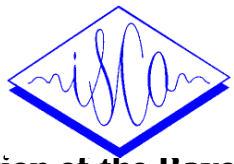

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems

Joaquin Gonzalez-Rodriguez[(1)], Javier Ortega-Garcia[(1)] and J.J. Lucena-Molina[(2)]

[(1)] Speech and Signal Processing Group (ATVS)
Audiovisual & Communications Dpt. (DIAC)
Universidad Politecnica de Madrid (UPM)
email: jgonzalez@diac.upm.es

[(2)] Laboratorio de Acustica e Imagen
Servicio de Policia Judicial
Direccion General de la Guardia Civil

Abstract

In this paper, excellent results are provided in the calculation of likelihood-ratios in real forensic conditions within the bayesian framework for the evaluation of speech evidences with a GMM-based speaker recognition system. Reported experiments have been performed with speakers from the Ahumada/Gaudí database, where 249 (122 male and 127 female) acted as reference population for the evaluation of the intervariability in each speech evidence, and the remaining 30 multisession male speakers acted as true/false suspects. Different GMM models have been trained from telephone recording sessions with different selections of the test files simulating different real forensic conditions. Results are provided in the form of likelihood ratios (LR) and are summarized in the form of Tippet plots, which are used to validate LR-based systems. All reported experiments have been performed with "IdentiVox" software, a tool for forensic speaker recognition that is actually been tested with real cases at Guardia Civil labs.

1. Introduction

During the last decades, an intense debate has taken place into the forensic community in order to achieve a common framework for the evaluation of evidence and its interpretation to the court. Fortunately, the Bayesian approach is now firmly established as a theoretical framework for any forensic discipline [1], even though it is not universally accepted. However, its application is not clear enough in every forensic area with the exception of DNA profiling, where likelihood ratios (LR) are "easily" obtained because of the discriminating power of the technique and the close relation between DNA profiles and populations (people). As an example, there are a lot of problems establishing (and updating) populations of cloth fibers, tool marks or paint traces in other forensic disciplines.

In this Bayesian framework, the roles of the scientist and the judge/jury are clearly separate, because the court wants to know the odds in favor of the prosecution proposition (C), ("the suspect has committed the crime"), given the circumstances of the case (I), and the observations made by the forensic scientist (E). These odds in favor of C are obtained from:

$$O(C|E,I) = \frac{Pr(E|C,I)}{Pr(E|\overline{C},I)} \cdot O(C|I)$$

Expressed in words, the Posterior odds = Likelihood ratio x Prior odds, where the prior odds concern to the court (background information relative to the case) and the likelihood ratio is provided by the forensic scientist. As a reference, Evett [1] propose a scale of likelihood ratios (LR) in the framework of DNA analysis with their respective linguistic qualifier suggesting the strength of verbal support for the evidence:

LR	*Verbal equivalent*
1 to 10	Limited support
10 to 100	Moderate support
100 to 1000	Strong support
Over 1000	Very strong support

Recently (Avignon, 1998) [2], the roles of speaker verification, speaker identification and type I and II error reporting have been properly criticized as alternatives to provide conclusions to the court. In that contribution, the use of the Bayesian approach is recommended because "assists scientists to assess the value of scientific evidence, help jurists to interpret scientific evidence, and clarify the respective

roles of scientists and of members of the court". In this way, the scientist alone cannot infer the identity of the speaker from the analysis of the scientific evidence, but gives the court the likelihood ratio of the two competing hypothesis (usually C, the questioned recording was made by the suspect, and $\overline{C}$, the questioned recording was not made by the suspect).

2. Likelihood ratio computation in speaker recognition

However, there is no closed solution to the problem of likelihood ratio (LR) computation. While it is assumed that the numerator of the LR calls for an assessment of the intra-variability of the system, and the denominator is the random match probability, those can be obtained from objective or subjective measures over relative frequencies in the relevant population.

Mewly and Drygajlo [3] propose a solution to this problem using automatic speaker recognition techniques. In their proposal, we have first to select the adequate population (usually from linguistic analysis or background knowledge), building speaker models (GMMs) with the selected individuals. We have also to record speech from the suspect, building a suspect speaker model and obtaining some reference utterances (SC: speech controls) that will be used later. The key issue here is the computation of the probability distributions (pdf.- probability density functions) of inter and intra-variability, where the speech evidence, that is, the likelihood of the questioned recording with the suspect model, will be referenced.

The speaker intravariability is computed as the distribution, assumed to be gaussian, of the likelihoods of the speech controls (reference recordings from the suspect) with the suspect model (*blue* pdf –monomodal- in the following figure). The intervariability is obtained as the (multimodal) pdf of the likelihoods of the questioned recording with the models of the (selected) reference population (*red* distribution in the following figure). Finally, the LR value is obtained as the quotient of the amplitudes of both distributions at the evidence likelihood (*black* line), as shown in figure 1. However, the experimental results presented in [3] were discouraging when their system was tested with real telephone speech.

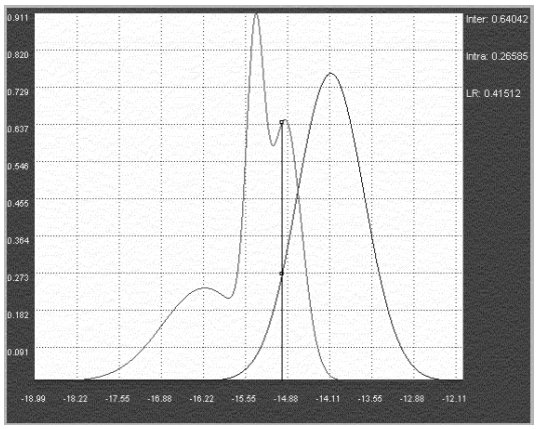

Figure 1.- Example of LR computation with an impostor audio file (LR=0.41), as provided by the system

3. Evaluation of LR-based systems

In order to test the abilities of systems providing their results in the form of LR values, some system calibration experiments have to be performed. Tippet [4], and later Evett [5], provides us a useful representation for between-source comparisons in any forensic discipline, representing proportion of cases with "LR values greater than …". Then, we will draw in Tippet plots (see figures 3, 4 and 5 as examples) simultaneously two curves, one for the C hypothesis (the system must provide high LR values) and another for the $\overline{C}$ hypothesis (the system must provide low LR values). In this way, for any x-axis value each curve shows proportion of cases with LR greater than x. Then, for greater separation between curves we will have higher discrimating power and then better systems (in an ideal system the curves should adjust to the upper-right and lower-left margins of the plot).

4. System design

IdentiVox software [8] is a multitask MDI Windows application developed with MS Visual C++. We have developed a classes library, programmed in ANSI C++, oriented to the development of automatic speaker recognition (and also other biometric) applications. The IdentiVox system performs, in a visual environment, speaker (and population) modeling, identification, threshold setting and verification, but also computes LR values from just the speech evidence, the suspect speech and a reference population, which can be easily built and selected with the population management functions of the system.

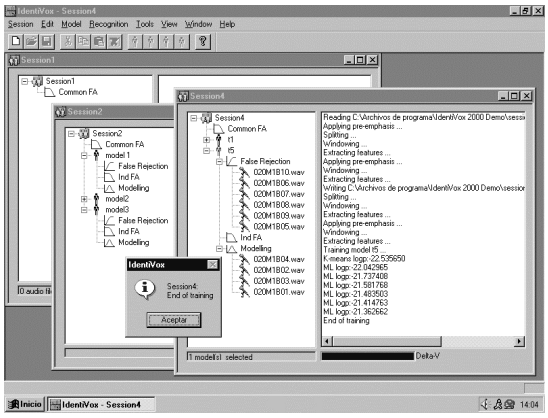

Figure 2. IdentiVox workspace and working sessions.

5. Experiments and results

In order to test our GMM-based system [6] in LR computations, some experiments have been designed (all described experiments have been performed with the *IdentiVox* software [8]). All the speech data have been obtained from the telephone sessions of Ahumada/Gaudí database [7]. Telephone speech 32 ms. windows (50% overlapped) are parameterized with 8MFCC + 8Δ + 8ΔΔ with CMN channel compensation. Speaker models are obtained via ML training with 32 gaussian mixtures from 1 minute of read speech. Speech controls (SC) and test files (TF) from the questioned recording are obtained from the 10 phonetically balanced read-phrases and the 10 digits strings for each corresponding session (3 to 5 seconds per test).

As we need separate suspects and population sets, we use the impostors monosession telephone recordings from Gaudí/Ahumada database, using 122 male and 127 female speakers as reference population for any LR value computation. Suspects speech have been obtained from the multisession male subcorpus (Ahumada), using in the reported experiments 30 speakers as "true" and as "false" suspects, upper and lower curves respectively in Tippet plots (by the time of writing this paper, results with 30 multisession women from Gaudí are being obtained, and the whole evaluation with the multisession speech from Gaudí/Ahumada – 103 male and 87 female speakers – will be available shortly).

Three different telephone tasks have been designed, simulating different usual conditions in forensic work:

Task T1: all speech comes from the same recording, which appears when the suspect acknowledges his own voice in a long conversation except in one or several utterances.

Task T2: suspect acknowledges multiple (irrelevant) conversations and one or several of them have to be tested (the speech evidence). In this task, we can perform multisession training.

Task T3: only one known recording is available from the suspect (typically recorded after his/her arrest in controlled conditions) and the speech evidence (one or several calls) comes from some time ago.

In order to simulate these situations, three different experiments have been designed, where the test files of each tested speakers acts both as true suspect with his model and as false suspect with the remaining (29) models. The results are shown in figures 4, 5 and 6 respectively (all reported experiments have been performed independently in the Acoustics and Image Forensic Laboratory of Guardia Civil, the spanish police institution equivalent to Carabinieri in Italy or Gendarmerie in France):

T1: all speech (train and test) comes from the first session, SC: 5 phrases and 5 digits-strings, TF: the remaining 5 phrases and 5 digits-strings.

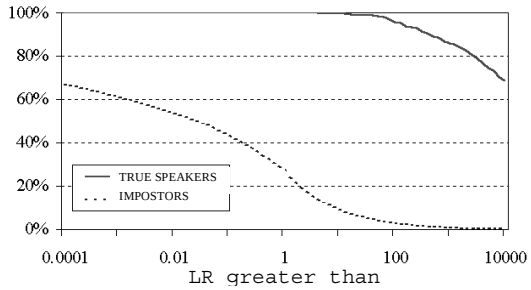

Figure 3.- C and $\overline{C}$ curves for task T1.

T2: training speech from sessions 1 and 2, SC: 5 different phrases from each session, and test speech from the third session, TF: 10 phrases.

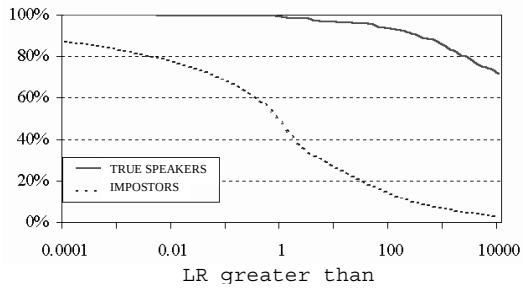

Figure 4.- C / $\overline{C}$ curves for task T2

T3: training speech from session 1, SC: 10 phrases, and test speech from sessions 2 and 3, TF: 5 different phrases from each session.

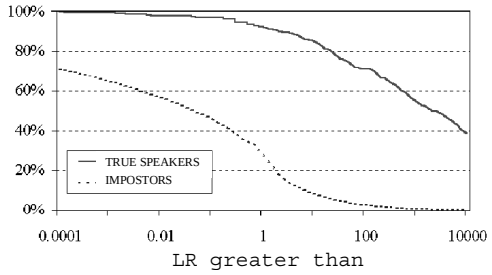

Figure 5.- $C / \overline{C}$ curves for task T3.

As an example, for a questioned recording and suspect model from the same recording (e.g., a long conversation where the suspect just not recognize his voice in certain phrases), in figure 3 we can see that a LR value greater than 100 will correctly reinforce the prosecution proposition.

Multiple postprocessing of this raw LR data is possible (mean, pdf distribution, discarding N-better and M-worst result, etc.), because multiple short length tests are usually possible from the same questioned recording, which will surely improve the results. However, every individual LR test is reported here in order to see the better and worst results of the system in any tested condition.

6. Conclusion

In this paper, a theoretically sound approach based in the Bayesian framework has been applied to the forensic work to be developed with speech evidences, obtaining high discriminating abilities when the system is tested with real multisession telephone speech with a big reference population in every LR test. While the system can obviously be improved, it provides a useful tool for the forensic scientist in order to provide to the court objective information relative to the speech evidence independently of the circumstances of the case.

ACKNOWLEDGEMENTS

Authors wish to thank all people from Speech and Signal Processing Group (ATVS) of the Universidad Politecnica de Madrid and Acoustics and Image Forensic Laboratory of Guardia Civil, and remark specially Marta Garcia-Gomar and Oscar Garcia-Ledesma from ATVS-UPM, directly involved in the development of the IdentiVox tool, and to Juan-Jesus Diaz-Gomez from Guardia Civil for its extensive testing and continuous suggestions.

7. References

[1] I.W. Evett, “Towards a Uniform Framework for Reporting Opinions in Forensic Science Casework”, Science & Justice 1998: 38(3), pp. 198-202.

[2] C. Champod and D. Mewly, “The Inference of Identity in Forensic Speaker Recognition”, Speech Communication, vol. 31, pp. 193-203, 2000.

[3] D. Mewly and A. Drygajlo, “The Influence of the Telephone Network in Automatic Forensic Speaker Recognition”, Proc. of Second European Academy of Forensic Science Meeting, Cracow (Poland), September 2000.

[4] Tippet C.F. et al., “The evidential value of the comparison of paint flakes from sources other than vehicles”, Journal of the Forensic Science Society, vol. 8, pp. 61-65, 1968.

[5] I.W. Evett and J.S Buckleton, “Statistical Analysis of STR (short tandem repeat) data”, Advances in Forensic Haemogenetics, A. Carracedo, B. Brickmann, and W. Bär, Editors. 1996, Springer-Verlag: Heidelberg, pp. 79-86.

[6] J. Gonzalez-Rodriguez, J. Ortega-García et al., “ATVS-UPM Results and System Description”, Site presentation at NIST 2001 Speaker Recognition Evaluation, MITAGS, Baltimore (USA), May 2001.

[7] J. Ortega-Garcia, J. Gonzalez-Rodriguez and V. Marrero-Aguiar (2000), “AHUMADA: a Large Speech Corpus in Spanish for Speaker Characterization and Identification”, Speech Communication, vol. 31 (3), June 2000, pp. 255-264, Ed. Elsevier Science.

[8] J. Gonzalez-Rodriguez, J. Ortega-Garcia, and J.J. Lucena-Molina, “IdentiVox: a PC-Windows Tool for Text-Independent Speaker Recognition in Forensic Environments”, Proc. of ENFSI Second European Academy of Forensic Science Meeting, Cracow (Poland), September 2000.