.

[66] D. Meuwly, *Speaker recognition in forensic sciences—the contribution of an automatic approach*, Ph.D. thesis, Institut de Police Scientifique et de Criminologie, Université de Lausanne, Lausanne, Switzerland, 2001.

[67] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253–1257, 1962.

[68] R. H. Bolt, F. S. Cooper, E. E. David Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, "Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes," *J. Acoust. Soc. Amer.*, vol. 47, pp. 597–612, 1970.

[69] R. H. Bolt, F. S. Cooper, E. E. David Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens, "Speaker identification by speech spectrograms: Some further observations," *J. Acoust. Soc. Amer.*, vol. 54, pp. 531–534, 1973.

[70] O. Tosi, H. Oyer, W. Lashbrook, C. Pedrey, and W. Nash, "Experiment on voice identification," *J. Acoust. Soc. Amer.*, vol. 51, no. 6, pp. 2030–2043, 1972.

[71] B. E. Koenig, "Spectrographic voice identification: A forensic survey," *J. Acoust. Soc. Amer.*, vol. 79, no. 6, pp. 2088–2090, 1986.

[72] H. Nakasone and S. D. Beck, "Forensic automatic speaker recognition," in *2001: A Speaker Odyssey—The Speaker Recognition Workshop*, pp. 139–142, Crete, Greece, June 2001.

[73] J. E. Paul et al., "Semi-Automatic Speaker Identification System (SASIS) — Analytical Studies," Final Report C74-11841501, Rockwell International, 1975.

[74] E. Bunge, "Speaker recognition by computer," *Philips Technical Review*, vol. 37, no. 8, pp. 207–219, 1977.

[75] H. Hollien, "SAUSI," in *Forensic Voice Identification*, pp. 155–191, Academic Press, NY, USA, 2002.

[76] H. Nakasone and C. Melvin, "C.A.V.I.S.: (Computer Assisted Voice Identification System)," Final Report 85-IJ-CX-0024, National Institute of Justice, 1989.

[77] M. Falcone and N. de Sario, "A PC speaker identification system for forensic use: IDEM," in *Proc. 1st ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 169–172, Martigny, Switzerland, April 1994.

[78] S. Furui, "Recent advances in speaker recognition," in *Audio- and Video-Based Biometric Person Authentification*, J. Bigun, G. Chollet, and G. Borgefors, Eds., vol. 1206 of *Lecture Notes in Computer Science*, pp. 237–252, Springer-Verlag, Berlin, 1997.

[79] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[80] A. F. Martin and M. A. Przybocki, "The NIST speaker recognition evaluations: 1996–2001," in *2001: A Speaker Odyssey—The Speaker Recognition Workshop*, pp. 39–43, Crete, Greece, June 2001.

[81] B. Robertson and G. A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, John Wiley & Sons, Chichester, UK, 1995.

[82] K. R. Foster and P. W. Huber, *Judging Science: Scientific Knowledge and the Federal Courts*, MIT Press, Cambridge, Mass, USA, 1997.

[83] I. W. Evett, "Towards a uniform framework for reporting opinions in forensic science casework," *Science & Justice*, vol. 38, no. 3, pp. 198–202, 1998.

[84] C. G. C. Aitken, "Statistical interpretation of evidence/Bayesian analysis," in *Encyclopedia of Forensic Sciences*, J. A. Siegel, P. J. Saukko, and G. C. Knupfer, Eds., vol. 2, pp. 717–724, Academic Press, NY, USA, 2000.

[85] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM)," in *2001: A Speaker Odyssey—The Speaker Recognition Workshop*, pp. 145–150, Crete, Greece, June 2001.

[86] J. Gonzalez-Rodriguez, J. Ortega-Garcia, and J.-J. Lucena-Molina, "On the application of the Bayesian approach to real forensic conditions with GMM-based systems," in *2001: A Speaker Odyssey—The Speaker Recognition Workshop*, pp. 135–138, Crete, Greece, June 2001.

[87] C. F. Tippet, V. J. Emerson, M. J. Fereday, et al., "The evidential value of the comparison of paint flakes from sources other than vehicles," *Journal of the Forensic Science Society*, vol. 8, pp. 61–65, 1968.

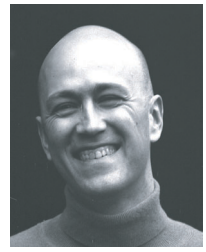
Frédéric Bimbot after graduating as a Telecommunication Engineer in 1985 (ENST, Paris, France), he received his Ph.D. degree in signal processing (speech synthesis using temporal decomposition) in 1988. He also obtained his B.A. degree in linguistics (Sorbonne Nouvelle University, Paris III) in 1987. In 1990, he joined CNRS (French National Center for Scientific Research) as a Permanent Researcher, worked with ENST for 7 years, and then moved to IRISA (CNRS & INRIA) in Rennes. He also repeatedly visited AT&T Bell Laboratories

between 1990 and 1999. He has been involved in several European projects: SPRINT (speech recognition using neural networks), SAM-A (assessment methodology), and DiVAN (audio indexing). He has also been the Work-Package Manager of research activities on speaker verification in the projects CAVE, PICASSO, and BANCA. From 1996 to 2000, he has been the Chairman of the Groupe Francophone de la Communication Parlée (now AFCP), and from 1998 to 2003, a member of the ISCA board (International Speech Communication Association, formerly known as ESCA). His research focuses on audio signal analysis, speech modeling, speaker characterization and verification, speech system assessment methodology, and audio source separation. He is heading the METISS research group at IRISA, dedicated to selected topics in speech and audio processing.

Jean-François Bonastre has been an Associate Professor at the LIA, the University of Avignon computer laboratory since 1994. He studied computer science in the University of Marseille and obtained a DEA (Master) in artificial intelligence in 1990. He obtained his Ph.D. degree in 1994, from the University of Avignon, and his HDR (Ph.D. supervision diploma) in 2000, both in computer science, both on speech science, more precisely, on speaker recognition. J.-F. Bonastre is the current President of the AFCP, the French Speaking Speech Communication Association (a Regional Branch of ISCA). He was the Chairman of the RLA2C workshop (1998) and a member of the Program Committee of Speaker Odyssey Workshops (2001 and 2004). J.-F. Bonastre has been an Invited Professor at Panasonic Speech Technology Lab. (PSTL), Calif, USA, in 2002.

Corinne Fredouille obtained her Ph.D. degree in 2000 in the field of automatic speaker recognition. She has joined the computer science laboratory LIA, University of Avignon, and more precisely, the speech processing team, as an Assistant Professor in 2003. Currently, she is an active member of the European ELISA Consortium, of AFCP, the French Speaking Speech Communication Association, and, of ISCA/SIG SPLC (Speaker and Language Characterization Special Interest Group).

Guillaume Gravier graduated in Applied Mathematics from Institut National des Sciences Appliquées (INSA Rouen) in 1995 and received his Ph.D. degree in signal and image processing from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, in January 2000. Since 2002, he is a Research Fellow at Centre National pour la Recherche Scientifique (CNRS), working at the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), INRIA, Rennes. His research interests are in the fields of speech recognition, speaker recognition, audio indexing, and multimedia information fusion. Guillaume Gravier also worked on speech synthesis at ELAN Informatique in Toulouse, France, from 1996 to 1997 and on audiovisual speech recognition at IBM Research, NY, USA, from 2001 to 2002.



Ivan Magrin-Chagnolleau received the Engineer Diploma in electrical engineering from the ENSEA, Cergy-Pontoise, France, in June 1992, the M.S. degree in electrical engineering from Paris XI University, Orsay, France, in September 1993, the M.A. degree in phonetics from Paris III University, Paris, France, in June 1996, and the Ph.D. degree in electrical engineering from the ENST, Paris, France, in January 1997. In February 1997, he joined the Speech and Image Processing Services laboratory of AT&T Labs Research, Florham Park, NJ, USA. In October 1998, he visited the Digital Signal Processing Group of the Electrical and Computer Engineering Department at Rice University, Houston, Tex, USA. In October 1999, he went to IRISA (a research institute in computer science and electrical engineering), Rennes, France. From October 2000 to August 2001, he was an Assistant Professor at LIA (the computer science laboratory of the University of Avignon), Avignon, France. In October 2001, he became a Permanent Researcher with CNRS (the French National Center for Scientific Research) and is currently working at the Laboratoire Dynamique Du Langage, one of the CNRS associated laboratories in Lyon, France. He has over 30 publications in the area of audio indexing, speaker characterization, language identification, pattern recognition, signal representations and decompositions, language and cognition, and data analysis, and one US patent in audio indexing. He is an IEEE Senior Member, a Member of the IEEE Signal Processing Society, the IEEE Computer Society, and the International Speech Communication Association (ISCA). http://www.ddl.ish-lyon.cnrs.fr/membres/imc/index.html.

Teva Merlin is currently a Ph.D. candidate at the computer science laboratory LIA at the University of Avignon.

Javier Ortega-García received the M.S. degree in electrical engineering (Ingeniero de Telecomunicación), in 1989; and the Ph.D. degree "cum laude" also in electrical engineering (Doctor Ingeniero de Telecomunicación), in 1996, both from Universidad Politécnica de Madrid, Spain. From 1999, he was an Associate Professor at the Audio-Visual and Communications Engineering Department, Universidad Politécnica de Madrid. From 1992 to 1999, he was an Assistant Professor also at Universidad Politécnica de Madrid. His research interests focus on biometrics signal processing: speaker recognition, face recognition, fingerprint recognition, online signature verification, data fusion, and multimodality in biometrics. His interests also span to forensic engineering, including forensic biometrics, acoustic signal processing, signal enhancement, and microphone arrays. He has published diverse international contributions, including book chapters, refereed journal, and conference papers. Dr. Ortega-García has chaired several sessions in international conferences. He has participated in some scientific and technical committees, as in EuroSpeech'95 (where he was also a Technical Secretary), EuroSpeech'01, EuroSpeech'03, and Odyssey'01—The Speaker Recognition Workshop. He has been appointed as General Chair at Odyssey'04—The Speaker Recognition Workshop to be held in Toledo, Spain, in June 2004.

Dijana Petrovska-Delacrétaz obtained her M.S. degree in Physics in 1981 from the Swiss Federal Institute of Technology (EPFL) in Lausanne. From 1982 to 1986, she worked as a Research Assistant at the Polymer Laboratory, EPFL. During a break to raise her son, she prepared her Ph.D. work, entitled "Study of the mechanical properties of healed polymers with different structures," that she defended in 1990. In 1995, she received a grant for women reinsertion of the Swiss National Science Foundation. That is how she started a new research activity in speech processing at the EPFL-CIRC, where she worked as a Postdoctoral Researcher until 1999. After one year spend as a Consultant in AT&T Speech Research Laboratories and another year in Ecole Nationale Superieure des Telecommunications (ENST), Paris, she worked as a Senior Assistant at the Informatics Department, Fribourg University (DIUF), Switzerland. Her main research activities are based on applications of data-driven speech segmentation for segmental speaker verification, language identification, and very low-bit speech coding. She has published 20 papers in journals and conferences, and held 3 patents.

Douglas A. Reynolds received the B.E.E. degree and Ph.D. degree in electrical engineering, both from Georgia Institute of Technology. He joined the Information Systems Technology Group at the Massachusetts Institute of Technology Lincoln Laboratory in 1992. Currently, he is a Senior Member of the technical staff and his research interests include robust speaker identification and verification, language recognition, speech recognition, and speech-content-based information retrieval. Douglas has over 40 publications in the area of speech processing and two patents related to secure voice authentication. Douglas is a Senior Member of IEEE Signal Processing Society and has served on the Speech Technical Committee.