Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition

*Andrzej Drygajlo, Didier Meuwly*and Anil Alexander*

Speech Processing Group, EPFL, Lausanne, Switzerland
*The Forensic Science Service, Birmingham, U.K.
andrzej.drygajlo@epfl.ch

Abstract

The goal of this paper is to establish a robust methodology for forensic automatic speaker recognition (FASR) based on sound statistical and probabilistic methods, and validated using databases recorded in real-life conditions. The interpretation of recorded speech as evidence in the forensic context presents particular challenges. The means proposed for dealing with them is through Bayesian inference and corpus based methodology. A probabilistic model – the odds form of Bayes' theorem and likelihood ratio – seems to be an adequate tool for assisting forensic experts in the speaker recognition domain to interpret this evidence. In forensic speaker recognition, statistical modelling techniques are based on the distribution of various features pertaining to the suspect's speech and its comparison to the distribution of the same features in a reference population with respect to the questioned recording. In this paper, the state-of-the-art automatic, text-independent speaker recognition system, using Gaussian mixture model (GMM), is adapted to the Bayesian interpretation (BI) framework to estimate the within-source variability of the suspected speaker and the between-sources variability, given the questioned recording. This double-statistical approach (BI-GMM) gives an adequate solution for the interpretation of the recorded speech as evidence in the judicial process.

1. Introduction

The forensic application of speaker recognition technology is one of the most controversial issues within the wide community of researchers, experts, operators and police institutes. During the last 30 years many teams of engineers, pattern recognition experts and computer programmers have failed to create a reliable forensic technique for forensic speaker recognition, although several systems for commercial applications, mostly speaker verification, were developed at that time. The main reason for this failure is that methodological aspects concerning automatic identification of speakers in criminalistics and the role of forensic expert has not been investigated untill recently [1].

The forensic expert's role is to testify to the worth of the evidence by using, if possible, a quantitative measure of this worth. It is up to other people (the judge and/or the jury) to use this information as an aid to their deliberations and decision [2].

The first goal of this paper is to investigate the ways of interpreting evidence within the field of forensic speaker recognition.

The second goal is the assessment of the methodology developed and related automatic speaker recognition techniques, as well as of the databases used. Forensic experts should give the court an evaluation, which illustrates the performance of the system, its discrimination value and its robustness to mismatched recording conditions. The objective is to propose some directions for standardization of assessment procedures in the field of forensic speaker recognition.

Different methods can be applied to determine if the unknown voice of the questioned recording (trace) belongs to the suspected speaker (source). The most persistent real-world challenge in this field is the variability of speech. There is within-speaker (within-source) variability as well as between-speakers (between-sources) variability. Consequently, forensic speaker recognition methods should provide a statistical-probabilistic evaluation, which attempts to give the court an indication of the strength of the evidence, given the estimated within-source variability and the between-sources variability.

The objective of this paper is the definition and the implementation of a largely automatic system for forensic speaker recognition based on statistical and probabilistic methods and validated using databases recorded in real-life conditions. The intended final objective outcome of this paper is not to promote one method against another, but is to make available to the legal and investigative bodies a methodology based on a scientific approach with rigorous experimental background, which is independent of the automatic speaker recognition method chosen.

In this paper, Bayes' theorem and a corpus-based methodology to interpret evidence are adopted for speaker recognition. An automatic speaker recognition method based on Gaussian Mixture Modelling (GMM) is used in a Bayesian interpretation (BI) framework as an example [3]. The methodology proposed in this paper needs three different databases for the calculation and the interpretation of the evidence.

2. Voice as evidence

A forensic expert has to interpret evidence material in the course of a criminal investigation. In the case of questioned recording (trace), the evidence does not consist in speech itself, but in the quantified degree of similarity between speaker dependent features extracted from the trace, and speaker dependent features extracted from recorded speech of a suspect, represented by his/her model [4]. In the case of statistical modelling, the speech of the suspected speaker can be represented by Gaussian mixture model (GMM). The principal structure for calculating the value of evidence (*E*) and its place in the whole recognition/interpretation process is presented in Fig. 1.

The calculated value of evidence does not allow the forensic expert alone to make an inference on the identity of the speaker. The suspect's voice can be recognized as the recorded voice of the trace, to the extent that the evidence supports the hypothesis that the questioned, and the suspect's recorded voices are from the same person versus the other hypothesis that they are not. As no ultimate set of speaker specific features is present or detected in speech, the recognition process remains in essence a statistical process based on models of speakers and collected data, which depend on a large number of design decisions.

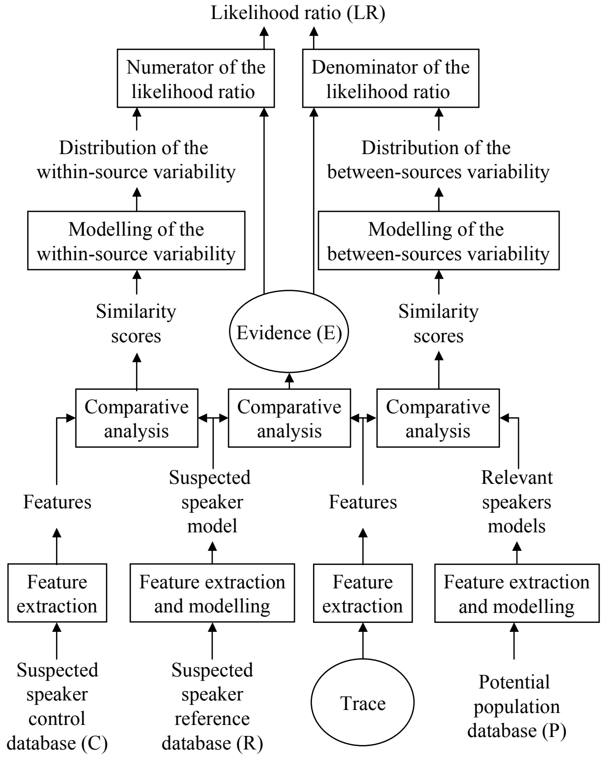
Likelihood ratio (LR)
Numerator of the likelihood ratio
Denominator of the likelihood ratio
Distribution of the within-source variability
Distribution of the between-sources variability
Modelling of the within-source variability
Modelling of the between-sources variability
Similarity scores
Evidence (E)
Similarity scores
Comparative analysis
Comparative analysis
Comparative analysis
Features
Suspected speaker model
Features
Relevant speakers models
Feature extraction
Feature extraction and modelling
Feature extraction
Feature extraction and modelling
Suspected speaker control database (C)
Suspected speaker reference database (R)
Trace
Potential population database (P)

Figure 1: Principal structure for the calculation and interpretation of the evidence

3. Bayesian interpretation of evidence

We propose to discuss forensic speaker recognition in the light of current state-of-the-art interpretation of forensic evidence based on the concept of identity used in criminalistics, a clear understanding of the inferential process and the respective duties of the actors involved in the judicial process, judges and forensic experts.

In forensic science, identity of source cannot be known with certainty, and therefore must be inferred [2]. The inference of identity can be seen as a reduction process, from an initial population to a restricted class, or, ultimately, to unity. Recently, an investigation concerning the inference of identity in forensic speaker recognition has shown the inadequacy of the speaker verification and speaker identification (in closed set and in open set) techniques already proposed to assess the evidence in this field [1]. These techniques are clearly inadequate for forensic purposes, because they force the forensic expert to make yes or no decisions, which are devolved upon the court.

In this paper, Bayes' theorem and a corpus-based methodology to interpret evidence are adopted for speaker recognition. The discussion benefits from the research work related to other forensic fields (e.g. fingerprint, fibers, DNA or glass trace evidence) [2].

The odds form of Bayes' theorem shows how new data (questioned recording) can be combined with prior background knowledge (prior odds (province of the court)) to give posterior odds (province of the court) for judicial outcomes or issues (Eq. 1). It allows for revision based on new information of a measure of uncertainty (likelihood ratio of the evidence (province of the forensic expert)) 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.

$$\overset{\text{posterior knowledge}}{\underset{\substack{\textit{posterior odds}\\ \text{(province of the court)}}}{\frac{p(H_0\,|\,E)}{p(H_1\,|\,E)}}} = \overset{\text{new data}}{\underset{\substack{\textit{likelihood ratio}\\ \text{(province of the expert)}}}{\frac{p(E\,|\,H_0)}{p(E\,|\,H_1)}}} \cdot \overset{\text{prior knowledge}}{\underset{\substack{\textit{prior odds}\\ \text{(province of the court)}}}{\frac{p(H_0)}{p(H_1)}}} \qquad (1)$$

This hypothetical-deductive reasoning method, based on the odds form of the Bayes theorem, allows to evaluate the likelihood ratio of the evidence, that leads to the statement of the degree of support for one hypothesis against the other. The ultimate question relies on the evaluation of the probative strength of this evidence provided by an automatic speaker recognition method.

4. Strength of evidence

The strength of the evidence is the result of the interpretation of the evidence, expressed in terms of the likelihood ratio of two alternative hypotheses. The principal structure for the calculation and the interpretation of the evidence is presented in Fig. 1. It includes the collection (or selection) of the databases, the automatic speaker recognition and the Bayesian interpretation.

4.1. Double statistical approach (BI-GMM)

The Bayesian interpretation (BI) methodology proposed in this paper needs a two-stage statistical approach. The first stage consists in modelling multivariate feature data using GMMs. The second stage transforms the data to a univariate projection based on modelling the similarity scores. The GMM method is not only used to calculate the evidence by comparing the questioned recording (trace) to the GMM of the suspected speaker (source), but it is also used to produce data necessary to model the within-source variability of the suspected speaker and the between-sources variability of the potential population of relevant speakers, given the questioned recording. The interpretation of the evidence consists of calculating the likelihood ratio using the probability density functions (pdfs) of the variabilities and the numerical value of evidence.

4.2. Databases

The information provided by the analysis of the questioned recording (trace) leads to specify the initial reference population of relevant speakers (potential population) having voices similar to the trace, and, combined with the police investigation, to focus on and select a suspected speaker. The methodology proposed in this paper needs three databases for the calculation and the interpretation of the evidence: 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 modelling the variability of the speech of all the potential relevant sources, using the automatic speaker recognition method. It allows evaluating the between-sources variability given the questioned recording, which means the distribution of the similarity scores that can be obtained, when the questioned recording is compared to the speaker models (GMMs) of the potential population database. The calculated between-sources variability pdf is then used to estimate the denominator of the likelihood ratio $p(E|H_1)$. Ideally, the technical characteristics of the recordings (e.g. signal acquisition and transmission) should be chosen according to the characteristics analyzed in the trace. Practically, the recording of such a large-scale speech database is a long and expensive procedure. A solution is to select an existing database relevant for this purpose, but recorded in mismatched conditions, that could be a valid proposition for common languages. If an existing database is available, the recording procedure can be limited to a selection of the relevant utterances, used to create speaker models of the P database. Since the recording conditions of the trace and the P database are 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, however, with the suspected speaker to evaluate her/his within-source variability, when the utterances of this database are compared to the suspected speaker model (GMM). This calculated within-source variability pdf is then used to estimate the numerator of the likelihood ratio $p(E|H_0)$. The recording of the C database should be constituted of utterances as far as possible equivalent to the trace, according to the technical characteristics, as well as to the quantity and style of speech.

The method proposed has been exhaustively tested in mock forensic cases corresponding to real caseworks using the IPSC Polyphone database, which consists of telephone quality in Swiss French [5]. In an example presented in Fig. 2, the strength of evidence, expressed in terms of likelihood ratio gives LR = 9.165 for the evidence value E = 9.94, in this case. The important point to be made here is that the estimate of the LR is only as good as the modelling techniques and databases used to derive it. In the example, the kernel density estimation technique was used to estimate pdfs from the data representing similarity scores [2].

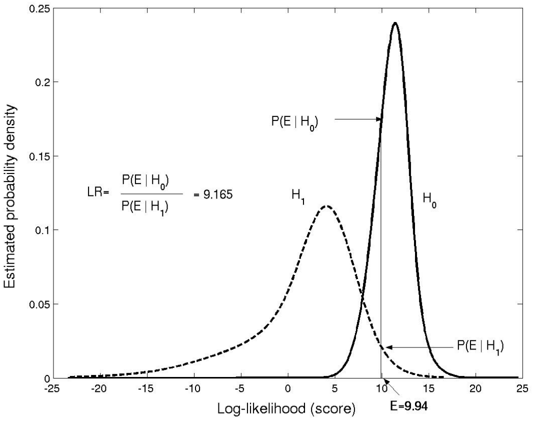

Figure 2: Graphical representation of the likelihood ratio (*LR*) estimation given the value of evidence *E*.

5. Evaluation of the strength of evidence

The likelihood ratio (*LR*) summarizes the statement of the forensic expert in the casework. However, the greatest interest to the jurists is the extent to which the *LRs* correctly discriminate "same speaker and different-speaker" pairs under operating conditions similar to those as regards to the case in hand. As was made clear in the US Supreme Court decision in Daubert case (Daubert v. Merrell Dow Pharmaceuticals, 1993) it should be criterial for the admissibility of scientific evidence to know to what extent the method can be, and has been, tested.

The principle for evaluation of the strength of evidence proposed in this paper consists in the estimation and the comparison of the likelihood ratios that can be obtained from the evidence E, on the one hand when the hypothesis H_0 is true (the suspected speaker truly is the source of the questioned recording) and, on the other hand, when the hypothesis H_1 is true (the suspected speaker is truly not the source of the questioned recording). The performance and reliability of an automatic speaker recognition method is evaluated by repeating the experiment described in the previous sections, with several speakers being at the origin of the questioned recording, and by representing the results using experimental (histogram based) probability distribution plots such as probability density functions $P(LR(H_i) = LR)$ (Fig. 3), cumulative distribution functions $P(LR(H_i) < LR)$ (Fig. 4), Tippett plots $P(LR(H_i) > LR)$ (Fig. 5), or their combinations.

The way of representation of the results in Fig. 5 is the one proposed by Evett and Buckleton in the field of interpretation of the forensic DNA analysis [6]. The authors have named this representation "Tippett plot", referring to the concepts of "within-source comparison" and "between-sources comparison" defined by Tippett *et al.*. These probability distributions are also known as reliability functions.

In order to test the BI-GMM system under operating conditions of the given mock case, 15 male speakers were chosen as suspects from the IPSC Polyphone database. For each suspect, 4 traces of duration 12-15 seconds were selected.

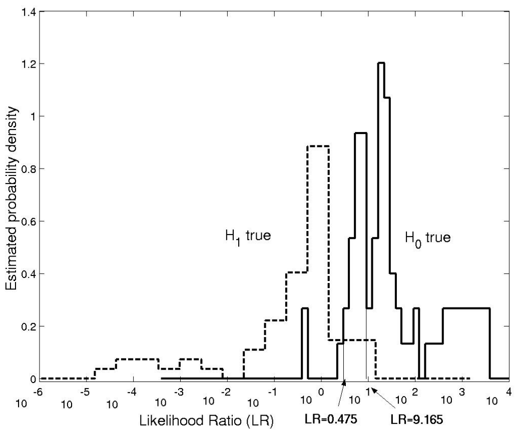

Figure 3: Estimated probability density functions of *LRs*.

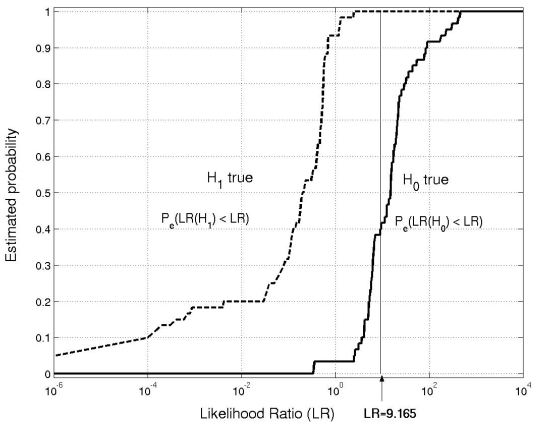

Figure 4: Cumulative distribution functions.

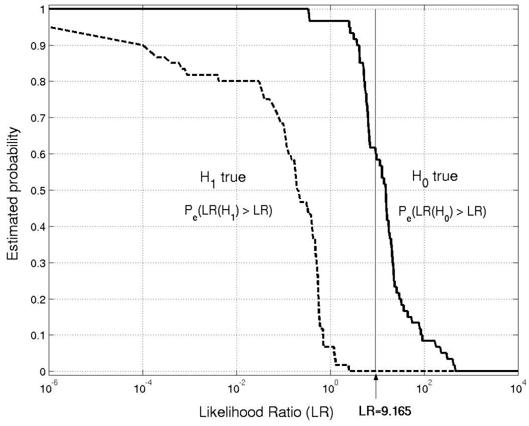

Figure 5: Tippett plots (reliability functions).

In order to create the R databases, 7 recordings of 2-3 minutes duration were chosen for each suspect. The C databases were created using 18 recordings of 10-15 seconds duration for each of them. This allowed us to have 60 cases when the hypothesis H_0 is true and 60 other cases when the hypothesis H_1 is true. In each case, the P database used was a subset of 100 speakers of the Swiss-French Polyphone database. In the test, GMMs with 32 Gaussian pdfs each and RASTA-PLP features were used. The results of this test do give some confidence in the method. For example, from Fig. 5 one can learn that *LR* exceeds 1 in 96% of cases when H_0 is true, and in only 6% of cases when H_1 is true. The likelihood ratio of 9.165, calculated in the case, is included in this working space. This value is exceeded in 60% of cases when H_0 is true, and in 0% of cases when H_1 is true. These observations indicate to the court how the *LR* quoted can be considered as reliable regarding the automatic speaker recognition methods and databases used.

6. Conclusions

The main objective of this work is to establish a robust methodology for forensic automatic speaker recognition based on statistical methods. This paper gives step-by-step guidelines for the calculation of the evidence, its strength and the evaluation of this strength under operating conditions of the casework. In the paper, an automatic method using the Gaussian mixture models (GMMs) and the Bayesian interpretation (BI) framework, which represents neither speaker verification nor speaker identification, was proposed for the forensic speaker recognition task. This method, using a likelihood ratio to indicate the strength of the evidence of the questioned recording, measures how this recording scores for the suspected speaker model, compared to relevant non-suspect speaker models. This method was developed in order to find an adequate solution for the interpretation of voice recording as scientific evidence in the judicial process. Given the approximate nature of the *LR* we also proposed a way to evaluate how reliable the estimate is.

7. References

[1] Champod, C., Meuwly, D., “The Inference of Identity in Forensic Speaker Identification”, *Speech Communication*, vol. 31, 2000, pp. 193-203.

[2] Aitken C., *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 1995.

[3] Meuwly D., Drygajlo, A., “Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM)”, *Proc. 2001: A Speaker Odyssey, The Speaker Recognition Workshop,* Crete, Greece, June 2001, pp. 145-150.

[4] Rose, P., *Forensic Speaker Identification*, Taylor & Francis, London, 2002.

[5] Meuwly, D., *Reconnaissance de locuteurs en sciences forensiques: l’apport d’une approche automatique,* Ph.D. thesis, IPSC, University of Lausanne, 2001.

[6] Evett, I., Buckleton, J., “Statistical Analysis of STR Data”, *Advances in Forensic Haemogenetics,* vol. 6, Springer-Verlag, Heidelberg, 1996, pp. 79-86.