2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM)

Didier Meuwly — Andrzej Drygajlo***

*Institut de Police Scientifique et Criminologie, University of Lausanne, Switzerland
**Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne
Didier.Meuwly@ipsc.unil.ch — Andrzej.Drygajlo@epfl.ch

Abstract

The goal of this paper is to establish a scientifically founded methodology for forensic automatic speaker recognition. The interpretation of recorded speech as evidence in the forensic context presents particular challenges. The means proposed in the paper for dealing with them is through Bayesian inference. This leads to the formulation of a likelihood ratio measure of evidence which weighs the evidence in favor of two competing hypotheses: 1) the suspected speaker is the source of the questioned recording (trace), 2) the speaker at the origin of the questioned recording is not the suspected speaker. The state-of-the-art automatic recognition system using Gaussian mixture model (GMM) is adapted to the Bayesian interpretation (BI) framework with the models of the within-source variability of the suspected speaker and the between-source variability of 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. Examples provided are for telephone quality speech recordings that account for a very large proportion of all forensic material for speaker recognition.

1. Introduction

A common directive for research and development of objective, automatic techniques in forensic speech processing is needed in order to meet present needs and imminent challenges of the criminalistic real world, such as the widespread use of cellular telephones and other modern voice communication systems. During the last twenty years many teams of engineers, pattern recognition experts and computer programmers have failed to create a reliable forensic technique, and ultimately, a computer based device, 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 in criminalistics and the role of forensic expert has not been investigated till recently.

In this paper, a new automatic approach using the GMMs and a Bayesian framework, which represents neither speaker identification nor speaker verification, was introduced for the forensic investigation task. This method, using a likelihood ratio to indicate the value of the evidence, measures how the questioned recording scores for the suspect speaker model, compared to relevant non-suspect speaker models.

2. Inference of identity

2.1. Principle

In criminalistics, the identification process seeks individualisation [1]. Identifying a person or an object means that it is possible to distinguish this person or object from all others on the surface of the Earth. The forensic individualisation process can be seen as a reduction process beginning from an initial population to a single person. Recently, an investigation concerning the inference of identity in forensic speaker recognition has shown the inadequacy of the main solutions proposed to assess the evidence in this field. The concept of identity underlying the verification and the identification tasks (in closed-set and in open-set) does not correspond to the concept of identity accepted in forensic science.

In addition, the use of this concept forces the forensic scientist to deal°—°without being aware°—°with prior and/or posterior probability ratios on the issue of identification itself, whereas these are assessments pertaining only to the court [2].

2.1.1. *Speaker verification*

Speaker verification used for forensic speaker recognition is the process of accepting or rejecting the identity of a suspected speaker as the source of the questioned recording. It is a discrimination task. The decision of discrimination between the questioned recording and the suspected speaker recording depends on a threshold. Discrimination is interpreted as a rejection and non-discrimination as an acceptation.

The above concept of identity does not correspond to the definition of the forensic individualization; if the random match probability is not null (corollary of the threshold), the conclusion the suspect is identified is inadequate and misleading.

Moreover, it must be pointed out that the threshold is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert. But jurists will interpret this threshold as an expression of the criminal standard beyond reasonable doubt . Would jurists accept that the concept of reasonable doubt on the identification of a suspect escapes their province and that the threshold is imposed onto the court by the scientist°? The response in the doctrine is negative, as expressed by the members of the Panel on Statistical Assessments as Evidence in Courts: [ ] the law may establish different thresholds for what is sufficient evidence in a case from those that statisticians would normally require in drawing conclusions. Clearly, the law must prevail and the statistician must adjust to the law s standards. Put another way, it is the utility function of the court that is appropriate, not the utility function of the statistician [3, p. 141].

Therefore, speaker verification is clearly inadequate for forensic purposes, because it forces the scientist to adopt a role and to make decisions which are devolved upon the court [2].

2.1.2. Speaker identification

Speaker identification used for forensic speaker recognition is the process of determining from which of the suspected speakers the questioned recording comes. It is a classification task. In fact, the classification cannot take place in a closed set of speakers (closed-set identification) because the assessment of the credibility of the exhaustiveness of the number of suspects is outside the duties of the expert; it is a judicial matter pertaining to the court. In addition, it seems particularly unfair to disclose only the identity of the best candidate without providing the evidence obtained for the others, not necessary only from the closed set of speakers.

To overcome this default, the classification should then take place in an open set of speakers (open-set identification), but such a framework still implies a final discrimination decision based on a threshold and suffers from the same conceptual drawbacks as the verification task [2].

2.1.3. A new paradigm

In general, the court wants to know the odds that the suspected speaker has produced the questioned recording, given the circumstances of the case and the observations made by the forensic scientist. In other words, the court looks for the odds on an issue stating that the suspected speaker is the source of the trace (questioned recording) versus its alternative stating that the source of the questioned recording is not the suspected speaker.

The adequate interpretation of the value of the evidence provided by an automatic speaker recognition method needs to consider the statistical value obtained in a particular framework, namely the Bayesian framework, which conversely helps forensic scientists, jurists and members of the jury in reaching their conclusions. As pointed out by Lewis, who in 1984 proposed the use of Bayes theorem in speaker identification, evidence does not consist uniquely of scientific data [4]. The forensic individualization process is best explained at present by the hypothetical-deductive method [5; 6], but science can only provide additional information to assist an answer that must be ultimately arrived at inductively.

2.2. Bayesian interpretation (BI)

2.2.1. Principle

The preliminary research work proves that a probabilistic model — the Bayes theorem—is an adequate tool for assisting scientists to assess the value of scientific evidence. It helps jurists to interpret scientific evidence and to clarify the respective roles of scientists and of members of the court [6].

The Bayesian model allows the revision based on new information of a measure of uncertainty about the truth or falsity of an issue. This approach shows how new data (questioned recording) can be combined with prior background knowledge (prior odds) to give posterior odds for judicial outcomes or issues.

2.2.2. Calculation of the evidence

The evidence E is the result of the comparative analysis of the speaker dependent features (x) extracted from the questioned recording (X), with the speaker dependent features (y) extracted from utterances of the suspected speaker (Y).

2.2.3. Concept of likelihood ratio

The Bayesian model shows how an *a priori* likelihood ratio between two competitive hypotheses, H_1 and H_2, can evolve to an *a posteriori* likelihood ratio of these two hypotheses, after the analysis of the questioned recording. H_1 represents the hypothesis that the suspected speaker is the source of the questioned recording while H_2 represents the hypothesis that the source of the questioned recording is another speaker; by definition H_1 and H_2 are mutually exclusive.

The likelihood of E is evaluated when the hypothesis H_1 is verified and when the hypothesis H_2 is verified. The ratio between these two likelihood values, the likelihood ratio (LR) is defined as the numerical value that allows for revision based on the new information E of the *a priori* probability ratio (prior odds) to the *a posteriori* probability ratio (posterior odds) of the two hypotheses H_1 and H_2:

posterior odds = likelihood ratio × prior odds

$$\frac{p\left(H_1 \mid E\right)}{p\left(H_2 \mid E\right)} = \frac{p\left(E \mid H_1\right)}{p\left(E \mid H_2\right)} \times \frac{p\left(H_1\right)}{p\left(H_2\right)}$$

a posteriori probability ratio	likelihood ratio (LR)	*a priori* probability ratio

The likelihood ratio (LR) indicates the strength of the evidence in the two competitive hypotheses (H_1 and H_2) and summarizes the statement of the forensic scientist.

2.2.4. Corpus based methodology

In most of the cases the questioned recording is a voice recorded through the telephone network. This trace is provided either from an anonymous call or from a wiretapping. The aural analysis of the trace, generally conducted by a speech scientist or a translator, consists of the determination of the spoken language and a subjective qualification of the speech (accent, timbre, assumption on the gender, etc.). This aural information and the technical analysis of the questioned recording ensure to define the population of the relevant speakers and, combined to police investigation, to focus on a suspected speaker.

On the basis of this trace, 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 first database, named *potential population database* (P), is a large-scale database used to model the speech of the speakers of the relevant population with the automatic speaker recognition method. It is used to evaluate the between-source variability of the questioned recording, that means the distribution of the similarity scores that can be obtained when the questioned recording is compared to the speakers of the potential population. Finally, the model of the between-source variability is used to calculate the denominator of the likelihood ratio: $p\,(E \mid H_2)$.

The second database, named *suspected speaker reference database* (R), is recorded with the suspected speaker to model its speech with the automatic speaker recognition method. This suspected speaker model is used to calculate the evidence, when the model is compared to the questioned recording.

The third database, named *suspected speaker control database* (C), is recorded with the suspected speaker to evaluate its within-source variability when this database is compared to the suspected speaker model. Finally, the model of the within-source variability is used to calculate the numerator of the likelihood ratio: $p\,(E \mid H_1)$.

2.2.5. Interpretation of the evidence

The interpretation of the evidence is performed in two steps. The first one consists in modeling the distribution of the similarity scores resulting from the evaluation of the within-source variability of the suspected speaker and the between-source variability of the questioned recording.

The second step is the evaluation of the evidence regarding the two competitive hypotheses $p^{\circ}(E \mid H_1)$ and $p^{\circ}(E \mid H_2)$. The calculation of $p^{\circ}(E \mid H_1)$ leads to the measure of the probability of E in the model of the within-source variability of the suspected speaker. The calculation of $p^{\circ}(E \mid H_2)$ leads to the measure of the probability of E in the model of the between-source variability of the questioned recording.

The method proposed has been exhaustively tested in a mock forensic case corresponding to a normal casework using databases recorded at IPSC [7].

3. Automatic speaker recognition (GMM)

The speaker recognition system chosen for the following experiment is based on a text-independent automatic speaker recognition method using Gaussian Mixture Models (GMMs) [8].

The statistical Gaussian mixture model (GMM) has several attributes that make it well suited for speaker modelling in forensic applications. The parametric modelling capabilities of the GMM allow it to model any arbitrarily shaped probability density function (pdf) with a weighted sum of M component Gaussian densities. Each component density is a D-variate Gaussian pdf with mean vector and covariance matrix. In the case of a diagonal covariance matrix, which is the case in our approach, the GMM is given by:

$$p(x|\lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D} b_i\left(x_j, \mu_{ji}, \sigma_{ji}^2\right)$$

where μ_{ji} is the mean and σ_{ji}^2 is the variance of D-dimensional feature vector component x_j; p_i are the mixture weights of M component Gaussian densities. In this experiment the feature vector consists of 12 perceptual linear prediction coefficients (PLP).

The means, variances and mixture weights represent the parameters of the speaker model λ. Maximum likelihood speaker model parameters are estimated using the iterative Expectation-Maximization (EM) algorithm.

The Gaussian mixture density is shown to provide a smooth approximation to the long-term distribution of feature vectors obtained from speaker utterances. Furthermore, by modeling the underlying short-term features, the GMM characterizes the short-term variations of a person's voice and so is capable of high discrimination performance for short utterances. Therefore, the use of GMMs for automatic speaker recognition in forensic applications is justified since the evidence material often consists of short speech segments. The GMM is easily interpretable by forensic experts. It is also computationally efficient and can easily be implemented.

Given the feature-vector sequence (x_t, t = 1, , T) of speech utterance and the speaker model, the similarity score S is computed as log-likelihood:

$$S = \frac{1}{T} \sum_{t=1}^{T} \log p\left(x_t|\lambda\right).$$

4. Experiment: calculation of the evidence

The calculation of the evidence includes the collection of the questioned recording, the definition of the potential population and the selection of the potential population database (P), the selection of a suspected person, the recording of the suspected person reference database (R) and the use of the GMM automatic speaker recognition method.

4.1. Simulated questioned recording

The simulated questioned recording used in this experiment is an anonymous call of 10 seconds length, transmitted through the public switched telephone network (PSTN). The signal to noise ratio of the trace is about 40 dB. The aural analysis of the trace let suppose that the speaker is a Swiss-French male person speaking without perceptible voice disguise.

4.2. Selection of the potential population database (P)

The potential population that should be considered is the population of the French part of Switzerland that counts about 1 million of male speakers. This population is modeled with a subset of 1000 male speakers of the Swiss French Polyphone database of Swisscom'. Each speaker of the database has recorded one speech session of 100 to 140 seconds, constituted of read and spontaneous speech. These 1000 speech sessions are used to calculate 1000 Gaussian mixture speaker models.

4.3. Selection of a suspected person

We consider that the police investigation has permitted to focus on a suspected person, a Swiss-French male speaker; in this experiment, the suspected person truly is the source of the simulated questioned recording.

4.4. Recording of the suspected speaker reference database (R)

This database is recorded with the suspected person to model his speech. During two months, six speech sessions of 100 to 140 seconds have been recorded through the PSTN, in the same way that it has been done for the ˙°Swiss

French Polyphone°¨ database. The sessions are used to calculate six Gaussian mixture speaker models ($\lambda_1 : \lambda_6$).

4.5. Calculation of the evidence

The first speaker model (λ_1) is used to calculate the evidence with the automatic speaker recognition method GMM, in comparing it with the simulated questioned recording. The result obtained is a similarity score of 6. This similarity score is the value of the evidence E = 6.

5. Experiment: interpretation of the evidence

The interpretation of the evidence includes the recording of the suspected person control database (C), the evaluation of the within-source and between-source variabilities and the calculation of the likelihood ratio.

5.1. Recording of the suspected person control database (C)

This database is recorded with the suspected person to evaluate his within-source variability. One speech session, during which the speaker has described a set of 30 pictures spontaneously and has simulated 5 phone discussions, has been recorded through the PSTN. This speech session has been segmented manually into utterances of 1 to 30 seconds of length that represent 35 utterances.

5.2. Evaluation of the within-source variability

The six speaker models ($\lambda_1 : \lambda_6$) are compared to the C database with the automatic speaker recognition method GMM to evaluate the within-source variability of the speech of the suspected speaker. As result of the comparison, the method delivers 210 similarity scores that can be represented by a histogram in Fig. 1:

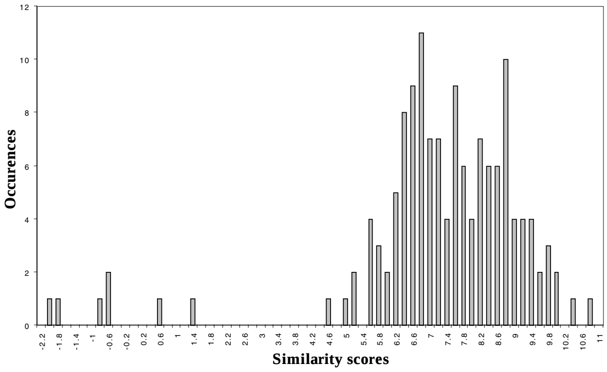

Figure 1*: Histogram given by GMM when calculating the scores of the within-source variability of the suspected speaker*

The distribution of the within-source variability scores is multimodal and cannot be estimated by a common law of distribution. For this reason, the data themselves are used as the source of the probability density function (pdf). In forensic science, Aitken [6] has proposed the application of the kernel density estimation (KDE), described by Silverman [9]. In this case, the estimation of the pdf is possible with KDE, because the distribution of the data is smooth enough to give the estimation represented in Fig. 2:

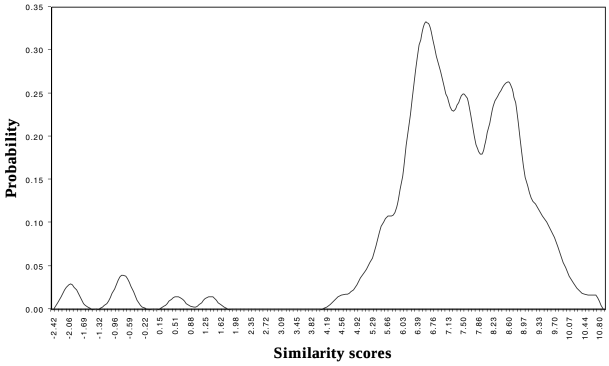

Figure 2*: Model of the within-source variability with kernel density estimation*

5.3. Evaluation of the between-source variability

The between-source variability of the questioned recording is evaluated by comparing the trace with the 1000 speaker models of the P database. As result of the comparison, the method delivers 8000 similarity scores that can be represented by a histogram in Fig. 3:

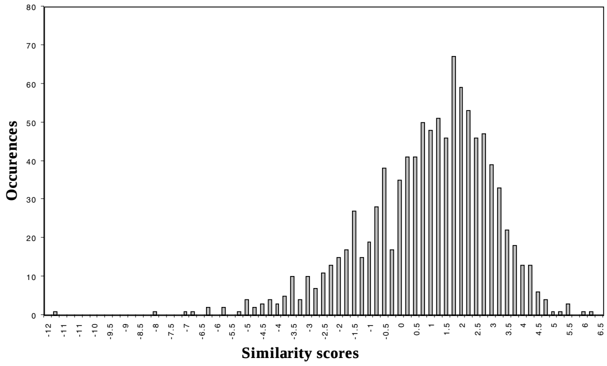

Figure 3*: Histogram given by GMM when calculating the scores of the between-source variability of the trace*

As the distribution of the between-source variability scores is multimodal and cannot be estimated by a common law of distribution, the pdf is estimated using KDE, as it is made for the within-source variability in section 5.2. It is represented in Fig 4:

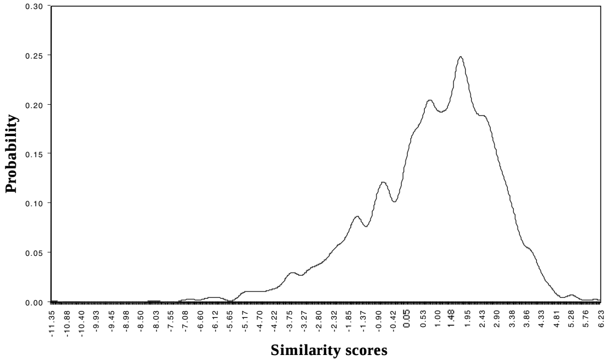

Figure 4*: Model of the between-source variability with kernel density estimation*

5.4. Evaluation of the likelihood ratio

The evaluation of the likelihood ratio results from the calculation of p (E | H_1) / p (E | H_2), for the evidence value of 6 (E = 6).

The value of p (E | H_1) is calculated for the evidence (E = 6) in the model of the within-source variability. This probability, represented in Fig. 5, is equal to 0.15:

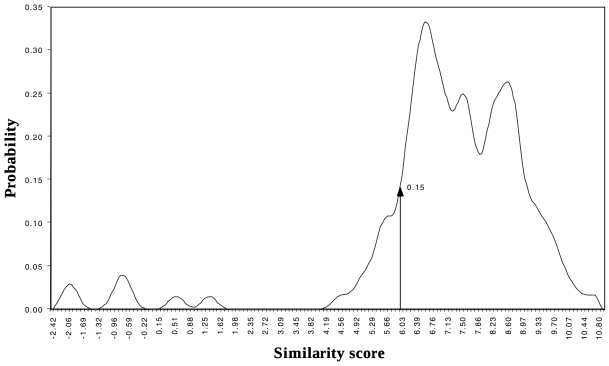

Figure 5: *Graphical representation of p (E | H_1)*

The value of p (E | H_2) is calculated for the evidence (E = 6) in the model of the between-source variability. This probability, represented in Fig. 6, is equal to 0.002:

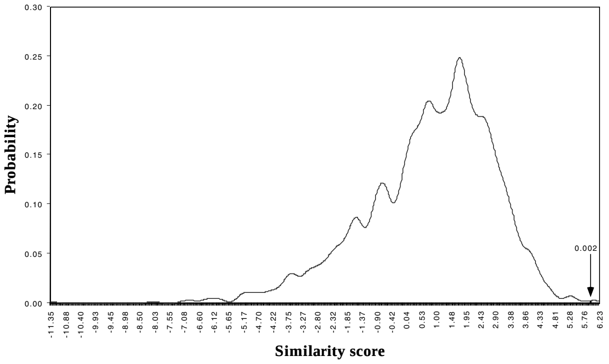

Figure 6: *Graphical representation of p (E | H_2)*

The interpretation of the evidence gives a likelihood ratio of 75 (0.15 / 0.002). This result, represented in Fig. 7, means that the forensic analysis of the questioned recording allows to revise the prior odds defined by the judge in multiplying them by 75:

$$\text{prior odds} \times 75 = \text{posterior odds}$$

6. Evaluation of the performance of forensic automatic speaker recognition method

6.1. Principle

The performance of an automatic speaker recognition method can be evaluated repeating the experiment described in sections 4 and 5 with several speakers.

The principle of the evaluation 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_1 is verified (the suspected speaker truly is the source of the questioned recording) and, on the other hand, when the hypothesis H_2 is verified (the suspected speaker is truly not the source of the questioned recording).

6.2. Tippett plots

The way of representation of the results is the one proposed by Evett and Buckleton in the field of interpretation of the forensic DNA analysis. The authors have named this representation Tippett plot , refering to the concepts of *within-source comparison* and *between-source comparison* defined by Tippett *et al.* [10, 11].

The horizontal axis is graduated with increasing values of likelihood ratios while the vertical axis indicates the estimated probability that the result of the experiment exceeds a given value of LR. The Tippett plot includes two curves: the first one shows the evolution of the estimated LR when the hypothesis H_1 is verified and the second one shows the evolution of the estimated LR when the hypothesis H_2 is verified.

6.3. Example

For evaluating the forensic automatic speaker recognition method presented in the previous sections, 48 speaker models and eight simulated traces were recorded by eight Swiss-French male speakers (six speaker models and one questioned recording per person). Six values of evidence were calculated for each speaker with the automatic speaker recognition method, comparing the questioned recording to the six models. On the whole, 48 values of evidence were calculated for the eight speakers. These 48 values are the values of evidence when the hypothesis H_1 is verified. For each of these 48 values a LR is estimated, following the method explained in this paper. The distribution of these LRs is illustrated by the grey curve in Fig. 8.

Then, the eight questioned recordings are compared to the speaker models of the potential population database. In this case the database is the subset of 1000 male speakers of the ˙ Swiss French Polyphone ¨ database of Swisscom' used in section 4.2. On the whole, 8000 values of evidence are calculated for the eight questioned recordings. These 8000 values are the values of evidence when the hypothesis H_2 is verified. For each of these 8000 values a LR is estimated, following the method explained in this paper. The distribution of these LRs is illustrated by the black curve in Fig. 8.

This way of presentation illustrates simultaneously the performance of the automatic speaker recognition method when the one or the other of the two alternative hypotheses H_1 or H_2 is verified.

7. Conclusion

In the paper, a new 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 value 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 new method was developed in order to find an adequate solution for the interpretation of voice recording as scientific evidence in the judicial process.

The corpus based methodology introduced in the paper provides a coherent way of assessing and presenting this kind of scientific evidence. The paper gives step by step guidelines for the calculation of the evidence and its assessment taking into account the models of the within-source variability of the suspected speaker and the between-source variability of the questioned recording.

In this paper we also presented how to use Tippett plots to study the performance of the automatic speaker recognition and the adequacy of the databases used in forensic speaker recognition application. In this way the Tippett plots contribute to the evaluation of the forensic utility of the method and databases.

8. References

[1] H. Tuthill, Individualization: Principles and Procedures in Criminalistics. Salem: Lightning Powder Co., 1994.

[2] C. Champod and D. Meuwly, The inference of identity in forensic speaker recognition, Speech Communication, vol. 31, pp. 193-203, 2000.

[3] S. E. Fienberg, The Evolving Role of Statistical Assessments as Evidence in the Courts, vol. 1., New York: Springer-Verlag, 1989, pp. 357.

[4] S. R. Lewis, Philosophy of Speaker Identification, presented at Police Applications of Speech and Tape Recording Analysis —Proceedings of the Institute of Acoustics, 1984.

[5] B. Robertson and G. A. Vignaux, Interpreting Evidence: Evaluating Forensic Science in the Courtroom , John Wiley & Sons, Chichester, 1995.

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

[7] D. Meuwly, M. El-Maliki, and A. Drygajlo, Forensic Speaker Recognition Using Gaussian Mixture Models and a Bayesian Framework, presented at 8th COST 250 Workshop: "Speaker Identification by Man and by Machine: Directions for Forensic Applications", Ankara, Turkey, 1998, p 52 — 55.

[8] D. A. Reynolds, Automatic Speaker Recognition Using Gaussian Mixture Speaker Model, The Lincoln Laboratory Journal, vol. 8, pp. 173 - 191, 1995.

[9] B. W. Silverman, Density estimation for statistics and data analysis. London: Chapman and Hall, 1986.

[10] C. Tippett, V. Emerson, M. Fereday, F. Lawton, A. Richardson, L. Jones, and S. Lampert, The Evidential Value of the Comparison of Paint Flakes from Sources other than Vehicles, J. Forensic Sci. Soc., pp. 61 - 65, 1968.

[11] I. W. Evett, J. S. Buckelton, Statistical analysis of STR data , in: 'Advances in Forensic Haemogenetics' (eds: Carraredo, A., B. Brinkmann, and W. B r) vol. 6, 1996, pp. 79 - 86.

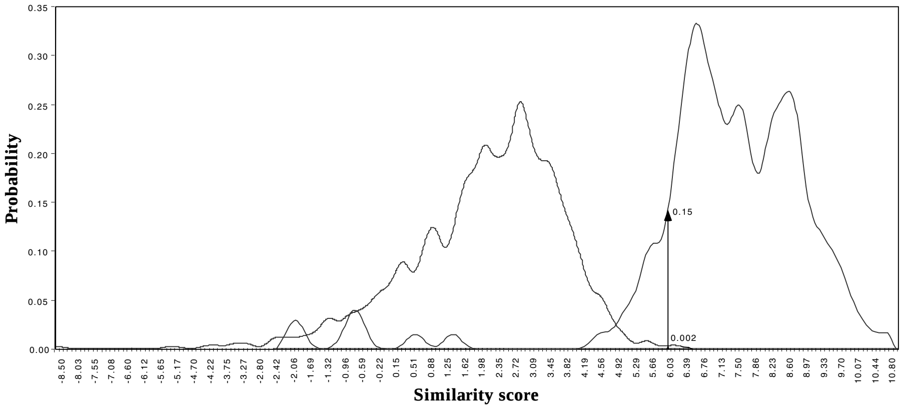

Figure 7:

Graphical representation of the likelihood ratio:

$p(E \mid H_1) / p(E \mid H_2)$

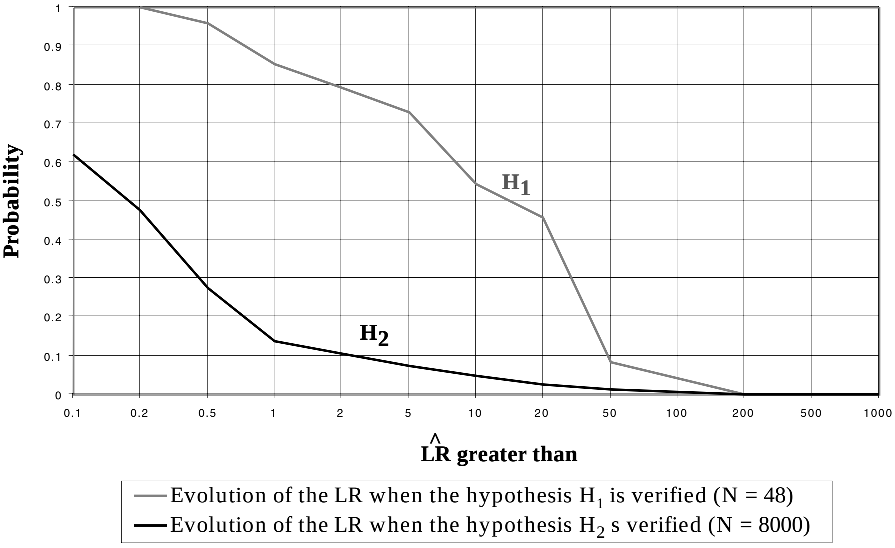

Figure 8:

Tippett plot