FORENSIC IDENTIFICATION REPORTING USING AUTOMATIC SPEAKER RECOGNITION SYSTEMS

J. Gonzalez-Rodriguez, J. Fierrez-Aguilar and J. Ortega-Garcia

ATVS (Speech and Signal Processing Group)
Dpt. Audio-Visual and Communication Engineering
Universidad Politecnica de Madrid (SPAIN)
jgonzalez@diac.upm.es

ABSTRACT

In this paper, we will show how any speaker recognition system can be adapted to provide its results according to the bayesian approach for evidence analysis and forensic reporting. This approach, firmly established in other forensic areas as fingerprint, DNA or fiber analysis, suits the needs of both the court and the forensic scientist. We will show the inadequacy of the classical approach to forensic reporting because of the use of thresholds and the suppression of the prior probabilities related to the case. We will also show how to assess the performance of those forensic systems through Tippet plots. Finally, an example is shown using NIST-Ahumada eval'2001 data, where the speaker recognition abilities of our system are assessed through DET plots, using then these raw scores as evidences into the forensic system, where relative to populations we will obtain the corresponding likelihood ratios values, which are assessed through Tippet plots.

1. INTRODUCTION

In this work, we will deal with the issue of how forensic scientists must report to the judge/jury their conclusions when speaker recognition techniques are used. In this sense, we will firstly note the difference from system characterization, that is, the identification abilities of the technique in use, to the characterization of the forensic system that will provide objective results to the Court. This is the key point of this contribution as forensic scientists must never arrogate the role of the judge/jury in taking decisions, and must know how to submit their results in order to comply with all the conditions of the judicial procedures, converting the system identification scores in meaningful values useful to the Court.

While commercial speaker recognition system performance, oriented to acceptance or rejection decisions, is widely assessed through different classical decision-based criteria, as type I and II errors or ROC and DET plots, an intense debate among forensic practitioners in "identification of the source" areas (as fingerprint, DNA, etc.) have taken place during the last decade in order to achieve a common framework for the evaluation of evidence and its interpretation to the court, and then how to assess the performance of forensic systems. Nowadays, the bayesian approach is firmly established as a theoretical framework for any forensic discipline. Forensic systems provide their results in the form of Likelihood Ratios according to this approach, being assessed, from the large experience gained in DNA-based person identification, through Tippet plots. In this paper, we will show the different nature of the outputs that automatic recognition systems must provide respectively in commercial and forensic approaches, even if the systems use the same core technology, and subsequently the need for different assessment tools specially suited for their corresponding applications.

2. ASSESSMENT OF SPEAKER RECOGNITION TECHNOLOGY

The objective of typical commercial speaker recognition systems is to accept true users and to reject impostors, usually minimizing some type of cost function. As the same technology could work for different systems in different operating conditions, it is usual to show all possible operating points. This has been done classically in detection tasks by means of ROC curves, showing the tradeoff of missed detections (false rejections) and false alarms (false acceptances).

However, as speaker recognition system performances increase, comparison of systems have become extremely difficult with this representation, as curves from different systems are extremely close to the lower left corner. This problem was overcome with the introduction of the DET (Detection Error Tradeoff) curve [1], which allows an almost linear representation of system performances, permitting easy observation of system contrasts (sample DET plots are shown in *figure 3*).

We want to note that this type of performance assessment (ROC/DET) perfectly suits the objective of the assessed systems, that is, to accept or reject users, because it directly shows both types of possible errors (missed detections and false alarms). Additionally, the core technology being used within any forensic system can also be assessed through ROC/DET curves or EER values, as has been shown in the literature [2][3].

3. CLASSICAL FORENSIC REPORTING

In the last years, the value of the different types of forensic evidence (even traditionally firmly established areas as fingerprint identification) have been severely attacked, questioning their scientific status, as is shown in influential books [4][5] and papers [6] in the field, specially "...after several highly publicized miscarriages of justice in which forensic expertise played a crucial role" [7].

Classically, there have been two different approaches to forensic reporting in "individualization of the source" areas, which includes areas as fingerprint, voice, face, signature, DNA, tool marks, paint, glass, fibers, and firearms. The first approach has been to provide just "identification" or

"exclusion/elimination" decisions, which results in a very high percentage of non-reporting cases. This approach has two main drawbacks: the first one is related with the use of subjective thresholds, specially in forensic conditions, as these techniques does not provide absolute identifications, where all the system/technique can provide is a score or a probability. Then, if the forensic scientist/system takes the subjective decision of identification or exclusion/rejection, he will be ignoring the prior probabilities related to the case (independent of the evidence under analysis), usurping the role of the court in taking this decision, as "... the use of thresholds is in essence a qualification of the acceptable level of reasonable doubt adopted by the expert" [8], even if these thresholds are adopted from objective measurements. The second drawback is the large amount of non-reporting cases that this identification/exclusion process induces, when "... there is no logical reason to suppress probability statements ... because ... any piece of evidence is relevant if it tends to make the matter which requires proof more or less probable than otherwise" [8].

The second classical approach to forensic reporting in this area consists in the use of a verbal scale of identification probabilities (typically "identification" / "very probable" / "probable" / "not conclusive" / "elimination"). This approach falls in the same errors as has just been noted, as it makes use of several subjective thresholds, and again ignores the prior probabilities (or usurp the judge/jury role if assign them) relative to every case.

4. BAYESIAN ANALYSIS OF FORENSIC EVIDENCE

Fortunately, the bayesian (or Likelihood-Ratio -LR-) approach is now firmly established as a theoretical framework for any forensic discipline [9][10]. As an example, there are eight Working Groups (DNA, Fibers, Fingerprint, Firearms, Handwriting, Tool Marks, Paint and Glass, Speech and Audio) in ENFSI (European Network of Forensic Science Institutes) dealing with individualization of the source. All of them [11], in discussions open also to non-European participants, have dealt or are dealing with the bayesian approach, looking for common standards and procedures.

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 (LR):

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

is provided by the forensic scientist. As a reference, in [9] a scale of likelihood ratios (LR) in the framework of DNA analysis is proposed with their respective linguistic qualifier suggesting the strength of verbal support for the evidence.

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" [8]. 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 voice was made by the suspect, and its opposite, it was not made by the suspect).

This likelihood ratio (LR), or Bayes factor, must be determined by the forensic scientist. In order to compute these numerator and denominator probabilities, population data need to exist in order to determine objective probabilities. For score-based systems, as all automatic speaker recognition techniques, speech databases are needed in order to model the distribution of measurements, both within and between sources, as this LR is in this case a ratio of probability density functions, rather than a ratio of probabilities.

Moreover, the bayesian approach allows to combine different types of evidence present in the process (blood type, fingerprint,...) and even the incorporation of subjective probabilities related to uncertain events, as shown in [10].

5. LR COMPUTATION IN FORENSIC SPEAKER RECOGNITION

In this section, we will show how any speaker recognition system can be turned into a bayesian forensic system. However, there is no closed solution to the problem of likelihood ratio (LR) computation, especially in the process of selection of the involved populations and its associated characteristics. 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, they can be obtained from objective or subjective measures over relative frequencies in the relevant population.

In [12] a solution to this problem for forensic speaker recognition is proposed using automatic speaker recognition techniques (*figure 1*). In this proposal, we have first to select the

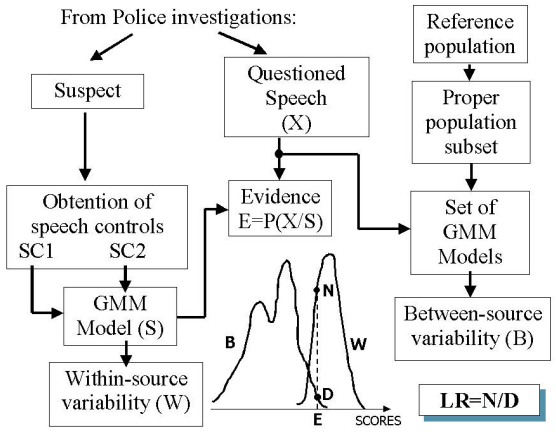

Fig. 1. Likelihood Ratio computation in Forensic Analysis of Speech Evidences.

adequate population (usually from linguistic analysis or background knowledge), building speaker GMMs [13] with the selected individuals. We have also to record speech from the suspect, building a suspect speaker model (GMM) with a part of it, and obtaining some reference utterances (SC: speech controls) that will be used to estimate the statistical distribution standing for the within-source variability.

Within-source distributions are usually assumed to be gaussian, as is obtained from the likelihoods of the speech controls (reference recordings from the suspect) with the suspect model. However, the between-source distribution cannot assumed to be gaussian as a reference population is involved. This estimation is performed in [12] using kernel density estimation, which gives a detailed model of the actual histogram. In our proposal [14], the between-source distribution estimation is performed with a multigaussian estimate (single dimension GMM) in order to avoid excessive details in the distribution, as the selected population (usually hundreds or thousands of speakers) is representing all possible speakers relative to the case (language, dialect, sex,..).

6. ASSESSMENT OF FORENSIC SPEAKER RECOGNITION SYSTEMS

In order to test the abilities of systems providing their results in the form of LR values, some assessment experiments have to be performed. In [15], a useful representation for between-source comparisons in any forensic discipline, the so-called Tippet plot, is provided, representing proportion of cases with "LR values greater than…". Then, we will draw in Tippet plots (*figure 2*) simultaneously two curves, one for the C hypothesis (the voice belongs to the suspect – target), where the system must provide high LR values (LR>>1), and another one for the opposite hypothesis (the voice does not belong to the suspect – non-target), where the system must provide low LR values (LR<<1). In this way, for any x-axis value each curve shows proportion of cases with LR greater than x. Then, the greater the separation between curves, the higher the discriminating power and the better the system (in an ideal system the curves should adjust respectively to the upper-right and lower-left margins of the plot). Additionally, good performance in LR values close to one is highly desired, that is, target LRs greater than one and non-target LRs smaller than one.

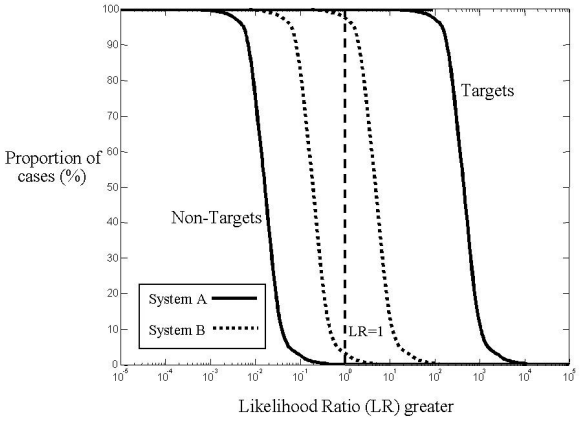

Figure 2. Example of Tippet curves for two competing systems.

7. FORENSIC EVALUATION WITH NIST-AHUMADA EVAL'2001 DATA

In this section, we will show the close relations and significant differences in the assessment of speaker recognition systems when used in commercial or forensic applications. An interesting example is presented here, where the authors will show the convenience and mutual relationship of both DET and Tippet curves in different environments but with the same data. We have used the NIST-Ahumada [16] data of year 2001 evaluation, which will be used to assess respectively the technology of our research group (ATVS-UPM), as to be used in any commercial/decision-oriented application, and the forensic system we have developed, according to the bayesian approach, based in this technology.

In *figure 3* we show the performance of our GMM-UBM implementation with the eval'2001 NIST-Ahumada data in an extended version of the "all" condition (every two 30 s. test file per speaker is tested with all 103 male models). In eval'2001 workshop, the authors presented [17] an UBM MAP-adapted GMM system with Tnorm, with a basic coefficient vector of 8 MFCC+delta+delta-delta. In *figure 2* this basic system has been improved suppressing the delta-delta coefficients and increasing the basic vector size to 12 or 19:

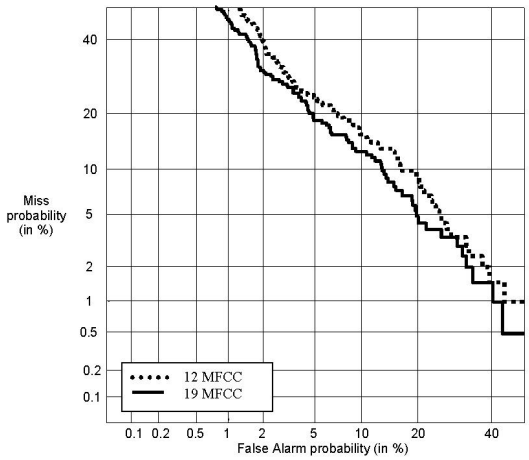

Figure 3. DET-plots for two versions (12 MFCC or 19 MFCC) of ATVS-system with NIST-Ahumada eval'2001data.

As can be seen, the 19 MFCC system performs better than the 12 MFCC one, assessed from a DET curve closer to the origin of coordinates (note that the best NISTeval'01 reported system with these data was just slightly better than this 19MFCC system). However, if we want to use any of these two systems in a forensic application, apart from the theoretical problems exposed previously (subjective thresholds and suppression of prior probabilities), the operating point of the system should have a very low (or even null) false acceptance rate. As an example, a false alarm probability of 1% (even this could be not accepted by the court) would mean a miss detection rate greater than 40%, which leads to a extremely high non-reporting rate. Does it mean that we cannot use this state-of-the-art speaker recognition technology in most forensic cases?

In this experiment, the same eval'2001 raw scores have been used to compute LR values, in order to show the

performance of a GMM-based forensic system. As we have just available in this dataset one speech file per speaker to build a model, and two test files per speaker, we will always use one of the files as test file, and the other one will be used as speech control. This is the information needed to estimate the within-source variability distribution (as just one likelihood is available, it will be used as mean value of a single-gaussian distribution with variance that of all speakers with his own test files). For the computation of every single Likelihood Ratio, we have selected as reference population the remaining 102 male speakers. Then, the between-source variability is obtained as the distribution of the likelihoods of every test file with all non-target models. Once we have the two distributions available for every test file, we compute the LR values and summarize them in the Tippet plots of *figure 4* for both systems (12/19 MFCC). Every Tippet plot is composed of two curves, target speakers (103x2=206 trials) and non-target speakers (103x2x102= 21012 trials):

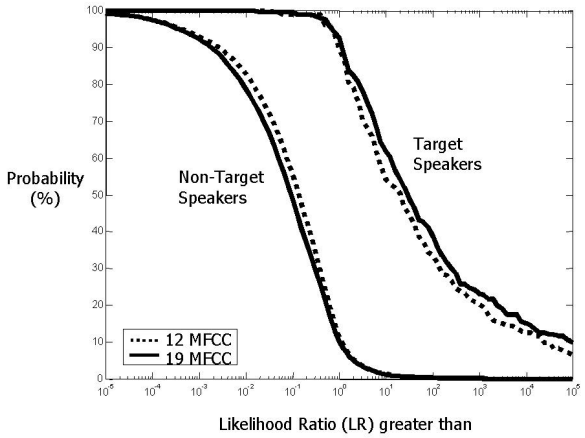

Figure 4. Tippet plots with NIST-Ahumada eval'2001 data.

As can be seen, the better the system the greater the separation between target and non-target curves for each system. But we want to note that, independently of the selected system (12/19 MFCC), we can provide a meaningful LR value for every single file or forensic case. From the results shown in the Tippet curves, the proposed system will certainly strengthen the prosecution hypothesis with target-speakers (LR>1) and will attenuate it for non-target speakers (LR<1), complying with the desired performance. Moreover, the system is not assuming any prior probability nor taking any decision, which corresponds to the court, and just limits its role to reinforce or attenuate the prosecution hypothesis.

8. CONCLUSION

We have shown in this contribution how any speaker recognition system can be adapted to work in the forensic environment according to the bayesian approach. Additionally, the roles of ROC/DET and Tippet plots in commercial and forensic applications have been clarified. While ROC/DET curves assess system/technology performance, they cannot be used to provide conclusions to the court as acceptance or rejection of speakers is not the objective of forensic speaker recognition. We have shown how easily a GMM-based system can be adapted to provide LR values according to the bayesian approach, allowing the court to take into consideration the prior probabilities of the case, and even to combine it with other types of evidence (DNA…). Finally, an interesting example have been presented with NIST-Ahumada eval'2001 data, comparing the roles of DET and Tippet plots for assessment respectively of a speaker recognition technology and a forensic system in the bayesian approach.

9. REFERENCES

1. Martin, A. et al., "The DET curve in assessment of detection task performance", Proc. EuroSpeech'97, pp. 1895-1898, Rhodes (Greece), 1997.
2. Nakasone, H., and Beck, S., "Forensic Automatic Speaker Recognition", Proc. of Odyssey'2001 Speaker Recognition Workshop, pp. 139-144, Crete (Greece), 2001.
3. Gonzalez-Rodriguez, J. et al., "IdentiVox: a PC-Windows Tool for Text-Independent Speaker Recognition in Forensic Environments", Proc. of ENFSI (European Network of Forensic Science Institutes) Meeting, Cracow (Poland), September 2000.
4. Robertson, B. et al., Interpreting Evidence–Evaluating Forensic Science in the Courtroom, Wiley, UK, 1995.
5. Foster, K.R. & P.W. Huber, Judging Science: Scientific Knowledge and the Federal Courts, MIT Press, 1997.
6. Taroni, F., Aitken, C.G.C., "Forensic Science at Trial", *Jurimetrics Journal* 37, 327-337, 1997.
7. Broeders, A.P.A., "Forensic Speech and Audio Analysis: the State of the Art in 2000 AD", Proc. of SEAF-2000, Ed. J. Ortega-García, Madrid (SPAIN), 2000.
8. Champod, C., "The Inference of Identity in Forensic Speaker Recognition", *Speech Comm.*, 31, 193-203, 2000.
9. Evett, I.W., "Towards a Uniform Framework for Reporting Opinions in Forensic Science Casework", *Science & Justice* 1998: 38(3), pp. 198-202.
10. Aitken, C.G.C., "Statistical Interpretation of Evidence /Bayesian Analysis", Encyclopedia of Forensic Sciences, pp. 717-724, Academic Press, 2000.
11. Meuwly, D., "Current Discussions of the ENFSI-WG About the Use of the Bayesian Approach for the Interpretation of Evidence", ENFSI Speech and Audio Group Meeting, Paris (France), 2001.
12. Meuwly, D., Drygajlo, A., "Forensic Speaker Recognition based on a Bayesian Framework and Gaussian Mixture Modeling", Proc. of Odyssey'2001 Speaker Recognition Workshop, Crete (Greece), 2001.
13. D.A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Models", *Speech Communication*, vol. 17, pp. 91-108, Elsevier, 1995.
14. Gonzalez-Rodriguez, J. et al., "On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems", Proc. of Odyssey'2001 Speaker Recognition Workshop, pp. 135-138, Crete (Greece), 2001.
15. 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.
16. Ortega-Garcia, J., Gonzalez-Rodriguez, J. et al., "AHUMADA: a Large Speech Corpus in Spanish for Speaker Characterization and Identification", *Speech Communication*, vol. 31, pp. 255-264, 2000.
17. Gonzalez-Rodriguez, J. et al., "ATVS-UPM Results and Presentation at NIST'2001 Speaker Recognition Evaluation", Linthicum Heights, Maryland (USA), 2001.