

Chapter 2

Automatic Speaker Recognition for Forensic Case Assessment and Interpretation

Andrzej Drygajlo

Abstract Forensic speaker recognition (FSR) is the process of determining if a specific individual (suspected speaker) is the source of a questioned voice recording (trace). The forensic expert's role is to testify to the worth of the voice evidence by using, if possible, a quantitative measure of this worth. It is up to the judge and/or the jury to use this information as an aid to their deliberations and decision. This chapter aims at presenting research advances in forensic automatic speaker recognition (FASR), including data-driven tools and related methodology, that provide a coherent way of quantifying and presenting recorded voice as biometric evidence, as well as the assessment of its strength (likelihood ratio) in the Bayesian interpretation framework, compatible with interpretations in other forensic disciplines. Step-by-step guidelines for the calculation of the biometric evidence and its strength under operating conditions of the casework are provided in this chapter. It also reports on the European Network of Forensic Science Institutes (ENFSI) evaluation campaign through a fake (simulated) case, organized by the Netherlands Forensic Institute (NFI), as an example, where an automatic method using the Gaussian mixture models (GMMs) and the Bayesian interpretation (BI) framework were implemented for the forensic speaker recognition task.

2.1 Introduction

Speaker recognition is the general term used to include all of the many different tasks of discriminating one person from another based on the sound of their voices. Forensics means the use of science or technology in the investigation and establishment of facts or evidence in the court of law. The role of forensic science is the provision of information (factual or opinion) to help answer questions of importance to investigators and to courts of law. Forensic speaker recognition (FSR) is the process of determining if a specific individual (suspected speaker) is the source of a questioned voice recording (trace). This process involves the comparison of record-

A. Drygajlo (✉)
EPFL Speech Processing and Biometrics Group,
UNIL School of Criminal Justice, Swiss Federal Institute of Technology Lausanne (EPFL),
University of Lausanne (UNIL), Lausanne, Switzerland
e-mail: andrzej.drygajlo@epfl.ch

ings of an unknown voice (questioned recording) with one or more recordings of a known voice (voice of the suspected speaker) [17, 50].

There are several types of forensic speaker recognition [50, 51]. When the recognition employs any trained skill or any technologically-supported procedure, the term technical forensic speaker recognition is often used. In contrast to this, so-called naïve forensic speaker recognition refers to the application of everyday abilities of people to recognize familiar voices.

The approaches commonly used for technical forensic speaker recognition include the aural-perceptual, auditory-instrumental, and automatic methods [51]. Aural-perceptual methods, based on human auditory perception, rely on the careful listening of recordings by trained phoneticians, where the perceived differences in the speech samples are used to estimate the extent of similarity between voices [44, 45]. The use of aural-spectrographic speaker recognition can be considered as another method in this approach. The exclusively visual comparison of spectrograms in what has been called the “voiceprint” approach has come under considerable criticism in the recent years [9, 35]. The auditory-instrumental methods involve the acoustic measurements of various features such as the average fundamental frequency, articulation rate, formant centre-frequencies, etc., and comparisons of their statistical characteristics [51].

Forensic automatic speaker recognition (FASR) is an established term used when automatic speaker recognition methods are adapted to forensic applications. In automatic speaker recognition, the deterministic or statistical models of acoustic features of the speaker’s voice and the acoustic features of questioned recordings are compared [17].

2.1.1 *Forensic Automatic Speaker Recognition (FASR)*

Biometrics is the science of establishing identity of individuals based on their biological and behavioral characteristics [32]. FASR offers data-driven biometric methodology for quantitative interpretation of recorded speech as evidence. Despite the variety of characteristics, the biometric processing chain that measures biometric differences between people have essentially the same architecture and many factors are common across several biometric modalities. This generic processing chain of biometric recognition, in particular automatic speaker recognition, starts from signal sensing, passes through features extraction and their modeling and ends at the stage of features against model comparison and interpretation of similarity scores (Fig. 2.1). Biometrics based FASR, presented in this chapter, is a relatively recent application of digital speech signal processing and pattern recognition for judicial purposes and particularly law enforcement.

Results of FASR based case assessment and interpretation may be of pivotal importance at any stage of the course of justice, be it the very first police investigation or a court trial. In the police *investigative mode*, abduction, is at the root of investigation [31]. Abductive reasoning follows a process of generating likely explanations, testing these with new observations and eliminating or re-ranking the expla-

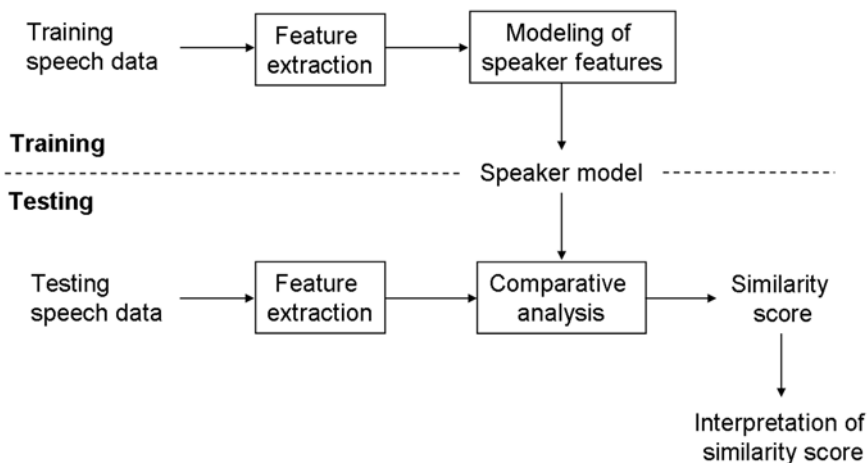


Fig. 2.1 Generic processing chain of automatic speaker recognition

nations. In this way, the investigator arrives at the best explanation of the observations, continually refining that view as further observations are made. In the forensic *evaluative mode* for a court trial, an opinion of evidential weight, based upon case specific propositions (hypotheses) and clear conditioning information (framework of circumstances) should be provided for use as evidence in court [31]. If there are two, mutually exclusive, competing propositions, exhaustive in the framework of circumstances of the case, then the odds form of Bayes' theorem can be used. The evaluative opinion of the forensic expert should be based around an assessment of a likelihood ratio of the observations given specific individual propositions (hypotheses) for the scientific findings. In the sequel of this chapter, the FASR application is limited to the evaluative mode of forensic case assessment and interpretation.

2.1.2 Overview of European Research on FASR in Case Assessment and Interpretation

The first published proposal that the likelihood-ratio framework be adopted for forensic voice comparison appears to have been made by Lewis [34]. This clearly had very little effect on the research community because it was not shown how such an idea could be implemented in practice. There was then more than decade-long time period before the idea appeared in publication again, this time showing an implementation. In April 1998 Meuwly, El-Maliki, and Drygajlo presented the paper entitled "Forensic Speaker Recognition Using Gaussian Mixture Models and a Bayesian Framework" at the COST-250 Workshop on Speaker Recognition by Man and by Machine: Directions for Forensic Applications [40]. They described the rationale for the use of the likelihood-ratio framework for forensic voice comparison, and described the design and results of tests of a Gaussian-Mixture-Model (GMM) system which calculated likelihood ratios. A substantial forensic opinion argument which

has had a greater impact on the research community was made by Champod and Meuwly, initially at the Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C) in April 1998, with a subsequent journal article published in 2000 [15]. This paper drew on the existing literature on the evaluation and interpretation of forensic evidence in fields such as DNA to make a lucid argument for its adoption in forensic voice comparison. Meuwly and Drygajlo also described the application of the likelihood-ratio framework to forensic voice comparison at the Congrès Français d'Acoustique in September 2000 [38]. At the International Speech Communication Association (ISCA) “A Speaker Odyssey, The Speaker Recognition Workshop” in June 2001, papers describing forensic automatic speaker recognition systems, using likelihood ratio and GMMs, were presented by Meuwly and Drygajlo, as well as by González-Rodríguez, Ortega-García, and Lucena-Molina [24, 39]. At ISCA's Interspeech conference in 2003 Drygajlo organized the first special session on forensic speaker recognition in the history of this conference [20, 23, 42]. Then, at Interspeech 2005 a tutorial on forensic automatic speaker recognition was presented by Drygajlo, and at Interspeech 2008 a keynote address was given by González-Rodríguez in which the likelihood-ratio framework was a central focus.

At two successive Interpol Forensic Science Symposia, in 2001 and 2004, Broeders presented reviews of developments in forensic voice comparison from 1998 to 2001 and 2001 to 2004 respectively [11, 12]. In both reports he discussed the need for forensic voice comparison evidence to be evaluated using the likelihood-ratio framework, and noted that a number of automatic systems could output likelihood ratios. At the Interpol Forensic Science Symposium in 2007, in the review on forensic audio and visual evidence, the following opinion was expressed by Jessen: “At least since 2004 forensic automatic speaker recognition has outgrown the initial developmental stages and is now a mature speech technological discipline in which there is solid knowledge about the ranges of recognition rates that can be obtained with this method. There also seems to be broad agreement as to which essential components in the three stages feature extraction, feature modelling and the calculation of distances a speaker recognition system must have as well as how the evidence is evaluated in a Bayesian approach to forensic decision making” [7, 8].

Important journal articles describing the likelihood-ratio framework and its use for the calculation of data-based likelihood ratios in forensic automatic speaker recognition were published by the European research groups in the middle of the last decade [4, 10, 17, 25, 26], and some important Ph.D. theses in the domain were completed by Meuwly in 2000 [36], Alexander in 2005 [2] and Ramos in 2007 [46]. A special chapter entitled “Forensic Evidence of Voice” by Drygajlo was introduced in the Encyclopedia of Biometrics in 2009 [19].

The 20 years, between 1984 and 2004, of pioneering research work allowed for carrying out a collaborative exercise in the Expert Working Group for Forensic Speech and Audio Analysis (FSAAWG) within the ENFSI (European Network of Forensic Science Institutes) by the Netherlands Forensic Institute (NFI), which has shown that there is increasing interest in using the automatic and auditory-instrumental, approaches to forensic voice comparison within the framework of Bayesian interpretation of forensic evidence [14, 18]. This chapter reports only on the automatic approach used for that collaborative exercise.

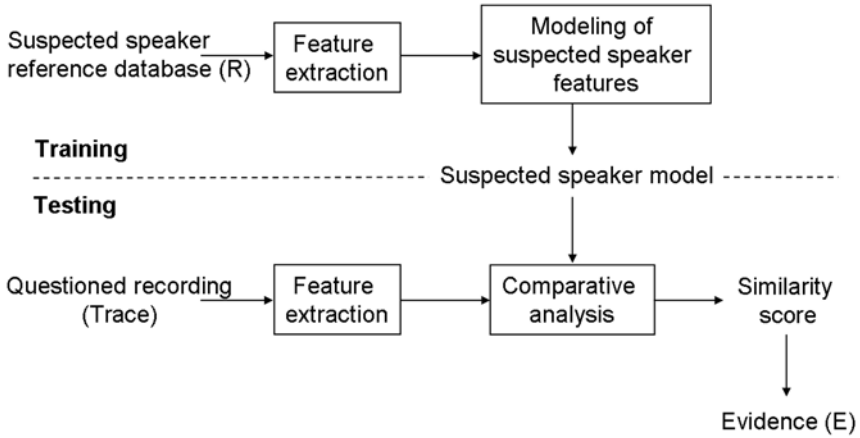


Fig. 2.2 Processing chain for calculating biometric evidence E

2.2 Voice as Biometric Evidence

The ongoing paradigm shift [41, 43, 52] in the evaluation and presentation of evidence in the forensic sciences which deal with the comparison of the quantifiable properties of samples of known and questioned origin is a shift towards requiring that evidence be evaluated and presented in a logically correct manner and that the reliability of the results be demonstrable. This approach needs biometric methods for recognition of individuals based on their biological and behavioural characteristics, as a common practice [16, 32, 53].

When using forensic automatic speaker recognition (FASR) the goal is to identify whether an unknown voice of a questioned recording (trace) came from a suspected speaker (source). Consequently, the *biometric evidence consists of the quantified degree of similarity between speaker-dependent features extracted from the trace and speaker-dependent features extracted from recorded speech of a suspect, represented by his or her model* [19, 20, 50] (Fig. 2.2).

To compute the evidence, the processing chain (Fig. 2.2) based on the generic biometric processing chain of automatic speaker recognition may be employed [20]. However, the calculated value of evidence does not allow the forensic expert alone to make an inference on the identity of the speaker.

As no ultimate set of speaker specific features is present or detected in speech, the recognition process remains in essence a statistical-probabilistic process based on models of speakers and collected data, which depend on a large number of design decisions. Information available from the acoustic features and their evidentiary value depend on the speech organs and language used [44]. The various speech organs have to be flexible to carry out their primary functions such as eating and breathing as well as their secondary function of speaking. The number and flexibility of the speech organs result in a high number of “degrees of freedom” when producing speech. These “degrees of freedom” may be manipulated at will or may be subject to variation due to external factors such as stress, fatigue, health, and so

on. The result of this plasticity of the vocal organs is that no two utterances from the same individual are ever identical in a physical sense. Moreover, no two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to this, the linguistic mechanism (language) driving the vocal mechanism is itself far from invariant. Each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, emphasis, choice of vocabulary and so on. Speaker recognition thus involves a situation where neither the physical basis of a person's speech (the vocal organs) nor the language driving it, are constant.

2.2.1 *Features*

The feature extraction module (Fig. 2.2) transforms the raw speech data into feature vectors in which speaker-specific properties are emphasized and statistical redundancies suppressed [33]. In the training mode (feature extraction and modeling modules), a suspected speaker model is created (trained) using the feature vectors. In the comparative analysis (testing) mode, the feature vectors extracted from the tested utterance (questioned recording) are compared against the suspected speaker model to give a similarity score of the evidence (E).

From the viewpoint of their physical interpretation, features commonly used for automatic speaker recognition are based on the various speech production and perception models. The speech signal continuously changes due to articulatory movements, and therefore, the signal must be broken down in short frames. Short-term spectral features, as the name suggests, are computed from short frames of about 20–30 ms in duration. Within this interval, the signal is assumed to remain stationary. These acoustic feature vectors are usually descriptors of the short-term spectral envelope which is an acoustic correlate of the resonance properties of the supralaryngeal vocal tract [6]. Thus some form of spectral envelope based features is used in most speaker recognition systems even if they are dependent on external recording conditions, e.g., Mel-Frequency Cepstral Coefficients (MFCCs) or Relative SpecTraI Perceptual Linear Prediction (RASTA-PLP) coefficients [22, 28, 29].

2.2.2 *Speaker Models*

Automatic speaker recognition systems can be text-dependent and text-independent. By using acoustic feature vectors extracted from a given speaker's training utterance, a speaker model is trained and stored into the recognition system as a reference. In text-dependent systems [27], suited for cooperative users, the model is utterance-specific and it includes the temporal dependencies between the feature vectors. In text-independent systems, there are no constraints on the words which the speakers are allowed to use [33]. Thus, the reference (what are spoken in training) and the test (what are uttered for testing) utterances may have completely different linguistic content, and the recognition system must take this phonetic mismatch into account. In forensic

applications, a text-independent automatic speaker recognition system is preferable to a text-dependent one, since the speakers can be considered non-cooperative as they do not specifically wish to be recognized. Text-independent recognition is the much more challenging of the two tasks. In general, phonetic variability represents one adverse factor to accuracy in text-independent speaker recognition.

Classical speaker models can be deterministic or statistical [13], also known as nonparametric and parametric models, respectively. In deterministic models, training and test feature vectors are directly compared with each other with the assumption that either one is an imperfect replica of the other. The amount of distortion between them represents their degree of dissimilarity. Dynamic time warping (DTW) and vector quantization (VQ) are representative examples of deterministic models for text-dependent and text-independent recognition, respectively [22].

In statistical models, each speaker is modeled as a probabilistic source with an unknown but fixed probability density function. The training phase is to estimate the parameters of the probability density function from a training sample. Comparison is usually done by evaluating the likelihood of the test utterance with respect to the model. The hidden Markov model (HMM) and the Gaussian mixture model (GMM) are the most popular statistical models for text-dependent and text-independent recognition, respectively [48].

In summary, a speaker is characterized by a speaker model such as DTW, VQ, HMM or GMM. At comparison analysis (testing), an unknown voice is first represented by a collection of feature vectors, and then evaluated against the speaker models [33].

Thus, the most persistent real-world challenge in this field is the variability of speech. There is within-speaker (within-source) variability, between-speakers (between-sources) variability and differences in recording session conditions for training and testing. Consequently, using any of the feature extraction techniques and any of the speaker models (deterministic or statistical), forensic speaker recognition methods should provide a statistical-probabilistic evaluation, which attempts to give the court an indication of the strength of the evidence, given the estimated within-source variability and the between-sources variability [20, 51], and this evaluation should be compatible with other interpretations in other forensic disciplines [21, 26, 36, 37]. The Bayesian interpretation framework, using a likelihood ratio concept, offers such interoperability. At a high level of abstraction, Bayesian data analysis is extremely simple, following the same, basic recipe: via Bayes' Theorem, we use the data to update prior beliefs about unknowns [30]. There is much to be said on the implementation of this procedure in any specific application, e.g., forensic speaker recognition, and these details are the subject of the present chapter.

2.3 Bayesian Interpretation of Biometric Evidence to Satisfy Evidentiary Requirements

To address the variability of speech, a probabilistic model [1], Bayesian inference [15] and data-driven approaches [20] appear to be adequate. In FASR statistical techniques the distribution of various features extracted from a suspect's speech is

compared with the distribution of the same features in a reference population with respect to the questioned recording. The goal is to infer the identity of a source [1], since it cannot be known with certainty.

The inference of identity can be seen as a reduction process, from an initial population to unity [37]. Recently, an investigation concerning the inference of identity in forensic speaker recognition has shown the inadequacy of the speaker verification and speaker identification (in closed set and in open set) techniques for forensic applications [15].

Speaker verification and identification are the two main automatic techniques of speech recognition used in commercial applications. When they are used for forensic speaker recognition they imply a final discrimination decision based on a threshold. Speaker verification is the task of deciding, given a sample of speech, whether a specified speaker is the source of it. Speaker identification is the task of deciding, given a sample of speech, who among many speakers is the source of it. Therefore, these techniques are clearly inadequate for forensic purposes, because they force the forensic expert to make decisions which are devolved upon the court. Consequently, the state-of-the-art speaker recognition algorithms using dynamic time warping (DTW) and hidden Markov models (HMMs) for text-dependent tasks, and vector quantization (VQ), Gaussian mixture models (GMMs), ergodic HMMs and others for text-independent tasks have to be adapted to the Bayesian interpretation framework which represents an adequate solution for the interpretation of the evidence in the judicial process [1, 49].

The court is faced with decision-making under uncertainty. In a case involving FASR it wants to know how likely it is that the speech samples of questioned recording have come from the suspected speaker. The answer to this question can be given using the Bayes' theorem and a data-driven approach to interpret the evidence [20, 49, 50].

The odds form of Bayes' theorem shows how new data (questioned recording) can be combined with prior background knowledge (prior odds) to give posterior odds for a judicial outcome (Eq. 1.1). This allows the forensic expert to revise the odds measure of uncertainty based on new information, by calculating the likelihood ratio of the evidence given the pair of competing hypotheses (propositions), e.g.: H_0 -the suspected speaker is the source of the questioned recording, H_1 -the speaker at the origin of the questioned recording is not the suspected speaker.

posterior knowledge		new data		prior knowledge	
$\frac{p(H_0 E)}{p(H_1 E)}$	=	$\frac{p(E H_0)}{p(E H_1)}$	\cdot	$\frac{p(H_0)}{p(H_1)}$	
<i>posterior odds</i>		<i>likelihood ratio</i>		<i>prior odds</i>	(1.1)
(province of the court)		(province of the expert)		(province of the court)	

This reasoning method, based on the odds form of the Bayes' theorem, allows evaluating the likelihood ratio of the evidence that leads to the statement of the degree

of support for one hypothesis against the other. As a result, the suspect's voice can be recognized as the recorded voice of the trace, to the extent that the evidence supports the hypothesis that the questioned and the suspect's recorded voices were generated by the same person (source) rather than the hypothesis that they were not the same person.

The value of a likelihood ratio depends critically on the choices one makes for describing the hypotheses. The hypotheses proposed in this section are not the only ones possible [3]. The numerator of the likelihood ratio can be considered a similarity term, and the denominator a typicality term [51]. In calculating the strength of evidence, the forensic scientist must consider not only the degree of similarity of the evidence with respect to the suspect, but also its degree of typicality with respect to the relevant population [41, 50].

The ultimate question relies on the evaluation of the probative strength of the voice evidence provided by an automatic speaker recognition method [25].

In developing an opinion based on Bayesian interpretation of evidence, the forensic expert has to utilize some form of inference process (from observations to the source). This evaluative opinion of evidential weight based upon the estimation of a likelihood ratio, should be based upon case specific propositions (hypotheses) and clear conditioning information (framework of circumstances) that is provided for evidential use in court [31].

2.4 Calculating the Strength of Biometric Evidence

The strength of the forensic evidence of voice is the result of the interpretation of the evidence, expressed in terms of the likelihood ratio given two alternative hypotheses. The principal processing chain for the interpretation of the evidence is presented in Figs. 2.3, 2.4 and 2.5 [20].

The methodological approach based on a Bayesian interpretation (BI) framework, presented in this chapter, is independent of the automatic speaker recognition method chosen, but the practical solution presented here as an example uses text-independent speaker recognition system based on Gaussian mixture model (GMM) [39].

The GMM method is not only used to calculate the evidence by comparing the questioned recording (trace) to the GMM of the suspected speaker (source), but it is also used to produce data necessary to model the within-source variability of the suspected speaker (Fig. 2.3) and the between-sources variability of the potential population of relevant speakers (Fig. 2.4), given the questioned recording. The Bayesian interpretation of the evidence consists of calculating the likelihood ratio using the probability density functions (pdfs) of the within-source and between-sources similarity scores and the single score E representing the value of evidence (Fig. 2.5).

The information provided by the analysis of the questioned recording (trace) leads to specify the initial reference population of relevant speakers (potential popu-

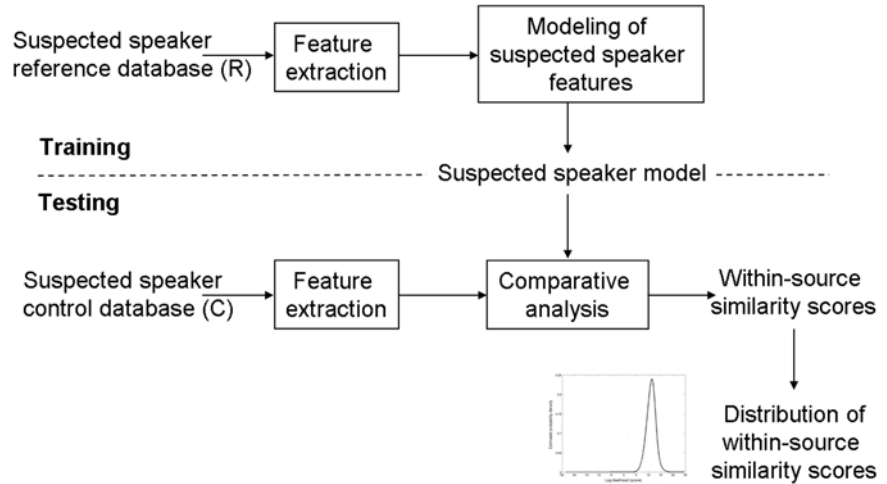


Fig. 2.3 Processing chain for calculating within-source similarity scores and their distribution

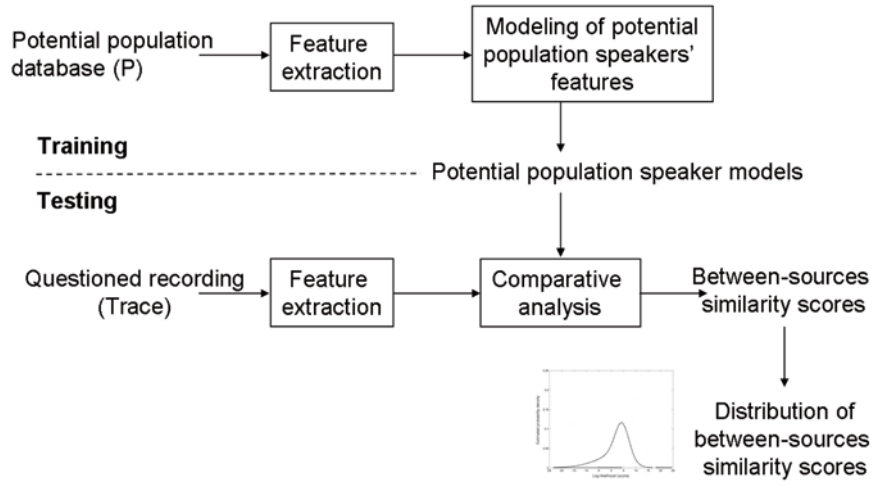
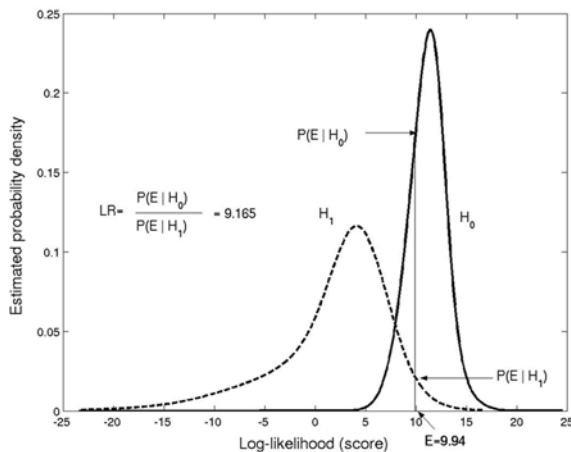


Fig. 2.4 Processing chain for calculating between-sources similarity scores and their distribution

lation) having voices similar to the trace, and, combined with the police investigation, to focus on and select a suspected speaker. The methodology presented needs three databases for the calculation and the interpretation of the evidence: the potential (relevant) population database (P), the suspected speaker reference database (R) and the suspected speaker control database (C) [39].

The potential population database (P) is a database for modeling the variability of the speech of all the potential relevant sources, using the automatic speaker recognition method. It allows evaluating the between-sources variability given the

Fig. 2.5 The likelihood ratio (LR) estimation given the value of the evidence E and the probability density functions (pdfs) of the within-source and between-sources similarity scores



questioned recording, which means the distribution of the similarity scores that can be obtained, when the questioned recording is compared to the speaker models (GMMs) of the potential population database. The calculated between-sources variability pdf is then used to estimate the denominator of the likelihood ratio $p(E|H_1)$. Ideally, the technical characteristics of the recordings (e.g. signal acquisition and transmission) should be chosen according to the characteristics analyzed in the trace.

The suspected speaker reference database (R) is recorded with the suspected speaker to model his/her speech with the automatic speaker recognition method. In this case, speech utterances should be produced in the same way as those of the P database. The suspected speaker model obtained is used to calculate the value of the evidence, by comparing the questioned recording to the model.

The suspected speaker control database (C) is recorded with the suspected speaker to evaluate her/his within-source variability, when the utterances of this database are compared to the suspected speaker model (GMM). This calculated within-source variability pdf is then used to estimate the numerator of the likelihood ratio $p(E|H_0)$. The recording of the C database should be constituted of utterances as far as possible equivalent to the trace, according to the technical characteristics, as well as to the quantity and style of speech.

The basic method proposed has been exhaustively tested in mock forensic cases corresponding to real caseworks [2, 14, 36]. In an example presented in Fig. 2.5, the strength of evidence, expressed in terms of likelihood ratio gives $LR=9.165$ for the evidence value $E=9.94$, in this case. This means that it is 9.165 times more likely to observe the score E given the hypothesis H_0 than H_1 . The important point to be made here is that the estimate of the LR is only as good as the modeling techniques and databases used to derive it. In the example, the kernel density estimation technique was used to estimate pdfs from the data representing similarity scores [1, 2].

The likelihood ratio (short form— LR) summarizes the statement of the forensic expert in the casework.

2.5 ENFSI-NFI Speaker Recognition Evaluation Through a Fake (Simulated) Case

When the Expert Working Group for Forensic Speech and Audio Analysis within the European Network of Forensic Science Institutes (ENFSI) was formed in 1997, one of its main goals was to gain insight into the different methods that are employed in the field of speaker recognition within these institutes. In 2004, a collaborative evaluation exercise was constructed at the Netherlands Forensic Institute (NFI) with English material that was recorded especially for this purpose [14]. Twelve reports were returned by the start of 2005, together with the results of all measurements that had been done and a completed questionnaire asking about the experience of the expert, the time spent on the collaborative exercise, the software that was used, etc. In this paper, the collaborative evaluation exercise is described, and a summary of the results using automatic speaker recognition method is presented based on the case report [5].

2.5.1 Formulation of the Case Key Issue

Twelve audio recordings were provided, by the Netherlands Forensic Institute (NFI) as part of a fake case evaluation, consisting of two reference recordings *R1* and *R2*, and ten questioned recordings *Q1–Q10*. The ten questioned recordings consisted of conversations between two speakers, i.e., each containing the speech of a known speaker and an unknown speaker. The two reference recordings consisted of two conversations between a known speaker and a suspected speaker.

The aim of the analysis was to determine whether the speech of the unknown speaker in each of the questioned recordings was produced by the suspected speaker in the reference recordings. The case key issue was phrased as follows: “The question in this case is whether the speaker referred to as ‘NN-male’ in the questioned material is the same person as the suspect, the speaker referred to as ‘Peter’ in the (uncontested) reference material.

2.5.2 The Case Recordings

2.5.2.1 Original Format

1 CD-ROM with 12 recordings in 16 kHz, 16-bit Linear PCM wave files were provided. According to the accompanying documentation, these recordings were recorded directly from a phone line onto a Digital Audio Tape (DAT), at 44 kHz and then down-sampled to 16 kHz and later transferred to a computer using Cool Edit Pro 2.0.2. Detailed transcriptions of the recordings with the corresponding dates, time and telephone numbers were also provided.

2.5.2.2 Preprocessing of the Case Recordings

Preprocessing, consisting of segmentation of the audio into the speech of individual speakers and removal of non-speech regions, was performed in order to prepare the recordings for the databases creation. It was ascertained by NFI, that all of the recordings provided were performed with a fixed telephone network and that there was no mobile (GSM) channel effect in the recording conditions of all the recordings in the case. Because of this, no attempt at compensating for mismatched conditions was made.

The preprocessing chain constituted of the three following steps:

- *Acquisition and Down-sampling*: Acquisition was unnecessary as the files were already in digital format. However, in order to maintain consistency with the other databases used for comparison, it was necessary to down-sample the audio files to 8 kHz, 16-bit Linear PCM files using Cool Edit Pro 2.0.
- *Segmentation*: The questioned recordings and the reference recordings were in the form of conversations between two speakers. In order to compare the speech of individual speakers it was necessary to segment each of the conversations. This segmentation was performed aurally, with the help of the transcripts provided. Zones of overlap, laughs, and other anomalies were discarded.
- *Removal of Non-Speech Regions*: The recordings were passed through a voice activity detector (VAD) [47], which separates speech and non-speech regions, using instantaneous signal to noise ratio (SNR). The non-speech regions of the recording contain information about the conditions of ambient noise present in the recording and no speaker-dependent information. Removal of these non-speech regions better allows for speaker specific characteristics to be considered, when modeling voice.

2.5.3 Databases

The methodology of Bayesian interpretation of voice as biometric evidence, presented in previous sections, was used as a means for calculating the value of evidence and its strength. This methodology requires, in addition to the questioned recordings (Q), the use of three databases: a suspected speaker reference database (R), a suspected speaker control database (C) and a potential population database (P). The set of recordings obtained, along with their durations, is presented in Table 2.1.

The two recordings R1 and R2, called by NFI reference recordings of the suspected speaker were divided into reference recordings of database R and control recordings of database C. The files R01 Peter.wav and R02 Peter.wav were further segmented into:

- two reference files R01_Peter_Ref1.wav (2 m 17 s) and R02_Peter_Ref1.wav (2 m 19 s) (R database)
- seven control recordings R01_Peter_C01.wav, R01_Peter_C02.wav, R01_Peter_C03.wav, R01_Peter_C04.wav, R02_Peter_C01.wav, R02_Peter_C02.wav and R02_Peter_C03.wav of 20 s each (C database)

Table 2.1 Individual speakers segments and their durations

No.	Source original recording	Speaker segmented recordings analyzed	Length of segmented recording (s)
1	Q1.wav	Q01_Eric.wav	169.46
2	Q1.wav	Q01_NN_Male.wav	172.28
3	Q2.wav	Q02_Eric.wav	20.73
4	Q2.wav	Q02_NN_Male.wav	11.51
5	Q3.wav	Q03_Eric.wav	91.38
6	Q3.wav	Q03_NN_Male.wav	57.59
7	Q4.wav	Q04_Eric.wav	298.23
8	Q4.wav	Q04_NN_Male.wav	279.03
9	Q5.wav	Q05_Eric.wav	25.59
10	Q5.wav	Q05_NN_Male.wav	15.86
11	Q6.wav	Q06_Eric.wav	132.09
12	Q6.wav	Q06_NN_Male.wav	88.57
13	Q7.wav	Q07_Eric.wav	10.23
14	Q7.wav	Q07_NN_Male.wav	6.39
15	Q8.wav	Q08_Eric.wav	26.62
16	Q8.wav	Q08_NN_Male.wav	15.86
17	Q9.wav	Q09_Eric.wav	32.76
18	Q9.wav	Q09_NN_Male.wav	16.89
19	Q10.wav	Q10_Eric.wav	33.53
20	Q10.wav	Q10_NN_Male.wav	18.68
21	R1.wav	R01_Jos.wav	109.01
22	R1.wav	R01_Peter.wav	432.29
23	R2.wav	R02_Jos.wav	44.79
24	R2.wav	R02_Peter.wav	197.62

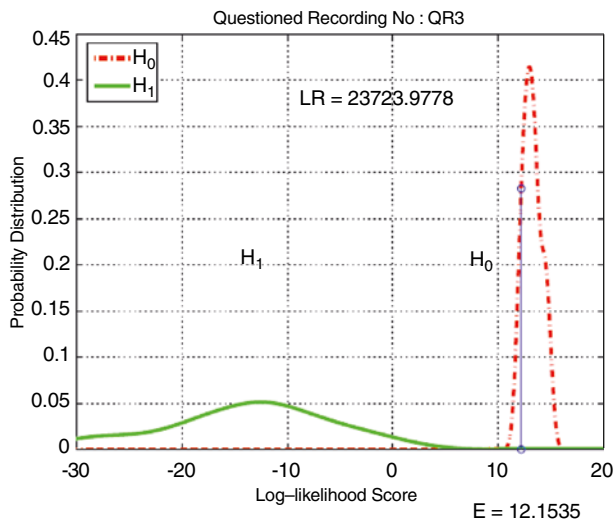
The potential population database (P) used is the PolyCOST 250 database. We have used 73 speakers from this database. This database was chosen among the available databases because it was found to be best suited to the case, especially in the language (English spoken by European speakers) and technical conditions (fixed European telephone network) under which the reference recordings of the suspect were made.

2.5.4 Calculation of the Value of Evidence and Its Strength

The following processing was then applied to each questioned recording Q_n and the R, C and P databases:

- Feature extraction and creation of models of the speakers' voices: Extraction of 12 RASTA-PLP features for each analysis frame and creation of a statistical model by means of a 64 component Gaussian mixture model (GMM)
- Calculation of the evidence score *E*: Comparison between the questioned recording Q_n and GMM of the suspected speaker (created using database R)

Fig. 2.6 The likelihood ratio (LR) estimation for Case 3 (Questioned recording Q03_NN_Male.wav)



- Within-source similarity scores: Comparison between GMM of the features of the reference recording (R) and the features of the control recordings of the suspected speaker (C)
- Between-sources similarity scores: Comparison between the features of the questioned recording Qn and the GMMs of the voices of the speakers from the database representing the potential population (P)
- Calculation of the strength of evidence: Calculation of the likelihood ratio (LR) by evaluating the relative likelihood ($p(E|H_0)/p(E|H_1)$) of observing the evidence score (E) given the hypothesis that the source of the questioned recording is the suspect (H_0) and the likelihood of observing the evidence score given hypothesis that someone else in the potential population was its source (H_1). Kernel density estimation was used to calculate the probability densities of distribution of scores for each of the hypotheses.

Each of the ten questioned recordings (Q1, Q2, ..., Q10) is considered as a separate case (Case 1, Case 2, ..., Case 10).

2.5.4.1 Example: Case 3

For Case 3 we consider the question: Is Peter in the reference recordings (R1 and R2) the same speaker as the unknown speaker in the recording Q3?

In Fig. 2.6 the distribution of scores for H_0 obtained when comparing the features of the suspected speaker control recordings (C database) of the suspected speaker, Peter, with the two statistical models of his speech (created using files from the R database) is represented by the red dotted line. The distribution of scores for H_1 obtained by comparing the segment of the questioned recording Q3, corresponding to the unknown speaker (Q03_NN_Male), with the Gaussian mix-

Table 2.2 Likelihood ratios and conclusions for all ten cases

Questioned Recording	Biometric Evidence (E)	Likelihood Ratio (LR)	Ground Truth	Conclusion Statement
Q1	10.86	6.56	different	inconclusive
Q2	11.20	163.41	same	inconclusive
Q3	12.15	23723.98	same	correct
Q4	12.68	21720.97	same	correct
Q5	13.51	11631.8	same	correct
Q6	11.63	329.0	same	correct
Q7	12.48	38407.33	same	rejected
Q8	10.68	0.660	different	inconclusive
Q9	12.92	3033.47	same	correct
Q10	7.19	4.36×10^{-23}	different	inconclusive

ture models of the speakers of the potential population database (P) is represented by the green solid line. The average score (E), represented by the point on the log-likelihood score axis in Fig. 2.6, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 12.15. A likelihood ratio of 23,723.98, obtained in Fig. 2.5, means that it is 23,723.98 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

We also observe that this score of E , is statistically significant (at a 5% statistical significance level) in the distribution of scores corresponding to hypothesis H_0 .

2.5.4.2 Example: All Cases

A summary of the results of the automatic speaker recognition for all ten cases is presented in Table 2.2. For each case we consider the question "Is the speaker, in the reference recordings R1 and R2, the same speaker as the unknown speaker in the questioned recording Qn?"

The conclusions with respect to each of the ten questioned recordings Qn have in each case been placed on the scale of conclusions that the expert uses. In Table 2.2, they are designated as correct or incorrect (the strength of the biometric evidence (E) is given by LR), inconclusive, or rejected. The results for Q3, Q4, Q5, Q6 and Q9 are correct. The latter category (rejected) includes the results for recordings that are judged to be too short (Q7), and the category inconclusive includes results that are not statistically significant (Q1, Q2, Q8, Q10). This means that the statistical significance analysis does not allow us to progress the case in any direction. The conclusions of the remaining participants of the ENFSI-NFI speaker recognition evaluation through a fake case are presented in [14] for comparison.

2.6 Summary

We discussed some important aspects of forensic speaker recognition, focusing on the necessary statistical-probabilistic framework for both quantifying and interpreting recorded voice as biometric evidence. Methodological guidelines for the calculation of the evidence and its strength under operating conditions of the casework were presented. As an example, an automatic method using the Gaussian mixture models (GMMs) and the Bayesian interpretation (BI) framework were implemented for the forensic speaker recognition task. The BI method represents neither speaker verification, nor speaker identification. These two recognition techniques cannot be used for the task, since categorical, absolute and deterministic conclusions about the identity of source of evidential traces are logically untenable because of the inductive nature of the process of the inference of identity. The method, using a likelihood ratio to indicate the strength of the biometric evidence of the questioned recording (trace), measures how this recording scores for the suspected speaker model compared to relevant non-suspect speaker models.

This chapter also reports on the first ENFSI evaluation campaign through a fake case, organized by the Netherlands Forensic Institute (NFI), as an example, where the proposed automatic method was applied. The aim of the case assessment and interpretation was to determine whether the recordings of unknown speakers, in the ten questioned recordings, were produced by the suspect (suspected speaker) present in the reference recordings. Note that the given conclusions take into consideration the likelihood ratios as well as other factors such as the length and content of the recordings and the statistical significance of the results. These factors may influence the statement of the conclusions coming from the likelihood ratio only. In such a case, the forensic expert can change the statement to inconclusive, e.g., if the statistical significance level is too low.

Statistical evaluation of voice as biometric evidence, and particularly probabilistic Bayesian methods such as calculation of likelihood ratios based on automatic speaker recognition methods, have been criticized, but they are the only demonstrably rational means of quantifying and evaluating the value of voice evidence available at the moment.

References

1. Aitken C, Taroni F (2004) *Statistics and the evaluation of evidence for forensic scientists*. Wiley, Chichester
2. Alexander A (2005) *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*. Ph.D. thesis, EPFL
3. Alexander A, Drygajlo A (2004) Scoring and direct methods for the interpretation of evidence in forensic speaker recognition. In: 8th international conference on spoken language processing (ICSLP 2004), Jeju, Korea, pp 2397–2400
4. Alexander A, Dessimoz D, Botti F, Drygajlo A (2005) Aural and automatic forensic speaker recognition in mismatched conditions. *Int J Speech Lang Law* 12(2):214–234

5. Alexander A, Drygajlo A, Botti F (2005) NFI: speaker recognition evaluation through a fake case. Case Report, EPFL-UNIL, Lausanne
6. Arcienega M, Alexander A, Zimmermann P, Drygajlo A (2005) A Bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition. InterSpeech, Lisbon
7. Association of Forensic Science Providers (2009) Standards for the formulation of evaluative forensic science opinion. *Sci Justice* 49:161–164
8. Bijhold J, Ruifrok A, Jessen M, Geradts Z, Ehrhardt S, Alberink I (2007) Forensic audio and visual evidence 2004–2007: a review. 15th INTERPOL forensic science symposium, Lyon, France
9. Bolt RH et al (1979) On the theory and practice of voice identification. National Academy of Sciences, Washington
10. Botti F, Alexander A, Drygajlo A (2004) On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition. *Forensic Sci Int* 146S:S101–S106
11. Broeders A (2001) Forensic speech and audio analysis, Forensic Linguistics 1998 to 2001: a review. 13th interpol forensic science symposium Lyon, INTERPOL, France, pp D2-53–D2-54
12. Broeders A (2004) Forensic speech and audio analysis, Forensic Linguistics 2001 to 2004: a review. 14th interpol forensic science symposium Lyon, INTERPOL, France, pp 171–188
13. Campbell J (1997) Speaker recognition: a tutorial. *Proc IEEE* 85(9):1437–1462
14. Cambier-Langeveld T (2007) Current methods in forensic speaker identification: results of a collaborative exercise. *Int J Speech Lang Law* 14(2):223–243
15. Champod C, Meuwly D (2000) The inference of identity in forensic speaker identification. *Speech Commun* 31(2–3):193–203
16. Dessimoz D, Champod C (2008) Linkages between Biometrics and Forensic Science. In: Jain A, Flynn P, Ross A eds *Handbook of biometrics*. Springer, New York, pp 425–459
17. Drygajlo A (2007) Forensic automatic speaker recognition. *IEEE Signal Process Mag* 24(2):132–135
18. Drygajlo A (2009) Statistical evaluation of biometric evidence in forensic automatic speaker recognition. In: Geradts ZJ, Franke KY, Veenman CJ eds *Computational forensics*. Springer, Berlin, pp 1–12
19. Drygajlo A (2009) Forensic Evidence of Voice. In: Li SZ ed *Encyclopedia of biometrics*. Springer, New York, pp 1388–1395
20. Drygajlo A, Meuwly D, Alexander A (2003) Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. 8th European conference on speech communication and technology (Eurospeech 2003), Geneva, Switzerland, pp 689–692
21. Evett I (1986) A Bayesian approach to the problem of interpreting glass evidence in forensic science casework. *J Forensic Sci Soc* 26(1):3–18
22. Furui S (1997) Recent advances in speaker recognition. *Pattern Recognit Lett* 18(9):859–872
23. Gfroerer S (2003) Auditory-instrumental forensic speaker recognition. 8th European Conference on Speech Communication and Technology (Eurospeech 2003). Geneva, Switzerland, pp 705–708
24. González-Rodríguez J, Ortega-García J, Lucena-Molina JJ (2001) On the application of the Bayesian framework to real forensic conditions with GMM-based systems. A Speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece
25. Gonzalez-Rodriguez J, Drygajlo A, Ramos-Castro D, Garcia-Gomar M, Ortega-Garcia J (2006) Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Comput Speech Lang* 20(2–3):331–355
26. Gonzalez-Rodriguez J, Rose P, Ramos D, Toledano DT, Ortega-Garcia J (2007) Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Trans Audio Speech Lang Process* 15(7):2104–2115
27. Hébert M (2008) Text-dependent speaker recognition. In: Benesty J, Sondhi M, Huang Y eds *Springer handbook of speech processing*. Springer, Heidelberg, pp 743–762

28. Hermansky H (1994) RASTA processing of speech. *IEEE Trans Speech Audio Process* 2(4):78–589
29. Huang X, Acero A, Hon H-W (2001) *Spoken Language Processing*. Prentice Hall PTR, Upper Saddle River
30. Jackman S (2009) *Bayesian analysis for the social sciences*. Wiley, Chichester
31. Jackson G, Jones S, Booth G, Champod C, Evett I (2006) The nature of forensic science opinion—a possible framework to guide thinking and practice in investigations and in court proceedings. *Sci Justice* 46:33–44
32. Jain AK, Flynn P, Ross AA, eds (2008) *Handbook of Biometrics*. Springer, New York
33. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. *Speech Commun* 52:12–40
34. Lewis SR (1984) Philosophy of speaker identification, police applications of speech and tape recording analysis. *Proc Inst Acoust* 6(1):69–77
35. Meuwly D (2000) Voice analysis. In: Siegel J, Knupfer G, Saukko P eds *Encyclopedia of forensic sciences*, Academic Press, London, pp 1413–1421
36. Meuwly D (2001) *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. PhD dissertation, University of Lausanne, Lausanne, Switzerland
37. Meuwly D (2006) Forensic individualisation from biometric data. *Sci Justice* 46(4):205–213
38. Meuwly D, Drygajlo A (2000) Reconnaissance automatique de locuteurs en sciences forensiques: Modélisation de la variabilité intralocuteur et interlocuteur. 5ème Congrès Français d'Acoustique, Lausanne, pp 522–525
39. Meuwly D, Drygajlo A (2001) Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM). *A Speaker Odyssey, The Speaker Recognition Workshop*, Crete, pp 145–150
40. Meuwly D, El-Maliki M, Drygajlo A (1998) Forensic speaker recognition using Gaussian mixture models and a Bayesian framework. *COST-250 workshop on speaker recognition by man and by machine: directions for forensic applications*, Ankara, Turkey, pp 52–55
41. Morrison G (2009) Forensic voice comparison and the paradigm shift. *Sci Justice* 49:298–308
42. Nakasone H (2003) Automated speaker recognition in real world conditions: controlling the uncontrollable. *European conference on speech communication and technology (Eurospeech 2003)*, Geneva, Switzerland, pp 697–700
43. National Research Council (2009) *Strengthening forensic science in the United States: a path forward*. National Academies Press, Washington
44. Nolan F (1983, reissued 2009) *The phonetic bases of speaker recognition*. Cambridge University Press, Cambridge
45. Nolan F (2001) Speaker identification evidence: its forms, limitations, and roles. *Conference on Law and Language: Prospect and Retrospect*, Levi (Finnish Lapland), pp 1–19
46. Ramos Castro D (2007) *Forensic Evaluation of the Evidence using Automatic Speaker Recognition Systems*. Ph.D. thesis, Universidad Autonoma de Madrid, Madrid, Spain
47. Renevey P, Drygajlo A (2001) Entropy based voice activity detection in very noisy conditions. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, pp 1887–1890
48. Reynolds D, Quatieri T, Dunn R (2000) Speaker verification using adapted Gaussian mixture models. *Digital Signal Process* 10(1):19–41
49. Robertson B, Vignaux G (1995) *Interpreting evidence. Evaluating forensic science in the courtroom*. Wiley, Chichester
50. Rose P (2002) *Forensic speaker identification*. Taylor and Francis, London
51. Rose P (2006) Technical forensic speaker recognition: evaluation, types and testing of evidence. *Comput Speech Lang* 20(2–3):159–191
52. Saks MJ, Koehler JJ (2005) The coming paradigm shift in forensic identification science. *Science* 309:892–895
53. Wayman J et al (eds) (2005) *Biometric systems: technology, design and performance evaluation*. Springer, New York



<http://www.springer.com/978-1-4614-0262-6>

Forensic Speaker Recognition
Law Enforcement and Counter-Terrorism
Neustein, A.; Patil, H.A. (Eds.)
2012, XXI, 540 p., Hardcover
ISBN: 978-1-4614-0262-6