Scoring and Direct Methods for the Interpretation of Evidence in Forensic Speaker Recognition

Anil Alexander, Andrzej Drygajlo

Swiss Federal Institute of Technology Lausanne (EPFL)
Lausanne, Switzerland
alexander.anil@epfl.ch, andrzej.drygajlo@epfl.ch

Abstract

In forensic speaker recognition, the strength of evidence is estimated using the likelihood ratio, which is the relative probability of observing the evidence, given the hypothesis that the suspect is the source of the questioned recording and the hypothesis that anyone else in a relevant potential population is its source. In order to calculate the likelihood ratio we use two approaches; one, directly using the likelihoods returned by the Gaussian Mixture Models (GMMs), and the other by modeling the distributions of these likelihood scores and then deriving the likelihood ratio on the basis of these score distributions. The former approach is used implicitly in speaker verification systems, although in forensic speaker recognition, the latter is preferred as it does not depend on the automatic speaker recognition technique used. However, both these methods have their advantages and disadvantages. In this paper, we propose statistical representations in order to evaluate the strength of evidence in each of these two methods.

1. Introduction

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.

At the heart of the forensic automatic speaker recognition system is the creation of a statistical model for the features of each speaker, testing the features of other utterances against this model and obtaining the likelihood that this utterance could have come from this speaker. Parameterization techniques such as Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) [1] are commonly used in order to extract features from speech. Statistical models such as Gaussian Mixture Models (GMMs) are used to estimate the probability density function of the distribution of features of the speaker. GMM is a useful method to estimate the probability distribution for multivariate as well as univariate data.

The calculation of probability densities of the speech features is dealt with at two levels; one, at the multidimensional feature space where the likelihoods of the multivariate feature vectors are estimated, as well as in the univariate level where the likelihood scores (derived in the multivariate level) for different hypotheses are modeled. A speaker recognition system returns the likelihood of whether a given utterance came from the model created for the speaker. Because of this, we have two different approaches for calculating likelihood ratios. The first method takes the likelihood values returned by the system, and directly uses them to evaluate the likelihood ratio. The second method determines the likelihood ratio using the probability distributions of these likelihood scores. In this paper we present these two approaches and discuss the evaluation of the strength of evidence with respect to each of them. We discuss statistical representations that can be used in order to evaluate the strength of evidence and the reliability of the results.

2. Methods for estimating likelihood ratio

The odds form of Bayes theorem (Eq. 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 the province of the forensic expert.

$$\frac{p(H_0|E)}{p(H_1|E)} = \frac{p(E|H_0)}{p(E|H_1)} \cdot \frac{p(H_0)}{p(H_1)} \tag{1}$$

In this section we present two methods that the forensic expert can use in speaker recognition in order to estimate the likelihood ratio: scoring method and direct method.

2.1. 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).

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 the trace.

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken
Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

- 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 (intra-variability) of his voice.

A brief summary of the methodology proposed in [2] to calculate a likelihood ratio for a given trace is as follows (illustrated in Fig. 1) :

- The features of trace are compared with the statistical models 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.
- 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 *within-source* and *between-sources* distributions at the point E. This is illustrated in Fig. 1 where, for an example forensic case, a likelihood ratio of 9.165 is obtained for $E = 9.94$ using the scoring method. The same example case will be considered in the direct method.

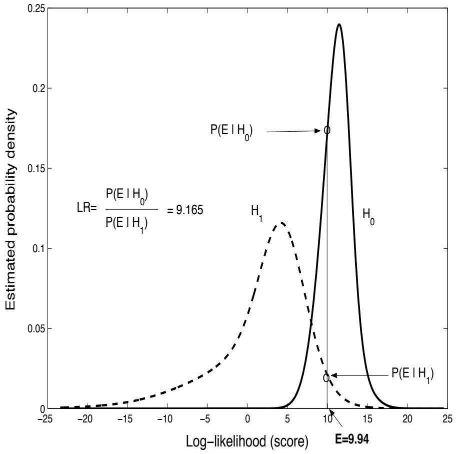

Figure 1: Illustration of the estimation of LR using the scoring method

Mathematically,

$$\begin{aligned} LR_{scoring} &= \frac{p(E|H_0)}{p(E|H_1)} \qquad (2)\\ &= \frac{p(L(X_T|\lambda_{R_i})|H_0)}{p(L(X_T|\lambda_{R_i})|H_1)}\\ for & \quad i = 1, ..., N_R \end{aligned}$$

where,

- N_R (Eq. 2) , N_C and N_P (Eq. 3 & 4) 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. 2, are modeled using probability density functions:

$$\begin{aligned} H_0 &= P(L(X_{C_i}|\lambda_{R_j})) \qquad (3)\\ for & \quad i = 1, ..., N_C\\ for & \quad j = 1, ..., N_R\\ H_1 &= P(L(X_T|\lambda_{P_i})) \qquad (4)\\ for & \quad i = 1, ..., N_P \end{aligned}$$

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

In this approach, the likelihood scores returned by speaker recognition system are used as indices in a Bayesian interpretation framework. This method of estimating likelihoods and likelihood ratios is similar to the kind used in the analysis of glass where the refractive index is used in order to calculate the likelihood ratio. In the case of glass analysis, the refractive index only signifies a particular value which is a property of the glass, and does not signify any likelihood of the piece of glass belonging to one type of glass or the other.

2.2. Direct Method

In the direct method, the likelihood ratio is the relative probability of observing the features of the trace in a probability distribution of the features of the suspect and 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).

A discussion of how to calculate 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), and the resulting score is 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.

$$LR_{direct} = \frac{\sqrt[N_R]{\prod_{i=1}^{N_R} p(X_T|\lambda_{R_i})}}{\sqrt[N_P]{\prod_{j=1}^{N_P} p(X_T|\lambda_{P_j})}} \qquad (5)$$

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).

In the forensic case example considered in Fig.1, a likelihood ratio of 1155 is estimated using the direct method.

2.3. 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 are affected by mismatched recording conditions of the databases involved. In the direct method, compensation of the mismatch must 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 [3]. 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. 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 sciences, a likelihood ratio of one is significant, 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 one, implies

$$p(X|\lambda_{R_i}) = p(X|\lambda_{P_j}) \quad \forall i, j \qquad (6)$$

and in the scoring method, this implies

$$p(L(X|\lambda_{R_i})|H_0) = p(L(X|\lambda_{R_i})|H_1) \qquad (7)$$
$$for \quad i = 1, ..., N_R$$

We can see in the direct method, that a likelihood ratio of one is obtained when the statistical models of the suspect and the potential population represent similar voices. An LR of one 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.

3. Estimating the significance of the strength of evidence

As discussed in the previous section, various automatic speaker recognition techniques, can give different likelihood ratios for the same case. Often, this depends on the algorithm on which the system is based. The likelihood ratio is dependent on the strengths or weaknesses of the system at hand. It is important thus, to know what the behavior of the system generally is, and to see what kind of likelihood ratios it returns.

The significance of the strength of evidence can be evaluated by estimating and comparing the likelihood ratios that are obtained for the evidence E when the hypothesis H_0 is true, i.e., the suspected speaker is indeed the source of the questioned recording and when the hypothesis H_1 is true, i.e., the suspected speaker is not the source of the questioned recording. By creating 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. Once we have 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.

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 $IPSC01$ Polyphone database. For each of these speakers 4 traces of duration 12-15 seconds were selected. The R database was created using 7 recordings 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 probability distribution plots such as the probability density functions $P(LR(H_i) = LR)$ (Fig. 2) and Tippett plots $P(LR(H_i) > LR))$ (Figs. 4 and 5). 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 is called the Tippett Plot. This representation has been used in the interpretation of the results of forensic DNA analysis [4]. The Tippett plot can be 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 considered in order to evaluate the strength of evidence.The following z test [5] can be used:

$$Z_{H_0} = \frac{LR_E - \mu_{LR_{H_0}}}{\sigma_{LR_{H_0}}} \qquad (8)$$

$$Z_{H_1} = \frac{LR_E - \mu_{LR_{H_1}}}{\sigma_{LR_{H_1}}} \qquad (9)$$

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.

For the case shown in Fig. 1, the likelihood ratios for E, using the direct method and the scoring method, are 1155 and

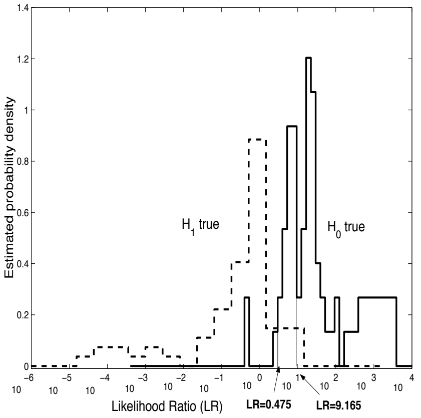

Figure 2: Probability density plot of LRs (scoring method)

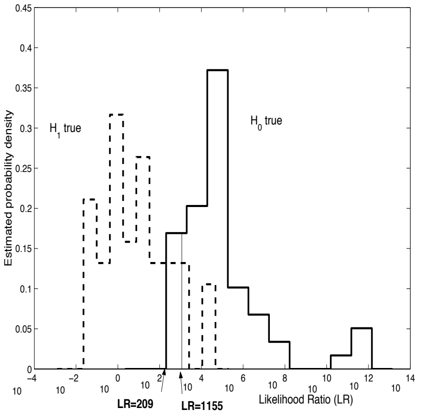

Figure 3: Probability density plot of LRs (direct method)

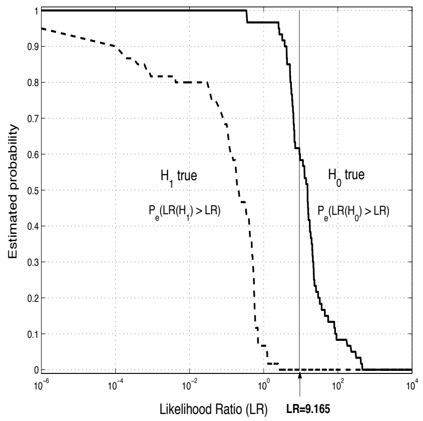

Figure 4: Tippett Plot (scoring method)

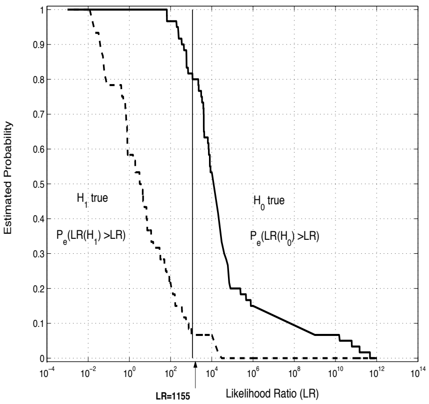

Figure 5: Tippett Plot (direct method)

9.165 respectively. In Figs. 2 and 3 the probability density functions of the scoring method and direct method LRs are presented. The points 0.475 and 209, in Figs. 2 and 3 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.

4. Conclusion

Two approaches in forensic automatic speaker recognition, one directly using the likelihoods returned by the Gaussian Mixture Models (GMMs), and the other by modeling the distribution of these likelihood scores and then deriving the likelihood ratio on the basis of these score distributions have been presented. Statistical representations to evaluate the strength of evidence and to compare the two methods have been presented.

5. References

[1] B. Gold and N. Morgan, *Speech and Audio Signal Processing: processing and perception of speech and music*. New York: John Wiley & Sons, 2000.

[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, F. Botti, and A. Drygajlo, “Handling Mismatch in Corpus-Based Forensic Speaker Recognition,” in *Proceedings of 2004: A Speaker Odyssey*, Toledo, Spain, 2004, to be published.

[4] I. Evett and J. Buckleton, “Statistical analysis of STR data,” *Advances in Forensic Haemogenetics*, vol. 6, pp. 79–86, 1996.

[5] C. Aitken, *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, 1997.