FORENSIC AUTOMATIC SPEAKER RECOGNITION USING BAYESIAN INTERPRETATION AND STATISTICAL COMPENSATION FOR MISMATCHED CONDITIONS

THÈSE N^{O} 3367 (2005)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Institut de traitement des signaux

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Anil ALEXANDER

Bachelor of Technology in Computer Science and Engineering, Indian Institute of Technology, Madras, Inde
et de nationalité indienne

acceptée sur proposition du jury:

Dr A. Drygajlo, directeur de thèse
Prof. P. Margot, rapporteur
Dr P. Rose, rapporteur
Prof. J.-P. Thiran, rapporteur

Lausanne, EPFL
2005

to my parents

Acknowledgements

My heartfelt thanks to my thesis advisor Dr. Andrzej Drygajlo who gave me the opportunity to work on the fascinating topic of forensic speaker recognition. His kindness, dedication, hardwork and attention to detail have been a great inspiration to me. I would particularly like to thank him for all his help in patiently and carefully correcting this thesis manuscript. I am also grateful to the members of my jury (Prof. Pierre Margot, Dr. Philip Rose and Prof. Jean-Phillippe Thiran) for reviewing my work and for their valuable comments and to Prof. Daniel Mlynek for acting as the president of the jury.

My most special thanks to my parents and my brother. Mummy and Daddy, I thank you for believing in me and helping me believe in myself; for giving me the dream and keeping the dream alive against all odds. Aju, you are my not-so-little brother, sounding board, counselor and best friend all rolled into one. Thank you for always being there.

My deepest gratitude goes to my lovely fiancée. Thank you Faye, for coming into my life, for all the love and support you have given me and for seeing this thesis through with me. This thesis would not have taken the shape it has now, without your help.

I would also like to say a big 'thank you' to all my friends and colleagues for their friendship and support that has seen me through the highs and lows of these last four years. In particular, I thank:

- Guillermo for being the best friend I have had these four years and for our lunch discussions on life, love and what the thesis should not be.
- My officemates Plamen, Kris and Jonas who have made my years at the laboratory an enjoyable and unforgettable experience: Plamen for his unique sense of humour and his advice and help; Kris and Jonas for their stimulating discussions and their sometimes radically different viewpoints on life.
- The Institut de Police Scientifique (and in particular Prof. Pierre Margot and Prof. Christophe Champod), at the University of Lausanne for kindly allowing me to work with them and to discover the world of forensic science.

- The Swiss National Science Foundation for providing funding for the research done in this thesis.

- Didier Meuwly, whose pioneering work in applying Bayesian interpretation to forensic automatic speaker recognition provided the basis for my research.

- Filippo Botti, for being an excellent friend and collaborator, and for all the enlightening and productive scientific discussions and arguments we have had on forensic speaker recognition and other subjects over the years.

- Damien Dessimoz, for being one of the kindest and nicest people I have worked with, hardworking, always ready to help and an inexhaustible source of references. Special thanks to him for translating the abstract of this thesis into French.

- Anne Schwan, for her warmth, motherly love and affection, for the innumerable lunches and dinners, wonderful hikes and trips, for the company I look forward to every Thursday, and making her place a home away from home, especially during the time I was writing my thesis.

- Reva Schwartz from the United States Secret Service for her helpful inputs and her enthusiastic support in my academic and professional endeavors.

- My close friend, neighbor and photography mentor Arunan for all his support, help, swimming sessions with Santanu, and his encouragement over these years.

- Two Mathews who really helped me a lot, especially when I just embarked on this adventure are Jacob and Magimai-Doss, with their advice and readiness to help in the various parts of this thesis process.

Above all, I thank God for guiding and taking care of me every step of the way.

Abstract

Nowadays, state-of-the-art automatic speaker recognition systems show very good performance in discriminating between voices of speakers under controlled recording conditions. However, the conditions in which recordings are made in investigative activities (e.g., anonymous calls and wire-tapping) cannot be controlled and pose a challenge to automatic speaker recognition. Differences in the phone handset, in the transmission channel and in the recording devices can introduce variability over and above that of the voices in the recordings. The strength of evidence, estimated using statistical models of within-source variability and between-sources variability, is expressed as a likelihood ratio, i.e., the probability of observing the features of the questioned recording in the statistical model of the suspected speaker's voice, given the two competing hypotheses: the suspected speaker is the source of the questioned recording and the speaker at the origin of the questioned recording is not the suspected speaker. The main unresolved problem in forensic automatic speaker recognition today is that of handling mismatch in recording conditions. Mismatch in recording conditions has to be considered in the estimation of the likelihood ratio.

The research in this thesis mainly addresses the problem of the erroneous estimation of the strength of evidence due to the mismatch in technical conditions of encoding, transmission and recording of the databases used in a Bayesian interpretation framework.

We investigate three main directions in applying the Bayesian interpretation framework to forensic automatic speaker recognition casework. The first addresses the problem of mismatched recording conditions of the databases used in the analysis. The second concerns introducing the Bayesian interpretation methodology to aural-perceptual speaker recognition as well as comparing aural-perceptual tests performed by laypersons with an automatic speaker recognition system, in matched and mismatched recording conditions. The third addresses the problem of variability in estimating the likelihood ratio, and several new solutions to cope with this variability are proposed.

Firstly, we propose a new approach to estimate and statistically compensate for the effects of mismatched recording conditions using databases, in order to estimate

parameters for scaling distributions to compensate for mismatch, called 'scaling databases'. These scaling databases reduce the need for recording large databases for potential populations in each recording condition, which is both expensive and time consuming. The compensation method is based on the principal Gaussian component in the distributions. The error in the likelihood ratios obtained after compensation increases with the deviation of the score distributions from the Gaussian distribution. We propose guidelines for the creation of a database that can be used in order to estimate and compensate for mismatch, and create a prototype of this database to validate the methodology for compensation.

Secondly, we analyze the effect of mismatched recording conditions on the strength of evidence, using both aural-perceptual and automatic speaker recognition methods. We have introduced the Bayesian interpretation methodology to aural-perceptual speaker recognition from which likelihood ratios can be estimated. It was experimentally observed that in matched recording conditions of suspect and questioned recordings, the automatic systems showed better performance than the aural recognition systems. In mismatched conditions, however, the baseline automatic systems showed a comparable or slightly degraded performance as compared to the aural recognition systems. Adapting the baseline automatic system to mismatch showed comparable or better performance than aural recognition in the same conditions.

Thirdly, in the application of Bayesian interpretation to real forensic case analysis, we propose several new solutions for the analysis of the variability of the strength of evidence using bootstrapping techniques, statistical significance testing and confidence intervals, and multivariate extensions of the likelihood ratio for handling cases where the suspect data is limited.

In order for forensic automatic speaker recognition to be acceptable for presentation in the courts, the methodologies and techniques have to be researched, tested and evaluated for error, as well as be generally accepted in the scientific community. The methodology presented in this thesis is viewed in the light of the Daubert (USA, 1993) ruling for the admissibility of scientific evidence.

Version abrégée

A l'heure actuelle, les systèmes de reconnaissance automatique de locuteur possèdent d'excellentes performances lorsqu'il s'agit de discriminer entre des voix de locuteurs acquises dans des conditions contrôlées. Pourtant, les conditions dans lesquelles la police effectue ses enregistrements (dans des cas d'appels anonymes ou d'écoutes téléphoniques), ne peuvent être contrôlées et sont donc un défi difficile pour la reconnaissance automatique de locuteurs. Des différences dans les combinés téléphoniques, dans le canal de transmission et dans l'appareil d'enregistrement peuvent introduire des variations plus importantes que celles des voix sur les enregistrements. La force probante de la preuve, estimée à l'aide de modèles statistiques des intra- et inter-variabilités de la source, est exprimée sous la forme d'un rapport de vraisemblance, i.e., est-ce plus probable que l'enregistrement de question (trace) a été produit par un locuteur suspecté ou par toute autre personne appartenant à la population pertinente. Le principal problème non résolu dans la reconnaissance automatique de locuteurs en sciences forensiques est la manière de traiter les conditions d'enregistrement différentes. Les conditions d'enregistrement différentes doivent être considérées dans l'estimation du rapport de vraisemblance.

Ce travail de thèse traite essentiellement du problème lié à l'estimation erronée de la force probante de la preuve due aux conditions différentes d'encodage, de transmission et d'enregistrement des bases de données utilisées dans un canevas d'interprétation bayesienne.

Nous avons approfondi trois directions principales en appliquant le canevas d'interprétation bayesienne à des cas forensiques de reconnaissance automatique de locuteurs. La première direction traite du problème des conditions différentes d'enregistrement des bases de données utilisées lors des analyses. La deuxième direction concerne l'introduction de la méthodologie d'interprétation bayesienne à la reconnaissance auditive de locuteurs et la comparaison entre les capacités de reconnaissance vocale de personnes sans connaissance en phonétique et d'un système automatique, dans des conditions semblables et différentes. La troisième direction traite du problème de l'incertitude dans l'estimation du rapport de vraisemblance, et plusieurs approches novatrices pour faire face à cette incertitude sont proposées.

Premièrement, nous avons proposé une approche novatrice pour estimer et compenser statistiquement les effets des conditions différentes d'enregistrement des bases de données, afin d'estimer les paramètres pour balancer les distributions, dans le but de compenser pour les conditions différentes appelées 'scaling databases'. Ces scaling databases réduisent le besoin d'enregistrer de grandes bases de données représentant les populations potentielles dans toutes les conditions d'enregistrement, ce qui demande beaucoup de temps et d'argent. La méthode de compensation est basée sur la gaussienne principale dans les distributions. L'erreur dans le rapport de vraisemblance obtenu après compensation augmente avec la déviation des distributions des scores de la distribution gaussienne. Nous proposons des directives pour la création d'une base de données pouvant être utilisée afin d'estimer et de compenser les conditions différentes, et avons créé un prototype de cette base de données dans le but de valider notre méthodologie de compensation.

Deuxièmement, nous avons analysé les effets des conditions différentes d'enregistrement sur la force probante de la preuve, pour la reconnaissance vocale de personnes sans connaissance en phonétique et d'un système automatique. Nous avons introduit la méthodologie d'interprétation bayesienne à la reconnaissance auditive par des profanes et nous avons estimé les rapports de vraisemblance correspondants. Nous avons pu observer expérimentalement que dans des conditions semblables pour les enregistrements de question et de référence, le système automatique obtient de meilleures performances par rapport à l'approche auditive. Cependant, dans des conditions différentes, le système automatique de référence a obtenu des performances comparables ou légèrement moins bonnes par rapport à l'approche auditive. En adaptant le système de reconnaissance de référence à des conditions différentes, le système obtient des performances comparables, voire meilleures, par rapport à l'approche auditive.

Troisièmement, dans l'application de l'interprétation bayesienne à des analyses de cas forensiques réels, nous proposons plusieurs approches novatrices pour l'analyse de la variabilité de la force probante de la preuve, en utilisant des techniques de 'bootstrapping', des tests statistiques de signification et des intervalles de confiance, ainsi que des extensions multivariées du rapport de vraisemblance pour traiter des cas dans lesquels les données du suspects sont limitées.

Afin que la valeur probante de la preuve, obtenue dans l'analyse de cas de reconnaissance automatique de locuteurs en sciences forensiques, puisse être présentée à la cour, les méthodologies et les techniques utilisées doivent pouvoir être testées, le taux d'erreurs engendré par la méthode doit être connu, et la méthode doit faire l'objet d'une acceptation générale de la communauté scientifique. La méthodologie présentée dans cette thèse est évaluée à la lumière des critères de l'affaire Daubert (USA, 1993), qui fixe les normes et standards au niveau de l'admissibilité des preuves scientifiques dans un tribunal.

Contents

List of Figures

List of Tables

1 Introduction

Genesis Chapter 27, Verses 23-25

Jacob went close to his father Isaac, who touched him and said, "The voice is the voice of Jacob, but the hands are the hands of Esau." He did not recognize him, for his hands were hairy like those of his brother Esau; so he blessed him."Are you really my son Esau?" he asked.

"I am," he replied.

Then he said, "My son, bring me some of your game to eat, so that I may give you my blessing."

1.1 Recognizing individuals by their voices

Human speech is a complex signal, an outcome of the influence of several physiological, psychological and environmental factors. The acoustic signal we produce when we speak is determined by, among others, the physiology of the vocal tract and articulators, childhood language and dialect acquisition, regional traits, and training from life experiences. These features give a distinctive "identity" to the speech of different individuals.

In order to tell two voices apart, it is necessary to consider characteristic features based on which the voices can be discriminated. When comparing recordings of human speech, it is necessary to actually compare the two voices, minimizing the influence of the external factors that have affected the recordings.

Modern communication takes the form of electronic mail and messaging, voice and video. Speech still remains one of the principal means of communication, and ad-

vances in telecommunication and recording technology have meant that an increasing amount of speech is transmitted through telecommunication networks. These transmission networks as well as the recording devices have their influence on the recording of the voice. The recorded speech is a manifestation of the voice of an individual on a physical medium. The difficulty of telling voices apart from such recorded speech is then compounded by the effects of transmission and recording.

1.1.1 The importance of determining the identity of a speaker in forensic applications

Forensic speaker recognition can be deemed necessary in several situations involving the courts, the police and investigative agencies, as well as private organizations. For instance, in courts, it is often necessary to find the source of a voice in the recording in question or to settle a challenge about the alleged source of a questioned recording. Earwitness identification by naive listeners may be erroneous, and expert evaluation of the evidence can go against the testimony of the listener [Rose, 2002, :2]. The police may use wire-tapping or body-wire techniques to collect information about suspected persons, and it is then necessary to identify the speakers of the sections of interest in these recordings. It is also often vital to identify and profile the speaker in cases of threats and warnings received by the police and investigative agencies. In counter-terrorism investigations, often, threats and propaganda recordings take the form of audio and video recordings (where the speaker cannot be seen), and it is necessary to identify or verify the claimed identity of the speaker. Even private organizations and companies may require analysis for purposes of internal inquiry in cases of harassment allegations [Koolwaaij and Boves, 1999], corruption, etc.

The voice differs from other biometric characteristics of the human being, like fingerprints or DNA, as it changes over time (the short term and the long term), depends on the health and emotional state of the speaker and can be altered, at will, by the speaker (disguise or impersonation) [Bonastre et al., 2003]. In addition, the voice differs from DNA or fingerprints in its functionality and in the extent of its variability.

1.2 Admissibility of scientific evidence in courts

In the United States of America, requirements for the admissibility of scientific evidence in federal courts have been laid down in the Daubert criteria, based on the Supreme Court pronouncement, in 1993, in Daubert vs Merell Dow Pharmaceuti-

cals*. The previously accepted standard for scientific expert opinions is the Frye† standard, where it is stated that expert opinion, based on a scientific technique, is inadmissible unless the technique is 'generally accepted' as reliable in the relevant scientific community. It was also stated that expert opinion based on a methodology that diverges significantly from the procedures accepted by recognized authorities in the field cannot be shown to be generally accepted as reliable. This 'general acceptance test' was superseded, after 70 years of its application, by the Federal Rules of Evidence‡. The difference between the Frye and Daubert standards is that according to Frye, admissible evidence may be based on a method generally accepted in the scientific community, but according to the Daubert ruling, the inquiry is not limited to this but also depends on a demonstration, in an 'objectively verifiable way', that the evidence is based on a reliable scientific method [Loevinger, 1995]. The judge has a screening role between the testifying expert witnesses and the jury, to ensure the reliability and relevance of the expert testimony.

The criteria set out in the Daubert ruling, in order to determine whether a testimony is based on scientific theories, reasoning or methodology that is reliable and valid, are summarized as [Loevinger, 1995] follows:

- Whether the theory can be tested and has been tested.
- Whether the technique has been published or subjected to peer review.
- Whether the technique has a known or potential rate of error in application.
- Whether standards exist and are maintained to control the operation of the technique.
- Whether the technique is generally accepted within the relevant scientific community.
- Whether the technique is based on facts or data of a type reasonably relied on by experts in the field.
- Whether the technique has a probative value that is not outweighed by the dangers of unfair prejudice, confusion of issues or misleading the jury.

For forensic speaker recognition to be admissible in the federal courts (in the United States), it would need to satisfy the requirements laid down in the Daubert ruling. It is worthwhile considering how the methodology and techniques used for

*in Daubert v Merrell Dow Pharmaceuticals 43 F 3d 1311; 125 L Ed (2d) 469; 509 US 579; 113 S Ct 2786 (1993)

†Frye v United States (1923) 293 Federal Reports (1st series) 1013,1014 (CA)

‡US Federal Rules of Evidence, U.S. Government Printing Office, Washington, 2004

forensic speaker recognition satisfy the Daubert criteria for their admissibility as scientific evidence. A discussion of how the methodologies presented in this thesis can be compatible with the Daubert criteria is presented in Chap. 6.

1.3 Speech as evidence

Audio and speech analysis plays an important role in forensic investigation and has led to the development of a range of analysis techniques including audio intelligibility enhancement, authentication of audiotape recordings, transcription, analysis of disputed recordings and speaker recognition.

Intelligibility enhancement

Audio enhancement techniques such as noise reduction and band-pass filtering, and amplification of the audio signal are used in order to improve the intelligibility of audio recordings. Noise reduction can be performed in order to suppress background and general broadband noise, although this may result in a degradation of original signal quality. It can be helpful in reducing tape hiss, microphone background noise or any noise that is constant throughout the duration of the recording. A band-pass filter can be applied in order to limit the input waveform within the frequency band where bulk of the information related to the speech is present (e.g., 300Hz - 3400 Hz). Simply increasing the playback volume (i.e. applying gain) is sometimes useful in making the recordings more audible, without the risk of saturation that would occur if the input to the recorder had simply been set to a high volume.

Forensic transcription of speech

Transcripts of audio recordings often appear as evidence in cases. Recordings that require legal transcription are often made as part of a surveillance operation; e.g., wiretapping performed by the police or undercover recordings made by hidden recording devices of poor quality. The content of these recordings is sometimes subject to dispute as the interpretation of an entire statement can hinge on a few key phrases or words which could have been misheard. Phonetic transcription requires the transcriber to carefully listen to the recording and transfer it into a symbolic representation from which this speech can be reproduced to the satisfaction of the original speaker, the transcriber, a native speaker of the tongue, or another speaker. This replicability is important in forensic issues as the transcription then becomes verifiable and can be checked by another transcriber [Rose, 2002, :37]. Phonetic transcription for forensic purposes requires awareness of linguistic, phonetic, acoustic, and psycholinguistic issues [Fraser, 2003].

Each transcription requires several rounds of careful listening to the same recording, and lengthy recordings can take a considerable amount of time to process. In addition, the conditions of recording may call for audio enhancement.

Audiotape authenticity verification

Establishing whether the audio recordings are indeed authentic and have not been tampered with, is often of importance to the analysis (speaker recognition, transcription, etc.) that follows. Authenticity verification is primarily concerned with audiotape recordings and has been of interest for several years now. Hollien [1990] proposes a set of requirements that should be fulfilled in order to consider an audiotape recording as authentic:

- It must have captured all of the audio events that occurred during the entire target period, and this period of interest must include the entire series of happenings relevant to the intelligence contained in the recording.
- It must be shown that the tape recording has not been interrupted in any manner.
- None of the sections of the recording should have been removed.
- Nothing should have been added to the recording.

If any of these conditions are not satisfied, Hollien considers that the recording is not authentic.

Critical listening and waveform examination can reveal whether the questioned tape contains any audio event 'signatures' or whether there are portions of the recording that are markedly different from the rest. These audio events include the mark made by the recording head on the audiotape at the start of the recording, when the recording is paused and when the recording is stopped. Along with the presence of acoustic anomalies, generally, the presence and character of such event signatures is one of the most important features relied on in forming an opinion. Spectrographic analysis can also be used in order to determine whether any segments from a recording have been recorded over or modified. The uniqueness of the patterns produced by the spectral distribution and intensity of the acoustic events permits the copied segments to be detected either visually or statistically [Gruber et al., 1993]. In addition to this analysis, microscope- and crystal-based techniques are used to identify the physical marks of the audio events on the magnetic tape [Koenig, 1990].

Although audiotape authentication has been an important discipline in forensic audio analysis, it has been losing its relevance in the recent years, with a multitude of new audio recording instruments, media and formats. Digital recorders have become

widely available and are inexpensive. In addition, audio-editing software makes the manipulation of audio recordings relatively easy. Although it is still possible to detect whether audio recordings have been tampered with, forgeries done by experienced or competent manipulators may be impossible to detect. Also, if the questioned recordings are copies of the original recording, then it is indeed difficult to establish its integrity (although, on its own, this does not imply that it has been tampered with).

1.4 Forensic speaker recognition: aural-perceptual, auditory-instrumental and automatic approaches

The approaches commonly used for speaker recognition include the aural-perceptual, the auditory instrumental, and the automatic approaches.

Aural-perceptual (also known as auditory analysis) methods basically rely on the careful listening of recordings by trained phoneticians, where the perceived differences in the speech samples are used to estimate the extent of similarity between voices. Early approaches to aural perceptual analysis included the careful analysis of dialectal and sociolectal features, speech defects and voice quality, but the adequacy of these measures was called into question and came under strong criticism [Künzel, 1998]. While these characteristics may prove inadequate for speaker recognition, they are useful in *speaker profiling*. In addition to these differences, speakers differ in their rate of speech, with their intonation pauses, their articulation and their diction. Also, higher level characteristics, such as idiomatic and linguistic characteristics, as well as the prosody of their speech, are indicative of the speaker's identity. With this approach, a subjective probability of the similarity of the two voices can be established. The aural-perceptual approach has its limitations, and in traditional phonetic analysis, it is used mainly to extract features of interest, which are then analyzed using the auditory-instrumental approach.

The auditory-instrumental approach involves the acoustic measurements of various parameters such as the average fundamental frequency (F_0), articulation rate, the pitch contour, the spectral energy, etc. The means and variances of these parameters are compared. The use of spectrograms, for speaker recognition can be considered as another method in this approach [Bolt et al., 1970, 1973]. The use of spectrograms in what has been called the 'voiceprint' approach, has come under considerable criticism in the recent years.

Forensic automatic speaker recognition is an established term used when automatic speaker recognition methods are adapted to forensic applications. In automat-

ic speaker recognition, the statistical models of acoustic parameters of the speaker's voice and the acoustic parameters of questioned recordings are compared. The quantified degree of similarity between speaker-dependent features extracted from the questioned recording (or trace) and speaker-dependent features extracted from the recorded speech of a suspect, represented by his/her model, is calculated in order to evaluate the evidence [Drygajlo et al., 2003]. In forensic automatic speaker recognition systems, the strength of such evidence is calculated as the relative probability of observing the features of the questioned recording in the statistical model of the suspected speaker's voice and in the statistical models of voices in a potential population. This corresponds to a speaker recognition system which performs a comparison of voices and determines how more likely it is that the questioned voice is from one source as compared to other sources. It can be remarked that speaker recognition is a general term used to include all of the many different tasks of discriminating people based on the sound of their voices.

The results from the aural-perceptual approach, where the interpretation is based on subjective probabilities, from the auditory-instrumental approach, where interpretation is based on subjective and statistical probabilities, and from the automatic approach, where interpretation is based on statistical probabilities, can be considered together to give a combined interpretation of the strength of the evidence [Majewski and Basztura, 1996; Meuwly, 2000].

1.5 Automatic speaker recognition and the task of forensic speaker recognition

Automatic speaker recognition is an attractive option for forensic speaker recognition tasks because forensic cases often include large amounts of audio data which are difficult to evaluate within the time constraints of an investigation or analysis required by the courts. Automatic speaker recognition has been shown to perform with a high accuracy in controlled conditions. Automatic recognition can complement the traditional aural-perceptual and semi-automatic speaker recognition techniques used in forensic speaker recognition. Aural-perceptual as well as semi-automatic (auditory-instrumental) speaker recognition require a high degree of mastery of a language and its nuances, and experience in extracting and comparing relevant characteristics. As modern criminal activity often spans several countries, there may be cases in which there is a need to analyze speech in languages where sufficient expertise is unavailable.

In recent years, there has been increasing interest in using automatic speaker recognition techniques for forensic tasks and several research groups around the world have been working on this problem [Meuwly et al., 1998; Koolwaaij and Boves, 1999;

Meuwly, 2001; Pfister and Beutler, 2003; Drygajlo et al., 2003; Gonzalez-Rodriguez et al., 2004; Botti et al., 2004a; Beck et al., 2004; Alexander et al., 2004; Niemi-Laitinen et al., 2005; Campbell et al., 2005]. Expressing conclusions in a forensic context requires that the methods used are understandable and that the results are interpretable by the courts. Conventional speaker verification techniques are not directly applicable in forensic analysis, and it is necessary to adapt them to requirements of the courts.

1.6 Bayesian methodology in estimating the strength of evidence

One of the reasons for the difficulties in using speech as an identifying characteristic is the variability of the speech not only for a single speaker (within-source variability) but also between different speakers (between-sources variability). Normally, characteristic features of speakers' voices show less variation for the same speaker than the variation of the same features within a population of speakers.

In the Bayesian approach, the interpretation of evidence must take place within a framework of circumstances, and to interpret the evidence, it is necessary to consider at least two propositions [Evett, 1998]. Instead of addressing the likelihood of just one hypothesis (e.g., the questioned sample comes from the suspected person), the expert should consider the likelihood of the evidence given at least one competing hypothesis (e.g., the questioned sample comes from some other person). The ratio of likelihoods can be understood as ratio of the similarity and typicality of the questioned recording with that of the suspected speaker and the potential population of speakers, i.e., the similarity between the questioned recording and the voice of the suspected speaker and the typicality of the questioned recording characteristics appearing in a relevant population.

1.6.1 Mismatched recording conditions and their effect on the strength of evidence

In forensic speaker recognition casework, the recordings analyzed often differ because of telephone channel distortions, ambient noise in the recording environments, the recording devices, as well as their linguistic content and duration. These factors may influence aural, instrumental and automatic speaker recognition. In many cases, the forensic expert does not have a choice in defining the recording conditions for the suspect and questioned recordings, as these recordings are provided by the police or the court, and additional recordings cannot be made. The Bayesian approach, based

on interpretation of evidence, relies heavily on the use of databases in order to evaluate the strength of the evidence. If there is a mismatch in the technical (encoding and transmission) and acoustic conditions between the recordings of the databases used, comparisons between them can lead to erroneous or misleading results, and therefore, it is of utmost necessity to reduce and quantify the effect of the mismatch.

The problem of mismatched recording conditions in a data-driven Bayesian interpretation framework is particular to automatic speaker recognition applied to forensic casework. In this thesis, we focus on forensic automatic speaker recognition and the effect of mismatched recording conditions of the databases used on the strength of evidence.

1.7 Objectives of the thesis

The main goal of this thesis is to measure and compensate for the effects of mismatch that arise in forensic case conditions due to the technical (encoding and transmission) and acoustic conditions of the recordings of the databases used, and to quantify the uncertainty that it introduces.

1.8 Major contributions of the thesis

The major contributions of this thesis, listed in the order of their importance, are:

- Proposal of a methodology to estimate and statistically compensate for the differences in recording conditions of the databases used in forensic speaker recognition within a Bayesian interpretation framework, at the level of score distributions.

- Proposition of guidelines for the creation of a forensic speaker recognition database that can be used in order to perform forensic casework in mismatched conditions, as well as the creation of two prototypes of this database (the IPSC-02 and the IPSC-03 databases).

- Study and analysis of mismatched technical conditions in training and testing phases of speaker recognition, and their effect on human-aural and automatic forensic speaker recognition.

- Proposition of a multivariate estimation of the strength of evidence in forensic automatic speaker recognition, called the 'direct method', which can be used instead of the univariate estimation, called the 'scoring method', when the suspect data available is limited.

- Analysis of the variability of the strength of evidence in forensic speaker recognition based on a bootstrapping technique and statistical significance analysis.
- Evaluation of the use of auxiliary features such as conditional dependencies of prosodic and spectral envelope features, using a Bayesian network approach, to deal with mismatched conditions in forensic applications.

1.9 Organization of the thesis

This thesis is organized as follows:

- *Chapter 1*: Introduction and presentation of the objectives and contributions of the thesis.
- *Chapter 2*: Discussion of the state-of-the-art forensic speaker recognition techniques (including feature extraction and modeling techniques), as well as forensic interpretation of speech evidence and issues concerning real forensic case conditions.
- *Chapter 3*: Presentation of the adaptation of Bayesian interpretation to real forensic case conditions.
- *Chapter 4*: Analysis of the influence of mismatched recording conditions on aural and automatic estimates of the strength of evidence.
- *Chapter 5*: Estimation of the influence of mismatched recording conditions and the effect of statistical compensation for mismatched conditions on the strength of evidence.
- *Chapter 6*: Evaluation of the mismatch compensation techniques presented in Chapter 5.
- *Chapter 7*: Discussion and review of the main ideas presented in the thesis.
- *Chapter 8*: The summary and conclusion of the thesis, with a discussion of possible extensions of the present work.
- *Appendix*: Case report of the Netherlands Forensic Institute (NFI) speaker recognition evaluation through a fake case, as well as descriptions of the three forensic speaker recognition databases created for the validation of the methods.

Forensic Automatic Speaker Recognition

2

As discussed in Chap. 1, the human voice is closely related to the identity of its owner and communicates information not only about the intended message but also about the emotional and physiological state of the speaker, as well about his/her social, educational and regional background. The human voice is rich in information about the speaker including but not limited to his identity, which is useful in forensic analysis to gain insight into the individual characteristics of the person who is speaking. Automatic speaker recognition methods are increasingly used for the identification of speakers and have been shown to be highly accurate in controlled recording conditions. Although automatic speaker recognition techniques have shown promise in being used in the forensic analysis of voice, they require adaptation to the often uncontrolled and adverse forensic conditions, and interpretation of results in a way that would be understood and acceptable to the courts.

In this chapter, we discuss forensic automatic speaker recognition, dealing with the requirements of forensic analysis, a Bayesian framework for interpretation of evidence, existing methods for automatic speaker recognition and their adaptation to the requirements of the Bayesian framework. It consists of three parts, the first one dealing with Bayesian and frequentist approaches to forensic recognition as well as a corpus-based methodology for the interpretation of voice evidence, the second dealing with state-of-the-art algorithms and methods used for automatic speaker recognition and statistical modeling and how they are adapted to an operational forensic automatic speaker recognition system, and the third dealing with automatic speaker recognition approaches to mismatched recording conditions.

2.1 Bayesian and frequentist approaches to forensic recognition

Two common approaches to the interpretation of evidence have been the *Bayesian* and *frequentist* with many experts choosing to express conclusions based on either strategy. The traditional or orthodox approach of frequentist evaluation has been challenged by the proponents of the more recent Bayesian approach. In this section, we present a discussion about these two approaches and their relationship with the analysis of speaker recognition evidence.

Robertson and Vignaux [1995] state that the main difference between Bayesian and frequentist approaches is that the Bayesian one discusses the probability of hypotheses, while the frequentist approach discusses only the probability of data. They cite an example of the Bayesian proponent being willing to discuss the probability that it will rain tomorrow, taking into account other information, including meteorological data, the color of the sky, etc., while the frequentist may only consider what proportion of days it rained in the years before in this particular period.

In the frequentist approach, a statistical inference, based on the testing of a single hypothesis, is used. This hypothesis is that the sample belongs to certain data-set and it is compared to a 'null hypothesis' that suggests that the data occurred by chance. In this approach, the probability of a certain event happening by chance is estimated. If indeed this probability is very small (and below a certain pre-decided threshold), it is considered that the null hypothesis can be rejected and that the result supports the hypothesis that is being tested.

As presented in [Robertson and Vignaux, 1995, :114], the orthodox approach to comparing evidence in forensic science follows three steps:

1. *Considering if the two samples can match:*
   Characteristic features of the two samples are compared, and the difference or similarity between the samples is evaluated. This measurement is limited by the accuracy of the measuring instrument. There is, thus, an uncertainty in the measurement process as well as the inherent differences between the samples that do indeed come from the same source. If the measurements of a feature considered, from two samples are different, it is estimated how probable it would be to obtain this difference, had the two samples come from the same source. The null hypothesis is that the two samples could have come from the same source. If the difference is not significant (at some arbitrary level of significance), it is decided that a match has been obtained.

2. *If a match is obtained, considering the probability that this match could have occurred by chance:*

In this step, the odds of obtaining the 'match' result by chance are evaluated, i.e., what is the probability of obtaining a result that the two samples (voices, DNA, glasses, etc.) have the same source if any two samples had been chosen at random from the population. For this step, it is necessary to have databases of samples, literature or other information about the characteristic or feature of interest, in order to estimate the frequencies of observing this characteristic in samples from a population.

3. *Formulation and expression of a conclusion as to whether the two samples have the same source:*
   Since the match has been decided in Step (1), it is compared only with the odds that the match has occurred by chance. Thus, the probability of observing the match is assigned to 1. Often, an arbitrary threshold is chosen, below which the expert states that it is not possible to conclude that they have the same source, and above which an assertion is made that the two samples have the same source, with statements that vary in their strength such as 'it is possible', 'it is very possible', etc.

This approach has come under criticism [Robertson and Vignaux, 1995; Aitken and Taroni, 2004; Champod and Meuwly, 2000] from the proponents of the Bayesian approach. The problem with this approach is that the expert has to assume the priors (which is not based on the conditions of the case, but often due to some convention). A so-called 'uninformative prior' is often used, which assumes that each of the possible explanations is equally possible. Also, the alternative hypothesis, i.e., the odds against a match by chance, may not be the appropriate alternative hypothesis in a case. Other difficulties with this approach include the 'fall-off-the-cliff effect' (i.e., if the rejection of the hypothesis is at some arbitrary threshold, values falling just short of this threshold would be rejected while others just at the threshold would be accepted), the match criteria (what exactly would constitute a match), the rejection of evidence even though it favors a hypothesis (even at a low significance level, it may still be possible that a given value may favor one hypothesis more than the other), overvaluation of evidence (just because a sample satisfies a significance test, the probability of obtaining this value, given the hypothesis is treated as one), and the transposition of the conditional (a trap of thinking, where the fact that a certain evidence value is more probable, given a certain hypothesis, leads to thinking that this hypothesis itself must by highly probable).

In the Bayesian approach, the interpretation of evidence must take place within a framework of circumstances and to interpret the evidence, it is necessary to consider at least two propositions [Evett, 1998]. Instead of addressing the probability of just one hypothesis, e.g., the questioned sample comes from the suspected person, the

expert should consider the probability of the evidence, given at least another competing hypothesis, e.g., the questioned sample comes from any other person, and thus assess the extent to which each one of them is supported by the evidence. The odds form of Bayes theorem (Eq. 2.1) shows how new data (questioned recording) can be combined with prior background knowledge (prior odds) to give posterior odds for judicial outcomes or issues. It allows for revision based on new information of a measure of uncertainty (likelihood ratio of the evidence (E)) which is applied to the pair of competing hypotheses: H_0 - the suspected speaker is the source of the questioned recording, H_1 - the speaker at the origin of the questioned recording is not the suspected speaker. The prior and posterior odds are the province of the court and only the likelihood ratio (LR) is the province of the forensic expert.

$$\underbrace{\frac{p(H_0|E)}{p(H_1|E)}}_{\text{posterior odds}} = \underbrace{\frac{p(E|H_0)}{p(E|H_1)}}_{\text{LR}} \cdot \underbrace{\frac{p(H_0)}{p(H_1)}}_{\text{prior odds}}. \tag{2.1}$$

The likelihood ratio (LR) is a measure of the support that the evidence E lends to each of the competing hypotheses, and is a measure of the strength of the evidence. It is the ratio of the likelihoods of observing the evidence given the two hypotheses.

The court would like to determine the posterior odds with respect to the hypotheses, or the ratio of the probabilities of the two hypotheses, given the evidence. It is not possible to determine the posterior odds without knowing the prior odds of the hypotheses. Often, the police, lawyers and the courts expect the expert to formulate the results of the analysis in this form. An example of such a request is, for instance, 'Is the speaker in the reference recordings the same speaker as the unknown speaker in the questioned recording ? *'. The prior odds of the analysis are determined by the circumstances of the case, and normally the expert does not have access to this information (this lack of access is even desirable). By considering the prior odds of the hypotheses as well as the likelihood ratio of the evidence, given the two hypotheses, it is possible to determine the posterior odds, based on which the court will make it's decision. The expert should limit the statement of conclusions to the probabilities of the evidence given the two hypotheses. Rose [2005] suggests that the expert's conclusion of an analysis could be worded as follows: *'There are always differences between speech samples, even from the same speaker. In this particular case, I estimate that you would be about 1000 times more likely to get the difference between the offender and suspect speech samples had they come from the same speaker than from different speakers. This, prior odds pending, gives moderately strong support to the prosecution hypothesis that the suspect said both samples'.*

*The questions in the NFI fake case evaluation presented in Appendix D, are in the same 'posterior' style, while the analysis and responses estimate only the likelihood ratio

Certain authors like Cook et al. [1998] have proposed that the Bayesian methodology should be used even at the stage of assessment of a forensic case. They describe a model for assessing and interpreting forensic cases, where the prosecution proposition as well as a defence proposition, against which it will be tested, are established, the evidence expected if the defense hypothesis is true, as well as the evidence expected if the prosecution hypothesis is true, are considered, and finally, an estimation of the likelihood ratios expected is presented to the court or agency requiring this analysis. This is useful in determining the probative value of the evidence or the likelihood ratio.

The Bayesian interpretation methodology and the courts

Although the Bayesian interpretation methodology provides a logical and coherent framework for the analysis of evidence, with a clear division of the roles of the expert and the courts, its acceptance in the courts has been gradual. Some of the criticism against using Bayesian interpretation of evidence concern the practical aspects for juries to incorporate this in their assessment of a case. The element of judgement of the expert in determining some of the numerical values in calculating the likelihood ratio may be concealed. Jurors do not evaluate evidence using a formula but with their knowledge and commonsense, and individual jurors may differ in their estimation of what probabilities should be attached to each piece of evidence, thus complicating the combination of these estimates during deliberations. Also, the Bayesian interpretation framework has come under criticism as being complex and requiring the jury to go into realms that are not their proper task *.

Bayesian interpretation in forensic automatic speaker recognition

The forensic expert's role is to testify to the worth of the evidence by using, if possible, a quantitative measure of this worth. Consequently, forensic automatic speaker recognition methods should provide a statistical-probabilistic evaluation which attempts to give the court an indication of the strength of the evidence, given the variability of speech.

Comparing an unknown questioned recording with a recording of the suspected speaker gives a numerical assessment of the distances between them or a subjective opinion expressing their similarities or differences. The probative value of this similarity has to be evaluated by assessing the probabilities of these observations against competing hypotheses [Champod and Meuwly, 2000]. The strength of this evidence

*R v Doheny. Court of Appeal Criminal Division. No. 95/5297/Y2; 1996. 'the introduction of Bayes' theorem, or any similar method, into a criminal trial plunges the jury into inappropriate and unnecessary realms of theory and complexity deflecting them from their proper task'

is the result of the interpretation of the evidence, expressed in terms of the likelihood ratio of two alternative hypotheses.

For instance, the two competing hypotheses can be:

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

We discuss how this likelihood ratio can be estimated using a corpus-based Bayesian methodology for speaker recognition. This methodology for forensic speaker recognition was presented in [Meuwly, 2001; Drygajlo et al., 2003]. The Bayesian methodology requires, in addition to the trace, the use of three databases: a suspect reference database (R), a suspect control database (C) and a potential population database (P). When the performance of the system is being evaluated, it is also necessary to use a database of traces (T) .

- The P database contains an exhaustive coverage of recordings of all possible voices satisfying the hypothesis: *anyone chosen at random from a relevant population could be the source of the trace.* These recordings are used to create models to evaluate the between-sources variability (inter-variability) of the potential population with respect to the trace.
- The R database contains recordings of the suspect that are as close as possible (in recording conditions and linguistically) to the recordings of speakers of P, and it is used to create the suspect speaker model as is done with models of P.
- The C database consists of recordings of the suspect that are very similar to the trace and is used to estimate the within-source variability (intravariability) of his voice.

A brief summary of the methodology proposed in [Meuwly, 2001; Drygajlo et al., 2003] to calculate a likelihood ratio for a given trace is as follows (schema illustrated in Fig. 2.1) :

- The features of the trace are compared with the statistical model of the suspect (created using database R), and the resulting score is the statistical evidence value E.
- The features of the trace are compared with statistical models of all the speakers in the potential population (P). The distribution of log-likelihood scores indicates the between-sources variability of the potential population, given the trace.

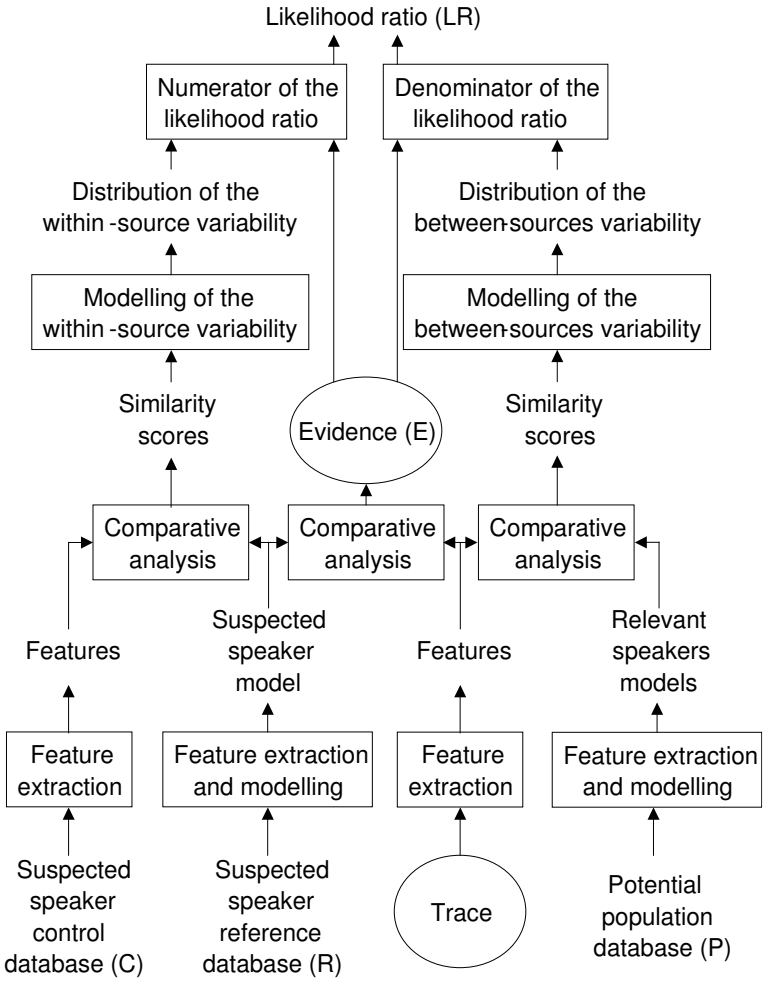


Figure 2.1: *Schema for the evaluation of the likelihood ratio presented in [Drygajlo et al., 2003]*

- The control database (C) recordings of the suspect are compared with the models created with R for the suspect, and the distribution of the log-likelihood scores gives the suspect's within-source variability.

- The likelihood ratio (i.e., the ratio of support that the evidence E lends to each of the hypotheses) is given by the ratio of the heights of the probability densities of the *within-source* and *between-sources* distributions at the point E. This is illustrated in Fig. 2.2 where, for an example forensic case, a likelihood ratio of 9.165 is obtained for $E = 9.94$.

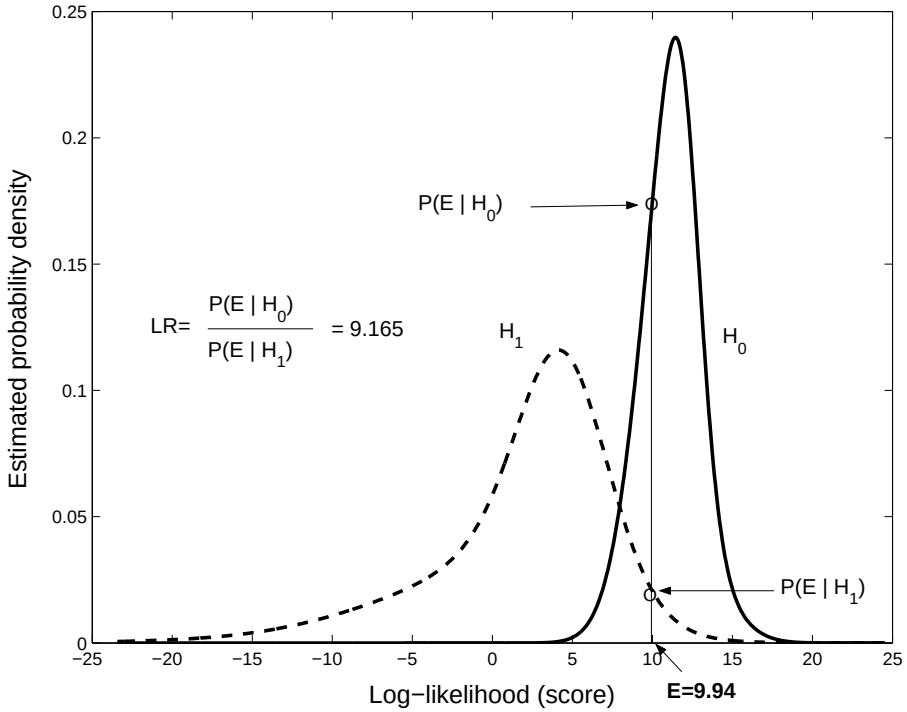


Figure 2.2: *Illustration of the estimation of LR using the corpus-based Bayesian interpretation method*

Estimating the significance of the strength of evidence

The strength of evidence (likelihood ratio) can be evaluated by estimating and comparing the likelihood ratios that are obtained for the evidence E in mock cases, when hypothesis H_0 is true and when the hypothesis H_1 is true. By creating mock cases which correspond to each of these hypotheses and calculating the LRs obtained for each of them, the performance and reliability of the speaker recognition system can be evaluated. In this way, it is possible to obtain two distributions; one for the hypothesis H_0 and the other for the hypothesis H_1. With these two distributions, it is possible to find the significance of a given value of LR that we obtain for a case, with respect to each of these distributions.

In order to measure the overall performance of the forensic speaker recognition methods, the cases can be separated into those where it was known that the suspected speaker was the source of the questioned recording and those where it was known that the suspected speaker was not the source of the questioned recording. These results are represented using cumulative probability distribution plots called Tippett plots. The Tippett plot represents the proportion of the likelihood ratios greater than a given LR, i.e., $P(LR(H_i) > LR)$, for cases corresponding to the hypotheses H_0 and H_1 true. The separation between the two curves in this representation is an indication of the performance of the system or method, with a larger separation implying better performance than a smaller one. This method of representation of performances was proposed by Evett and Buckleton [1996], in the field of interpretation of forensic DNA analysis. This representation has been named the Tippett plot in [Meuwly, 2001], referring to the concepts of within-source comparison and between-sources comparison defined by Tippett et al. [1968]. An example of one such Tippett plot can be seen in Fig. 2.3.

This performance evaluation method is different from the speaker verification domain where the task is to compare two recordings and ascertain whether they have the same or different sources. Normally, a threshold is used in the verification domain to decide whether the two recordings come from the same source. The Receiver Operating Characteristic (ROC), was used to compare the recognition performances of different systems. The ROC curve plots the false acceptance rate or false match rate (with respect to a threshold score) on the horizontal axis, with the correct acceptance rate plotted on the vertical. Martin et al. [1997] introduced a variant of this curve called the DET (Detection Error Tradeoff) curve where the false acceptance rate (FAR) and the false rejection rate (FRR) are plotted on the horizontal and vertical axes respectively. In the forensic domain, it is not acceptable to use such a threshold, and measures such as the DET curves and the equal error rates (EER) can only be used to measure the performance of the systems. An example of one such DET plot can be seen in Fig. 2.4.

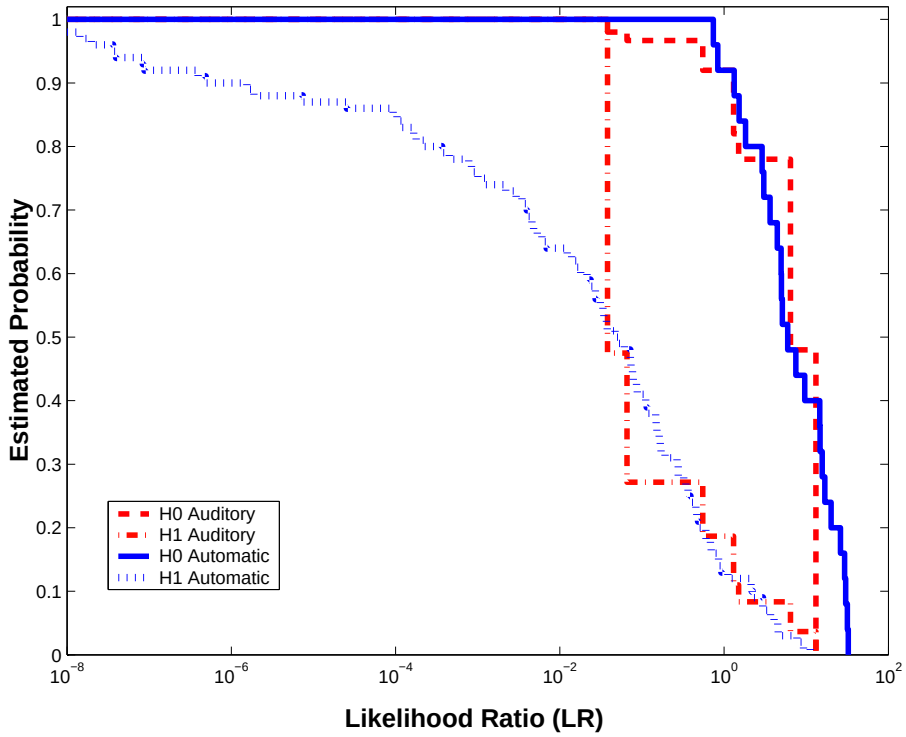


Figure 2.3: *An example of a Tippett plot*

2.2 Automatic speaker recognition

Speaker recognition is traditionally classified into speaker identification and speaker verification. *Speaker identification* is the process of determining from a set of speakers, which particular speaker a certain utterance comes from, and *speaker verification* involves determining whether a given utterance actually belongs to the speaker who claims to have produced it. Basically, identification and verification vary in the size of the number of possible alternatives for the speaker who is at the source of the test utterance [Rosenberg, 1976; Doddington, 1985; Furui, 1997]. In speaker verification, there is a single speaker whose voice is compared with that of the test utterance, and a decision is made either to accept or to reject this speaker as the source of the test utterance. In speaker identification however, the test utterance is compared with a population of possible candidates who could be at the source of this utterance. Speaker identification can be performed in two test scenarios, i.e., *open* and *closed set*, that differ in their possible outcomes. In closed set identification, one speaker among the set of possible speakers in the database is the source of the test recording, while in open set identification, it is possible that *none* of the speakers in the database are at the source of the test recording.

Automatic speaker recognition is further divided on the basis of the content of

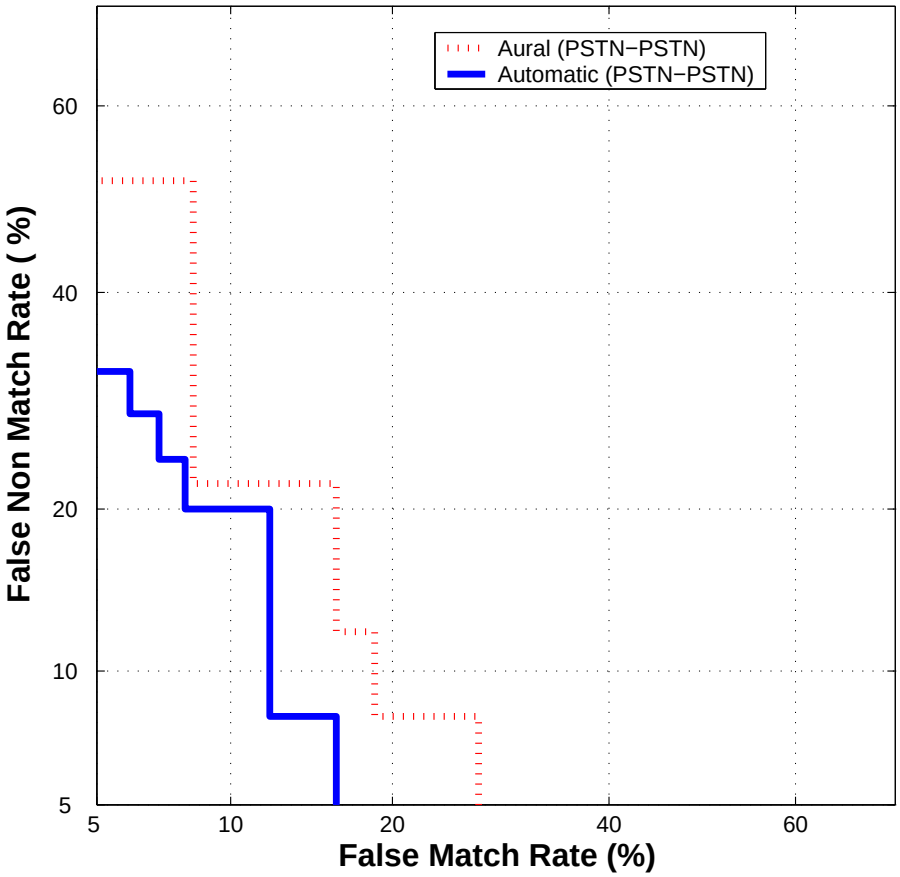


Figure 2.4: *An example of DET Plot*

the utterances used for training and testing the recognizer into *text dependent* and *text independent* methods. In *text-dependent* speaker recognition, a specific utterance (keyword or sentence) from the speaker is used for training the system, and it is necessary to use the very same utterance for testing his identity. Template matching techniques are often used for text-dependent speaker recognition. This utterance could be a PIN code, password, etc. This type of recognition is popular with access control systems. In *text-independent* speaker recognition, the training and testing utterances can be different, i.e., the speaker is not required to speak the same utterance for both training and testing. Text-independent speaker recognition is more suitable for forensic tasks.

2.2.1 Features used for speaker recognition

Ideally, the features chosen for speaker recognition must satisfy the following criteria [Wolf, 1972]:

- *Should have a lower within-speaker (within-source) variability and a relatively higher between speakers (between-sources) variability*: If the variation of a feature for a speaker is equal or more than the variation of the same feature across other speakers, it will not be useful for the recognition of the speaker. The ratio of within-source and between-sources variation for a feature is also known as the F-ratio [Fukunaga, 1990] and a higher F-ratio is desirable for the features used in speaker recognition. Several acoustic features are chosen on this basis [Rose, 2003, :3060].

- *Should be stable over time*: The speech of an individual undergoes very short-term variations (of the order of seconds, where the way words are pronounced or emphasized changes according to the conversational context), short-term variations (emotional state, stress, fatigue, illnesses like a cold or laryngitis) as well variations in the longer term (due to age, state of health, etc.). In practice, it is difficult to find features that are truly stable in both the short and long term.

- *Should be difficult to disguise or mimic*: A characteristic that the speaker is successfully able to mask or change, at will, is not a good choice for a characteristic feature.

- *Should be robust to transmission and noise*: This is an important consideration as environmental noise and the modern communication networks for transmission of the voice leave their mark on the signal.

- *Should be relatively easy to extract and measure, and should occur frequently in the speech samples*:

  A sufficient number of examples of the characteristic should exist (in all the samples compared), as conclusions drawn on the similarity should be statistically significant.

In addition to these often-cited criteria, Rose [2002] proposes that *each parameter measured must be maximally independent of the other*. If similarities are observed in more than one set of features, then the relation between these two features must be considered. If the conditional dependency between features can be estimated reliably, it is possible to use features that are not strictly mutually independent. However, without careful consideration of the independence of features, the similarities seen may, in effect, be considered more than once. Rose [2002] also presents a classification of forensic phonetic parameters into acoustic or auditory features and linguistic or non-linguistic features. There had been sharp divisions in the forensic speaker recognition community about using only acoustic or only auditory features, and

nowadays, an approach combining aural and acoustic methods is generally accepted. In automatic speaker recognition, acoustic parameters are extracted from the speech signal. In this thesis, we use short-term acoustic features (Sec. 2.2.1).

The morphology of the vocal tract, made up of the pharynx, the oral cavity and the nasal cavity, influences the voice. The human speech production mechanism is driven by the acoustic excitation of the vocal tract (Fig. 2.5), by airflow from the lungs, which pass through the vocal cords. The resonances of the vocal tract are called *formants*. The speech produced can be classified into *voiced* and *unvoiced sounds* which differ in the kind of excitation of the vocal tract required to produce them. For instance, voiced sounds (e.g., vowels, nasal sounds) are produced by excitation of the vocal cords with nearly periodic vibrations while unvoiced sounds, like fricatives are the result of an excitation source with a constant power density spectrum (like white noise). The glottal pulse shape, the fundamental frequency, the position of the articulators, the response of the vocal tract to voiced and unvoiced excitation, all contribute to the distinctiveness of a speaker's voice. The likely shape of the vocal tract can be approximately estimated from the analysis of the spectral shape of the voice signal [Campbell, 1997]. In automatic speaker recognition, the coefficients representing the sounds, taking into consideration the vocal tract shape and excitation are parameterized and used as features.

Speech parameterization using short-term feature extraction methods

Speech parameterization is used to represent the speech signal into a redundant compact form. In short-term feature extraction, speech is analyzed in short segments or windows, under the assumption that within a window of typically 20-30 milliseconds, the signal can be assumed to be stationary. In addition, each of these analysis windows overlap, e.g., by 25% or 50%. A set of coefficients calculated for the analysis frame represents the feature vector.

The acoustic features that are commonly used for feature extraction include, Mel-frequency cepstral coefficients (MFCCs) [Davis and Mermelstein, 1980], linear-frequency cepstral coefficients, linear prediction cepstral (LPC) coefficients [Makhoul, 1975], and perceptual linear prediction (PLP) cepstral coefficients [Hermansky, 1990]. Some of the channel compensation techniques used in order to make these features robust to the channel mismatch and noise include RASTA processing [Hermansky, 1994] (for PLP), cepstral mean substraction (CMS) and the missing feature approach (for noise) [El-Maliki, 2000]. An evaluation of the various features for robust speaker identification is presented in [Reynolds, 1994].

The MFCCs-based feature extraction technique basically includes windowing the signal, applying the fast Fourier transform (FFT), taking the log of the magnitude and then warping the frequencies on a Mel scale, followed by applying the inverse

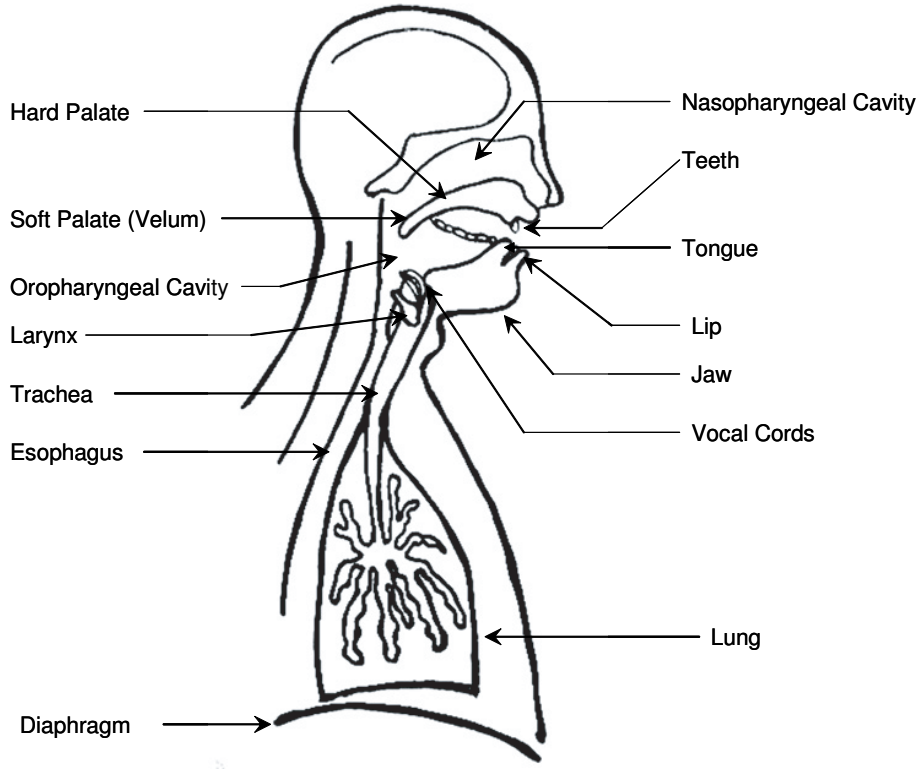


Figure 2.5: *Vocal tract and articulators (Illustration by Faye Carasco)*

FFT. MFCCs have been demonstrated to show good performance in speaker recognition. The Mel scale is based on the human perception of sound frequencies that is nonlinear and places less emphasis on higher frequencies. Sometimes, especially in speech recognition applications, the time derivative of the Mel-cepstra called *delta* or *delta-delta* features are used in order to include information about the temporal aspect of the features. Removing the mean of the cepstral features reduces the effect of channel distortions. Perceptual linear prediction analysis (PLP) was introduced by Hermansky [1990]. This technique is inspired by the properties of human perception. In this method, a convolution of the short-term power spectrum of speech, with a critical-band masking pattern, is followed by resampling, preemphasis with an equal-loudness curve and compression using a cubic root nonlinearity to give an all-pole model that is consistent to phenomena associated with human speech perception. RASTA-PLP stands for 'RelAtive SpecTrAl' and is a feature extraction technique that is particularly useful when using telephone speech, as it is more robust to channel distortions. It is an improvement on the PLP method. A description of the RASTA-PLP feature extraction is presented in Sec. 2.2.5.

2.2.2 Modeling algorithms used for automatic speaker recognition

Campbell [1997] defines two types of models, i.e., template models and stochastic models. These models differ in the way that the pattern matching is performed. Template models follow a deterministic pattern matching method wherein test vectors are evaluated on the assumption that they are inexact replicas of the template, and a distance is calculated between the template frames and the observation or test frames. Stochastic modeling uses probabilistic pattern matching, and the assumption that is tested is that the test observation and the training data were created from the same underlying process. In stochastic modeling, the conditional probability or likelihood of the test observation, given the model, is used as a measure of the similarity of the test and training frames.

Template modelling: Dynamic Time Warping

Dynamic Time Warping (DTW) is a template model used in text-dependent speaker recognition. Basically, it involves comparing a sequence of M (X_1,...,X_M) vector templates with another sequence of N (Y_1,...,Y_M) vectors by calculating the accumulated distance between these two sequences. It performs a 'constrained, piece-wise linear mapping', aligning the two signals and minimizing the distance between them [Campbell, 1997]. In aligning the signals, which may not be of the same lengths (because of the timing differences in human speech), the time axes (either for one or both the signals) are 'warped', and thus the name dynamic time warping. This warping compensates for speaker-rate variability. If the two signals are identical, then the path between them is diagonal, and thus, distance between them is minimal. The distance measure is the accumulated deviation from the diagonal. This text-independent method has limited application in forensic problems, although it is useful in order to compare utterances containing the same words.

Template modeling: Vector Quantization

Vector quantization [Gray, 1984] is a form of template modeling that uses several templates in order to represent different analysis frames of the speech signal. A clustering algorithm is used in order to group together frames of speech into a vector code-book that is then used to represent the speaker who is enrolled into the verification system. Some of the earliest work in speaker recognition, using vector quantization techniques, was done by Soong et al. [1985] who observed that a set of short-time raw feature vectors of a speaker could be used directly to represent acoustical, phonological or physiological characteristics of that speaker, if the training set included sufficient variation.

Speaker-based vector quantization estimates the partitioning of the feature vector space into non-overlapping, convex partitions, and all the vectors within a partition are represented by a centroid vector. The code-book generation is done by finding a partitioning that minimizes the average distortion over the whole training data. The set of these clusters is called the vector code-book, which represents the speaker. These code-books are of considerably lower dimensions than the signal that is being modeled. In applications where memory usage is an issue, vector quantization is particularly useful, because the size of the code-book is small. The code-book size, which can be arbitrary (provided it is smaller than the number of data points), has been observed to be a factor in the accuracy of the algorithm. Increasing the size of the code-book has been observed to decrease the error rates in speaker recognition. As with other pattern recognition methods, increasing the number of parameters required to model the data, might result in learning the data and not the general characteristics of the speaker. When a test utterance is compared with the code-book of the speaker, a distance can be calculated from the frames of the test utterance to the centers of the code vectors representing the speaker. The match score is the distance between the test vector and the minimum distance code word in the code-book. For a code-book (C), the match score z of N frames of speech $(x_1, ..., x_N)$ is given by

$$z = \sum_{i=1}^{N} \min_{x_c \in C} (d(x_i, x_c)). \tag{2.2}$$

This distance measures the similarity between the test utterance and the model of the speaker enrolled.

Stochastic Modeling: Gaussian Mixture Modeling

For text-independent speaker recognition, one of the most widely used techniques to model the probability density of the input features is the Gaussian mixture model (GMM) [Reynolds, 1992]. While more complicated likelihood functions like the hidden Markov models (HMM) can also be used for both text-dependent and text-independent speaker recognition, these models have not shown to be better than GMMs [Bimbot et al., 2004]. The GMM can be considered to be an HMM with a single state.

The likelihood function for a feature vector x (D-dimensional) is given as:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x). \tag{2.3}$$

This likelihood, is a weighted sum of each of the M Gaussian component densities $p_i(x)$. The individual mixture weights (w_i) sum up to 1. Each Gaussian mixture model (λ) is characterized by three parameters, i.e., the means (μ), variances (Σ) and weights (w). For an M-component GMM, the statistical model corresponds to, $\lambda = (\mu_i, \Sigma_i, w_i)$ for $i = (1, ..., M)$.

The likelihood corresponding to the ith Gaussian component is given by:

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}. \tag{2.4}$$

An iterative expectation-maximization (EM) algorithm is used to estimate the maximum likelihood model for the training feature vectors [Dempster et al., 1977]. The estimates of the Gaussian mixture model parameters are improved with each iteration, with the likelihood of the model, given the training feature vectors, increasing with each iteration until convergence. The training can be stopped when the additional change in likelihood for each iteration is below a certain threshold or a certain predetermined number of iterations have been made.

In the estimation of the model parameters, it is possible to choose, either full covariance matrices or diagonal covariance matrices. Normally, diagonal covariance matrices are used because it is computationally less intensive to use them than to use full covariance matrices, because of the repeated inversions of these matrices.

The log-likelihood (also referred to as the log-likelihood score or the score in this thesis), for a sequence of N feature vectors $X = (x_1, ..., x_N)$ with respect to a model λ, is given by:

$$\log p(X|\lambda) = \frac{1}{N}\sum_{j=1}^{N} \log p(x_j|\lambda). \tag{2.5}$$

This is the average log-likelihood score, and since it is averaged over all the feature frames, it is independent of the duration of the test vector. GMM is an *unsupervised training* algorithm and is ideal for text-independent speaker recognition as it models only the distribution of the features from the speech and does not consider the temporal sequencing of the features.

Maximum a posteriori (MAP) adaptation

Gaussian mixture models for a speaker's voice can be trained using the modeling described in the Sec. 2.2.2. For this, it is necessary that sufficient training data is available in order to create a model of the speech. Another way of estimating a statistical model, which is especially useful when the training data available is of short duration, is by using maximum a posteriori adaptation [Gauvain and Lee, 1994] of a

background model trained on the voices of several other speakers. This background model is a large GMM that is trained with a large amount of data which encompasses the different kinds of speech that may be encountered by the system during training. These different kinds may include different channel conditions, composition of speakers, acoustic conditions, etc. Often, the order of these GMMs can be between 512 and 2048. A summary of the adaptation steps presented in [Bimbot et al., 2004] is given below.

For each mixture i from the background model, $Pr(i|x_t)$ is calculated.

$$Pr(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)}. \tag{2.6}$$

and

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \quad e^{\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'(\Sigma_i)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}}. \tag{2.7}$$

Using $Pr(i|x_t)$, sufficient statistics are calculated for the weight, mean and variance parameters as follows:

$$n_i = \sum_{t=1}^{T} Pr(i|x_t). \tag{2.8}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t. \tag{2.9}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t) x_t^2. \tag{2.10}$$

These new statistics calculated from the training data are then used adapt the background model, and the new weights ($\hat{\omega}_i$), means ($\hat{\mu}_i$) and variances ($\hat{\sigma}_i^2$) are given by:

$$\hat{\omega}_i = \left[\alpha_i n_i/T + (1-\alpha_i)\omega_i\right]\gamma. \tag{2.11}$$

$$\hat{\mu}_i = \alpha_i E_i(x) + (1-\alpha_i)\mu_i. \tag{2.12}$$

$$\hat{\sigma}_i^2 = \alpha_i E_i(x^2) + (1-\alpha_i)(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2. \tag{2.13}$$

The adaptation coefficient α_i controls the balance between the old and new model parameter estimates. A scale factor γ is used, which ensures that all the new mixture weights sum to 1.

$$\alpha_i = \frac{n_i}{n_i + r}, \tag{2.14}$$

where r is a fixed relevance factor, which determines the extent of 'mixing' of the old and new estimates of the parameters. Low values for α_i ($\alpha_i \rightarrow 0$), will result in new parameter estimates from the data to be de-emphasized, while higher values ($\alpha_i \rightarrow 1$) will emphasize the use of the new training data-dependent parameters.

2.2.3 Alternative techniques used in speaker recognition

Artificial Neural Networks

Neural networks are *unsupervised* classifiers used to learn complex mappings between inputs and outputs, and have been used for speaker recognition [Farrell et al., 1994; Oglesby, 1989]. The multilayer perceptron (MLP) is a popular neural network consisting of one input layer, one or more hidden layers and one output layer. The training of MLPs is usually performed using an iterative gradient algorithm known as back-propagation [Bishop, 1995]. The MLP can be trained to estimate the posterior probabilities of different classes. The MLP estimates the posterior probability $P(q_k|x_n)$ for a test vector x_n belonging to the class q_k. For a group of N speakers, an MLP-based speaker recognition system can be constructed as follows. The input to the MLP are the feature vectors corresponding to a speaker i, and the output is a target vector of size N whose ith component is labeled as 1, and all the other $N-1$ elements are labelled as 0. One of the difficulties in training MLPs with large speaker populations is that for a given speaker, most of the training vectors have '0' labels, and only a few have '1' labels, and because of this, the MLP is biased more towards '0' for every test vector [Farrell et al., 1994].

The neural network learns a nonlinear mapping between the features of the speaker and the identity of the speaker (represented in the target vector), and when a test vector is given as input, the most likely target label, given this observation, is calculated. Since the neural network directly estimates the posterior probability of a given class ($P(q_k|x_n)$) the likelihood of the observation given the class $P(x_n|q_k)$ can be estimated if the prior probability $P(q_k)$ and $P(x_n)$ are known. One difficulty in using neural networks is that the posterior probabilities cannot be directly used in the Bayesian interpretation framework, where we would like to evaluate $P(x_n|q_k)$, and in forensic casework, we may not have access to the prior probabilities, $P(q_k)$ and $P(x_n)$.

2.2.4 Classifying speakers according to the difficulty in recognizing them

In early work in text independent speaker recognition [Li and Porter, 1988], the issue of certain speakers affecting the recognition accuracy more than others has been reported. They reported that for 10 second test durations, 90% of the errors were caused by 30% of the population. This issue of certain speakers accounting for a disproportionately high number of false acceptances or rejections have led to speakers being classified into different groups. This classification is of interest in the design of a forensic speaker recognition database, as every database will contain a proportion of speakers who are easy to classify and others who are difficult to classify. Doddington et al. [1998] has proposed an interesting classification of speakers into different categories on the basis of the difficulty of recognizing them using an automatic system. Each category of speakers has been assigned to a different animal.

- **Sheep**: They represent the majority of the population, and are the default speaker type. Most automatic systems for verification and recognition perform reasonably well for them with a low number of false acceptances and false rejections.

- **Goats**: The goats represent a section of the speakers who are particularly difficult to recognize. In the evaluation of speaker recognition performance, the goats account for a disproportionate share of false rejection. This set of speakers is of particular interest as they are consistently not recognized correctly, and this implies that both verification systems and recognition systems will be affected by their presence.

- **Lambs**: The speakers whose voices are most easily imitated by another speaker are called lambs. The presence of lambs tends to increase the false acceptance rate. These characteristics of the speech extracted for this set of speakers are also easily observed in imposter speakers, and this problem could potentially represent a system weakness [Doddington et al., 1998].

- **Wolves**: The wolves are the speakers whose voices are complementary to the lambs, in the sense that the characteristic features of their voices are exceptionally similar to the features of other speakers. Their speech is often likely to be recognized as that of some other speaker. Wolves would account for a disproportionately large share of the false acceptances. Like the lambs, these speakers also represent a potential system weakness.

2.2.5 Using automatic speaker recognition techniques in forensic automatic speaker recognition

In this section, we describe how we can put together algorithms for feature extraction, modeling of features, as well as statistical modeling of scores, to create a forensic automatic speaker recognition system that is adapted to the Bayesian interpretation framework described in Meuwly [2001]; Drygajlo et al. [2003].

Automatic speaker recognition consists of mainly five steps, namely, preprocessing of the input speech file, extraction of features, modeling, comparison of the features and models, and the interpretation of the results. The interpretation can be a decision in the case of speaker verification systems, and a likelihood ratio estimation in the case of a forensic speaker recognition system. A schema of this processing is presented in Fig 2.6.

Preprocessing

The input file is pre-processed, and the quality of the audio recording is checked in order to determine whether it can be used for further analysis. Careful examination of the spectrogram of the audio and listening to the recording, along with measures like the signal-to-noise ratio (SNR), may indicate what portions of the recording maybe useful for analysis and what should be discarded. If the analysis deems it necessary, filtering, noise reduction, amplitude normalization, etc. can be performed. Removing the zones of silence (non-speech segments), as they do not contain any speaker-dependent information, has been shown to improve the accuracy of recognition, and this can also be performed in the preprocessing phase [Reynolds, 1992].

Preprocessing: Voice Activity Detection

Voice activity detection has been widely used for variable-rate speech coding, speech recognition and speech enhancement [Sohn et al., 1999], and has been of particular interest in telephone transmissions as silence constitutes large part of the transmission time, in each direction, for telephone calls [Srinivasan and Gersho, 1993].

In our experiments, time-varying estimates of the signal-to-noise ratio (SNR) are used in order to separate speech and non-speech zones using Murphy's algorithm [Reynolds, 1992, :119]. The noise energy floor is continuously estimated by tracking the minimum values of the frame energy. This noise energy floor is slowly raised and compared with the energy of each successive frame. If the energy of the current frame is lower than that of the noise floor estimate, then the noise floor estimate is updated to the energy of the current frame. An instantaneous SNR is calculated as a ratio between the noise floor calculated and the signal energy of the current frame. If the instantaneous SNR is above a certain threshold for several consecutive frames,

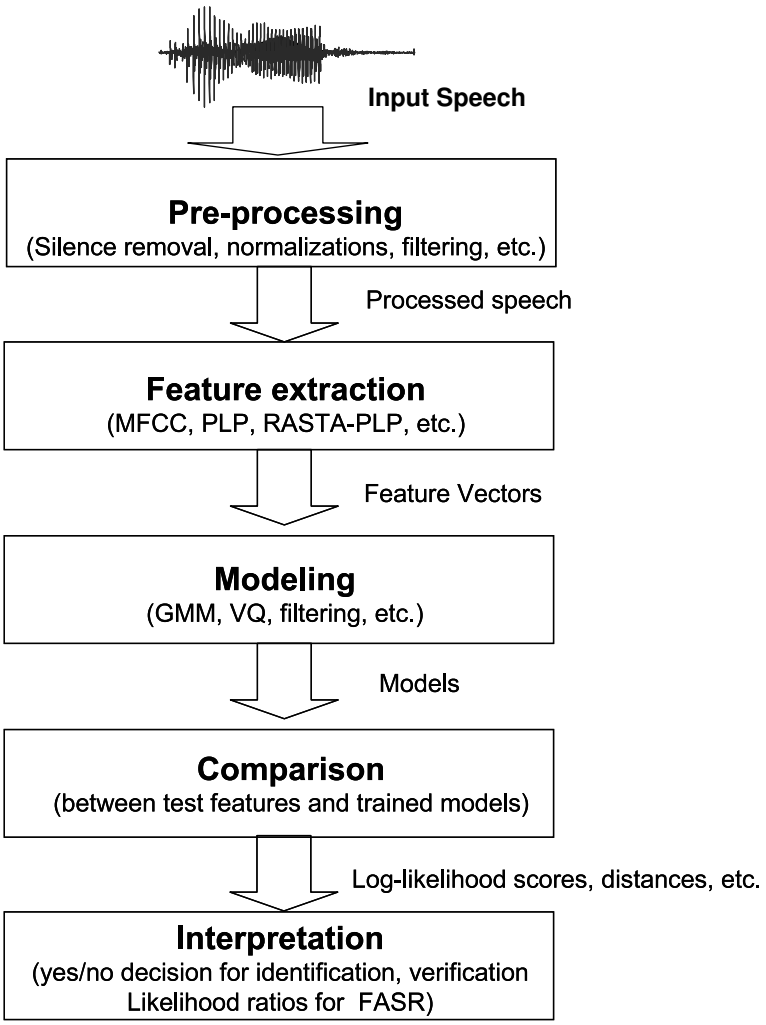


Figure 2.6: *Stages in the automatic speaker recognition process*

the frames are classified as speech segments. In marking the non-speech and speech segments, a decision about the markings is not made until a sufficient number of such frames have been observed. These frames are then marked as speech and non-speech retrospectively, i.e., once a marking decision (speech or non-speech) has been made, a certain number of earlier frames is also similarly marked as speech or non-speech.

These time-varying estimates of the SNR are used in order to separate speech and silence zones. The frame size is 160 samples (for a sampling rate of 8000 Hz), with a shift of 50 samples, and the minimum SNR in order to make a decision of speech or non-speech is 5dB. Five frames before and after the frame that induced the decision that a speech segment is present are also marked as speech.

The steps involved in this pre-processing phase are summarized as follows:

- An estimate of the noise energy floor that is present in the signal is calculated

using the first few frames.

- The noise floor estimate is raised slightly by a constant factor (say a 1 % increase).
- The energy of each successive frame is calculated and compared with the last noise floor estimate.
- If the energy of the current frame is lower than that of the noise floor estimate, then the noise floor estimate is updated to the energy of the current frame.
- The instantaneous SNR is calculated as a ratio between the noise floor we have calculated and the signal energy of the current frame.

$$\mathrm{SNR}_{inst}(t) = -10 \log \frac{\text{frame energy}(t)}{\text{noise energy floor}(t)} \tag{2.15}$$

- If the instantaneous SNR is above a certain threshold we consider that it could be a speech segment. If this behavior is repeated for several consecutive frames, i.e. high instantaneous SNRs are observed over several frames, the frames are classified as speech segments.
- Similarly for non-speech frames, if the instantaneous SNR is below a certain threshold we consider that it could be a non-speech segment. If this behavior is repeated for several consecutive frames, i.e. low instantaneous SNRs are observed over several frames, the frames are classified as non-speech segments.
- In marking the non-speech and speech segments, a certain number of hangover frames are considered, where a decision about the markings is not made until sufficient number of such frames have been observed. These hangover frames are marked as speech and non-speech retrospectively, i.e., once a marking decision (speech, non-speech) has been made, a certain number of earlier frames also similarly marked as speech or non-speech.

Feature Extraction: RASTA-PLP feature extraction

In the experiments in this thesis, RASTA-PLP features were used because of their robustness to the distortions caused by different communication channels. The idea behind RASTA is the suppression of factors in the speech signal that are changing slowly. Similarities have been drawn between the insensitivity of human audition to slow changes in the spectral signals, e.g., steady background noise does not seem to have a severe adverse effect on communication between speakers [Hermansky, 1994]. In RASTA-PLP, stationary convolutional influences of the steady state spectral

factors are reduced by the band-pass or high-pass filtering of each spectral trajectory. The RASTA band-pass filtering is performed on the logarithmic spectrum (linearly transformed logarithmic spectrum) or the spectrum compressed by a non-linearity of $ln(1 + const \cdot x)$. Influences of linear distortions and additive noise in speech are alleviated.

RASTA-PLP feature extraction performs the following for each analysis frame as described in [Hermansky, 1994]:

- Computation of the critical-band power spectrum.
- Transformation of the spectral amplitude through a compressing static non-linear transformation.
- Filtering the time trajectory of each of the transformed spectral components.
- Multiplication by the equal loudness curve and then raising it to the power 0.33 (to simulate the power law of hearing).
- Computation of an all-pole model of the resulting spectrum.

In our experiments, we have used RASTA-PLP feature vectors with an analysis window size of 20 milliseconds, a window step size of 10 milliseconds, at a sampling frequency of 8000 Hz (almost all the recordings considered are in telephone quality and hence were sampled at 8000 Hz), with 12 coefficients to be calculated for every frame.

Statistical Modeling : GMM

In the experiments presented in this thesis, the number of Gaussians used ranged between 32 and 64, depending on the task and the amount of data available for enrollment.

In Fig. 2.7, a 12 component GMM model of 12 MFCC features are presented. The thick outer curve for each feature represents the outer boundary of the GMM, and the gray background represents the features that are being modeled. The component Gaussians whose individual contributions, represented by the thinner curves within the gray area, are added to give the total likelihood represented by the thick outer curve.

2.2.6 Kernel-density estimation for modeling between- and within-source variability of the voice

Kernel density estimation [Silverman, 1986] can be used to estimate probability density functions, as long as the distributions are fairly smooth. Probability distributions

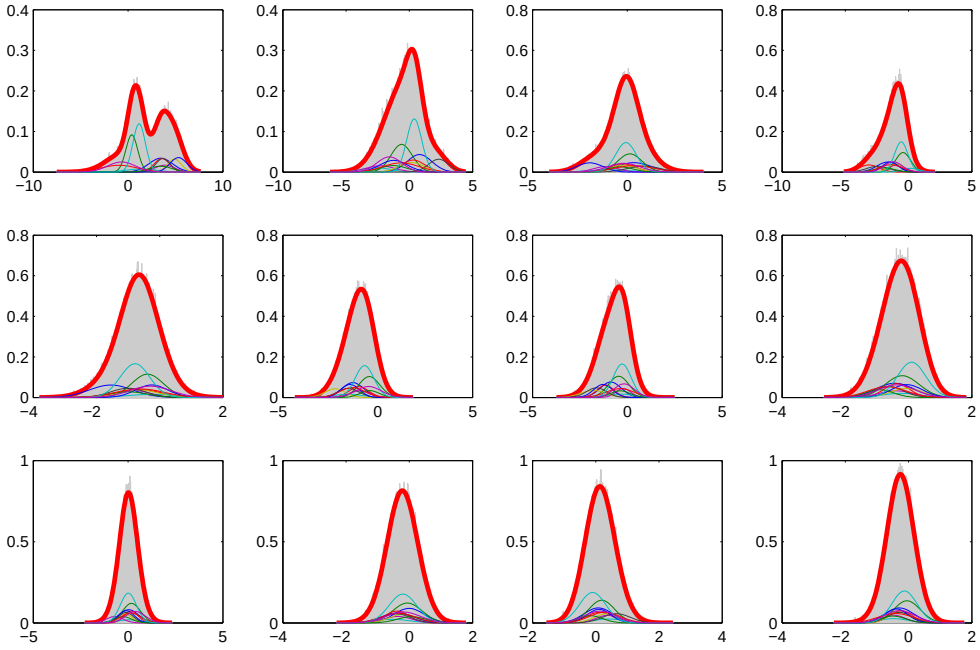


Figure 2.7: *A 12 component Gaussian mixture model of 12 MFCC features a speaker's voice*

are often described using histograms, where the occurrence of a sample within an interval is represented by a rectangular block. For histograms, there is a tradeoff between the width of the interval and the resolution of the probability density function. If the bins are too wide, then the histogram will not be representative of the underlying distribution, and if the bins are too narrow, then a lot of the 'detail' is modeled, and the general nature of the underlying features is lost. In kernel density estimation, this rectangular block is replaced by a kernel function, one of which is positioned over each sample. The probability density estimate for a point is obtained by adding all the individual probability density estimates over all the samples in the data set. Similar to the difficulty in using histograms in order to plot a probability density function due to the decision that has to be made about the bin size or interval width, for kernel density estimation, the spread of each Gaussian kernel has to be chosen. Similarly, if the variances of the Gaussian kernels chosen are large, then the resulting probability density function (pdf) curve is very smooth. For small values of variance, the pdf curve is spiky [Aitken and Taroni, 2004, :331].

For a dataset D, containing k elements, $D = x_1, ..., x_k$, where s is the sample variance, θ denotes the mean value of the element and λ is the smoothing parameter (this smoothing parameter λ has to be chosen). Aitken and Taroni [2004] suggest

that it is possible to choose λ subjectively, based on the best visual fits or from the scientist's personal experience.

The kernel density function is given by:

$$K(\theta|x_i,\lambda) = \frac{1}{\lambda s\sqrt{2\pi}} exp\left(-\frac{1}{2}\frac{(\theta - x_i)^2}{\lambda^2 s^2}\right). \tag{2.16}$$

The estimated probability density function is given by:

$$f(\theta|D,\theta) = \frac{1}{k}\Sigma_{i=1}^{k} K(\theta|x_i,\lambda). \tag{2.17}$$

The distribution of between-sources and the within-source variability scores can be modeled using kernel density estimation, especially when it is seen that each of these distributions are not normal. Kernel density estimation has been used for various forensic analyses such as DNA profiling, elemental composition of glass, cat hairs, etc. In our experiments, kernel density estimation has been used to model both within-source and between sources variability.

2.2.7 Double Statistical Model

The starting point in forensic automatic speaker recognition is the recordings, from which features are extracted and statistical models are created. The scores obtained by comparing the features and the statistical models are subsequently used to derive models of within-source and between-sources variabilities. In an individual case, in order to estimate the strength of evidence, a double statistical model is used to model the multivariate distribution of acoustic features and the univariate distributions of scores pertaining to between-sources and within-source variabilities. In order to model the multivariate distribution of features, Gaussian mixture modeling (GMM) is used, and kernel density estimation is used in order to model the within-source and between-sources score distributions.

In order to consider the average performance of the automatic system, probability distribution plots of the likelihood ratios obtained across several cases can be used. These probability distribution plots of the LRs are thus multi-case compared with the single LR of the double statistical model used in an individual case. Tippett plots represent the cumulative density plots of these probability distributions of LRs in cases where H_0 is true and H_1 is true. While they provide valuable insight into the *average* performance of the system in cases similar to a given case, each individual case may have a unique set of conditions that makes it difficult to compare the LR obtained to the corresponding Tippett plots.

2.3 Influence of modern communication networks and recording conditions on automatic speaker recognition

In [Bimbot et al., 2004], the authors have cited the 'lack of robustness to channel variability and mismatched conditions' as the 'biggest impediment' to the widespread use of automatic speaker recognition technology. Most automatic systems rely mainly on features related to the spectra, and thus, are also very dependent on the transmission channel and the noise which affect the spectrum greatly. The transmission channel and signal coding effects, ambient noise and line echoes all contribute to the degradation of the performance of conventional speaker recognition techniques. Mismatches between training and testing speech files, which are commonly encountered in forensic cases, contribute also to the significant decrease in performance compared to match channel conditions [Dunn et al., 2001; Nakasone and Beck, 2001].

2.3.1 Other groups working on this problem of mismatched conditions in forensic automatic speaker recognition

The effect of recording conditions on forensic speaker recognition has been an important research concern. The MIT Lincoln Laboratory group have approached this problem, proposing a framework for confidence estimation that takes into consideration, in addition to the scores, information such as the channel type, the SNR and the duration of utterances used [Campbell et al., 2005]. In order to obtain the confidence intervals, they use a neural networks based solution, with a multi-layer perceptron. The ATVS (Speech and Signal Processing Group), from the Universidad Autonoma de Madrid, has proposed methods such as Target-Dependent Likelihood Ratio Alignment (TDLRA) that help reduce the effect of mismatched conditions [Gonzalez-Rodriguez et al., 2004; Gonzales-Rodriguez et al., 2005]. The TDLRA calculates an alignment factor based on the Tippett plots created with enrollment data. This normalization attempts to be conservative, adjusting the proportion of H_1 cases so that only a predefined percentage will have an LR above 1. This conservative measure is claimed to ensure the 'presumption of innocence'. The authors have reported that this method had a positive effect on the performance of the system when the recording conditions of the potential population was mismatched, as the normalization between suspects performed by TDLRA led to the better performance of the system. Over the years, there has been a lot of research in the automatic speaker recognition community on the front-end features that are robust to channel and noise conditions.

Statistical compensation methods have been used in automatic speaker recognition in order to handle the problem of mismatched recording conditions, at the level of features, statistical modeling of the features, and at the level of scores. In the following sections (Sec. 2.3.2 and 2.3.3), we discuss statistical compensation and normalization techniques at the level of features, models and score distributions.

2.3.2 Compensation of channel mismatch with feature mapping and speaker model synthesis

Reynolds [2003] has proposed both feature mapping as well as speaker model synthesis to handle the issue of mismatch. In the speaker model synthesis (SMS) approach, a speaker model corresponding to one set of channel conditions is synthesized from another channel-dependent speaker model. The starting point for this synthesis is a root GMM that is trained using channel-independent data (aggregation of data from several different channel types). This model is then adapted, using channel-dependent data to each of the channel conditions. Since the root GMM was adapted to these channel conditions, there is still a correspondence between the Gaussians of the channel-dependent models and the root model. Thus, the transformation between channel-dependent models as well as the channel-independent models can be calculated by estimating the mean shift and the variance scaling of each of the Gaussians.

For training utterance in an 'unseen' channel condition, a synthetic model can be created in matched conditions. Mean shifting and variance scaling is applied to the model parameters in order to adapt a certain channel-dependent model into another channel-dependent model. A transformation (T) of parameters for a Gaussian component i, from condition $C1$ to condition $C2$ is given as:

$$T_i^{C1\to C2}(\omega_i) = \frac{\omega_i^{C2}}{\omega_i^{C1}}. \tag{2.18}$$

$$T_i^{C1\to C2}(\mu_i) = \mu_i + (\mu_i^{C2} - \mu_i^{C1}). \tag{2.19}$$

$$T_i^{C1\to C2}(\sigma_i) = \sigma_i \cdot \frac{\sigma_i^{C2}}{\sigma_i^{C1}}, \tag{2.20}$$

where ω, μ, σ represent the mixture weights, means and standard deviations respectively.

Thus, this method involves the following steps:

- The system detects the most likely channel-dependent background model, and MAP adaptation [Reynolds, 1995] using the training data is performed.

- Along with this, models for the same speaker in other conditions are synthesized using the mean shift and variance scaling. It is possible, thus, to obtain models of the same speaker corresponding to several different channel conditions.

- During testing, the most likely channel-dependent background model is used along with the synthesized speaker model (corresponding to the channel) in order to calculate the likelihood.

Feature Mapping

In the feature mapping approach, channel-dependent features are mapped into a common channel-independent feature space. As in the case of speaker model synthesis (SMS), in this approach as well, transformations and mappings are learnt for the difference in channels.

The approach of adaptation in the feature domain is attractive because it is independent of the modeling technique or structure. This approach, discussed in [Reynolds, 2003] involves the following steps:

- Data from several different channels is used in order to train a channel-independent GMM.

- This channel-independent GMM is then adapted using channel-dependent data. The transformations of the model parameters from the channel-independent to the channel-dependent model can then be used to create mapping functions for the features.

- The top one decoded Gaussian, i.e., the Gaussian mixture component with the maximum contribution to the total probability i.e., the component i where $i = argmax_{1 \leqq j \leqq M} \omega_j^{C1} p_j^{C1}$, is selected.

- Feature mapping is then performed. If x is a feature from the space modeled by the channel-dependent GMM in condition $C1$ and $i = argmax_{1 \leqq j \leqq M} \omega_j^{C1} p_j^{C1}$, where $p_j^{C1}(x) = N(\mu_i^{C1}, \sigma_i^{C1})$ is the jth mixture component of the GMM in condition C1, the feature mapping from a channel-dependent feature x to the channel-independent feature y is then given by:

$$y = M_i^{C1 \to C2}(x) = (x - \mu_i^{C1}) \cdot \frac{\sigma_i^{C2}}{\sigma_i^{C1}} + \mu_i^{C2}. \tag{2.21}$$

This feature mapping is used to transform $x \sim N(\mu_i^{C2}, \sigma_i^{C2})$ into $y \sim N(\mu_i^{C1}, \sigma_i^{C1})$ [Reynolds, 2003].

- For a given test utterance, the most likely channel-dependent model is detected, and the feature adaptation of the test utterance is done, mapping the features into a so-called channel-independent space.

- Subsequent testing is done using the speaker model created with a feature-mapped test utterance and the root channel-independent GMM as the universal background model. Once feature mapping is performed, any modeling technique can be used, and we are not bound to Gaussian mixture modeling.

The distribution scaling techniques presented in this thesis are similar to this feature mapping technique, although they are applied to the score domain, and they directly consider means and covariances of the score distributions for adaptation as opposed to the multi-Gaussian approach, choosing the most 'dominant' or principal Gaussian.

2.3.3 Score normalization techniques to handle mismatch

The distortion of the distributions of features due to mismatch results in the log-likelihood scores also correspondingly being affected. In Sec. 2.3.2 we discuss feature and model-level techniques for the compensation of mismatch. At the score level, some of the attempts to overcome handset mismatch include using score normalization techniques like the $Z-norm$, $H-norm$ and $T-norm$. Normalization techniques result in the transformation of the scores and their distributions into a score space, where distortion effects are reduced. Normalization techniques have been used in speaker verification in order to reduce particular effects of specific speakers and test recordings [Auckenthaler et al., 2000]. Early work in score normalization in speaker recognition was done by Li and Porter [1988], and several normalization methods have been proposed based on this work.

If prior information about the handset type, the language spoken, the sex of the speakers, etc. is known, this information can be used in order to enhance the performance of the verification. However, in forensic recognition, this additional information may not be easily available.

For mismatched conditions, these normalizations (or the lack of them) can have serious consequences on the evaluation of the likelihood ratio. The score distributions of both H_0 and H_1 are important for the evaluation of the LR. Scaling and normalizing the distributions of scores is a common technique in speaker verification and results in a reduction in the equal error rate (EER). In the Bayesian interpretation of automatic speaker recognition results, scaling and normalization can also serve to better the accuracy of recognition results, but they also have an effect on the estimate of the LRs.

In the following section, several normalization methods are presented. Li and Porter [1988], observed that although the true speaker distributions were relatively stable, the imposter distributions varied widely. They proposed to normalize all scores relative to the mean and standard deviation of the distribution of imposter scores.

Z-Norm

In zero normalization (or Z-norm), a model-based score normalization is performed. The model of a speaker is compared with a set of imposters and an imposter score distribution specific to this speaker is estimated [Auckenthaler et al., 2000]. The mean and the variance of the imposter score distribution is then used to normalize the scores.

$$Sc_{\text{z-norm}} = \frac{Sc - \mu_{\text{imposter}}}{\sigma_{imposter}}, \tag{2.22}$$

where μ_{imposter} and σ_{imposter} are the mean and variance for the imposter distribution, Sc_{z-norm} is the normalized score of the distribution. The imposter distribution is normalized to a mean of 0 and a standard deviation of 1. Z-norm score normalization has been used extensively for speaker verification [Bimbot et al., 2004].

T-Norm

Here, the mean (μ_N) and the standard deviation (σ_N) of the log-likelihood scores of the imposter distributions (when the two recordings do not come from the same source) are used as parameters for normalization equation Eq. 2.23. All the scores are then normalized with respect to the imposter distribution.

$$Sc_{i_{\text{T-norm}}} = \frac{Sc_i - \mu_N}{\sigma_N}. \tag{2.23}$$

This normalization is a test-centric normalization, where a given test recording is compared with a set of models (similar to the cohort-based approach). This method is used in order to avoid acoustic mismatch between the test utterances and the trained speaker models [Auckenthaler et al., 2000]. Navratil and Ramaswamy [2003] have further explored the effect of the T-norm on the DET curves. They claim that under certain assumptions, the T-norm performs a gaussianization of the individual true and imposter score populations.

H-Norm

This normalization is very similar to the z-norm technique, and is used to deal with handset mismatch between training and testing. Handset-dependent recordings from

imposters are compared to each speaker model, and the normalization parameters are calculated from this. During testing, it is necessary, that the type of handset used for the test recording is known. In this case, the parameters for score normalization can be applied.

D-Norm

In D-norm proposed by Ben et al. [2002] they present a method to generate pseudo-imposter data using the world model. They use a Monte-Carlo-based method to obtain client and imposter data using the client and world models.

The normalized score is given by

$$L_\lambda(X) = \frac{L_\lambda(X)}{KL2(\lambda, \bar{\lambda})}, \tag{2.24}$$

where $KL2(\lambda, \bar{\lambda})$ is the estimate of the symmetrized Kullback-Leibler distance between the client and the world models (distances estimated using the Monte-Carlo generated data). The advantage of using D-norm is that there is no need for additional normalization data apart from the world model.

HT-Norm

This normalization is similar to the T-norm and is based on the idea of handset-dependent normalization, i.e., H-norm. In this normalization, parameters are calculated by testing each input speech signal with handset-dependent speaker models. Thus, for each handset type, corresponding normalization parameters are calculated. During testing, it is necessary to determine which handset is used for the claimed speaker model and apply the normalization corresponding to that handset [Bimbot et al., 2004]. This means that the HT-norm is also expensive in computation, and the normalization parameters have to be determined during testing [Auckenthaler et al., 2000].

C-Norm

In order to deal with cellular data mismatch, when there are several telephones present, Reynolds introduced this normalization during the NIST 2002 to deal with cellular data which belonged to several different cellular phones and unidentified handsets [Bimbot et al., 2004]. This method was proposed in order to handle the problem of several different cellular telephones and involved the blind clustering of normalization data, followed by calculation of the means and variances of the imposter distribution (similar to H-norm).

Application of score-normalization techniques to forensic speaker recognition

We tested adaptation of the T-norm technique used in the domain of speaker verification in [Botti et al., 2004b]. The idea behind this normalization is to transform all the scores obtained by comparing a given test utterance with all the models in the potential population and with the suspect's model to a mean of 0 and a variance of 1. Application of this technique resulted greater separation between the curves corresponding to LR in H_0 cases and H_1 cases. Most of the score normalization techniques are imposter-centric i.e., the normalization parameters are calculated on the basis of the imposter score distributions. For forensic tasks, this dependence on only the imposter distribution for normalization may not be desirable.

2.3.4 Handling mismatch using auxiliary features in a Bayesian Networks approach

Bayesian networks have been used in forensic casework, in general, and have also recently been applied to the problems of speaker recognition [Arcienega and Drygajlo, 2003; Richiardi et al., 2005]. Typical forensic cases have complex frameworks, involving many variables having several dependencies among themselves. It is useful for the forensic scientist to have a logical framework in which he can understand all the dependencies which exist among the different characteristics of evidence. Bayesian networks are a method of inference where uncertainty is handled in a logically rigorous but simple way. These represent a general method of inference for scientific evidence. An inference problem such as speaker recognition, for either investigative or judicial purposes, can be broken down into smaller problems which can be solved separately and then combined to provide a solution. Bayesian networks can be used both in recognition as well as the interpretation and, thus, are of interest to forensic speaker recognition.

For the specific problem of handling mismatch in speaker recognition, Bayesian-network-based approaches have been seen to show promise. In [Arcienega et al., 2005], we have proposed a method using Bayesian networks [Jensen, 2001; Pearl, 1988] to combine prosodic features with those of the spectral envelope in order to reduce the effects of channel mismatch. This method was evaluated using both speaker verification as well as forensic speaker recognition. The Bayesian network approach presented showed its capacity to exploit the information carried by the additional features (pitch and voicing status) in order to improve the recognition scores in mismatched conditions. The pitch, carrying information about the speaker's identity, which had already been proved to be strongly robust to noise [Arcienega and Drygajlo, 2003], was also shown to be robust to channel distortions. Convolutional modifications to

the speech such as the ones introduced by PSTN or GSM channels may severely affect spectral envelope features but have almost no influence on the pitch. Incorporating these prosodic features, using a Bayesian network, has been shown to improve the performance of both speaker verification and forensic speaker recognition systems in mismatched training and testing conditions.

10 speakers of the IPSC-03 database (see Appendix C) were used to build a background model, and 20 other speakers were taken as mock suspects for performing the speaker recognition itself in three different recording conditions namely the public switched telephone network (PSTN), the Global System for Mobile Communication (GSM) and acoustic direct room recordings. A summary of the results obtained for speaker verification performance in [Arcienega et al., 2005] are presented in Tables 2.1 and 2.2.

Table 2.1: EERs when the training data is speech recorded through PSTN

Speech used for Tests	**GMM-UBM EER** [%]	**BN system EER** [%]
PSTN	4.8	3.3
GSM	42.3	31.9
Room	37.5	22.5

Table 2.2: EERs when the training data is speech recorded in the calling room.

Speech used for Tests	**GMM-UBM EER** [%]	**BN System EER** [%]
Room	1.8	1.0
GSM	22.8	18.9
PSTN	25.8	20.4

The performance can be represented using cumulative probability distribution plots such as the Tippett plots $P(LR(H_i) > LR))$ (Figs. 2.8 and 2.9).

In Fig. 2.8, we observe that when the Bayesian network (BN) system is applied to mismatched conditions (using the PSTN recording for training and the room recording for testing), there is a considerable increase in the separation between the two curves, indicating a better performance.

Similarly, in Fig. 2.9, when the Bayesian network (BN) system is applied to mismatched conditions (using the PSTN recording for training and the GSM recording for testing), there is an improvement in the performance, although it is not as marked as in the case presented in Fig. 2.8.

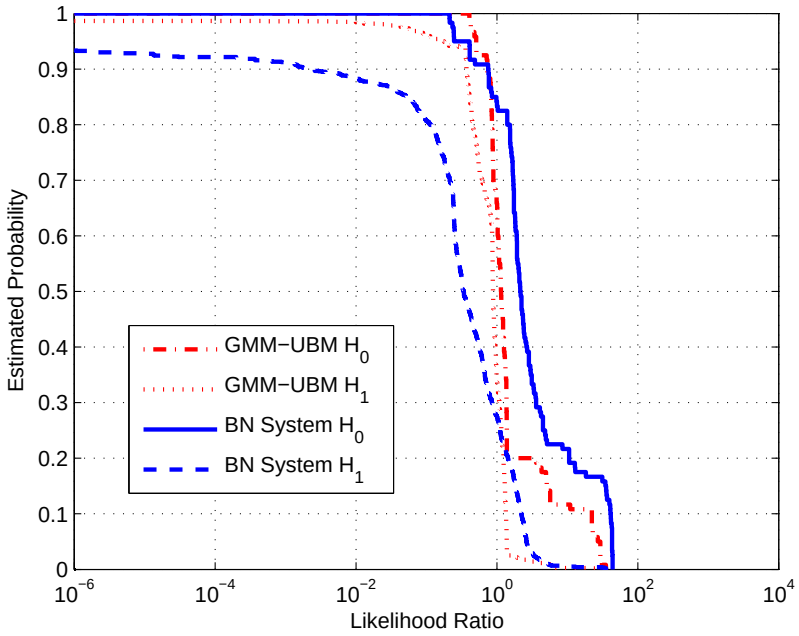


Figure 2.8: *Tippett plot PSTN-Room*

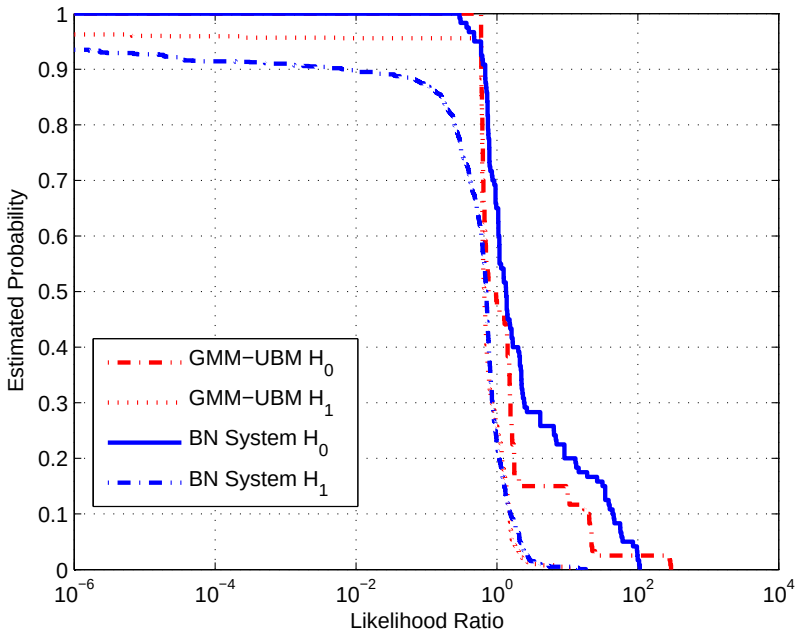


Figure 2.9: *Tippett plot PSTN-GSM*

Thus, incorporating prosodic features like the pitch, using a Bayesian network, has been shown to improve the performance of both speaker verification and forensic speaker recognition systems in mismatched training and testing conditions.

2.4 Summary

In this chapter, forensic automatic speaker recognition, using a Bayesian framework for interpretation of evidence, was presented. State-of-the-art techniques and approaches to handling mismatched conditions in automatic speaker recognition are discussed. The main points presented in this chapter include:

- Automatic speaker recognition techniques have to be adapted to the often uncontrolled and adverse forensic conditions, and results have to be interpreted in a way that would be understood and acceptable to the courts.
- The likelihood ratio (LR) is a measure of the support that the evidence E lends each of the two competing hypotheses and is a measure of the strength of the evidence.
- The prior and posterior odds are the province of the court, and only the likelihood ratio is the province of the forensic expert.
- The likelihood ratio can be estimated using a corpus-based Bayesian methodology for speaker recognition, where the within-source variability of the suspected speaker and the between-sources variability of the questioned recording within a potential population, is considered.
- In order to measure the performance of the forensic speaker recognition methods, cumulative probability distribution plots called Tippett plots which represent the proportion of the likelihood ratios greater than a given LR i.e., $P(LR(H_i) > LR)$, for cases corresponding to the hypotheses H_0 and H_1, can be used.
- In order to measure the performance of systems in the speaker verification domain, the Detection Error Tradeoff (DET) curves and Receiver Operating Characteristic (ROC) curves can be used. These curves plot match and non-match rates, and implicitly use thresholds which are undesirable for forensic comparisons.
- The acoustic features that are commonly used in automatic speaker recognition include Mel-frequency cepstral coefficients, linear-frequency cepstral coefficients, linear prediction cepstral coefficients, perceptual linear prediction

cepstral coefficients, and the modeling techniques include dynamic time warping, vector quantization, Gaussian mixture modeling, Hidden Markov Models (HMMs) and artificial neural networks. The spectrum-based features are sensitive to transmission channel distortions and noise.

- The steps involved in automatic speaker recognition include the preprocessing of the input speech file, extraction of features, modeling of features, comparison of the features and models, and the interpretation of the results. The interpretation can be a decision in the case of speaker verification systems and a likelihood ratio estimation in the case of a forensic speaker recognition system.

- Certain speakers are more difficult to recognize using automatic systems, and the varying degrees in this difficulty have led to the classification of speakers into goats, sheep, wolves and lambs.

- To reduce the effect of mismatched recording conditions, statistical compensation techniques can be applied at the level of features, models and scores. At the score level, some of the attempts to overcome handset mismatch include using score normalization techniques like the Z-norm, H-norm and T-norm.

- Incorporating prosodic features like the pitch, using a Bayesian network, has been shown to improve the performance of both speaker verification and forensic speaker recognition systems in mismatched transmission channel conditions of training and testing data.

Bayesian interpretation and the strength of evidence in forensic automatic speaker recognition

3

3.1 Introduction

In this chapter, the Bayesian interpretation methodology applied to forensic automatic speaker recognition casework is presented. We consider the requirements of this approach and present how to estimate the strength of evidence in different cases. We propose methods to deal with cases with sufficient and insufficient data to estimate the within-source variability of the voice of the suspected speaker, different methods for evaluating the strength of evidence, analysis of the variability in the strength of evidence, as well as complementary information that should be presented along with the strength of evidence to the courts.

The Bayesian interpretation methodology is gaining acceptance as a logical and coherent way of analysing forensic evidence. There has been hesitation on the part of the forensic experts to present their results in a Bayesian framework, and they have favored, a reporting style where they would try to answer only whether or not the piece of evidence in question came from the suspect. A conclusion of an analysis, typical of this style is, 'I am not/reasonably/absolutely sure that the voice on the tape is that of the suspected speaker'. This is due, in part, to the tendency of lawyers, police and, to a certain extent, judges and juries to expect the results of a forensic analysis in a binary, 'yes'/'no' kind of framework. In [Evett, 1998], Evett highlights the pitfalls of reporting evidence using this kind of approach. He goes on to define the principles of presenting forensic evidence as follows:

- Interpretation of evidence must take place within a framework of circumstances.
- To interpret the evidence, it is necessary to consider at least two propositions, i.e., it is not appropriate to speculate about the truth of any hypothesis without considering the possibility of an alternative hypothesis.
- It is necessary for the scientist to consider the probability of the evidence, given each of the stated propositions.

Bayesian interpretation provides an elegant framework for a forensic expert to present evidence to the courts. The Bayesian interpretation framework used in forensic automatic speaker recognition applies these principles to reporting the strength of evidence. The evidence is always evaluated with respect to two competing propositions, and the results are presented as a likelihood ratio.

3.2 Bayesian interpretation methodology applied to automatic speaker recognition

In the following section, a Bayesian interpretation methodology for the interpretation of automatic speaker recognition evidence is presented. Two situations that are often observed in forensic investigations, dealing with sufficient or insufficient amounts of speech data, are discussed. In certain cases, such as when the suspected speaker cooperates with the investigation and agrees to be recorded under controlled conditions by the expert, or when wire-tapping of the telephone conversations of the suspected speaker is performed by the police over a period of time, there is a sufficient amount of speech data from the suspected speaker. Often, it is also observed that in forensic investigations, the forensic expert has only a questioned recording and a single recording of a suspect. He is required, by the police or the courts, to evaluate whether the voice in both these recordings comes from the same person, in spite of having insufficient data to estimate the variability of the suspected speaker's voice. Both these situations appear often in casework, and two methodologies have been proposed to handle such cases.

3.2.1 Bayesian interpretation in cases with sufficient suspect reference data

When there is sufficient data in order to estimate the intra-variability of the suspected speaker's speech, the approach described in detail in Chap. 2.1 can be used.

The two hypotheses considered are:

- H_0 (the suspect is the source of the questioned recording.)
- H_1 (another speaker from the relevant population is the source of the questioned recording.)

The information provided by the analysis of the questioned recording (trace) leads to the choice of a reference population of relevant speakers (potential population) having voices similar to the trace. Three databases are used in this methodology, namely, the potential population database (P), the suspected speaker reference database (R) and the suspected speaker control database (C). The potential population database (P) is a database for modeling the variability of the speech of all the potential relevant sources, using the automatic speaker recognition method. It allows for evaluation of the between-sources variability, given the questioned recording, i.e., the distribution of the similarity scores that can be obtained, when the questioned recording is compared to the speaker models (GMMs) of the potential population database [Drygajlo et al., 2003].

Since the recording conditions of the trace and the P database are often different, it is necessary to record two different databases with the suspected speaker. The suspected speaker reference database (R) is recorded with the suspected speaker to model his/her speech with the automatic speaker recognition method. In this case, speech utterances should be produced in the same way as those of the P database. The suspected speaker model obtained is used to calculate the value of the evidence by comparing the questioned recording to the model. The suspected speaker control database (C) is recorded, with the suspected speaker to evaluate her/his within-source variability when the utterances of this database are compared to the suspected speaker model (GMM). The recordings of the C database should be similar to the trace in their technical recording conditions and linguistic content.

3.2.2 Bayesian interpretation in cases with insufficient suspect reference data

In many cases only one recording of the suspect is available due to the nature of the investigation, e.g. when it is not possible to have additional recordings of the suspect's voice, since it may alert him to the fact that he is being investigated. As a consequence, it is not always possible to evaluate the within-source variability of the suspect with this single recording. However, as this is a recurring problem in forensic speaker recognition, it is necessary to define an interpretation framework for evaluating the evidence in the absence of additional control recordings (presented in [Botti et al., 2004a]).

When it is not possible to obtain additional suspect recordings, we cannot evaluate the within-source variability of the suspect as only one suspect utterance is available.

In this situation, the methodology proposed for forensic automatic speaker recognition, based on a Bayesian approach [Meuwly, 2001; Drygajlo et al., 2003] (described in detail in Chapter 2), cannot be applied.

We propose to handle this situation, by again using a corpus-based Bayesian approach, with a set of hypotheses different from the ones in the approach with sufficient data to estimate the within-source variability of the suspected speaker.

The two hypotheses considered are :

- H_0 - the two recordings compared have the same source.
- H_1 - the two recordings compared have different sources.

There is a subtle difference in the two sets of hypotheses. By considering the scores obtained in similar conditions when H_0 or H_1 is true, it is possible to evaluate the evidence score (E). This evaluation attempts to answer the question of how often would we observe a similarity score equal to E, when we compare two recordings that do indeed come from the same source, and when we compare two recordings that do indeed come from different sources. By similar conditions, we mean that although the speech content may be different, the effect of channel distortions, noise and recording conditions should be similar across the recordings.

Let us use the following notation: suspected speaker recording SR, questioned recording QR, and the evidence score E (obtained by comparing SR and QR). In order to create comparisons in similar cases, two databases are required, namely the SDB, or the Speakers database, and the TDB, namely the Traces database. The SDB should contain mock suspects in recording conditions similar to that of the suspected speaker. Similarly, the TDB should contain mock traces in the recording conditions of the questioned recording (QR). For instance, if the trace corresponds to a recording of a cellular telephone and the recording of the suspect comes from a fixed telephone, the databases required would be a mock trace database of cellular recordings (TDB) and a reference database of corresponding speakers using a fixed telephone (SDB).

From these two databases, it is possible to construct two sets of mock cases, with a speaker model chosen from the SDB and a questioned recording chosen from the TDB as follows:

1. Cases where two recordings coming from the same source are compared (H_0 cases).
2. Cases where two recordings coming from different sources are compared (H_1 cases).

These comparisons are illustrated in Fig. 3.1

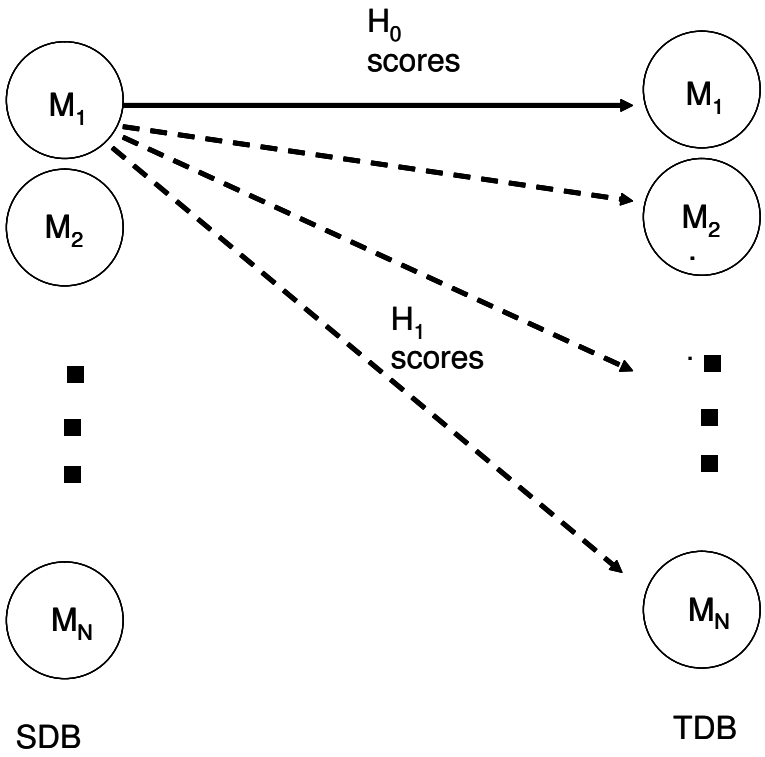


Figure 3.1: *H_0 and H_1 scores obtained comparing each recording from the speakers database (SDB) and the traces database (TDB)*

For each mock case, we obtain the similarity score. The probability distributions of all the scores of H_0 cases and of H_1 cases are then plotted. The H_0 curve represents the distribution of scores that we can expect when a trace belonging to a speaker, within the conditions of the case, is compared to the speaker. The H_1 curve represents the distribution of scores that we can expect when a trace that does not belong to a speaker, within the conditions of the case, is compared to the speaker.

Then, the expert has to compare the *real* trace from the case with the suspect's model, and obtains a score for E.

Note that in this methodology, it is assumed that the TDB is recorded in similar conditions to the questioned recording QR, and the SDB is recorded in conditions similar to that of the suspect recording SR. It can be observed that in similar conditions, when two recordings coming from the same source are compared, scores obtained (H_0 scores) are in the same higher range. Similarly, scores obtained comparing recordings of different speakers (H_1 scores), are in the same range, on an average, lower than scores obtained in H_0 cases. This assumption implies that we consider similar within-source variability for different speakers, which is not entirely correct because the voices of different speakers may vary differently. Although differences can ex-

ist between speakers, scores obtained by comparing speakers' models with their own voices (H_0 cases) are in a similar range that is distinct from the scores obtained by comparing their voices with someone else's (H_1 cases). We consider differences between speakers' within-source variability as not significant compared to differences between scores of H_0 cases and scores of H_1 cases. While such differences are, for the most part, negligible, especially in well-matched conditions, this assumption heavily depends on the degree of mismatch between training and test recording conditions.

Additionally, in order to verify the assumption of compatibility between recordings of the case and the mock databases, it is possible to calculate scores obtained by testing the models of speakers of the mock database SDB against the QR and against each recording of TDB. With these two sets of comparisons, we have two sets of values for H_1. Statistical significance testing can be applied to see whether these distributions are compatible [Alexander et al., 2004]. Similarly, in order to evaluate the compatibility between SDB and SR, we can compare, statistically, the scores obtained by testing the recordings of TDB against SR and SDB. Statistical significance testing can be applied to these two sets of H_1 values to see whether these distributions are compatible. If distribution scores show incompatibility between the recordings, then the forensic expert has to decide either to select or record more compatible SDB and TDB databases, or decide not to do the case using this methodology, or apply statistical compensation techniques.

3.3 Multivariate and univariate approaches to estimating the strength of evidence

In automatic speaker recognition, the features that are considered representative of the speaker's identity are modeled using statistical techniques, and the likelihood of observing these features is estimated. The estimation of the likelihood ratio, as explained in the Sec. 3.2.1 and 3.2.2, can be done at two distinct levels; at the level of the features such as MFCCs, RASTA-PLP, etc. that are modeled and at the level of the scores pertaining to each hypothesis. At the lower level of the feature vectors, it is typical to observe multivariate feature vectors that are extracted, per analysis window, from the audio signal. In text-independent speaker recognition, the ensemble of these feature vectors is considered representative of the voice of the speaker, provided there is a sufficient amount of audio data that covers the range of the speaker's speech. The probability of observing the features of a recording in the statistical model of the speaker is represented as the likelihood score. For computational convenience, these scores are represented in the log domain and are called log-likelihood scores. For the estimation of the likelihood ratio, with respect to

the two competing hypotheses, statistical models of the distributions of these scores are calculated. These log-likelihood scores are univariate values, and the distributions of these scores represent each of the hypotheses. Thus a two tier calculation of likelihoods is done, first at the feature level, and the second at the level of the score distributions. This representation is convenient as the univariate representation of scores can easily be visualized. However, it is also possible to use a direct approach to calculating the strength of evidence instead of this two-tier calculation, evaluating the likelihood ratio directly from the distributions of the features of the voices.

Thus, the calculation of probability densities of the speech features can be dealt with at two levels; one, at the multidimensional feature space, where the likelihoods of the multivariate feature vectors are estimated, as well as at the univariate level, where the likelihood scores (derived in the multivariate level) for different hypotheses are modeled. It is thus possible to define the likelihood ratio for these two levels.

The first method takes the likelihood values returned by the system and directly uses them to evaluate the likelihood ratio. A similar likelihood ratio estimation, using phonetic-acoustic features, is presented in [Rose, 2003, :4111] and [Rose, 2002]. The second method determines the likelihood ratio using the probability distributions of these likelihood scores. The first method is called the *scoring method*, and the second one, the *direct method*. These two approaches discussed below are illustrated in Fig. 3.2.

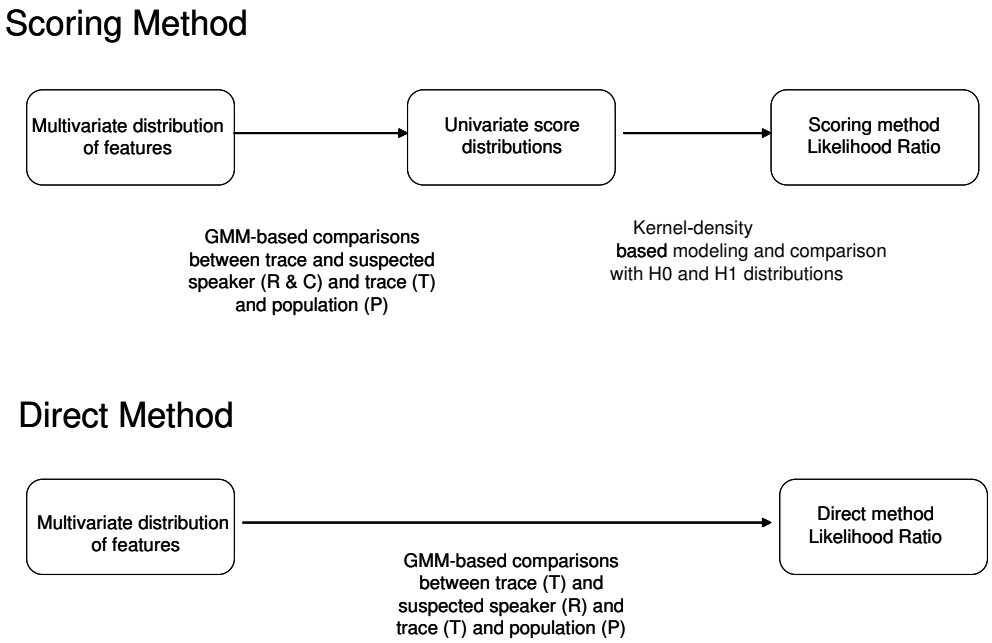


Figure 3.2: *Scoring and direct methods for estimating the strength of evidence*

The scoring approach is data-driven and requires a considerable amount of sus-

pected speaker data, in terms of the suspected speaker reference and control database.

3.3.1 Direct method

In the direct method, the likelihood ratio is defined as the relative probability of observing the features of the trace in a probability distribution of the features of the suspect and of observing the same features in the probability distribution model of any other speaker from a potential population. The direct method, in the Bayesian methodology, requires, in addition to the trace, the use of two databases: the suspect reference database (R) and the potential population database (P).

The calculation of the likelihood ratio for a given trace in the direct method is as follows:

- The features of the trace are compared with the statistical models of the suspect (created using database R), to obtain the evidence value (E).
- The trace is compared with statistical models of all the speakers in the potential population (P). Considering the log-likelihood score as the log of the likelihood that the trace came from the statistical model of the speaker, we calculate the likelihood that the trace could have come from any speaker of the potential population.

Mathematically, the LR in the direct method is the ratio of the average (geometric mean) likelihood of the features in the trace appearing in the statistical models of the features of the suspect and the average likelihood of the features of the trace appearing in the statistical models of the features of the speakers in the potential population.

$$LR_{direct} = \frac{\sqrt[N_R]{\prod_{i=1}^{N_R} p(X_T|\lambda_{R_i})}}{\sqrt[N_P]{\prod_{j=1}^{N_P} p(X_T|\lambda_{P_j})}}, \tag{3.1}$$

where

- X_T are the features of the questioned recording (T)
- λ_{R_i} is the ith statistical model of the suspect created from the R database which contains N_R recordings, and λ_{P_j} is the model of the jth speaker of the P database (containing a total of N_P speakers).

We illustrate the likelihood ratio, using the case example presented in Chap. 2, Fig. 2.2, where the suspected speaker was indeed the source of the trace (H_0 case). A likelihood ratio of 1155 is estimated using the direct method.

The direct method is especially useful when the recordings available from the suspected speaker are not very long, and insufficient to extract both suspect reference and control databases.

3.3.2 Scoring method

In the scoring method, the LR is defined as the relative probability of observing a score E in the distribution of scores that represent the variability of the suspect's speech and the distribution of scores that represent the variability of the potential population speech with respect to the questioned recording (trace).

Mathematically,

$$\begin{aligned} LR_{scoring} &= \frac{p(E|H_0)}{p(E|H_1)} \\ &= \frac{p(L(X_T|\lambda_{R_i})|H_0)}{p(L(X_T|\lambda_{R_i})|H_1)} \\ &for \quad i = 1, ..., N_R, \end{aligned} \tag{3.2}$$

where

- N_R, N_C and N_P (Eq. 3.3) are the number of recordings in the suspect reference database (R), the suspect control database (C) and the potential population database (P) respectively,
- X_T are the features of the questioned recording (T), X_{C_i} are the features of the ith recording in the control database (C),
- λ_{R_i} is the ith statistical model of the suspect created from the R database which contains N_R recordings, and λ_{P_j} is the model of the jth speaker of the potential population (P) database containing a total of N_P speakers),
- $L(X|\lambda)$ is the likelihood of observing distribution of features X given a statistical model λ.

Further, hypotheses H_0 and H_1 in Eq. 3.3, are modeled using probability density functions:

$$\begin{aligned} H_0 &= P(L(X_{C_i}|\lambda_{R_j})) \\ &for \quad i = 1, ..., N_C \\ &for \quad j = 1, ..., N_R, \end{aligned} \tag{3.3}$$

$$\begin{aligned} H_1 &= P(L(X_T|\lambda_{P_i})) \\ &for \quad i = 1, ..., N_P. \end{aligned} \tag{3.4}$$

Note: p represents a likelihood and P represents a probability distribution

We illustrate the likelihood ratio using the same case example considered in the direct method (presented in Chap. 2, Fig. 2.2), where the suspected speaker was indeed the source of the trace (H_0 case). A likelihood ratio of 9.165 is estimated using the scoring method.

In this approach, the log-likelihood scores returned by the speaker recognition system are used as indices in a Bayesian interpretation framework. The log-likelihood scores are a measure of the similarity of the features of the recording and the features used to create the statistical model. Thus, the meaning of this *score* is more than that of an index and implicitly includes a measure of the extent of similarity.

This method of estimating likelihoods and likelihood ratios is similar to the kind used in the forensic analysis of glass, where the refractive index is used in order to calculate the likelihood ratio. In the case of glass analysis, for instance, the refractive index, is a property of the glass and is different from a *score*. This index does not refer to the similarity or differences between different glasses as a score would.

Comparison of the Direct and Scoring Methods

The scoring method is a general basis of interpreting the strength of evidence, and it is used in the interpretation of evidence in several types of forensic analysis. The direct method, however, can be applied only in cases where the results of the analysis are likelihoods.

Both methods can be affected by mismatched recording conditions of the databases involved. In the direct method, compensation of the mismatch can be attempted either in the acoustic feature space or in the statistical modeling of the features. In the scoring method, statistical compensation of mismatch can be applied to the scores using databases in different conditions with which the extent of mismatch can be estimated [Alexander et al., 2004]. The direct method does not require the use of all the three databases (P, R, C) used in the scoring method, and relies only on P and R. It is also less computationally intensive than the scoring method, and requires less suspected speaker speech data. We have observed in our experiments that, the range of the values of likelihood ratio has less variation in the scoring method than in the direct method, and a much higher LR is obtained for the direct method than for the scoring method.

In forensic analysis, a likelihood ratio of 1 important, as it implies the point at which neither the hypothesis (H_0) nor the hypothesis (H_1) can be supported more than the other. In the direct method, a likelihood ratio equal to 1, implies

$$p(X|\lambda_{R_i}) \quad = p(X|\lambda_{P_j}) \quad \forall i, j \tag{3.5}$$

and in the scoring method, this implies

$$\begin{aligned} p(L(X|\lambda_{R_i})|H_0) &= p(L(X|\lambda_{R_i})|H_1) \\ for &\quad i = 1, ..., N_R \end{aligned} \tag{3.6}$$

We can see, in the direct method, that a likelihood ratio of 1 is obtained when the statistical models of the suspect and the potential population represent similar voices. An LR of 1 in the scoring method will imply that the score E obtained by comparing the features of the trace with the model of the suspect, is equally probable in each of the distributions of scores corresponding to the hypotheses H_0 and H_1.

Both these methods were tested with several simulated forensic cases in order to analyse and compare the strength of evidence. In order to test the methods, 15 male speakers were chosen from the IPSC-01 Polyphone database. For each of these speakers, four traces of duration 12-15 seconds were selected. The R database was created using seven recordings, each of 2-3 minutes duration, for each suspect. The C databases were created using 32 recordings of 10-15 seconds for each of them. Using this test database, it was possible to create 60 mock cases when the hypothesis H_0 is true and 60 mock cases when the hypothesis H_1 is true. For each case, the P database was a subset of 100 speakers of the Swiss-French Polyphone database. The GMM-based automatic system used 32 Gaussian pdfs to model a speaker.

The experimental results can be represented using the probability distribution plots such as the probability density functions $P(LR(H_i) = LR)$ (Fig. 3.3) and Tippett plots $P(LR(H_i) > LR))$ (Figs. 3.5 and 3.6). The Tippett plots are used to indicate to the court how strongly a given likelihood ratio can represent either of the hypotheses H_0 or H_1.

The significance probability, which may be thought of as providing a measure of compatibility of data with a hypothesis, may be also considered in order to evaluate the strength of evidence. The following z test [Aitken, 1997] can be used:

$$Z_{H_0} = \frac{LR_E - \mu_{LR_{H_0}}}{\sigma_{LR_{H_0}}} \tag{3.7}$$

$$, Z_{H_1} = \frac{LR_E - \mu_{LR_{H_1}}}{\sigma_{LR_{H_1}}}, \tag{3.8}$$

where $\mu_{LR_{H_0}}$, $\mu_{LR_{H_1}}$, $\sigma_{LR_{H_0}}$ and $\sigma_{LR_{H_1}}$ are the means and standard deviations corresponding to H_0 true and H_1 true distributions, and LR_E is the likelihood ratio of the case considered.

The probability P, which is derived from the z value, is the probability of observing the likelihood ratio obtained or any value higher, in cases where the suspect is the source of the trace. These probabilities are similar to the probabilities represented

on the Tippett plot for LR_E. The z test calculates the significance of LR_E under the assumption of normality of the scores, while the Tippett plot directly represents the significance of LR_E without making this assumption.

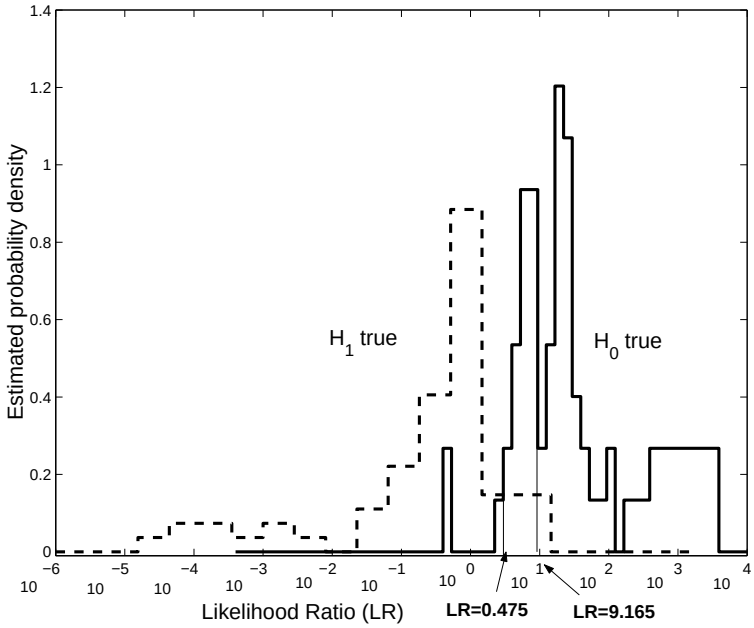


Figure 3.3: *Probability density plot of LRs (scoring method)*

For the case shown in Fig. 2.2, the likelihood ratios for E, using the direct method and the scoring method, are 1155 and 9.165 respectively. In Figs. 3.3 and 3.4, the probability density functions of the scoring method and direct method LRs are presented. The points 0.475 and 209, in Figs. 3.3 and 3.4, represent the intersection of the probability distribution functions of likelihood ratios corresponding to H_1 true and H_0 true cases. Z_{H_1} values (estimated in the log domain) for the scoring and the direct method are 1.1624 and 1.4716 respectively, which correspond to 5.4% and 7.08% probabilities of observing these likelihood ratios, or greater in cases where the hypothesis H_1 is true. Z_{H_0} values (estimated in the log domain), for the scoring and the direct method, are -0.3110 and -0.6995 respectively, which correspond to 62.2% and 75% probabilities of observing these likelihood ratios or greater in cases where the hypothesis H_0 is true.

These results indicate that for the example case considered, both the direct method and the scoring method estimated LRs that are typically observed in cases where the suspect is indeed the source of the trace. The direct method has a lower proportion of cases exceeding this LR, for the hypothesis H_1 and H_0, than the scoring method.

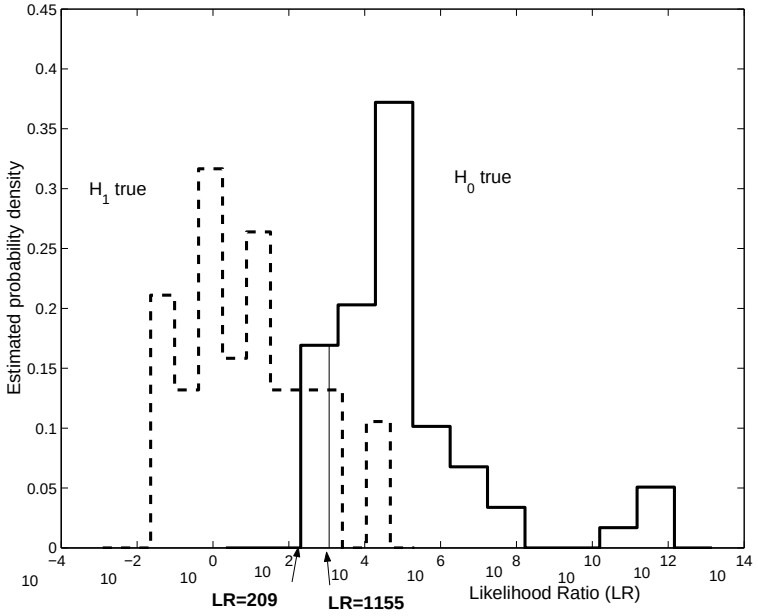


Figure 3.4: *Probability density plot of LRs (direct method)*

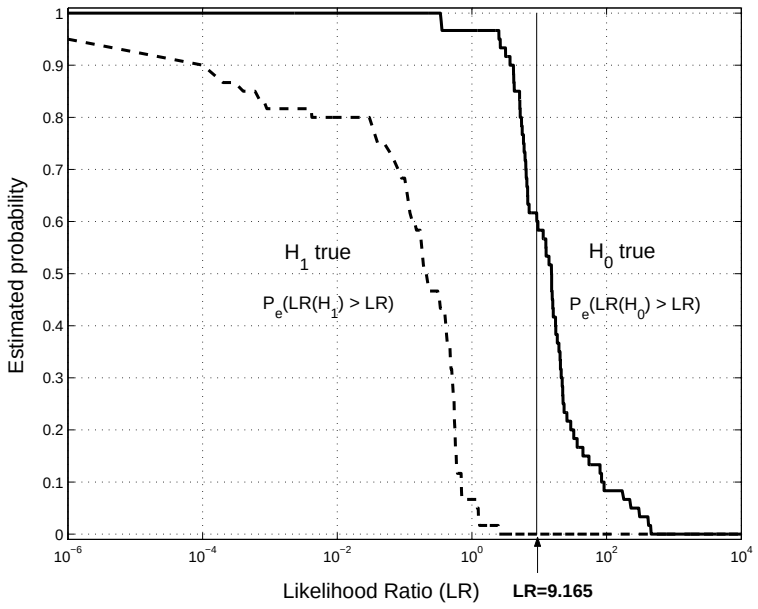


Figure 3.5: *Tippett Plot (scoring method)*

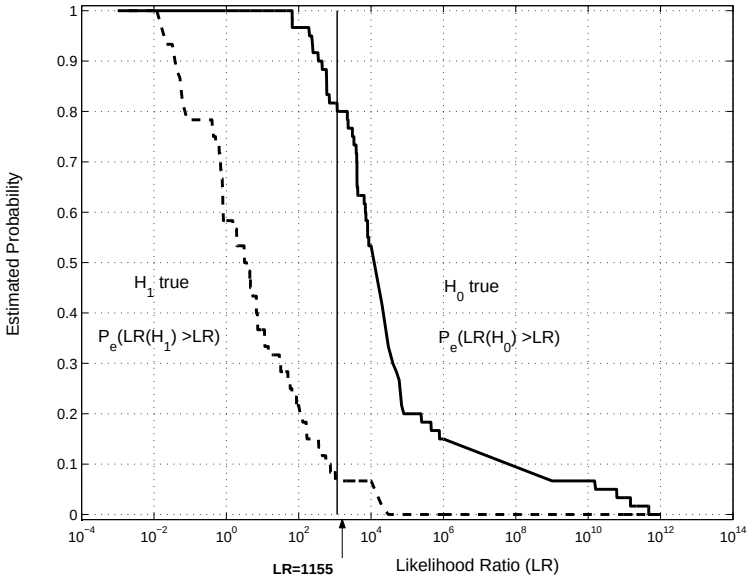


Figure 3.6: *Tippett Plot (direct method)*

3.4 Variability estimation of the strength of evidence

In this section we consider another issue in the estimation of the strength of evidence, which is the extent of variability within an LR obtained. While the LR may support one hypothesis over the other, it is important to consider how much variation there could be within a given estimate of the LR. If for instance the LR shows support for one hypothesis, but if the variability in the LR is considerable, it may even be possible that this LR would support the competing hypothesis as well.

In forensic automatic speaker recognition, the likelihood ratio is influenced by the recording conditions of the case (transmission channel, noise, and recording devices), the choice of the databases used (the extent of mismatched recording conditions between the databases)[Alexander et al., 2004], the mathematical modeling (used to model the score distributions) and the duration of the recordings available for the case.

The influence of these factors contributes to the uncertainty in the estimation of the likelihood ratio. For the evidence score E obtained in a case, the expert has to calculate the extent of this uncertainty. He must be able to present his assessment of the strength of evidence to the court, taking into consideration and incorporating this uncertainty.

Let us consider the two example cases of the estimation of the likelihood ratio

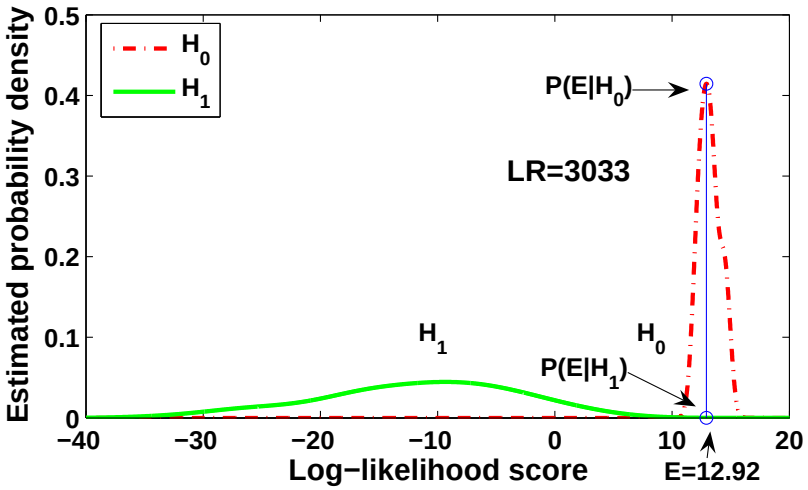


Figure 3.7: *Case 1: Illustration of the estimation of LR*

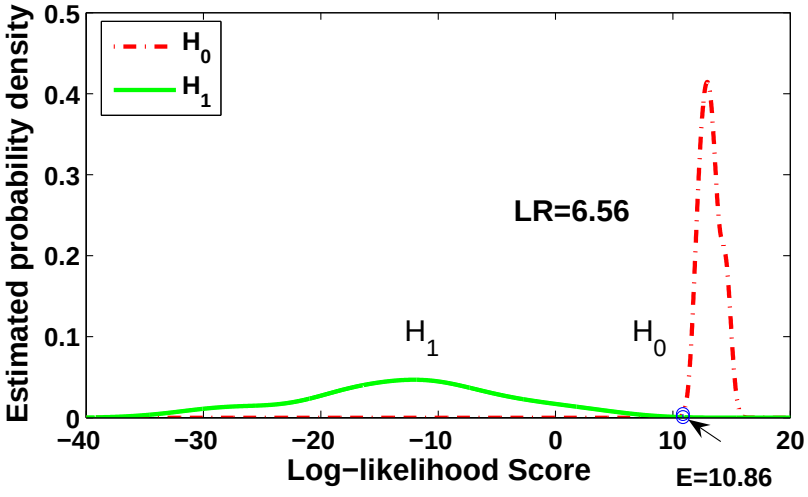


Figure 3.8: *Case 2: Likelihood ratio estimated on the tails of the score distributions of H_0 and H_1*

presented in Figs. 3.7 and 3.8. In these figures, the score (E), represented by a point on the log-likelihood score axis is obtained by comparing the questioned recording with the Gaussian mixture models of the suspect's speech. The LR is evaluated by taking the ratio of the heights of the score distribution H_0 ($p(E|H_0)$) and the score distribution H_1 ($p(E|H_0)$) at the point E. In these cases, kernel density estimation has been used to obtain the probability density functions for the two score distributions [Aitken and Taroni, 2004]. In *Case 2*, (Fig.3.8), at the point E on the log-likelihood axis, we observe that the corresponding probability densities, $p(E|H_0)$ and $p(E|H_1)$, are both very small. This implies that at this point, scores corresponding to each of the hypotheses are less likely to be observed. The estimates of the probability densities $p(E|H_0)$ and $p(E|H_1)$ depend greatly on the mathematical modeling of each the two score distributions.

In addition, although the LR of 6.56 lends support to the hypothesis H_0, we observe that the point E, at which the likelihood ratio has been evaluated, corresponds to the tails of the distributions of H_0 and H_1 scores. If we check whether it is likely that this evidence score E could indeed have come from the distribution of scores corresponding to H_0, using statistical significance testing (with a significance level of 5%), we find that this score is not significant. As a result, we have an inconsistency that the likelihood ratio 6.56, at E, supports the hypothesis H_0, but E is not significant with respect to the H_0 score distribution.

Thus, there can be two situations where the expert has to deal with sparse data, within the context of a given case, for the calculation of the LR.

- The likelihood ratio lends support to one of the hypotheses H_0 or H_1, for a certain evidence score (E), but this score is located on the outlier of the distribution of scores for this hypothesis and is not relevant at a certain significance level.

- The likelihood ratio calculated for a certain evidence score (E) corresponds to the ratio of extremely small probability densities with respect to each or both of the score distributions (H_0 and H_1). The sparsity of data in these regions makes the LR very sensitive to the modeling of the probability densities, and a small change in either of the score distributions (H_0 or H_1) or the evidence score (E) could result in a big change in the LR.

3.5 Subsets-bootstrapping of the strength of evidence

In order to estimate the uncertainty of the likelihood ratio, particularly for regions on the log-likelihood score axis where the distribution of scores is sparse and $p(E|H_0)$ and

$p(E|H_1)$) are very small, it is possible to use a bootstrapping technique to calculate the variations in the likelihood ratio estimates. These regions of low probability density of scores ($p(E|H_0)$ and $p(E|H_1)$) suffer from high variations in the likelihood ratio estimates for small changes in the evidence score (E) or the distributions of H_0 and H_1 scores. For these regions, the expert will not be able to estimate a reliable likelihood ratio. The expert then has the option of disregarding these zones or avoiding a precise likelihood ratio estimate and providing an interval of possible LR values.

Bootstrapping techniques have been proposed in [Bolle et al., 2004], for biometric systems, where the match-score accuracy is expressed as the receiver operating characteristic (ROC) curve. They have proposed that confidence intervals or margins of error should be provided for this curve, to determine whether the differences in accuracy between systems are actually statistically significant. In forensic speaker recognition, the ROC curve is only used to compare performances between different systems and cannot be used for the interpretation of evidence. The subsets bootstrapping technique for the estimation of the likelihood ratio can be performed as follows:

- From each of the H_0 and H_1 score distributions, subsets are chosen using 'leave one out' (i.e., removal of one element per run, with replacement).
- From each of these subsets, the probability distributions for H_0 and H_1 are recalculated, and a new estimate for the likelihood ratio (LR) is calculated for the evidence E.
- Confidence intervals for the bootstrapped LRs are calculated using all subsets, and this range is presented as the LR for the evidence E.

For all the bootstrapped scores of E, confidence intervals at 95% significance are obtained. That is, it can be said, with 95% of certainty, that the true value of the likelihood ratio lies within the computed confidence interval. The 95% bootstrap confidence interval at the point E is computed by generating several different bootstrap estimates for the LR at E. In Fig. 3.9, we observe the variation of this likelihood due to the bootstrapping procedure for a given evidence score (E).

$$LR_E \in [\mu_{LR_b} - 1.96\sigma_{LR_b(E)}, \mu_{LR_b(E)} + 1.96\sigma_{LR_b(E)}] \tag{3.9}$$

where LR_E is the range of the new bootstrapped likelihood ratio at the point E, $\mu_{LR_b(E)}$ and $\sigma_{LR_b(E)}$ are the means and standard deviations of the subset bootstrapped LRs calculated at E. As likelihood ratios cannot extend below 0, the lower limit of confidence interval is constrained by 0. Note that the interval calculated as 1.96 times the standard deviation from the mean corresponds to a 95% confidence interval.

3.5.1 Verbal equivalents for the likelihood ratio intervals

Verbal equivalents serve to explain the likelihood ratio to judges or juries who may find the numerical likelihood scale difficult to grasp. An equivalent verbal expression scale of the likelihood ratio has been proposed in [Evett, 1998] for DNA analysis. Although this verbal scale was proposed in the context of DNA analysis, this scale is also used in other fields of forensic analysis to express the strength of evidence and has also been applied to expressing the results in forensic speaker recognition. This scale, which is based on intervals, is presented in Table 3.5.1.

Likelihood Ratio	**Verbal expression**
1 to 10	Limited support for H_0
10 to 100	Moderate support for H_0
100 to 1000	Strong support for H_0
1000 and above	Very strong support for H_0
Below 1	Corresponding support for the competing hypothesis H_1

Table 3.1: Likelihood ratios and their verbal equivalents

There are some difficulties with this scale, as discussed in [Champod and Evett, 2000], such as there is a need to explain the meaning of this scale to the jury and the

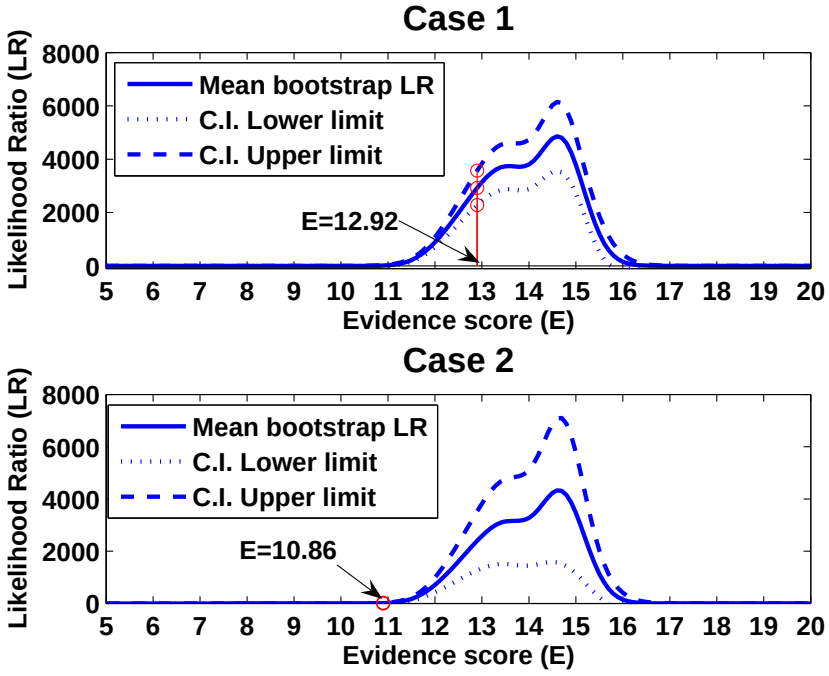


Figure 3.9: *Bootstrapped likelihood ratio for the two example cases*

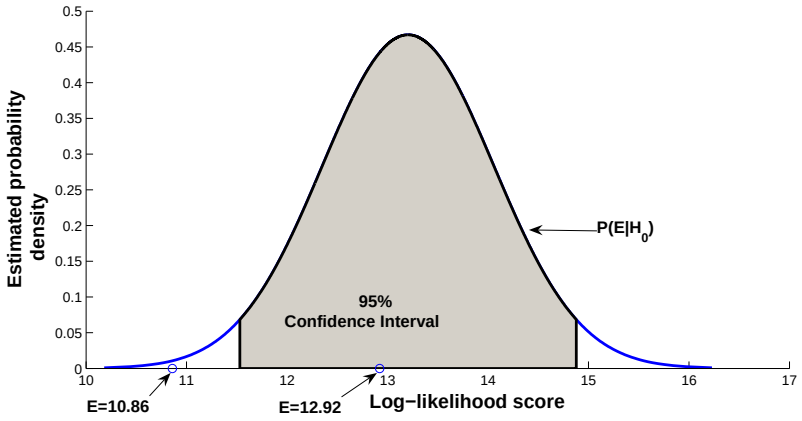


Figure 3.10: *Confidence intervals for the H_0 score distribution*

court, this scale is stepwise while the nature of likelihood ratios is continuous, and this scale does not distinguish between LRs above 10,000 or very small LRs. They suggest that forensic scientists associate either the numerical expression of the LRs or the intervals of the LRs with the verbal scale. Associating both the numerical LRs as well as their confidence intervals to the verbal scale will allow the expert to explain the strength of the evidence and the uncertainty in its estimation.

The confidence intervals (discussed in the previous section), for the LR can help determine in what range of the verbal equivalents it is in, for example if the confidence interval for a given likelihood ratio of 120 is between 60 and 180, then the verbal equivalent would be 'between limited and moderate support for the hypothesis H_0.

3.6 Significance testing of the evidence

Another means of estimating the reliability of an estimated likelihood ratio is to use classical significance testing for each of the hypotheses. If the likelihood ratio lends more support to a certain hypothesis than the other, this implies that it is more likely to observe a certain score of evidence given this hypothesis than the competing one. The significance probability, which may be thought of as providing a measure of compatibility of data with a hypothesis, may also be considered in order to evaluate the strength of evidence [Aitken and Taroni, 2004]. Both the likelihood ratio and the significance probabilities can be expected to show similar trends for a given evidence score (E), when calculated in significant regions of the score distributions. In general, if the likelihood ratio supports the hypothesis H_0, the score (E) must be significant with respect to the distribution of scores corresponding to H_0.

In the example *Case 1* (Fig. 3.7), the likelihood ratio estimated (3033) clearly

supports the hypothesis H_0, as the evidence score $E = 12.92$ is in the middle of the H_0 score distribution and in tails of the H_1 score distribution. In the example *Case 2* (Fig. 3.8), although the likelihood ratio estimated (6.56) is greater than 1 and supports the hypothesis H_0, the evidence score $E = 10.86$ is on the tails of both the distributions of H_0 and H_1.

If the likelihood ratio supports the hypothesis H_0, but the significance probability of a score E in the distribution scores pertaining to the hypothesis H_0 is small, this is an indication that the likelihood ratio estimate may be inaccurate. The evaluation of the significance probabilities should follow the estimation of the likelihood ratio, in order to determine whether a likelihood ratio is reliable.

In Fig. 3.10, we compare the evidence scores, both of which lead to LRs supporting the H_0 hypothesis, using the two-tailed *z-test* for statistical significance, with the null hypothesis that *the score E comes from the score distribution of H_0*. The upper and lower limits for the confidence interval of 95% for H_0 score distribution are 11.50 and 14.87 respectively. The significance probability of the score $E = 12.92$, using the two tailed z-test, is 0.74 (we accept the null hypothesis), and for $E = 10.86$, it is 0.006 (we reject the null hypothesis). Thus, in the first case ($E = 12.92$), both the likelihood ratio and the statistical significance analysis allow us to progress the case in the direction of support for the hypothesis that the suspected speaker is indeed the source of the questioned recording, but in the second case ($E = 10.86$), in spite of a likelihood ratio greater than 1, *the statistical significance of the evidence does not allow us to progress the case in either direction.*

3.6.1 Possible likelihood ratio outcomes in a case

It is useful for both the forensic expert and the court, to know what the range of likelihood ratio values could be for a given case. It is possible to calculate the likelihood ratio, along with the bootstrapped confidence intervals, for the entire range of possible values of E and trace the evolution of the range of possible likelihood ratio values for a given case. This evolution of the LR, along with the lower and upper confidence intervals, is presented in Fig. 3.9. In a case where the two score distributions overlap considerably, it follows that we will not obtain high likelihood ratios in support of either hypothesis. However, if there is good separation between the two distributions, high likelihood ratios can be expected.

If the maximum possible likelihood ratio offers, for instance, only limited support for the hypothesis H_0, it is possible that the expert and the courts decide that there is insufficient evidentiary value to continue with the analysis. Also, in cases where, due to the specific conditions of the analysis such as mismatched recording conditions and the accuracy of the automatic recognition system, the range of possible likelihood

ratios is limited, it is necessary to present this information in addition to the likelihood ratio.

3.7 Complementary measures to the strength of evidence

While the LR provides an estimate of the strength of the evidence with respect to the two hypotheses, it does not consider the risk of errors. In a score based system (such as the GMM-based automatic recognition system), a higher score implies a greater similarity between two samples. When we consider the evolution of the LR with respect to the evidence score E, we observe that it rises to a maximum and then decreases (Fig. 3.11). The LR decreases even when evidence score is very high (which implies that the questioned recording and the suspected speaker's voice were indeed very similar). If we are interested in the relative risk of making an error with respect to choosing a certain hypothesis on the basis of the likelihood ratio, it is necessary to consider an additional measure. For instance, if a certain likelihood ratio (say 10) is obtained for an evidence score E, it is of interest to know what the risk of errors would be in 'choosing' the H_0 hypothesis to be true if E was used as a decision threshold, i.e., what would be the risk of making errors if we had a system that would accept comparisons obtaining a similarity score of greater than E as coming from the speaker, while rejecting all scores that obtain a similarity score of less than E. Note that, 'choosing' a certain hypothesis to be true departs from the Bayesian interpretation approach where the expert should evaluate only the probability of the evidence given the hypothesis. The error ratio provides a link between automatic speaker recognition techniques where the match and non-match rates for a given threshold are considered.

The *error ratio* (ER) (presented in [Botti et al., 2004a]) is the proportion of cases for which recordings from the same source were wrongly considered to come from different sources, divided by the proportion of cases in which recordings from different sources were wrongly considered to be from the same source, if E is used as a threshold in a hypothesis test for match or non-match.

For example, for a given value of E, an ER of 10 means that the risk of making an error by excluding a suspect is 10 times higher than the risk of identifying him as the source, if the score of E is used to decide.

$$ER = \frac{P(NonMatch|H_0, E)}{P(Match|H_1, E)} = \frac{FNMR_E}{FMR_E}. \tag{3.10}$$

If the score of the trace, given the suspect model, is E, then the error ratio (ER) is:

$$ER = \frac{\int_{-\infty}^{E} p(x|H_0)dx}{\int_{E}^{\infty} p(x|H_1)dx}. \quad (3.11)$$

Note that although the DET (Detection Error Tradeoff) curve [Martin et al., 1997] shows the relative evolution of False Match with False Non-Match probabilities, it cannot be used to estimate the extent of relative risk of error *for a given E*, since the score does not appear explicitly in the plot, and hence, DET is useful only for the comparison and evaluation of system performance [Nakasone and Beck, 2001]. In the DET curve, it can be remarked that *False Alarm probability* corresponds to the False Match Rate and the *Miss probability* corresponds to the False-Non Match Rate.

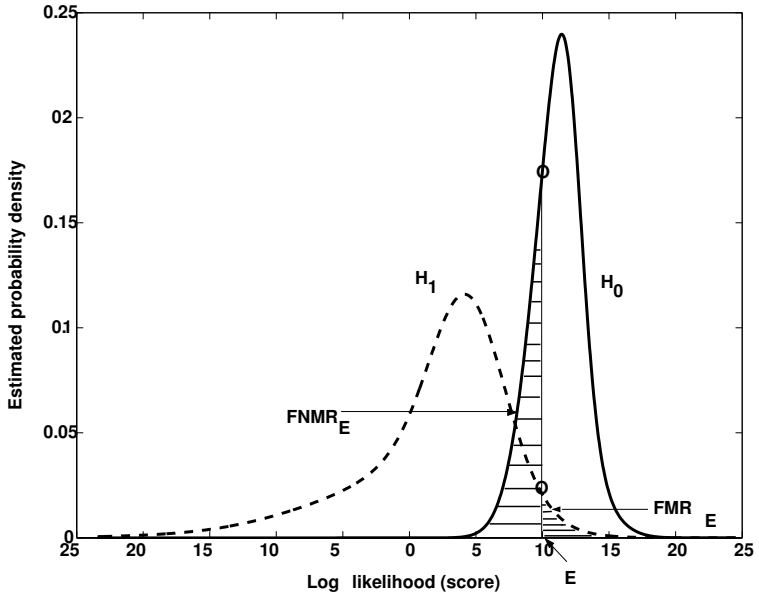


Figure 3.11: *The areas corresponding to the numerator and denominator of the ER, (the $FNMR_E$ and FMR_E*

Let us examine how the ER is calculated and what it means. The numerator of the ER corresponds to the area under the H_0 distribution curve below E. This refers to the percentage of comparisons with a score smaller than the score E in which the source of the two compared recordings is the same. With the increase of the score of E (moving along the match score axis), the higher this percentage is, the lower the risk is in deciding in favor of the hypothesis H_0. In fact, if there are more comparisons between pairs of recordings (having the same source) giving a smaller score than E, it implies that the E of the case is actually a strong match.

Similarly, the denominator of the ER is the area under the H_1 curve above the score of E. It is the proportion of comparisons which have obtained a greater score than E for which the source of the two compared recordings is not the same. As this proportion of cases increases, so does the support for choosing H_1, and the risk involved in choosing H_0 is higher.

Again, with the increase of the score of E, we come to the point where there are only very few comparisons between recordings coming from different sources giving a higher score than E, and the conviction that our value does not belong to the H_1 distribution becomes stronger. Actually, if no comparison between recordings coming from different sources has given a value higher than E, this supports the hypothesis that E belongs to the H_0 distribution. An illustration of the relative evolution of the LR and ER is presented in Fig. 3.12 using an example case created from the IPSC-02 database.

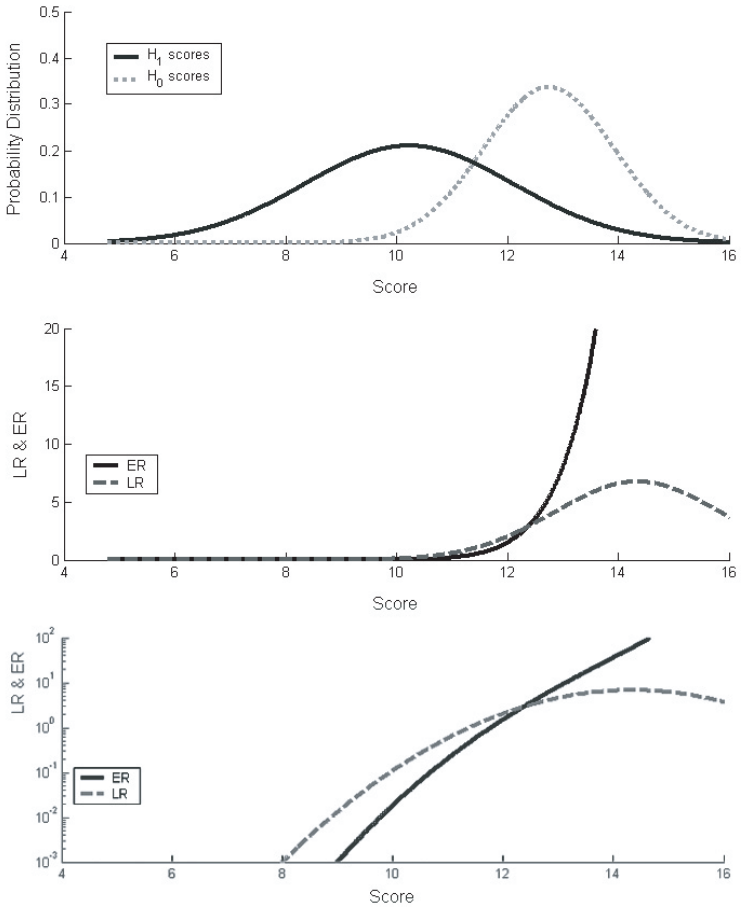


Figure 3.12: *Relative evolution of the LR and the ER for a case*

If the decision threshold in a speaker verification system is set to E, the system

cannot determine whether this trace comes from the suspect or not. This means that for this particular case, the verification system would neither accept nor reject H_0. This is desirable as at this threshold E, no binary decision has to be made for the case. However, the expert can evaluate how good E would be as a decision threshold for similar cases for which he knows whether H_0 or H_1 is true. If the expert tries to evaluate the performance of a system based on this threshold (E) with all the cases, in his experience, similar to the given case, he will be able to obtain the False Match Rate (FMR) and False Non-Match Rate (FMNR) for this value of E. If he obtains a very high FMR and a very low FNMR, he knows that the risk involved in supporting the H_0 hypothesis would be high. Similarly, if he obtains a low FMR and a high FNMR, implying a high ER, he can conclude that, based on his experience with cases under similar conditions, there is lower risk involved in accepting H_0. From the perspective of making a decision, it is desirable to know the relative risk of choosing one hypothesis over the other, based on past experience with cases in similar conditions, using this system. Both the likelihood ratio and the error ratio present complementary information, and we do not suggest the exclusive use of one method or the other.

When LR=1, the probability of observing the score of E given one hypothesis (H_0) is the same as given the competitive hypothesis (H_1). The strength of evidence is 1, and neither of the hypotheses can be favored. When the total error (FMR+$FNMR$) is minimum, it implies that at this point LR=1. ER, at this point, can be different from 1, since the risk of error that we would have if we took a decision in favor of one hypothesis might be higher than the risk of error taking a decision in favor of the other hypothesis. ER will be equal to 1 when, for a given value of E, the risk of error in choosing either of the hypotheses is equal.

$$FMR_E + FNMR_E = minimum \quad => \quad \mathbf{LR = 1} \tag{3.12}$$

$$FMR_E = FNMR_E \quad for \quad \mathbf{ER = 1}. \tag{3.13}$$

Additionally, the ER is less sensitive than the LR to artifacts of the modeling of each of the distributions. The LR is calculated as a ratio of two heights, and its estimation is sensitive to the mathematical modeling of the probability density function (pdf) of H_0 and H_1 and artifacts (especially in the tails of the pdf) can lead to erroneous estimation of LR, as the ER is calculated as a ratio of areas, and it is less sensitive to these artifacts.

3.8 Summary

In this chapter we have considered practical issues in the implementation of the Bayesian interpretation methodology for forensic automatic speaker recognition casework and the evaluation of results. We presented methods to deal with cases with sufficient and insufficient suspected speaker data, different methods for evaluating the strength of evidence, analysis of the variability in the strength of evidence as well as complementary information that should be presented along with the strength of evidence to the courts.

In order to present the results of forensic speaker recognition analysis to the court for a given case, the expert should evaluate the following:

- The likelihood ratio, accompanied by a bootstrapped confidence interval and the range of possible values of the strength of evidence.
- The statistical significance of the evidence score obtained with respect to each of the hypotheses.
- Complementary information such as probability distribution plots for cases in similar conditions like Tippett plots, as well as the relative error in choosing either hypothesis.
- The equivalent of the range of likelihood ratio scores on the verbal scale, and an explanation of the meaning of the scale.

The influence of mismatched recording conditions on aural and automatic estimates of the strength of evidence

4

In this chapter, we discuss the influence of mismatched recording conditions on aural and automatic speaker recognition, in the context of forensic analyses, using a Bayesian framework for the interpretation of evidence. Differences in the phone handset, in the transmission channel and in the recording tools have been shown to introduce variability and mismatch between two recordings in a case. In these conditions, the accuracy of the various automatic speaker recognition systems can be severely impaired. Human aural recognition is also affected by recording and environmental conditions, and judgements on the identity of speakers become less reliable in adverse conditions. It is of interest, therefore, to compare the performance of aural and automatic recognition in mismatched conditions common to forensic cases.

Perceptual tests were performed by non-experts and their performance was compared with that of an automatic speaker recognition system. These experiments were performed with 90 phonetically untrained subjects (or laypersons). Several forensic cases were simulated, using the Polyphone IPSC-02 database (see Appendix B), varying in linguistic content and technical conditions of recording *. The strength of evidence for both humans and the baseline automatic system is obtained by calculating likelihood ratios using perceptual scores for humans and log-likelihood scores

*This database was recorded in collaboration with Quentin Rossy [Rossy, 2003] from the Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne.

for the automatic system. A methodology analogous to Bayesian interpretation in forensic automatic speaker recognition is applied to the perceptual scores given by humans, in order to estimate the strength of evidence. The degradation of the accuracy of human recognition in mismatched recording conditions is contrasted with that of the automatic system under similar recording conditions. The conditions considered are fixed telephone (PSTN), cellular telephone (GSM) and noisy speech in forensically realistic conditions. The perceptual cues that human subjects use to pick out differences in voices are studied along with their importance in different recording conditions. While automatic speaker recognition was seen to have higher accuracy in matched conditions of training and testing, its performance degrades significantly in mismatched conditions. Aural recognition accuracy was also observed to degrade from matched conditions to mismatched conditions, and in mismatched conditions, the baseline automatic systems showed comparable or slightly degraded performance compared to the aural recognition systems. The baseline automatic system, with adaptation for noisy conditions, showed comparable or better performance than aural recognition. The higher level perceptual cues used by human listeners in order to recognize speakers are discussed. The possibility of increasing the accuracy of automatic systems using the perceptual cues that remain robust to mismatched recording conditions is also discussed. The work presented in this chapter was done in collaboration with Damien Dessimoz [Dessimoz, 2004] from the Institut de Police Scientifique.

4.1 Mismatched recording conditions and aural and automatic speaker recognition

Human beings use aural, linguistic and other background knowledge to perform this recognition. In the forensic context (in crimes such as kidnappings, rape, burglary, etc.), it is sometimes required that listeners identify the voice that they hear [Kerstholt et al., 2003]. Speaker recognition in forensic cases has often to be accomplished in difficult conditions, where distortions could be introduced either due to the transmission (telephone channels, ambient noise, etc.), the system (such as the telephone handset used, the recording instrument, etc.) or the speaker (disguise, stress, emotion, etc.). In this section, the extent to which aural and automatic recognition are adapted to environmental and channel conditions, and the relative performance of aural and automatic recognition in conditions of mismatch is analyzed, along with ways to adapt for mismatch. These experiments were done in order to compare and contrast the speaker recognition capabilities of ordinary untrained subjects (naive, unfamiliar speaker recognition) with those of an automatic speaker recognition sys-

tem.

Bayesian interpretation methods should provide a statistical-probabilistic evaluation which attempts to give the court an indication of the strength of the evidence (likelihood ratio), given the estimated within-speaker (within-source) variability of the suspected speaker's voice, and the between-speakers (between-sources) variability of the questioned recording, given a relevant potential population [Drygajlo et al., 2003]. Recently, there have been suggestions by forensic phoneticians to express the outcome of the aural and instrumental phonetic approaches as a Bayesian likelihood ratio. They have further qualified it as the logically correct way of quantifying the strength of identification evidence and suggested that it should constitute the conceptual basis for forensic-phonetic comparisons [Rose, 2002]. Corpus-based methodologies are used in the Bayesian interpretation framework in automatic recognition, using databases to evaluate the within-speaker variability and between-speakers variability (when it is possible to obtain sufficient recordings of the suspected speaker as well as that of a relevant potential population).

As discussed in Sec. 3.2.2, in many cases, only one recording of the suspect is available due to the nature of the investigation, e.g. when it is not possible to have additional recordings of the suspect's voice, as it may alert him to the fact that he is being investigated. It is often necessary to perform one-to-one comparisons of the questioned recording and the recordings of the suspect's voice. As a consequence, it is not always possible to evaluate the within-source variability of the suspect with this single recording. However, since this is a recurring problem in forensic speaker recognition, an interpretation framework for evaluating the evidence, even in the absence of additional control recordings, has been investigated [Botti et al., 2004a].

The automatic speaker recognition system relies on its feature extraction and statistical modeling algorithm, just as the subjects rely on their experience in extracting features specific to the given speaker and build and memorize a 'model' of the identity of the speaker's voice. Thus, the experience and ability of the human brain to extract speaker-specific information and its memory of a particular voice identity can be compared to the automatic system's feature extraction and statistical modeling algorithms. This analogy is extended further by allowing the subjects to listen to the recordings as many times as they would like, before converging to their memorized model of the identity of the speaker, just as the statistical modeling algorithm is allowed to converge to a statistical model of the identity of the speaker after several iterations (Sec. 2.2.2).

In order to simulate forensic case conditions, it was necessary to select a test database from which mock cases could be created. In this study, the Polyphone IPSC-02, which is a forensic speaker recognition database, was chosen. This database contains speech from 10 speakers in three different recording conditions and

two languages. The three recording conditions include transmission through a PSTN (Public Switched Telephone Network) and a GSM (Global System for Mobile communications) network and recordings made using an analog tape-recorder answering machine. French and German are the two main languages present in this database. For more details about this database, see Appendix B.

In our experiments, we have used a subset of five speakers from this database, using speech recorded through a PSTN and GSM network, from speakers whose mother tongue is French. For one part of the test, artificially generated white noise was added to the PSTN recordings at a signal-to-noise (SNR) ratio of 10 dB. Five short segments of spontaneous speech (between 10 and 20s) for each speaker in PSTN, GSM and noisy conditions were used for the mock questioned recordings and five longer recordings (90s) were used as reference recordings for the mock suspected speaker. Some studies indicate that using languages different from the listener's mother tongue influence the accuracy of recognition [Yarmey, 1995], and hence, all the test recordings used were in French (the mother tongue of all the test subjects). The mock questioned recordings chosen were simulated forensic cases with undisguised imitations of hoaxes and threats. The test database (of voices from the five speakers), from which the mock questioned recording and the mock suspected speaker's voice for each comparison had been selected, consisted of speakers who belonged to the same geographical region (the French-speaking part of Switzerland) and were all male university students. However, care was taken to ascertain that none of these test speakers were familiar to the 90 subjects who participated in this test, although they came from the same geographical region.

In phonetic terms, this recognition is termed naive, unfamiliar speaker recognition, where the listeners are tested on their ability to recognize a voice with minimal prior exposure. It has been reported that familiar speaker recognition, where the listener is familiar with the voice of the speaker because they have been exposed to it sufficient number of times, shows significantly lower error rates than unfamiliar, naive recognition, often with half the percentage of errors [Rose, 2002, :98].

4.2 Estimating the strength of evidence in aural and automatic speaker recognition

A total of 90 French-speaking subjects were chosen for the experiments, each performing 25 comparisons, with no limitation on the number of times they could listen to a particular recording. None of these speakers had any formal training in phonetics. A seven-level verbal scale was established, ranging from *I am certain that the speakers in the two recordings are different* (corresponding to a score of 1) to *I am*

certain that the speakers in the two recordings are the same (corresponding to a score of 7). The option *I am unable to make any judgement whether the speakers were the same or different* was placed at the middle of the scale (corresponding to a score of 4). The seven-level scale was chosen because of studies that suggest that it is difficult for humans to differentiate accurately between more than seven levels of comparison for a certain stimulus [Miller, 1956]. These scores and their verbal equivalents are presented in Table 4.1.

Table 4.1: Perceptual scores and their verbal equivalents

Score	**Verbal Equivalent**
1	I am sure that the two speakers are not the same
2	I am almost sure that the two speakers are not the same
3	It is possible that the two speakers are not the same
4	I am unable to decide
5	It is possible that the two speakers are the same
6	I am almost sure that the two speakers are the same
7	I am sure that the two speakers are the same

The subjects were required to listen to the suspect reference recording and the questioned recording, and to estimate their similarity on the verbal scale. The order in which the comparisons were presented was randomized in order to minimize the effect of memorization (by which the subjects would remember the voices heard in previous comparisons, and use this knowledge in order to make their decision). All the tests in these experiments were performed with a single computer (the same sound card), using exactly the same set of headphones. At the end of the recognition session, the subjects were also asked to list the factors that they considered important when making their decisions. Each test session corresponded to approximately one hour, with a total of 90 hours spent in order to perform the aural comparisons.

In the automatic system, feature extraction was performed using the RASTA-PLP [Hermansky, 1994] technique, and statistical modeling of these features was performed using a Gaussian Mixture Modeling (GMM) based classifier with 32 Gaussian mixture components [Reynolds and Rose, 1995]. The zones where speech was not present were removed in a pre-processing phase, in order to avoid including non-speech information in the statistical models of the speakers' voices. During the testing phase, features were extracted from the test utterances and compared with the statistical model of the speaker created in the training phase. The likelihood of observing the features of the test recording in the statistical model of the suspected speaker's voice was calculated. The logarithm of the likelihood obtained was used as a score for the comparison of the test utterance and the model. These scores were normalized, per training utterance, to a mean of 0 and a unit standard deviation in order to standardize scores across

speakers.

The total computation time taken to perform feature extraction, training and testing for all the comparisons was approximately three hours.

From the human subjects as well as the automatic speaker recognition system, a set of scores is obtained as results for these mock cases. The scores from the human subjects are discrete, and range from 1 to 7, while the automatic system returns scores that are on the continuous scale and can take any value in the continuous range of results, with a mean of 1 and standard deviation of 0. In order to be able to interpret the similarity scores (E) in a Bayesian interpretation framework, it is necessary to compare them with respect to the two competing hypotheses, H_0 - the suspected speaker is the source of the questioned recording, and H_1 - the speaker at the origin of the questioned recording is not the suspected speaker.

Both the automatic speaker recognition system and the human subjects give scores for the comparison between the two recordings, which indicate how similar the two recordings are to each other. In the forensic automatic speaker recognition case, the evidence (E) is the degree of similarity, calculated by the system, between the statistical model of the suspect's voice and the features of the trace [Drygajlo et al., 2003]. In the forensic aural recognition case, however, an estimate of the similarity score (E) between the suspected speaker's voice and the questioned recording is given by a subject on a verbal perceptual scale.

In order to evaluate the strength of evidence, the forensic expert has to estimate the likelihood ratio of the evidence, given the hypotheses that two recordings have the same source (H_0) and that the two recordings have a different source (H_1). The evaluation of the likelihood ratio of the evidence E allows us to calculate the degree of support for one hypothesis against the other as follows:

$$LR = \frac{p(E|H_0)}{p(E|H_1)}. \tag{4.1}$$

Note that in forensic automatic speaker recognition, the expert should be able to evaluate whether the tools he uses perform well with the recordings, whether incompatibilities between the databases that he uses can affect the estimation of the strength of evidence and whether compensations can be performed to reduce the effects of such incompatibilities [Alexander et al., 2004].

Here, it is necessary to calculate the strength of the evidence for the discrete and the continuous case in aural and automatic recognition respectively.

The likelihood ratio, as illustrated in Fig. 4.1, is the ratio of the heights of the distributions of scores for hypotheses H_0 and H_1, at E. The likelihood ratio for the perceptual scores given by the test subjects can be calculated in a similar way using

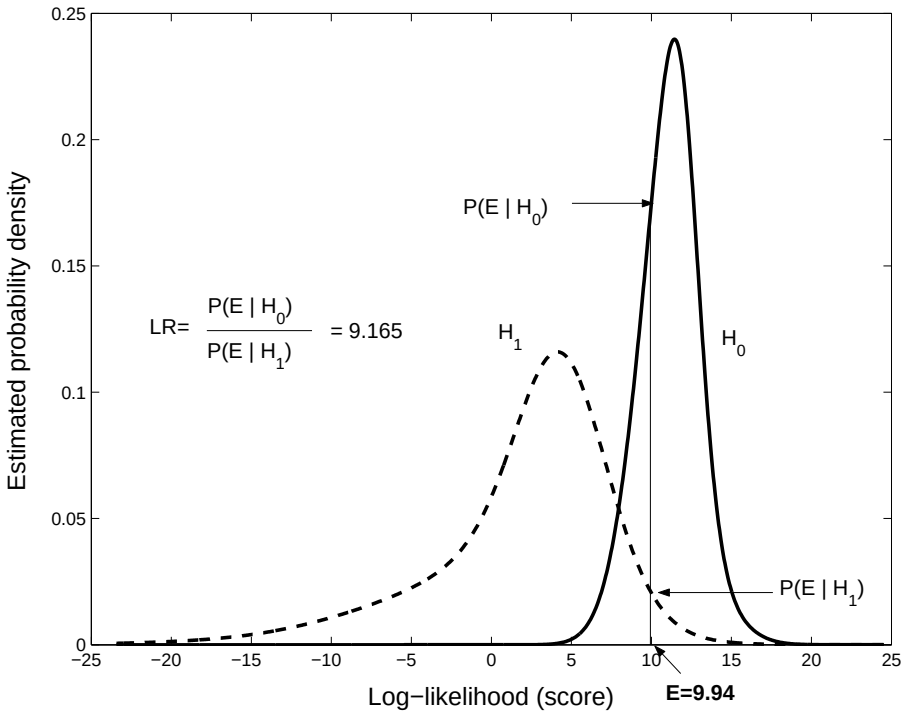


Figure 4.1: *Estimation of Likelihood Ratio using automatic speaker recognition scores*

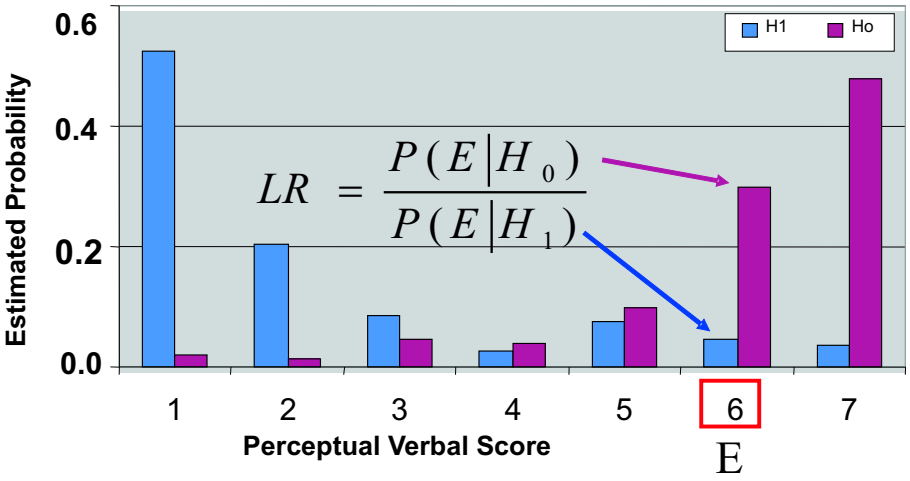


Figure 4.2: *Estimation of Likelihood Ratio using aural speaker recognition scores*

Bayesian interpretation. Since the scores obtained in this perceptual scale are discrete scores (from 1 to 7) and not continuous as in the case of log-likelihood scores returned by the automatic system, a histogram approximation of the probability density can be calculated. The histograms corresponding to each of the hypotheses are obtained from the perceptual scores corresponding to the mock cases where H_0 and H_1 are true, and calculating the frequency of appearance of these values (from 1 to 7) for all

the scores on the perceptual scale.

The frequency of appearance is calculated in the following way. Each score (from 1 to 7) is considered as an interval, and the number of times a certain score is obtained, when one of the hypotheses is known to be true, is used to calculate the frequency of appearance of that score. The area corresponding to each of the histograms is then normalized, so that the total area under each of the histograms is unity, by dividing it with the total number of scores per hypothesis. The likelihood ratio would then be the relative heights, on the histogram, of the two hypotheses, at the point E (shown in Fig. 4.2).

4.3 Comparing the strength of evidence in aural and automatic methods

The strength of evidence can be evaluated by estimating and comparing the likelihood ratios that are obtained for the evidence E in mock cases, where hypothesis H_0 is true and when the hypothesis H_1 is true. By creating mock cases which correspond to each of these hypotheses and calculating the LRs obtained for each of them, the performance and reliability of the speaker recognition system can be evaluated. In this way, we get two distributions; one for the hypothesis H_0 and the other for the hypothesis H_1. With these two distributions, it is possible to find the significance of a given value of LR that we obtain for a case, with respect to each of these distributions.

In order to measure the performance of each of the speaker recognition methods, the cases described in the previous section are considered, separating them into those where it was known that the suspected speaker was the source of the questioned recording and those where it was known that the suspected speaker was not the source of the questioned recording. These results are represented using cumulative probability distribution plots called Tippett plots.

As presented earlier, a Tippett plot represents the proportion of the likelihood ratios greater than a given LR, i.e., $P(LR(H_i) > LR)$, for cases corresponding to the hypotheses H_0 and H_1. The separation between the two curves in this representation is an indication of the performance of the system or method, with a larger separation implying better performance than a smaller one.

4.4 Comparison in matched and mismatched conditions in terms of Bayesian interpretation of evidence

The experimental results are represented using the Tippett plots $P(LR(H_i) > LR)$ (Figs. 4.3 and 4.4). The integration of probability distribution, which can be used to represent how many cases are above a given value of likelihood ratio with respect to each hypothesis, H_0 or H_1, can be used to indicate to the court how strongly a given likelihood ratio can represent either of the hypotheses.

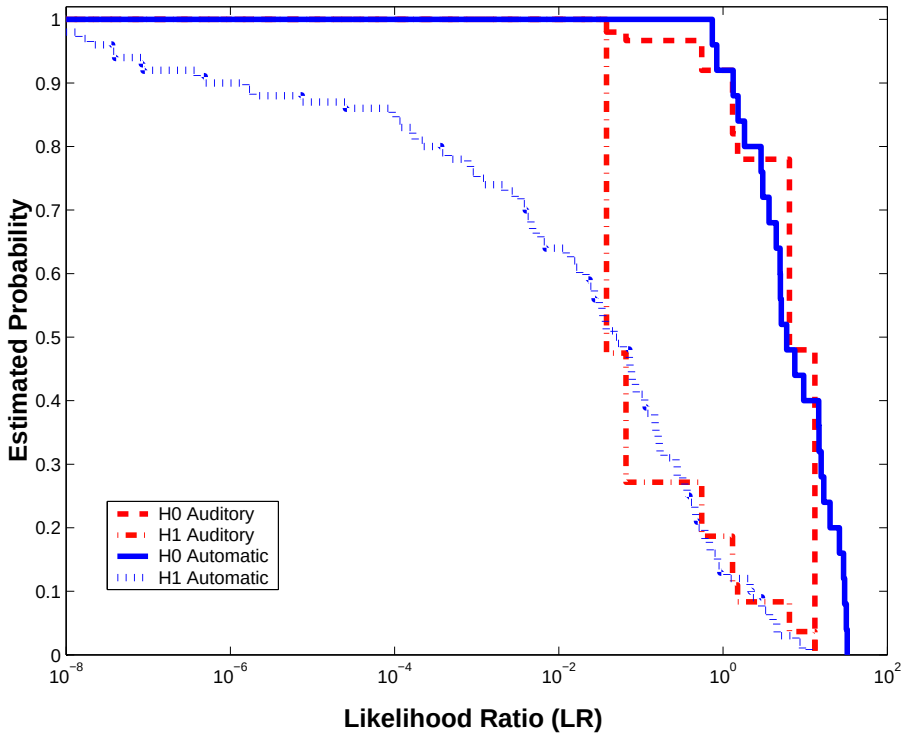


Figure 4.3: *Tippett plot in matched condition (PSTN-PSTN) for aural and automatic recognition*

We observe, in the Tippett plots for both the aural and automatic systems, that in matched conditions, the likelihood ratios for the H_0 and H_1 hypotheses are well separated. In Fig. 4.3, the aural and automatic likelihood ratios are presented for matched conditions, i.e., when both the suspected speaker's speech as well as the questioned recording were made in PSTN transmission conditions. In this plot, we observe that both the aural and automatic systems show good separation between the two curves. The curves of the automatic system are slightly more separated than those of the aural recognition. Here, we can observe that in matched conditions, the automatic recognition performed better than the aural recognition.

However, in the Tippett plots for both the aural and automatic systems, we observe that in mismatched conditions, the likelihood ratios for the H_0 and H_1 hypotheses are not as well separated as in the matched case. In Fig. 4.4, the aural and automatic likelihood ratios are presented for mismatched conditions, i.e., when the suspected speaker's speech was made in PSTN transmission conditions and the questioned recording was in noisy-PSTN conditions. In this plot, we observe that both the aural and automatic systems show degraded performance, with the curves of the aural system showing more separation than those of the automatic recognition. This implies that in mismatched conditions, the aural recognition performed better than the automatic recognition.

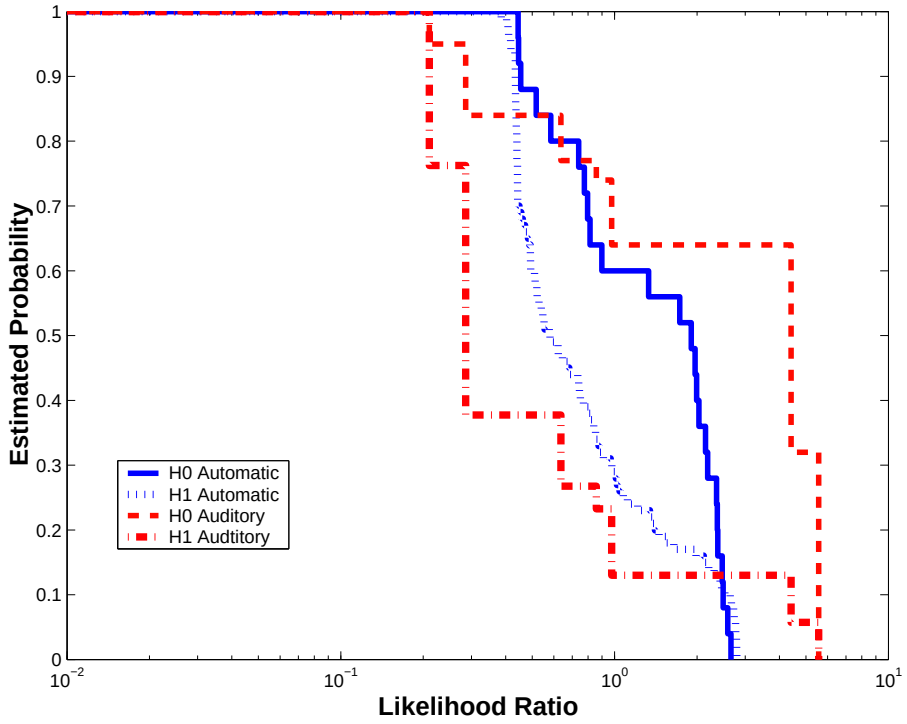


Figure 4.4: *Tippett plot in mismatched conditions (PSTN-Noisy PSTN) for aural and automatic recognition*

In Fig. 4.5, the aural likelihood ratios and the likelihood ratios of the automatic system adapted to mismatched conditions, i.e., when the suspected speaker's speech was in PSTN transmission conditions and the questioned recording in noisy-PSTN conditions, are presented. The enhancement applied in the automatic system was using an algorithm for spectral subtraction, in order to reduce the effects of the noise [Martin, 1994]. This method helped boost the performance of the automatic system when it was applied at the pre-processing phase of the recognition. Here, we observe a better separation between the curves corresponding to the automatic recognition. This performance is similar to the performance of the aural recognition.

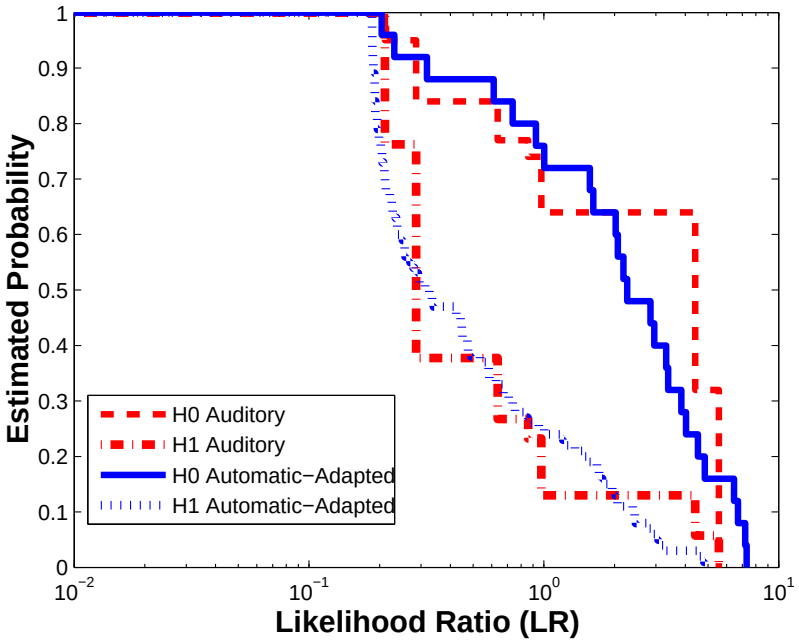


Figure 4.5: *Tippett plot in adapted mismatched conditions (PSTN-Noisy PSTN) for aural and automatic recognition*

4.5 Comparison in matched and mismatched conditions in terms of Bayesian decision theory

In order to compare the performance of automatic speaker verification systems, often, Receiver Operator Curves (ROC) or Detection Error Tradeoff (DET) curves are used [Martin et al., 1997]. The DET curve plots the relative evolution of False Match Rate (FMR) and False Non-Match Rate (FNMR) when using a decision point. The Equal Error Rate (EER) is the point at which the FMR is equal to the FNMR and is used in order to evaluate the performance of automatic systems. Although the forensic speaker recognition task does not use a threshold, as in the speaker verification approach, it is informative to compare measures existing in the speaker verification domain to measures used in forensic speaker recognition, in order to ascertain whether the same trends can be observed.

Note that the DET curve for aural recognition is not as smooth as that for the automatic recognition. This is because of the limited number of discrete perceptual verbal scores (from 1 to 7) and not because of the number of comparisons used to plot the DET curves.

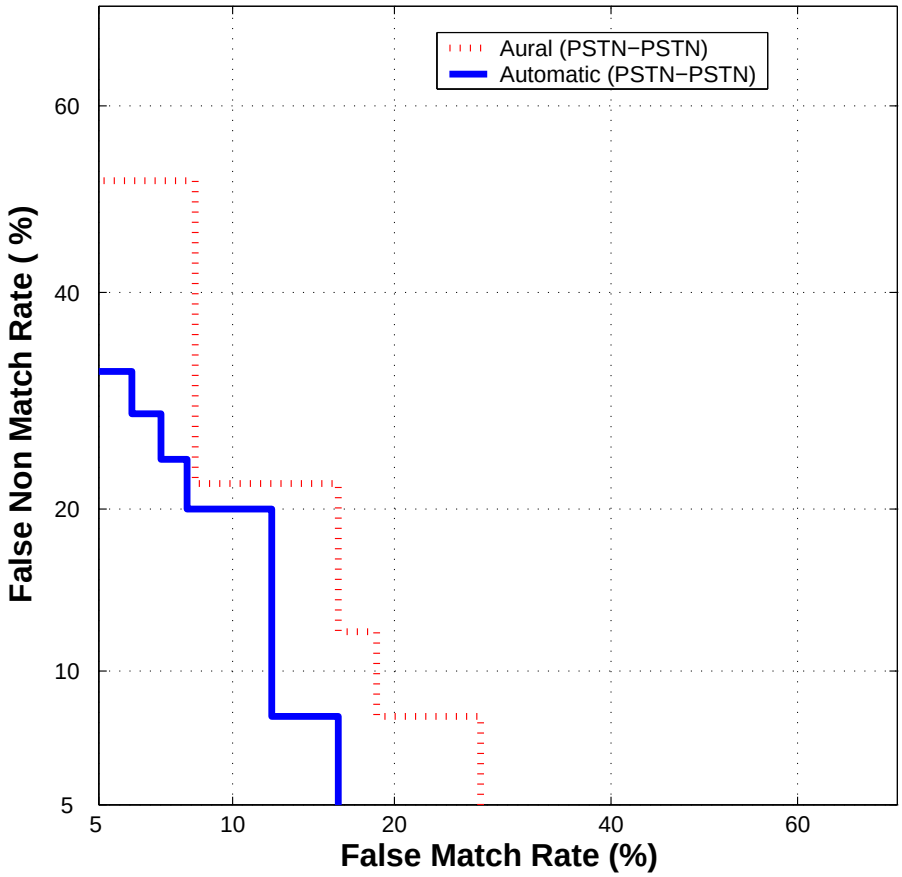


Figure 4.6: *DET Plot for comparison between the aural and the automatic recognition (PSTN-PSTN)*

Matched conditions

In Figs.4.6 and 4.7, we observe the relative performance of aural and automatic recognition in matched conditions (PSTN-PSTN) and (GSM-GSM) respectively. Automatic speaker recognition is seen to outperform aural speaker recognition, in matched conditions, with EERs as low as 4% and 12% for GSM-GSM and PSTN-PSTN comparisons respectively. In general, both aural and automatic recognition perform better in matched conditions than in conditions of mismatch.

Mismatched conditions

In Figs.4.8 and 4.9, we observe the relative performance of aural and automatic recognition in mismatched conditions, PSTN-GSM and PSTN-PSTN noisy respectively. In noisy conditions, human aural recognition is better than the baseline automatic system not adapted to noise. However, in mismatched recording conditions, automatic speaker recognition systems show accuracies comparable to aural recognition when

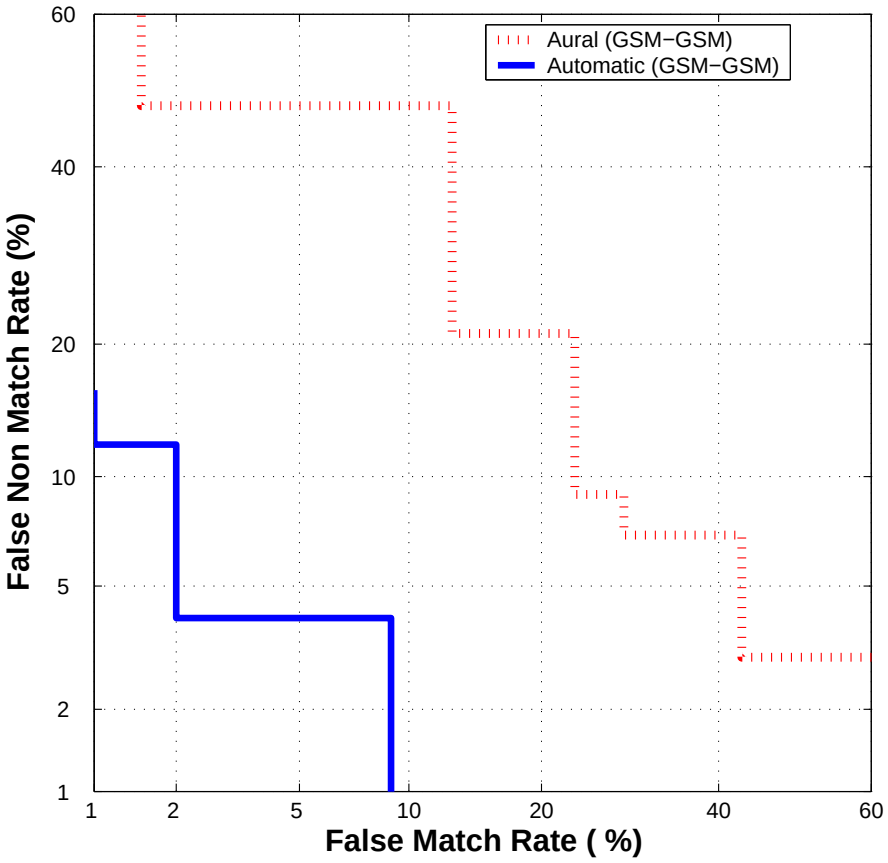


Figure 4.7: *DET Plot for comparison between the aural and the automatic recognition (GSM-GSM)*

the telephone channel conditions are changed. In mismatched conditions, the baseline automatic recognition system shows a similar or slightly degraded performance compared to the aural recognition. It is thus important that the baseline automatic system is adapted for the changes in conditions.

In Fig. 4.10, the DET curves, after adaptation for noisy conditions, are shown. Note that a reduction in the error rates is observed in the DET curves (Fig. 4.10) after this adaptation is performed. This shows a similar trend as in the Tippett plot (Fig. 4.5), where the separation between the curves for automatic recognition was reduced after applying adaptation for mismatch.

From the previous section, we conclude that aural recognition can outperform baseline automatic recognition in mismatched recording conditions. To a large extent, this can be attributed to the human auditory capability to adapt to different conditions and to use extra information that is not explicitly modeled in the automatic system. Apart from the speech signal, human recognition relies on other external cues that are difficult to model by automatic methods.

For instance, when humans recognize other speakers, they include high-level speech

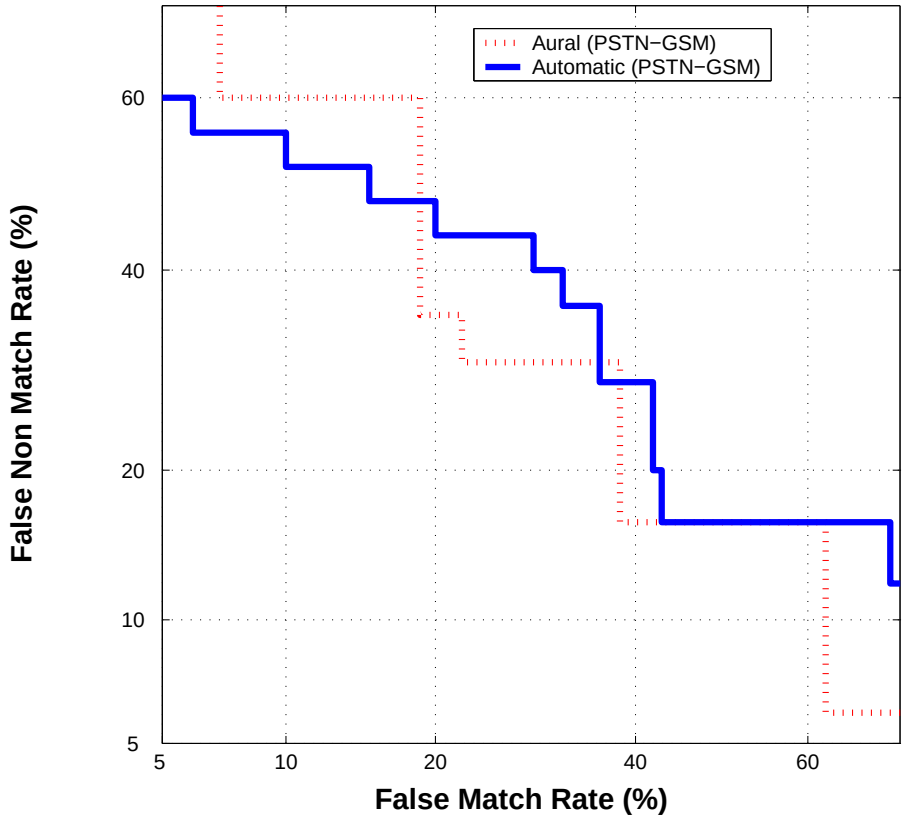


Figure 4.8: *DET Plot for comparison between the aural and the automatic recognition (PSTN-GSM)*

features about the identity of the speaker such as mannerisms in their speech, certain anomalous words that they use, their accent, pauses, speaking rate, etc. For example, during telephone calls, although the voice heard at the receiver end of the listener may be very different from the original voice of the speaker, most people are able to identify the caller quite easily. Often, listeners are able to make the transition from a rough idea of the identity of the speaker to pinpointing exactly who he or she is simply by paying attention to the high-level data such as the vocabulary chosen, the pronunciation of certain words and using other background information that one may have about the caller. These features are largely independent of the conditions of recording and are the same across conditions. Since these cues are largely relied upon, the drop in accuracy of aural speaker recognition, when there is a change in conditions, remains small. However, for the automatic system, which relies mainly on the low-level speech information and does not incorporate higher level information, the changes in conditions may make a large difference in its accuracy.

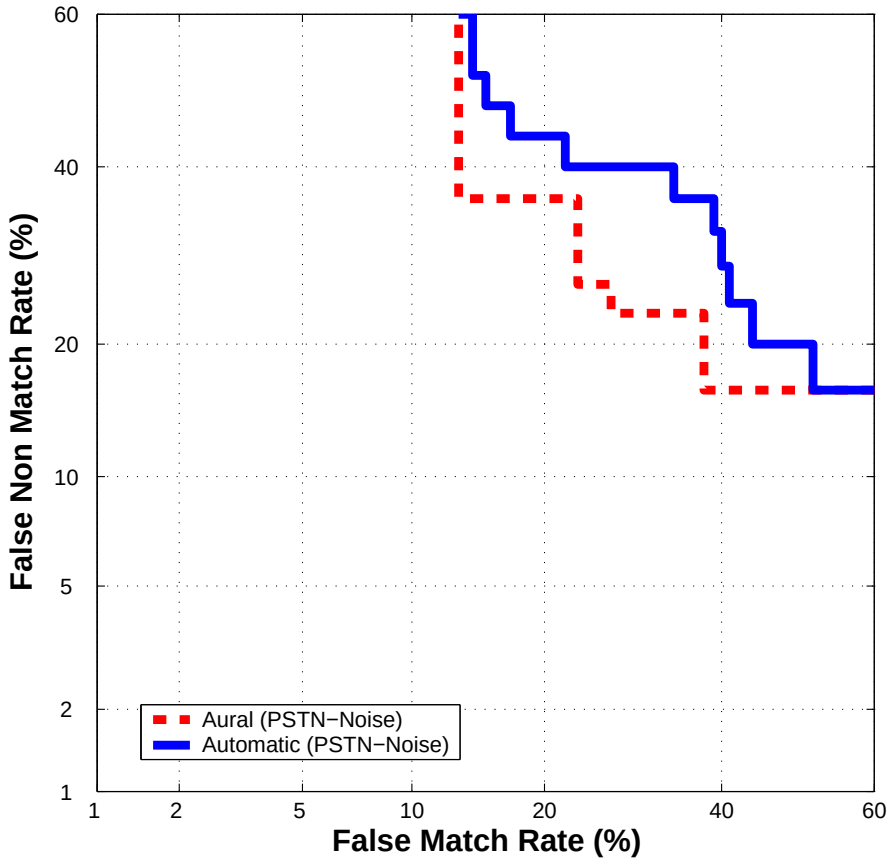


Figure 4.9: *DET Plot for comparison between the aural and the automatic recognition (PSTN-Noisy PSTN)*

4.6 Perceptual cues used by laypersons

Human beings depend largely on perceptual cues in the speech, such as pronunciation, word choice and speech anomalies [Schmidt-Nielsen and Crystal, 2000]. There have been studies which attempt to quantify these various aural and perceptual means that laypersons use in order to identify speakers [Voiers, 1964].

In our study, we were able to identify the factors important to each of the subjects in order to recognize speakers, i.e., the accent, the timbre, intonation, the rate of speech, speech defects or anomalies, breathing, the loudness, similarity to known voices and their intuition. These criteria were obtained by asking the subjects, at the end of each experiment session, what factors they considered in recognizing the source of the questioned recording. The subjects were allowed to describe the characteristics which they thought they used in order to recognize speakers, in their own words, and no prompts were given. These perceived characteristics were compiled and grouped. Since none of the subjects had any training in phonetics, the words were interpreted and categorized.

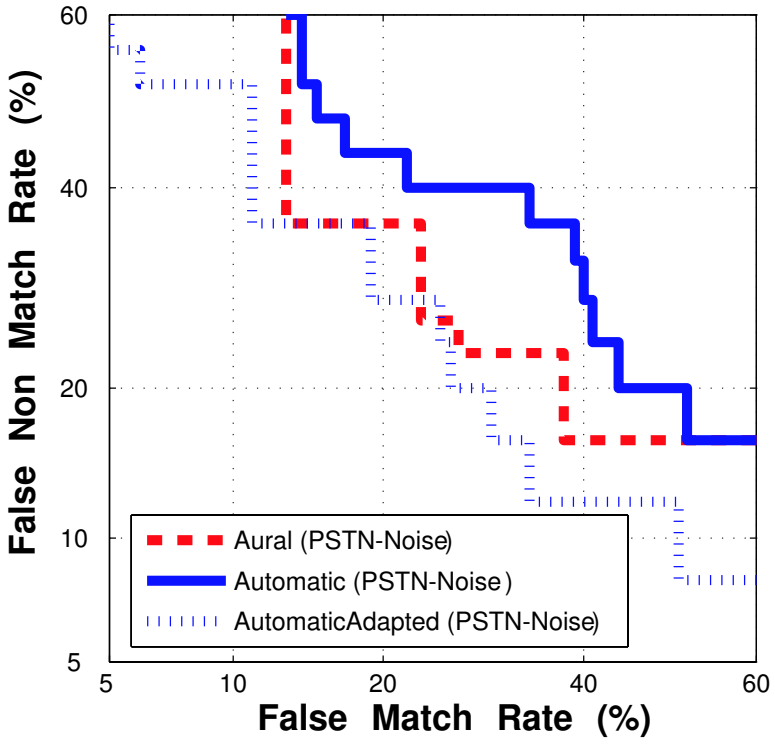


Figure 4.10: *DET Plot for comparison between the aural and the automatic recognition (PSTN - Adapted Noisy PSTN)*

Table 4.2: Relative importance of perceptual cues

Criteria	PSTN-PSTN (%)	PSTN-GSM (%)	PSTN-Noise (%)	GSM-GSM (%)
Accent	34	31	30	26
Timbre	25	25	22	22
Intonation	16	24	18	21
Rate of speech	9	7	12	12
Speech anomalies (defects)	6	7	8	5
Breathing	5	0	2	2
Volume	3	0	0	0
Imagined physiognomy	3	0	2	2
Similarity to known voices	0	2	0	2
Intuition	0	4	6	9

In Table 4.2 , we have presented these factors and their relative importance to the subjects in each of the different conditions of recording. The recording conditions have been varied in order to study the differences in the perceptual cues that human beings use to recognize different speakers.

We observe that the main characteristics that humans depend upon, in all the three conditions are mainly the accent, intonation, timbre, rate of speech and speech anomalies. We have tried to identify these perceptual cues used by the subjects using our interpretation of their responses, in order to understand which of these cues can be incorporated into the automatic system in order to improve its performance.

- Accents: Accents can vary widely regionally as well as socially. In this study, a subset of the Polyphone IPSC-02 database was chosen, so that the speakers would have a Swiss French accent, but regional variations within this population were not considered. All the test speakers had similar social and educational backgrounds. In spite of these constraints, the accents were the most important feature that the subjects claimed to use when recognizing the speakers. Accents are difficult to model explicitly in automatic systems, although it is possible to build background statistical models that represent a particular accent or dialect, and use it to normalize the automatic recognition scores.

- Intonation : Speakers use intonation to encode syntactic features by modifying their pitch [Rose, 2002], as well as to signal information about the emotional state of the speaker (e.g., anger, boredom, depression, etc.). Intonation may include differences in pitch level (e.g. high and low voices) and in pitch movement (e.g. monotonous and melodic voices). Some of the subjects believed they used intonation as a criterion in comparing the recordings belonging to different speakers. Pitch has been successfully incorporated in automatic speaker recognition [Arcienega and Drygajlo, 2003]. It is extremely difficult to incorporate intonation information indicative of the abstract emotional state of the speaker in the automatic system, although the use of suprasegmental features like pitch shows some promise.

- Rate of speech: Speakers differ from each other in terms of the rate at which they speak and other speech durational factors (pauses, time taken for a word or phrase, etc.), and this is useful information in ascertaining the identity of a speaker. There have been studies in forensic phonetics, in which temporal information about the speech and their use as forensic parameters for speaker recognition were considered [Hollien, 1990; Künzel, 1997]. In some of our preliminary experiments modeling the normalized rate of speech, we concluded that by normalizing the rate of speech and using information about the transition from silence to speech zones, it is indeed possible to increase the accuracy of automatic recognition. However, it is necessary to have sufficient amounts of recordings to derive these statistics. This is often not the case when we consider forensic cases with short questioned recordings. In these tests, the questioned

recordings were too short to derive any meaningful statistics about the rate of speech.

- Timbre: Timbre is the subjective correlate of all those sound properties that do not directly influence pitch or loudness. These properties include static and dynamic aspects of the voice quality (laryngeal or supralaryngeal), vowel quality (formant structure) and the temporal evolution of the speech spectral power distribution. The subjects believed that this was a factor they considered in judging the identity of the speaker. Although arguably, this information is present in the low level speech information, the timbre is not explicitly quantified as a parameter in automatic speaker recognition.

- Speech anomalies: Because of the rule-based nature of language, speakers of a language can usually discern realizations of speech that deviate from 'normal speech' well [Rose, 2002]. While certain deviations of speech may be reasonable or acceptable in one language, they may not be acceptable in other languages. These speech- and voice-pathological deviations are often used as clues to the identity of a person in aural recognition. Speech anomalies are difficult to model in an automatic system as these are often linguistic in nature. Baseline text-independent speaker recognition, which is used in the automatic recognition system, does not allow us to take such information into consideration.

4.7 Relative importance of the perceptual cues

The relative importance of each of these main characteristics remains very similar across different recording and environmental conditions implying that human perception of speaker identity mainly depends on characteristics that are robust to conditions. This is in stark contrast to the baseline automatic speaker recognition system which depends heavily on the conditions of recording. Considering these additional factors is of importance in severely degraded conditions as is often the case in forensic casework. The human auditory system is able to adapt to the effects of masking by noise and other distortions [Bregman, 1990]. Consequently, in the automatic system, it is necessary to explicitly adapt the recognition process to each of the conditions.

4.8 Summary

In this chapter, perceptual speaker recognition tests were performed with laypersons and their performance was compared with that of a baseline automatic speaker recognition system. It was observed that in matched recording conditions of suspect and

questioned recordings, the automatic systems showed better performance than the aural recognition systems. In mismatched conditions, however, the baseline automatic systems showed comparable or slightly degraded performance compared to the aural recognition systems. The extent to which mismatch affected the accuracy of human aural recognition in mismatched recording conditions was similar to that of the automatic system under similar recording conditions. Thus, the baseline automatic speaker recognition system should be adapted to each of the mismatched conditions in order to increase its accuracy, as was observed with adaptation to noisy conditions. The adapted system shows comparable or better performance than aural recognition in the same conditions. The perceptual cues that human listeners rely upon in order to identify speakers were analyzed. The accuracy of automatic systems can be increased using these perceptual cues that remain robust to mismatched conditions.

This scheme of comparison can also be adapted in order to analyze the performance of the aural perception of persons with phonetic training. It should be stressed that in its current framework, this comparison relies only on the auditory abilities of the listener and does not take into account any other instrumental techniques. Thus, future study would essentially be comparing the automatic system with the auditory recognition abilities of the phonetician. It would evaluate the performance of phoneticians and automatic systems, in different conditions, and would indicate methods of combining the aural perceptive approach of trained subjects with that of the automatic system, depending on the conditions of the case. It is to be noted that while phoneticians also use aural-perceptive analysis of the recordings in forensic casework, they mainly use aural-instrumental techniques to note and extract features which can be used to derive a likelihood ratio.

Statistical compensation techniques for handling mismatched conditions

5

5.1 Introduction

In forensic speaker recognition casework, the recordings analyzed often differ because of telephone channel distortions, ambient noise in the recording environments, the recording devices, as well as in their linguistic content and duration. These factors may influence aural, instrumental and automatic speaker recognition. The scores obtained comparing recordings of speakers in the same recording conditions (also known as matched conditions), are different from the scores obtained comparing the same speakers in different recording conditions (known as mismatched conditions). In this chapter, we discuss a methodology to estimate the mismatch in recording conditions that arises in forensic cases, to quantify the uncertainty that is introduced due to different recording conditions and to compensate for mismatched recording conditions. The main issues treated in this chapter include:

- Analysis of how and why mismatched recording conditions pose an important problem for forensic automatic speaker recognition.
- Estimation of the extent of mismatch within and across databases used for analysis.
- Estimation and statistical compensation for mismatch using representative databases in similar conditions:
  - Creation of a forensic speaker recognition database to deal with mismatch and the requirements for this database.

– Use of this database to estimate and compensate for the change in the strength of evidence because of mismatched recording conditions.

5.2 Mismatched recording conditions and the Bayesian interpretation methodology

The data-driven Bayesian methodology for automatic speaker recognition requires, in addition to the questioned recording, the use of a suspect reference database (R), a suspect control database (C) and a potential population database (P), as described in chapters 2 and 3. In this data-driven approach, mismatch between the databases, due to the transmission conditions, recording devices, noise, linguistic content and the duration of the recordings, can influence the evaluation of the strength of the evidence.

Ideally, the three databases described must be in the same recording conditions for the forensic automatic speaker recognition to be accurate. For every case, each of these databases has to be chosen and verified carefully for compatibility in the recording conditions. If there is a possibility for the expert to record the suspect in controlled conditions, it is necessary that this acquisition is done, so that the two databases (R and C) thus obtained are compatible with the potential population database and with the questioned recording respectively. Indeed, in the creation of a suspect reference database (R), every effort is taken to record the database in approximately the same way as the potential population database (P). Similarly, in order to create the suspect control database (C), care must be taken to ensure that it is recorded in conditions similar to those of the questioned recording [Meuwly, 2001; Drygajlo et al., 2003]. However, in practice, it is extremely difficult to reproduce conditions *identical* to that of potential population database or to obtain control recordings in similar conditions as that of the questioned recording.

Generally, the forensic expert is faced with two situations in creating the suspect reference and control databases. One possibility is that he is supplied with the recordings from the police (or the court) using their recording equipment, in which case, he has little control over the conditions of recording. The other possibility is that he can perform the acquisition of the suspect's voice using his own recording equipment, in controlled conditions. However, even this does not ensure complete compatibility with the potential population database which has a large number of speakers recorded using several different telephones, in different transmission and background noise conditions.

Important assumptions in the corpus-based Bayesian interpretation methodology, presented in [Meuwly, 2001; Drygajlo et al., 2003], concern the recording conditions as

well as the linguistic content of the databases used in the analysis. The assumptions in this methodology are:

- The suspected speaker reference database (R) is recorded in similar conditions of recording and linguistic content to the potential population recordings (P).
- The suspected speaker control database (C) is recorded in similar conditions as well as linguistic content to the trace (questioned recording) (T).

In speaker recognition casework, it is often difficult to satisfy these assumptions. Practically, it is not often possible to obtain databases that are identical to the trace under consideration. For instance, if a trace received is recorded in conditions the expert has not come across before, it may be difficult to procure a relevant potential population database (as they may not be available or easily recorded) or a suspected speaker control database recorded in the same conditions as those of the questioned recording. In this chapter, we assume, according to the Bayesian interpretation methodology, that it is possible to record the R and C databases of the suspected speaker, and it is possible to determine the conditions for such recordings.

We consider a situation where the suspect databases, R and C, are recorded in conditions of the trace and do not correspond to the recording conditions of the potential population database. According to the Bayesian interpretation methodology described, the trace (T) is compared to the model of the suspect, and a value for E is calculated. Then, the trace is compared to all the speaker models of the potential population database to estimate the *between-sources* variability. The *within-source variability* is estimated by comparing the suspect control database (C) with the suspect model of the reference database (R). It should be remembered that since both C and R come from the suspect, their comparison should give scores that correspond to the H_0 distribution. Also, since we can be reasonably sure that the real source of the trace is not part of the potential population, comparison between the trace and the potential population gives scores that correspond to the H_1 distribution.

In a case where there is mismatch between P and the other databases, it is possible, for instance, that the technical conditions of the trace and those of the recordings used to create the models of the suspect are more similar to each other than to the technical conditions of the potential population database. This similarity of conditions may wrongly suggest a similarity between the voices.

Sometimes, this mismatch problem is two-fold, i.e., not just between databases, but also within databases. Mismatch within a database can severely affect the accuracy of speaker recognition, and the discriminative power of the system is diminished. Mismatch across databases may give misleading results which do not depend on the similarity of voices but on the similarity in the recording conditions. In the following section (Sec. 5.3), the detection of mismatch within and across databases is discussed.

5.3 Estimating discrimination and mismatch within a database

As the first step to any recognition task, it is necessary to check whether the feature extraction and classification algorithms that the forensic expert possesses are capable of discriminating between speakers under the given conditions. It is known that incompatible recordings within a database can result in diminished speaker recognition accuracy [Dunn et al., 2001; Reynolds et al., 2000]. Thus, before starting a speaker recognition task with a certain chosen database, it is necessary to verify whether the system is indeed able to discriminate efficiently between speakers under these conditions.

If the scores obtained when the suspected speaker is indeed the source of the questioned recording and the scores obtained when he is not the source are different and well separated, it implies that the system returns discriminative scores under these conditions. The distribution of likelihood scores for cases in which the speaker is truly the source of the test utterance should be approximately in the same range, likelihood scores for cases in which the speaker is not the source of the test utterance should be in another range of values, and there should be a good separation between these two ranges. The extent of separation between the scores of the two hypotheses is a measure of the discrimination of the recognition system on the database. In order to quantify this discrimination, we introduce the following measure:

$$DC = \left| \frac{\mu_{H_0} - \mu_{H_1}}{\sigma_{H_0} + \sigma_{H_1}} \right|, \tag{5.1}$$

where μ_{H_0} is mean of the scores when H_0 is true, μ_{H_1} is the mean of the scores where H_1 is true, σ_{H_0} is the standard deviation of the scores if H_0 is true and σ_{H_1} is the standard deviation of the scores if H_1 is true.

The discrimination coefficient (DC) is a measure of the distance between the two distributions. The DC allows us to quantify the separation between the two distributions. It is up to the expert to decide what value of DC would be acceptable for a database. For instance, if the DC is 1 or below, this would imply that the several values that belong to the distribution H_1 could also have come from the distribution H_0, and thus, the system shows poor discrimination with this database. Similarly, if DC is between 1 and 2, this implies moderate to good discrimination, and above 2 would imply very good discrimination between speakers.

In order to calculate the DC, we performed comparisons using a 194-speaker subset of the Swisscom Polyphone database, using a GMM-based classifier, with 32 Gaussian mixture components and 12 RASTA-PLP coefficients. The same test was performed using 39 speakers of the NIST 2002 FBI database [Beck, 1998; Nakasone

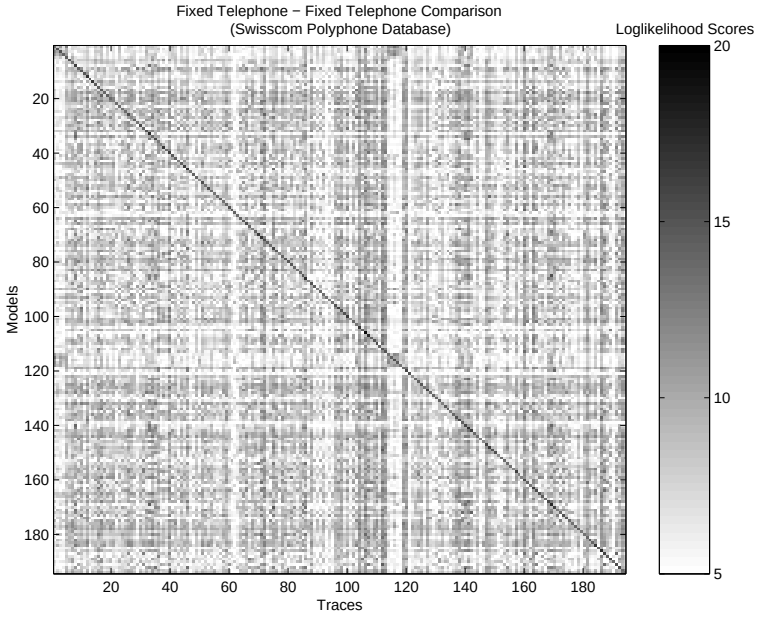


Figure 5.1: *Illustration of the discrimination on a subset of the Swisscom Polyphone database (PSTN)*

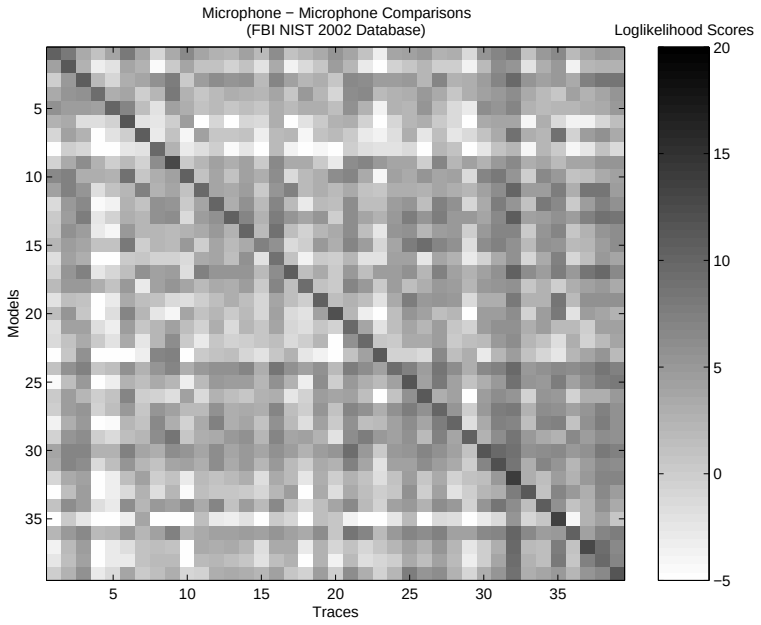


Figure 5.2: *Illustration of the discrimination on a subset of the FBI NIST 2002 database (Microphone)*

and Beck, 2001]. If we tabulate these results in a matrix, with the models of different speakers on one axis and the test utterances from these speakers in the same order on the other axis, the diagonal elements should have the largest score in each row. This is because these elements represent scores for H_0 true, and higher scores imply higher similarity between the suspect's models and the test features. Figs. 5.1 and 5.2 present a likelihood score gray-scaled graphic of the likelihoods returned (i.e., each square represents the score the test utterance of the speaker, in that particular column, obtained when comparing it with the model of the speaker in the row where it was found). The values of DC obtained were 1.7248 and 1.7556 for the FBI and Swisscom databases respectively.

In speaker verification tasks, in order to evaluate the performance of the automatic system, DET curves and EERs are used [Martin et al., 1997]. In order to calculate the DET curves and EERs, it is necessary to make a large number of mock client and mock imposter comparisons, in order to see whether the False Acceptance Rates (FAR) and the False Rejection Rates (FRR) are satisfactory for the use to which the automatic speaker verification system is put.

In our case, if it is possible to similarly generate a large number of comparisons corresponding to cases where the suspected speaker was the source of the trace and when the suspected speaker was not the source of the trace, then these measures can also be used to judge whether the recognition performs well under these conditions. With the resulting scores, it is possible to calculate the error rates and the DET curves. The use of such measures requires caution as, semantically, measures such as FRR, FAR and EER imply the use of a threshold and a binary decision, and if used in the Bayesian interpretation framework, should only be used in order to evaluate the performance of the system.

5.4 Measuring mismatch across databases

Often, the choice of a database, such as the potential population database P, is based on criteria such as language, sex and perceived technical conditions of the recording. However, this kind of choice is not always guaranteed to be sufficient (especially if the recordings of the suspect are very different from those of the potential population). One of the difficulties that a forensic expert faces is in determining whether a given database can be used with a new case or whether this choice will actually lead to the likelihood ratio being underestimated or overestimated.

We illustrate this data mismatch with the two following experiments:

- We choose 12 speakers from the potential population database (P) (Swisscom Polyphone database) and 12 speakers from a test database (IPSC-02 Polyphone

database), recorded according to the methodology specified in [Meuwly and Drygajlo, 2001]. We then create models for each of the speakers from both the databases and select traces for each of them. In order to test the assumption that the two databases are compatible, we compare each trace with all the models we have trained, for both the databases.

- A similar experiment is performed using the FBI NIST 2002 Speaker Recognition database. Here, two sets of 8 speakers are chosen under different conditions (telephone and microphone recordings) as potential population databases (P), and models are trained for each of these speakers and compared with test utterances from the same speakers. Comparisons are thus performed within a given recording condition as well as between the two conditions. If we obtain the same range of results for comparisons within a database as well as for comparisons across databases, we can conclude that the databases are compatible.

Let us consider cases in which H_1 is true. As illustrated in Figs. 5.3 and 5.4, there are four possible zones from which we obtain scores corresponding to H_1 true. These are the two zones within each database where the suspect is not the source of the trace, and two zones of comparison across the two database where the suspect is not the source of the trace. It can be observed that there is a clear difference between the range of H_1-true scores within and across databases.

In ideal conditions where there is no mismatch, all these values have to be in the same range. We propose to use a simple statistic to verify whether all the sets of H_1 scores have values that are in the same range [Aitken, 1997]. This is the statistic for a *large sample test concerning the difference between two means.*

Let us suppose we want to see whether we can reject the null hypothesis that there is no difference between the two separate distributions.

- Null Hypothesis : μ_1-μ_2= 0
- Alternative Hypothesis : μ_1-$\mu_2 \neq 0$

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{5.2}$$

where μ_1 is the mean of the first distribution, μ_2 is the mean of the second distribution, σ_1 and σ_2 are the standard deviations of the first and second distributions, and n_1, n_2 are the number of data-points in each of the distributions. For example, setting the level of significance at $\alpha = 0.05$, for z we should have $z > 1.96$ if the distributions are not similar or compatible. If this is not the case, we cannot attribute the differences between the two distributions to chance.

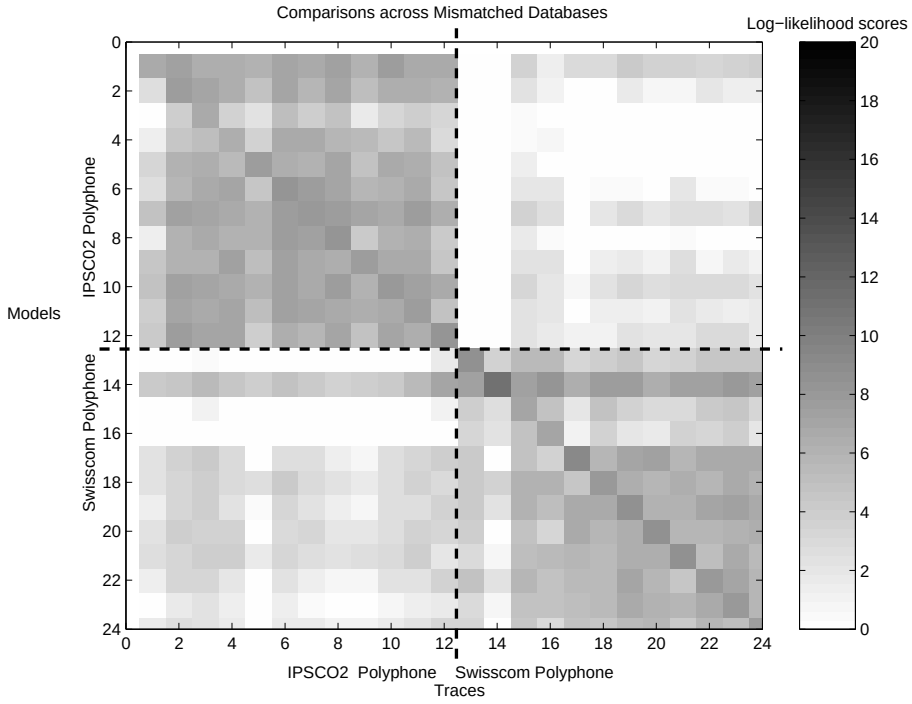


Figure 5.3: *Comparisons across incompatible databases: Swisscom Polyphone and IPSC-02*

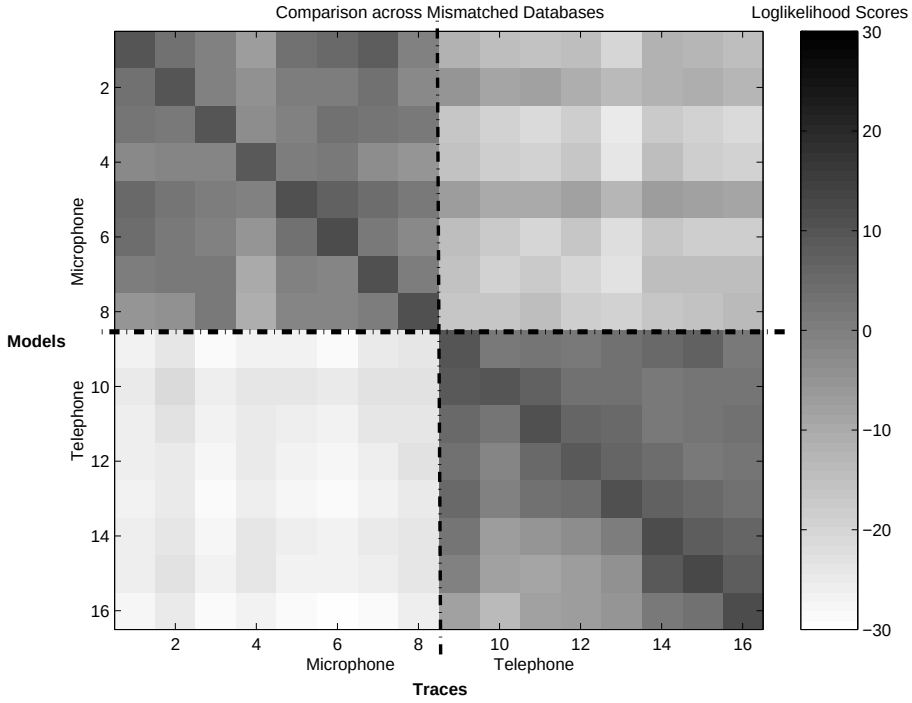


Figure 5.4: *Comparisons across incompatible databases: FBI NIST 2002 Microphone and Telephone*

If this statistic is not satisfied, the specialist should try to use the other databases he has at his disposal and find one that is compatible. Here, he has the option of choosing not to proceed with the analysis of the case using the Bayesian interpretation methodology or to statistically compensate for the mismatched conditions (presented in Section 5.5).

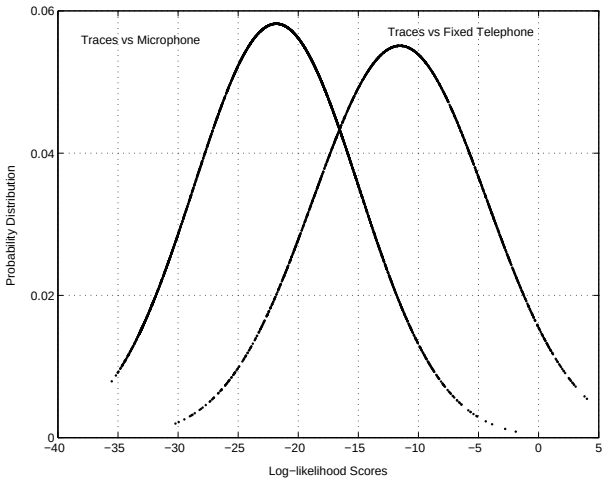


Figure 5.5: *Distribution of scores for the comparison of traces with the population database in two different conditions : fixed telephone and microphone*

Thus, if an existing mismatch is undetected, it is likely that the uncompensated usage of the Bayesian interpretation framework gives erroneous results. After detecting a mismatch, the expert has a choice of selecting another compatible database, if possible, deciding not to analyze the case in the Bayesian interpretation framework or performing statistical compensation for the mismatched conditions. Detecting and compensating mismatches between databases helps to reflect, more accurately, the similarity or dissimilarity of the real speech contained in the recordings.

5.4.1 Mismatch and the corpus-based Bayesian interpretation methodology

As discussed in Section 5.2, recordings obtained from the police do not often match conditions under which the potential population database was recorded, and it may not be possible to additionally record the voice of the suspect in matched conditions. Each of the databases considered can contribute to the problem of mismatch, and the corresponding scores for E as well as the H_0 and H_1 distributions obtained may give the under- or over-estimation of the likelihood ratio.

We make an assumption that the recording conditions of each of the databases is known precisely and that it is possible to record the suspected speaker R and C

databases accordingly in a certain known condition. This assumption is satisfied if the expert is able to obtain, from the police or investigating agency, information on what conditions the recordings were made in or compare the recordings with databases in known conditions. Also, enumerating the assumptions presented in 5.2 we have:

1. The suspected speaker reference database (R) is recorded in similar conditions to the potential population recordings (P).

2. The suspected speaker control database (C) is recorded in similar conditions to the trace (T).

In order to illustrate the effect of this mismatch, we consider each of the situations where mismatch is possible. In each case, the strength of the evidence (or the likelihood ratio) is affected differently. In order to understand the nature of the mismatch, we enumerate, in Table 5.1, the possible situations in which mismatched conditions can affect Bayesian interpretation of evidence, considering only two recording conditions (C_1 and C_2). Note that not all these possible situations are actually encountered in practice and have been listed only for completeness. In practical cases, the recordings could be made in GSM telephone, PSTN telephone or noisy conditions. If more conditions are considered, the complexity of the mismatch is compounded further.

Table 5.1: Mismatched recording conditions in the databases used and the assumptions in the Bayesian interpretation methodology

Sit. No	P	R	C	T	Satisfied assumptions
1	C_1	C_1	C_1	C_1	1 and 2
2	C_1	C_1	C_1	C_2	1
3	C_1	C_1	C_2	C_1	1
4	C_1	C_2	C_1	C_1	2
5	C_2	C_1	C_1	C_1	2
6	C_1	C_1	C_2	C_2	1 and 2
7	C_1	C_2	C_2	C_1	neither
8	C_1	C_2	C_1	C_2	neither

In [Meuwly, 2001], it is assumed that the reference (R) and control (C) databases can be recorded in the same conditions, or that R is in the same condition as P, and C is in the same condition as T. Therefore, situations #3, #4, #7 and #8 can be considered less likely to occur, as the conditions of R and C database can be suitably chosen by the expert in order to avoid these situations. Let us now consider situations #1,#2, #5 and #6.

Situation #1 is the ideal situation, of matched conditions, when all the speakers are recorded in exactly the same conditions. Unfortunately, in real forensic conditions, cases with perfectly matching conditions are not frequently encountered. The

Bayesian interpretation methodology has been demonstrated to work well when the conditions of recording do not suffer from mismatch [Alexander et al., 2005]. In this situation, with E and the H_0 and H_1 score distributions, the likelihood ratio is given by Eq. 5.3.

$$LR = \frac{p(E|H_0)}{p(E|H_1)} \tag{5.3}$$

In situation #2, i.e., recording conditions of P,C,R and T are C_1,C_1,C_1,C_2, the trace was recorded in a condition different from all the other databases in the analysis, i.e., the P, R, and C databases are all in the same condition. In this situation, the distribution that is most affected due to the mismatch is the $H_{1\leftrightarrow}$ distribution, since the trace in condition C_2 is compared to all the speakers from the potential population in condition C_1. Also, since the questioned recording is not in the same recording conditions as the P, R, C databases, the evidence score $E_{\leftrightarrow}$ (trace compared with R database) will also be affected by the mismatch.

Thus, due to mismatch,

- The hypothetical matched H_1 distribution is shifted to the distribution $H_{1\leftrightarrow}$.
- The hypothetical matched E score is shifted to the score $E_{\leftrightarrow}$

The likelihood ratio (LR) is also distorted to $LR_{\leftrightarrow}$.

The distorted likelihood ratio takes the form,

$$LR_{\leftrightarrow} = \frac{p(E_{\leftrightarrow}|H_0)}{p(E_{\leftrightarrow}|H_{1\leftrightarrow})}. \tag{5.4}$$

Because of the variation in E and H_1, the LR can vary within a range of values. Thus, the LR cannot be represented by a single value but by a range of possible values. This variation in the LR directly affects the strength of the evidence. The LR changes from support of the H_0 hypothesis to support of the opposing hypothesis H_1 at point 1. If the range of $LR_{\leftrightarrow}$ values extends both below and above 1, this would mean that within the variation of the LR, neither hypothesis can be supported.

The distributions, therefore, can take the following form, as shown in Fig. 5.6.

In situation #5, where the P database is in condition C_2, the R database is in C_1, the C database is in C_1 and T in C_1. This situation, however, may often be encountered in forensic cases as discussed in [Alexander et al., 2004].

Figs. 5.7 and 5.8 illustrate the pitfall of applying Bayesian interpretation when the conditions of the potential population database are incompatible with the conditions

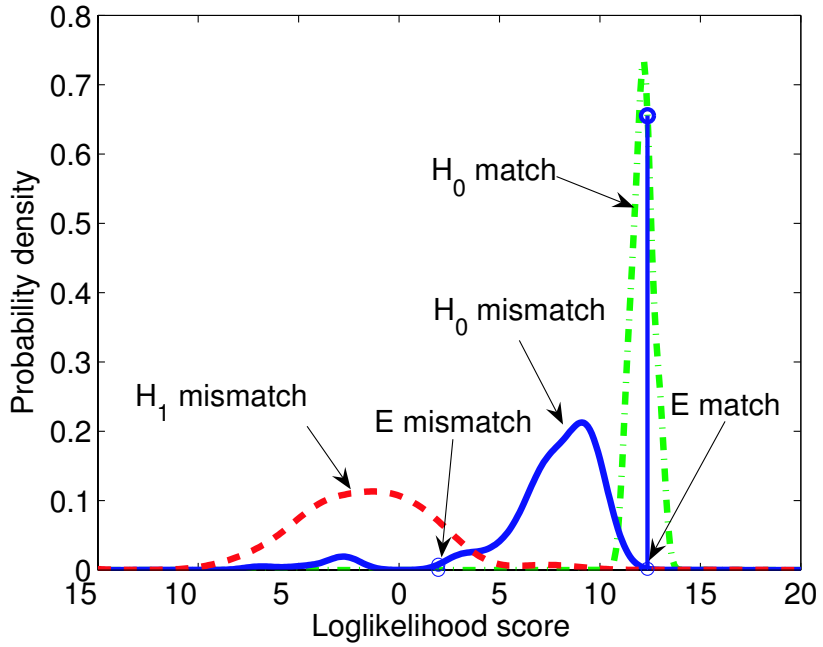


Figure 5.6: *Distribution of scores for mismatched conditions: T in PSTN and P,C,R in GSM conditions*

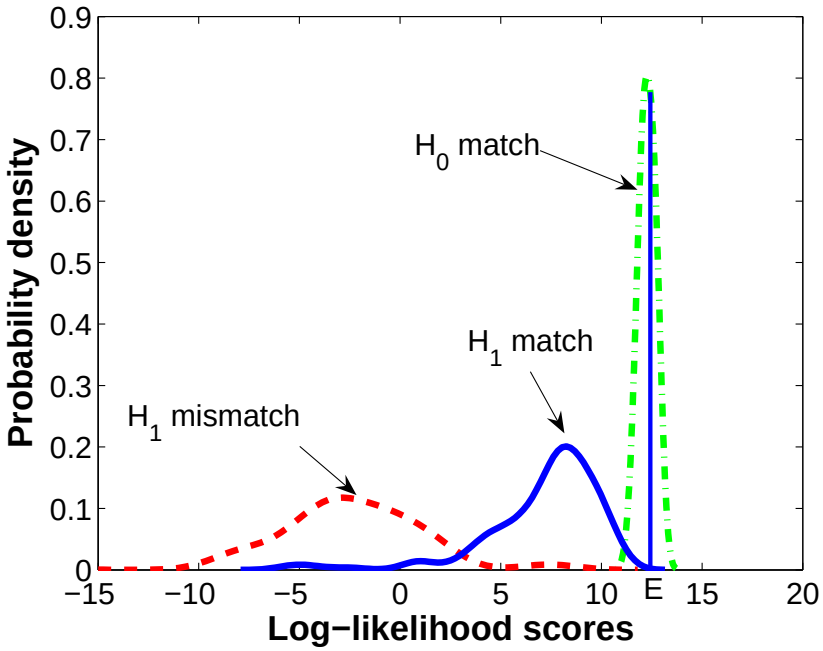


Figure 5.7: *Distribution of scores for mismatched conditions: P database in the GSM recording condition with R, C and T in PSTN conditions (all from the IPSC03 database)*

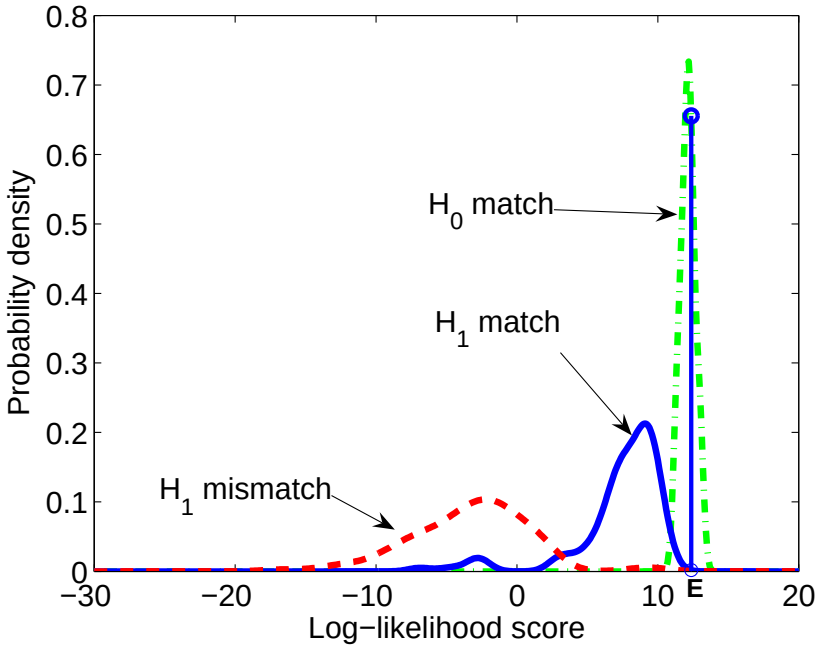


Figure 5.8: *Distribution of scores for mismatched conditions: P database in PSTN recording condition and R,C and T in GSM conditions (all from the IPSC03 database)*

under which the suspected speaker is recorded. Comparing a trace with two potential population databases, in different conditions, often gives two distinctly separated distributions of H_1-true scores. Since $P(E|H_1)$ directly depends on the distribution of H_1 scores, the denominator of the likelihood ratio depends directly on the conditions of the potential population chosen.

The likelihood ratio can shift from a position of supporting the H_0 hypothesis to a position of supporting the H_1 hypothesis, as a consequence of mismatch. This can have significant consequences if the likelihood ratios are reported without indicating the possible effects of a mismatched potential population database.

In situation #6, where P,R,C,T=C_1,C_1,C_2,C_2, both the suspected speaker control data and the questioned data are in the same conditions. Although there is mismatch in this situation, the important assumption that the trace and the control are in the same condition is satisfied. Thus, if we consider the equation for the distorted likelihood ratio, it is:

$$LR_{\leftrightarrow} = \frac{p(E_{\leftrightarrow}|H_0{\leftrightarrow})}{p(E_{\leftrightarrow}|H_1{\leftrightarrow})} \tag{5.5}$$

Situation #6 is illustrated by using an example case where H_0 is true in Fig. 5.9.

The P and R databases are in the same recording conditions (PSTN), and C and T are in the same recording conditions (GSM).

In [Meuwly, 2001], this is cited as a possible solution for when the potential population database is in different recording conditions from the trace, i.e., to record the R database in the same conditions as the P database, and the C database in the same conditions as the T database. In Fig. 5.9, the H_0 and H_1 distributions are obtained from comparisons of recordings corresponding to this scenario (mismatched conditions) as well as the distributions that would have been obtained if all the recordings had been in matched conditions. There is a shift and difference in shape for the H_0 and H_1 score distributions corresponding to matched and mismatched conditions. In mismatched conditions, while the likelihood ratio for E is still greater than 1 (2.27), we observe that there is a large overlap between the two distributions. With such overlap, it is difficult to discriminate between cases where H_0 is true and H_1 is true. The likelihood ratio obtained in these scenario corresponds to 2.27, while in matched recording conditions, the likelihood ratio should correspond to 350 (Fig. 5.9).

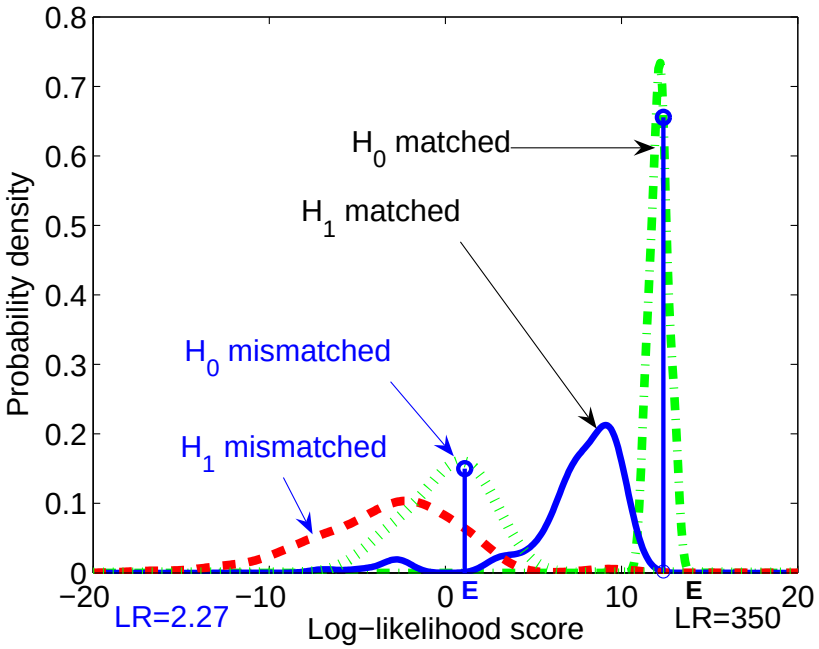


Figure 5.9: *Distribution of scores for mismatched conditions: P and R databases in PSTN recording condition and C and T in GSM conditions*

Between the different situations described, one of the most difficult problems is that faced in situation #5, where the relevant potential population is in mismatched recording conditions. An example of such a case is as follows; the police perform wiretapping of a suspect's mobile telephone, and over the course of the investigation,

they collect a certain amount of recordings of his speech using the recording equipment available to them. They also have a recording of interest made (also using their wiretapping equipment) from another mobile telephone, which they would like to confirm as belonging to the suspected speaker. Now, the forensic expert handling the analysis requires a relevant potential population database that corresponds to the case recording in language of utterance, sex of the speaker, etc. Often, he may choose available corpora of different languages to use as a potential population. However, it is very difficult to find this relevant potential population in the recording condition that he requires (in this case, GSM conditions). Thus, it may be necessary to perform the analysis using a P database that is in mismatched conditions, along with a statistical evaluation and compensation of the mismatch.

5.5 Compensation for mismatch across databases

In the previous section (Sec. 5.4.1), the effect of mismatched recording conditions on the evaluation of the likelihood ratios in the Bayesian interpretation methodology has been presented. It has to be ascertained whether the recording conditions of the databases used are mismatched and whether the expert should proceed with the analysis. As discussed earlier, if mismatched recording conditions are observed within or across databases during the analysis of a case, the expert should try to use other databases that he has at his disposal and find one that is compatible with the trace. Note that he also has the option of choosing not to proceed with analysis of the case using the Bayesian interpretation methodology or to statistically compensate for the mismatched conditions.

In order to measure the extent of change that the acoustic mismatch introduces, we can compare the trace recording not only with the speaker models in the condition of the trace, but with the smaller database that adequately represents different conditions of recording, with the same set of speakers. This database consists of recordings of speakers in several different recording conditions, in order to detect and compensate for mismatched conditions. When the trace recording is compared with this smaller database, in two different conditions, we can estimate the probability distribution for each of these sets of scores for each condition. Since the set of speakers is the same, this represents the shift in the probability distribution of scores mainly due to the recording conditions. The design of such a database, from which both the variability introduced by mismatched conditions can be estimated, is described in Sec 5.5.2.

By comparing the trace with models in the two conditions of interest, we can estimate to what extent the difference in the acoustic conditions of this database would result in a shift in the probability distribution of the scores. For instance,

in Fig. 5.5 the distributions of the scores of the comparisons of traces with two conditions (fixed telephone and microphone) are shown. We see that although these two distributions show similar variances, their means are shifted. If the potential population is in either of these conditions, a distribution scaling, as described in Section 5.5.1, can be applied to reduce the bias of the likelihood ratio.

In the following section, we consider handling mismatch at the level of the univariate distribution of scores. We compare matched and mismatched comparisons involving two recording conditions and establish means for compensating for this mismatch.

5.5.1 Estimation and compensation of mismatch using distribution scaling

The following method is used to obtain the parameters for the 'shift' in the score distributions between mismatched conditions and to apply these transformations to the score distributions calculated as a result of mismatched conditions.

Let us consider a situation in which the only potential population available for a case is recorded in a condition different from that of the suspect reference, control databases and the trace. It is not feasible to record a new potential population database in the conditions of the case, because this is expensive and time-consuming. However, the between-sources variability (H_1) distribution of scores is affected by mismatch and shifted from what it 'would have been' in matched conditions. In order to estimate what the between-sources variability distribution would have been, we propose to use a smaller sub-database that contains the same speakers, in both the conditions of the case (R, C and T databases) as well as in the condition of the potential population P. The trace is compared with the models of the two sets of speakers in these two conditions, and two distributions of scores are obtained, i.e., those obtained by comparing the questioned recording with the sub-database in matched conditions as well as in mismatched conditions. Since the same set of speakers is used in each of the databases and since their linguistic content is very similar, the 'shift' in scores is speaker-independent and depends solely on the difference in recording conditions. An estimate of this 'shift' can then be applied to the between-sources variability scores obtained with the potential population database in mismatched conditions, in order to estimate what the between-sources variability would be in matched conditions.

Basically, in order to estimate the 'shift' in the score distributions between conditions C_1 and C_2, the following steps are performed:

- Select (or record) a database which has a set of speakers in two different conditions, C_1 and C_2, with the same linguistic content and of approximately the

same duration. This database is a sub-database of a scaling database (S) which contains a set of the same speakers in different recording conditions.

- Train statistical models of the sub-databases of the scaling database (S) in both conditions (C_1 and C_2).
- Compare the features of the questioned recording with the models of the speakers in both conditions. Let the scores obtained be scores (X_1) and (X_2), for conditions C_1 and C_2 respectively.
- Estimate a distribution scaling that transforms the scores obtained in one condition to those of the second condition. The estimation of this scaling is discussed this section.
- Apply the compensation, using the transformation estimated in the previous step, to the scores' distributions that were shifted due to mismatched recording conditions.

This is illustrated for the case of a potential population database in mismatched conditions in Fig. 5.10. In order to perform this compensation, we basically need an estimate for the distribution scaling parameters as well as a database with the same speakers in different recording conditions from which these parameters can be estimated. The derivation of the scaling parameters as well as how to record a database to handle mismatch is discussed in this section.

Estimation of statistical transformations depending on the difference in conditions

In order to calculate the shift brought about due to mismatched conditions, we select two databases (S_{C_1} and S_{C_2}) containing an identical set of speakers but recorded in two different conditions. One of these databases should be in the recording conditions of the database in question, D_Q. We estimate the changes due to the difference in conditions, using this sub-database, and apply it to comparisons with the database D_Q.

Thus, it is necessary to estimate the transformation distribution of scores obtained using database D_Q, from the condition C_1 to C_2.

Let us consider the following theorem from probability theory:

Theorem 1 *If a random variable x has mean μ and variance σ^2 then the random variable $x_\star = a_1 + a_2 x$ ($a_2 \geq 0$) has the mean $\mu_\star$ and variance $\sigma_\star^2$, where $\mu_\star = a_1 + a_2\mu$ and $\sigma_\star^2 = a_2^2\sigma^2$*

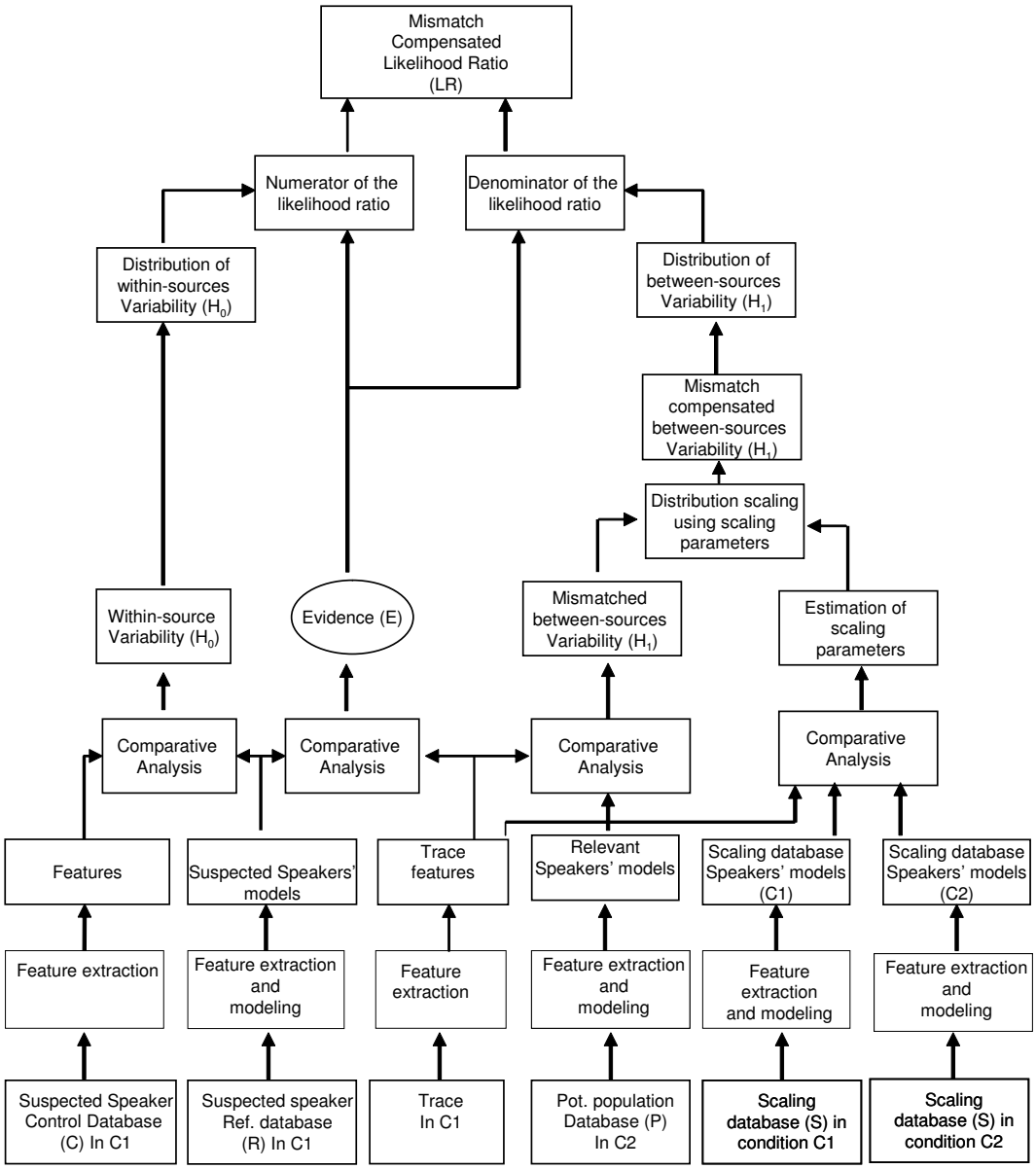


Figure 5.10: *Schema for handling a case when mismatch has been detected between the P database and the R, C and T databases*

Proof:

This theorem is proved for a continuous distribution [Kreyszig, 1999, 1076-1077]. $f(x)$ represents the probability density, and $\int_{-\infty}^{\infty} f(x)dx = 1$. From the theorem, $x_\star = a_1 + a_2 x$. Also, $f(x)dx = f_\star(x_\star)dx_\star$ for the differentials dx and $dx_\star$.

$$\begin{aligned}\mu_\star = \int_{-\infty}^{\infty} x_\star f_\star dx_\star &= \int_{-\infty}^{\infty} (a_1 + a_2 x) f(x) dx \\ &= a_1 \int_{-\infty}^{\infty} f(x)dx + a_2 \int_{-\infty}^{\infty} x f(x) dx \\ &= a_1 + a_2 \mu \end{aligned} \tag{5.6}$$

The first integral ($\int_{-\infty}^{\infty} f(x)dx$) equals one and the second integral ($\int_{-\infty}^{\infty} x f(x)dx$) equals μ.

Additionally,

$$\begin{aligned} x_\star - \mu_\star &= a_1 + a_2 x - (a_1 + a_2 \mu) \\ &= a_2 (x - \mu) \end{aligned} \tag{5.7}$$

Solving for the variances, and using Eq. 5.7 we obtain:

$$\sigma_\star^2 = \int_{-\infty}^{\infty} (x_\star - \mu_\star)^2 f_\star(x_\star) dx_\star = a_2^2 \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = a_2^2 \sigma_2^2 \tag{5.8}$$

Now, we consider the two score distributions which correspond to the scores obtained comparing a test observation with training data in two different conditions, C_1 and C_2, as X_1 and X_2, with μ_1 and σ_1 as the mean and standard deviation of the X_1 score distribution in condition C_1, and μ_2 and σ_2 as the mean and standard deviation of the X_2 score distributions in condition C_2.

A transformation of scores would estimate the change in scores from condition C_1 to C_2.

If

$$X_2 = a_1 + a_2 X_1, \tag{5.9}$$

then, from the above theorem we have,

$$\mu_2 = a_1 + a_2 \mu_1 \tag{5.10}$$

$$\sigma_2^2 = a_2^2 \sigma_1^2 \tag{5.11}$$

Solving for a_1 and a_2

$$a_1 = \mu_2 - \frac{\sigma_2}{\sigma_1}\mu_1 \tag{5.12}$$

$$a_2 = \frac{\sigma_2}{\sigma_1} \tag{5.13}$$

Thus, we have the distribution scaling equation:

$$X_2 = (X_1 - \mu_1)\frac{\sigma_2}{\sigma_1} + \mu_2 \tag{5.14}$$

For instance, if the potential population database (P) is in a condition different from the other databases in the analysis, the scores can be scaled. A scaling, from a mismatched condition to the condition corresponding to the case, is applied to the potential population scores. The mismatched potential population scores are scaled by the measure in Eq. 5.15. With this scaling, we shift the mean of the score distribution obtained when potential population is compared to the trace, towards the mean of the matched H_1 score distribution (obtained using a scaling database) and scale its standard deviation to reflect the standard deviation of the matched H_1 score distribution.

$$f(X) = \left(X - \mu_{H_{1C_1}}\right) \cdot \frac{\sigma_{H_{1C_2}}}{\sigma_{H_{1C_1}}} + \mu_{H_{1C_2}} \tag{5.15}$$

where $\mu_{H_{1C_1}}$ and $\sigma_{H_{1C_1}}$ are the mean and standard deviation of the H_1 score distribution in condition 1, $\mu_{H_{1C_2}}$ and $\sigma_{H_{1C_2}}$ are the mean and standard deviation of the H_1 score distribution in condition 2.

Let us consider the two cases presented in Fig. 5.7 (P in GSM, R,C,T in PSTN recording conditions) and Fig. 5.8 (P in PSTN, R,C,T in GSM recording conditions), and apply the distribution scaling described above.

20 speakers from the IPSC-03 database (See Appendix C), in $PSTN$ and GSM conditions, were used in order to estimate the parameters for the distribution scaling. We obtain mismatch compensated H_1 score distributions (H_1 comp. represented by a dotted line in Figs. 5.11 and 5.12). Also the H_1 score distributions that would be hypothetically obtained in matched conditions are shown. Note that in a real case where there was a mismatch in the recording conditions of the P database, it would not be possible to know what the score distribution *would be* in matched conditions. We are able to estimate this distribution by selecting both the P databases from the IPSC-03 database to simulate matched conditions and mismatched conditions. The H_1 score distribution after applying compensation for mismatch is similar to the H_1 score distribution obtained in matched conditions, in each case.

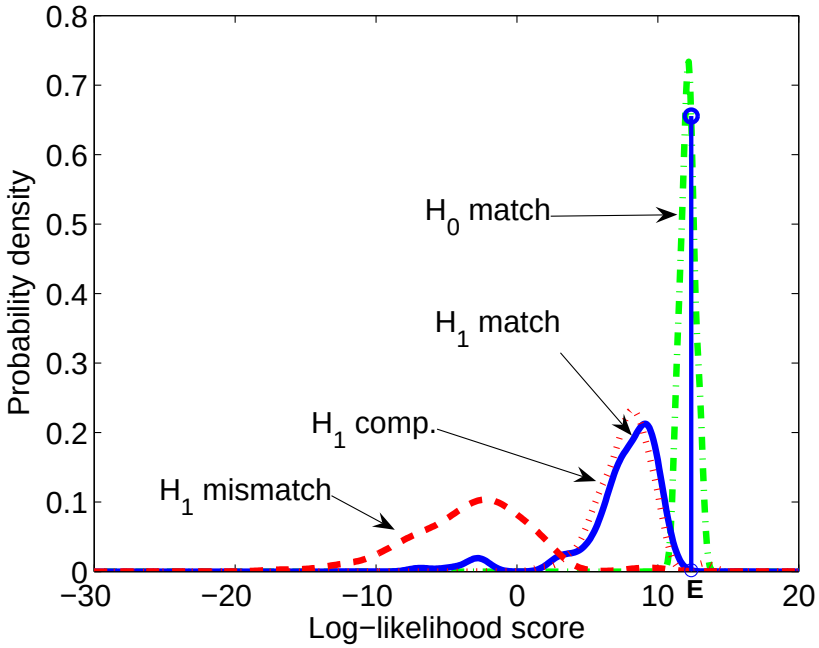


Figure 5.11: *Statistical compensation for mismatch: P in PSTN recording condition and R,C and T in GSM conditions*

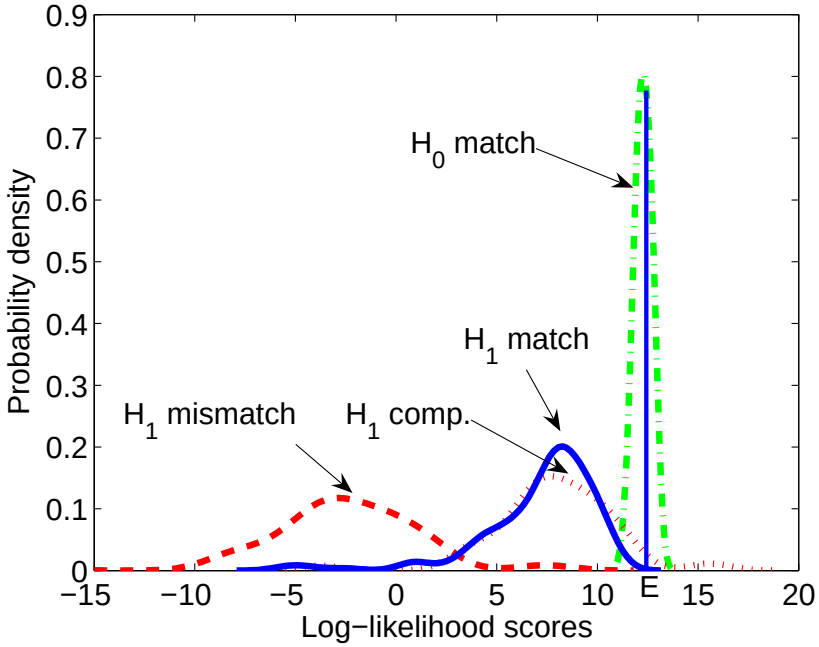


Figure 5.12: *Statistical compensation for mismatch: P in GSM recording condition and R,C and T in PSTN conditions*

A comprehensive evaluation of the statistical compensation of mismatch, using several cases and considering different recording conditions is presented in Chap. 6.

5.5.2 Methodology to construct a database to handle mismatch

As discussed in the previous section (Sec. 5.5.1), in order to perform statistical compensation, we require a database with the same speakers in different recording conditions from which these statistical compensation parameters can be estimated.

In order to record or obtain a database for forensic casework, the first consideration should be the requirements of typical case analysis. Some practical considerations involved in recording a database for forensic speaker recognition include that the recordings must,

- simulate real-world conditions closely.
- contain similar-sounding subjects as far as possible. i.e., same language and accent, preferably the same sex [Rose, 2002, :96].
- include different technical conditions (recording conditions, transmission channels).
- include a statistically significant number of speakers.
- have recordings of sufficient duration to create statistical models of speakers.

Thus, for a forensic laboratory that considers acquiring or recording a database to aid their analyses, the important considerations in investing in such a forensic database should include:

- **Language**

  The main language (or languages) in which the majority of the cases that come for analysis is an important consideration in the creation of a database. Often, a single language may not be sufficient for analysis, as modern crime spans countries and borders with ease (e.g., drug traffic, terrorism, etc.). Also, in countries where the population is multi-ethnic and has diversity in dialects and languages, there is a need to handle speaker recognition in many different languages and dialects.

- **Technical Conditions**

  The technical conditions of the audio recordings (i.e., due to recording devices or transmission channels) depend greatly on the telecommunication network

and hardware, both of which have seen significant advances in recent years. Each new technology adds to the complexity of forensic audio analysis. The technical areas that require the attention of the forensic analyst include:

- **Recording instruments**: In typical forensic audio analysis, recordings obtained from the courts can come in a variety of formats and can be recorded using different recording instruments. Audio-tapes, which are still commonly encountered in casework, are giving way to micro-cassettes and digital recorders. A modern forensic audio laboratory should be prepared to accept a number of different formats and recording media. A clear understanding of the effects of the different recording instruments and media is necessary for the analysis. A database should thus contain a representative number of recordings in each of these conditions.
- **Transmission channel**: Recent advances in telecommunications have offered a variety of possibilities for the transmission of voice, apart from the traditional PSTN network, such as mobile telephony (Global System for Mobile communication (GSM), Universal Mobile Telecommunications System (UMTS)) and Voice over Internet Protocol (VoIP). Since the speech analyzed may have been transmitted through any of the above transmission channels, in any comparison, it is necessary to determine what the transmission channels exactly are and the extent of the influence of these on the speaker recognition and other audio analysis. The forensic database should include all the different transmission channels of interest.

- **Diversity of the speakers in the databases**

  Speakers differ in the difficulty for their recognition by the automatic systems. Doddington et al. [1998] have classified speakers into goats, sheep, wolves, and lambs on the basis as discussed in Sec. 2.2.4. A large population of speakers will contain several speakers in each category, and it is necessary that the database created/selected for the recognition task should reflect this fact. It should not contain, for instance, only speakers with distinctive voices. The sample of the subjects chosen should be random enough to choose a representative sample of the population it represents.

- **Size**

  The likelihood ratio concerns itself with the proposition of the sources, namely the suspected speaker as the source of the questioned recording, and the population. Both these databases, require that the data used to model the suspected speaker as well as the population are sufficient to capture their general characteristics. The denominator of the likelihood ratio considers the hypothesis that

the questioned recording could have come from anyone else from a relevant potential population. What the relevant population is, in a case, can be decided as a function of the case conditions. It is possible for the forensic expert to consult with the court in order to determine the relevant potential population that is meaningful in a case. Similarly, it is necessary to have sufficient amount of suspect data (both reference and control data) in order to be able to estimate the variability of the speech of the suspected speaker. Thus, determining the size of each of the databases used is important in order to address the various 'propositions of sources'.

- **Requirements from the Bayesian interpretation methodology**

  The Bayesian interpretation methodology requires, in addition to the trace, the use of three databases: a suspect reference database (R) a suspect control (C) and a potential population database (P) as well as a database of traces T when performance of the system is being evaluated. Thus, for use in the Bayesian interpretation framework, ideally, this database must contain a suspect reference (R) and control (C) databases as well as trace database (T) for *every* speaker. The potential population database (P) can be constituted by using the R databases of all the individual speakers in the database.

These conditions are of primary concern in creating/selecting the hypothetical forensic speaker recognition database. Ideally, the forensic laboratory must have a database containing a large number of speakers (sufficient to represent a potential population) recorded in a wide variety of conditions. In practice, however, such a database is not easily obtained. There have been efforts to record a large database in several different languages, in various channel conditions, by the United States Secret Service (USSS) and the Federal Bureau of Investigation (FBI) [Beck et al., 2004]. It is possible, however, to build a collection of smaller databases, along with the potential population database, corresponding to different case conditions. This collection of databases will contain a considerably smaller number of speakers than the potential population but will be sufficiently large to derive statistics to compensate for mismatched conditions.

A prototype forensic speaker recognition database

In this section, we propose a methodology for the creation of a database consisting of recordings of speakers in several different forensically relevant conditions, in order to detect and compensate for mismatched conditions.

The steps in recording this database include:

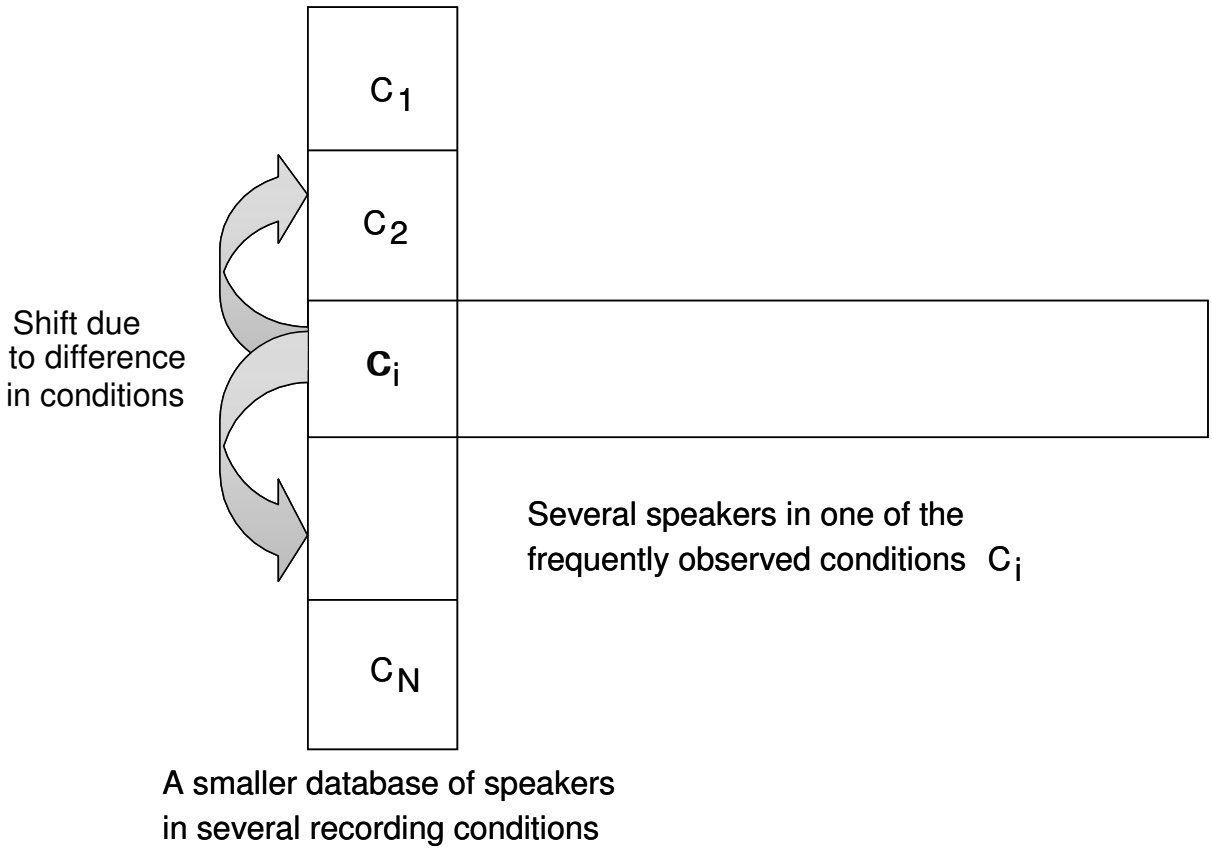


Figure 5.13: *Schema of a forensic database to handle mismatch*

- Record one or more databases in the most commonly encountered recording condition, which contains a sufficient number of speakers that can serve as a potential population database.

- Using a subset of speakers of this first database, record several smaller databases in different recording conditions which contain a sufficient number of speakers from which statistics for scaling can be calculated.

Thus we have a 'T'-shaped database Fig. 5.13, with a greater number of speakers in a single 'base' condition (let this be the most common condition) and several smaller 'mismatch-compensation databases', in each of the individual conditions. A schema of the use of such a scaling database in handling mismatch is illustrated in Fig. 5.10.

A forensic speaker recognition database was created at the Institut de Police Scientifique, University of Lausanne, and the Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne as a prototype of a speaker recognition database created according to the methodology described above. This database includes 73

male speakers in three different recording conditions, i.e., PSTN, GSM and a calling-room acoustic recording using digital recorder. These recordings are of controlled and uncontrolled speaking modes, made in controlled conditions, in a quiet room, and consist of over 4800 audio files totalling over 40 hours. It contains forensically realistic questioned recordings, in the form of prompted and freely made threatening telephone calls, using the same telephone instruments and networks for each condition, for the every speaker, to minimize the effects of individual handsets as well as those of digital recordings at the source and after having passed through a transmission channel. This forensically realistic database can be used for aural, instrumental and automatic speaker recognition in order to estimate and compensate for mismatch in recording conditions.

Deciding when the data in a database is 'sufficient'

The choice of a potential population database is central to Bayesian interpretation of evidence. As discussed in the previous section, both the size and the recording conditions of the suspected speaker control database and the potential population database are important factors that have to be considered in order to evaluate the source-level propositions correctly.

This choice, necessitates the notion of *data sufficiency*, i.e., how much data is necessary before we can conclude that the characteristics of the individual or the population have been adequately represented.

Let us consider a characteristic feature that is of interest to the case analysis. This characteristic can be any distinguishing feature of the individual or the population (e.g., height, weight, salary etc.). In order to decide that we have sampled a sufficient number of individuals to have a general idea about the characteristics of the population, it would be necessary that whatever subset of database is chosen, the difference between the characteristics of that subset and the characteristics of the theoretical relevant population is minimal.

In order to do this, let us consider the characteristics of such a hypothetical database, that is truly representative of the population. If it is indeed representative of the features of the population, then adding one more speaker to it will not significantly change the general characteristics of the features observed. In other words, if it is decided that the features of N individuals are sufficiently representative of a population, then adding one more individual to the set, i.e., $N+1$ should not significantly change the properties of the features observed. Similarly, in order to construct this database, we consider an incremental strategy, where the number of individuals represented in the database is increased one by one, until the characteristics of the features considered do not change significantly. Examples of how the number of speakers in the scaling database can be determined, by considering the 'relative

change' for an increase in the number of speakers in the scaling database, is presented in Sec. 6.1.5.

Distribution scaling and the Gaussian assumption

Both the distribution scaling presented as well as the various score normalization techniques (presented in Sec. 2.3.3) use the means and standard deviations of the score distributions as parameters, and therefore rely on the Gaussian assumption. It is necessary to closely observe the data in order to ascertain whether it is indeed Gaussian, and sufficient data has been considered in estimation of these parameters.

It is possible to validate how well the mismatch-compensated score distributions are representative of what the distributions would have been in matched conditions, by using data that is in both conditions. Using a test-bed of identical speakers in two different conditions, varying only in the mismatched condition of interest, it is possible to calculate the real shift of the score distribution due to conditions, as well as the hypothesized mismatch-compensated score distribution. Tests such as the divergence test as well as the *Bhattacharya distance* can be used to perform this analysis.

The measure should consider the shift in means as well as the change in the distributions. The *Bhattacharya distance* satisfies both these criteria and is the upper bound on the Bayesian error between two normally distributed classes. In an ideal case, i.e., if the mismatch-compensated distribution corresponds exactly to the real distribution, the *Bhattacharya* distance will go to 0.

For normal probability density functions (pdf), the *Bhattacharya distance* between the distribution of scores in the first class and the second class corresponds to:

$$d_B^2 = \frac{1}{2}\ln\frac{|\frac{C_1+C_2}{2}|}{|C_1|^{\frac{1}{2}}|C_2|^{\frac{1}{2}}} + \frac{1}{8}(\mu_1-\mu_2)^T\left(\frac{C_1+C_2}{2}\right)^{-1}(\mu_1-\mu_2) \tag{5.16}$$

This distance is indicative of how well the mismatch-compensated score distribution reflects the real conditions. However, what is of greater interest to forensic automatic speaker recognition is the effect the scaling has on the strength of evidence or the likelihood ratio. We need to know if the hypothesized estimate of the likelihood ratio is close to the likelihood ratio that could be obtained in matched conditions. An evaluation of the performance of the statistical compensation presented in this chapter, using mock cases constructed from the IPSC03 database, is discussed in Chap. 6.

5.6 Handling a case

Consider that a forensic expert has received a recording of a suspect and a questioned recording for which the court would like to determine the source. Let us also assume that for this case, the expert has a cooperative suspect of whose voice a sufficiently long recording can be made. This allows the expert to create models and estimate the within-source variability in the suspect's voice. This assumption is consistent with those required by the Bayesian interpretation framework proposed in [Meuwly and Drygajlo, 2001]. It is also assumed that it is possible to determine the conditions in which each of the databases is recorded. For example, a case in which the suspected speaker reference (R) as well as control recordings (C) are known to have been made through a fixed (PSTN) line, but the questioned recording (T) and the relevant potential population (P) have been recorded in cellular (GSM) and fixed (PSTN) telephone conditions. Although this information may not always be available to the expert, it is possible to detect or even request the court or the police to provide this information. A schema of how a case, in which mismatch has been detected between the P database (which is in recording condition C2) and the R, C and T databases (which are in recording condition C1), is handled, is presented in Fig. 5.14.

The steps involved in handling such a case (illustrated in Fig. 5.14) are as follows:

- Obtain the conditions of recording of each of the databases, either by requesting for this information or by identifying them automatically.
- Ascertain whether significant mismatch exists within or across databases, by comparing recordings from each of the databases with each other. Statistical significance testing can be used to check whether the score distributions thus obtained are similar at a certain level of significance.
- If a mismatch is detected between the databases, then
  - select another, more compatible potential population database, or
  - investigate the possibility of recording the suspect in conditions compatible with the databases available, and
  - if neither of the above options are possible, either decide not to analyze the case using the Bayesian interpretation framework or apply statistical compensation to the mismatched conditions' scores.
- The following steps are carried out if statistical compensation of the mismatched conditions is to be performed:
  - If a mismatched comparison involves a comparison between two recordings in two conditions (say trace features in condition C_1 and potential

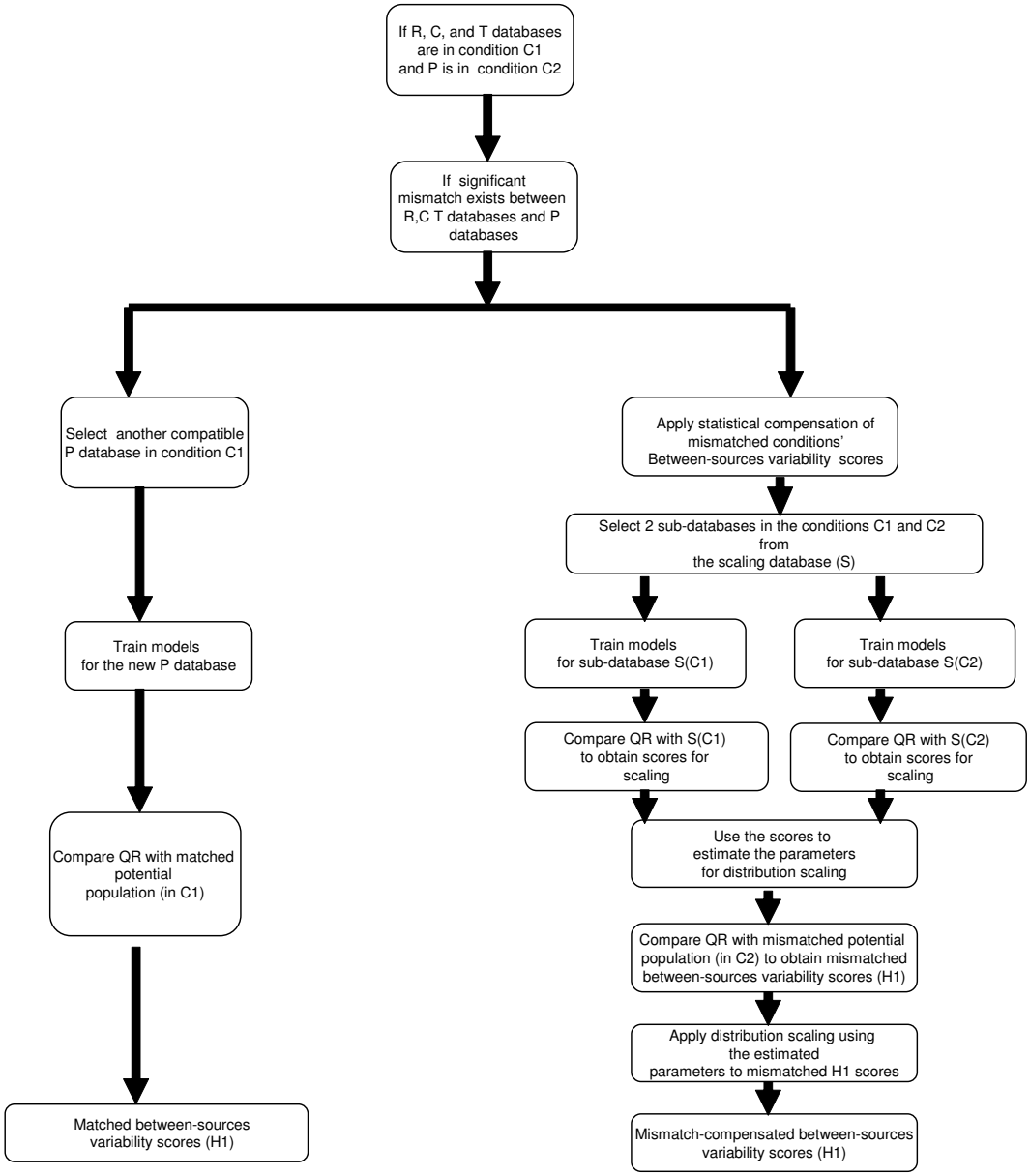


Figure 5.14: *Detecting and compensating for mismatch between P database and the R, C and T databases*

population models in condition C_2), choose two sub-databases of the scaling database which contain speech of the same speakers, with the same linguistic content and manner of speaking, in recording conditions C_1 and C_2.

– Compare the trace recordings in condition C_1 with the subsets of the scaling database in conditions C_1 and C_2 for two set of scores that will be obtained in the mock matched and mismatched conditions respectively.

– Estimate the distribution scaling parameters from the scores in mock matched and mismatched conditions, as described in Sec. 2.3.3.

– Using the statistics of H_1-true distributions in the database, in the two conditions, apply the scaling presented in Eq. 5.15 (illustrated in Fig. 5.14) to the potential population scores in the Bayesian interpretation framework and derive the compensated likelihood ratios.

Evaluation of statistical compensation techniques

6

In Chapter 5, the detection of mismatch in the recording conditions of the databases used and statistical compensation techniques for handling mismatched conditions, using a scaling database, have been presented. In this chapter, we present an evaluation of the statistical compensation method presented. The evaluation of the method is discussed in two sections using data from:

- The IPSC-03 database (Appendix C) which contains recordings of 73 speakers including the relevant suspect reference (R), suspect control (C) and trace (T) databases for each speaker in 3 different recording conditions,

- The Netherlands Forensic Institute (NFI) speaker recognition evaluation through a fake case (2005) along with the IPSC-03 database and the PolyCOST 250 database [Hennebert et al., 2000].

In Section 6.1 of the evaluation, over 2700 mock cases, each using R, C, P, and T databases selected from the IPSC-03 database, that simulate cases in matched and mismatched recording conditions are created. Mismatched conditions are simulated for each mock case by choosing databases containing the same speakers in recording conditions other than that of the mock case. The number of speakers in the scaling database that is sufficient to estimate compensation parameters is estimated. Statistical compensation of the mismatched score distributions is performed using the parameters calculated using the scaling database (S), and the likelihood ratios obtained after compensation are evaluated using cumulative probability density plots called Tippett plots.

In Section 6.2 of this evaluation, in addition to the suspect control (C) and reference (R) recordings extracted from the ten cases that are part of the NFI speaker recognition evaluation, three sets of P databases are selected from the IPSC-03 database (in PSTN, GSM and room acoustic conditions) to simulate matched and mismatched recording conditions. A comparison is made with the results obtained using these databases and original evaluation results for which the PolyCOST database in $PSTN$ conditions had been used as a potential population database in matched conditions [Hennebert et al., 2000].

6.1 Evaluation using simulated case examples from the IPSC-03 database

The IPSC-03 database was recorded by the Institut de Police Scientifique (IPS), University of Lausanne, and the Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne (EPFL), between January and June 2005. This database was recorded in collaboration with Zimmermann [2005] from the Institut de Police Scientifique. It contains speech of 73 male speakers in three different recording conditions and several different controlled and uncontrolled speaking modes. The speakers are males, aged between 18 and 50, and are university educated, speaking French. The majority of the speakers are undergraduate and graduate students of between 18 and 30, who have similar educational backgrounds. The recording conditions of this database include speech transmitted through a public switched telephone network (PSTN), global system for mobile communications (GSM) network, as well as calling room acoustic recordings using a digital recorder. Details about this database are presented in Appendix C.

Using subsets of the IPSC-03 database, a set of mock cases has been constructed to evaluate the influence of mismatched recording conditions on the likelihood ratio and the effect of applying statistical compensation for mismatch. 150 cases where H_0 is known to be true and 150 cases H_1 is known to be true are thus constructed. In order to construct the 150 cases, 50 different speakers were used, and 3 cases were considered per speaker. For each of these mock cases, four databases are required, namely, the suspected speaker reference database (R), the suspected speaker control database (C), the potential population database (P) and the trace database (T). In addition to these four databases, we also use a set of 20 speakers as a scaling database, from which the compensation for mismatch can be estimated.

For each speaker in the IPSC-03 database, we have:

- An R database consisting of two recordings of approximately 2-3 minutes duration comprising read speech and spontaneous responses to questions.

- A C database consisting of three recordings of 15-30 seconds comprising spontaneous speech, read speech and a simulated dialogue.

- A T database consisting of three recordings of 15-30 seconds comprising simulated threats (read from cue cards).

For each case, subsets of the IPSC-03 database were chosen to represent the P, R, C databases and the trace (T). A potential population database (P) consisting of 50 speakers was selected from the IPSC-03 database. For each of the 50 speakers, two files, similar in content to the R database, of 2-3 minutes length, were used to create the P database. As the entire IPSC-03 database consists of 73 speakers, for every mock case, 50 speakers (other than the suspected speaker) are chosen as the potential population. In order to simulate mismatch, the same set of 50 speakers was selected in each recording condition ($PSTN$, GSM and room acoustic). For instance, for a mock case in which all the databases were in PSTN conditions (for the matched case), mismatch would be simulated by using the same set of 50 speakers who are used as the potential population in the matched case, and using their recordings in GSM or in the room acoustic conditions.

The scaling database (S) consists of 20 speakers in three different conditions, i.e., PSTN, GSM and acoustic-room recording. For each speaker, in the scaling database, we have two files which are recorded exactly like the reference databases (2-3 minutes, read and spontaneous responses to questions) in each recording condition. These files are used to create two models for this speaker per condition.

The partitioning of the IPSC-03 database for the experiments is shown in Fig. 6.1.

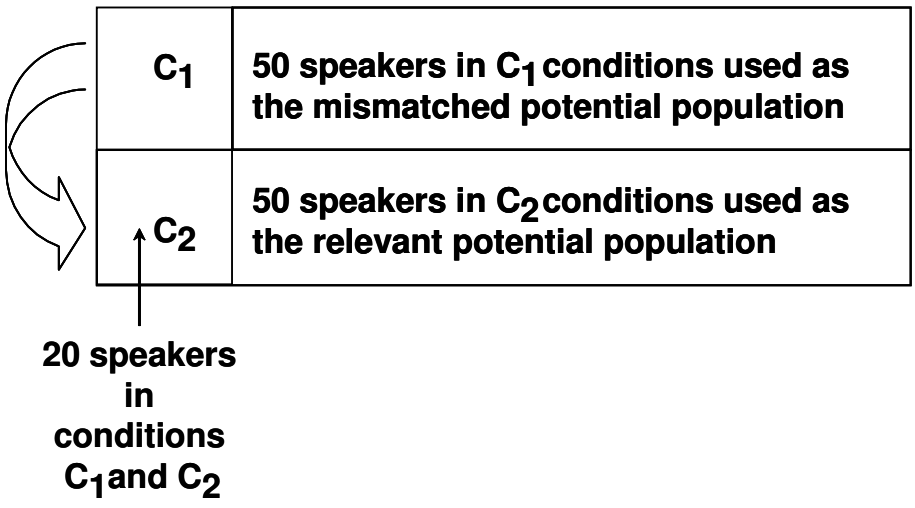


Figure 6.1: *Partitioning of the IPSC-03 database for the evaluation*

6.1.1 Evaluation of the strength of evidence in matched recording conditions

First, let us consider the performance of the forensic automatic speaker recognition system in matched recording conditions. This allows us to evaluate what, ideally, should be the performance of the system, in controlled conditions, where there is no mismatch. The effect of the three different recording conditions on the strength of evidence can also be evaluated.

For this evaluation, we created 150 cases where H_0 is true, i.e., the suspected speaker is indeed the source of the questioned recording, and 150 cases where H_1 is true, i.e., the suspected speaker is not the source of the questioned recording. We then evaluated the LRs for each case and plotted the cumulative density of the LRs using Tippett plots. The separation of the two curves is a good indication of the performance of the system. In the Bayesian interpretation framework, the likelihood ratio of 1 is important, as it represents the turning point between the support for one hypothesis and the other. The point where $LR = 1$ can serve as a reference to evaluate the proportion of cases where H_0 was true, which had an LR greater than 1 (should be close to 100% ideally), and the proportion of cases where H_1 was true, which had an LR greater than 1 (should be close to 0% ideally).

Performance in matched recording conditions using $PSTN$ recordings

The Tippett plot for matched conditions using mock cases in $PSTN$ conditions is shown in Fig. 6.2. In matched recording conditions, 97.3% of the cases corresponding to H_0 true (intercept of the $LR = 1$ with the H_0 curve) obtain an LR greater than 1. (Note that in the figure, the x-axis has a logarithmic scale, and this $LR = 1$ is represented by 10^0.)Similarly, 6.5% of the cases corresponding to H_1 true (intersection of the $LR = 1$ with the H_1 curve) obtain a likelihood ratio above 1. The LRs obtained for H_1 true, range from close to 0 to 6. The LRs obtained for H_0 true, range between 1 and 10^9.

Performance in matched recording conditions using GSM recordings

The Tippett plot for matched conditions using mock cases, in GSM conditions, is shown in Fig. 6.3. In matched recording conditions, 93.3% of the cases corresponding to H_0 true obtain an LR greater than 1. Similarly, 6.6% of the cases corresponding to H_1 true obtain a likelihood ratio above 1. The LRs obtained for H_1 true, range from close to 0 to 24. The LRs obtained for H_0 true, range between 1 and 10^8.

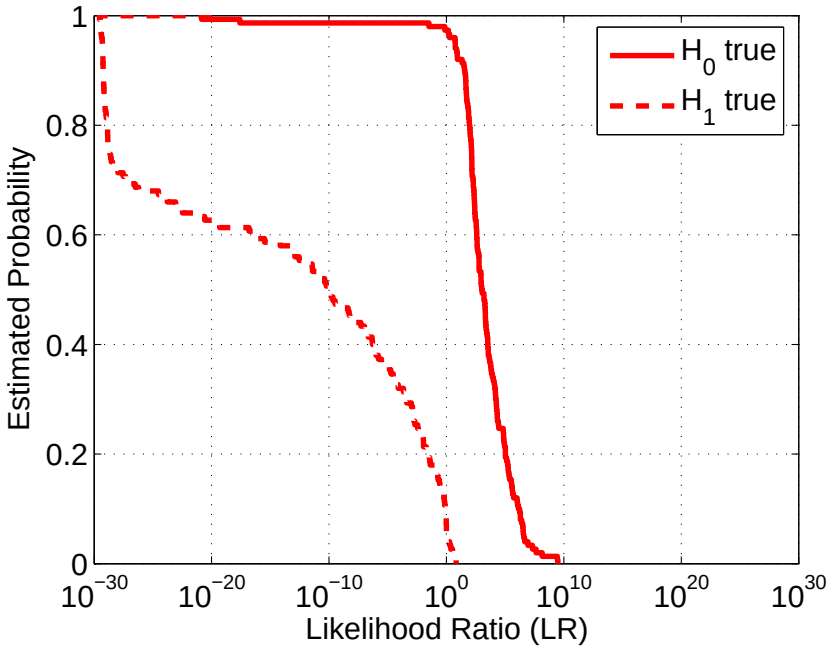


Figure 6.2: *Tippett plot for matched conditions (PSTN-PSTN)*

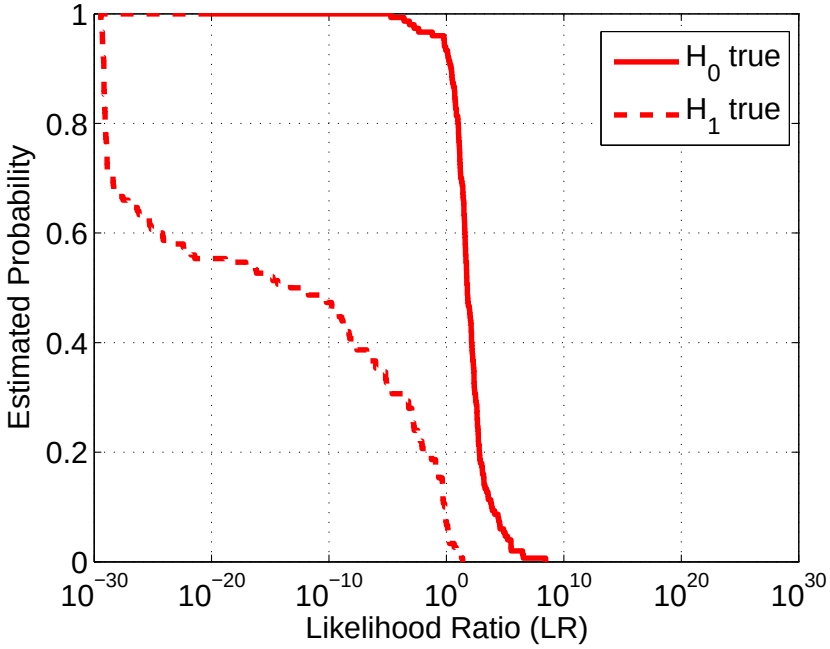


Figure 6.3: *Tippett plot for matched conditions (GSM-GSM)*

Performance in matched recording conditions using room acoustic recordings

The Tippett plot for matched conditions using mock cases, in room acoustic conditions, is shown in Fig. 6.4. In matched recording conditions, 91.3% of the cases

corresponding to H_0 true, obtain an LR greater than 1. Similarly, 3.4% of the cases corresponding to H_1 true obtain a likelihood ratio above 1. The LRs obtained for H_1 true, range from close to 0 to 3. The LRs obtained for H_0 true, range between 1 and 10^{17}.

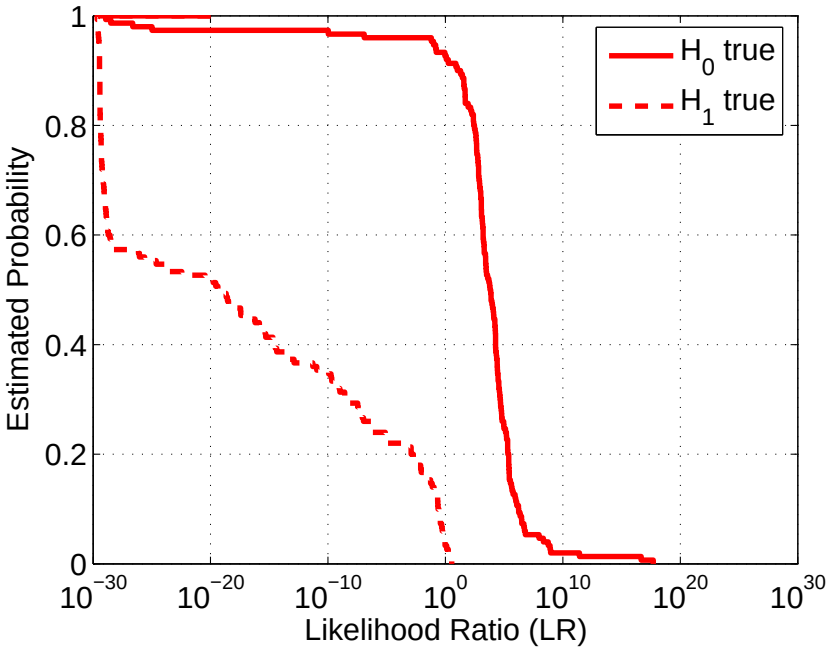


Figure 6.4: *Tippett plot for matched conditions (Room acoustic -Room acoustic)*

Performances in matched conditions

In general, as we have seen, the performance of the forensic automatic speaker recognition system in matched conditions is very good, with a good separation between the H_0 and H_1 curves, and with a small percentage of cases where the LRs are contrary to the ground truth of the case. We observe that while the performances in matched conditions are comparable across the three conditions, the best performances (using an LR of 1 as a reference point), correspond to the PSTN condition (if we consider the sum of the proportion of cases corresponding to H_1 true, with an LR greater than 1, and the proportion of cases corresponding to H_0 true, with an LR less than 1). While in room-acoustic conditions, there is only a small proportion of cases (3%) corresponding to H_1 true that obtain a likelihood ratio above 1, a bigger proportion of cases (9%) obtain a likelihood ratio above 1 for H_0 true.

6.1.2 Evaluation of the strength of evidence in mismatched recording conditions

Let us now consider the performance of the automatic speaker recognition system in mismatched recording conditions. The effect of using databases recorded in three different conditions ($PSTN$,GSM, and room acoustic) on the strength of evidence can be evaluated. The different 'pairs' of mismatched conditions are evaluated and compared.

With the three different recording conditions for the databases, there are six possible combinations in which mismatch can occur, comparing:

- R, C and T databases in $PSTN$ and P database in GSM or room acoustic recording conditions.
- R, C and T databases in GSM and P database in $PSTN$ or room acoustic recording conditions.
- R, C and T databases in room acoustic recording and P database in $PSTN$ or GSM recording conditions.

For each of these combinations, we construct 300 cases, consisting of 150 cases where H_0 is true and 150 cases where H_1 is true.

Performance in mismatched recording conditions using mock cases in $PSTN$ conditions

Consider the situation in which the R, C and T databases are recorded in PSTN conditions, and the P database is mismatched, in room acoustic conditions. The Tippett plot for the 300 mock cases, evaluated when the mismatched potential population is recorded in room acoustic conditions, is given in Fig. 6.5. In mismatched recording conditions, 100% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 53% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1. This implies that more than half of the cases in which H_0 are known to be true, would be wrongly evaluated with an LR greater than 1.

Similarly, the Tippett plot for the 300 mock cases, evaluated when the mismatched potential population is in GSM, is shown in Fig. 6.6. In mismatched recording conditions, 96.6% of the cases corresponding to H_0 true obtain an LR greater than 1, and 17% of the cases corresponding to H_1 true have obtained a likelihood ratio above 1. Although the effect of mismatch is less than in the case where the P database was in room acoustic conditions, the proportion of cases wrongly evaluated to have an LR greater than 1 is still high.

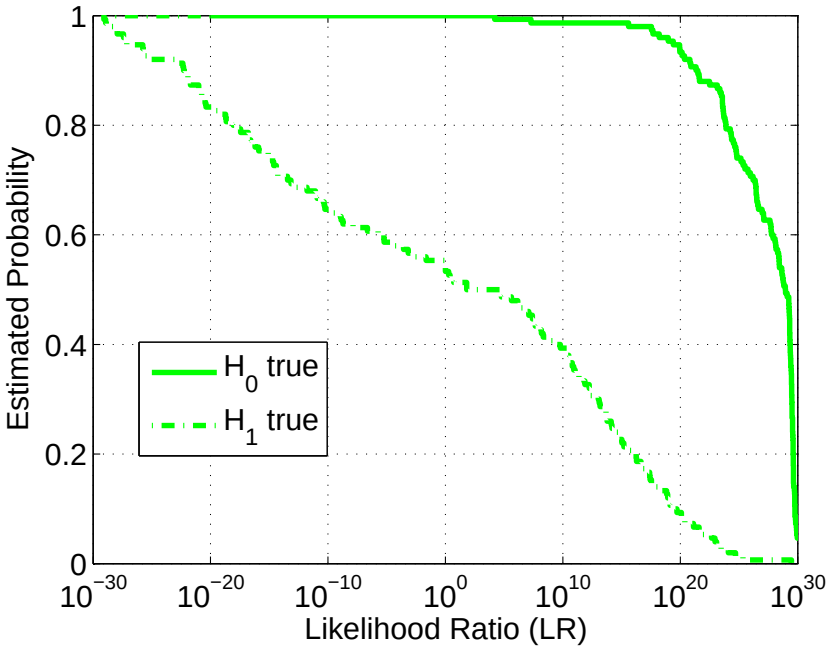


Figure 6.5: *Tippett plot: R,C,T in PSTN conditions, P in room acoustic conditions*

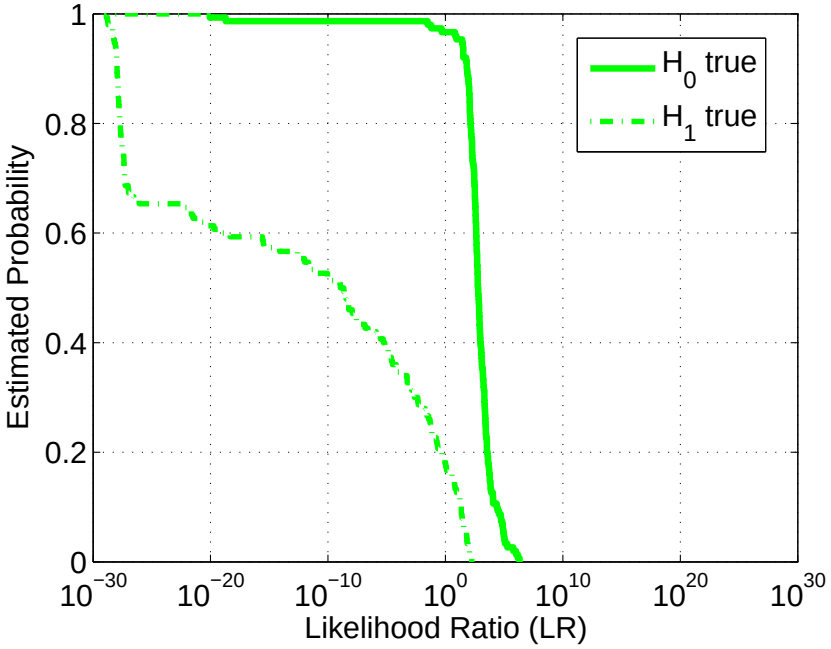


Figure 6.6: *Tippett plot: R,C,T in PSTN conditions, P in GSM conditions*

Performance in mismatched recording conditions using mock cases in *GSM* conditions

Consider the case in which the R, C and T databases are in GSM conditions, and the P database is mismatched, in room acoustic conditions. The Tippett plot for

the 300 mock cases, evaluated when the mismatched potential population is in room acoustic conditions, is shown in Fig. 6.7. In mismatched recording conditions, 100% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 56% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1. Again, this implies that more than half of the cases, in which H_0 is known to be true, would be wrongly evaluated with an LR greater than 1.

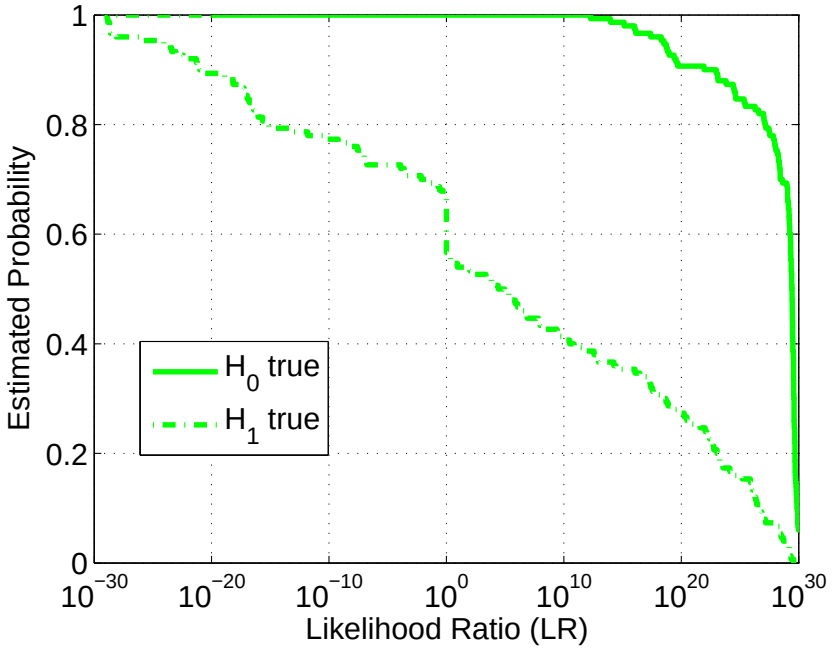


Figure 6.7: *Tippett plot: R,C,T in GSM conditions, P in room acoustic conditions*

Similarly, the Tippett plots for the 300 mock cases, evaluated when the mismatched potential population is in PSTN conditions is given in Fig. 6.8. In mismatched recording conditions, 97.3% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 20.6% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1. While the effect of mismatch is less than in the case where the P database was in room acoustic conditions, the proportion of cases wrongly evaluated to have an LR greater than 1 is still high.

Performance in mismatched recording conditions using mock cases in room acoustic conditions

Finally, we consider the situation in which the R, C and T databases are in room acoustic conditions, and the P database is mismatched, in $PSTN$ recording conditions. The Tippett plot for the 300 mock cases, evaluated when the mismatched

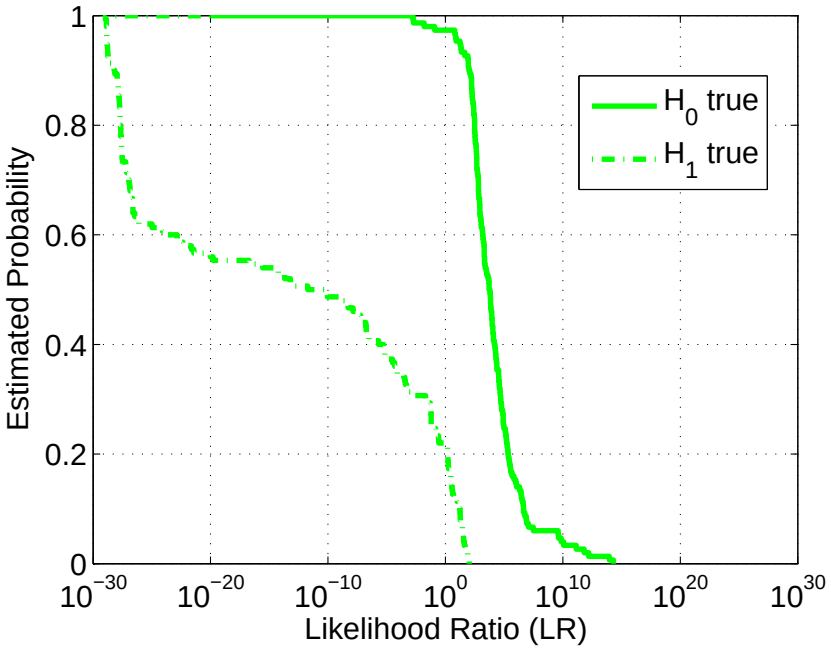


Figure 6.8: *Tippett plot: R,C,T in GSM conditions, P in PSTN conditions*

potential population is in room acoustic conditions, is given in Fig. 6.9. In mismatched recording conditions, 97.3% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 36% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1.

The corresponding Tippett plot for the 300 mock cases, evaluated when the mismatched potential population is in GSM conditions, is given in Fig. 6.10. In mismatched recording conditions, 97.5% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 30% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1. For both these cases, about a third of H_1 cases is wrongly evaluated, with LRs above 1, which is quite high.

6.1.3 Performances in mismatched conditions

Thus, in general, the performance of the forensic automatic speaker recognition system in mismatched conditions is poor, especially when compared to its performance in matched conditions. There is a good separation between the H_0 and H_1 curves, but the percentage of cases where the LRs are contrary to the ground truth of the case is quite high, ranging from 20% to over 50%. It can also be observed that the Tippett plots obtained due to the same pair of mismatched conditions are similar (e.g., the Tippett plot for R,C,T in PSTN and P in GSM is similar to the Tippett plot for R,C,T in GSM and P in PSTN).

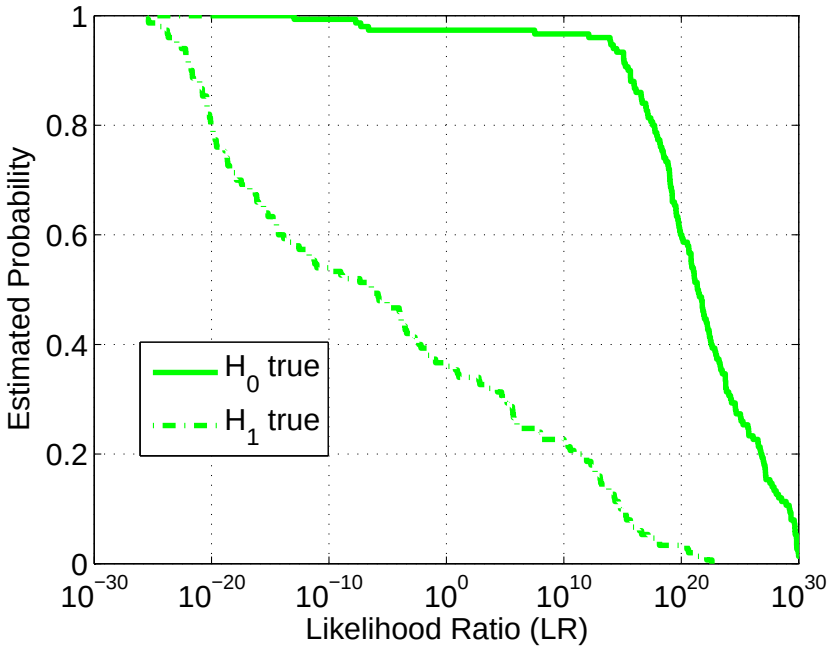


Figure 6.9: *Tippett plot: R,C,T in room acoustic conditions, P in PSTN conditions*

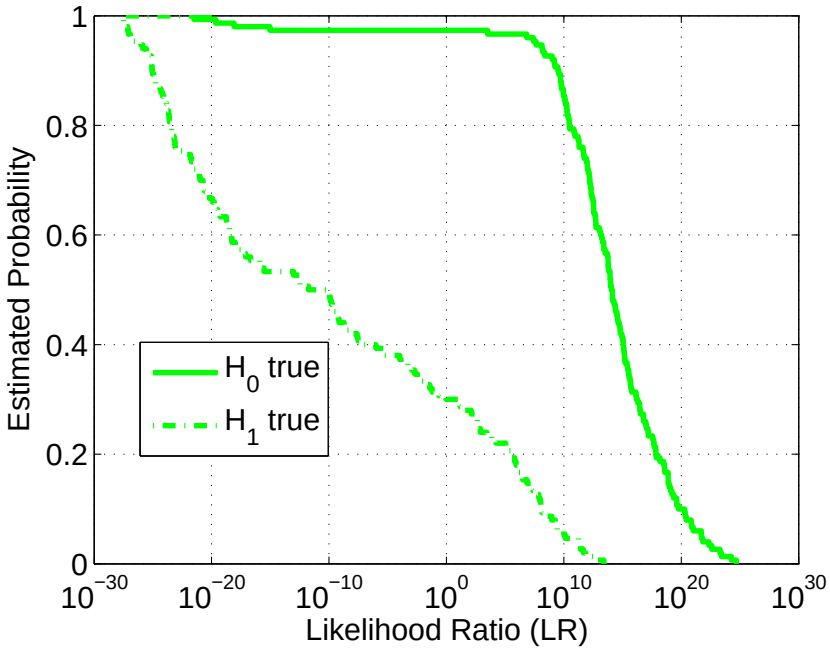


Figure 6.10: *Tippett plot: R,C,T in room acoustic conditions, P in GSM conditions*

For all conditions of mismatch, most of the cases corresponding to H_0 true have LRs greater than 1, which means that only few of H_0 cases have an LR that is wrongly evaluated, but this is overshadowed by the fact that a large number of cases corresponding to H_1 true also have a likelihood ratio greater than 1. Both the LRs corresponding to H_0 and H_1 true, are greater than the corresponding LRs for H_0 and H_1 true in matched conditions. However, for H_1 true cases, this can mean that the LRs go above 1, which consequently implies that the system would wrongly conclude that, given the evidence, it is more likely that the suspected speaker is the source of the questioned recording than any other speaker in the population, when this is not the case. Arguably, judicially fatal errors could result, if the effects of mismatch are not considered.

6.1.4 Statistical compensation for mismatched conditions

So far, we have observed that in matched conditions, the automatic system performs well, and the LRs obtained are largely compatible with the reality or 'ground truth' of the case. In mismatched conditions, the LRs obtained are often not compatible with the ground truth. Hence, statistical compensation techniques should be able to compensate for the effects of mismatched conditions and the LRs obtained after compensation must be more compatible with the ground truth.

To analyze the effect of compensation for mismatched conditions, we use a scaling database. As illustrated in Fig. 6.1, the IPSC-03 database was partitioned into 50 speakers as a potential population database and 20 speakers as the scaling database for mismatch compensation. From the IPSC-03 database, we extract the PSTN, GSM and acoustic recordings made with a digital recorder, as the three databases that could be used as simulated potential population databases. All the three sub-databases contain the same set of speakers, speaking French, recorded in three conditions. The likelihood ratios for mismatched conditions, as well as the compensation for mismatch, are estimated for each of the recordings. Other databases (such as the PolyCOST, the IPSC-01 and the IPSC-02 databases) were not used for scaling, because they either do not share the same set of speakers in different conditions or do not have the number of speakers as required for this evaluation exercise. Unless the same set of speakers is chosen, the differences in the LR, cannot solely be attributed to the difference in conditions but also due to the difference between the speakers. By using the same set of speakers, whose speech has approximately the same linguistic content and duration as that in the scaling database, the differences can be considered to be mainly due to the recording conditions.

6.1.5 Determining the size of the scaling database (S)

The scaling database was used to estimate the parameters for compensation which are used in Eq. 5.15. The scaling database contains the same speakers in several different conditions and is much smaller than the P database.

It is necessary to determine how many speakers are sufficient in order to estimate the 'shift' in the distribution parameters due to the differences in recording conditions. Larger scaling databases would evidently give better estimates of the shift (it is of interest to determine how many speakers, at least, it should contain in order to reliably estimate the compensation) and would give likelihood ratios that are similar to those that can be obtained in matched conditions. In order to do this, we consider the likelihood ratios estimated in a single case and increase the size of the scaling database gradually, observing the effect of this on the likelihood ratio. We also consider the effect of increasing the size of the scaling database for several cases, and observe the change in the global performance with the increasing scaling database size.

Size of the scaling database and its effect on the LR

We consider two cases, one in which the suspect was indeed the source of the trace (H_0), and the other in which the suspect was not the source of the trace (H_1). For each of these cases, mismatch is introduced by choosing the P database (GSM) in conditions different to the R, C and T databases (PSTN).

The scaling database (S) is chosen with the same speakers, in the two conditions in the case, i.e., GSM and PSTN. The trace is compared to the recordings of the speakers in the S database, and the scaling parameters are estimated. Compensation is then applied to the (mismatched) scores that were obtained comparing the trace with the P database, giving a new estimate of the between-sources score variation, and a new likelihood ratio is estimated. The number of speakers in the S database is gradually increased from one speaker to the size of the entire IPSC-03 database (which is used as a potential population), and the evolution of LR is observed. The effect of increasing the size of the S database, for the two cases, is shown in Figs. 6.11 and 6.12 respectively.

For the case where H_0 is known true (Fig. 6.11), we observe that the mismatched LR (32000) is much higher than the corresponding LR in matched conditions (114). With a small increase in the number of speakers in the scaling database, we observe a sharp decrease in the LR from the mismatched LR and close to the LR in matched conditions (50 to 70). While the compensated LR does not converge to the matched LR, the relative change in the LRs is small even with just 10 speakers. It is to be noted that while the compensation resulted in an estimate of LR more representative

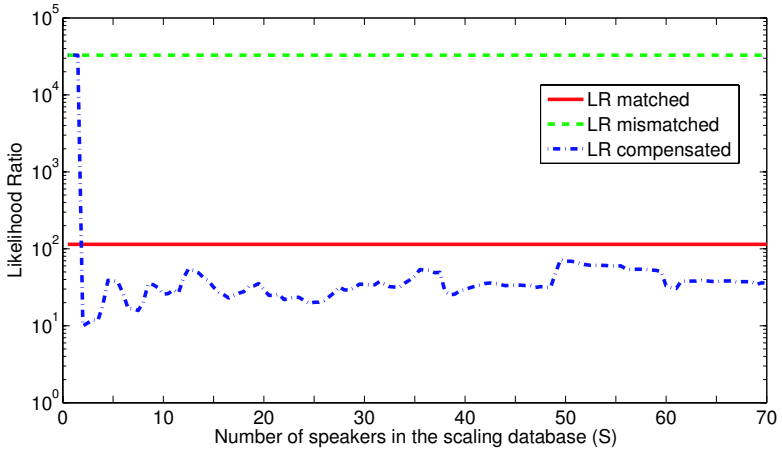


Figure 6.11: *The variation in the compensated likelihood ratio for increasing number of speakers of the scaling database (S) where H_0 is true*

of a matched LR, there is still a certain error in its estimation.

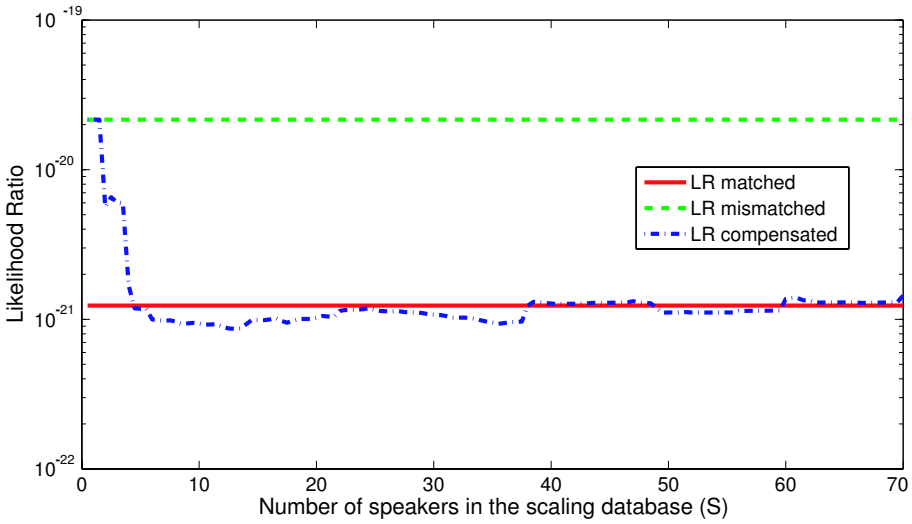


Figure 6.12: *The variation in the compensated likelihood ratio for increasing number of speakers of the scaling database (S) where H_1 is true*

For the case where H_1 is known true (Fig. 6.12), similarly, we observe that the mismatched LR (10^{-19}) is much higher than the corresponding LR in matched conditions (10^{-21}). As in the earlier case, as we increase in the number of speakers in the scaling database, a sharp decrease in the LR from the mismatched LR and close to the LR in matched conditions (10^{-21}) can be observed. Again, while the compensated LR does not converge to a particular value, the relative change in the LRs is small even with just 10 speakers.

Thus, in considering the two example cases for H_0 true and H_1 true, increasing the size of the scaling database beyond 10-20 speakers does not significantly change the estimates of the LR. It is necessary, however, to perform this test over all the mock cases created in order to determine globally how many speakers would be sufficient to reliably estimate the parameters for the compensation and beyond which the estimates of the compensated LR do not change significantly.

Size of the scaling database and its effect on the global performance of the system

The likelihood ratio of 1, which is the turning point for support of one hypothesis over the competing hypothesis, should ideally have a minimum proportion of H_1 cases and a maximum proportion of H_0 obtaining LRs above it. By choosing the reference point of an LR of 1, we determine how many speakers would be necessary for compensation such that the proportion of LRs corresponding to H_0 true and H_1 true, with LRs above 1, after compensation, are similar to those obtained in matched conditions.

In order to estimate the number of speakers that would be sufficient in the scaling database, across several cases, we use a set of 300 cases created using 50 speakers, and we use a scaling database (S) whose size is gradually increased.

In Fig. 6.13, the proportion of LRs greater than 1, where the recording condition of the cases is GSM and the mismatched potential population is in PSTN, is shown. We gradually increase the size of the scaling database (S) from 1 till 20.

In Figs. 6.13 and 6.14, we observe that the proportion of cases where H_1 is true and the LRs were higher than 1, drops from the mismatched conditions close to the proportion of cases in matched conditions very rapidly, even when only 6-10 speakers are considered in the S database.

Similarly, the proportion of cases where H_0 is true and the likelihoods were higher than 1 drops from the mismatched conditions to close to the proportion of cases in matched conditions very rapidly, even after only 6-10 speakers are considered in the S database.

Further, we consider pairs of mismatched conditions and observe whether a certain pair of conditions would require more data than the other. In Figs. 6.15, 6.16, 6.17 and 6.18 and we observe that even with a change in the pair of mismatched conditions under consideration, the compensation can be considered to have a stable effect on the LRs for around 20 speakers.

Considering each of the mismatched conditions above and the size of the S database, we have decided to use 20 speakers as the scaling database and 50 speakers for the mismatched potential population database.

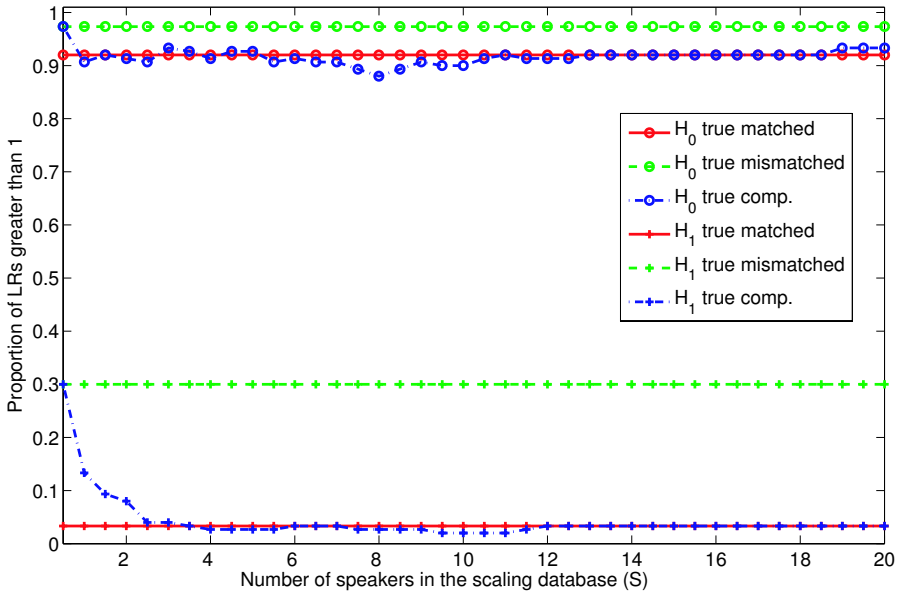


Figure 6.13: *The proportion of LRs greater than 1 for each of the hypotheses (R,C,T in room acoustic conditions, P in GSM conditions)*

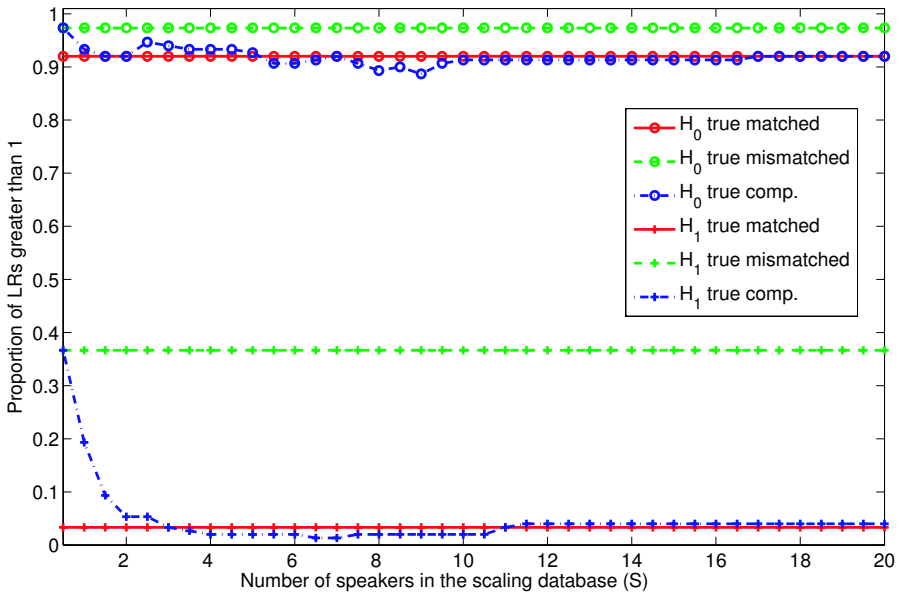


Figure 6.14: *The proportion of LRs greater than 1 for each of the hypotheses (R,C,T in room acoustic conditions, P in PSTN conditions)*

Applying statistical compensation using the estimated scaling parameters

Mock cases in *PSTN* conditions

Let us first consider the case in which the R, C and T databases are in $PSTN$ conditions, and the P database is mismatched and recorded in room acoustic or in

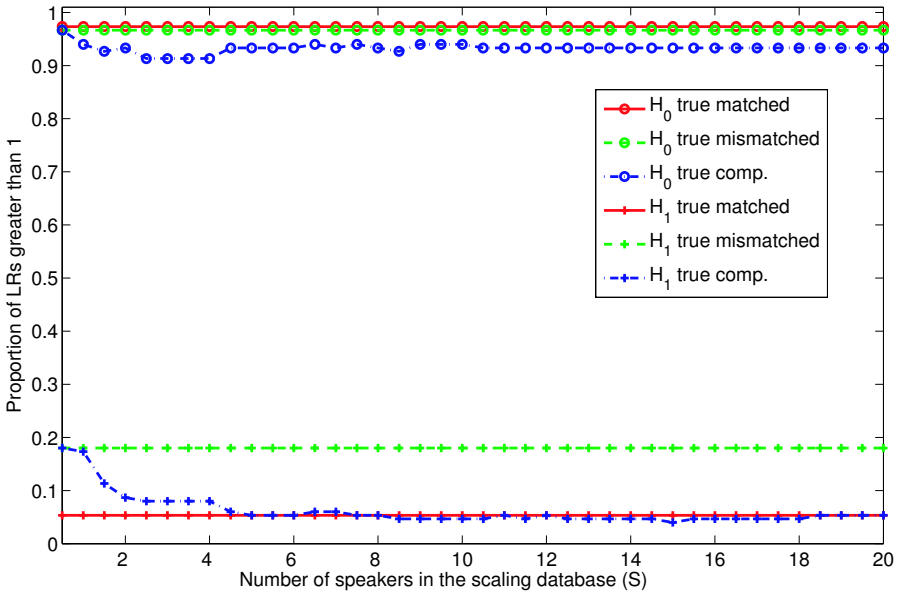


Figure 6.15: *The proportion of LRs greater than 1 for each of the hypotheses (R,C,T in PSTN conditions, P in GSM conditions)*

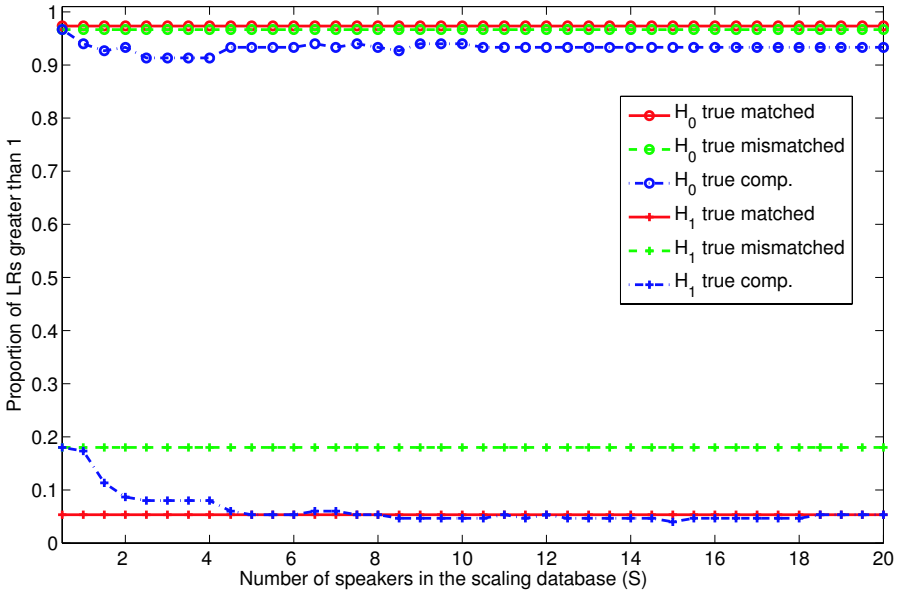


Figure 6.16: *The proportion of LRs greater than 1 for each of the hypotheses (R,C,T in PSTN conditions, P in room acoustic conditions)*

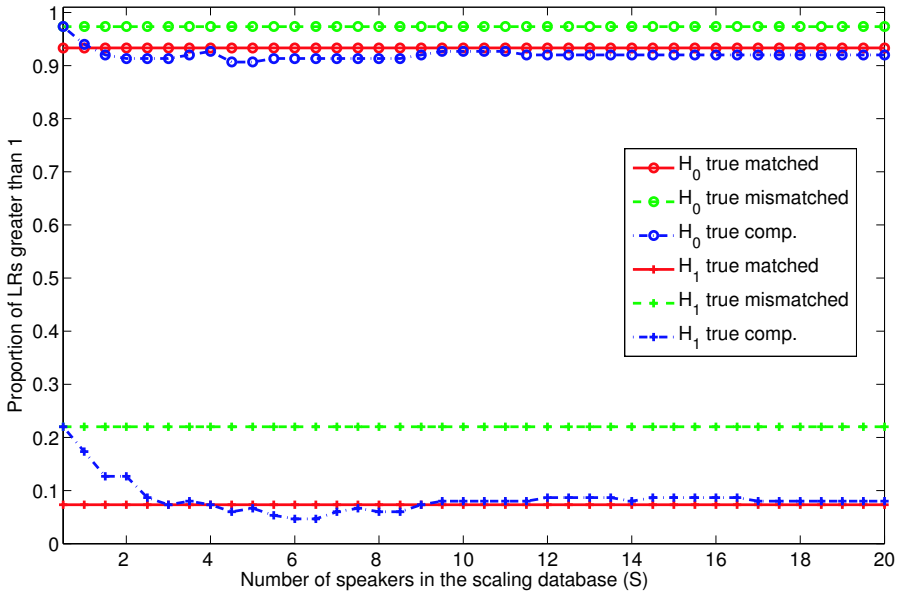


Figure 6.17: *The proportion of LRs greater than 1 for each of the hypotheses (R,C,T in GSM conditions, P in PSTN conditions)*

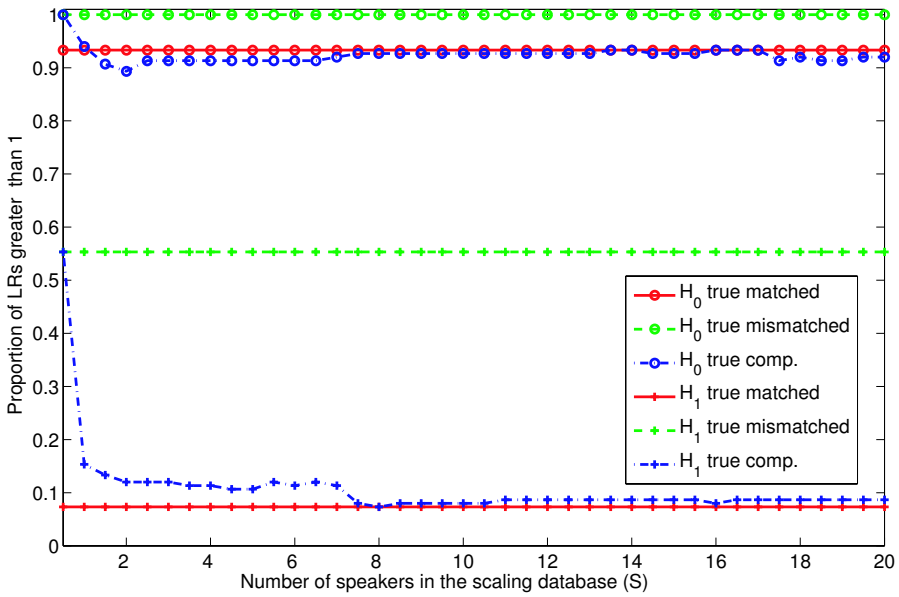


Figure 6.18: *The proportion of LRs greater than 1 for each of the hypotheses (R,C,T in GSM conditions, P in room acoustic conditions)*

GSM conditions:

1. *Room acoustic conditions*: The Tippett plot for the 300 mock cases evaluated, when the mismatched potential population is room acoustic conditions, is given in Fig. 6.19. Before compensation was applied, 100% of the cases corresponding to H_0 true, obtained an LR greater than 1, and 53% of the cases corresponding to H_1 true, obtained a likelihood ratio above 1. After compensation for mismatch, 95.5% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 4% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1.

2. *GSM conditions*: Similarly, the Tippett plot for the 300 mock cases, evaluated when the mismatched potential population is in GSM, is given in Fig. 6.20. Before compensation was applied, 96.6% of the cases corresponding to H_0 true, obtained an LR greater than 1, and 17% of the cases corresponding to H_1 true, obtained a likelihood ratio above 1. After compensation for mismatch, 93.3% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 7% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1.

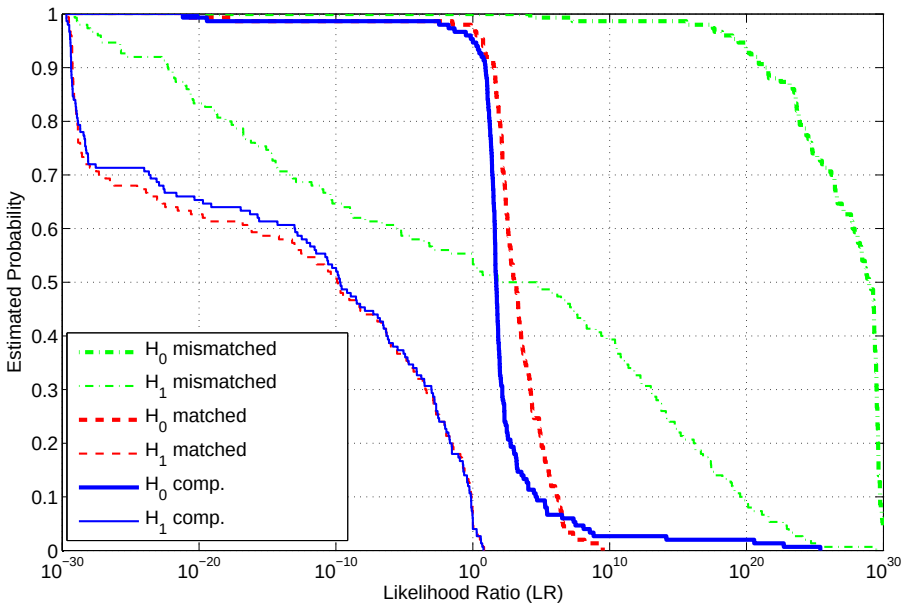


Figure 6.19: *Tippett plot after statistical compensation: R,C,T in PSTN conditions, P in room acoustic conditions*

Mock cases in *GSM* conditions

Consider the case in which the R, C and T databases are recorded in GSM conditions, and the P database is mismatched, and recorded in room acoustic conditions or PSTN

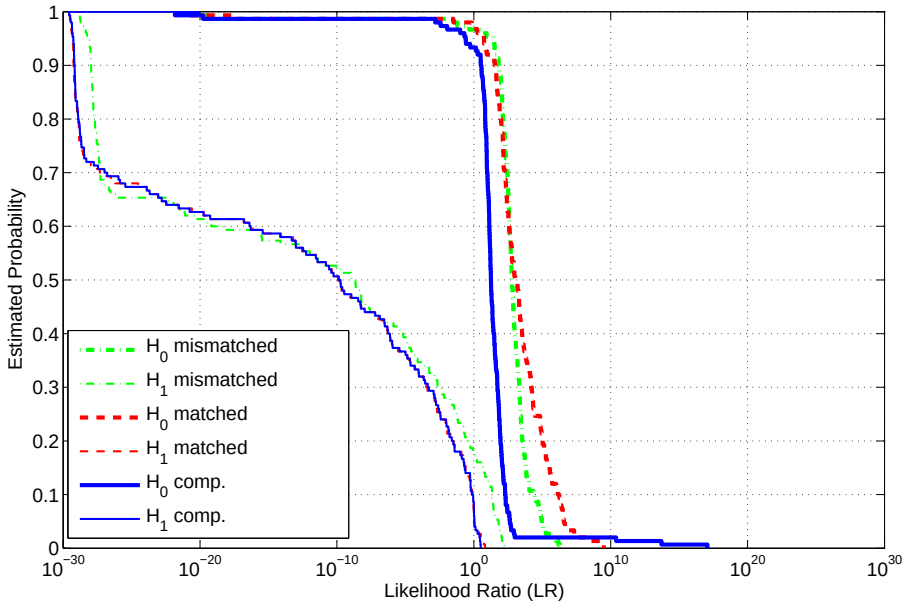


Figure 6.20: *Tippett plot after statistical compensation: R,C,T in PSTN conditions, P in GSM conditions*

conditions.

1. *Room acoustic conditions*: The Tippett plot for the 300 mock cases, evaluated when the mismatched potential population is room acoustic conditions, is given in Fig. 6.21. Before compensation was applied, 100% of the cases corresponding to H_0 true, obtained an LR greater than 1, and 56% of the cases corresponding to H_1 true, obtained a likelihood ratio above 1. After compensation for mismatch, 90.6% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 8.6% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1.

2. *PSTN conditions*: Similarly, the Tippett plot for the 300 mock cases, evaluated when the mismatched potential population is PSTN conditions, is given in Fig. 6.22. Before compensation was applied, 97.3% of the cases corresponding to H_0 true, obtained an LR greater than 1, and 20.6% of the cases corresponding to H_1 true, obtained a likelihood ratio above 1. After compensation for mismatch, 91.3% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 8% of the cases corresponding to H_1 true obtain a likelihood ratio above 1.

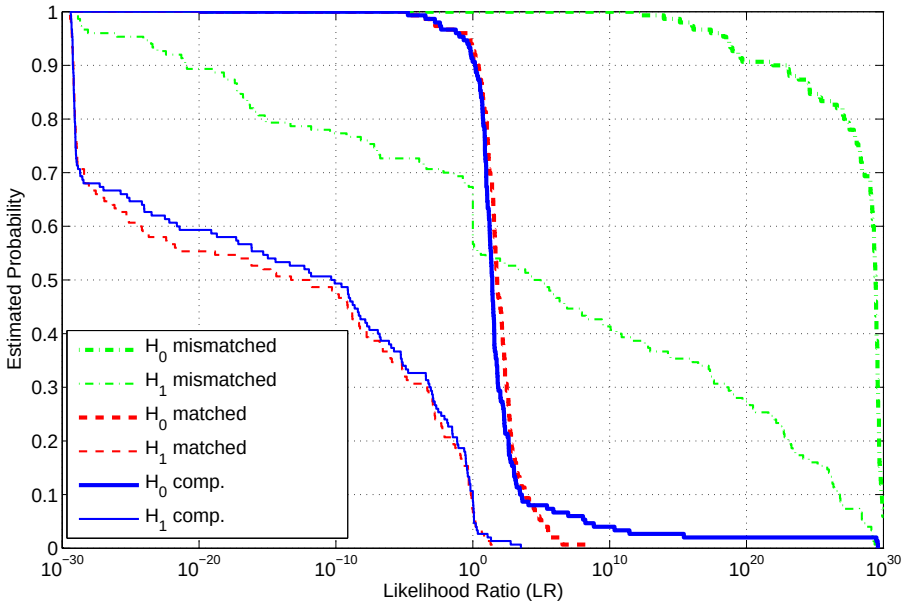


Figure 6.21: *Tippett plot after statistical compensation: R,C,T in GSM conditions, P in room acoustic conditions*

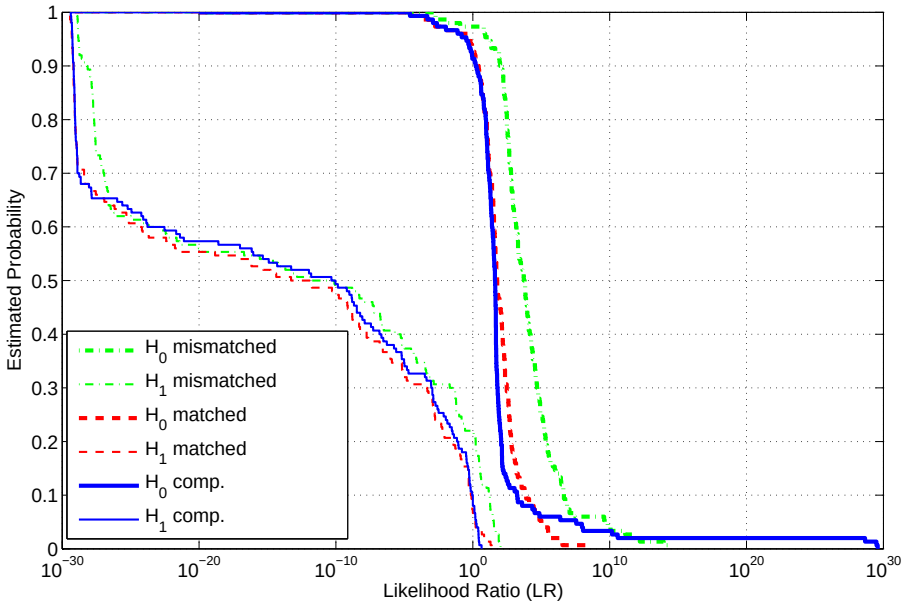


Figure 6.22: *Tippett plot after statistical compensation: R,C,T in GSM conditions, P in PSTN conditions*

Mock cases in room acoustic conditions

Finally, considering the case in which the R, C and T databases are in room-acoustic conditions, and the P database is mismatched, and in GSM or PSTN conditions.

1. *PSTN conditions*:The Tippett plots For the 300 mock cases, evaluated when the mismatched potential population is PSTN conditions is given in Fig. 6.23. Before compensation was applied, 97.3% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 36% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1. After compensation for mismatch, 92% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 4% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1.

2. *GSM conditions*: The corresponding Tippett plot for the 300 mock cases, evaluated when the mismatched potential population is GSM conditions, is given in Fig. 6.24. Before compensation was applied, 97.5% of the cases corresponding to H_0 true, obtained an LR greater than 1, and 30% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1. After compensation for mismatch, 93.4% of the cases corresponding to H_0 true, obtain an LR greater than 1, and 3.3% of the cases corresponding to H_1 true, obtain a likelihood ratio above 1.

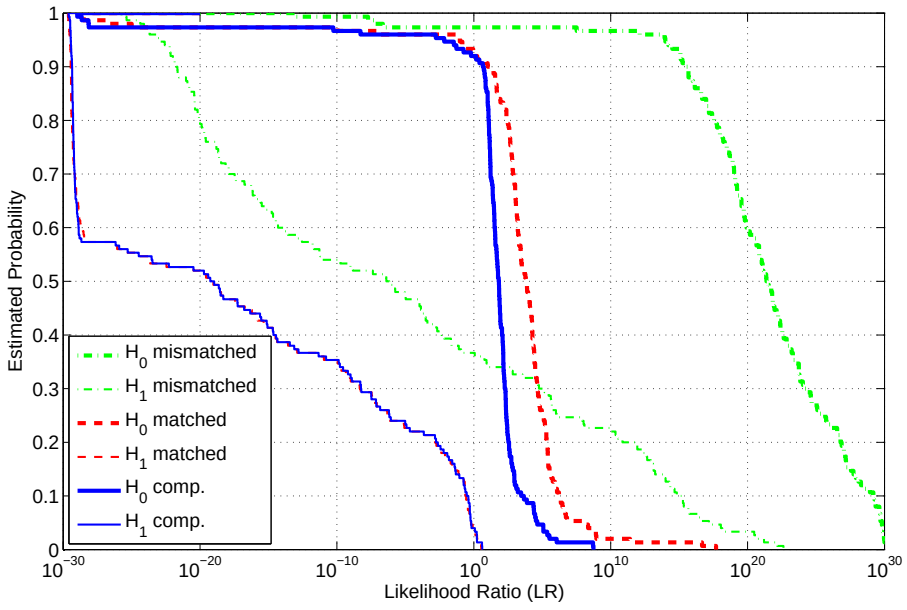


Figure 6.23: *Tippett plot after statistical compensation: R,C,T in room acoustic conditions, P in PSTN conditions*

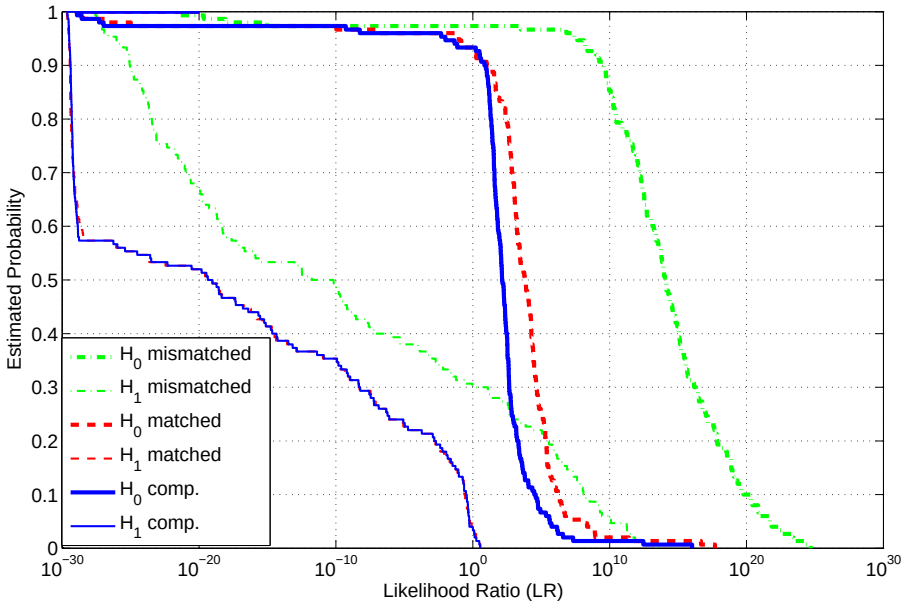


Figure 6.24: *Tippett plot after statistical compensation: R,C,T in room acoustic conditions, P in GSM conditions*

6.1.6 Comparison with recording suspect reference and control databases in conditions of the potential population and trace databases

Let us consider another set of mock cases in which the trace database (T) is in GSM conditions and the potential population database (P), available to the expert for this case is in PSTN conditions. This can happen if the expert receives a questioned recording which is in GSM conditions, and it is possible for him to record the suspect in recording conditions of his choice. However, for reasons such as language, sex of the speaker, etc., he observes that he has a relevant potential population database that is in only in $PSTN$ conditions. The expert then has two choices:

1. To record the suspect reference (R) in the same conditions as the P database and control (C) databases in the same conditions as the questioned recording (GSM).

2. To record the suspect reference (R) and control (C) databases in the same conditions as the questioned recording (GSM), and to apply statistical compensation using the scaling database.

We evaluate these two possibilities, by creating several cases similar to this using the IPSC-03 database. The Tippett plot for this comparison is illustrated in Fig.

6.25.

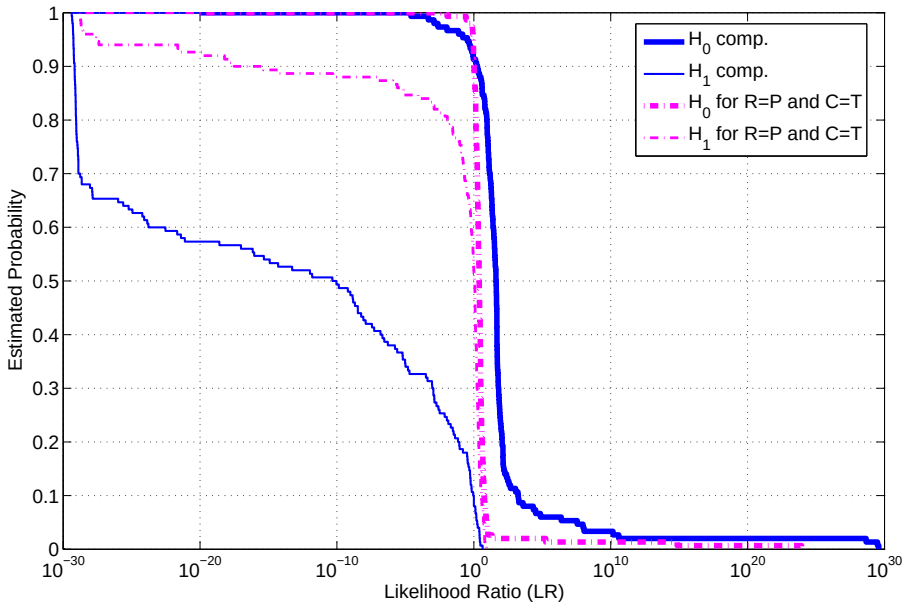


Figure 6.25: *Comparison of handling mismatch using statistical compensation and by recording R in the same conditions as P (PSTN), and C in the same conditions as T (GSM)*

We observe that the separation between the mismatch-compensated H_0 and H_1 curves is greater than the separation between the H_0 and H_1 curves when the suspect reference and control databases were recorded in similar conditions to the potential population and trace respectively. If the expert had instead chosen to record the suspect reference and control databases in the same conditions as the questioned recording, and applied statistical compensation for the mismatched conditions, this would correspond to Fig. 6.22. We observe that in this experiment, it is better for the expert to record the R and C databases in the same conditions as the questioned recording T (GSM), and to use the P database in $PSTN$ conditions as well the scaling database in order to compensate for the mismatch.

6.2 NFI speaker recognition evaluation through a fake case

In this section, we use data from the Netherlands Forensic Institute (NFI) Speaker Recognition Evaluation (using a fake case) and from the IPSC-03 database. This forensic speaker recognition evaluation was conducted in 2004-2005 in order to com-

pare the methods used by different forensic institutes belonging to the European Network of Forensic Science Institutes (ENFSI).

Twelve audio recordings were provided by the NFI as part of a fake case evaluation, consisting of 2 reference recordings and 10 questioned recordings. The 10 questioned recordings consist of conversations between two speakers, i.e., each containing the speech of a known speaker and an unknown speaker. The 2 reference recordings consist of two conversations between a known speaker and a suspected speaker. The aim of the test was to determine whether the speech of the unknown speaker, in each of the questioned recordings, was produced by the suspected speaker in the reference recordings. The ground truth of the evaluation was subsequently released by the NFI [Cambier-Langeveld, 2005].

In the original conditions of the fake case evaluation, there was no explicit mismatch in the conditions, i.e., the transmission channel condition for all the recordings was PSTN. The potential population database (P) used in the original evaluation was a subset of the PolyCOST database [Hennebert et al., 2000] containing 73 European speakers, recorded in PSTN conditions, speaking in English. This, however, does not consider possible mismatches due to differences in handsets and recording instruments within the PolyCOST potential population database.

In this section, we will introduce mismatched conditions into the analysis by choosing potential population databases that are mismatched in channel and language. For a detailed analysis and report of this case, in *matched* conditions, using the PolyCOST database, refer to the [Appendix D]. For this part of the analysis we use the IPSC-03 database, and evaluate how mismatch would affect LRs obtained, and the effect of compensating for mismatch on the outcome of the evaluation. The suspected speaker reference database (R) and the suspected speaker control database (C) were extracted from the NFI fake case evaluation data recorded PSTN conditions. The LRs due to match, mismatch and after compensation will be viewed in the light of the ground truth released by the NFI.

The questioned recordings in each sub-case have been named Q1 through to Q10. The ground truth was as follows:

- In the cases corresponding to recordings Q2,Q3, Q4, Q5, Q6, Q7 and Q9, the suspected speaker was indeed the source of the questioned recording.
- In the cases corresponding to recordings Q1, Q8 and Q10, the suspected speaker was not the source of the questioned recording.

Using a potential population recorded in GSM conditions

In this part of the evaluation, we use the speech of 50 speakers from the IPSC-03 database, recorded in GSM conditions, as the mismatched potential population

database. The recordings of 20 other speakers, in GSM and $PSTN$ conditions, were used as the scaling database.

A summary of the LRs obtained in matched and mismatched conditions is presented in Table 6.1. In the table, the column 'Correct' refers to whether the LR obtained in the case was compatible (or correct) with the ground truth, e.g., if the LR obtained is 4.48 (implying that it is 4.48 times more likely that the suspected speaker is at the source of the trace, than any other speaker of the potential population) and for this questioned recording, the ground truth corresponds to the fact that the suspect was not the source of the questioned recording, then this is 'incompatible' with the ground truth. The graphical estimation of the LRs, with and without mismatch, as well as after compensation, is shown in Figs. 6.2(a-t).

Table 6.1: Simulating mismatched conditions using GSM recordings as the potential population database (P) in the NFI fake case evaluation

Trace No.	Mismatched LR (P in GSM)	Correct	Compensated LR (GSM to PSTN)	Correct	Matched LR (P in PSTN)	Correct
Q1	4.481	×	0.0595	✓	0.1844	✓
Q2	1319.3	✓	7.1379	✓	2015.7	✓
Q3	2.62E+08	✓	89.778	✓	1.42E+07	✓
Q4	7.05E+05	✓	88.908	✓	2.38E+07	✓
Q5	1.24E+07	✓	36.98	✓	3.13E+07	✓
Q6	27561	✓	10.313	✓	8715.5	✓
Q7	3.71E+06	✓	36.708	✓	3.33E+06	✓
Q8	8.2393	×	0.017012	✓	0.19434	✓
Q9	2.77E+06	✓	34.414	✓	4.46E+06	✓
Q10	0.00067337	✓	3.41E-08	✓	5.22E-07	✓

It can be observed that under the conditions of mismatch, all the cases in which the suspect was indeed the source of the trace have LRs greater than 1. However, in two of the three cases where the suspected speaker was not the source of the trace, LRs above 1 are obtained.

This behavior is also observed in Tippett plots corresponding to mismatched conditions (Fig. 6.6), where P was in GSM conditions and R,C and T databases were in PSTN conditions. In this situation, the majority of H_0 true cases observing cases obtaining an $LR > 1$. However in the case of H_1 true, we observed that the approximately 20% of the cases obtain a LR above 1. Similarly, while all the H_0 cases in this evaluation have LRs greater than one, it is the H_1 cases, where the LRs do not correspond to the ground truth.

After applying compensation, using the scaling database, all the seven cases corresponding to H_0 true have an LR greater than 1, and the three cases corresponding to H_1 true, have an LR less than 1. Although the LRs obtained in compensated

conditions are not the same as those obtained in matched conditions, they show the same trends and match correctly with the ground truth.

Using a potential population recorded in room acoustic conditions

Similarly, for the same ten cases, the same R, C and T databases are used, and in order to simulate mismatch, we use the speech of 50 speakers from the IPSC-03 database recorded in room acoustic recording conditions as the mismatched potential population database. 20 other speakers, in room acoustic and PSTN conditions, were used as the scaling database.

A summary of the LRs obtained in matched and mismatched conditions is presented in Table 6.2. The graphical estimation of the LRs, with and without mismatch, as well as after compensation, are shown in Figs. 6.2(a-t).

Again, it can be observed that under the conditions of mismatch, all the cases in which the suspect was indeed the source of the trace have LRs greater than 1. However, in two of the three cases where the suspected speaker was not the source of the trace, we obtain LRs above 1.

Table 6.2: Simulating mismatched conditions using room-acoustic recordings as the potential population database (P) in the NFI fake case evaluation

Trace No.	Mismatched LR (P in room -acoustic)	Correct	Compensated LR (room-acoustic to PSTN)	Correct	Matched LR (P in PSTN)	Correct
Q1	4.49E+06	×	0.15087	✓	0.1844	✓
Q2	2.58E+07	✓	181.25	✓	2015.7	✓
Q3	2.83E+08	✓	4740.3	✓	1.42E+07	✓
Q4	4.00E+08	✓	1.10E+07	✓	2.38E+07	✓
Q5	3.37E+08	✓	5.21E+07	✓	3.13E+07	✓
Q6	1.12E+08	✓	2091.3	✓	8715.5	✓
Q7	3.68E+08	✓	13737	✓	3.33E+06	✓
Q8	1.54E+06	×	0.021398	✓	0.19434	✓
Q9	4.15E+08	✓	8.64E+06	✓	4.46E+06	✓
Q10	~ 0 (5.42E-17)	✓	~ 0 (4.34E-24)	✓	~ 0 (2.83E-23)	✓

In this instance as well, applying compensation, using the scaling database, all the seven cases corresponding to H_0 true have an LR greater than 1, and the three cases corresponding to H_1 true, have an LR less than 1. Although the LRs obtained in compensated conditions are not the same as those obtained in matched conditions, they show the same trends and match correctly with the ground truth.

Using the PolyCOST 250 database potential population

For the NFI evaluation, the PolyCOST database had been chosen as the potential population (see Appendix D for the original evaluation report). This database was chosen among the available databases because it was found to be best suited to the case, especially in the language (English spoken by European speakers) and technical conditions (fixed European telephone network) under which the reference recordings of the suspect were made. The main difference between this database and the IPSC-03 database are in the language (English instead of French) and the different recording conditions (PSTN instead of PSTN, GSM, and room acoustic).

A summary of the *LR*s obtained in matched and mismatched conditions is presented in Table 6.3.

Table 6.3: LRs obtained using the PolyCOST database, in English, as the potential population

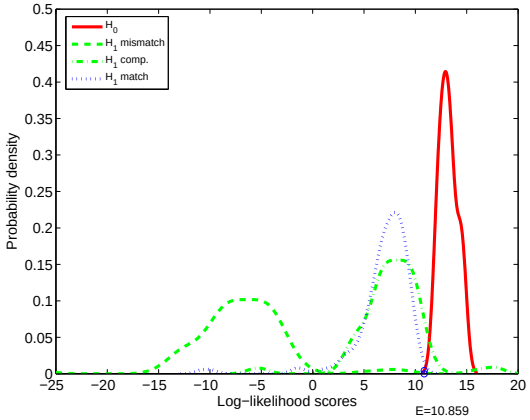
Trace No.	LR (P in English)	Correct
Q1	6.56	×
Q2	163.41	✓
Q3	23723.98	✓
Q4	21720.97	✓
Q5	11631.8	✓
Q6	329.0	✓
Q7	38407.33	✓
Q8	0.660	✓
Q9	3033.47	✓
Q10	~ 0 (4.36 x 10-23)	✓

In this analysis, all the cases in which the suspect was indeed the source of the trace have *LR*s above 1. However, in one of the three cases where the suspected speaker was not the source of the trace, we obtained *LR*s above 1*. This highlights the fact that the choice of the potential population can affect the outcome of the analysis, and that the strength of evidence depends on the recording conditions and language of the potential population database.

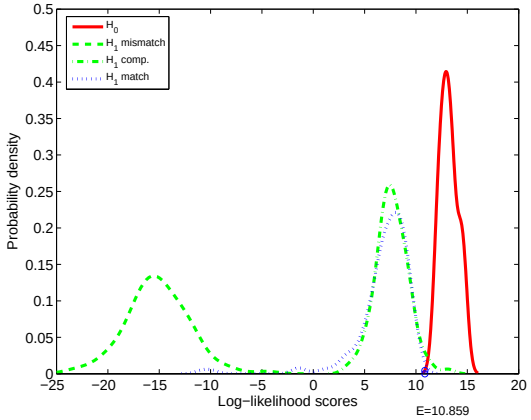
The *LR*s for the matched conditions, using the IPSC-03 database as well as the PolyCOST database, show the same trends. The questioned recordings for which relatively higher *LR*s are obtained for the IPSC-03 PSTN database, also obtain rel-

*Note that in the real evaluation, this LR for $QR1$ was not considered significant (at a 5% significance level), and it was concluded that, 'although the likelihood ratio constitutes limited support for the hypothesis questioned recording (QR1) and the reference recordings from the suspected speaker, have the same source, the statistical significance analysis does not allow us to progress the case in any direction'.

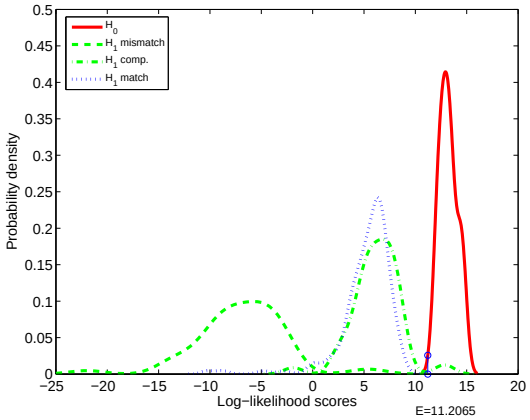
atively higher LRs, using the PolyCOST database. The questioned recordings for which relatively lower LRs are obtained for the IPSC-03 PSTN database, also obtain relatively lower LRs using the PolyCOST database.



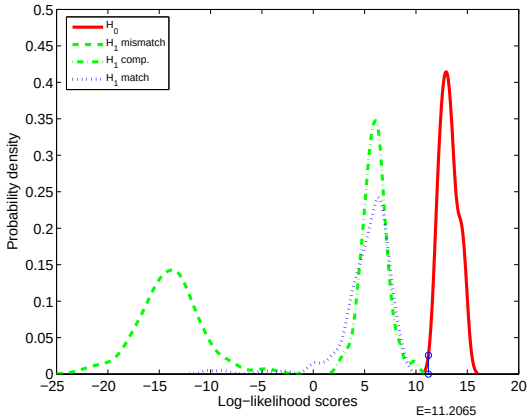
(a) Questioned recording 1 (H_1 true): Potential Population in GSM condition



(b) Questioned recording 1 (H_1 true): Potential Population in acoustic room recording condition condition



(c) Questioned recording 2 (H_1 true): Potential Population in GSM condition

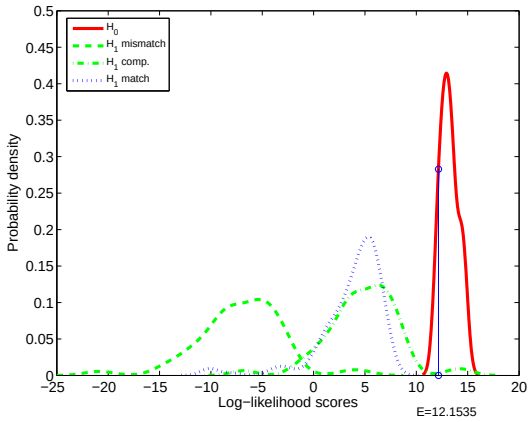


(d) Questioned recording 2 (H_0 true): Potential Population in acoustic room recording condition condition

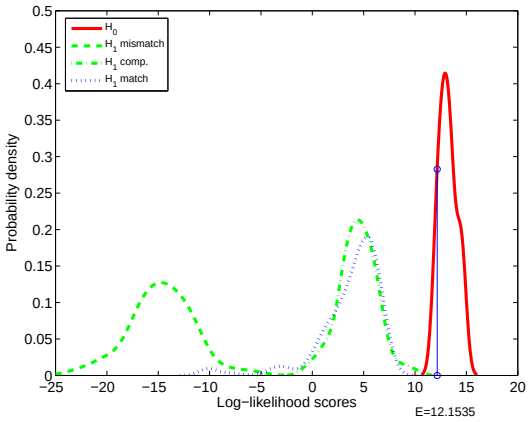
6.3 Summary

In this chapter, an evaluation of the statistical compensation methods proposed earlier was performed, using several simulated cases in mismatched conditions as well as with individual cases from the Netherlands Forensic Institute (NFI) Speaker Recognition Evaluation through a fake case.

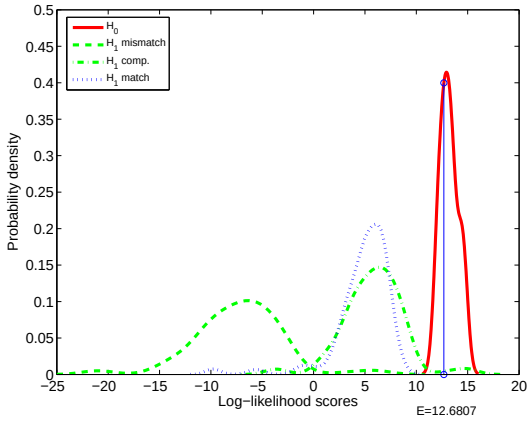
The conclusions from this evaluation are:



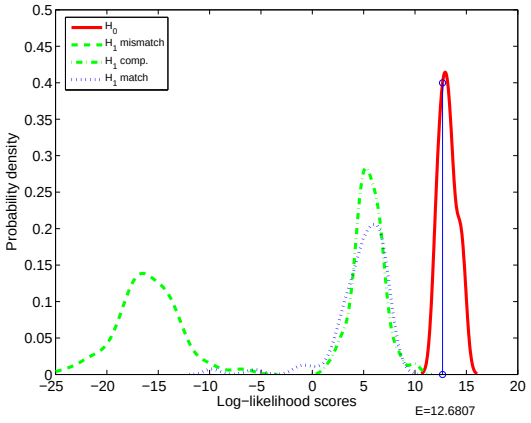
(e) Questioned recording 3 (H_0 true): Potential Population in GSM condition



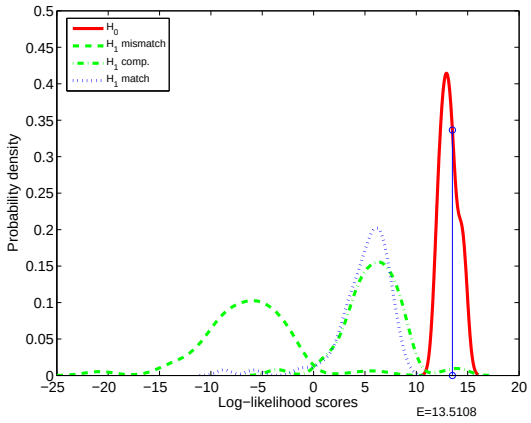
(f) Questioned recording 3 (H_1 true): Potential Population in acoustic room recording condition condition



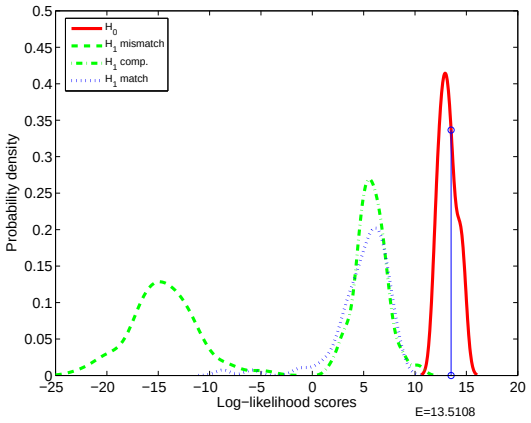
(g) Questioned recording 4 (H_0 true): Potential Population in GSM condition



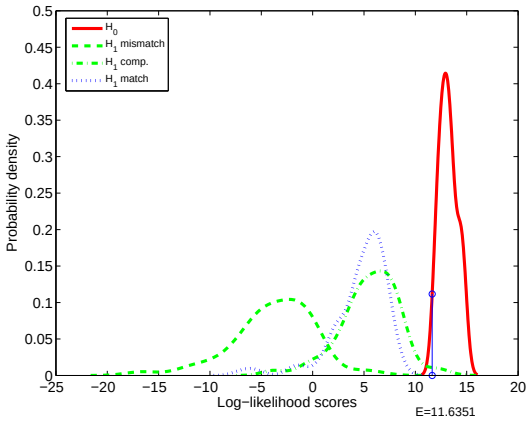
(h) Questioned recording 4 (H_1 true): Potential Population in acoustic room recording condition condition



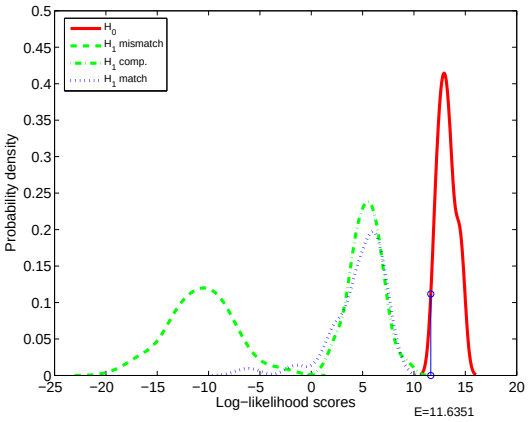
(i) Questioned recording 5 (H_0 true): Potential Population in GSM condition



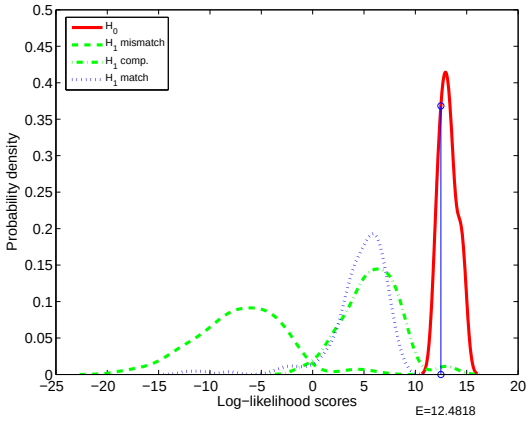
(j) Questioned recording 5 (H_1 true): Potential Population in acoustic room recording condition condition



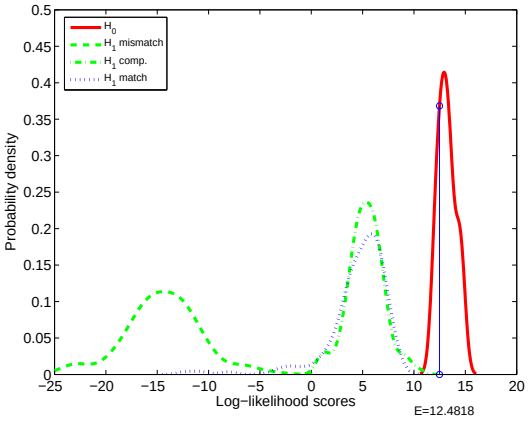
(k) Questioned recording 6 (H_0 true): Potential Population in GSM condition



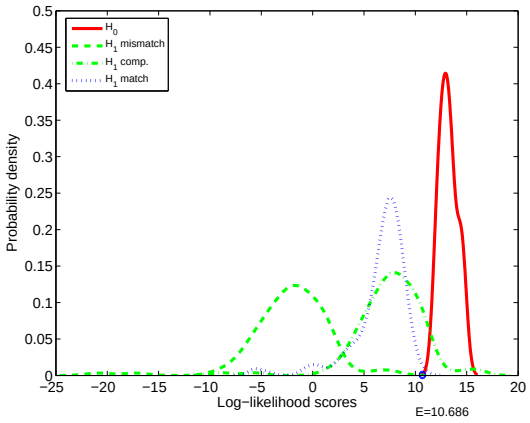
(l) Questioned recording 6 (H_1 true): Potential Population in acoustic room recording condition condition



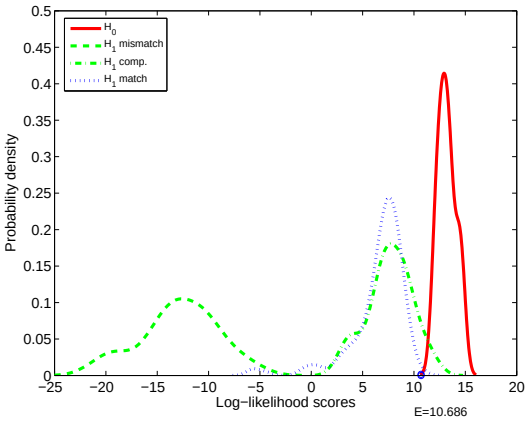
(m) Questioned recording 7 (H_0 true): Potential Population in GSM condition



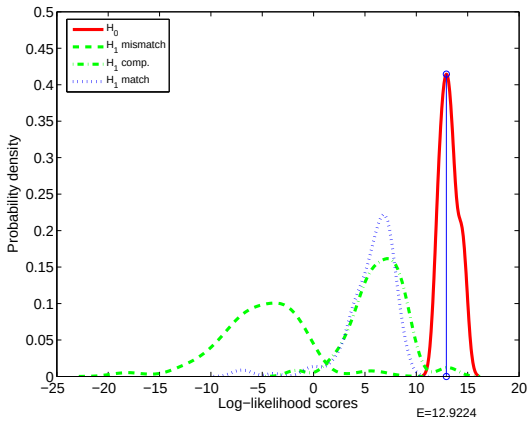
(n) Questioned recording 7 (H_1 true): Potential Population in acoustic room recording condition condition



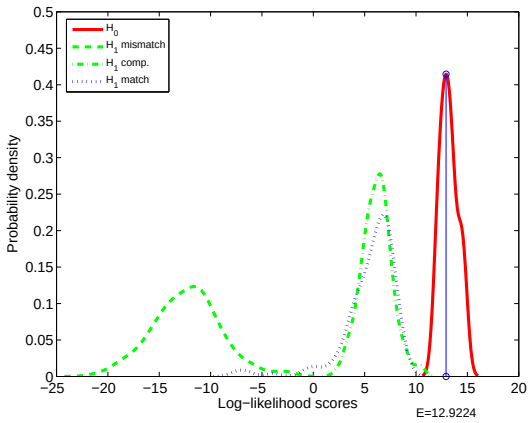
(o) Questioned recording 8 (H_1 true): Potential Population in GSM condition



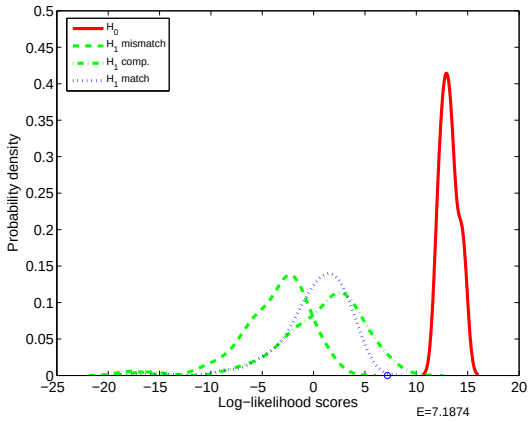
(p) Questioned recording 8 (H_1 true): Potential Population in acoustic room recording condition condition



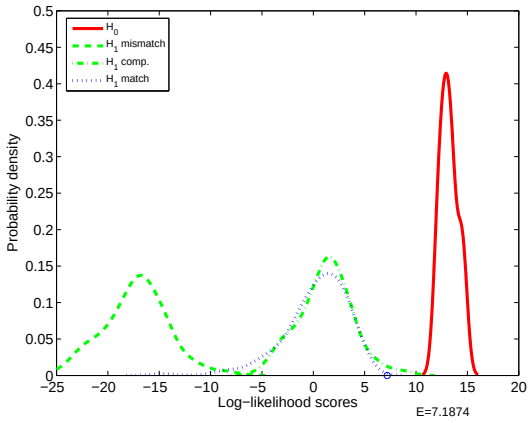
(q) Questioned recording 9 (H_0 true): Potential Population in GSM condition



(r) Questioned recording 9 (H_1 true): Potential Population in acoustic room recording condition condition



(s) Questioned recording 10 (H_0 true): Potential Population in GSM condition



(t) Questioned recording 10 (H_1 true): Potential Population in acoustic room recording condition condition

- The automatic speaker recognition system shows good performance in matched conditions, with a small proportion of H_1-true cases obtaining an LR greater than 1 and only a small proportion of H_0-true cases obtaining an LR less than 1.
- The system shows poor performance in mismatched conditions, with a high proportion of H_1-true cases obtaining an LR greater than 1 and a small proportion of H_0-true cases obtaining an LR less than 1.
- A scaling database (S) can be used in order to estimate the parameters for the compensation of the mismatched conditions, if the conditions of mismatch are known.
- The number of speakers required for the scaling database was determined by choosing an LR of 1 as a reference point and by compensating for mismatch using a gradually increasing number of speakers in the scaling database, until there is relatively no change in the proportions of cases obtaining an LR above 1, when H_0 and H_1 are true.
- It was experimentally shown that size of the scaling database need not be very large (of the order of tens of speakers).
- There is an error in the estimation of the compensated LRs, when compared to the LRs obtained in matched conditions, but the compensated LRs correspond more closely to matched conditions than mismatched conditions.

7 Discussion

In this chapter, the main ideas proposed in the thesis are reviewed, and the advantages and limitations of the methods proposed are discussed. The extent of the applicability of Bayesian interpretation methodology, with statistical compensation for mismatched recording conditions in real forensic cases, and the admissibility of scientific evidence based on these methods is considered.

The main problem this thesis focuses on is the effect of mismatched recording conditions on the evaluation of the strength of evidence. Mismatched recording conditions are shown to adversely affect the strength of evidence in both aural-perceptual and automatic speaker recognition methods. The estimation and statistical compensation of mismatch, using databases representative of the mismatched conditions, have been proposed. The other problems that have been considered are related to the adaptation of the Bayesian interpretation methodology to the requirements of real forensic casework. Some of the difficulties observed in our experience with real forensic speaker recognition cases, in applying the Bayesian interpretation methodology presented in [Meuwly, 2001] to real forensic casework, have motivated most of the research work presented in this thesis. The challenges included:

- Selecting a potential population database that corresponded to the case in language and recording conditions, and handling cases where this was not possible.

- Handling limited suspect data, i.e., when only a single recording of a suspect and a single questioned recording was available, hence making it difficult to estimate the within-sources variability of the suspected speaker's voice.

- Using complementary information (e.g. Tippett plots, risk of errors in choosing either hypothesis, etc.) to the likelihood ratio presented to the courts.
- Estimating confidence intervals on the likelihood ratios obtained, especially when the likelihood ratios were evaluated on the tails of the distributions of scores for the two hypotheses.

7.1 Handling mismatch within the Bayesian interpretation framework

One of the advantages of Bayesian interpretation using likelihood ratios is that it provides a logical framework for the evaluation of evidence in the light of competing propositions, which is very similar to the adversarial system typical to the courts. The Bayesian interpretation methodology is important because it provides a general framework for the interpretation of the strength of several kinds of forensic evidence. The success of Bayesian interpretation depends on its elegant and clearly understandable framework which can be explained to juries and judges in a logical, coherent way.

Likelihood ratios, used in Bayesian interpretation in forensic analysis, depend on the relative prevalence of the distinguishing characteristics of the sample (voice, fingerprint, DNA, glass, etc.) in question, with respect to the sample it is being compared to, as well as to a relevant population of such samples. In evaluating the likelihood ratio, it is important that the comparisons are based on characteristics of the sample under analysis and not on the conditions in which measurements of these characteristics were acquired. The problem of mismatched conditions of the acquisition of samples is general to many fields of forensic analysis, and it is necessary that the experts take this mismatch into account in their analysis and when formulating their conclusions. Sample acquisition in mismatched conditions can pose a serious challenge to the implementation of the Bayesian interpretation (B.I.) framework in practical conditions of forensic analysis. If they are not considered, the strength of evidence can be inaccurate or even erroneous. This, in turn, could perhaps result in serious miscarriages of justice. Any methodology that tries to compensate for mismatch within the Bayesian interpretation framework should also be logical and easy to understand and explain to the courts. The method chosen for compensation of mismatch must therefore be simple, and the choice of such a method, which is acceptable from the technical standpoint as well as easily interpretable by the courts, is a difficult one.

The method to evaluate and compensate for mismatched conditions, as presented in this thesis (Chap. 5) seeks to address the need for handling mismatch within the data-driven Bayesian interpretation framework which can be explained by the expert

and is easily interpretable for the courts. The method for statistical compensation essentially involves estimating the extent of shift in distribution parameters that mismatched conditions introduce, using scaling databases, and statistically compensating for the mismatch in an individual case. This method of distribution scaling was chosen over more sophisticated techniques, such as compensation using multi-Gaussians (Sec. 2.3.2), because the additional benefit in accuracy would be offset by the difficulty in interpretation and understanding of these methods in the courts.

The statistical compensation method moves away from traditional signal processing approaches to handling channel and mismatch at the feature extraction front-end or in the statistical models of the speech. Instead, it is applied at the level of scores and has the advantage of being universally applicable across different automatic speaker recognition systems. These scores could represent distances, similarity measures, likelihoods, etc. In keeping with the Bayesian interpretation framework, this technique allows for any feature extraction or modeling technique to be used. In fact, just as the Bayesian interpretation methodology can be directly applied to other fields of forensic analysis, statistical compensation techniques can also be applied to mismatched conditions in other fields of forensic analysis.

This thesis considers mismatch in the context of recording conditions and suggests a compensation methodology. However, the compensation is also generally applicable for other types of mismatch between recordings. For example, if the language of recording for a certain case is French, in certain technical conditions, and the relevant potential population, in the same technical conditions, is available only in English, then the influence of language mismatch between French and English can be estimated and compensated for in a similar manner, using a scaling database. Some of the main issues related to the methodology for statistical compensation for mismatched recording conditions are discussed below:

Knowledge of the recording conditions of the databases used

One of the important assumptions in the method proposed for the compensation of mismatch, using scaling databases containing speakers in each of the mismatched conditions, is that the conditions of recording in the case are known precisely. This means the expert should know under what conditions the recordings in the case were made. In a case where mismatch is known to occur between two conditions, C_1 and C_2, it is necessary that the scaling database should be recorded in exactly the same C_1 and C_2 conditions. If statistical compensation has to be applied for scores, it is necessary to know, precisely, the recording conditions of the case and the databases the expert has at his disposal. It is also possible for the expert to analyze and perform tests on the recordings in order to estimate what conditions they are in. Although this information can be requested or tests can be performed in order to ascertain the

recording conditions, accurate information about these recording conditions may not often be available to the forensic expert, and in such cases, it may be difficult to use this method. Secondly, this method is limited by the fact that the *same* speakers have to be recorded in several different conditions, and it should then be possible to record these speakers again, in the future, when new unseen conditions are encountered.

Data sufficiency

In our experiments (Sec. 6.1.5), we have observed that the number of speakers in the scaling database does not need to be very high (in the order of tens) to facilitate estimation of the shift in distribution parameters. Thus, it is possible, for instance, for a forensic audio laboratory to use a relatively small set of speakers who are recorded in a variety of commonly observed conditions, and to also have the additional possibility of recording them in previously unobserved conditions when certain analyses demand it. Since the number of speakers required for this is relatively small, this set could be, for instance, the staff of the laboratory who can be recorded in previously unseen conditions. These speakers can thus constitute the scaling database.

Limits to using statistical compensation techniques

There are certain limits to using statistical compensation for mismatch as this does not constitute a perfect mapping between scores obtained in one condition to another. While statistical compensation allows us to calculate likelihood ratios that are closer to the likelihood ratios obtained in matched conditions than the corresponding likelihood ratios in mismatched conditions, they are not *exactly* the same as the likelihood ratios in matched conditions. The compensation method is an approximate transformation. We observe that while the compensations based on means and variances are satisfactory for the majority of the samples in the distributions, there is still a certain extent of error (Sec. 6.1.5), especially with distributions that deviate from the Gaussian distribution significantly.

Methodology for creating corpora handling mismatch

Since databases (suspect control, suspect reference and potential population databases) play a central role in the Bayesian interpretation methodology, the creation and selection of the databases have to be done carefully. Apart from casework requirements, for the evaluation of the performance of the recognition system, it is necessary to have databases that are forensically realistic. Protocols for the creation of such databases that can be used, in practice, for forensic casework are also necessary.

Three databases have been used in this thesis, namely IPSC-01 [Meuwly et al., 2003] (see Appendix A, the IPSC-02 and the IPSC-03 (see Appendix B and C). Two

of these databases were recorded during the course of this thesis and were created specifically for use in the Bayesian interpretation methodology. We have also proposed a methodology for the creation of a speaker recognition database consisting of recordings of speakers in several different forensically relevant conditions, in order to detect and compensate for mismatched conditions. A prototype of a speaker recognition database (IPSC-03) was created according to this methodology, which is a forensically realistic database, and it can be used for aural, instrumental and automatic speaker recognition, in order to estimate and compensate for mismatch in recording conditions. The creation of each database gave insights into the difficulty in creating databases that can be used for forensic speaker recognition casework. For instance, the IPSC-02 database was rich in the variety of recording conditions (PSTN, GSM, and analog tape-recording on an answering machine) as well as in the languages used (mainly French and German, as well a few other languages that the subjects were familiar with). However, this database was small in size, containing only 10 speakers. Although, this database provided insight into the influence of recording conditions and language, it was necessary to record a database with a larger number of speakers. The IPSC-03 database was recorded according to the methodology presented in Section 5.5.2, in order to validate the compensation for mismatched conditions. This database contained a sufficiently large number of speakers (73) from which statistics could be reasonably calculated.

7.1.1 The influence of mismatch on aural and automatic speaker recognition

We have studied the influence of mismatched recording conditions on both automatic as well as aural-perceptual speaker recognition, and observed that mismatch constitutes a serious problem for automatic and aural-perceptual recognition (Chap. 4). The human aural-perceptual recognition already adapts to mismatch in conditions, and the automatic system can benefit from compensation for mismatched conditions. The use of scaling databases of the same speakers in a variety of conditions, where the effects of the recording conditions on various parameters can be estimated, can be of use to aural-perceptual and phonetic-acoustic methods of analysis. In phonetic-acoustic analysis, the extent of variability introduced by the change in conditions, e.g., variability in the formant frequencies due to a certain telephone transmission [Byrne and Foulkes, 2004], can be measured and compensated for.

We have also presented a method of converting subjective opinions of the listeners performing a recognition task (e.g., I am sure that the two speakers are the same) to that of a Bayesian likelihood ratio. Considering the hypotheses that the two recordings could have the same source and that the two recordings have a different

source, we are able to map these subjective opinions onto likelihoods and, thereby, derive the strength of evidence on their basis. The scheme of comparison presented in Chap. 4 can also be adapted in order to analyze the performance of the aural perception of persons with phonetic training. It could serve to evaluate the performance of phoneticians and automatic systems, in different conditions, and would indicate methods of combining the aural-perceptive approach of trained subjects with that of the automatic system, depending on the conditions of the case. While this presents a good way of highlighting the weaknesses and combining the strengths of each of the compensation methods, in conducting a large number of tests of aural recognition, in several different conditions, with each phonetician (required to obtain statistically valid comparisons), comparing the results with those of the automatic system will prove difficult.

7.1.2 Adapting the Bayesian interpretation framework to real forensic conditions

While mismatch was one of the main problems encountered in applying the Bayesian interpretation methodology in practice, other areas, not exclusively related to mismatch, were also considered in this thesis. These include handling cases with limited suspect data*, presenting complementary information that was not available in the likelihood ratio, such as the risk of errors for a given value of evidence E, and the confidence interval for a given likelihood ratio. In this section, we discuss and review these other methods.

A general interpretation framework for cases in which only limited suspect data was available was presented. When the suspect data is limited, the strength of evidence can be evaluated with respect to the hypotheses, H_0 (the two recordings have the same source) and H_1 (the two recordings have different sources). These hypotheses are different from the hypotheses used in [Meuwly, 2001], i.e., H_0 - the suspected speaker is the source of the questioned recording, and H_1 - the speaker at the origin of the questioned recording is not the suspected speaker. The within-source variability of individual speakers can be approximated by the average within-source variability from representative databases. The main assumption in this framework is that although differences can exist between speakers, scores obtained by comparing speakers' models with their own voices (H_0 cases) are in a similar (higher) range that is distinct from the range of scores (lower) obtained by comparing the speakers' own voices with someone else's models (H_1 cases). The differences between speakers' within-source variability is assumed not to be significant compared to the differences

*which could not be handled directly with the existing Bayesian interpretation methodology which depends heavily on databases and the availability of suspect reference and control data

between scores obtained in H_0 cases and scores obtained in H_1 cases. While such differences are, for the most part, negligible in well-matched conditions, in mismatched conditions, this assumption may not hold.

We used the Error Ratio (ER) which takes into consideration the relative risk of error for an evidence score in choosing either of the hypotheses. The ER is the proportion of cases for which recordings from the same source would be wrongly considered to come from different sources, divided by the proportion of cases in which recordings from different sources were wrongly considered to be from the same source, if E is used as a threshold in a hypothesis test for match or non-match. This is complementary information to the likelihood ratio, and it can be provided to the courts so long as it is not confused with the strength of the evidence which is evaluated using the likelihood ratio.

The strength of evidence in forensic speaker recognition can vary and is often better represented using a range of possible values than a precise number (Sec. 3.4). This variability is analyzed using statistical significance analysis and a subset bootstrapping technique. Confidence intervals for the likelihood ratio estimates can be derived from the variability analysis, and these intervals have been presented in the context of equivalent verbal scales used for reporting likelihood ratios to the courts. We proposed that in addition to the likelihood ratio, the statistical significance of the evidence score obtained with respect to each of the hypotheses, a bootstrapped confidence interval for a range of possible values of the strength of evidence, as well as the equivalent of the range of likelihood ratio scores on the verbal scale and an explanation of the meaning of the scale should also be presented. This analysis is particularly important because there can be cases in which, for instance, while the likelihood ratio supports a certain hypothesis (likelihood ratio greater than 1), the confidence interval includes values both above and below 1. This implies that the variability analysis does not allow us to support either hypothesis, given the evidence. Statistical significance testing is particularly important to avoid computing likelihood ratios in regions of the probability distribution where there is sparse data, especially in the tails of each of the score distributions. When the probability at the tails of the distributions tends to 0, the likelihood ratios calculated at these points may tend to infinity, and thus, exaggerated likelihood ratios may be obtained in these regions. Significance testing for the evidence, with respect to each of the distributions, can thus help avoid erroneous and exaggerated values of the likelihood ratio.

7.2 Admissibility of forensic automatic speaker recognition based on Bayesian interpretation and statistical compensation in courts

In order for forensic automatic speaker recognition analysis-based evidence to be acceptable for presentation in the courts, the methodologies and techniques have to be researched, tested and evaluated for error, as well as be generally accepted in the speaker recognition community. Let us consider again the requirements that the Daubert ruling (described in Sec. 1.2) sets out for the admissibility of scientific evidence.

- Whether the theory can be tested or whether there is proof of testing of the underlying hypothesis upon which the analysis technique is based.

  Forensic automatic speaker recognition using Bayesian interpretation and compensation for mismatch lends itself to testing for errors, using databases of cases that are created in similar conditions as those of the case. These methods have been tested with databases specifically created to simulate conditions in forensic casework, from which mock cases in similar conditions can be constructed. However, the conditions in which these methods have been tested are still 'controlled', to some extent, and do not represent all the variability that is encountered in real forensic cases. The performance of the recognition system before and after adaptation for mismatch, for cases in similar conditions, can be shown using the Tippett plots as proof of testing of technique.

- Whether the technique has been published or subjected to peer review.

  The Bayesian interpretation methodology for automatic speaker recognition has been published in international journals [Champod and Meuwly, 2000], a doctoral thesis [Meuwly, 2001] and several conferences (COST 250 workshop 1998, European Academy of Forensic Science (EAFS) 2000, EAFS 2003, Eurospeech 2003, Speaker Odyssey 2004, Interspeech 2005) have dedicated sessions or tutorials on forensic speaker recognition [Meuwly et al., 1998; Drygajlo et al., 2003; Pfister and Beutler, 2003; Gonzalez-Rodriguez et al., 2004; Alexander et al., 2004; Botti et al., 2004a; Campbell et al., 2005].

  The method of statistical compensation for handling mismatched recording conditions has been presented to both the automatic speaker recognition community (Speaker Odyssey 2004) [Alexander et al., 2004] and to forensic phoneticians (IAFPA conference 2005). Some automatic systems have been built by universities and private companies, based on this methodology, such as Agnitio in Spain (BatVOX) and EPFL-IPSC (Aspic).

- Whether the technique has a known or potential error rate in application.

  Knowing the potential error rate of a method can often be very important in establishing the admissibility of certain evidence. The Tippett plots show the judge what the known or potential error rate would be in cases where the method is applied under similar conditions. These Tippett plots which represent the cumulative density functions of the likelihood ratios in cases where the suspected speaker was truly the source of the questioned recording and in those when the suspected speaker was not the source of the questioned recording, can be used to illustrate how effective the methods used by the experts have been in similar cases.

  In the case of mismatch and the compensation methods that have been proposed, Tippett plots have been presented to show that the compensation helps reduce the effects of mismatch for several cases, both when H_0 is true and H_1 is true.

  Also, there have been evaluations like the NFI-TNO 2003 evaluation [Bouten and van Leeuwen, 2004] wherein several state-of-the-art systems were evaluated using real forensic data. However, this evaluation was styled similar to the NIST evaluations for speaker verification. The lowest equal error rate (EER) obtained in this evaluation was 12.1% (15-second test segments and 60-second training segments). While these evaluations are basically for speaker verification, they evaluate the performance of the automatic speaker verification system in forensic case conditions.

- Whether standards exist and are maintained to control the operation of the technique.

  Forensic speaker recognition is an important part of forensic audio analysis, and there exist many universities and government forensic laboratories that are working in this and related areas. Some of these laboratories have joined together to establish organizations that can maintain common guidelines and standards for analysis, as well as to test and accept various techniques.

  The European Network of Forensic Science Institutes, Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG) includes members representing forensic and university laboratories from 22 countries in Europe, as well as the USA. This group is concerned with the problem of forensic speech and audio in general, and this includes speaker recognition, speaker profiling, voice lineups, audio transcription, audiotape authentication, etc. One of the important concerns of ENFSI-FSAAWG in speaker recognition, is to provide a standardized evaluation of different methodologies and reporting strategies. The NFI

fake case evaluation discussed earlier is one such collaborative exercise where different experts were given the same material and the analysis methodologies and reports were compared. Such testing can help in maintaining standards and to control different operating techniques. Another such organization is the International Association for Forensic Phonetics and Acoustics (IAFPA). This organization defines standards for its members, in their reporting on casework, and requires them to be aware of the limitations of techniques that they use. Although the IAFPA is primarily concerned with forensic phonetics, forensic automatic speaker recognition now has a growing presence at their annual conferences.

- Whether the technique is generally accepted within the relevant scientific community.

  The Bayesian interpretation methodology is gradually gaining acceptance as the logical way of presenting evidence to the forensic speaker recognition community. Speaker recognition itself has generated a fair amount of controversy ever since its early use. The popular perception of speaker recognition, perhaps influenced by the media, is that it is possible to uniquely identify a speaker by his voice, and the scientific community is making attempts to dispel such myths. In [Bonastre et al., 2003], the authors caution that 'although solutions do exist to some constrained recognition applications, there is no scientific process that enables one to uniquely characterize a person's voice or to identify, with absolute certainty, an individual from his or her voice'. The absence of any unique characterization of the speaker's voice fits well with the Bayesian interpretation methodology which requires that both within-source and between-sources variability of any characterization should be considered.

  Compensation for mismatched conditions using score normalizations has been used in automatic speaker recognition for several years. As discussed earlier, the statistical compensation method using representative databases has been presented to the automatic speaker recognition community as well as to forensic phoneticians. However, it must gain widespread acceptance before this requirement from the Daubert ruling can be satisfied completely.

- Whether the method is based on facts or data of type reasonably relied on by experts in the field.

  A variety of different algorithms and features is used for automatic speaker recognition. For instance, feature extraction techniques such as MFCCs, RASTA-PLP, LPC, etc. are established techniques relied upon in the speech and speaker recognition community, while modeling techniques such as vector

quantization, Gaussian mixture modeling, dynamic time warping (DTW), etc. are also widely accepted. The choice of the algorithm and features depends on the nature of the analysis task, as a certain combination of features and modeling algorithms may be more suitable for a given set of conditions than another. A systematic analysis of results is possible by considering similar cases in the same conditions, where the outcome was known, and it is thus possible to choose the most promising combination of algorithms and features that are suitable for the analysis.

- Whether the evidence presented using this methodology will have a probative value that is not outweighed by the dangers of unfair prejudice, confusion of issues or misleading the jury.

  The Bayesian interpretation of evidence is an elegant framework for the evaluation of evidence in the context of competing hypotheses in a judicial adversarial system. The propositions chosen for each of the hypotheses are made according to the conditions of the case; in order to avoid prejudice, each proposition (e.g., the suspected speaker is at the source of the questioned recording) is counterweighed by a competing proposition (e.g., someone else in the population is at the source of the questioned recording). The Gaussian mixture models lend themselves directly to evaluating the likelihood of observing the features of the questioned recording in the statistical model of the suspected speaker's voice. The multivariate features are reduced to univariate log-likelihood scores (in the scoring method), and the evaluation of the evidence is reduced to a simple hypothesis comparison problem which is easier to understand and visualize. The compensation method proposed is applied in the univariate log-likelihood score domain, where distribution scaling is also easier to visualize and understand.

7.3 Summary

In this chapter, the advantages and limitations of the methods proposed in this thesis were reviewed. The salient points discussed include:

- Mismatched recording conditions adversely affect the strength of evidence, both in the use of aural-perceptual as well as automatic speaker recognition systems.
- The methodology presented for the compensation of mismatch, using representative databases, fits well in the Bayesian interpretation framework and can be easily interpreted and visualized.
- The statistical compensation proposed has a certain amount of error, especially with distributions that significantly deviate from the Gaussian distribution.

- Experimentally, we have observed that it is not necessary for the number of speakers in the scaling database to be very high, although a larger number of speakers would give a more accurate estimate of the shift in score distributions due to recording conditions.

- An important assumption in the method proposed for the compensation of mismatch using scaling databases is that mismatched conditions of recording in the case are known precisely.

- When suspect data is limited, the strength of evidence can be evaluated with respect to the hypotheses, H_0 (the two recordings have the same source) and H_1 (the two recordings have different sources). The within-source variability of individual speakers can be approximated by the average within-source variability of several speaker databases.

- The strength of evidence in forensic speaker recognition is variable and is often better represented using a range of possible values than a precise number.

- The methodology presented in this thesis can be viewed in the light of the Daubert ruling for the admissibility of scientific evidence.

8 Conclusion

In this thesis we propose a new approach to estimate and compensate for the effects of mismatch arising in forensic automatic speaker recognition casework in a corpus-based Bayesian interpretation framework, due to the technical conditions of encoding, transmission and recording of the databases used. Mismatch in the recording conditions between databases used in analysis can lead to erroneous or misleading estimates of the strength of evidence, and it is of utmost necessity to quantify and reduce the uncertainty introduced in the likelihood ratios, in order to avoid possible miscarriages of justice.

We investigated two main directions in applying the Bayesian interpretation framework to problems in forensic speaker recognition, the first concerning the problem of mismatched recording conditions of the databases used in the analysis and the second concerning the Bayesian interpretation as applied to real forensic case analysis.

The main contributions related to handling mismatched recording conditions of the databases used in forensic speaker recognition, within a Bayesian interpretation framework, include a methodology for estimating and statistically compensating for the differences in recording conditions, guidelines for the creation of a forensic speaker recognition database that can be used in order to perform forensic casework in mismatched conditions and the analysis of mismatched technical conditions in training and testing phases of speaker recognition and their effect on human-aural and automatic forensic speaker recognition.

The contributions related to Bayesian interpretation applied to real forensic case analysis include the analysis of the variability of the strength of evidence using bootstrapping techniques, statistical significance testing and confidence intervals, handling

cases where the suspect data is limited, and using complementary information to the likelihood ratio about the risk of errors involved in choosing a certain hypothesis.

8.1 Handling mismatched recording conditions in the Bayesian interpretation framework

8.1.1 Handling mismatch in recording conditions in corpus-based forensic speaker recognition

If an existing mismatch between databases is undetected, it is likely that the uncompensated usage of the Bayesian interpretation framework gives erroneous results. After detecting a mismatch, the expert has a choice of selecting another database compatible with the questioned recording (if possible), deciding not to analyze the case in the Bayesian interpretation framework or performing statistical compensation for the mismatched conditions. Detecting and compensating mismatches between databases helps to reflect more accurately the similarity or dissimilarity of the voices in the recordings.

We have introduced a methodology to estimate and statistically deal with differences in recording conditions of the databases used at the level of scores. This statistical compensation for mismatch is applied at the level of scores, as opposed to the level of features or of models. This method of statistical compensation is thus universally applicable to the many different combinations of feature extraction and modeling techniques used in automatic speaker recognition.

Under the assumption that the recording conditions of the questioned recording as well as potential population database are known precisely, it is possible to use databases for scaling score distributions, named 'scaling databases'. These scaling databases contain a set of the same speakers in various different recording conditions, in order to estimate parameters for scaling distributions to compensate for mismatch. In our experiments, we have observed that the number of speakers in the scaling database does not need to be very high (in the order of tens) to estimate the shift in parameters. While statistical compensation allows us to calculate likelihood ratios that are closer to the likelihood ratios obtained in matched conditions than the corresponding likelihood ratios in mismatched conditions, they are not *exactly* the same as the likelihood ratios in matched conditions. The compensation method is an approximate transformation based on means and variances and there is still a certain extent of error, especially with distributions that deviate from the Gaussian distribution significantly.

Although primarily the Bayesian interpretation framework is used for the evalu-

ation of evidence in court, it is also a valuable tool for investigatory purposes. The methodology of compensating mismatch is equally important for both these purposes in order to avoid results that are affected by differences in recording conditions. When it is not possible to use suspect reference recordings that are recorded in exactly the same conditions as those of the potential population, compensation reduces the effect of mismatch, and estimates scores that could have been obtained if the suspect reference and potential population databases had been recorded in the same conditions.

8.1.2 Methodology for creating databases to handle mismatch

We have proposed a methodology to create a forensic speaker recognition database in order to estimate the mismatch in recording conditions that arises in forensic cases, to compensate for its effects and to quantify the uncertainty that is introduced due to changing conditions. This database contains:

- One or more databases in the most commonly encountered recording condition, which contains a sufficient number of speakers that can serve as a potential population database.
- A subset of speakers present in these databases, are used to record several smaller databases in different recording conditions which contains a sufficient number of speakers from which distribution scaling parameters representative of the difference in recording conditions can be calculated.

Forensically realistic databases for evaluation of forensic speaker recognition casework have been created according to the requirements of the Bayesian interpretation methodology in order to validate the methods proposed. Two of the databases used in this thesis, namely the IPSC-02 and the IPSC-03 [see Appendix B and C] were recorded during the course of this thesis and were created as specifically adapted for use in the Bayesian interpretation methodology. In order to perform statistical compensation, we require such a database with the same speakers in different recording conditions, from which these statistical compensation parameters can be estimated.

8.1.3 Mismatched recording conditions and their effect on aural and automatic forensic speaker recognition

We have analyzed mismatched technical conditions, and their effect on forensic aural and automatic speaker recognition. With perceptual speaker recognition tests performed with laypersons as well as a baseline automatic speaker recognition system, it was observed that in matched recording conditions of suspect and questioned

recordings, the automatic systems showed better performance than the aural recognition systems. In mismatched conditions, however, the baseline automatic systems showed a comparable or slightly degraded performance compared to the aural recognition systems. The extent to which mismatch affected the accuracy of human aural recognition in mismatched recording conditions was similar to that of the automatic system, under similar recording conditions. Thus, the baseline automatic speaker recognition system should be adapted to each of the mismatched conditions in order to increase its accuracy. The adapted automatic system shows comparable or better performance than aural recognition in the same conditions. The perceptual cues that human listeners rely upon, in order to identify speakers, were analyzed. It was suggested that the accuracy of automatic systems can be increased using these perceptual cues that remain robust to mismatched conditions.

8.2 Applying Bayesian interpretation methodology to real forensic conditions

8.2.1 Scoring method and direct methods for the evaluation of the likelihood ratio

In addition to the univariate evaluation of the likelihood ratio, using the scores, we have shown that the multivariate approach of directly using the likelihood of observing features, given the statistical models can also be used for the estimation of the strength of evidence in forensic automatic speaker recognition, especially when the available suspect data is limited. These two approaches, named the Direct Method and the Scoring Method, differ in that one directly uses the likelihoods returned by the Gaussian Mixture Models (GMMs) and the other models the distribution of these likelihood scores and then derives the likelihood ratio on the basis of these score distributions. Statistical representations using probability distributions like the Tippett plots to evaluate the strength of evidence and to compare the two methods were also presented.

8.2.2 Bayesian interpretation in cases with sufficient and insufficient suspect reference data

A general interpretation framework to handle forensic cases using automatic speaker recognition, both in situations where there is a limited duration of recordings for the suspected speaker as well when the length of the recordings is sufficient to estimate the within-speaker variability, has been used. In many cases only one recording of

the suspected speaker is available due to the nature of the investigation. When suspect data is limited, the within-source variability of individual speakers can be approximated by the average within-source variability from databases that are similar in recording conditions to the databases used in the case.

8.2.3 Analysis of the variability of the strength of evidence

The strength of evidence, or the likelihood ratio, in forensic speaker recognition was shown to depend on various influences in the analysis, such as the approximation in the mathematical modeling of the score distributions, the choice of speakers in the potential population and the choice of the recordings to estimate the within-source variability of suspected speaker's voice. The likelihood ratio is seen to be variable and is often better represented using a range of possible values than a single number. This variability has been analyzed using a statistical significance analysis and a subset bootstrapping technique. Confidence intervals for the likelihood ratio estimates have been derived, and these intervals have been presented in the context of equivalent verbal scales used for reporting likelihood ratios to the courts. Considering the effect of this variability in the likelihood ratios for a given case, the expert should evaluate the following:

- The statistical significance of the evidence score obtained with respect to each of the hypotheses.
- The likelihood ratio, accompanied by confidence interval which gives a range of possible values of the strength of evidence.
- The equivalent of the range of the likelihood ratio on the verbal scale, and an explanation of the meaning of the scale.

8.2.4 Complementary measures to the strength of evidence

While the likelihood ratio provides an estimate of the strength of the evidence with respect to the two hypotheses, it does not consider the risk of errors in choosing either one. We have used a complementary measure, the Error Ratio (ER), to the likelihood ratio for interpretation of the evidence, which takes into consideration the relative risk of error for an evidence score in choosing either of the hypotheses. It is the proportion of cases for which recordings from the same source would be wrongly considered to come from different sources, divided by the proportion of cases in which recordings from different sources were wrongly considered to be from the same source, if the evidence score E is used as a threshold in a hypothesis test for match or non-match. Additionally, the error ratio is less sensitive than the likelihood ratio to artifacts of

the modeling each of the distributions as it is a ratio of areas and not a ratio of heights. While likelihood ratio measures the strength of the evidence, the error ratio gives complementary information, about the quality of the match in a score-based framework (given the trace and the databases), to interpret the observed value of the evidence. The likelihood ratio and the error ratio cannot be used instead of each other as they have different meanings and different evolutions.

8.3 Future directions

The work presented in this thesis can be extended in several directions:

- Combining the strength of evidence using aural-perceptive and acoustic-phonetic approaches (aural-instrumental) of trained phoneticians with that of the likelihood ratio returned by the automatic system. In combining the strength of evidence, it is necessary to take into consideration, whether the characteristics considered in aural-instrumental and automatic methods are statistically independent, and if this is not the case, this interdependence should be taken into account when evaluating the overall likelihood ratio.

- One of the assumptions made in the statistical compensation methodology, is that information regarding conditions of recording of the databases used in the analysis is known to the expert or can be requested for. However, this may not always be the case, and it can be necessary for him to determine the recording conditions by himself. In such situations, methods for automatically determining the recording conditions of the recordings in a case can prove very useful.

- Databases for both casework and research in forensic speaker recognition are necessary in the data-driven Bayesian approach. While we have presented guidelines for the creation of a database to handle the problem of mismatched recording conditions, there is a need for protocols for creating more general, forensically realistic databases, to handle the various kinds of situations encountered in forensic speaker recognition casework.

- In this work, the mismatched conditions considered were those of the recording conditions. However, mismatch in languages, and the linguistic content can also affect the strength of evidence, and it is necessary to estimate and compensate for the uncertainty they introduce in the strength of evidence.

Description of the Polyphone IPSC-01 database

A

This database has been recorded with 32 Swiss-French speakers; 8 pairs of females whose voices sound similar and 8 pairs of males whose voices sound similar. Each of the individuals in the pair where related to each other, like mother and daughter, sisters (including twins), father and son, brother (including twins). Details about the contents of this database are available in [Meuwly, 2001] and [Meuwly et al., 2003]. In summary, it contains:

- The *suspected speaker reference data set (R)*, used to model the speech of the suspected source of the trace.
- The *suspected speaker control data set (C)*, used to evaluate the within-source variability of the suspected source of the trace.
- The *forensic traces data set (T)*, used to reproduce traces recorded in forensic conditions.

The details of each of these three databases are as follows:

- *Suspected speaker(s) reference database (R)*: This database contains 7 reference samples (80 to 120 seconds of read and spontaneous speech) per speaker, recorded over a period of one to three months. 6 of these samples are recorded through the PSTN network and 1 over the Global System for Mobile Communication (GSM) network.
- *Suspected speaker(s) control data set (C)*: Contains 13 to 47 samples of spontaneous speech (1 to 28 seconds) depending on the speaker. All of these samples are recorded through the PSTN in a single session.

- *Trace data set (T)*: Contains recordings made under realistically reproduced forensic conditions. It is constituted by 6 samples of spontaneous speech per speaker (5 through PSTN and 1 through GSM), 1 anonymous call without disguise per speaker (through PSTN), and 1 anonymous call with free disguise per speaker (through PSTN).

In addition, the potential population data set (P), used to model the between-sources variability of all the potential sources of the trace. This database was chosen from the Swisscom Swiss-French Polyphone database that contains one session each of read and spontaneous speech (80 to120 seconds) for 4500 speakers (2500 females and 200 males) recorded through the public switched telephone network (PSTN). This database is available from the European Languages Resources Association (ELRA).

Additionally, one of the spontaneous speech samples for every speaker, recorded through the PSTN is used to as input to generate 8 noisy samples with different signal-to-noise ratios: 30, 24, 18, 12, 9, 6, 3 and 0 dB.

This database was recorded by Didier Meuwly [Meuwly et al., 2003]at the Institut de Police Scientifique, University of Lausanne in collaboration with the Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne.

Description of the Polyphone IPSC-02 database

B

This is a database of 10 speakers, five of whom are native French speakers and the other five are native German speakers. All the speakers are also bilingual, speaking both French and German. For each speaker, a complete set of the suspected speaker reference database (R),a suspected speaker control database (C),and a suspected speaker trace database have been created (T), in each of three different recording conditions, i.e., PSTN and GSM recorded digitally, and PSTN recorded on analog tape using an answering machine.

Most of the recordings were contemporaneous with a majority of the sessions completed in a day. The entire set of recordings for this database was made during a five-day period between January 22, 2003, and January 27, 2003.

Technical Specification

The specifications of the handsets used are as follows:

- *PSTN telephone handset :* ASCOM®Eurit 133 wireless RNIS (*DECT* standard)
- *GSM telephone handset:* Siemens M50 with the Swiss telephone operator, Orange®.

The remote recordings were made both on a digital voice recording server as well as on an analog telephone answering machine.

- The *digital recordings* were made on a Sun®UltraSparc®, with a SunISDN-BRI/SBITM sound acquisition card, recording format G.711 Logiciel Sunlink ISDN 1.0 from Sun Microsystems.
- The *analog tape recordings* were made on a Panasonic ®Auto-Logic EASA Phone (microprocessor control) Model KX-T1442B5, from the Matsushita Electric Ind. Co. Ltd.

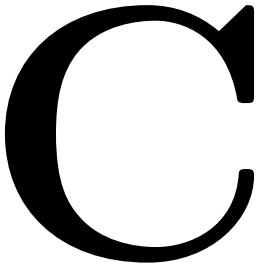
Nomenclature

1. Sex (M/F)
2. Speaker number
3. Language: FR = French/GE = German
4. Database type R = Reference / C = Control / T = Trace.
5. Recording session number
6. Kind of network (F = fixed network (PSTN) C = cellular network (GSM), A = analog tape-recording)
7. Recording number (e.g., suspected speaker control recordings 01 to 20).
8. Speaking style: (AN: Anonymous Normal, AD: Anonymous Disguised, NO: Normal, SP: Spontaneous, DL: Dialog)

e.g., M01FRR1F01_AN.WAV

This work was recorded as part of a study on the simulation of real cases for the recognition of speakers done by Quentin Rossy [Rossy, 2003], in collaboration with the Speech Processing and Biometrics Group, EPFL.

Description of the Polyphone IPSC-03 database



This database for forensic speaker recognition was recorded by the Institut de Police Scientifique (IPS), University of Lausanne, and the Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne (EPFL). It contains speech from 73 male speakers, in three different recording conditions and several different controlled and uncontrolled speaking modes. This database was recorded between January and June 2005.

The recordings for the database were made in controlled conditions, in a quiet room, located in the IPS and Ecole des Science Criminélles (ESC) building of UNIL.

The recording conditions of this database include transmission through a public switched telephone network (PSTN), a global system for mobile communications (GSM) network as well as calling-room acoustic conditions. The recording room contained a fixed line (PSTN) and mobile (GSM) telephone, and they both used the Swisscom ®telephone network provider.

The fixed line telephone instrument was a 'Meridian, Northern Telecom ®', and the mobile handset was a Nokia ®8310.

All the telephone calls were made from the recording room to an ISDN server located at the Signal Processing Institute, EPFL. The European ISDN (DSS1) transmission standard was used, and an answering machine application was used to record the calls. The transmitted speech was sampled at 16,000 Hz and recorded as 16-bit linear PCM Microsoft WAV files.

Along with these recordings, a third recording condition was simulated using a microphone and recorder placed directly in the recording room. The subjects spoke into a Sony ®electret condenser microphone (CARDIO ECM-23), placed at a distance

of about 30cm from the mouth of the speaker and connected to a Sony ®portable digital recorder (ICD-MS1). This speech was recorded in MSV format, at a sampling rate of 11,025 Hz. The cues were presented to the subjects in the form of a printed Microsoft®PowerPoint presentation (in order to avoid introducing the sound of a computer in the room), and care was taken to ensure that the recording room was free of any additional sound-adsorbing material.

The recorded speakers were male, aged between 18 and 50, with a majority being university-educated students, assistants (between 18 and 30 years of age) and faculty from within IPS and EPFL. All the utterances were in French.

The recordings of telephonic speech were made in two sessions with each of the subjects, the first using the PSTN (fixed) recording condition and the second using the GSM (mobile) recording condition. Additionally, two direct recordings, per speaker, were made on the microphone-recorder (digital) system described above. The length of each of these recordings was 10-15 minutes. Thus, four recordings were obtained, per speaker, in three different recording conditions (one in PSTN, one in GSM and two in room acoustic conditions). These conditions were called *Fixed*, *Cellular* and *Digital* respectively.

In addition to the actual text to be read out, the cue sheets contained detailed instructions for completing the task of speaking in three distinct styles. The first of these was the *normal* mode which involved simply reading printed text. The second *spontaneous* mode involved two simulated situations of a death threat call and a call to the police informing them of the presence of a bomb in a toilet. The third (*dialog*) mode involved reading a text in the tone of a conversation. The recordings (MSV format on the recorder and WAV on the answering machine) were edited with CoolEdit Pro 2®and grouped into various "subfiles" as described below.

The database contains 11 traces, 3 reference and 3 control recordings, grouped as the T, R and C sub-databases respectively.

- The T database consists of 9 files with read text and 2 *spontaneous* files. These 9 files are edited into three groups of 3 files, each having similar linguistic content. The *spontaneous* files are the simulations of calls as described earlier.
- The R database consists of recordings of read text only. Two of these recordings are identical one to the other. The content of the R database is similar to the IPSC-01 and IPSC-02 databases.
- The C database consists of recordings in the three different modes described above, viz., the *normal*, *spontaneous* and *dialog*.

A total of 73 speakers was recorded for this database, and it should be noted that the recordings for 63 of these are complete, with the four sets of recordings, i.e.,

PSTN, GSM, and two sets of acoustic room recordings. For the remaining 10 are only partially complete and for whom the fourth set of acoustic-room recordings, not available (for technical reasons).

The lengths of the recordings vary from a few seconds for the shortest (T40 and T50) to approximately two minutes for the longest (R01 and R02). This represents a total recording time of approximately 40 to 45 hours.

The nomenclature of the files can be described with an example:

Speaker No.1, in the $fixed$ condition, for the file $Control02$, with speech in the *normal* mode, is called $M001FRFC02\_NO.wav$.

The individual parts of the filename represent:

- M the sex of the speaker - Male
- 001 the chronological 'number' of the speaker; this number goes from 001 to 073.
- FR the language of speech - French
- C to denote it is a *control* recording. This is replaced by 'T' for the *trace* and by 'R' for the $reference$ recordings.
- 02 the number of the subfile. This number can take the values 10, 11, 12, 20, 21, 22, 30, 31, 32 for the T recordings; 00, 01 and 02 for the R recordings; and 01, 02 and 03 for the C recordings.
- NO the mode of the speech referring to the $Normal$ speaking mode. These letters are replaced by SP for the *spontaneous* and by DL for *dialog* mode.

The layout of this database is illustrated in Fig. C.1.

Credits

This database was recorded by Philipp Zimmerman, Damien Dessimoz, and Filippo Botti of the Institut de Police Scientifique, University of Lausanne, and Anil Alexander and Andrzej Drygajlo of the Signal Processing Institute, EPFL [Zimmermann, 2005].

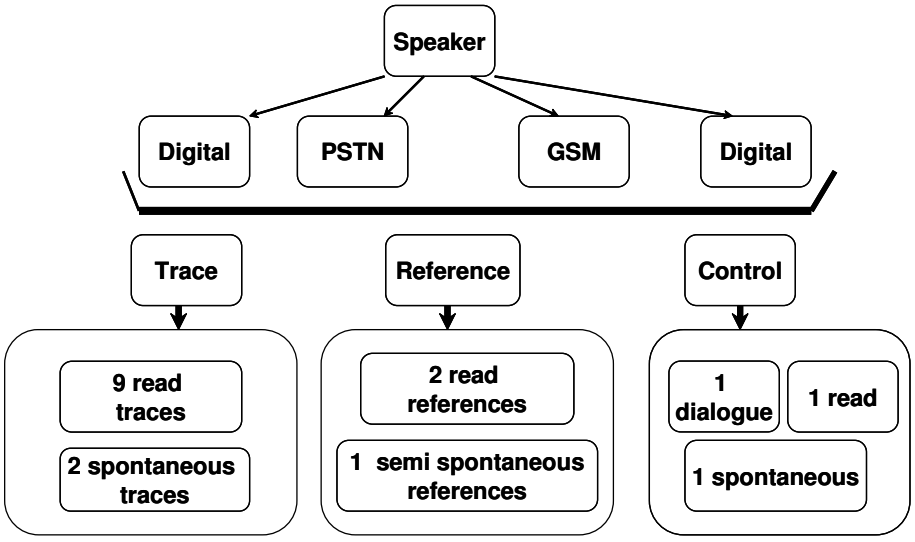


Figure C.1: *Layout of the IPSC03 database*

Netherlands Forensic Institute Speaker recognition evaluation

D

D.1 NFI speaker recognition evaluation through a fake case

Twelve audio recordings were provided, by the Netherlands Forensic Institute (NFI) as part of a fake case evaluation, consisting of 2 reference recordings and 10 questioned recordings. The 10 questioned recordings consist of conversations between two speakers, i.e., each containing the speech of a known speaker and an unknown speaker. The 2 reference recordings consist of two conversations between a known speaker and a suspected speaker.

The aim of the analysis is to determine whether the speech of the unknown speaker in each of the questioned recordings was produced by the suspected speaker in the reference recordings.

D.1.1 Correspondence

- **July 31, 2004**: Discussion and agreement by Anil Alexander and Andrzej Drygajlo to participate in the Fake Case Evaluation conducted by the Netherlands Forensic Institute, presented at the annual conference of the International Association for Forensic Phonetics and Acoustics, 2004, Helsinki, Finland.

- **November 2004**: A CD-ROM containing 12 audio recordings, in digital format along with corresponding transcripts for each of the recordings was received at the Speech Processing and Biometrics Group at Lausanne, Switzerland.

- **February 10, 2005**: Queries related to the recording conditions and content of the questioned recordings and the two reference recordings were submitted to NFI via e-mail. We requested for confirmation whether the two reference recordings contained the same suspected speaker. We also inquired whether it was possible to record the suspected speaker under controlled conditions at our laboratory.

- **February 14, 2005**: Response to the queries, detailing the recording conditions of all the recordings as the digital recordings of speech that had passed through a fixed telephone channel was received from NFI. It was confirmed that it was the suspected speaker present in both the reference recordings were the same person. It was expressed that recording the suspected speaker under controlled conditions was not possible.

D.1.2 Description of items submitted for analysis

1 CD-ROM with 12 recordings in 16 KHz, 16-bit Linear PCM wave files. According to the accompanying documentation, these recordings were recorded directly from a phone line onto a Digital Audio Tape (DAT), at 44 kHz and then down-sampled to 16 kHz and later transferred to a computer using Cool Edit Pro 2.0. 2. Transcripts of all the audio recordings. Detailed transcriptions of the recordings with the corresponding dates, time and telephone numbers.

D.1.3 Plan of Work

The evaluation process constituted of the two following steps:

- **Pre-processing**: Pre-processing consisting of segmentation of the audio into the speech of individual speakers, removal of non-speech regions and selection of reference and control recordings was performed, in order to prepare the recordings for the analysis. We ascertained, through email correspondence with NFI, that all of the recordings provided were performed with a fixed telephone and that there was no mobile (GSM) channel effect in the recording conditions of all the recordings in the case. Because of this, no attempt at compensating for mismatched conditions was made.

- **Analysis**: The Bayesian interpretation methodology [Drygajlo et al., 2003; Meuwly, 2001]for the forensic automatic speaker recognition, as explained in the Section Methodology, was used for this case. The result of the comparison between the model of the suspect speaker and the questioned recording (called evidence (E)) is evaluated given two hypotheses: H_0 - the suspect is the

source of the questioned recording H_1 - anyone else in the relevant population is the source. The result of this evaluation is the strength of evidence, which is expressed as a likelihood ratio (LR).

A second analysis was performed using another approach to the Bayesian interpretation methodology [Botti et al., 2004a] as a control experiment. The recordings used consisted exclusively of recordings obtained from the case. The recordings in the case were used in order to estimate the distributions of scores obtained when comparing two recordings have the same speaker as source, as well as when two recordings that have two different speakers as their sources. Details of this analysis are presented in Appendix 2.

Pre-Processing

1. **Acquisition and Down-sampling**: Acquisition was unnecessary as the files were already in digital format. However, in order to maintain consistency with the other databases used for comparison, it was necessary to down-sample the audio files to 8 kHz, 16-bit Linear PCM files using Cool Edit Pro 2.0.

2. **Segmentation**: The questioned recordings and the reference recordings were in the form of conversations between two speakers. In order to compare the speech of individual speakers it was necessary to segment each of the conversations. This segmentation was performed aurally, with the help of the transcripts provided. Zones of overlap, laughs, and other anomalies were discarded.

3. **Removal of Non-Speech Regions**: The recordings were passed through a voice activity detector (VAD), which separates speech and non-speech regions, using instantaneous signal to noise ratio (SNR). This algorithm is described in [Reynolds, 1992]. The non-speech regions of the recording contain more information about the conditions of ambient noise present in the recording and no speaker-dependent information. Removal of these non-speech regions better allows for speaker specific characteristics to be considered, when modelling voice.

4. **Selection of Reference and Control Recordings**: The reference recordings of the suspected speaker were divided into reference recordings that would be used for training statistical models of his speech and control recordings that are compared with the models of the suspected speaker, in order to estimate the within-source variability of his speech [section Methodology].

D.1.4 Methodology

In the first part of the analysis, the Bayesian interpretation methodology presented in [Drygajlo et al., 2003; Meuwly, 2001], was used as a means of calculating the strength of evidence.

This Bayesian methodology requires, in addition to the questioned recording, the use of three databases: a suspected speaker reference database (R), a suspected speaker control database (C), a potential population database (P), and the questioned recording database (T).

- The P database contains an exhaustive coverage of recordings of all possible voices satisfying the hypothesis (H_1): anyone chosen at random from a relevant population could be the source of the trace. These recordings are used to create models of speakers to evaluate the between-sources variability given the trace.

- The R database contains recordings of the suspected speaker that are as close as possible (in recording conditions and linguistically) to the recordings of speakers of P, and it is used to create the suspected speaker's model, exactly as is done with models of P.

- The C database consists of recordings of the suspected speaker that are very similar to the trace, and it is used to estimate the within-source variability of the voice of the suspected speaker.

A brief summary of the steps required in order to calculate the evidence and its strength for a given trace is as follows:

1. The trace is compared with the statistical model of the suspect (created using database R), and the resulting score is the evidence score (E).

2. The trace is compared with statistical models of all the speakers in the potential population (P). The distribution of obtained scores is an estimation of the between-sources variability of the trace with the potential population.

3. The suspected speaker control database (C) recordings are compared with the models created with the suspected speaker reference database (R) for the suspect, and the distribution of scores obtained estimates the suspect's within-source variability.

4. The likelihood ratio (LR) (i.e., the ratio of support that the evidence (E) lends to each of the hypotheses), is given by the ratio of the heights of the within-source and between-sources distributions at the point E.

In this analysis, the potential population database (P) used is the PolyCOST 250 database [Hennebert et al., 2000]. We have used 73 speakers from this database in the analysis. The suspected speaker reference database consisted of two reference files of duration 2m 19s, and 2m 17s. The suspected speaker control databases consisted of 7 recordings of 20 seconds duration each. This database contains mainly European speakers who speak both in English and in their mother-tongue. This database was chosen among the available databases because it was found to be best suited to the case, especially in the language (English spoken by European speakers) and technical conditions (fixed European telephone network) under which the reference recordings of the suspect were made.

Note that an accurate estimation of the likelihood ratio in a Bayesian framework is possible, only if the technical conditions of the suspected speaker reference (R) and potential population (P) databases are identical, and the suspected speaker control database (C) was recorded in the same conditions as the questioned recording. More explicitly, following assumptions have to be satisfied, 'The suspected speaker control database and the questioned recording were recorded in similar conditions.' The suspected speaker reference database and the potential population database were recorded in similar recording conditions.

In practice, it can be observed that it is very difficult to satisfy all these requirements. Incompatibilities in the databases used can result in under-estimation or over-estimation of the likelihood ratio [Alexander et al., 2004]. In order to consider the recording conditions of the case, in the control experiment presented in Appendix 2, we evaluate the strength of evidence using recordings exclusively from the case, and check whether the choice of the databases used significantly affect the results obtained.

D.1.5 Technical Analysis

During the preprocessing step the recordings were segmented into recordings of individual speakers. Each of the segmented recordings contains only the speech of a specific speaker. This segmentation was performed aurally with the help of the transcripts provided. The set of recordings obtained, along with their durations, is presented in Table D.1.

The files R01_Peter.wav and R02_Peter.wav were further segmented into:

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

Table D.1: Individual speakers segments and their durations

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

R02_Peter_C02.wav and R02_Peter_C03.wav each of 20s each (C database).

The P database used is the COST 250 database [Hennebert et al., 2000]. We have used 73 speakers from this database in the analysis.

The following analysis procedure was then applied to the R, C and P databases thus created:

- Analysis and creation of models of the speakers voice: Extraction of 12 RASTA-PLP features for each analysis frame [Hermansky, 1994] and creation of a statistical model by means of a 64 component Gaussian mixture model (GMM) Reynolds [1995].

- Within-source variability estimation: Comparison between statistical model of the features of the reference recording and the features of the control recordings of the suspected speaker.

- Between-sources variability estimation: Comparison between the features of the questioned recording and the statistical models of the voices of the speakers from the database representing the potential population.
- Calculation of the evidence score: Comparison between the questioned recording and the model of the suspected speaker.
- Calculation of the strength of evidence: Calculation of the likelihood ratio by evaluating the relative likelihood ($\frac{p(E|H_0)}{p(E|H_1)}$) of observing the evidence score (E) given the hypothesis that the source of the questioned recording is the suspect (H_0) and the likelihood of observing the evidence score given hypothesis that someone else in the potential population was its source (H_1) [Robertson and Vignaux, 1995]. Kernel density estimation was used to calculate the probability densities of distribution of scores for each of the hypotheses.

D.1.6 Results of the Technical Analysis

Each of the ten questioned recordings ($Q1, ..., Q10$) is considered as a separate case. For each case, we consider the question, 'Is the speaker Peter, in the reference recordings $R1$ and $R2$, the same speaker as the unknown speaker in the questioned recording (Qn)?'

The following databases are used in each case:

- Potential Population Database (P): PolyCOST 250 Database,
- Reference Database (R): 2 reference recordings from the recordings R1 and R2, belonging to the suspected speaker, Peter
- Control Database (C): 7 control recordings from the recordings R1 and R2 belonging to the suspected speaker.

The distribution of scores for H_0 obtained when comparing the features of the suspected speaker control recordings (C database) of the suspected speaker, Peter, with the two statistical models of his speech (created using files from the R database) is represented by the red dotted line.

Case 1

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q1$?

Trace Database (T): Q01_NN_Male.wav containing the speech of the unknown speaker from the questioned recording Q1.

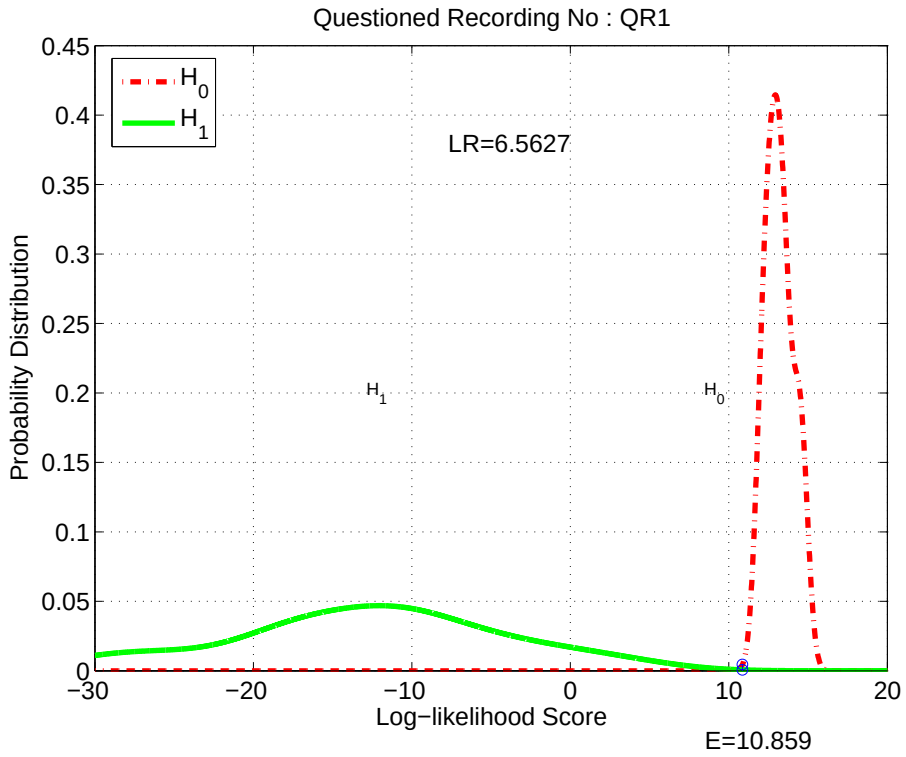


Figure D.1: *Case 1: Questioned recording Q01_NN_Male.wav*

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording Q1, corresponding to the unknown speaker (Q01_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.1.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.1, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 10.86.

A likelihood ratio of 6.56, obtained in Fig. D.1, means that it is 6.56 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

However, in the figure above, we observe that this value corresponds to the tails of the distributions of H_0 and H_1 scores (which is not statistically significant, at a 5% significance level). It is not possible to affirm that this score (E) would be representative of either of the distributions.

Case 2

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q2$?

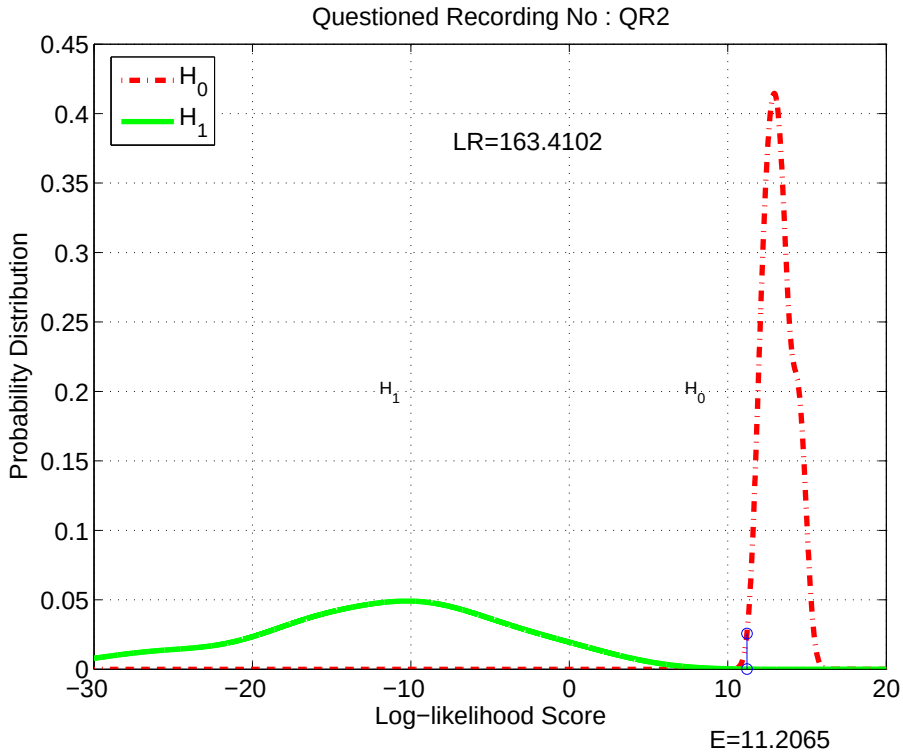


Figure D.2: *Case 2: Questioned recording Q02_NN_Male.wav*

Trace Database (T): Q02_NN_Male.wav containing the speech of the unknown speaker from the questioned recording Q2.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording $Q2$, corresponding to the unknown speaker (Q02_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.2.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.2, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 11.21.

A likelihood ratio of 163.41, obtained in Fig. D.2, means that it is 163.41 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

However, in the figure above, we observe that this value corresponds to the tails of the distributions of H_0 and H_1 scores (which is not statistically significant, at a 5% significance level). It is not possible to affirm that this score (E) would be representative of either of the distributions.

Case 3

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q3$?

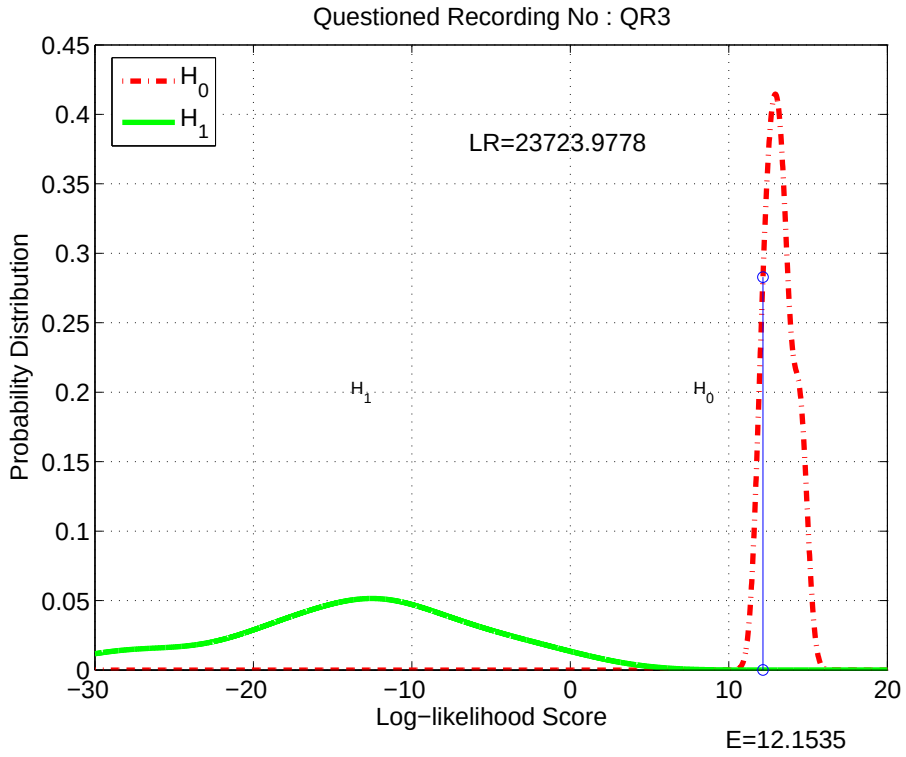


Figure D.3: *Case 3: Questioned recording Q03_NN_Male.wav*

Trace Database (T): Q03_NN_Male.wav containing the speech of the unknown speaker from the questioned recording $Q3$.

Discussion

The distribution of scores for H1 obtained by comparing the segment of the questioned recording Q3, corresponding to the unknown speaker (Q03_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.3.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.3, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 12.15.

A likelihood ratio of 23723.98, obtained in Fig. D.2, means that it is 23723.98 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

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

Case 4

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q4$?

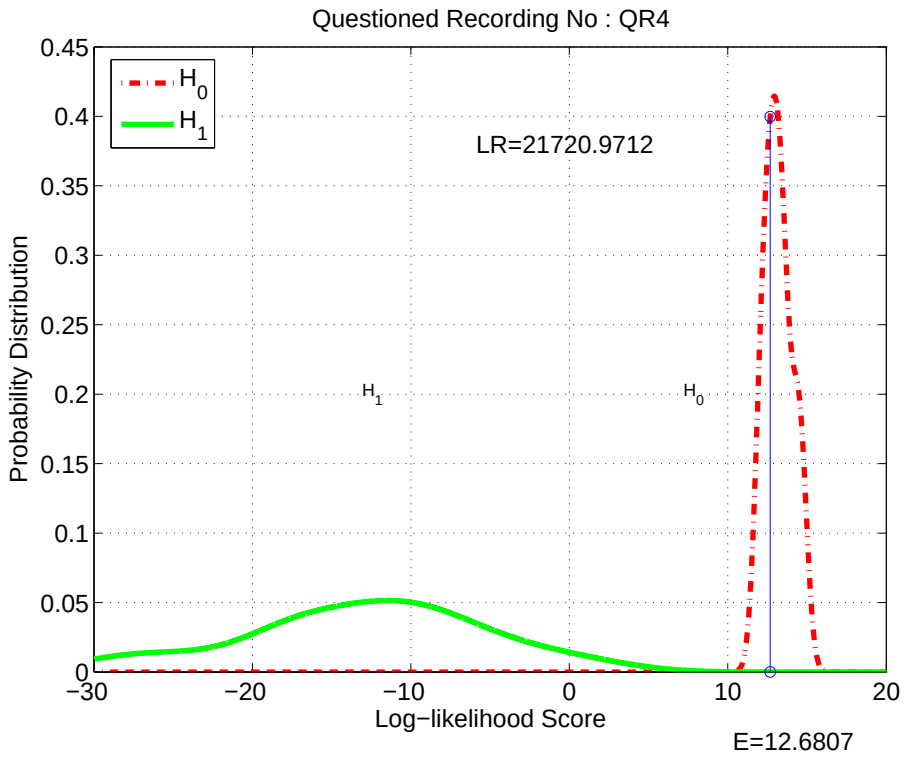


Figure D.4: *Case 4: Questioned recording Q04_NN_Male.wav*

Trace Database (T): Q04_NN_Male.wav containing the speech of the unknown speaker from the questioned recording $Q4$.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording Q_4, corresponding to the unknown speaker (Q04_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.4.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.4, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 12.68

A likelihood ratio of 21720.97, obtained in Fig. D.4, means that it is 21720.97 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

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

Case 5

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q5$?

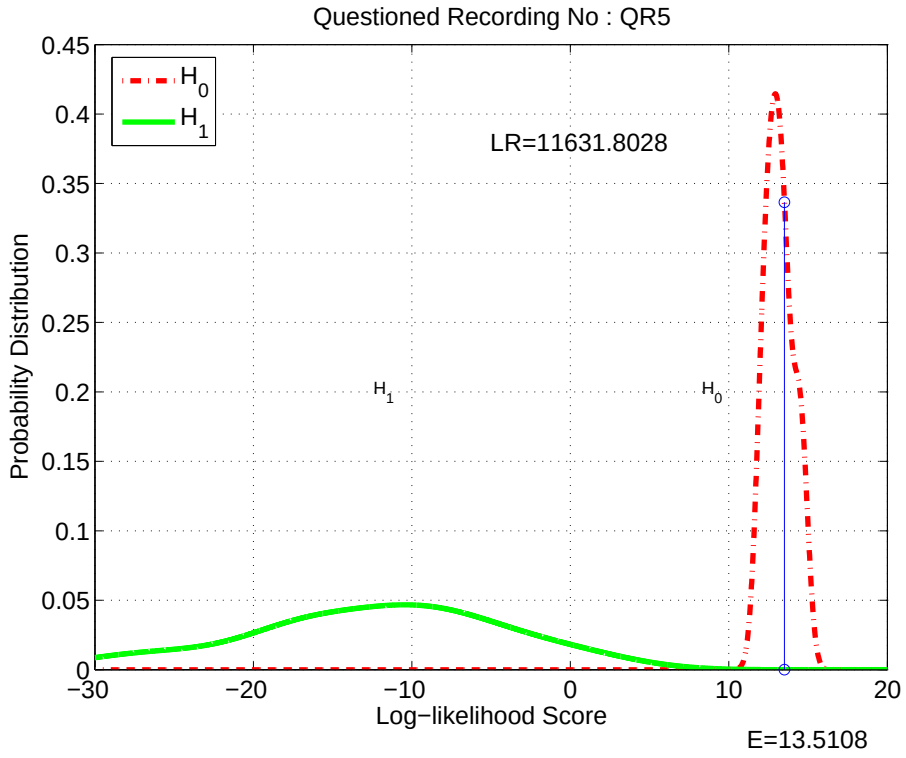


Figure D.5: *Case 5: Questioned recording Q05_NN_Male.wav*

Trace Database (T): Q05_NN_Male.wav containing the speech of the unknown speaker from the questioned recording $Q5$.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording $Q5$, corresponding to the unknown speaker (Q05_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.5.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.5, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 13.51.

A likelihood ratio of 11631.8, obtained in Fig. D.5, means that it is 11631.8 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

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

Case 6

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q6$?

Trace Database (T): Q06_NN_Male.wav containing the speech of the unknown speaker from the questioned recording $Q6$.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording $Q6$, corresponding to the unknown speaker (Q06_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.6.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.6, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 11.64.

A likelihood ratio of 329.0, obtained in Fig. D.6, means that it is 329.0 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

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

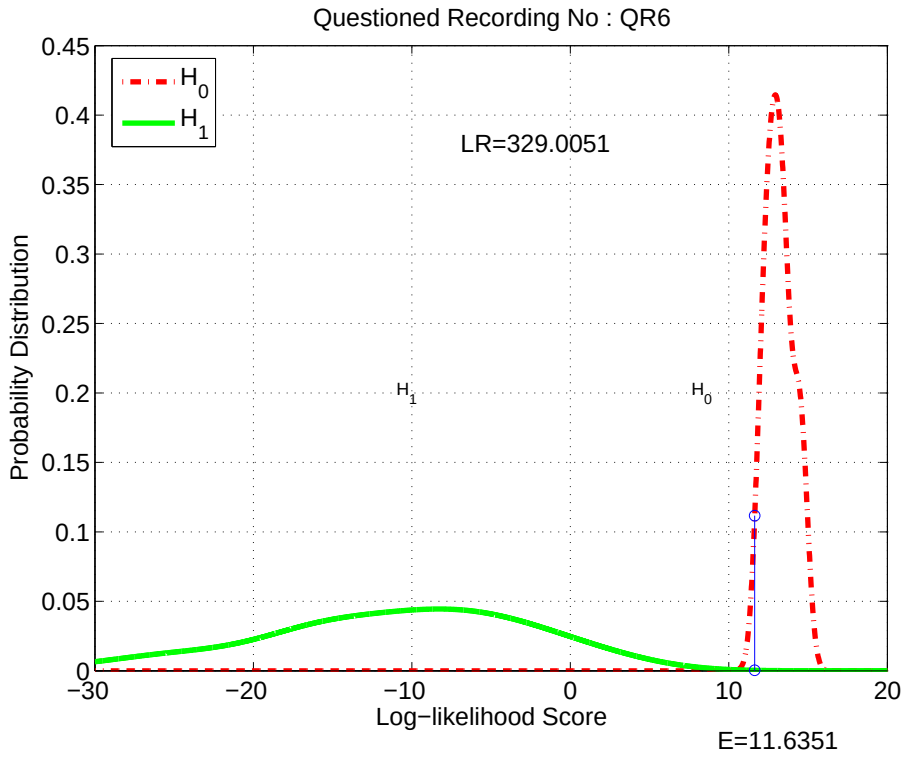


Figure D.6: *Case 6: Questioned recording Q06_NN_Male.wav*

Case 7

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q7$?

Trace Database (T): Q07_NN_Male.wav containing the speech of the unknown speaker from the questioned recording $Q7$.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording $Q7$, corresponding to the unknown speaker (Q07_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.7.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.7, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 12.48.

A likelihood ratio of 38407.33, obtained in Fig. D.7, means that it is 38407.33 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

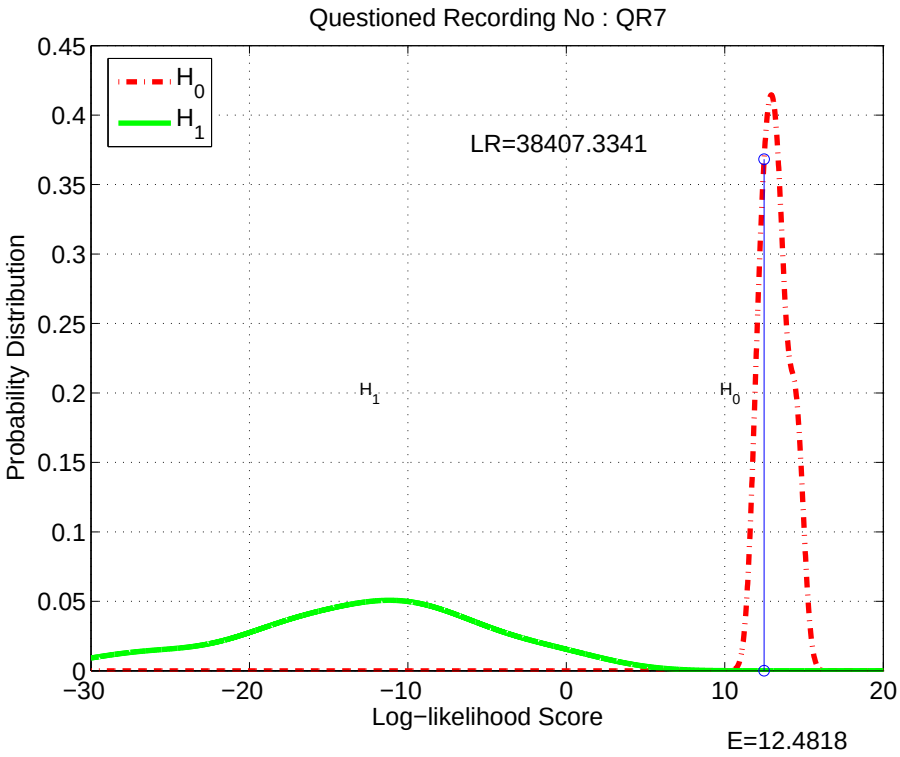


Figure D.7: *Case 7: Questioned recording Q07_NN_Male.wav*

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

It is to be noted however, that the length of the questioned recording Q07_NN_Male.wav is very short (6.4 seconds). This in turn reduces the relevance of the likelihood ratio.

Case 8

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q8$?

Trace Database (T): Q08_NN_Male.wav containing the speech of the unknown speaker from the questioned recording $Q8$.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording $Q8$, corresponding to the unknown speaker (Q08_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.8. The average score (E), represented by the point on the log-likelihood score axis in Fig. D.8, obtained by comparing the

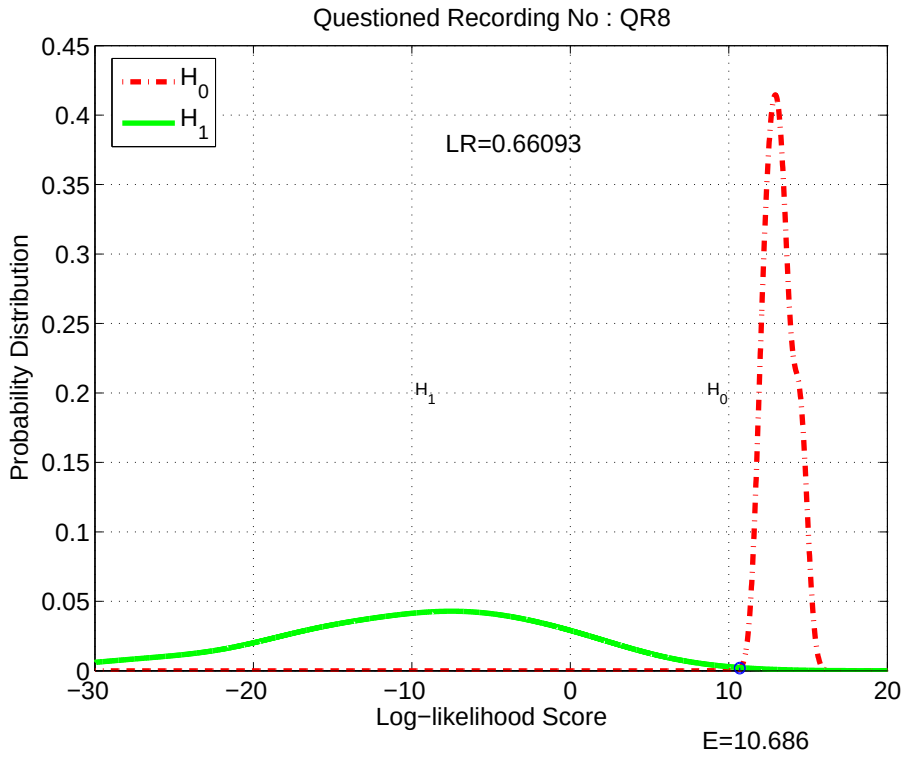


Figure D.8: *Case 8: Questioned recording Q08_NN_Male.wav*

questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 10.67.

The average score (E), represented by the blue line in Fig. D.8, obtained by comparing the questioned recording with the statistical models of the suspect speech, is 10.686.

A likelihood ratio of 0.660, obtained in Fig. D.8, means that it is 0.660 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

However, in the figure above, we observe that this value corresponds to the tails of the distributions of H_0 and H_1 scores (which is not statistically significant, at a 5% significance level). It is not possible to affirm that this score (E) would be representative of either of the distributions.

Case 9

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q9$?

Trace Database (T): Q09_NN_Male.wav containing the speech of the unknown

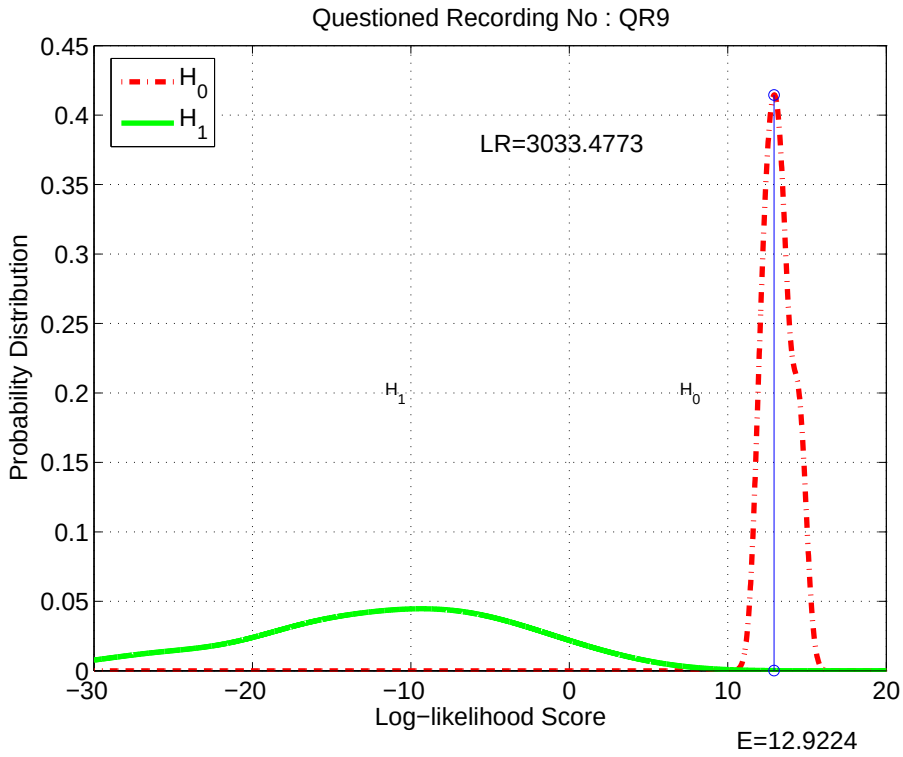


Figure D.9: *Case 9: Questioned recording Q09_NN_Male.wav*

speaker from the questioned recording $Q9$.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording $Q9$, corresponding to the unknown speaker (Q09_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.9.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.9, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 12.92.

A likelihood ratio of 3033.47, obtained in Fig. D.9, means that it is 3033.47 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

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

Case 10

Is Peter in the reference recordings ($R1$ and $R2$) the same speaker as the unknown speaker in the recording $Q10$?

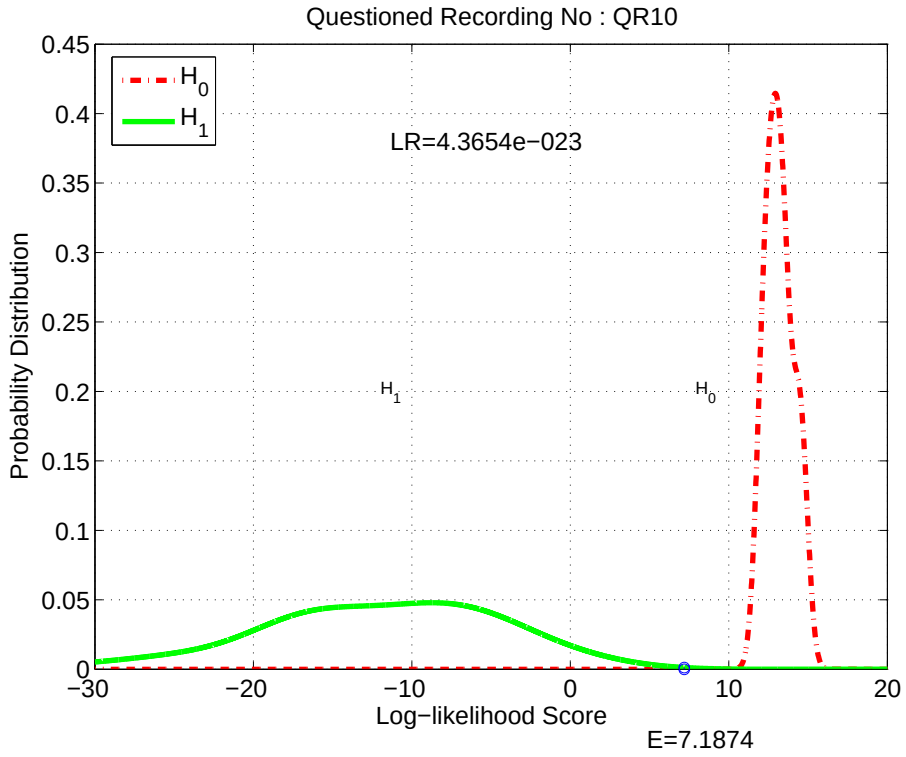


Figure D.10: *Case 10: Questioned recording Q10_NN_Male.wav*

Trace Database (T): Q10_NN_Male.wav containing the speech of the unknown speaker from the questioned recording $Q10$.

Discussion

The distribution of scores for H_1 obtained by comparing the segment of the questioned recording $Q10$, corresponding to the unknown speaker (Q10_NN_Male), with the Gaussian mixture models of the speakers of the potential population database (P) is represented by the green line in Fig. D.10.

The average score (E), represented by the point on the log-likelihood score axis in Fig. D.10, obtained by comparing the questioned recording with the Gaussian mixture models of the suspected speaker, Peter's speech is 7.18.

A likelihood ratio of 4.36 x 10-23, obtained in Fig. D.10, means that it is 4.36 x 10-23 times more likely to observe this score (E) given the hypothesis H_0 (the suspect is the source of the questioned recording) than given the hypothesis H_1 (that another speaker from the relevant population is the source of the questioned recording).

However, in the figure above, we observe that this value corresponds to the tails of the distributions of H_0 and H_1 scores (which is not statistically significant, at a 5% significance level). It is not possible to affirm that this score (E) would be representative of either of the distributions.

A summary of the results of the analyses is presented below in Table D.2.

Table D.2: LRs obtained using the PolyCOST database, in English, as the potential population

Trace No.	LR (P in English)	Correct
Q1	6.56	×
Q2	163.41	✓
Q3	23723.98	✓
Q4	21720.97	✓
Q5	11631.8	✓
Q6	329.0	✓
Q7	38407.33	✓
Q8	0.660	✓
Q9	3033.47	✓
Q10	4.36 x 10-23	✓

D.1.7 Conclusion

The aim of the analysis is to determine whether the recordings of unknown speakers, in the ten questioned recordings, was produced by the suspected speaker, named Peter in the reference recordings.

Note that the conclusions given below are a summary and synthesis of the results from the analyses taking into consideration the likelihood ratios as well as other factors such the length and content of the recordings and the statistical significance of the results. These factors may influence the statement of the conclusions from the likelihood ratio or do not allow us to progress the case in any direction.

1. Although the likelihood ratio constitutes limited support for the hypothesis that the recording of the unknown speaker (Q01_NN_Male) in the questioned recording ($Q1$) and the reference recordings from the suspected speaker, Peter ($R1$ and $R2$), have the same source, the statistical significance analysis does not allow us to progress the case in any direction.

2. Although the likelihood ratio constitutes strong support for the hypothesis that the recording of the unknown speaker (Q02_NN_Male) in the questioned recording ($Q2$) and the reference recordings from the suspected speaker, Peter ($R1$

and $R2$), have the same source, the statistical significance analysis does not allow us to progress the case in any direction.

3. The likelihood ratio obtained constitutes very strong support for the hypothesis that the recording of the unknown speaker (Q03_NN_Male) in the questioned recording ($Q3$) and the reference recordings from the suspected speaker Peter ($R1$ and $R2$), have the same source.

4. The likelihood ratio obtained constitutes very strong support for the hypothesis that the recording of the unknown speaker (Q04_NN_Male) in the questioned recording ($Q4$) and the reference recordings from the suspected speaker Peter ($R1$ and $R2$), have the same source.

5. The likelihood ratio obtained constitutes very strong support for the hypothesis that the recording of the unknown speaker (Q05_NN_Male) in the questioned recording ($Q5$) and the reference recordings from the suspected speaker Peter ($R1$ and $R2$), have the same source.

6. The likelihood ratio obtained constitutes strong support for the hypothesis that the recording of the unknown speaker (Q06_NN_Male) in the questioned recording ($Q6$) and the reference recordings from the suspected speaker Peter ($R1$ and $R2$), have the same source.

7. The likelihood ratio obtained constitutes very strong support for the hypothesis that the recording of the unknown speaker (Q07_NN_Male) in the questioned recording ($Q7$) and the reference recordings from the suspected speaker Peter ($R1$ and $R2$), have the same source. However, the duration of the recording is too short and diminishes the relevance of the likelihood ratio.

8. Although the likelihood ratio constitutes limited support for the hypothesis that the recording of the unknown speaker (Q08_NN_Male) in the questioned recording ($Q8$) and the reference recordings from the suspected speaker, Peter ($R1$ and $R2$), have a different source, the statistical significance analysis does not allow us to progress the case in any direction.

9. The likelihood ratio obtained constitutes very strong support for the hypothesis that the recording of the unknown speaker (Q09_NN_Male) in the questioned recording ($Q9$) and the reference recordings from the suspected speaker Peter ($R1$ and $R2$), have the same source.

10. Although the likelihood ratio constitutes very strong support for the hypothesis that the recording of the unknown speaker (Q10_NN_Male) in the questioned recording ($Q10$) and the reference recordings from the suspected speaker, Peter

($R1$ and $R2$), have a different source, the statistical significance analysis does not allow us to progress the case in any direction.

Appendix 2

This part of analysis was performed with each of the cases in order to see whether the likelihood ratios obtained using the PolyCOST 250 database would correspond to the likelihood ratios obtained using only data from the case. We have performed this analysis as control experiment. It is to be noted that in this analysis the amount of data used is limited and hence the conclusions drawn cannot be considered in the estimation of the strength of evidence except as complementary validation. The following hypotheses were evaluated with respect to the data obtained from the case:

- H_0 = The two recordings have the same source
- H_1 = The two recordings have different sources

From the recordings present in the case, it was possible to extract the speech of three distinct speakers, Eric, Jos, and Peter (according to the transcripts). The segments of the recordings belonging to each of these speakers had been separated according to the transcripts. It is possible, then, to use these speakers in order to estimate the scores that would be obtained if two recordings have the same source, as well as if two recordings have different sources. The advantage of using these speakers from the case is that we will be able to estimate these scores in exactly the same recording conditions as that of the case. The bias due to mismatched recording conditions can be avoided. However, the disadvantage is that the number of speakers as well as the number of recordings that can be used for each speaker, in this analysis, is limited. Still, the distributions of scores from this analysis can be used to compare the results obtained in the analysis with the relevant potential population. It can be verified as to whether the trends in the likelihood ratios, obtained for each of the cases, in the first analysis Bayesian interpretation framework, are similar to those obtained in the second analysis. Using different sections of the audio files we had for each of these speakers as reference and control recordings, it was possible to create a set of pair-wise comparisons, corresponding to each of the hypotheses (H_0 and H_1). Thus, it was possible to estimate the log-likelihood scores obtained when the two recordings came from the same speaker, in exactly the recording conditions corresponding to the case. Note that it was necessary to exclude the files corresponding to the speaker Jos, as there was an insufficient amount of data to create a reference model, as well as to obtain a number of control recordings of reasonable length.

The likelihood ratios from the estimated are summarized below: In these cases

Case No	Source recording	Questioned recording	Score	Likelihood Ratio	Verbal equivalent of the LR
1	Q1	Q01_NN_Male.wav	10.86	0.356	weak support H1
2	Q2	Q02_NN_Male.wav	11.20	0.770	weak support H1
3	Q3	Q03_NN_Male.wav	12.15	3.376	weak support H0
4	Q4	Q04_NN_Male.wav	12.68	5.885	weak support H0
5	Q5	Q05_NN_Male.wav	13.51	13.189	moderate support H0
6	Q6	Q06_NN_Male.wav	11.63	1.662	weak support H0
7	Q7	Q07_NN_Male.wav	12.48	4.834	weak support H0
8	Q8	Q08_NN_Male.wav	10.68	0.229	weak support H1
9	Q9	Q09_NN_Male.wav	12.92	7.404	weak support H0
10	Q10	Q10_NN_Male.wav	7.187	3.69x10-9	strong support H1

we observe that, similar trends in the likelihood ratios are observed in the control experiments to those in the first analysis method. We observe that for most of the cases, where relatively lower likelihood ratios are observed in the first analysis method, similar low values for the likelihood ratios are observed in the control analysis, and relatively higher likelihood ratios observed in the first analysis correspond to similar higher values in the second analysis. For the questioned recordings, $Q1$, $Q2$, $Q8$ and $Q10$, we observe lower values for the likelihood ratio in the second phase of analysis as in the first one. However, in the second phase results, we observe that the values of E do not suffer from being on the tails of the distributions H_0 and H_1, and are statistically significant (at a 5% significance) level. Thus these likelihood ratios (for $Q1$, $Q2$, $Q8$ and $Q10$) support the hypothesis H_1, that the questioned recording may have a different source from the reference recordings of the suspected speaker. We also observe that the likelihood ratios pertaining to the other questioned recordings in this second analysis supporting the hypothesis H0 (same source) are lower than the corresponding likelihood ratios in the first analysis, although they show the same trends.

Bibliography

Aitken, C. (1997). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, Chichester.

Aitken, C. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, Chichester.

Alexander, A., Botti, F., and Drygajlo, A. (2004). Handling Mismatch in Corpus-Based Forensic Speaker Recognition. In *Odyssey 2004, The Speaker and Language Recognition Workshop*, pages 69–74, Toledo, Spain.

Alexander, A., Drygajlo, A., and Botti, F. (2005). NFI: Speaker recognition evaluation through a fake case. Case report, Swiss Federal Institute of Technology, Lausanne (EPFL), Lausanne, Switzerland.

Arcienega, M., Alexander, A., Zimmerman, P., and Drygajlo, A. (2005). A bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition. In *Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*, pages 2009–2012, Lisbon, Portugal.

Arcienega, M. and Drygajlo, A. (2003). A Bayesian network approach for combining pitch and reliable spectral envelope features for robust speaker verification. In Kittler, J. and Nixon, M. S., editors, *Proc. 4th Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 78–85, Guildford, UK. Springer.

Auckenthaler, R., Carey, M. J., and Lloyd-Thomas, H. (2000). Score normalisation for text-independent speaker verification system. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 Speaker Recognition Workshop*, 10 (1-3):42–54.

Beck, S. D. (1998). *FBI Voice Database For Automated Speaker Recognition Systems, User's Manual for Training and Testing*. Federal Bureau of Investigation.

Beck, S. D., Schwartz, R., and Nakasone, H. (2004). A bilingual multi-modal voice corpus for language and speaker recognition (LASR) services. In *Odyssey 2004, The Speaker and Language Recognition Workshop*, pages 265–270, Toledo, Spain.

Ben, M., Blouet, R., and Bimbot, F. (2002). A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacrétaz, D., and Reynolds, D. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Walten street, Oxford.

Bolle, R. M., Ratha, N. K., and Pankanti, S. (2004). Error analysis of pattern recognition systems: the subsets bootstrap. *Comput. Vis. Image Underst.*, 93(1):1–33.

Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M., and Stevens, K. N. (1970). Speaker identification by speech spectrograms: A scientist's view of its reliability for legal purposes. *J. Acoustic. Soc. Am.*, 47(2):597–603.

Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M., and Stevens, K. N. (1973). Speaker identification by speech spectrogram: some further observations. *J. Acoust. Soc. Amer.*, 54(2):531–537.

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J. P., Reynolds, D. A., and Magrin-Chagnolleau, I. (2003). Auditory Instrumental Forensic Speaker Recognition. In *Proc. Eurospeech 2003*, pages 33–36, Geneva, Switzerland.

Botti, F., Alexander, A., and Drygajlo, A. (2004a). An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data. In *Proceedings of 2004: A Speaker Odyssey*, pages 63–68, Toledo, Spain.

Botti, F., Alexander, A., and Drygajlo, A. (2004b). On compensation of mismatched recording conditions in the bayesian approach for forensic automatic speaker recognition. *Forensic Science International*, 146(Supplement 1,2):S101–S106.

Bouten, J. S. and van Leeuwen, D. A. (2004). Results of the 2003 NFI-TNO forensic speaker recognition evaluation. In *Odyssey 2004, The Speaker and Language Recognition Workshop*, pages 75–82, Toledo, Spain.

Bregman, A. (1990). *Auditory Scene Analysis*. MIT Press, Cambridge Mass.

Byrne, C. and 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. (2005). Netherlands Forensic Institute (NFI), speaker recognition fake case evaluation. June 2-3, 2005, 8th Meeting of the ENFSI Expert Working Group for Forensic Speech and Audio Analysis (FSAAWG).

Campbell, J. (1997). Speaker recognition: A tutorial. *Proceedings of IEEE*, 85:1437–1462.

Campbell, W., Reynolds, D., Campbell, J., and Brady, K. (2005). Estimating and evaluating confidence for forensic speaker recognition. In *ICASSP*, volume 1, pages 717–720, Philadelphia, USA.

Champod, C. and Evett, I. W. (2000). Commentary on: Broeders, A. P. A. (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics, 6(2): 228-41. *Forensic Linguistics*, 7.2:238–243.

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

Cook, R., Evett, I., Jackson, G., Jones, P. J., and Lambert, J. (1998). A model for case assessment and interpretation. *Science & Justice*, 38(3):151–156.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society (B)*, 39:1–38.

Dessimoz, D. (2004). Reconnaissance de locuteurs : Comparaison de performances entre la reconnaissance auditive par des profanes et de systemes automatiques. Project report, Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, Switzerland.

Doddington, G. (1985). Speaker recognition - identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., and Reynolds, D. A. (1998). Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *5th International Conference on Spoken Language Processing (ICSLP)*, page 0608, Sydney, Australia.

Drygajlo, A., Meuwly, D., and Alexander, A. (2003). Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition. In *Proc. Eurospeech 2003*, pages 689–692, Geneva, Switzerland.

Dunn, R. B., Quatieri, T. F., Reynolds, D. A., and Campbell, J. (2001). Speaker recognition from coded speech in matched and mismatched conditions. In *2001: A Speaker Odyssey*, Crete, Greece.

El-Maliki, M. (2000). *Speaker verification with missing features in noisy environments*. PhD thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland.

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science and Justice*, 38(3):198–202.

Evett, I. W. and Buckleton, J. S. (1996). Statistical analysis of STR data : Advances in forensic haemogenetics. *Springer-Verlag*, 6:79–86.

Farrell, K. R., Mammone, R. J., and Assaleh, K. T. (1994). Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on speech and audio processing*, 2(1):194–205.

Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *Forensic Linguistics*, 10(2):203–226.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press, second edition.

Furui, S. (1997). Recent advances in speaker recognition. *Pattern Recognition Letters*, 18(9):859–872.

Gauvain, J. L. and Lee, C. H. (1994). Maximum a-posteriori estimation of multivariate Gaussian mixture observations. *IEEE Trans. Speech, Audio Processing*, 2:291–298.

Gonzales-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., and Ortega-Garcia, J. (2005). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language (to appear)*.

Gonzalez-Rodriguez, J., Ramos-Castro, D., Garcia-Gomar, M., and Ortega-Garcia, J. (2004). On robust estimation of likelihood ratios: the ATVS-UPM system at 2003 NFI/TNO forensic evaluation. In *Odyssey 2004, The Speaker and Language Recognition Workshop*, Toledo, Spain.

Gray, R. M. (1984). Vector quantization. *IEEE ASSP Magazine*, 1:4–29.

Gruber, J. S., Poza, F., and Pellicano, A. J. (1993). *Audio Recordings: Evidence, Experts and Technology*, volume 48 of *Am Jur Trials*. Lawyers Cooperative Publishing.

Hennebert, J., Melin, H., Petrovska, D., and Genoud, D. (2000). POLYCOST: A telephone-speech database for speaker recognition. *Speech Communication*, 31:265–270.

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

Hermansky, H. (1994). RASTA Processing of Speech. *IEEE Trans. on Speech, and Audio Proc*, 2(4):578–589.

Hollien, H. (1990). *The Acoustics of Crime, The New Science of Forensic Phonetics*. Number ISBN 0-306-43467-9. New York: Plenum Publishing Corporation.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Springer.

Kerstholt, J., Jansen, E., Amelsvoort, A. v., and Broeders, A. (2003). Earwitness line-ups: effects of speech duration, retention interval and acoustic environment on identification accuracy. In *Proc. Eurospeech 2003*, pages 709–712, Geneva, Switzerland.

Künzel, H. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics*, 4/1:48–83.

Künzel, H. J. (1998). Forensic speaker identification: A view from the crime lab. In *Proceedings of the COST Workshop on Speaker Recognition by Man and Machine*, pages 4–8, Technical University of Ankara, Ankara, Turkey.

Koenig, B. (1990). Authentication of forensic audio recordings. *Journal of the Audio Engineering Society*, 38(1/2):3–33.

Koolwaaij, J. and Boves, L. (1999). On decision making in forensic casework. *Forensic Linguistics, the International Journal of Speech, Language and the Law*, 6(2):242–264.

Kreyszig, E. (1999). *Advanced engineering mathematics*. John Wiley and Sons, Singapore, 8th edition.

Li, K. P. and Porter, J. (1988). Normalizations and selection of speech segments for speaker recognition scoring. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-88.*, volume 1, pages 595–598, New York, USA.

Loevinger, L. (1995). Science as evidence. *Jurimetrics*, 35(2):153–190.

Majewski, W. and Basztura, C. (1996). Integrated approach to speaker recognition in forensic applications. *Forensic Linguistics*, 3(1):50–64.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proc. IEEE*, 63:561–580.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET Curve in Assessment of Detection Task Performance. In *Proc. Eurospeech '97*, pages 1895–1898, Rhodes, Greece.

Martin, R. (1994). Spectral subtraction based on minimum statistics. In *EUSIPCO-94*, pages 1182–1185.

Meuwly, D. (2000). *Voice Analysis, in : Encyclopedia of Forensic Science*, pages 1413 – 1420. London: Academic Press Ltd.

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

Meuwly, D., Alexander, A., Drygajlo, A., and Botti, F. (2003). Polyphone-IPSC: A shared speakers database for evaluation of forensic-automatic speaker recognition systems. In *Forensic Science International*, volume 136, page 367, Istanbul, Turkey. Elsevier.

Meuwly, D. and Drygajlo, A. (2001). Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modeling (GMM). In *2001, A Speaker Odyssey: The Speaker Recognition Workshop*, pages 145–150.

Meuwly, D., El-Maliki, M., and Drygajlo, A. (1998). Forensic speaker recognition using Gaussian mixture models and a bayesian framework. In *8th COST 250 workshop: Speaker Identification by Man and by Machine: Directions for Forensic Applications*, pages 52–55.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits in our Capacity for Processing Information. *The Psychological Review*, 63:81–97.

Nakasone, H. and Beck, S. D. (2001). Forensic Automatic Speaker Recognition. In *2001: A Speaker Odyssey*, pages 139–144, Crete, Greece.

Navratil, J. and Ramaswamy, G. N. (2003). The awe and mystery of T-norm. In *Proc. Eurospeech 2003*, pages 2009–2012, Geneva, Switzerland.

Niemi-Laitinen, T., Saastamoinen, J., Kinnunen, T., and Fränti, P. (2005). Applying MFCC-based automatic speaker recognition to GSM and forensic data. In *2nd Baltic Conf. on Human Language Technologies (HLT'05)*, pages 317–322, Tallinn, Estonia.

Oglesby, J. Mason, J. (1989). Speaker recognition with a neural classifier. In *Proceedings First IEE International Conference on artificial Neural Networks*, volume 313, pages 306–309.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Fransisco: Morgan Kaufmann Publishers.

Pfister, B. and Beutler, R. (2003). Estimating the weight of evidence in forensic speaker verification. In *Proc. Eurospeech 2003*, pages 701–704, Geneva, Switzerland.

Reynolds, D. A. (1992). *A Gaussian mixture modeling approach to text-independent speaker identification*. PhD thesis, Georgia Institute of Technology, Atlanta, Georgia.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Commun.*, 17(1-2):91–108.

Reynolds, D. A. (2003). Channel robust speaker verification via feature mapping. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 2, pages 53–56.

Reynolds, D. A., Doddington, G. R., Przybocki, M. A., and Martin, A. F. (2000). The NIST speaker recognition evaluation - overview methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254.

Reynolds, D. A. and Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83.

Richiardi, J., Prodanov, P., and Drygajlo, A. (2005). A probabilistic measure of modality reliability in speaker verification. In *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing 2005*, pages 709–712, Philadelphia, USA.

Robertson, B. and Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley and Sons, Chichester.

Rose, P. (2002). *Forensic Speaker Identification*. Forensic Science. London & New York: Taylor and Francis.

Rose, P. (2003). *The technical comparison of forensic voice samples*. Freckelton I. and Selby H, series editors, Issue 99, Expert evidence. Sydney: Thomson Lawbook Company.

Rose, P. (2005). Technical forensic speaker recognition: evaluation, types and testing of evidence. *Computer Speech and Language (to appear)*.

Rosenberg, A. (1976). Automatic speaker verification: A review. *Proceedings of the IEEE*, 64(4):475–487.

Rossy, Q. (2003). Simulation de cas reels de reconnaissance de locuteurs au moyen du logiciel ASPIC. Project report, Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, Switzerland.

Schmidt-Nielsen, A. and Crystal, T. H. (2000). Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. *Digital Signal Processing*, 10:249–266.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Sohn, J., Kim, N., and Sung, W. (1999). A statistical model-based voice activity detector. *IEEE Signal Process. Lett.*, 6(1):1–3.

Soong, F., Rosenberg, A., Rabiner, L., and Juang, B. (1985). A vector quantization approach to speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 10, pages 387–390.

Srinivasan, K. and Gersho, A. (1993). Voice activity detection for cellular networks. In *Proc. IEEE Speech Coding Workshop*, number 85-86.

Tippett, C. F., Emerson, V., Fereday, M., Lawton, F., and Lampert, S. (1968). The evidential value of the comparison of paint flakes from sources other than vehicles. *Journal of Forensic Sci. Soc.*, 8:61–65.

Voiers, W. D. (1964). Perceptual bases of speaker identity. *Journal of the Acoustic Society of America*, 36(6):1065–1073.

Wolf, J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51:2044–2056.

Yarmey, A. D. (1995). Earwitness Speaker Identification. *Psychology, Public Policy, and Law*, 1(4):792–816.

Zimmermann, P. (2005). Analyse de l'influence des conditions d'enregistrement dans la reconnaissance automatique de locuteurs en sciences forensiques. Project report, Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, Switzerland.

Curriculum Vitae

URL: http://www.anilalexander.org
e-mail: alexander.anil@epfl.ch
Date of birth : June 16, 1977
Nationality : Indian

Anil ALEXANDER
Chemin de Veilloud 14
1024 Ecublens
Switzerland

Education

- *Ph.D. student in forensic speaker recognition* (since 2001), in the framework of the Swiss National Science Foundation project *'The Challenge of Forensic Speech Processing: Automatic Speaker Recognition for Criminal Investigations'*.

  Signal Processing Institute, School of Engineering, Swiss Federal Institute of Technology, Lausanne, Switzerland.

- *Postgraduate Course in Computer Science* Language and Speech Engineering (2002)
  Swiss Federal Institute of Technology, Lausanne, Switzerland

- *Doctoral School in Computer Science*, Department of Computer Science (2000-2001)
  Swiss Federal Institute of Technology, Lausanne, Switzerland

- *Bachelor of Technology (B.Tech.)* - Computer Science and Engineering (1996-2000)
  Indian Institute of Technology - Madras, Chennai, India

Professional Experience

- Research Assistant (since 2002), Speech Processing and Biometrics Group, Laboratoire d'Intelligence Artificielle Perceptive (LIAP), Signal Processing Institute
  School of Engineering, Swiss Federal Institute of Technology, Lausanne

- Research Assistant - Forensic Speaker Recognition (2001-2002), Institut de Police Scientifique et de Criminologie, Faculty of Law, University of Lausanne

- Forensic Audio Analysis Casework (since 2002)- Forensic expert in several cases dealing with forensic automatic speaker recognition, audiotape authentication, enhancement and legal transcription of poor quality audio recordings, and intelligibility enhancement

- Associate Software Engineer (2000) VERITAS, Pune, India

Awards and Distinctions

- 2nd Place, Best Poster, "On Estimation of the Strength of Evidence in Forensic Speaker Recognition" A. Alexander and A. Drygajlo, Fifth International Conference on Forensic Statistics, Venice, Italy, 2002

- Ranked in the top 0.1% in the Sciences at the All India Secondary School Examination National Merit Scholarship from the Department of Education, Government of India

Professional Affiliations

- International Speech Communication Association (ISCA)

- Institute of Electrical and Electronics Engineers (IEEE)

Computer Skills

- **Languages**: C, C++, Perl, MATLAB, Visual Basic, Pascal and Latex

- **Platforms**: Linux, Unix (Admin), DOS, Windows-XP/NT/95 (Admin), Solaris

- **Tools**: Matlab (Signal Processing, Statistics, System Identification Toolboxes), Adobe Audition / CoolEdit Pro, Wavesurfer (including scripting with the Snack Toolbox), Praat, Apache (Webserver)

Teaching Experience

- Swiss Federal Institute of Technology, Lausanne
  Teaching: Speech Processing Laboratory (Spring 2002, Spring 2003, Spring 2004)
  Supervision: Supervision of four students' Masters theses and several semester projects in Forensic Speaker Recognition

- Indian Institute of Technology, Madras (1999 to 2000)
  Undergraduate Course in Computer Science - Introduction to Computing

Publications

1. D.Meuwly, A. Alexander, A. Drygajlo, and F. Botti, "Polyphone-IPSC: A Shared Speakers Database for Evaluation of Forensic-automatic Speaker Recognition Systems," in *Forensic Science International*, vol. 136. Istanbul, Turkey: Elsevier, September 2003, p. 367.

2. A. Drygajlo, D. Meuwly, and A. Alexander, "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 689–692.

3. A. Alexander and A. Drygajlo, "Scoring and Direct Methods for the Interpretation of Evidence in Forensic Speaker Recognition," in *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP)*, Jeju, Korea, 2004, pp 2397-2400.

4. F. Botti., A. Alexander, and A. Drygajlo, "An Interpretation Framework for the Evaluation of Evidence in Forensic Automatic Speaker Recognition with Limited Suspect Data," in *Proceedings of 2004: A Speaker Odyssey*, Toledo, Spain, 2004, pp. 63–68.

5. A. Alexander, F. Botti, and A. Drygajlo, "Handling Mismatch in Corpus-Based Forensic Speaker Recognition," in *Proceedings of 2004: A Speaker Odyssey*, Toledo, Spain, May 2004, pp. 69–74.

6. A. Alexander, F. Botti, D. Dessimoz and A. Drygajlo, (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International*, 146(Supplement 1,2):S95–S99.

7. F. Botti, A. Alexander, and A. Drygajlo, (2004). . On compensation of mismatched recording conditions in the bayesian approach for forensic automatic speaker recognition. *Forensic Science International*, 146(Supplement 1,2):S101–S106.

8. A. Alexander, D. Dessimoz, F. Botti, and A. Drygajlo, Aural and Automatic Forensic Speaker recognition in Mismatched Conditions. to be published in *Forensic Linguistics, The International Journal of Speech, Language and the Law.*

9. M. Arcienega, A. Alexander, P. Zimmerman and A. Drygajlo, A Bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition. *Interspeech'2005*, September 2005, Lisbon, Portugal, pp. 2009–2012