Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer,
Isolde Wagner, Jos Vermeulen and Tuija Niemi

Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition

including
Guidance on the Conduct of Proficiency Testing and Collaborative Exercises

2015

With the financial support of the Prevention of and Fight against Crime Programme
European Commission - Directorate-General Home Affairs

A project funded by the EU ISEC 2011
Agreement Number: HOME/2011/ISEC/MO/4000002384

This project has been funded with support of the European Commission. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

Preface

This publication is the result of the European project “Methodological guidelines for semiautomatic and automatic speaker recognition for case assessment and interpretation”, chaired by Andrzej Drygajlo. It consists of two parts. Part 1 contains Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition and Part 2 contains Guidance on the Conduct of Proficiency Testing and Collaborative Exercises for Forensic Semiautomatic and Automatic Speaker Recognition. This project has been conducted in the framework of the ENFSI Monopoly Programme 2011 "Improving Forensic Methodologies across Europe (IFMAE)" within the context of the ENFSI Forensic Speech and Audio Analysis Working Group (FSAAWG).

The Leading Team of this project consists of:

Andrzej Drygajlo (Chair), EPFL and UNIL Institute of Forensic Science Switzerland
Stefan Gfroerer, Bundeskriminalamt Germany
Michael Jessen, Bundeskriminalamt Germany
Tuija Niemi, National Bureau of Investigation Finland
Dawid Niemiec, Central Forensic Laboratory of the Police Poland
Jos Vermeulen, Netherlands Forensic Institute
Isolde Wagner, Bundeskriminalamt Germany

As part of the project, several meetings were held across Europe (Opening Seminar in Lausanne, May 2013; Expert Workshop in Wiesbaden, November 2014; Dissemination Conference in Warsaw, September 2015) in which experts in the development and practical application of forensic speaker recognition were invited and asked for their input on various stages of the two documents. We are very grateful for their insightful comments. We also like to thank Rudolf Haraksim for his valuable contribution to this publication on issues of validation.

Contents

Part 1:

Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition

Part 2:

Guidance on the Conduct of Proficiency Testing and Collaborative Exercises for Forensic Semiautomatic and Automatic Speaker Recognition

Part 1:

Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition

1 Aims

Speaker recognition in general is the term used to denote the different ways of discriminating one person from another based on speech.

Forensic speaker recognition is speaker recognition adapted to the limitations of forensic speech material and the specific needs of reporting to and interacting with the court.

Forensic automatic speaker recognition (FASR) and forensic semiautomatic speaker recognition (FSASR) are established terms used when automatic and partially automatic speaker recognition methods are adapted to and used within a forensic setting.

The main purpose of the present document is to provide forensic experts of the Forensic Speech and Audio Analysis Working Group (FSAAWG) in the European Network of Forensic Science Institutes (ENFSI) with guidelines for best practice in FASR and FSASR for case assessment and interpretation.

Best-practice FASR and FSASR are embedded within the Bayesian interpretation framework. The task of the forensic expert in FASR and FSASR is to report to the court, given defined speech evidence and hypotheses, an indication of the strength of the evidence based on the analysis and comparison of the recordings used in the casework. This strength-of-evidence statement is made in the form of a likelihood ratio (LR).

The aims of this document are as follows:

- To provide a methodology for interpreting recorded speech as scientific evidence and establishing quantitative measures of evidence.

- To review existing methods of pre-processing, feature extraction, feature modelling and similarity scoring to be used for any specific speech data of the case, and to indicate how those methods are implemented for calculating a likelihood ratio as the strength of evidence.

- To define an interpretation framework based on Bayes' theorem and the likelihood ratio to be used in the FASR and FSASR domain independently of the baseline speaker recognition system.

- To establish methodology for the use of FASR and FSASR methods under operating conditions of casework, in particular, to handle the data involved in a case, to measure performance of a chosen method in specific conditions for evaluation and validation purposes, and to combine the outputs of FASR and FSASR with outputs of other methods (e.g., auditory-phonetic and acoustic-phonetic).

In line with the aims of this document, the objectives of respective chapters are:

- To delimit methodological aspects in speaker recognition that fall within the scope of these guidelines from those that are not addressed (chapter 2).

- To present a universal methodology of FASR and FSASR that can be applied independently of the specifics of a particular forensic speaker recognition case and independently of the specific methods used (chapter 3).

- To show how likelihood ratio based methods in FASR and FSASR can be validated according to validation criteria that are based upon performance characteristics and metrics (chapter 4).

- To address how each individual case can be described and assessed with respect to the applicability and application of FASR and FSASR methods (chapter 5).

- To provide guidance on how evaluation of the strength of evidence (likelihood ratio) under operating conditions of the case can be interpreted and reported to the court (chapter 6).

- To make recommendations on how the FASR or FSASR based case assessment and interpretation can be documented and reported to the mandating authority or party (chapter 7).

- To address best practice aspects regarding quality assurance in FASR and FSASR (chapter 8).

2 Scope

This document provides forensic experts with methodological guidelines for best practice in forensic automatic and semiautomatic speaker recognition. It does not recommend particular methods. The focus is on methodology, not methods.

2.1 Evaluative and investigative modes

FASR and FSASR are an application of speech signal processing, pattern recognition, machine learning and some elements of acoustic phonetics for legal and law enforcement purposes. They can be used in two different modes: evaluative and investigative.

- **Evaluative mode**. In the forensic evaluative mode for a court trial, a statement of evidential weight (strength of evidence), based upon case-specific hypotheses (propositions) and conditioning information, should be provided in court. The evaluative statement of the forensic expert should be based upon the assignment of a likelihood ratio (LR) of the observations given specific hypotheses (propositions) for the scientific findings.

- **Investigative mode**. In the police investigative mode, likely explanations of given observations are generated, these are tested based on new observations, and previous explanations are discarded or revised accordingly. In this cyclic manner, the investigator ultimately arrives at the best explanation of the observations.

In this document, the application is limited to the evaluative mode of forensic case assessment and interpretation.

2.2 Current approaches in forensic speaker recognition

A distinction can be made between naïve speaker recognition and technical speaker recognition. Naïve speaker recognition involves the application of intuitive abilities of listeners to recognise speakers, either familiar or unfamiliar to them. Technical speaker recognition involves the use of speech processing methods or phonetic or linguistic methods in order to arrive at findings relevant to the speaker recognition task.

One type of technical speaker recognition is the auditory-phonetic-and-acoustic-phonetic approach. This approach takes into account a broad range of speaker characteristics with proven speaker discriminatory information. Within that approach, auditory-phonetic methods are based on the auditory examination of recordings by trained phoneticians or linguists. Auditory speaker characteristics examined in recordings include auditory voice quality, filled pauses and other non-fluent behaviour, grammatical and lexical properties, regional features and several more. Acoustic-phonetic methods involve the acoustic measurement of various acoustic parameters such as mean fundamental frequency, formant centre frequencies and articulation rate. Based on auditory-phonetic-and-acoustic-phonetic analysis, similarity between questioned speaker and suspected speaker recordings as well as typicality of the findings in a relevant population are determined or estimated qualitatively or quantitatively.

Forensic automatic speaker recognition (FASR) and forensic semiautomatic speaker recognition (FSASR) is another type of technical speaker recognition, where automatic or partially automatic speaker recognition methods in their central processing (feature extraction, feature modelling, similarity scoring and calculation of likelihood ratio) are adapted to forensic applications.

Only technical speaker recognition and specifically only FASR and FSASR fall within the scope of this document.

When a forensic expert uses FASR or FSASR as well as the auditory-phonetic-and-acoustic-phonetic method, the findings from both should be reflected in the final conclusion.

2.3 Bayesian interpretation framework

The findings of FASR and FSASR can be interpreted within the Bayesian interpretation framework. This framework is based on the odds form of Bayes' theorem, which represents the following relationship:

$$\underbrace{\frac{P(H_0 \mid E)}{P(H_1 \mid E)}}_{\substack{\text{posterior knowledge} \\ \textit{posterior odds} \\ \text{(province of the court)}}} = \underbrace{\frac{p(E \mid H_0)}{p(E \mid H_1)}}_{\substack{\text{new data} \\ \textit{likelihood ratio} \\ \text{(province of the expert)}}} \times \underbrace{\frac{P(H_0)}{P(H_1)}}_{\substack{\text{prior knowledge} \\ \textit{prior odds} \\ \text{(province of the court)}}}$$

The odds form of Bayes' theorem shows how new data, reported as likelihood ratio (LR), can be combined with prior knowledge about the case (knowledge unrelated to the speech data) in order to arrive at posterior odds. Only the LR is the province of the forensic expert; the prior odds and posterior odds are the province of the court. The LR expresses the likelihood of the speech evidence under the two competing hypotheses, e.g.:

- H_0 - the suspected speaker is the source of the questioned recording
- H_1 - the suspected speaker is not the source of the questioned recording

With these hypotheses, the numerator of the LR, i.e., likelihood $p(E|H_0)$, corresponds to a numerical statement about the degree of similarity of the evidence with respect to the suspect and the denominator of the LR, i.e., likelihood $p(E|H_1)$, corresponds to a numerical statement about the degree of typicality with respect to the relevant population. The LR is the ratio between these two statements. With LR values larger than unity there is stronger support for the H_0 hypothesis and with LR values smaller than unity there is stronger support for the H_1 hypothesis.

The methodology using the Bayesian interpretation framework concerns speech pre-processing, feature extraction, feature modelling (to create speaker models), similarity scoring and calculation of a likelihood ratio (LR). The observed speech evidence (E) can be defined on the level of feature extraction or similarity scoring. Based on this choice, calculation of a LR can be performed by the "direct method" or the "scoring method", respectively.

The value of a likelihood ratio critically depends on the choices one makes for stating the observed speech evidence (E) and the hypotheses H_0 and H_1, with corresponding models of within-speaker and between-speaker speech variability. It also depends on several aspects of the speech analysis, including the non-automatic aspects of the data preparation stage, the kind of features, feature models and similarity scoring used, as well as the databases employed in the process. Consequently, the methodology using the Bayesian interpretation framework concerns also performance evaluation of LR methods for method validation and case-specific evaluation purposes.

The LR is a statement about strength of evidence. This differs from methods mostly found in the field of non-forensic speaker recognition (identification and verification), in which categorical decisions are made.

Such methods or other methods incompatible with the Bayesian interpretation framework fall outside the scope of this document.

2.4 Forensic conditions

Automatic speaker recognition systems can be very successful in discriminating between speakers when recording conditions are well controlled. In forensic settings, however, the conditions in which questioned speaker recordings are made remain largely uncontrolled and can constitute a challenge to methods in FASR and FSASR. Within-speaker variability of speech patterns (e.g., state of health, emotions) represents an undesirable factor (also called "intrinsic variability"). Technical and environmental effects such as variations in microphones, recording devices or background noise and speech signal distortion can introduce further variability beyond the within-speaker variability itself (also called "extrinsic variability"). Certain levels of variability of these sources are present in all cases, and there are differences in the combination of these factors between recordings. Beyond a certain level of those differences the term "mismatched conditions" is used. Addressing such mismatched conditions is an important task in FASR and FSASR. Mismatch may also include differences in the language spoken in the recordings or long-term temporal differences (non-contemporaneous speech).

Methods of automatic and semiautomatic speaker recognition that do not take into account the specific challenges of forensically realistic material are not covered by these guidelines.

3 Methodology of FASR and FSASR

3.1 Definitions

Forensic Automatic Speaker Recognition (FASR) refers to a method (or group of methods) of forensic speaker recognition that in its central processing stages operates automatically. The central processing stages consist of at least feature extraction, feature modelling, similarity scoring and LR computation. Pre-processing stages, such as voice activity detection, may also be included into the realm of automatic processing, but this is not a necessary requirement of FASR. FASR usually contains some level of manual input and supervision by a forensic expert. Involvement of the forensic expert is also necessary at the post-processing stages (i.e., the interpretation of the results) where the system outputs the final results in terms of likelihood ratios.

Forensic Semiautomatic Speaker Recognition (FSASR) refers to a method (or group of methods) of forensic speaker recognition that in its central processing stages operates partially automatically and partially with human intervention (henceforth "manually"). Specifically, manual processing occurs at the feature extraction level whereas automatic processing occurs at the feature modelling, similarity scoring and LR computation levels. Based on this characterisation, the crucial difference between FASR and FSASR lies at the feature extraction level: it operates automatically in FASR but manually in FSASR. The features typically used in FSASR derive from linguistics and phonetics in general and most commonly from acoustic phonetics. The input of the forensic expert at the feature extraction level includes the auditory identification and phonetic-phonological classification of sounds or prosodic units, their segmentation in the signal (location of beginning, end, centre, or other events), as well as the examination and, if necessary, correction of the results of, for example, automatic formant or pitch tracking routines.

3.2 Pre-processing

FASR and FSASR require a pre-processing stage. Although automatic procedures exist and might be used for some aspects of pre-processing (e.g., voice activity detection), usually most aspects cannot be automatised with sufficient accuracy and require the supervision and action of the forensic expert. Pre-processing includes the following tasks:

Audio format analysis, conversion, and digitisation: If necessary, submitted audio material has to be digitised or converted into an audio format that corresponds to the requirements of the FASR or FSASR method.

Speaker separation (also known as speaker diarisation): The speech of the relevant speaker (questioned or suspected) has to be separated from the speech of the irrelevant speaker if more than one speaker occurs in a recording.

Removal of pauses (also known as voice activity detection): Most FASR methods and some FSASR methods require the removal of speech pauses, which has to be ensured prior to the analysis.

Removal of local technical disturbances: Strong local disturbances in terms of background noise, signal distortions and similar other problems can interfere in the FASR and FSASR analysis and should be removed.

Removal of non-speech vocalisations and unusual speaking styles: Audio recordings may contain non-speech events such as laughter, cough, throat clearing, clicking, and breathing or unusual speaking styles such as falsetto voice, whispering, and shouting. The forensic expert can decide whether to remove these portions from the signal.

For any of the previous selection and removal processes, they should be documented using annotation files (label files) that are temporally aligned with the audio files. If speaker separation or pause removal is performed automatically, it should remain transparent (and if necessary correctible) which parts of the input file have been removed and which ones remain. Instead of altering the audio recording by removing sections it is also possible to leave the recording unaltered and use the labelling information to guide the analysis to the relevant portions, e.g., to those that are voice active, contain the relevant speaker and so forth.

Audio enhancement: If recommended by a particular method, audio enhancement may be used. It has to be shown in method validations that the effect of using enhancement improves performance of the FASR or FSASR system. It also has to be taken into account that in a legal context audio enhancement might be considered as manipulation of the observed evidence.

3.3 Features

Feature extraction is the first step in the central processing stage of FASR and FSASR. Although some features are typically used in FASR, such as MFCCs (Mel Frequency Cepstral Coefficients), and some features are typically used in FSASR (such as vowel formants), there are grey areas. For example, a FASR method might include the automatic (unsupervised) segmentation of vowels and tracking of vowel formants. Consequently, the features addressed here are not divided into the categories FASR and FSASR. The following selection presents the most frequently used features in FASR and FSASR.

Short-term spectral envelope features. Short-term spectral envelope features focus on the spectral shape of the speech signal as derived over a short portion (frame) of the signal. They aim at examining the influence of the vocal tract (frame by frame) while ignoring the influence of the voice source, in particular fundamental frequency. Short-term spectral envelope features are the most commonly used type of features in both FASR and FSASR. In FASR, short-term spectral envelope features frequently used are MFCCs (Mel Frequency Cepstral Coefficients), PLPCCs (Perceptual Linear Prediction Cepstral Coefficients) and LPCCs (Linear Prediction Cepstral Coefficients). In FSASR, short-term spectral envelope features occur in terms of formant frequencies and bandwidths, either on a vowel-specific basis or as long-term formants. Short-term spectral envelope features can be extracted globally (applying to the entire recording minus portions removed due to pre-processing) or locally (applying to certain segmented phonological categories such as vowels, fricatives or nasals). They can also be extracted statically and dynamically. Dynamic applications include the curve fitting of formant movements in diphthongs (or of formant-pattern trajectories generally, including in monophthongs) or the application of velocity and acceleration information in MFCCs calculated across multiple frames.

Fundamental frequency (f0). Included in this category are distribution parameters such as f0-average (mean, median and mode), f0-variability (standard deviation or the coefficient of variation, i.e., standard deviation divided by mean) or other parameters such as base level (i.e., various methods of capturing the lowest end of the f0 distribution) as well as the skewness and kurtosis of the distribution. Dynamics of f0 can be captured by velocity and acceleration features or curve-fitting methods of the intonation curve or the local tonal curve in tone languages. It is possible to combine short-term spectral envelope features with fundamental frequency features at the speaker modelling stage.

Amplitude. The smoothed amplitude curve can be treated in similar ways as the f0-curve, but is used less frequently due to irrelevant speaker-external influences on the amplitude values.

Duration. A duration-based feature is Articulation Rate (AR), which can be measured as the inverse of average syllable duration per recording excluding pauses. Parameters such as the percentage of the voiced or vocalic portions of a syllable are duration parameters in the domain of speech rhythm.

N-grams. The output from a phone (i.e., sound) recogniser or word recogniser can be used to construct N-grams, i.e., models of the sequencing of these units in groups of two (bigrams), three (trigrams) etc., and these N-grams can be used as speaker-discriminatory features.

The features listed here are predominantly acoustic. Further features which are based on auditory analysis or linguistics can increase the speaker-discriminatory power of the analysis. They can be covered by the auditory-phonetic-and-acoustic-phonetic approach (2.2) or by FASR or FSASR if they can be captured in terms of likelihood ratios.

3.4 Speaker modelling and similarity scoring

Feature modelling, similarity scoring and LR calculation stages follow the feature extraction stage in the central processing chain of speaker recognition. The value of a likelihood ratio depends critically on the choices one makes for describing the hypotheses, corresponding feature models, similarity scoring and comparative analysis for calculating likelihood ratios.

In text-dependent systems the speaker model is utterance-specific, i.e., it takes into account dependencies between the feature vectors that are tied to the particular words and their succession found in the utterance. Text-independent systems, on the other hand, are insensitive to the specific words and their sequencing occurring in an utterance. This means that text-independent systems must cope with the mismatch in linguistic-phonetic content that comes along with text-independence. The most widely used statistical model in text-dependent speaker recognition is the hidden Markov model (HMM) and the most widely used one in text-independent speaker recognition is the Gaussian mixture model (GMM). Deterministic models of dynamic time warping (DTW) and vector quantisation (VQ) can also be used in text-dependent and text-independent systems, respectively.

GMMs are the foundation of both the more classical GMM-UBM approach and of more recent approaches based on supervectors and i-vectors. In the supervector approach, the model parameters of the GMM are projected into high-dimensional space. The i-vector approach provides methods to reduce the dimensionality of supervectors while maintaining the essential speaker information. In order to calculate a likelihood ratio, GMMs or their super- or i-vector equivalent are needed not only for creating a suspected speaker model associated with the H_0 hypothesis, but also for creating a model of the relevant population associated with the H_1 hypothesis. The relevant population is often represented by the UBM (Universal Background Model).

GMMs have been used in both FASR and FSASR. In FSASR, the method of MVKD (Multivariate Kernel Density) is also frequently used. With the MVKD method, questioned speaker and suspected speaker are modelled with single Gaussians, whereas the relevant population is modelled with a Kernel Density function. Single Gaussians can be sufficient because of localised feature extraction, i.e., limited to specific vowel or consonant categories.

When comparing a questioned speaker recording with a suspected speaker recording, one commonly used procedure is to create a feature model of the suspected speaker and compare that model with the feature vectors of the questioned speaker ("data-to-model" type of comparison). It is also possible to calculate models for both questioned speaker and suspected speaker and to compare the models ("model-to-model" type of comparison). Such comparisons (data-to-model and model-to-model) allow obtaining similarity scores or directly likelihoods when using statistical models, e.g., GMMs or i-vectors.

3.5 Calculation of likelihood ratio (LR)

The calculation of a likelihood ratio (LR) depends on the speaker models and the similarity scoring used, as well as on the choice one makes for stating the hypotheses. There are two main methods commonly used for this purpose: the scoring method and the direct method.

3.5.1 Scoring method

The methodological approach based on the scoring method is independent of the features and speaker models chosen. This methodology needs a two-

stage processing. The first stage consists of calculating scores using any type of feature vectors (multivariate data) as well as any of the speaker models (e.g., VQ, GMM, i-vector) and similarity scoring. The second stage transforms the obtained similarity scores into two univariate distributions (probability density functions). The values of these functions represent likelihoods of scores given chosen hypotheses. This two-stage processing can be done in many ways depending on the choices one makes for stating the hypotheses.

If the hypotheses are stated as follows:

- H_0 - the suspected speaker is the source of the questioned recording,
- H_1- the suspected speaker is not the source of the questioned recording,

the H_1 hypothesis can be represented by the between-source distribution of similarity scores that result from comparing the feature vectors or model of the questioned speaker with the ones of several other speakers from the relevant population database, and the H_0 hypothesis can be represented by the within-source distribution of similarity scores as the result of comparing the feature vectors or model of the suspected speaker using its control database with the ones of the same speaker using its reference database.

In this case, the strength of evidence is represented by the likelihood ratio (LR) calculated as the ratio of likelihoods obtained from the distributions of within-source and between-source similarity scores for the single score E representing the value of the observed speech evidence. This score is obtained by comparing the feature vectors or model of the questioned speaker recording with the ones of the suspected speaker using the suspected speaker reference database.

If the hypotheses are stated as follows:

- H_0 - the suspected speaker recording and the questioned recording have the same source,
- H_1- the suspected speaker recording and the questioned recording have different sources,

the H_0 hypothesis can be represented by the distribution of similarity scores that result from same-speaker comparisons, and the H_1 hypothesis can be represented by the distribution of similarity scores as the result of different-speaker comparisons using the relevant population database in both cases.

In this respect, the LR, as the strength of evidence, can be calculated by dividing the likelihoods from the distributions of the same-source and different-source similarity scores for the single score E representing the

value of the observed speech evidence. This score is obtained by comparing the feature vectors or model of the questioned speaker recording with the ones of the suspected speaker. The scoring method is illustrated in Figure 1.

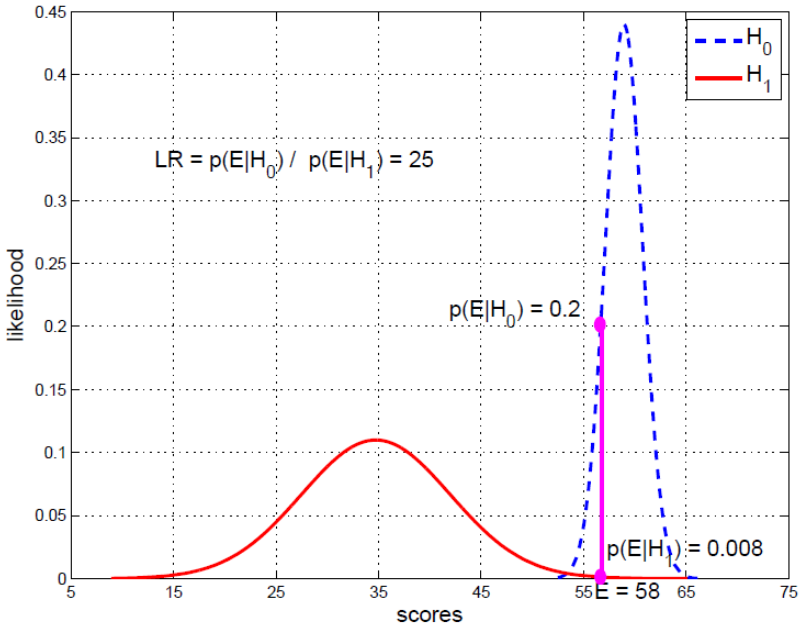

Figure 1: Illustration of the scoring method. In this example, the evidence score is E = 58. Dividing the likelihood value of H_0 distribution (right side of figure) by the likelihood value of H_1 distribution (left side of figure) for the observed speech evidence score E results in LR = 25.

The conclusion that the forensic expert provides to the court shall be related to the assigned likelihood ratio LR (strength of evidence), the observable speech evidence E and the hypotheses H_0 and H_1 under consideration.

Such a non-categorical opinion can be expressed, for example, as follows:

> A likelihood ratio of 25 (obtained in Figure 1), means that it is 25 times more likely to observe the speech evidence score (E) given the hypothesis H_0 (e.g., the suspected speaker recording and the questioned recording have the same source) than given the hypothesis H_1 (e.g., the suspected speaker recording and the questioned recording have different sources).

The judge or the jury in the court uses such an opinion for their deliberations and decision.

The MVKD method frequently used in FSASR (see 3.4) is based on a similar concept as the scoring method. In MVKD, the acoustic distance between questioned speaker and suspected speaker is compared against the distances found in the same-speaker comparisons and the different-speaker comparisons of a population. The former divided by the latter results in the LR. A difference from the scoring method is that MVKD takes place in the multidimensional space of the feature vectors and their models. The probabilistic linear discriminant analysis (PLDA) method within the i-vector modelling approach is similar to MVKD and can be applied for calculating LRs as well.

3.5.2 Direct method

The methodological approach based on the direct method needs a statistical model which can compute a likelihood value when feature vectors are compared against such a model. For example, GMMs or i-vectors combined with PLDA offer such a property.

The direct method uses two databases: the suspected speaker reference database and the relevant population database. They can be used to create two statistical models, e.g., GMM 1 – statistical model of the suspected speaker and GMM 2 – statistical model of the relevant population. The universal background model (UBM) trained with the relevant population database can also be used as GMM 2. In order to calculate the likelihood ratio, the multivariate evidence represented by the ensemble of feature vectors extracted from the questioned recording is compared to GMM 1 and GMM 2. The first comparison gives the similarity likelihood score (numerator of LR) and the second one the typicality likelihood score (denominator of LR). The LRs obtained with the direct method are often not well calibrated (3.6).

The conclusion using the direct method, to be provided to the court by the forensic expert, can be stated in a way similar to the conclusion expressed in the scoring method.

3.6 Calibration and fusion

Upon evaluating calibration (4.3) it might turn out that the likelihood ratios produced by a given method are not well calibrated, i.e., that they have a substantial degree of calibration loss. Hence, a distinction is made here between uncalibrated LRs and calibrated LRs. A FASR or FSASR method

might either output calibrated LRs right away or LRs that are not well calibrated. When the direct method to obtain LRs is used (3.5.2), calibration of the LR values is often necessary. When the scoring method is used the output might be well calibrated. Otherwise, the LR values resulting from the scoring method can also be calibrated.

In FASR and FSASR real world application scenarios, in particular in mismatched recording conditions, calibration should be taken with caution.

So far calibration has been discussed for single FASR and FSASR methods, but it is possible that information from several FASR and FSASR methods can be combined. One commonly used way of combining evidence are fusion methods where inputs from several systems are combined in a way that one likelihood ratio per trial is obtained. Logistic regression fusion is such a method that can provide calibration along with fusion.

3.7 Mismatched recording conditions

Methods vary regarding the ways they deal with so called mismatched conditions. A mismatch occurs when factors which are not related to the speakers' characteristics influence system results. An example of a mismatch is when the questioned recording is made from a telephone conversation in a noisy environment whereas the suspected speaker recording is from a microphone-recorded police interview in a reverberant room.

There exist various ways of reducing the mismatch effect on speaker recognition system results at different stages of the recognition process: feature extraction, feature modelling and similarity scoring. These include cepstral mean subtraction, cepstral mean and variance normalisation, feature warping, relative spectra (RASTA) filtering, T-norm, and Z-norm. Developments in joint factor analysis and the i-vector approach also provide effective means of mismatch compensation. Statistical compensation for mismatched recording conditions in the scoring method can be done using the principal Gaussian component compensation technique.

Compensation for mismatched conditions is an important but difficult task in FASR and FSASR. Currently, in forensic applications, it is not possible to remove mismatch effects completely, but only to reduce their influence.

3.8 Databases

FASR and FSASR methods require the use of several databases. Depending on the particular method they can include:

- Background population (e.g., for creating a UBM in the direct method or in the prior stages of the scoring method)
- Relevant population or Reference population (H_1-population in the scoring method)
- Suspected speaker reference and control database (H_0-population in the scoring method)
- Development database (for calibration and fusion)
- Validation database (for validation of LR methods)
- Evaluation database (for case-specific performance evaluation)

Some of the databases are part of the method, for example they are needed to calculate a likelihood ratio (in the example list, the first four items). Others (the last two items) do not belong to the method itself, but are needed for the LR method validation (chapter 4) and evaluation of LR methods on a case-specific basis (5.5 and chapter 6). The notion of "relevant population" (5.3) is not limited to the second item on the list but can apply to most or all of the items (see also chapter 6).

The final result of a speaker recognition case critically depends not only on the similarity between questioned and suspected speaker sample but also on the properties of all databases used in the casework. Depending on the method and the databases available or collected for the case, database selection can be performed by the forensic expert or can be conducted automatically under pre-defined criteria.

Some of the FASR and FSASR software applications have their own environment for the storage and management of database audio files, whereas in other cases database management is performed independently. In any of these situations, the management of the databases available to the forensic expert and their content needs to be documented carefully and the content held as up-to-date as possible. If the database is going to be shared with forensic experts from other laboratories or made available for research, issues of anonymisation must also be considered.

4 Method Validation

4.1 Introduction

The purpose of validation of a FASR or FSASR method is to measure its performance in specific conditions and to determine whether the performance satisfies a given validation criterion or several criteria.

The output of any FASR or FSASR system considered in this chapter is a likelihood ratio (LR) and the method to be validated is a LR method.

Validation should be carried out using a validation database. The intrinsic property of this database is the known ground truth regarding the source of origin of each of the recordings. When the ground truth of the recordings related to the origin of the speakers is known, there are two types of trials, e.g.:

a) Same speaker trials (SS trials)
b) Different speaker trials (DS trials)

In the process of empirical evaluation on the validation database the FASR or FSASR method outputs a LR for each of the SS and DS trials (see 3.5 on LR calculation).

Given the LRs and the ground truth regarding the SS and DS trials, it is possible to derive a number of performance characteristics and metrics as well as validation criteria.

Validation of a LR method in FASR or FSASR has to be performed with speech samples that are typical representations of the speech material the forensic laboratory is confronted with in everyday work. It is not sufficient to rely on method validations with data drawn from outside forensic contexts or from forensic samples that differ from those typically found in ones' own casework. Relying on figures given by commercial developers of FASR or FSASR systems regarding the performance of their systems in international speaker recognition tests is likewise considered insufficient. Instead, those systems have to undergo forensic-specific (understand: in-house) validation. For this purpose it is necessary to build and maintain validation databases with known ground truth of the same-speaker and different-speaker pairs, if possible based on previous casework (if the recording conditions are similar), or on data sets matching the casework conditions very closely.

4.2 Performance characteristics and metrics

It is possible to derive a number of performance characteristics and performance metrics associated with the characteristics, given the LRs and the ground truth regarding the SS and DS trials.

Performance characteristics (e.g., Tippett plots I and II as well as Proportions Trade-Off (PTO), Applied Probability Error (APE), Empirical Cross Entropy (ECE) plots) describe support for the correct hypothesis of a LR method.

Performance metrics provide a single numerical value that describes the performance in terms of e.g., accuracy, discriminating power and calibration of the LR method (probabilities of misleading evidence ($PMEH_0$ and $PMEH_1$), equal proportion probability (EPP), log-likelihood-ratio cost (Cllr)).

- *Accuracy – performance property which presents the closeness of agreement between an assigned LR value and ground truth status of the hypothesis.*
- *Discriminating power – performance property representing the capability of a given method to distinguish between SS and DS comparisons where different hypotheses are true.*
- *Calibration – performance property of a set of LR values in which perfect calibration implies reliable probabilistic interpretation of the comparisons for either proposition. The strength of evidence of well-calibrated LRs tends to increase with the discriminating power for a given method.*

Validation criteria present conditions related to the performance metrics that have to be met as a necessary condition for the method to be deemed as valid.

Natural logarithm transformation (logLR) is often applied to the LR values. The main advantage of the log transformation is the symmetrical scale with the center of symmetry at logLR = 0. The LRs supporting the hypothesis H_0 tend to have positive values and the LRs supporting the hypothesis H_1 tend to have negative values as shown in Figure 2.

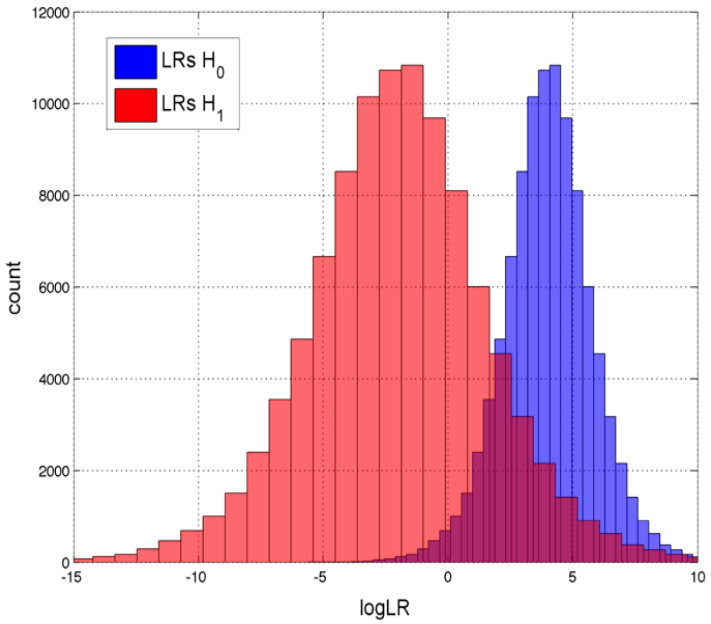

Figure 2: Distributions of logLRs supporting either H_1 or H_0 hypothesis

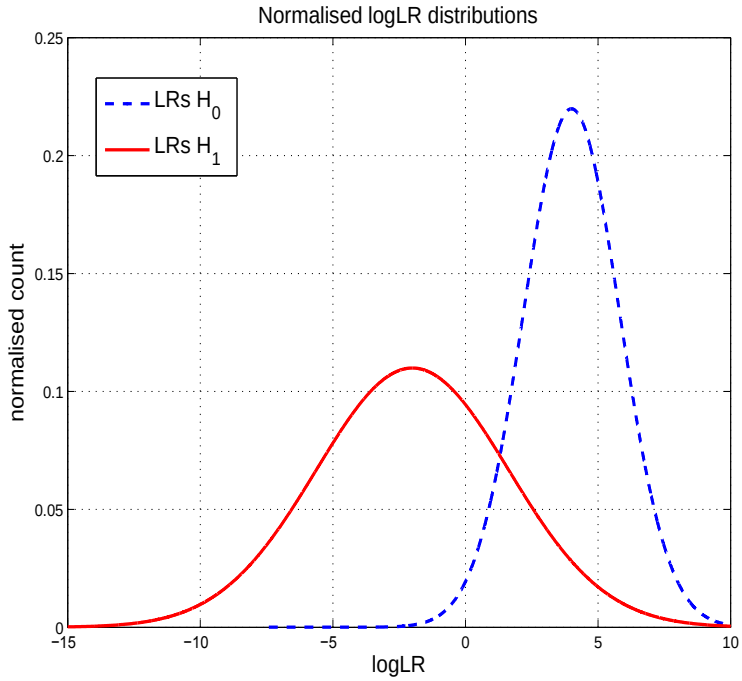

Figure 3: Normalised distributions of logLRs supporting either H_1 or H_0 hypothesis

4.2.1 Performance characteristic 1 – Tippett plots I

In order to move from the histograms (count vs. LR) of the LRs for H_1 and H_0 hypotheses to cumulative distribution functions (probability vs. LR), they have to be approximated by probability density functions and normalised so that the area under each of the curves is equal to 1 (Figure 3).

Following the normalisation of the LR distributions the cumulative distribution functions (CDFs) of the LRs can be computed. The cumulative distribution functions (Figure 4), unlike the normalised LR distributions (Figure 3), are plotted as the probability on the Y-axis vs. the LRs on the X-axis. They represent the cumulative proportion of LRs less than the LR value on the X-axis for the H_1 and H_0 hypotheses.

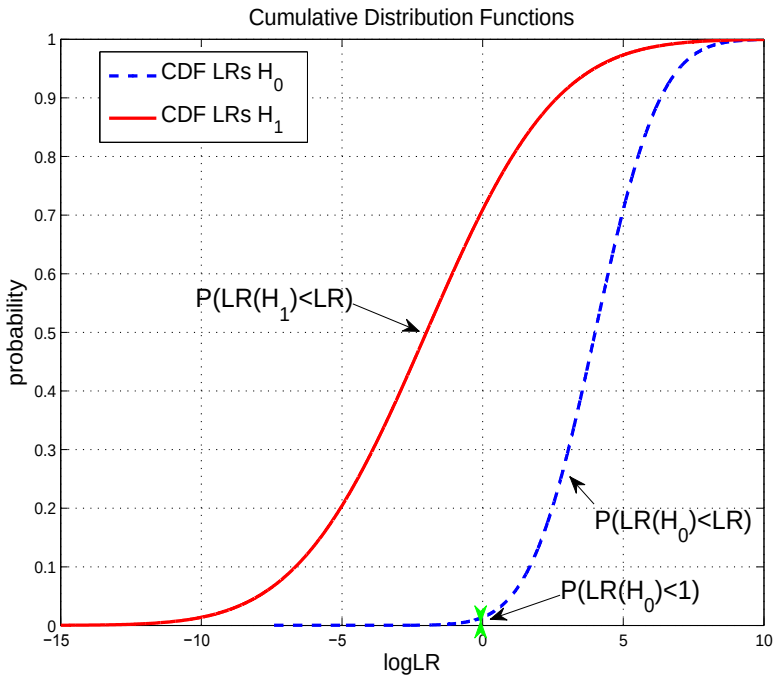

Figure 4: Cumulative Distribution Functions (CDFs) for H_1 and H_0 hypotheses

Tippett plots (Figure 5) show the inverse CDFs of LRs for both H_0 and H_1 hypotheses. They represent the cumulative proportion of LRs greater than the LR value on the X-axis for the H_1 and H_0 hypotheses.

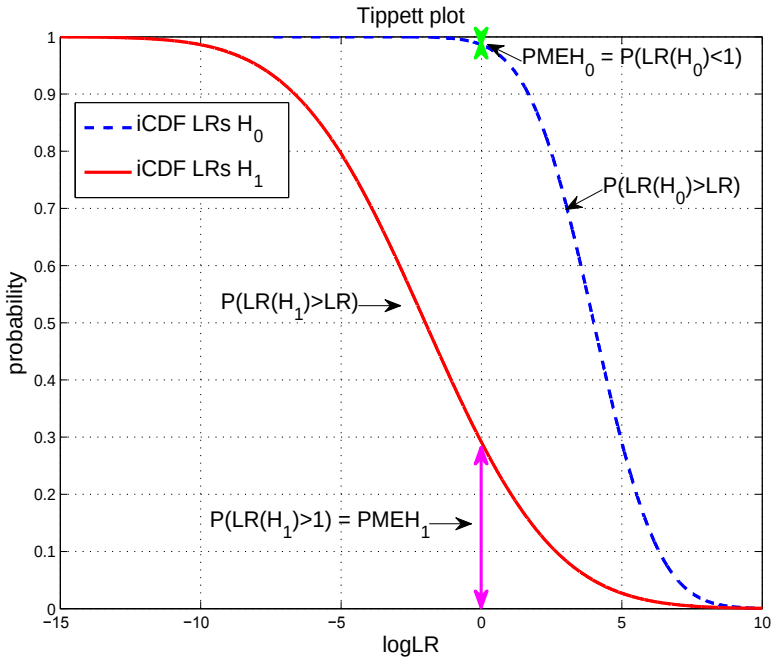

Figure 5: Tippett plots I (inverse cumulative distribution functions)

4.2.1.1 Performance metric 1 – Probabilities of misleading evidence ($PMEH_0$ and $PMEH_1$)

Probabilities of Misleading Evidence ($PMEH_0$ and $PMEH_1$) are associated with the Tippett plots representation. They are defined in the following way:

- $PMEH_0$: probability of misleading evidence in favour of the hypothesis H_1. The probability of all LRs that are smaller than 1, knowing the H_0 hypothesis is true $PMEH_0 = P(LR(H_0) < 1) = 1 - P(LR(H_0) > 1$
- $PMEH_1$: probability of misleading evidence in favour of the hypothesis H_0. The probability of all the LRs that are bigger than 1, knowing the H_1 hypothesis is true $PMEH_1 = P(LR(H_1) > 1)$

Based on their definitions the PMEs can be seen as a measure of accuracy. The corresponding PMEs in the example in Figure 5 are $PMEH_0 = 0.011$ and $PMEH_1 = 0.279$, at logLR = 0.

4.2.2 Performance characteristic 2 – Tippett plots II

Figure 6 shows a graphical representation of the CDF for H_0 $P(LR(H_0) < LR)$ and the inverse CDF (iCDF) for H_1 $P(LR(H_1) > LR)$. The advantage of this representation is that at the intersection of the CDF and iCDF we find the Equal Proportion Probability (EPP).

They represent:

- the cumulative proportion of LRs less than the LR value on the X-axis for the H_0 hypothesis; $P(LR(H_0) < LR)$
- the inverse cumulative proportion of LRs greater than the LR value on the X-axis for the H_1 hypothesis; $P(LR(H_1) > LR)$

The representation in Figure 6 shows Tippett plots II.

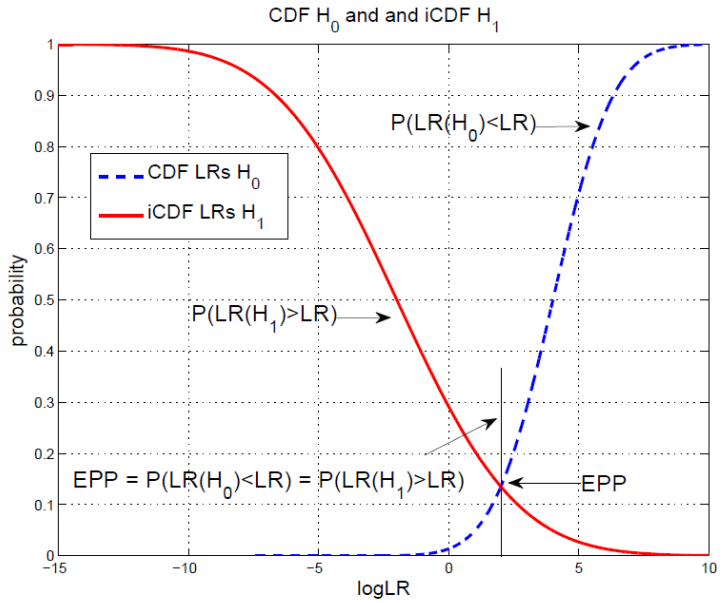

Figure 6: Tippett plots II

The corresponding EPP value in this case is EPP = 0.120 at logLR = 2 (Figure 6).

The proportions trade-off (PTO) curve can be obtained from Figure 6 by representing the probability of proportions $LR(H_0) < LR$ vs. probability of proportions $LR(H_1) > LR$ for likelihood ratios calculated for all the SS and DS trials (example shown in Figure 7). This curve can be represented on either linear or normal deviate scales. When presented on normal deviate scales, its linearity occurs when the logLRs are normally distributed.

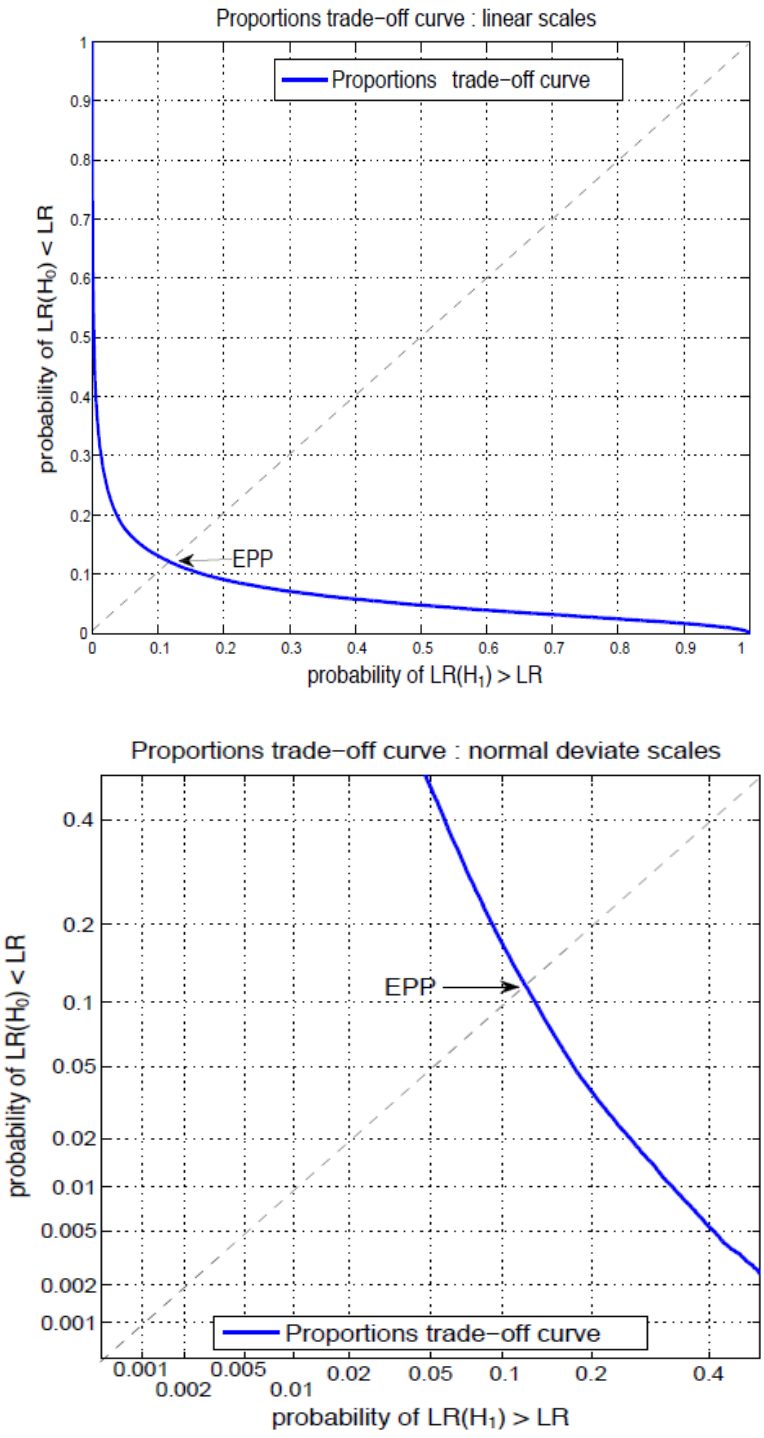

Figure 7: Proportions trade-off curves. Upper: linear scales; lower: normal deviate scales

4.2.2.1 Performance metric 2 – Equal Proportion Probability (EPP)

The EPP is present in the proportions trade-off curve at the intersection of this curve and the main diagonal as shown in Figure 7. In the example presented EPP = 0.120 for logLR = 2.

Based on the definitions in section 4.2 the EPP can be seen as a metric of discriminating power. The lower the EPP the better the discriminating capabilities of a FASR or FSASR system e.g., capacity of the system to discriminate between the SS and DS trials.

Tippett plots II can serve as an indicator of miscalibration. For a perfectly calibrated system the EPP is equal to both types of misleading evidence and is calculated for logLR = 0.

If comparing different FASR or FSASR methods or if comparing the same method based on data recorded in various conditions, the proportions trade-off (PTO) curve and the EPP compare the systems based on the discriminating power. The closer the curve and the EPP are to the zero value, the higher the discriminating power of the system.

4.3 Performance characteristics and metrics based upon ranges of prior probabilities

The support of the LRs to either of the hypotheses in the Bayesian decision process in the court is measured at the level of posterior probabilities. In order to compute the posterior probabilities, LRs, produced by a method under evaluation, need to be combined with prior probabilities, which are by definition unknown to the forensic expert. Since a particular prior is not known, the posterior probabilities are computed for a wide range of priors.

4.3.1 Performance characteristic 3 – Applied Probability of Error (APE)

Applied probability of error (APE) is proposed as a posterior-based performance characteristic (Figure 8). It is an interpretation of the total-error probability for a range of possible priors. In the APE the total probability of error P(error) is calculated by applying a logarithmic scoring rule to assign the expected costs (weights) to the probabilities of proportions in the following way:

$$P(\text{error}) = P(LR(H_0)<LR)(\log LR) \times P(H_0) + P(LR(H_1)>LR)(\log LR) \times P(H_1)$$

Using the logarithmic scoring rule small LR values correctly supporting the H_1 hypothesis and large LR values correctly supporting the H_0 hypothesis are assigned small cost, while the erroneous LR values supporting the wrong hypothesis in either case are assigned a large cost. The bigger the probabilities of proportions, the larger the cost assigned.

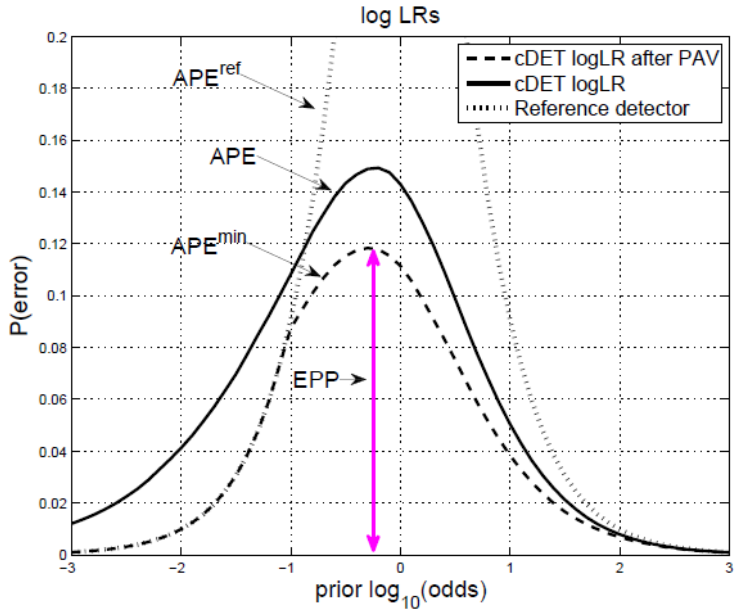

Figure 8: Applied probability of error (APE) plots

The solid curve shows the probability of error obtained when minimum cost decisions are made with the log-likelihood ratios calculated for each value of prior probability. The area under the solid line is equal to the log-likelihood-ratio cost (Cllr).

The dashed curve shows the probability of error following an invertible transformation that optimises Cllr without changing the discriminating power of the system (e.g., Pool Adjacent Violators (PAV)) of the likelihood ratios calculated for each value of prior probability. The area under the dashed curve is proportional to $Cllr^{min}$, which can be interpreted as the total discrimination error for a range of priors. The maximum of this curve corresponds to the EPP.

The dotted curve represents the probability of error for the reference detector for each value of a prior probability and the LR value of 1. The decisions of this detector depend only on the prior probabilities. Since the dotted curve of the APE plots represents a system which constantly outputs LR = 1, we can determine the "applicable range" of a speaker recognition system. The speaker recognition system under evaluation provides support

to the correct hypothesis as long as the Cllr curve stays below the curve of the reference detector. In Figure 8 this condition is violated for the prior log-odds smaller than -0.85.

4.3.1.1 – Performance metric 3 – log-likelihood-ratio cost (Cllr)

Log-likelihood-ratio cost (Cllr – area under the solid curve) is proposed as a metric of accuracy related to the applied probability of error.

$$Cllr = \frac{1}{2Nss}\sum_{iss} log_2\left(1+\frac{1}{LRi}\right) + \frac{1}{2Nds}\sum_{jds} log_2(1+LRj)$$

where *Nss* and *Nds* are respectively the number of likelihood ratios in the SS and DS dataset. The indices *iss* and *jds* denote summing over the SS and DS likelihood ratios.

As the accuracy of a speaker recognition system gets higher the Cllr tends towards zero.

Following the pool-adjacent-violators (PAV) transformation of the logLRs we can obtain the parameter $Cllr^{min}$ as discriminating power metric.

The calibration loss performance metric $Cllr^{cal}$ is obtained by subtracting $Cllr^{min}$ from Cllr. $Cllr^{cal}$ represents the difference in areas between the solid and dashed curves in the APE plots. The Cllr, $Cllr^{min}$ and $Cllr^{cal}$ values are shown in Figure 9.

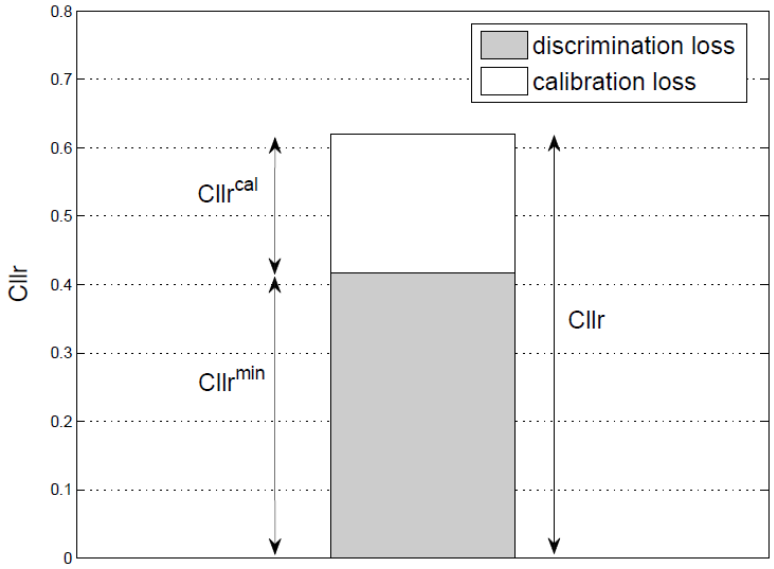

Figure 9: Cllr, Cllr^{min} and Cllr^{cal} values

In the example presented in Figure 9, Cllr is equal to 0.61, Cllr^{min} is equal to 0.41 and Cllr^{cal} is equal to 0.2.

4.3.2 Performance characteristic 4 – Empirical cross-entropy (ECE)

The empirical cross-entropy (ECE) is proposed as a measure of accuracy and calibration and as an alternative to the applied probability of error plots. ECE represents a loss of information due to the uncertainty in the propositions at each particular prior. The recognition system leads to high accuracy when the solid and dashed ECE curves (Figure 10) get close to zero. Cllr is the value of the ECE at the prior log odds equal to zero and Cllr^{min} is the value of the ECE^{min} at the prior log-odds equal to zero, which measures the discriminating power. While Cllr measures the average cost of decisions for all prior probabilities, ECE measures the information needed to support the correct hypothesis for a given set of LR values. The difference between Cllr and Cllr^{min} is attributed to the calibration loss Cllr^{cal}. In a well-calibrated system ECE and ECE^{min} are close together, which is not the case in Figure 10.

However, if the Cllr is also interpreted as a loss of information, then it is restricted to the prior odds of 1 (prior probabilities of 0.5). Therefore, depending on its interpretation as average cost or information loss, the Cllr is independent of the prior probabilities or not.

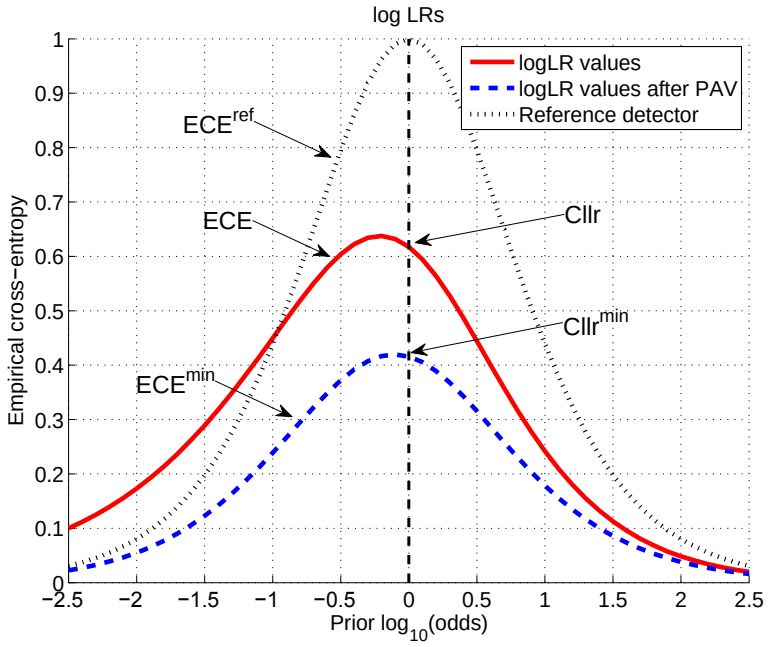

Figure 10: Empirical cross-entropy (ECE) plots

A reference detector that outputs the logLR values of zero is provided and displayed as a dotted line in ECE plots. The solid line represents the accuracy of the LRs as the function of prior log-odds, the dashed curve represents the discriminating power of the LRs and the difference between these two curves represents the calibration loss. As the ECE tends towards zero, the accuracy of the LR method improves. Likewise as ECE^{min} tends towards zero, the discriminating power of the LR method improves.

A range of application of the FASR or FSASR system can be deduced from the ECE plots as well. The speaker recognition system under evaluation provides support to the correct hypothesis as long as the solid curve stays below that of the reference system. In Figure 10 this condition is violated for the prior log-odds smaller than -0.85.

4.4 Validation criteria and validation decision

A validation criterion presents a requirement related to the performance metric that has to be met as a necessary condition for the LR method to be deemed as valid.

In order to specify this validation criterion the forensic expert should be aware of the functionality of a given method in a given (case-related) condition. Validation criteria can, for example, be formulated in the following way:

"$PMEH_0$ for the method under evaluation should be smaller than 0.05."
"$PMEH_1$ for the method under evaluation should be smaller than 0.3."
"EPP for the method under evaluation should be smaller than 0.13."
"Cllr for the method under evaluation should be smaller than 0.65."
"$Cllr^{min}$ for the method under evaluation should be smaller than 0.45."
"$Cllr^{cal}$ for the method under evaluation should be smaller than 0.05."

The validation criterion can be based on different sources (or a combination of these):

a) direct comparison with the state of the art
b) empirically based on the performance measures from previous (similar) cases
c) empirically based on the performance measures of previous method validations
d) based on the results obtained in the collaborative exercises from partner laboratories

Option a) the comparison with the state of the art does not seem to be appropriate as the commercial systems are not tested for a range of forensic conditions or related to the casework.

Option b) appears to be reasonable with respect to matching the forensic condition, however a situation may arise when "similar case(s)" were not previously evaluated. In such a case the forensic expert could refer to option d).

As with option b), option c) relies on in-house experiments and previously handled (in-house) cases. As in the previous option, when similar case(s) were not handled the forensic expert could refer to option d).

Option d) appears to be the most appropriate since there is potential to average the validation criterion over the collaborative exercise results from different participating laboratories.

The validation decision is a binary statement (pass/fail) regarding the fulfilment of a validation criterion related to the performance metric under evaluation.

If the validation criterion or the set of validation criteria is satisfied the validation has been successful. An example is shown in Table 1.

Table 1: Example of validation criteria and validation decisions

Performance characteristic	Performance metric type	Performance metric	Validation criterion	Validation decision
Tippett plots	Accuracy	$PMEH_0$ $PMEH_1$	< 0.05 < 0.3	Passed Passed
Proportions trade-off curve	Discriminating power	EPP	< 0.13	Passed
APE plots ECE plots	Accuracy Discriminating power Calibration	Cllr $Cllr^{min}$ $Cllr^{cal}$	< 0.65 < 0.45 < 0.05	Passed Passed **Failed**

5 Case Assessment

5.1 Introduction

In case assessment, the audio samples provided have to be examined with respect to a variety of characteristics of the speech recordings, including its net duration and various technical and contextual aspects of speech. Subsequently, these characteristics are compared to the requirements of the FASR and FSASR methods that are at the disposal of the forensic laboratory. If the requirements of any of the FASR and FSASR methods are met, this method or set of methods can be applied to the case at hand. If the requirements are not met, the corresponding methods cannot be applied.

5.2 Preparatory aspects

Before commencing examination of a case, the forensic expert should register the items submitted for examination along with the available information and assess them with regard to the requirements of the mandating authority or party. The forensic expert should, if necessary, clarify or re-define the requirements with the mandating authority or party. Questions asked should be documented and referred in the final report. In general, specific care should be taken with the transportation of speech evidence, e.g., compact discs (CDs), hard disc drives (HDDs), solid state drives (SSDs), to avoid any physical damage or exposure to strong electrical static, magnetic fields or large temperature variations. If the file format is compatible with the requirements of the FASR or FSASR system it can be used directly. If conversion is necessary it should not change the acoustic quality of the speech samples.

Information about file format history should be considered (3.2). Any indications of quality reductions that may have an effect on system performance have to be taken into account.

It might be advised, in case of a substantial compilation of submitted audio material, to enquire which items offer the best choice of speech data in terms of forensic value. The forensic expert should examine if and to what extent the request made by the mandating authority or party can be accommodated with the speech materials provided.

Since considerable human input is required in FSASR and some in FASR, the problem of cognitive bias has to be taken into consideration. Such bias

could become manifest e.g., in editing and/or multiple processing of case materials leading to different results. If multiple analyses have been carried out with the case material, it is mandatory therefore that the reasons are substantiated and the different results are reported. When examinations or test results are rejected, the reasons should also be reported (7.2).

5.3 Relevant population

The relevant population is the set of speakers chosen when formulating the H_1 hypothesis (also referred to as the different-speaker hypothesis or the alternative hypothesis) in forensic speaker recognition. Although possible in theory that the different-speaker hypothesis includes any possible speaker, this hypothesis is usually more restricted at least to speakers of the same sex, but often also to speakers of the same language and perhaps even the same language variety. The definition of the relevant population might also include criteria of the specific quantity and quality profile of the case (5.4).

Although some aspects of defining the relevant population can be fairly technical (sometimes to a point that different-speaker sets are selected automatically in a FASR system or that the same UBM is used for several cases), some other aspects may be of direct interest to the mandating authority or party, and it might get involved in the definition of the relevant population. If the mandating authority or party requests a specific relevant population, it needs to be assessed by the forensic expert whether this request can be met.

If the mandating authority or party does not request any specific relevant population it is necessary that the forensic expert defines a relevant population that is compatible with the circumstances of the case at hand and for which the necessary databases and other resources are available.

5.4 Quantity and quality profile of the forensic audio material

Whether FASR and FSASR methods can be applied meaningfully depends on the quantity (5.4.1) and technical quality (5.4.2) of the questioned speaker and suspected speaker recordings as well as on the contextual (behavioural and situational) aspects governing the production of the respective speech utterances (5.4.3). Another important aspect is whether the technical quality and the contextual aspects of the questioned speaker recordings and the suspected speaker recordings are compatible (recorded

in matching conditions) or whether they differ systematically (recorded in mismatched conditions) (5.4.4). Within a forensic case, these characteristics can occur in different combinations.

5.4.1 Quantity

One important factor is the duration of the speech sample from a speaker. Many FASR and FSASR methods require that the "net duration" (i.e., pure speech from the relevant speaker, with all irrelevant information removed or disregarded) is no shorter than about 15-30 seconds. There is no general rule about the amount of audio material necessary, and different methods might have different requirements. Ultimately, the minimum net duration required for a method has to be established with a method validation (chapter 4) or other tests (5.5).

5.4.2 Technical Quality

In addition to the duration factor, there are several factors concerning the technical quality of the audio material that can have an influence on the applicability and success of FASR and FSASR. These involve, but are not limited to the following:

- filter effects
- environmental noise (e.g., music, traffic, hiss, hum)
- quantisation noise
- reduced signal to noise ratio (SNR)
- distortions and artefacts
- drop outs and amplitude reductions
- echo, reverberation (delay effects)
- compression
- effects of audio enhancement where it has been applied

5.4.3 Contextual Aspects

In addition to the technical aspects of recording quality, there are a number of contextual (behavioural and situational) quality aspects that might have a limiting effect on FASR and FSASR methodology as well. These include:

- speaking under stress and emotion
- speaking under the influence of intoxication
- increased vocal loudness (effort); shouting
- speaking under the influence of the common cold or allergic reactions
- speaking while fatigued
- speaking with reduced intelligibility
- speaking styles other than spontaneous speech (e.g., reading, repeating)

Voice disguise is another source of quality limitation. Voice disguise can occur in the form of technical manipulations (e.g., manipulations of playback speed; pitch-shifting) or in the form of behavioural modifications (e.g., speaking in falsetto voice; speaking in slow, staccato rhythm; imitating a foreign accent).

5.4.4 Mismatched conditions

With any of the technical and contextual aspects mentioned so far there could be a situation of mismatched conditions. Mismatch means that for any given factor, the questioned speaker recording and the suspected speaker recording show different instantiations of that factor. For example, the questioned speaker recording might contain heavy street noise whereas the suspected speaker recording is made in a quiet environment. The utterances in the questioned speaker recording might also be spoken with a loud voice whereas the vocal effort in the suspected speaker recording is neutral.

Two further types of mismatch are non-contemporaneous speech and language mismatch. In non-contemporaneous speech there is a time delay between the date of the recording of the questioned speaker and the one of the suspected speaker. If the time delay is at the order of several years, significant changes in the voice patterns can have occurred, although some speakers also remain remarkably stable over time. Scenarios in which the time delay is at the order of just a few days, weeks or months are commonly referred to as session-mismatch or session variability. Session mismatch in the sense of short-term non-contemporaneousness occurs in almost all FSR cases, and a FASR or FSASR method has to be able to deal with it on a regular basis.

Language mismatch means that a different language is spoken in the questioned speaker recording than in the suspected speaker recording.

Language mismatch does not generally preclude the application of FASR and FSASR methods because vocal tract characteristics and prosodic phenomena can remain fairly stable across first and second language. However, the language structure itself can impose its influence on the features, for example the system of vowel phonemes and their phonetic implementation in a language has an influence on formant frequencies and MFCCs.

The extent to which the factors mentioned above influence FASR and FSASR has to be determined empirically through method validation, either without mismatch compensation or, if possible, with application of mismatch compensation methods (3.7).

5.5 Comparing the quantity and quality profile with FASR and FSASR requirements

After the quantity and quality profile of the case has been determined, it is necessary to compare this profile with the system requirements of any FASR or FSASR method that is at the disposal of the forensic laboratory.

If the quantity and quality profile of the case under examination meets the requirements of a system for which a method validation is available, the case can be processed with the FASR or FSASR method for which the system requirements are met.

If system requirements for a given FASR or FSASR method are not met, it can be considered whether a new database can be compiled or whether an existing database can be adapted and evaluated in a way that the quality and quantity profile of the case is met. In that case it is important that a test is performed on this new or modified test set and that performance characteristics and metrics are derived that are analogous to a full method validation (chapter 4). The only difference from a full method validation would be that such a more case-specific testing and evaluation does not contain a validation criterion.

5.6 Procedure if more than one FASR and FSASR method can be applied

If more than one FASR or FSASR method can be applied it should be considered whether fusion between these methods could be carried out (3.6). For fusion to be applicable, there has to be a development database

from which the fusion weights of the individual methods are determined. Alternatively, the fusion weights are determined based on cross validation from the same database that is used for the method validation or the case-specific evaluation.

6 Evaluation and Interpretation

6.1 Introduction

Evaluation of specific case results concerns the likelihood ratio (LR), which summarises the statement of the forensic expert in the casework and its comparison with the likelihood ratios that can be obtained from the observed speech evidence under operating conditions of the case, on the one hand when the hypothesis H_0 is true and, on the other hand, when the hypothesis H_1 is true (2.3 and 3.5). Evaluation should be carried out using a relevant population database of the case (3.5, 3.8 and 5.3). For evaluating the strength of evidence (LR), performance characteristics defined in chapter 4 can be used together with the case-specific likelihood ratio LR_{case}. In order to facilitate the interpretation of the strength of evidence statement made in the case, the court can be provided with Tippett plots I and II as well as performance metrics such as $PMEH_0$, $PMEH_1$, EPP, Cllr and their interpretations regarding the case-specific likelihood ratio LR_{case}.

Methodology concerning evaluation and interpretation of specific case findings is presented in section 6.2. Section 6.3 introduces considerations about how to combine the strength of evidence derived from FASR or FSASR with findings that fall outside the scope of FASR and FSASR methods but are nevertheless relevant to forensic speaker recognition.

6.2 Case-specific strength of evidence and its evaluation

When applied to specific speaker recognition based casework, a FASR or FSASR method allows a numerical statement about the strength of evidence (LR) that the method can provide. Such casework can involve measuring calibration of likelihood ratios. Calibrated LR is understood here as a likelihood ratio computed with a method aimed at improving the calibration metric (see 3.6 and chapter 4). Some methods for calculating LRs can produce well-calibrated results without the need for additional calibration.

Calibrated likelihood ratios are the most desirable strength-of-evidence statements and they are most compatible with the Bayesian interpretation framework. Calibrated likelihood ratios have a direct interpretation that can be reported and explained to the court. For example, if the likelihood ratio of a case has a value of 100 it is 100 times more likely to observe speech evidence supporting the H_0 hypothesis than the H_1 hypothesis given the

questioned speaker and suspected speaker recordings, as well as the relevant population database.

An example of a case-specific LR expressed as calibrated likelihood ratio (corresponding to the vertical magenta line) is shown along with Tippett plots II in Figure 11 (4.2.2).

Tippett plots provide a clear picture in terms of actual values of the LRs derived in the specific case for the relevant population of same-speaker and different-speaker trials under H_0 and H_1 hypotheses. General information about the accuracy of the system used in the case can be given by probabilities of misleading evidence $PMEH_0$ and $PMEH_1$, which represent accuracy metrics. The value of the discriminating power metric EPP (Equal Proportion Probability), corresponding to the crossing point of H_0 and H_1 plots, shows the discrimination of the system used. Alignment of this crossing point with logLR = 0 on the X-axis indicates that the system is calibrated.

The log-likelihood-ratio cost (Cllr), which assesses the overall performance of the system, can be used to confirm the accuracy performance of the system derived from the Tippett plots. $Cllr^{min}$ and $Cllr^{cal}$ can be used to confirm discriminating power and calibration loss of the case-specific system, respectively.

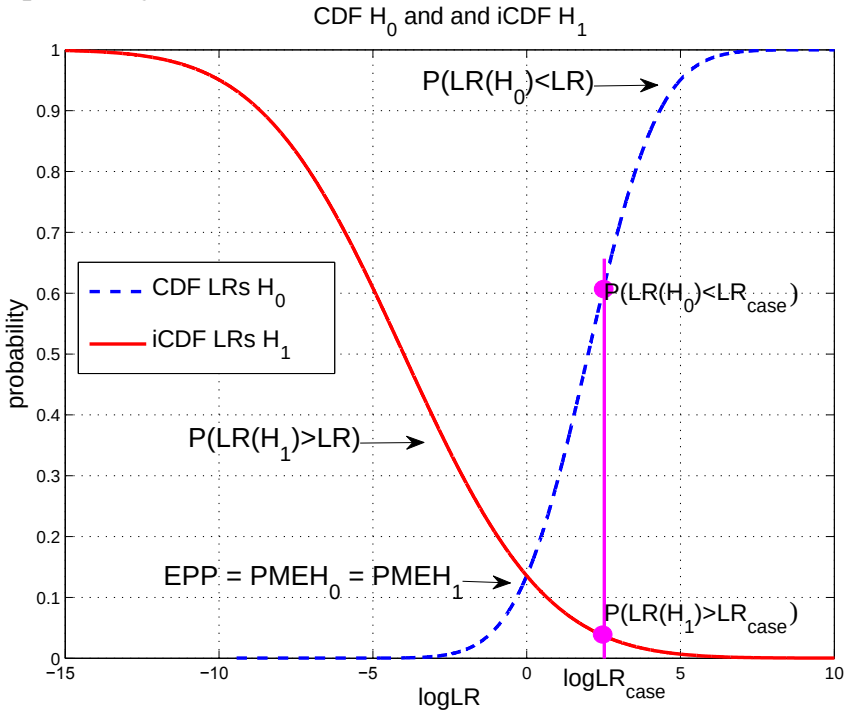

Figure 11: Relation between case-specific LR (LR_{case}) and Tippett plots II for calibrated LRs

For calibrated LRs, if the case-specific LR_{case} falls on the right side of logLR = 0, the case-specific speech evidence provides stronger support for the H_0 hypothesis than the H_1 hypothesis. In the example presented in Figure 11, the reported $logLR_{case}$ value is equal to 2.5. Such a value is reported to the court as the strength of evidence.

Within the performance characteristic (Tippett plots II) it is also possible to report what the probabilistic distance of case-specific LR_{case} is to the misleading evidence $PMEH_0$. This distance can be calculated as $P(LR(H_0) < LR_{case}) - PMEH_0$. It corresponds to the proportion of likelihood ratios $LR(H_0)$ greater than LR = 1 and smaller than LR_{case}.

It is furthermore possible to report the case-specific probabilistic error expressed as proportion of likelihood ratios $LR(H_1)$ greater than the LR_{case} value $P(LR(H_1) > LR_{case})$.

An analogous set of values can be reported if LR_{case} falls on the left side of logLR = 0.

Figure 12 shows Tippett plots II for a non-calibrated set of LRs. The non-calibrated nature of the data can be seen from the fact that the intersection point between the plots of $P(LR(H_1) > LR)$ and $P(LR(H_0) < LR)$ does not occur at or near logLR = 0 on the X-axis but occurs at some distance from it (near 2 in this example).

Similarly to calibrated LRs, for non-calibrated LRs general information about the accuracy of the system used in the case can be given by probabilities of misleading evidence $PMEH_0$ and $PMEH_1$, which represent accuracy metrics. The value of discrimination power metric EPP (Equal Proportion Probability), corresponding to the crossing point of H_0 and H_1 plots, shows the discriminating power of the system used. Clear misalignment of this crossing point with logLR = 0 on the X-axis indicates that the system is non-calibrated. As before with calibrated LRs, Cllr can be used to state the accuracy performance of the non-calibrated LRs. $Cllr^{min}$ and $Cllr^{cal}$ can be used to confirm discrimination power and calibration loss of the system, respectively. Calibration loss is greater in an uncalibrated than a calibrated system.

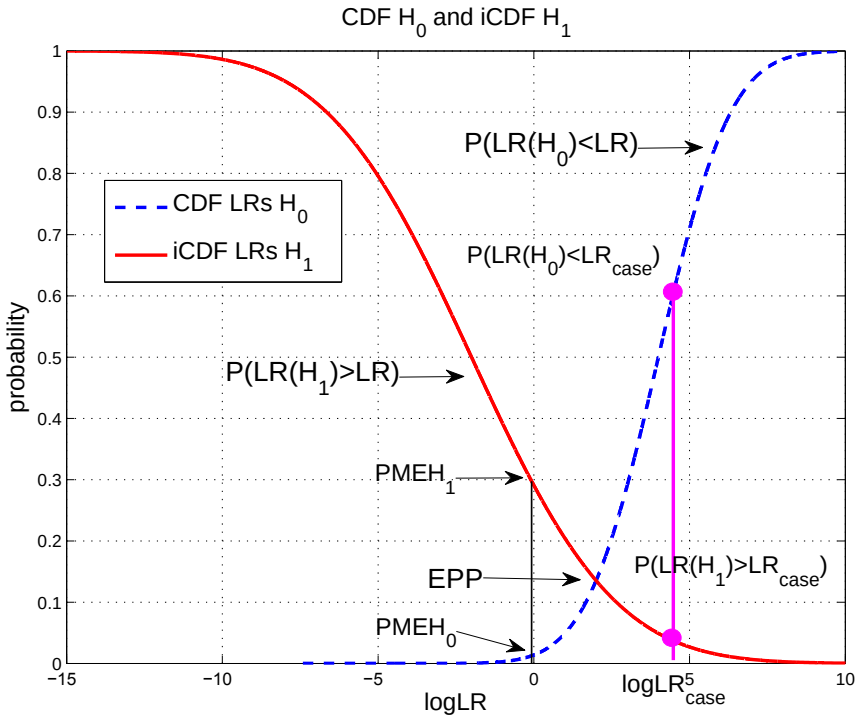

Figure 12: Relation between case-specific LR (LR_{case}) and Tippett plots II for non-calibrated LRs

An example of a case-specific LR expressed as non-calibrated likelihood ratio (corresponding to the vertical magenta line) is shown along with Tippett plots II in Figure 12.

Within the performance characteristic (Tippett plots II) it is also possible to report what the probabilistic distance of case-specific LR_{case} is to misleading evidence $PMEH_0$. This distance can be calculated as $P(LR(H_0) < LR_{case}) - PMEH_0$. The case-specific probabilistic error expressed as proportion of likelihood ratios $LR(H_1)$ greater than the LR_{case} value $P(LR(H_1) > LR_{case})$ can be calculated as well.

In the example shown in Figure 12, LR_{case} falls on the right side of logLR = 0. If it falls on the left side, an analogous set of values can be reported.

6.3 Combining the strength of evidence derived from FASR or FSASR with other examination results

Whereas the output of a FASR or FSASR method or a combination thereof allows a numerical strength of evidence statement, this is usually not possible with other methods of FSR coming from the domain of the auditory-phonetic-and-acoustic-phonetic approach. If the results from both

domains of FSR are combined, the outcome cannot be a numerical statement since the auditory-phonetic-and-acoustic-phonetic approach cannot provide this. The remaining options are verbal statements. If the outcome of the auditory-phonetic-and-acoustic-phonetic analysis is expressed as verbal statement, the combination with the quantitative LR by the FASR or FSASR system can be achieved verbally.

7 Case File and Reporting

7.1 Case file

A full record of the work undertaken and of the correspondence associated with the case at the time of the examination is contained within the case file.

Under certain circumstances the case file can be disclosed and made available for inspection at the forensic laboratory or elsewhere.

The exact requirements for case file content may vary with different legal systems or laboratory regulations. In any case the records should contain enough detail to enable another expert competent in FASR or FSASR to identify the previous work carried out and assess the findings independently.

Generally, the following information items should be included in the case file:

- a list of all items (e.g., speech recordings, files etc.) received
- information regarding recording equipment and method, file format and audio codec for digital speech items submitted
- if possible, information about whether an audio file has been re-converted or compressed prior to being submitted, i.e., whether the format or settings of the submitted audio file differs from the original recording
- for each item the name, address and status of the submitting party, the date of receipt, the method of delivery (e.g., via email, by handing over, by mail, etc.), any relevant accompanying information, and the labelling
- the nature of the casework requested by the mandating authority or party, i.e., the key issues in the case (e.g., FASR or FSASR using questioned speaker and suspected speaker recordings) and any conditioning information
- any competing hypotheses (propositions) expressed within the specific circumstances of the case
- all exchange of information within the laboratory and between the laboratory and the mandating authority or party about the case
- sufficient information to fully inform the mandating authority or party if speech recordings of suspected speakers were made by the

forensic expert or other personnel of the laboratory in question for FASR or FSASR purposes in the case
- information regarding equipment, file format and audio codec for FASR or FSASR purposes in the case
- the examination strategy,
    - i.e., what examinations (e.g., FASR or FSASR) and analyses have been performed, by whom, when, and in what order
    - rationale and information if speech items were edited for FASR or FSASR purposes
    - information concerning the speech samples that were used for FASR or FSASR purposes
    - information about pre-processing, feature extraction, speaker modelling and LR calculation methods, selected databases, system settings, version of the software and related details
- if examinations have been carried out with the assistance of scientific support staff, details of those who assisted, and brief details of the work performed by them, should be given
- information for the mandating authority or party if analyses were not continued after initial pre-assessment of the speech items (e.g., because items submitted were found to be not suited for FASR or FSASR analyses or because further scientific examinations were unlikely to differentiate between the hypotheses)
- observations and analysis results
- evaluation of the strength of evidence that the findings provide regarding the key issues and related hypotheses
- details of any generated administrative and technical review of the casework and analyses undertaken in the specific case
- final conclusion and the report given to the mandating authority or party
- method of return of items, by whom and when

7.2 Reporting

This section is intended to provide the basic framework of writing reports. The framework is not meant to be rigid, i.e., not all issues covered may be applicable in all circumstances. Those writing the reports or making statements have to make a professional judgement as to what is necessary regarding the circumstances and what meets the needs of the mandating authority or party (court, prosecution, or defence team).

Reports in FASR and FSASR should be made in standard format and style. Appendices may be used where there are tables of results or statistics or descriptions of technical nature that would interrupt the flow of the report or statement. Careful drafting and structuring is necessary to ensure all relevant information and observations are recorded. Reports must be written in a logical format and be understandable to a layperson, if possible.

In general, a report on FASR or FSASR should include the following chapters and information:

7.2.1 Administrative information

- sequentially numbered pages each bearing a laboratory case identifier, e.g., a unique alpha-numeric identification or code, and the reference of the mandating authority or party
- the forensic expert's full name and occupation
- the address of the forensic laboratory, i.e., the institution to which the forensic expert is affiliated
- information regarding qualifications and experience of the expert in order to provide the mandating authority or party with information for them to assess whether or not the examiner can be considered an expert in the context of the particular case
- the signature of the expert and the date
- the date(s) the speech material and other relevant items have arrived
- the name and status of the mandating authority or party submitting the materials and items
- a list of the speech items that were collected or submitted and examined
- a list of the speech items that were produced by the laboratory and examined
- a list of and comments covering speech items not examined and the justification for this
- method of return of items by whom and when
- a list of any items not returned but kept in the laboratory
- details of all relevant communication within the laboratory and between the laboratory and the mandating authority or party and others

7.2.2 Mandate and questions

- the mandate and the questions asked by the mandating authority or party
- the conditioning information used for the case analysis
- the purpose of the examination, i.e., a clear statement of the issues the examinations are intended to address
- the exact hypotheses (propositions) addressed, providing details of their definition

7.2.3 Method(s) and technical issues

- technical issues of FASR and FSASR, i.e., explanations of the scientific principles involved and the type of information that FASR and FSASR examinations can yield
- explanations of technical terms used in FASR and FSASR
- details of the examination, i.e., what kind of examinations have been carried out, which items and edits of speech samples were used
- descriptions of the databases used
- the FASR and FSASR software and hardware tools and versions used
- limitations of the method(s) used as such and as dictated by the circumstances of the case

7.2.4 Examination and results

- details of the examinations carried out, by whom, and their limitations
- analytical results obtained

7.2.5 Discussion and interpretation

- discussion of the strength of evidence and support that the findings provide for the hypotheses under investigation
- assessment and description of what the findings mean and what their significance is in the context of the case

7.2.6 Conclusion

- The conclusion should express the degree of support given by the FASR or FSASR findings to one hypothesis (proposition) against a competing alternative. It constitutes a non-categorical opinion of the forensic expert and should be stated in a stand-alone form.

- The conclusion should be formulated as a likelihood ratio (LR) or expressed as a verbal equivalent of a LR; see *ENFSI Guideline for evaluative reporting in forensic science, ENFSI, 2015*.

- If the conclusion is based on FASR or FSASR findings, the forensic expert has to make clear what kind of speech properties and their characteristics have been taken into account in the whole chain of processing (pre-processing, feature extraction, feature modelling (to create speaker models) and calculation of LR) and then the case-specific evaluation of the LR method.

- If in addition to FASR or FSASR there are findings from the domain of auditory-phonetic-and-acoustic-phonetic analyses they can be combined into a final verbal statement.

8 Quality Assurance

8.1 Aims

This chapter provides forensic experts with best practice regarding quality assurance in FASR and FSASR. It gives specific quality advice, requirements, and criteria.

8.2 Personnel

8.2.1 Key responsibilities

The key role in FASR and FSASR has the forensic expert responsible in a particular case for directing and performing the examination of the speech items submitted, recording speech items, providing the evidence, interpreting the findings and writing the final report.

8.2.2 Required competencies

In general, the required competencies to carry out the key responsibilities are defined in QCC-CAP-003 *Performance based standards for forensic science practitioners, ENFSI, 2004.*

The qualifications, competencies and experience that forensic experts require to carry out examinations involving FSASR and FASR depend on the demands of the various aspects of these examinations. Experts should be knowledgeable of all relevant aspects of all applicable fields involved, i.e., a forensic expert who uses formants in FSASR should have qualifications and understand what they are, how formants may be measured and how they may ultimately contribute to the outcome of the speaker recognition process.

8.2.3 Specific requirements (qualifications and experience)

Forensic laboratories should have written documents of the specific requirements concerning qualifications and experience for both FSASR and FASR forensic experts.

The following specific requirements, qualifications and areas of competence would be expected for the forensic experts in FASR and FSASR.

For both FSASR and FASR, a PhD or Master degree in speech and audio analysis-related sciences is strongly recommended.

Moreover, forensic experts should be accepted as experts in the FASR and FSASR field by fulfilling the following criteria of having:

- peer review and publication record
- knowledge of and ability to demonstrate the theories, technology and procedures applicable in FSASR and FASR
- competence in FSASR and FASR from casework
- knowledge and experience of the demands of the criminal justice system for the presentation of testimony as a forensic expert witness

8.2.4 Training

For training purposes laboratories should have:

- documented training programmes
- procedures for assessing and documentation that a trainee has achieved the level of competence required
- training and competence records for forensic experts and other laboratory personnel involved in FASR and FSASR

8.2.5 Competence assessment

In general, competence assessment can be accomplished through a combination of several means, including:

- practical tests which can include real FASR or FSASR casework
- written and oral examination
- mock court exercises
- FASR or FSASR casework conducted under close supervision
- an inventory of previous casework

Each trainee should be recognised as competent following successful completion of a competence assessment exercise as specified before being allowed to undertake independent case work.

8.2.6 Maintenance of competence

All personnel involved in FSASR and FASR should maintain their competency. Evidence in support of this should be available for periodic review.

Forensic experts should:

- participate actively and routinely in FSASR and FASR casework
- read literature containing relevant information
- take part in suitable seminars, meetings, training courses as well as research and development (R&D) projects
- be up-to-date with technical procedures and standards

8.3 Equipment

An equipment inventory should be maintained for all significant items used for FSASR and FASR examinations (e.g., manuals, software versions).

Only properly operating equipment (hardware and software) should be used in case work, and then only within the limits of its performance check.

The laboratory should have a list of:

- inventories of equipment held
- if applicable, records of maintenance operations
- names of persons responsible for the equipment

8.4 Accommodation and environmental conditions

Forensic laboratories active in FSASR and FASR examinations should be designed for efficient and effective working. Particular attention needs to be given to environment issues, especially the prevention of environmental noises interfering with auditory and acoustic examinations within the laboratory.

8.5 Methods or analysis protocols

Policy of local practices regarding FASR and FSASR methods and analysis protocols should be documented whenever possible as Standard Operating Procedures (SOPs) and included in a training programme. The SOPs should be reviewed and updated if necessary.

As a minimum requirement laboratories should have SOPs for:

- the examination methods and other processes used in FASR and FSASR
- quality control
- documenting and presenting results of examinations

8.6 Validation

Validation is the confirmation by examination, testing and the provision of a validation decision that the particular requirements for a specific intended use of FASR and FSASR methods are fulfilled.

The forensic laboratory should only use properly evaluated and validated methods and procedures (including software) for FASR and FSASR examinations as well as the evaluation and interpretation of their evidential significance in the context of the case.

Validation requires as a minimum that:

- there is an agreed requirement for the method or procedure
- the critical aspects of the examination method or procedure have been identified and the limitations defined
- the methods, materials and equipment used have been demonstrated to be fit for purpose in meeting the examination requirements
- there are appropriate quality control and quality assurance procedures in place for monitoring performance (e.g., proficiency tests (PTs) and collaborative exercises (CEs))
- the method or procedure is fully documented
- the examination results obtained are replicable
- the method or procedure has been subjected to independent assessment and, where novel, peer reviewed
- the forensic experts using the method or procedure have demonstrated that they are competent to do so

Where the methods or procedures have been validated elsewhere, the forensic laboratory should demonstrate that it can achieve the same quality of results in its own hands.

Software testing should involve testing its key functions and formulating a test procedure to ensure that it fully meets the examination requirements.

Testing should be documented and re-testing should take place on upgrades when they become available.

8.7 Case review

FASR and FSASR casework requires a review of the critical findings that make a significant contribution to the case together with both technical and management review aspects.

If possible, a second competent expert should cross-check:

- records of all communication within the laboratory and with external personnel
- details and results of all examinations and tests carried out
- critical findings
- draft and final statement as well as reports

Appendix 1: Annotated Bibliography

This annotated bibliography provides a selection of relevant literature and is organised according to the section names used in this document. Some of the publications refer directly to specific statements made in the main document (e.g., about specific methods proposed in the literature by certain authors), others are provided because they offer a general overview of the relevant subject area. Some terminological issues (such as synonyms among technical terms) are also included in this annotated bibliography. The references are given with name and year. The full spell-out of these references is provided in the following alphabetic list. The bibliography presented here is necessarily non-exhaustive, i.e., there are other relevant references that are not mentioned in this appendix. The annotated bibliography also covers the function of a glossary, which is not provided in this guideline document. Most of the terms used in the document are technical terms on FASR and FSASR and these are best explained by referring to the original literature.

1 Aims

Forensic speaker recognition (FSR) is used as a traditional cover term for forensic automatic speaker recognition (FASR), forensic semiautomatic speaker recognition (FSASR) and the auditory-phonetic-and-acoustic-phonetic approach (Meuwly et al. 1998; Campbell et al. 2009; Neustein & Patil 2012; Hansen & Hasan 2015). Other terms such as "forensic speaker comparison" (Foulkes and French 2012) and "forensic voice comparison" (Morrison 2010) are also in use.

2 Scope

2.1 Evaluative and investigative modes

A guide for thinking and practice in investigations (investigative mode) and in court proceedings (evaluative mode) are presented in Jackson et al. (2006) and Jackson at al. (2015) and are adapted for FASR in Drygajlo (2012).

2.2 Current approaches in forensic speaker recognition

The most recent, general tutorial on speaker recognition including forensic speaker recognition is published by Hansen & Hasan (2015). A survey on international practices in FSR is presented by Gold & French (2011). A recent survey of literature on FSR in general, including FASR, FSASR and the auditory-phonetic-and-acoustic-phonetic approach, is provided by Morrison & Enzinger (2013). Other reviews on FSR in general are presented by Meuwly (2001), Jessen (2008), Eriksson (2012), Drygajlo (2013), French & Stevens (2013) and Campbell (2014), and on some new

research directions by Drygajlo (2014). The term "technical speaker recognition" goes back to Nolan (1983) and Rose (2006). The term auditory-phonetic-and-acoustic-phonetic analysis (AuPA & AcPA) is proposed by Gold & French (2011). Reviews characterising the AuPA & AcPA approach include Nolan (1983, 1997), Hollien (2002), Gfroerer (2003), Watt (2010) and Foulkes & French (2012). Cambier-Langeveld (2007) presents the methods and results of a collaborative exercise on FSR in which the method was left open. Some of the participants used the auditory-phonetic-and-acoustic-phonetic approach, some used FASR and some FSASR. A description of FASR-methods applied to this collaborative exercise is provided by Drygajlo (2009a).

2.3 Bayesian interpretation framework

Literature on some general principles of the interpretation of forensic evidence in a Bayesian interpretation framework includes Evett (1998), Cook et al. (1998), Champod & Evett (2011), Aitken et al. (2010) and Jackson et al. (2015); see also the literature cited therein. Further introductions to the Bayesian framework include Robertson & Vignaux (1995) and Aitken & Taroni (2004) for forensics generally and Meuwly et al. (1998), Champod & Meuwly (2000), Meuwly & Drygajlo (2001), Rose (2002), Morrison (2009b, 2010) and Drygajlo (2011), with special emphasis on FASR and FSASR.

2.4 Forensic conditions

The distinction between intrinsic variability and extrinsic variability is explained in Hansen & Hasan (2015).

3 Methodology of FASR and FSASR

3.1 Definitions

Forensic automatic speaker recognition (FASR):

Texts providing introductions to and overviews of general automatic speaker recognition include Reynolds & Campbell (2008) and Kinnunen & Li (2010). Specific considerations to the forensic aspects of automatic speaker recognition are given by Drygajlo et al. (2003), Künzel & Gonzalez-Rodriguez (2003), Nakasone (2003), Alexander (2005), Drygajlo (2007), Ramos-Castro (2007), Gonzalez-Rodriguez & Ramos (2007), Gonzalez-Rodriguez et al. (2007) and Drygajlo (2011, 2012, 2015).

Forensic semiautomatic speaker recognition (FSASR):

Introductions to FSASR are provided by Rose (2002, 2003) and Morrison (2010). An overview of some earlier semiautomatic systems and approaches is given by Meuwly (2001) including further literature.

3.2 Pre-processing

Speaker separation:

Some references on speaker diarisation are Reynolds & Torres-Carrasquillo (2005), Tranter & Reynolds (2006), Castaldo et al. (2008), Avilés-Casco (2011) and Anguera et al. (2012).

Removal of pauses:

Overviews of Voice Activity Detection (VAD) are given by Mak & Yu (2014) and Sahidullah & Saha (2012).

One useful technique of annotation on several levels and annotation-based net speech extraction is presented in Boersma (2014).

3.3 Features

Short-term spectral envelope features:

Kinnunen & Li (2010) provide a review of short-term spectral envelope features (called short-term spectral features by them), with special emphasis on features typically used in FASR. Long Term Formant (LTF) analysis, a typical FSASR application, was mentioned first in Nolan & Grigoras (2005). Subsequent studies on LTF with some variations in scope and methodology include Becker et al. (2008, 2009), Moos (2010), Cao & Kong (2012), Gold (2014) and Jessen et al. (2014). Examples of classical local segment-based vowel formant measurements are Rose (2002), Rose (2010) and Morrison et al. (2011). Local application of cepstral coefficients to specific vowels and consonants (also called segmental cepstra) is addressed in Rose et al. (2003) and Rose (2013a). Studies in which the dynamics of formants have been investigated in a forensic context include McDougall (2004, 2006), Morrison (2009a, 2011) and Zhang et al. (2011, 2013a, 2013b). Franco-Pedroso et al. (2012) show that curve fitting can also be applied to segmental-cepstra trajectories.

Although formants are usually approached semiautomatically (i.e., hand-measured or based on manually corrected formant tracks), there have been studies in which formant frequencies are measured automatically (De Castro et al. 2009; Chen et al. 2009; Gonzalez-Rodriguez 2011; Franco-Pedroso et al. 2013; Jessen et al. 2014).

Harrison (2013) provides general information on formant measurements in FSR and presents experimental results on aspects such as the influence of the formant tracking software and the analysis settings.

Fundamental frequency:

Hudson et al. (2007) show the difference between different average statistics (mean, median, mode) on global f0 measures. Kinoshita et al. (2009) and Kinoshita & Ishihara (2014) demonstrate how mean, variance,

skewness, kurtosis and related global f0 distribution parameters can be used as dimensions in a multidimensional feature vector and modelled with the MVKD formula. The f0 base level, and how it can be derived from a global f0 distribution, is addressed in Lindh & Eriksson (2007).

A study and literature discussion on the effect of vocal loudness on f0 is presented in Jessen et al. (2005).

Reynolds et al. (2002) show how the dynamics of f0 can be captured with delta coefficients. Reynolds et al. (2002) also describe a tokenisation process by which the f0 contour is scanned for rising and falling portions. Further tokenisation methods involving f0 and other prosodic features are proposed in Shriberg et al. (2005) and Dehak et al. (2007). A local and dynamic application of forensic f0-analysis in a tone language is shown in Li & Rose (2012). The potential use of linguistically motivated intonation models in FSR is addressed in Kraayeveld (1997) and Leemann et al. (2014b). Rose (2013b) demonstrates how intonation can be analysed in a concrete FSR case.

Combination of short-term spectral envelope features with fundamental frequency features at the speaker modelling stage using a Bayesian network approach is described by Arcienega et al. (2005).

Amplitude:
Compared to fundamental frequency, amplitude is much less commonly used as a feature in FSR. An example is shown by Reynolds et al. (2002) who worked with GMM modelling of a four-dimensional feature vector, consisting of the f0-curve, the amplitude curve and the respective delta curves.

Duration:
Discussion of the forensic aspects of Articulation Rate and relevant AR-population statistics are provided by Jessen (2007), Cao (2011) and Gold (2014). The percentage of the syllable that is vocalic or the percentage of the syllable that is voiced, as well as other measures of speech rhythm and timing have been investigated with a focus on speaker-discriminatory information by Leemann et al. (2014a).

A general survey on prosodic and other higher-level features in automatic speaker recognition is presented by Shriberg (2007) and Kockmann et al. (2011).

N-grams:
N-gram modelling is discussed in Shriberg (2007) and Kinnunen & Li (2010).

A proposal of how auditory features can be treated within the likelihood ratio framework is presented by Aitken & Gold (2013).

3.4 Speaker modelling and similarity scoring

The distinction between deterministic and statistical models is summarised, along with further literature in Drygajlo (2011) and addressed empirically in Drygajlo & Ugnat (2012). Text-dependent speaker recognition is explained in Hébert (2008) and text-independent speaker recognition in Reynolds & Campbell (2008). Among the very earliest proposals to adapt the GMM approach to the requirements of forensics are Meuwly et al. (1998), Meuwly & Drygajlo (2001) and Drygajlo et al. (2003).The GMM-UBM approach along with MAP was presented in Reynolds et al. (2000); more recent overviews include Reynolds & Campbell (2008). Ramos-Castro et al. (2006) presents a MAP-based procedure coping with data sparsity (e.g., only one suspected speaker recording). Normalisation methods, supervectors and many other aspects of automatic speaker recognition have been summarised in Kinnunen & Li (2010). Joint Factor Analysis and the i-vector approach are presented in Kenny et al. (2005, 2006, 2007), Kanagasundaram et al. (2011) and Dehak et al. (2009, 2011).

Some studies in which the GMM-UBM approach has been applied to FSASR are as follows: on long-term formant analysis Becker et al. (2008, 2009), Becker (2012), Jessen et al. (2014); on vowel formant centre frequencies Rose & Winter (2010); on formant trajectories with curve fitting Morrison (2011); on segmental cepstra of fricatives Rose (2011); on segmental cepstra and pole-zero estimates of nasals Enzinger et al. (2011).

The term MVLR (Multivariate Likelihood Ratio) has been used in Gonzalez-Rodriguez et al. (2007) and the term MVKD (Multivariate Kernel Density) in Morrison (2011). The method is the same and it is based on the procedures proposed by Aitken & Lucy (2004). The MKVD formula is presented and its essentials explained by Rose (2013a). Some studies in which the MVKD formula has been used in FSASR are as follows: on vowel formant centre frequencies Gonzalez-Rodriguez et al. (2007), Rose (2010), Morrison et al. (2011), Rhodes (2013); on formant trajectories with curve fitting Morrison (2009a, 2011), Zhang et al. (2013a,b), Rhodes (2013), Hughes (2014); on f0-distribution parameters Kinoshita et al. (2009), Kinoshita & Ishihara (2014); on segmental cepstra Rose (2013a); on long-term formants and articulation rate Gold (2014); on spectral moments in fricatives Kavanagh (2012) (cf. Jongman et al. 2000 on spectral moments in fricatives from a general phonetic perspective).

3.5 Calculation of likelihood ratio (LR)

The difference between the scoring method and the direct method, as well as the significance of the relevant population database, the suspected

speaker reference database and the suspected speaker control database is explained by Alexander & Drygajlo (2004) and Drygajlo (2007, 2009a,b, 2011, 2012). Further empirical results and elaborations are presented in Alexander (2005) and Gonzalez-Rodriguez et al. (2006).

3.6 Calibration and fusion

Calibration has two different aspects, calibration as a process of converting an uncalibrated to a calibrated likelihood ratio and calibration as a goodness criterion of a test set in a validation. For example, Morrison (2013) focuses on the former aspect and Ramos & Gonzalez-Rodriguez (2013) on the latter. Further references addressing one or both of these aspects of calibration are Brümmer & du Preez (2006), van Leeuwen & Brümmer (2007, 2013), Ramos et al. (2013) and Brümmer & Swart (2014). Calibration as a goodness criterion is addressed in 4.2. A study using calibration in both deterministic and statistical models is presented in Drygajlo & Ugnat (2012).

The method of logistic regression fusion, as it is frequently used, was introduced by Brümmer et al. (2007). A recent tutorial of logistic regression calibration and fusion is provided by Morrison (2013). Zhang et al. (2011) present an example in which LRs derived from FASR are fused with those derived from FSASR. An example of using cross validation when a separate development database for calibration and fusion is missing is shown in Morrison (2011).

Aside from fusion there are other proposals of how to combine evidence from different methods. Whereas fusion can be considered as a "back-end" combination process, there is a "front-end" proposal by Gold (2014) that is based on correlation tests of different features.

Another proposal of combining results from different methods has been to use Principal Component Analysis (PCA) as a means of providing an alternative to fusion by combining information from different features (Nair et al. 2014).

3.7 Mismatched recording conditions

Mismatch compensation methods are explained in some of the literature listed in 3.4. The principal Gaussian component compensation technique is proposed and studied in Alexander (2005). Further references to mismatch compensation with special emphasis on FSR include Alexander et al. (2004, 2005), Botti et al. (2004) and Alonso Moreno & Drygajlo (2012). One way to address the mismatch problem in FASR is by including features that are generally less sensitive to mismatch than short-term spectral envelope features (Arcienega et al. 2005). Methods of mismatch compensation that are compatible with the i-vector approach are summarised by Hansen & Hasan (2015). Forensic case simulations and LR-

testing based on mismatched conditions and their compensation are presented in Enzinger & Morrison (2015) and Enzinger et al. (2016).

3.8 Databases

As mentioned for 3.5, the first three types of databases in the list are explained by Alexander & Drygajlo (2004) and Drygajlo (2007, 2009a,b, 2011, 2012). The use of a development database in order to perform calibration and fusion is explained in Morrison (2013). Ramos-Castro et al. (2007) show how reference populations can be selected automatically.

4 Method Validation

4.1 Introduction

The basis for this chapter are the ENFSI "Guidelines for the single laboratory validation of instrumental and human based methods in forensic science" (QCC-VAL-002, 2014) and a guideline for the validation of likelihood ratio methods presented in Meuwly et al. (2016).

Further information on validation in forensic science is given by Haraksim (2014) and Haraksim et. al. (2015). One of the earliest presentations of validation in FASR were given by Meuwly et al. (2003). Some performance metrics and characteristics including Cllr (Log-Likelihood-Ratio cost) and APE plots (Applied Probability of Error) are summarised and explained in van Leeuwen & Brümmer (2007).

Validation has to be performed with speech that is typical of the speech material the forensic laboratory is confronted with in everyday work. Examples of validation based on forensic data include van Leeuwen et al. (2006), Ramos et al. (2008), Becker (2012), Becker et al. (2010, 2012), Solewicz et al. (2012) and van der Vloed et al. (2014).

4.2 Performance characteristics and metrics

Tippett plots were introduced in Evett & Buckleton (1996). They were proposed for forensic speaker recognition in Meuwly & Drygajlo (2001). Tippett plots occur in two versions in the literature. Tippett plots I are more common in the FASR literature (e.g., Drygajlo et al. 2003; van Leeuwen & Brümmer 2007; Ramos-Castro 2007), whereas Tippett plots II are more commonly used in the FSASR literature (e.g., Morrison 2010; Rose 2010; Gold 2014; Hughes 2014).

There are also differences in the literature about whether Tippett plots are referred to in singular or plural form. Figure 5, for example, would be referred to as Tippett plots (suggesting there are two plots in that figure) by some authors, whereas others refer to it as a Tippett plot. Among the former authors are Gonzalez-Rodriguez & Ramos (2007) and Ramos et al. (2013), whereas among the latter are Alexander (2005) and Morrison

(2010). In the present document the plural form is used. The same applies to APE plots and ECE plots.

The probabilities of misleading evidence ($PMEH_0$ and $PMEH_1$) are adapted from Neumann et. al. (2007), Ramos-Castro (2007) and Gonzalez-Rodriguez & Ramos (2007).The proportions trade-off (PTO) curve is related to the detection error trade-off (DET) plot (Martin et al. 1997 and van Leeuwen & Brümmer 2007), but does not include the concept of an error.

4.3 Performance characteristics and metrics based upon ranges of prior probabilities

The applied probability of error (APE) is described in Brümmer & du Preez (2006) and the use of logarithmic scoring rules is addressed in detail in Ramos-Castro (2007). The empirical cross-entropy (ECE) plot is explained and argued for in Ramos-Castro (2007), Ramos & Gonzalez-Rodriguez (2013) and Ramos et al. (2013).

5 Case Assessment

5.2 Preparatory aspects

An overview of the topic of cognitive bias in general is given by Kahneman (2011). Many publications about cognitive bias as applied to forensic science have been presented by Itiel Dror (Dror 2013 is one of the recent examples). The speaker identification group at the Netherlands Forensic Institute since about ten years has developed a method called blind grouping with the intention of preventing confirmation bias, which is one type of cognitive bias (Cambier-Langeveld et al. 2014, including further relevant references).

5.3 Relevant population

Discussion of the concept of relevant population and different ways of defining it is provided by Morrison et al. (2012, 2014), Morrison & Stoel (2014), Hughes (2014) and Hughes & Foulkes (2015).

5.4 Quantity and quality profile of the forensic audio material

5.4.2 Technical Quality. An example of a topic that has been addressed in this category is the effect of telephone filtering on the measurement of formant frequencies. Several studies show that the effect is small or absent for the second and third formant but systematic for the first formant (raising in telephone relative to studio speech), especially for high vowels, which linguistically have a low first formant (Byrne & Foulkes 2004; Lawrence et al. 2008). The effect of telephone transmission on the measurement and LR-based testing of formant dynamics has been studied in Zhang et al. (2013b).

5.4.3 Contextual aspects. There are several studies on the phonetic effects of different contextual (behavioural and situational) aspects available in the literature. The following list contains a small selection: Influence of stress (Hansen & Patil 2007; Kirchhübel 2013; Kirchhübel et al. 2011; Roberts 2012), influence of alcohol consumption (Chin & Pisoni 1997), increased vocal effort (Jessen et al. 2005; Hansen & Patil 2007), fatigue (Vogel et al. 2010), voice disguise (Rose & Simmons 1996; Wagner & Köster 1999; Künzel et al. (2004); Eriksson 2010), effects of facial concealment (Fecher 2014).

5.4.4 Mismatched conditions. Long-term non-contemporaneous speech is studied by Künzel (2007), Kelly et al. (2012, 2013), Kelly & Harte (2015) and Rhodes (2013). The importance of taking into account short-term non-contemporaneousness (session mismatch) is shown in Enzinger & Morrison (2012).

Language mismatch is addressed in van Leeuwen et al. (2006), Brümmer et al. (2007), Künzel (2013) and van der Vloed et al. (2014).

Effects of vocal effort including shouted speech and ways to compensate for it are presented by Hanilci et al. (2013a), Hanilci et al. (2013b) and Pohjalainen et al. (2014).

6 Evaluation and Interpretation

Tippett plots, introduced in forensic speaker recognition in Meuwly & Drygajlo (2001) and Meuwly (2001), at the present time provide a standard representation for performance evaluation of LR methods in FASR and FSASR cases (e.g., Drygajlo et al. 2003; van Leeuwen & Brümmer 2007; Ramos-Castro 2007; Morrison 2010; Rose 2010; Gold 2014; Hughes 2014). Examples of tests that are adapted to the specific conditions of a case are given by Rose (2013b), Enzinger & Morrison (2015) and Enzinger et al. (2016). Normally, they are completed by performance metrics such as probabilities of misleading evidence (Ramos-Castro 2007 and Gonzalez-Rodriguez & Ramos 2007) and log-likelihood-ratio cost (Cllr) (van Leeuwen & Brümmer 2007).

A version of the proposal to use case-specific probabilistic error as a way of presenting case-specific strength of evidence results is expressed by Solewicz et al. (2013).

7 Case File and Reporting

The Forensic Science Service (2007) provides a thorough introduction into expert witness reporting issues. Some expressions therein have been adopted for this document. In general, the basis for this chapter is the *ENFSI Guideline for evaluative reporting in forensic science, ENFSI, 2015*.

8 Quality Assurance

In general the required competencies to carry out the key responsibilities are defined in QCC-CAP-003 *Performance based standards for forensic science practitioners, ENFSI 2004*. Some concepts are also borrowed from the FIT-2005-*01 Guidelines for best practice in forensic examination of digital technology, ENFSI 2009*.

Alphabetic list:

Aitken, C. G. G. & Lucy, D. (2004): Evaluation of trace evidence in the form of multivariate data. *Applied Statistics* 53: 109-122.

Aitken C. G. G. & Taroni, F. (2004): *Statistics and evaluation of evidence for forensic scientists*. 2nded. Chichester: Wiley.

Aitken, C. G. G, Roberts, P. & Jackson, G. (2010): Fundamentals of probability and statistical evidence in criminal proceedings. Guidance for judges, lawyers, forensic scientists and expert witnesses. Practitioner guide no 1, Royal Statistical Society.

Aitken, C. & Gold, E. (2013): Evidence evaluation for discrete data. *Forensic Science International* 230: 147-155.

Alexander, A. (2005): *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*. PhD dissertation, EPFL Lausanne.

Alexander A., Botti, F. & Drygajlo, A. (2004): Handling mismatch in corpus-based forensic speaker recognition. *Proceedings of ODYSSEY 2004* (Toledo), pp. 69-74.

Alexander A. & Drygajlo, A. (2004): Scoring and direct methods for the interpretation of evidence in forensic speaker recognition. *Proceedings of the International Conference* on *Spoken Language* Processing *(ICSLP)* (Jeju, Korea), pp. 2397-2400.

Alexander, A., Dessimoz, D., Botti. F. & Drygajlo, A. (2005): Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech, Language and Law* 12: 214-234.

Alonso Moreno, V. & Drygajlo, A. (2012): A joint factor analysis model for handling mismatched recording conditions in forensic automatic speaker recognition. *Proceedings of the International Conference on Biometrics (ICB 2012)* (New Delhi), pp. 484-489.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G. & Vinyals, O. (2012): Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing* 20: 356-370.

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. *Proceedings of INTERSPEECH 2005* (Lisbon), pp. 2009-2012.

Avilés-Casco, C. V. (2011): *Robust diarization for speaker characterization*. PhD dissertation, Universidad de Zaragoza.

Becker, T. (2012): *Automatischer forensischer Stimmenvergleich*. Norderstedt: Books on Demand.

Becker, T., Jessen, M. & Grigoras, C. (2008): Forensic speaker verification using formant features and Gaussian mixture models. *Proceedings of INTERSPEECH 2008* (Brisbane), pp. 1505-1508.

Becker, T., Jessen, M. & Grigoras, C. (2009): Speaker verification based on formants using Gaussian mixture models. *Proceedings of NAG/DAGA 2009 (*Rotterdam), pp. 1640-1643.

Becker, T., Jessen, M., Alsbach, S., Broß, F. & Meier, T. (2010): SPES: The BKA automatic voice comparison system. *Proceedings of ODYSSEY 2010* (Brno, Czech Republic), pp. 58-62.

Becker, T., Solewicz, Y., Jardine, G. & Gfrörer, S. (2012): Comparing automatic forensic voice comparison systems under forensic conditions. *Proceedings of the Audio Engineering Society (AES) 46th International Conference* (Denver, Colorado), pp. 197-202.

Boersma, P. (2014): The use of Praat in corpus research. In: J. Durand, U. Gut & G. Kristoffersen (eds.) *The Oxford handbook of corpus phonology*. Oxford: Oxford University Press. pp. 342-360.

Botti, F., Alexander, A. & Drygajlo, A. (2004): On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition. *Forensic Science International*, volume 146 Supplement 1, pp. S101-S106.

Brümmer, N. & du Preez, J. (2006): Application-independent evaluation of speaker detection. *Computer Speech and Language* 20: 230-275.

Brümmer, N., Burget, L, Černocký, J., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D., Matějka, P., Schwarz, P. & Strasheim, A. (2007): Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing* 15: 2072-2084.

Brümmer, N. & Swart, A. (2014): Bayesian calibration for forensic evidence reporting. *Proceedings of INTERSPEECH 2014* (Singapore), pp. 388-392.

Byrne, C. & Foulkes, P. (2004): The 'mobile phone effect' on vowel formants. *The International Journal of Speech, Language and the Law* 11: 83-102.

Cambier-Langeveld, T. (2007): Current methods in forensic speaker identification: Results of a collaborative exercise. *The International Journal of Speech, Language and the Law* 14: 223-243.

Cambier-Langeveld, T., van Rossum, M. & Vermeulen, J. (2014): Whose voice is that? Challenges in forensic phonetics. In: J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N. O. Schiller & E. van Zanten (eds.) *Above and beyond the segments: Experimental linguistics and phonetics*. Amsterdam: Benjamins. pp. 14-27.

Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., & Matrouf, D. (2009): Forensic speaker recognition: A need for caution. *IEEE Signal Processing Magazine* 26: 95-103.

Campbell, J. P. (2014): Speaker recognition for forensic applications. Keynote Address at *ODYSSEY 2014* (Joensuu, Finland).

Cao, H. & Wang, Y. (2011): A forensic aspect of articulation rate variation in Chinese. *Proceedings of the International Congress of Phonetic Sciences 17* (Hong Kong), pp. 396-399.

Cao, H. & Kong, J. (2012): Speech length threshold in forensic speaker comparison by using long-term cumulative formant (LTCF) analysis. *Second International Conference on Instrumentation & Measurement, Computer, Communication and Control* (IMCCC) (Harbin), pp. 418-421.

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P. & Vair, C. (2008): Stream-based speaker segmentation using speaker factors and eigenvoices. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)* (Las Vegas), pp. 4133-4136.

Champod, C. & Meuwly, D. (2000): The inference of identity in forensic speaker recognition. *Speech Communication* 31:193-203.

Champod, C. & Evett, I. W. (2011): Evidence interpretation: a logical approach. *Wiley Encyclopedia of Forensic Science*. Published Online: 15 Dec 2011, DOI: 10.1002/9780470061589.fsa122.

Chen, N. F., Shen, W., Campbell, J. & Schwartz, R. (2009): Large-scale analysis of formant frequency estimation variability in conversational telephone speech. *Proceedings of INTERSPEECH 2009* (Brighton), pp. 2203-2206.

Chin, S. B. & Pisoni, D. (1997): *Alcohol and speech*. San Diego: Academic Press.

Cook, R., Evett, I. W., Jackson, G., Jones, P. J. & Lambert, J. A. (1998): A model for case assessment and interpretation. *Science & Justice* 38:151-156.

De Castro, A., Ramos, D. & Gonzalez-Rodriguez, J. (2009): Forensic speaker recognition using traditional features comparing automatic and

human-in-the-loop formant tracking. *Proceedings of INTERSPEECH 2009* (Brighton), pp. 2343-2346.

Dehak, N., Dumouchel, P. & Kenny, P. (2007): Modeling prosodic features with Joint Factor Analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 15: 2095-2103.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P. & Dumouchel, P. (2009): Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. *Proceedings of INTERSPEECH 2009* (Brighton), pp. 4237-4240.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. & Ouellet, P. (2011): Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19: 788-798.

Dror, I. E. (2013): The ambition to be scientific: Human expert performance and objectivity. *Science & Justice* 53: 81-82.

Drygajlo, A. (2007): Forensic automatic speaker recognition. *IEEE Signal Processing Magazine* 24: 132-135.

Drygajlo, A. (2009a): Statistical evaluation of biometric evidence in forensic automatic speaker recognition. In: Z. J. Geradts, K. Y. Franke & C. J. Veenman (eds.) *Computational Forensics*. Berlin: Springer. pp. 1-12.

Drygajlo, A. (2009b): Forensic evidence of voice, In: S. Z. Li (ed.) *Encyclopedia of Biometrics*. Berlin: Springer. pp. 1388-1395.

Drygajlo, A. (2011): Voice: Biometric analysis and interpretation of. *Wiley Encyclopedia of Forensic Science*. Published Online: 15 Dec 2011, DOI: 10.1002/9780470061589.fsa1034.

Drygajlo, A. (2012): Automatic speaker recognition for forensic case assessment and interpretation. In: A. Neustein & H. A. Patil (eds.) *Forensic speaker recognition. Law enforcement and counter-terrorism*. Berlin: Springer. pp. 21-39.

Drygajlo, A. (2013): Forensic automatic speaker recognition: Theory, implementation and practice. Tutorial at *INTERSPEECH 2013* (Lyon).

Drygajlo, A. (2014): From speaker recognition to forensic speaker recognition. In: V. Cantoni, D. Dimov, & M. Tistarelli (eds.) *Biometric Authentication: First International Workshop, BIOMET 2014, Sofia, Bulgaria, Revised Selected Papers*. Berlin: Springer, pp. 93-104.

Drygajlo, A. (2015): On methodological guidelines for automatic speaker recognition for the purpose of case assessment and interpretation. Keynote speech at *European Academy of Forensic Science Conference (EAFS 2015)* (Prague).

Drygajlo, A., Meuwly, D. & Alexander, A. (2003): Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. *Proceedings of EUROSPEECH 2003* (Geneva), pp. 689-692.

Drygajlo, A. & Ugnat, L. (2012): Comparative evaluation of calibrated deterministic and statistical models for forensic automatic speaker recognition systems. Presentation at the *European Academy of Forensic Science Conference (EAFS 2012)*, The Hague.

Enzinger, E., Balazs, P., Marelli, S. & Becker, T. (2011): A logarithmic based pole-zero vocal tract model estimation for speaker verification. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)* (Prague), pp. 4820-4823.

Enzinger, E. & Morrison, G. S. (2012): The importance of using between session test data in evaluating the performance of forensic-voice comparison systems. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology* (Sydney), pp. 137-140.

Enzinger, E. & Morrison, G. S. (2015): Mismatched distances from speakers to telephone in a forensic-voice-comparison case. *Speech Communication* 70: 28-41.

Enzinger, E., Morrison, G. S. & Ochoa, F. (2016): A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice* 56: 42-57.

Eriksson, A. (2010): The disguised voice: Imitating accents or speech styles and impersonating individuals. In: C. Llamas & D. Watt (eds.) *Language and identities*. Edinburgh: Edinburgh University Press. pp. 86-96.

Eriksson, A. (2012): Aural/acoustic vs. automatic methods in forensic phonetic case work. In: A. Neustein & H. A. Patil (eds.) *Forensic speaker recognition. Law enforcement and counter-terrorism*. Berlin: Springer. pp. 41-69.

Evett, I. W. & Buckleton, J. S. (1996): Statistical analysis of STR data. In: A. Carracedo, B. Brinkmann & W. Bär (eds.) *Advances in forensic haemogenetics*, vol. 6. Berlin: Springer. pp. 79-86.

Evett, I. W. (1998): Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice* 38: 198-202.

Fecher, N. (2014): *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants*. PhD dissertation, University of York.

Forensic Science Service (2007): A guide to the production of CJA statements. FSS-GP-171.

Foulkes, P. & French, P. (2012): Forensic speaker comparison: A linguistic-acoustic perspective. In: L. M. Solan & P M. Tiersma (eds.) *Oxford handbook of language and law*. Oxford: Oxford University Press. pp. 557-572.

Franco-Pedroso, J., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J. & Ramos, D. (2012): Fine-grained automatic speaker recognition using

cepstral-trajectories in phone units. In: C. Donohue, S. Ishihara & W. Steed (eds.) *Quantitative approaches to problems in linguistics. Studies in honour of Phil Rose*. München: LINCOM. pp. 185-195.

Franco-Pedroso, J., Espinoza-Cuadros, F. & Gonzalez-Rodriguez, J. (2013): Formant trajectories in linguistic units for text-independent speaker recognition. *Proceedings of the International Conference on Biometrics (ICB 2013)* (Madrid), pp. 1-6.

French, P. & Stevens, L. (2013): Forensic Speech Science. In: M. Jones & R. Anne-Knight (eds.) *The Bloomsbury companion to phonetics*. London: Continuum. pp. 183-197.

Gfroerer, S. (2003): Auditory-instrumental forensic speaker recognition. *Proceedings of EUROSPEECH 2003* (Geneva), pp. 705-708.

Gold, E. (2014): *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. PhD dissertation, University of York.

Gold, E. & French, P. (2011): International practices in forensic speaker comparison. The *International Journal of Speech, Language and the Law* 18: 293-307.

Gonzalez-Rodriguez, J. (2011): Speaker recognition using temporal contours in linguistic units: the case of formant and formant-bandwidth trajectories. *Proceedings of INTERSPEECH 2011* (Florence), pp. 133-136.

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. *Computer Speech and Langua*ge 20: 331-355.

Gonzalez-Rodriguez, J. & Ramos, D. (2007): Forensic automatic speaker classification in the "coming paradigm shift". In: C. Müller (ed.) *Speaker classification I: Fundamentals, features, and methods*. Berlin: Springer. pp. 205-217.

Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T. & Ortega-Garcia, J. (2007): Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 15: 2104-2115.

Hanilci, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P. & Ertas, F. (2013a): Speaker identification from shouted speech: Analysis and compensation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)* (Vancouver), pp. 8027-8031.

Hanilci, C., Kinnunen, T., Rajan, P., Pohjalainen, J., Alku, P. & Ertas, F. (2013b): Comparison of spectrum estimators in speaker verification:

Mismatch conditions induced by vocal effort. *Proceedings of INTERSPEECH 2013* (Lyon), pp. 2881-2885.
Hansen, J. H. L. & Patil, S. (2007): Speech under stress: Analysis, modelling and recognition. In: C. Müller (ed.) *Speaker classification I: Fundamentals, features, and methods*. Berlin: Springer. pp. 108-137.
Hansen J. H. L., & Taufiq, H. (2015): Speaker recognition by machines and humans. *IEEE Signal Processing Magazine* 32: 74-99.
Haraksim, R. (2014): *Validation of likelihood ratio methods used in forensic evidence evaluation: Application in forensic fingerprints*. PhD dissertation, University of Twente, Enschede.
Haraksim, R., Ramos, D., Meuwly, D. & Berger, C. E. H. (2015): Measuring coherence of computer-assisted likelihood ratio methods. *Forensic Science International* 249: 123-132.
Harrison, P. T. (2013): *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurements*. PhD dissertation, University of York.
Hébert, M. (2008): Text-dependent speaker recognition. In: J. Benesty, M. M. Sondhi & Y. Huang (eds.) *Springer handbook of speech processing*. Berlin: Springer. pp. 743-762.
Hollien, H. (2002): *Forensic voice identification*. San Diego: Academic Press.
Hudson, T., de Jong, G., McDougall, K., Harrison, P. & Nolan, F. (2007): F0 statistics for 100 young male speakers of Standard Southern British English. *Proceedings of the International Congress of Phonetic Sciences 16* (Saarbrücken), pp. 1809-1812.
Hughes, V. (2014): *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. PhD dissertation, University of York.
Hughes, V. & Foulkes, P. (2015): The relevant population in forensic voice comparison: effects of varying delimitations of social class and age. *Speech Communication* 66: 218-230.
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. *Science & Justice* 46: 33-44.
Jackson, G., Aitken, C. & Roberts, P. (2015): Case assessment and interpretation of expert evidence. Guidance for judges, lawyers, forensic scientists and expert witnesses. Practitioner guide no 4.
Jessen, M. (2007): Forensic reference data on articulation rate in German. *Science & Justice* 47: 50-67.
Jessen, M. (2008): Forensic phonetics. *Language and Linguistics Compass* 2: 671-711.

Jessen, M., Köster, O. & Gfroerer, S. (2005): Influence of vocal effort on average and variability of fundamental frequency. *The International Journal of Speech, Language and the Law* 12: 174-213.

Jessen, M., Alexander, A. & Forth, O. (2014): Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software. *Proceedings of the Audio Engineering Society (AES) 54th International Conference* (London), pp. 28-35.

Jongman, A., Wayland, R. & Wong, S. (2000): Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America* 108: 1252-1263.

Kahneman, D. (2011): *Thinking, fast and slow*. London etc.: Penguin (published with Penguin Books 2012).

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S. & Mason, M. (2011): i-vector based speaker recognition on short utterances. *Proceedings of INTERSPEECH 2011*(Florence), pp. 2341-2344.

Kavanagh, C. M. (2012): *New consonantal acoustic parameters for forensic speaker comparison*. PhD dissertation, University of York.

Kelly, F., Drygajlo, A. & Harte, N. (2012): Speaker verification with long-term ageing data. *Proceedings of the International Conference on Biometrics (ICB 2012)* (New Delhi), pp. 478-483.

Kelly F., Drygajlo, A., & Harte, N. (2013): Speaker verification in score-ageing-quality classification space. *Computer Speech and Language* 27: 1068-1084.

Kelly F. & Harte, N. (2015): Forensic comparison of ageing voices from automatic and auditory perspectives. *The International Journal of Speech, Language and the Law* 22: 167-202.

Kenny, P., Boulianne, G., Quellet, P. & Dumouchel, P. (2005): Factor analysis simplified. *Proceedings o*f *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)* (Philadelphia), pp. 637-640.

Kenny, P., Boulianne, G., Quellet, P. & Dumouchel, P. (2006): Improvements in factor analysis based speaker verification. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP 2006)* (Toulouse), pp. 113-116.

Kenny, P., Boulianne, G. & Dumouchel, P. (2007): Joint Factor Analysis versus Eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 15: 1435-1447.

Kinnunen, T. & Li, H. (2010): An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication* 52: 12-40.

Kinoshita, Y., Ishihara, S. & Rose, P. (2009): Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker

recognition. *The International Journal of Speech, Language and the Law* 16: 91-111.

Kinoshita, Y. & Ishihara, S. (2014): Background population: how does it affect LR-based forensic voice comparison? *The International Journal of Speech, Language and the Law* 21: 191-224.

Kirchhübel, C., Howard, D. M. & Stedmon, A. W. (2011): Acoustic correlates of speech when under stress: Research, methods and future directions. *The International Journal of Speech, Language and the Law* 18: 75-98.

Kirchhübel, C. (2013): *The acoustic and temporal characteristics of deceptive speech*. PhD dissertation, University of York.

Kockmann, M., Ferrer, L., Burget L., Shriberg, E. & Černocky, J. (2011): Recent progress in prosodic speaker verification. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)* (Prague), pp. 4556-4559.

Kraayeveld, J. (1997): *Idiosyncrasy in prosody: Speaker and speaker group identification in Dutch using melodic and temporal information*. PhD dissertation, Catholic University of Nijmegen.

Künzel, H. J. (2007): Non-contemporary speech samples: Auditory detectability of an 11 year delay and its effect on automatic speaker identification. *The International Journal of Speech, Language and the Law* 14: 109-136.

Künzel, H. J. (2013): Automatic speaker recognition with cross-language speech material. *The International Journal of Speech, Language and the Law* 20: 21-44.

Künzel, H. J. & Gonzalez-Rodriguez, J. (2003): Combining automatic and phonetic-acoustic speaker recognition techniques for forensic applications. *Proceedings of the International Congress of Phonetic Sciences 17* (Barcelona), pp. 1619-1622.

Künzel, H. J., Gonzalez-Rodriguez, J. & Ortega-García, J. (2004): Effect of voice disguise on the performance of a forensic automatic speaker recognition system. *Proceedings of ODYSSEY 2004* (Toledo), pp. 153-156.

Lawrence, S., Nolan, F. & McDougall, K. (2008): Acoustic and perceptual effects of telephone transmission on vowel quality. *The International Journal of Speech, Language and the Law* 15: 161-192.

Leemann, A., Kolly, M.-J. & Dellwo, V. (2014a): Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International* 238: 59-67.

Leemann, A., Dellwo, V., Mixdorff, H., O'Reilly, M. & Kolly, M.-J. (2014b): Speaker-individuality in Fujisaki model f0 features: Implications for forensic voice comparison. *The International Journal of Speech, Language and the Law* 21: 343-370.

Li, J. & Rose, P. (2012): Likelihood ratio-based forensic voice comparison with F-pattern and tonal f0 from the Cantonese /ɔy/ diphthong. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology* (Sydney), pp 201-204.

Lindh, J. & Eriksson, A. (2007): Robustness of long time measures of fundamental frequency. *Proceedings of INTERSPEECH 2007* (Antwerpen), pp. 2025-2028.

Mak, M.-W. & Yu, H.-B. (2014): A study of voice activity detection techniques for NIST speaker recognition evaluations. *Computer Speech and Language* 28: 295-313.

Martin, A., Doddington, G., Kamm, T., Ordowski, M. & Przybocki, M. (1997): The DET curve in assessment of detection task performance. *Proceedings of EUROSPEECH 1997* (Rhodes, Greece), pp. 1895-1898.

McDougall, K. (2004): Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *The International Journal of Speech, Language and the Law* 11: 103-130.

McDougall, K. (2006): Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *The International Journal of Speech, Language and the Law* 13: 89-126.

Meuwly, D. (2001): *Reconnaissance automatique de locuteurs en sciences forensiques: l'apport d'une approche automatique*. PhD dissertation, EPFL Lausanne.

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, pp. 52-55.

Meuwly, D. & Drygajlo, A. (2001): Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM). *Proceedings of ODYSSEY 2001* (Crete), pp. 145-150.

Meuwly D., Goode, A., Drygajlo, A., Gonzalez-Rodriguez, J., & Lucena Molina, J., (2003): Validation of forensic automatic speaker recognition systems: Evaluation frameworks for intelligence and evidential purposes. Proceedings of the European Academy of Forensic Science Meeting 2003, Istanbul, in: *Forensic Science International*, Vol. 136/Suppl. 1, p. 364.

Meuwly, D., Haraksim, R. & Ramos, D. (2016): A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. To appear in *Forensic Science International*.

Moos, A. (2010): Long-Term Formant Distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician* 101/102: 7-24.

Morrison, G. S. (2009a): Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America* 125: 2387-2397.

Morrison, G. S. (2009b): Forensic voice comparison and the paradigm shift. *Science & Justice* 49: 298-308.

Morrison, G. S. (2010): *Forensic voice comparison*. In: I. Freckelton & H. Selby (eds.) *Expert evidence* (Chapter 99). Sydney: Thomson Reuters.

Morrison, G. S. (2011): A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model - universal background model (GMM-UBM). *Speech Communication* 53: 242-256.

Morrison, G. S. (2013): Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences* 45: 173-197.

Morrison, G. S. (2014): Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice* 54: 245-256.

Morrison, G. S., Zhang, C. & Rose, P. (2011): An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International* 208: 59-65.

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012): Database selection for forensic voice comparison. *Proceedings of ODYSSEY 2012* (Singapore), pp. 62-77.

Morrison, G. S. & Enzinger, E. (2013): Forensic Speech Science. In: 17th *Interpol International Forensic Science Managers Symposium* (Lyon), pp. 616-623.

Morrison, G. S. & Stoel, R. D. (2014): Forensic strength of evidence statements should preferably be likelihood ratios calculated using relevant data, quantitative measurements, and statistical models – a response to Lennard (2013) Fingerprint identification: How far have we come? *Australian Journal of Forensic Sciences* 46: 282-292.

Nair, B., Alzqhoul, E. & Guillemin, B. J. (2014): Determination of likelihood ratios for forensic voice comparison using Principal Component Analysis. The *International Journal of Speech, Language and the Law* 21: 83-112.

Nakasone, H. (2003): Automated speaker recognition in real world conditions: Controlling the uncontrollable. *Proceedings of EUROSPEECH 2003* (Geneva), pp. 697-700.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A. & Bromage-Griffiths, A. (2007): Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences* 52: 54-64.

Neustein, A. &. Patil, H. A. (eds.) (2012): *Forensic speaker recognition. Law enforcement and counter-terrorism*. Berlin: Springer.

Nolan, F. (1983): *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nolan, F. (1997): Speaker recognition and forensic phonetics. In: W. J. Hardcastle & J. Laver (eds.) *The handbook of phonetic sciences*. Oxford: Blackwell. pp. 744-767.

Nolan, F. & Grigoras, C. (2005): A case for formant analysis in forensic speaker identification. *The International Journal of Speech, Language and the Law* 12: 143-173.

Pohjalainen, J., Hanilçi, J., Kinnunen, T. & Alku, P. (2014): Mixture Linear Prediction in speaker verification under vocal effort mismatch. *IEEE Signal Processing Letters* 21: 1516-1520.

Ramos-Castro, D. (2007): *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD dissertation, Universidad Autónoma de Madrid.

Ramos-Castro, D., Gonzalez-Rodriguez, J., Montero-Asenjo, A. & Ortega-Garcia, J. (2006): Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation. *Proceedings of ODYSSEY 2006* (San Juan), pp. 1-5.

Ramos-Castro, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J. & Ortega-Garcia, J. (2007): Speaker verification using speaker- and test-dependent fast score normalization. *Pattern Recognition Letters* 28: 90-98.

Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J. & Lucena-Molina, J. J. (2008): Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish. *Proceedings of INTERSPEECH 2008* (Brisbane), pp. 1493-1496.

Ramos, D. & Gonzalez-Rodriguez, J. (2013): Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International* 230: 156-169.

Ramos, D., Gonzalez-Rodriguez, J., Zadora, G. & Aitken, C. (2013): Information-theoretical assessment of the performance of likelihood ratio models. *Journal of Forensic Sciences* 58: 1503-1518.

Reynolds, D., Quatieri, T. & Dunn, R. (2000): Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10: 19-41.

Reynolds, D., Andrews, W., Campbell, J., Navrátil, J., Peskin, B., Adami, A., Jin, Q., Klusáček, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D. & Xiang, B. (2002): SuperSID Project Final Report. Exploiting high-level information for high-performance speaker recognition. [http://www.clsp.jhu.edu/vfsrv/ws2002/groups/supersid/SuperSID_Final_Report_CLSP_WS02_2003_10_06.pdf]

Reynolds, D. & Torres-Carrasquillo, P. (2005): Approaches and applications of audio diarization. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)* (Philadelphia), pp. 953-956.

Reynolds, D. A. & Campbell, W. M. (2008): Text-independent speaker recognition. In: J. Benesty, M. M. Sondhi & Y. Huang (eds.) *Springer handbook of speech processing*. Berlin: Springer. pp. 763-781.

Rhodes, R. (2013): *Assessing non-contemporaneous forensic speech evidence: Acoustic features, formant frequency-based likelihood ratios and ASR performance*. PhD dissertation, University of York.

Roberts, L. S. (2014): *A forensic phonetic study of the vocal responses of individuals in distress*. PhD dissertation, University of York.

Robertson, B. & Vignaux, G. A. (1995): *Interpreting evidence. Evaluating forensic science in the courtroom*. Chichester etc.: Wiley.

Rose, P. (2002): *Forensic speaker identification*. London: Taylor & Francis.

Rose, P. (2003): *The technical comparison of forensic voice samples*. In: I. Freckelton & H. Selby (eds.) *Expert evidence* (Chapter 99). Sydney: Thompson Lawbook Co.

Rose, P. (2006): Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech and Language* 20: 159-191.

Rose, P. (2010): The effect of correlation on strength of evidence estimates in forensic voice comparison: uni- and multivariate likelihood ratio-based discrimination with Australian English vowel acoustics. *International Journal of Biometrics* 2: 316-329.

Rose, P. (2011): Forensic voice comparison with secular shibboleths – A hybrid fused GMM-multivariate likelihood ratio-based approach using alveolo-palatal fricative cepstral spectra. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)* (Prague), pp. 5900-5903.

Rose, P. (2013a): More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *The International Journal of Speech, Language and the Law* 20: 77-116.

Rose, P. (2013b): Where the science ends and the law begins: likelihood ratio-based forensic voice comparison in a $150 million telephone fraud. *The International Journal of Speech, Language and the Law* 20: 277-324.

Rose, P. & Simmons, A. (1996): F-pattern variability in disguise and over the telephone. Comparisons for forensic speaker identification. *Proceedings of the 6th Australian International Conference on Speech Science and Technology* (Adelaide), pp. 121-126.

Rose, P., Osanai, T. & Kinoshita, Y. (2003): Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based

segmental discrimination with a Bayesian likelihood ratio as threshold. *The International Journal of Speech, Language and the Law* 10: 179-202.

Rose, P. & Winter, E. (2010): Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analysis. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology* (Melbourne), pp. 42-45.

Sahidullah, M. & Saha, G. (2012): Comparison of speech activity detection techniques for speaker recognition, http://arxiv.org/pdf/1210.0297.pdf.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A. & Stolcke, A. (2005): Modeling prosodic feature sequences for speaker recognition. *Speech Communication* 46: 455-472.

Shriberg, E. (2007): Higher-level features in speaker recognition. In: C. Müller (ed.) *Speaker classification I: Fundamentals, features, and methods*. Berlin: Springer. pp. 241-259.

Solewicz, Y. A., Becker, T., Jardine, G. & Gfrörer, S. (2012): Comparison of speaker recognition systems on a real forensic benchmark. *Proceedings of ODYSSEY 2012* (Singapore), pp. 86-91.

Solewicz, Y. A., Jardine, G., Becker, T. & Gfrörer, S. (2013): Estimated intra-speaker variability boundaries in forensic speaker recognition casework. *Proceedings of Biometric Technologies in Forensic Science (BTFS)* (Nijmegen), pp. 31-33.

Tranter, S. E. & Reynolds, D. A. (2006): An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing* 14: 1557-1565.

Van der Vloed, D., Bouten, J. & Van Leeuwen, D. A. (2014): NFI-FRITS: A forensic speaker recognition database and some first experiments. *Proceedings of ODYSSEY 2014* (Joensuu, Finland), pp. 6-13.

Van Leeuwen, D. A, Martin, A. F., Przybocki, M. A. & Bouten, J. S. (2006): NIST and NFI-TNO evaluations of automatic speaker recognition. *Computer Speech and Language* 20: 128-158.

Van Leeuwen, D. A. & Brümmer, N. (2007): An introduction to application-independent evaluation of speaker recognition systems. In: C. Müller (ed.) *Speaker classification I: Fundamentals, features, and methods*. Berlin: Springer. pp. 330-353.

Van Leeuwen, D. & Brümmer, N. (2013): The distribution of calibrated likelihood-ratios in speaker recognition. *Proceedings of INTERSPEECH 2013* (Lyon), pp. 1619-1623.

Vogel, A. P., Fletcher, J. & Maruff, P. (2010): Acoustic analysis of the effects of sustained wakefulness on speech. *Journal of the Acoustical Society of America* 128: 3747-3756.

Wagner, I. & Köster, O. (1999): Perceptual recognition of familiar voices using falsetto as a type of voice disguise. *Proceedings of the*

International Congress of Phonetic Sciences 14 (San Francisco), pp. 1381-1384.

Watt, D. (2010): The identification of the individual through speech. In: C. Llamas & D. Watt (eds.) *Language and identities*. Edinburgh: Edinburgh University Press. pp. 76-85.

Zhang, C., Morrison, G. S. & Thiruvaran, T. (2011): Forensic voice comparison using Chinese /iau/. *Proceedings of the International Congress of Phonetic Sciences 17* (Hong Kong), pp. 2280-2283.

Zhang, C., Morrison, G. S., Ochoa, F. & Enzinger, E. (2013a): Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. *Journal of the Acoustical Society of America* 133, EL54-EL60.

Zhang, C., Morrison, G. S., Enzinger, E. & Ochoa, F. (2013b): Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – female voices. *Speech Communication* 55: 796-813.

Appendix 2: List of Abbreviations

APE:	Applied Probability of Error
AR:	Articulation Rate
CAP:	Competence Assurance Project (of the ENFSI QCC)
CDF:	Cumulative Distribution Function
CE:	Collaborative Exercise
Cllr:	Log-likelihood-ratio cost
$Cllr^{min}$:	Minimum Cllr
$Cllr^{cal}$:	Calibration loss
DET:	Detection Error Trade-off
DS:	Different Speaker
DTW:	Dynamic Time Warping
E:	Observed Evidence
ECE:	Empirical Cross Entropy
ECE^{min}:	Minimum ECE
EER:	Equal Error Rate
ENFSI:	European Network of Forensic Science Institutes
EPP:	Equal Proportion Probability
f0:	fundamental frequency
FASR:	Forensic Automatic Speaker Recognition
FSAAWG:	Forensic Speech and Audio Analysis Working Group (of ENFSI)
FSASR:	Forensic Semiautomatic Speaker Recognition
FSR:	Forensic Speaker Recognition
GMM:	Gaussian Mixture Model
H:	Hypothesis
HMM:	Hidden Markov Model
iCDF:	inverse Cumulative Distribution Function
JFA:	Joint Factor Analysis
LDA:	Linear Discriminant Analysis
LR:	Likelihood Ratio
$LR_{case:}$	case-specific Likelihood Ratio
LTF:	Long Term Formants
MAP:	Maximum A Posteriori
MFCC:	Mel Frequency Cepstral Coefficients
MVLR:	Multivariate Likelihood Ratio method (same as MVKD)
MVKD:	Multivariate Kernel Density method (same as MVLR)
P:	probability
p:	likelihood
PAV:	Pool Adjacent Violators
PCA:	Principal Component Analysis
PLDA:	Probabilistic Linear Discriminant Analysis

PLPCC:	Perceptual Linear Prediction Cepstral Coefficients
PME:	Probability of Misleading Evidence
PT:	Proficiency Test/Testing
PTO:	Proportions trade-off
QCC:	ENFSI Standing Committee for Quality and Competence
RASTA:	Relative Spectra
R&D:	Research and Development
SNR:	Signal to Noise Ratio
SOP:	Standard Operating Procedure
SS:	Same Speaker
T-norm:	Test normalisation
UBM:	Universal Background Model
VQ:	Vector Quantisation
Z-norm:	Zero normalisation

Part 2:

Guidance on the Conduct of Proficiency Testing and Collaborative Exercises for Forensic Semiautomatic and Automatic Speaker Recognition

1 Introduction

According to ISO/IEC 17025 accredited laboratories shall have quality control procedures for monitoring the validity of tests undertaken. This monitoring includes regular participation in proficiency tests (PTs), i.e., inter-laboratory comparisons used for the identification of the laboratory performance in tests and for observations of the laboratories' long-term performance. Collaborative exercises (CEs) can be used as additional means of inter-laboratory comparison, but address specific issues and can be open outcome procedures.

If PTs and CEs are carried out within the context of an accredited quality assurance program, they can be regarded as a valuable instrument to reflect the quality of the laboratories' test results, as described by ISO/IEC 17025. EWG (Expert Working Group) member laboratories should carry out PTs and CEs in accordance with the recommendations of the ENFSI documents *"Policy on proficiency tests and collaborative exercises within ENFSI"* (*QCC-PTCE-002)* and *"Guidance on the conduct of proficiency tests and collaborative exercises within ENFSI" (QCC-PT-001).*

For ENFSI member laboratories organised in the Forensic Speech and Audio Analysis Working Group (FSAAWG) using FASR and FSASR, inter-laboratory comparisons may not always be feasible because of crucial differences between laboratories. The differences are related in particular to specific procedures, parameters or requirements, e.g., the language involved, technical issues of the recorded speech or the amount of net speech needed for analysis, etc. For this reason single laboratory validation may serve as equivalent quality control, such as the regular use of different speaker corpora, or re-application of tests using the same or different systems. In general terms, schemes for single laboratory validation are described in the ENFSI document "*Guidelines for the single laboratory validation of instrumental and human based methods in forensic science" (QCC-VAL-002).*

2 Aims

The purpose of this document (Part 2 of this book) is to provide guidelines for the members of the ENFSI Expert Working Group for Forensic Speech and Audio Analysis (FSAAWG) as an aid on how to organise effective proficiency tests (PTs) and collaborative exercises (CEs) for FASR and FSASR. These guidelines apply to PTs and CEs conducted by an ENFSI

FSAAWG member laboratory and to PTs and CEs outsourced by the FSAAWG to an external provider.

The aim of inter-laboratory PTs and CEs as well as single laboratory validation in FASR and FSASR is to enable laboratories (to be referred to as "participants" in this document) that undertake forensic speaker recognition examinations using FASR and FSASR techniques to reflect and enhance the quality of their tests and demonstrate that they are fit for the purpose.

3 Principles of PTs and CEs in FASR and FSASR

The principles of Forensic Automatic and Semiautomatic Speaker Recognition methods are described in the document *"Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition"*(Part 1 of this book).

In line with the definitions set out in the *"Guidance on the conduct of proficiency tests and collaborative exercises within ENFSI",* a PT in FASR and FSASR evaluates the performance of participants in an inter-laboratory comparison against pre-established performance indices that contain a pass/fail criterion. These performance indices are called validation criteria. Validation criteria specify certain numerical values of performance metrics such as Probabilities of Misleading Evidence (PMEs), Equal Proportion Probability (EPP) or Log-likelihood-ratio cost (Cllr).

CE in FASR and FSASR is an inter-laboratory comparison that does not necessarily result in a performance assessment but rather focusses on specific issues of FASR and FSASR methodology or application, such as the effect of different extracted acoustic features, the impact of different populations or of different scoring and direct methods.

In FASR and FSASR a PT can also be combined with a CE in such a way that for participants with comparable systems and languages the trial run is a PT with a pass/fail criterion and for those with incomparable systems and/or languages the same trial run can be performed simultaneously in an open outcome procedure as a CE to compare and investigate to what extent the differences affect the examination results.

4 Reference Documents

The following reference documents provide information on the conduct of proficiency tests and collaborative exercises for FASR and FSASR. The guidelines are based on these documents:

- Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition (2015) (Part 1 of this book)
- ENFSI QCC-PTCE-002:2015 Policy on proficiency tests and collaborative exercises within ENFSI
- ENFSI QCC-PT-001:2014 Guidance on the conduct of proficiency tests and collaborative exercises within ENFSI
- ENFSI QCC-VAL-002:2015 Guidelines for the single laboratory validation of instrumental and human based methods in forensic science
- ISO/IEC 17025:2005 General requirements for the competence of testing and calibration laboratories

5 Definitions

For definitions and terminology refer to the documents *"Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition" (2015) and "Guidance on the conduct of proficiency tests and collaborative exercises within ENFSI" (QCC-PT-001).*

6 Responsibilities and Roles

6.1 Forensic Speech and Audio Analysis Working Group

The Forensic Speech and Audio Analysis Working Group (FSAAWG) is responsible for the provision and promotion of PTs and CEs in FASR and FSASR and for the selection of the Provider.

6.2 Provider and Coordinator

The Provider could be an ENFSI member laboratory or a commercial body or a public body. The Provider appoints the Coordinator. The Coordinator

has the responsibility for organising and managing a PT/CE scheme and should seek consulting from an Advisory Group throughout test administration and evaluation.

6.3 Advisory Group

Technical directions and advice should be provided by an Advisory Group made up of FSAAWG representatives.

The statutes of the Advisory Group are:

- to define the objective of the trial
- to specify the type and composition of the testing material (e.g., language spoken, speech style, duration of questioned speaker and suspected speaker recordings)
- to define the performance parameters that will be used (i.e., the performance metrics)
- to define the expected outcomes (i.e., the validation criteria)
- to advise on the best way to organise the exercise
- to prepare for the evaluation of participants' examination results
- to advise on the assessment of the results and the content of the feedback for the participants
- to represent the opinion of the ENFSI FSAAWG
- to provide instructions to the Coordinator on technical and other issues
- to assess the results obtained and examine the implications they have
- to help adapting new test routines
- to offer, if necessary, suggestions or solutions regarding problems of the test procedure

7 Trial Organisation

7.1 Timescale and structure

It is recommended that one PT or CE is organised per year by the FSAAWG network.

The structure is as follows:

- Creation, compilation and quality control of the speech test samples

- Dissemination of the speech test samples and instructions to the participants
- Invitation of the participants to analyse the speech test data and report the examination results to the Provider
- Analysis of the examination results and their comparison among the participating laboratories
- Compilation and distribution of reports to participants
- Review and identification of requirements for future tests

7.2 Frequency of Participation

Annual participation in a PT or CE, if available, is recommended for FSAAWG member laboratories.

7.3 Confidentiality

In order to guarantee confidentiality, singular laboratory reference codes (Lab-ID) are assigned to each participant. This permits results to be reported without revealing the identities of the participating laboratories.

7.4 Test Development

The Advisory Group continuously attempts to improve PT and CE schemes and to introduce new suggestions.

8 Trial Preparation

FSAAWG member laboratories may have different FASR and FSASR systems in use. Hence before specifying the type and composition of the test material and determining the performance metrics and validation criteria the Coordinator must ascertain information from prospective test participants about the characteristics of the audio material specified in their particular Standard Operation Procedures (e.g., language, duration, audio quality) as well as about their performance metrics and validation criteria (e.g., PMEs, EPP or Cllr at a given value or range of values).

Based on this information, the Coordinator in association with the Advisory Group makes a decision on the test material and the performance metrics

and validation criteria that reflect the dominant conditions among the prospective participants.

9 Preparation of Test Materials

Test materials in the FSAAWG consist of speech data and must be prepared in a controlled process. PTs and CEs should reflect casework conditions as closely as possible. Therefore speech test materials should be as similar as possible to those routinely tested by participating laboratories.

10 Participants' Results

10.1 Trial procedure

Participants are requested to examine the test materials in the same manner as routine samples. They should not give the trial any special treatment that would not be given in the same circumstances to casework. The analysis of the test materials should be performed – if possible – with the same settings and procedures as used for routine casework.

10.2 Data reporting

The PT and CE examination results obtained by means of FASR and FSASR have to be submitted to the Provider in a format that contains all relevant information.

It is mandatory that every participating laboratory provides a numerical examination result for each single FASR or FSASR trial.

Participants are requested to report their performance metrics (e.g., PMEs, EPP, Cllr) and validation criteria that they use to characterise their FASR or FSASR casework performance. This information is important as the results may be evaluated and reported according to these performance parameters.

Participants are requested to return their results by the given deadline.

Examination results received after the deadline are not evaluated and included in the feedback report.

11 Assessment of Performance

The laboratories´ examination results are collected and analysed by the Provider.

Satisfactory examination results are achieved if the PT meets the predefined validation criteria.

The outcome of a CE can provide useful information for further investigation and application or for a subsequent PT.

12 Feedback to Participants

Feedback reports should be completed within a short period and sent to participants by email no later than two months after the deadline for the results. The participants' results are identified by the Lab-ID. Certificates for each participant including the obtained performance results are provided with the feedback reports.

13 Examples

Four examples are constructed that give a general impression on how a PT or CE can be applied to FASR or FSASR methods.

13.1 Example 1: PT on FASR

At the outset the Coordinator requests from the prospective PT participants the system requirements of their methods, e.g., which speech stylistic and technical conditions are required, which speech signal duration is needed and which languages can be processed. The outcome of this inquiry shows that most prospective participants can process telephone conversations in a forensically realistic spontaneous speech style (including increased vocal loudness and emotional components). As for the language, only some participants mention that they are able to analyse languages other than their national language, but among those who do not, all would be able to analyse other languages if provided with a proper database. Asked for their performance metrics, most participants mention that EPP is used as one of

their performance metrics. As EPP-values for the specified PT conditions, they indicate to have a range of probabilities from 0.07 to 0.3 obtained in their internal method validations.

The Coordinator in consultation with the Advisory Group decides to compile a PT test set with 50 forensically realistic telephone conversation sides involving 25 different speakers with two non-contemporaneous speech samples each. The language spoken in the recordings is Turkish. Along with the test data a database of a 30 recordings spoken in Turkish is provided, each containing a different speaker. The Coordinator decides that in order to pass the PT, the EPP must be equal or lower than 0.25. In this validation criterion the few prospective participants who mention EPPs ranging from of 0.25 to 0.3 cannot be accommodated. The validation criterion is based on the fact that the great majority of prospective participants reported EPP values smaller than 0.2 and it is also based on consulting the research literature about which EPP might approximately be expected under the specified conditions. EPP of 0.25 instead of 0.2 is used as validation criterion because of the increased difficulty (among all or the great majority of participants) of analysing a language other than their national language.

After the test has been carried out and the Coordinator has received the examination results and the accompanying reports, it turns out that different participants have used the 30-speaker database differently, some to create a UBM (Universal Background Model), some as reference population and some in yet other ways. This is interesting information that can help inform future PTs or that can be further addressed in a CE.

13.2 Example 2: CE on FASR

Some prospective participants mention that they use Cllr as validation criterion in addition to or instead of using EPP. In order to calculate Cllr meaningfully, the test outputs have to be calibrated likelihood ratios (LRs). The Coordinator asks the participants which kind of data they need to produce calibrated LR. With respect to the situation addressed in Example 1, where no common-ground language can be found, the Coordinator asks whether the participants could report calibrated LRs if provided with the database of 30 different speakers mentioned in Example 1. Some of the participants answer affirmatively whereas others answer that in addition they need a development database for calibration purposes. Such a development database would require recordings from about 30 speakers, but this time with two recordings per speaker. This would be needed in

addition to the 30 single-recordings-per-speaker database. The Coordinator examines whether it is realistic to provide these data and concedes that this is not possible. Instead he arrives at the following decision: The PT contains as test set the one specified in Example 1 (50 Turkish-language recordings from 25 speakers – 2 recordings per speaker – as test material as well as single Turkish-language recordings from 30 speakers as additional database). For participants who can provide calibrated LRs based on this material, these are collected and EPP as well as the Cllr can be calculated by the Coordinator as part of the performance assessment process of the PT. For participants who cannot provide calibrated LRs based on this material, the Coordinator offers calibration of the submitted values using cross validation and calculates Cllr subsequently to this calibration stage.

Due to the differences that are to be expected by using different methods to arrive at calibrated LRs it is decided that the test will be offered as a CE. Based on the results and the different methods used it is planned that a variation of this test can be prepared for a PT in the future.

13.3 Example 3: PT on FSASR

The Coordinator makes an inquiry about the FSASR methods that are used among the prospective PT participants. From this inquiry it emerges that the FSASR method used most commonly is the measurement of formant frequencies in various vowel categories. Some respondents use one set of formant measurements (F1 to F3) at the centre of the vowel, others process all formant information from the left boundary of a vowel token to the right boundary. The former method is called the point method and the latter the interval method. Asked for their method used to calculate likelihood ratios some respondents mention that they use the MVKD (multivariate kernel density) approach, others say that they use the GMM-UBM approach. Another question of the Coordinator concerns the language analysed with the formant analysis method. The response from the prospective PT participants is that they only analyse their national language with this method. Finally, the Coordinator asks for the kind of performance metrics used with the formant analysis method and the typical values of these metrics. It turns out that all respondents use EPP among their performance metrics. EPP values found with the vowel phoneme /a/ are between 0.15 and 0.35 in natural conversation telephone speech. For most other vowel phonemes EPP is similar though generally higher and fusion between different vowel phonemes can improve EPP.

Based on this information the following PT is provided. Recordings involving 25 different speakers with two non-contemporaneous speech samples each are provided speaking in forensically realistic telephone conversations. The language spoken in the recordings is Dutch. In order to enable all non-Dutch-speaking participants to carry out formant analysis, the audio files are provided together with annotation files (also called label files) in which all tokens of the vowel phonemes /a/ and /i/ are marked from beginning to end. Based on this information the participant is able to identify the formants and measure their values. In addition to the test files, audio files from additional 25 speakers are provided, again speaking in two different sessions and with the same speech style as before and again provided with annotation files. The validation metric is EPP and the validation criterion is set at a value of 0.3, which is meant as a conservative value in order to accommodate many different methods.

When the results of the PT submissions are analysed it occurs that some participants have used the MVKD approach and others have used the GMM/UBM approach. All of those applying MVKD have used the point method and most of those applying GMM/UBM have used the interval method. Different use has also been made of the development database; users of MVKD used speaker pairs whereas most users of GMM-UBM only needed one file per speaker. Furthermore, there were differences in if and how the results for the two different vowels were combined. These methodological differences can be addressed in further CEs or PTs.

13.4 Example 4: CE on FSASR

A CE is conducted that focusses on different issues involving long-term formant analysis (LTF). The Coordinator makes an inquiry among prospective participants about whether they have used LTF before and if yes, which language they have worked on, which formant-tracking software and settings they have used, if and how they have calculated uncalibrated or calibrated LRs and which performance in terms of EPP and Cllr they have obtained. The response is diverse, ranging from participants who have not used LTF before to those who are specific about all aspects of LTF analysis and can provide EPP and Cllr values.

The Coordinator compiles a test data set with two recordings each of 22 male speakers of German from natural telephone conversations. The Coordinator furthermore provides a database of 35 natural conversation telephone speech recordings, each containing a different male speaker of German. Based on the responses from the participants the Coordinator in

association with the Advisory Group decides to design a CE that is maximally flexible regarding the methods and evaluation schemes:

- For laboratories that want to participate but have not used LTF before, instructions are provided on how LTF can be performed. This is possible using user-friendly open source software. They are free to analyse and submit as many samples as they want and they receive feedback from the Coordinator about their LTF results and how they compare to other results. This exercise does not necessarily involve LR calculations yet but can entirely focus on the feature extraction level of LTF.

- For participants who have used LTF before, but not in a FSASR context (i.e., without calculating LRs), they are requested to analyse all the 44 test samples and provide the examination results in the form of (corrected) formant track raw data. The Coordinator offers as a service to calculate LRs based on the examination results. Also the data of 35 speakers from natural conversation telephone recordings may be examined by the participants or, if the workload is too high, the Coordinator can use a pre-processed version of this database. Based on these LRs the Coordinator calculates EPP and Cllr (with prior cross-validation calibration if Cllr is reported). Each participant can compare these with the examination results from other participants and receives feedback about what results they would obtain if they processed their LTF analysis results in terms of FSASR.

- For participants who have used LTF before and did so within a FSASR context, they are requested to submit their LR values. Based on this delivery the Coordinator calculates EPP and Cllr. As a variation it is possible to address the question of language-dependence. If the participant works on another language than German they can process the CE German test data by using a database other than the German database provided with the CE. Ideally they would use the German database as well. It would then be possible to gain insight into which influence the use of different-language corpora have on LTF within a FSASR context.