Available online at www.sciencedirect.com
SCIENCE DIRECT®
Forensic Science International 146S (2004) S101–S106
Forensic Science International
www.elsevier.com/locate/forsciint

On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition

F. Botti[a,*], A. Alexander[b], A. Drygajlo[b]

[a]*Ecole des sciences criminelles, Institut de Police Scientifique, University of Lausanne, Switzerland*
[b]*Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne, Switzerland*

Abstract

This paper deals with a procedure to compensate for mismatched recording conditions in forensic speaker recognition, using a statistical score normalization. Bayesian interpretation of the evidence in forensic automatic speaker recognition depends on three sets of recordings in order to perform forensic casework: reference (*R*) and control (*C*) recordings of the suspect, and a potential population database (*P*), as well as a questioned recording (*QR*) [1]. The requirement of similar recording conditions between suspect control database (*C*) and the questioned recording (*QR*) is often not satisfied in real forensic cases. The aim of this paper is to investigate a procedure of normalization of scores, which is based on an adaptation of the Test-normalization (T-norm) [2] technique used in the speaker verification domain, to compensate for the mismatch. Polyphone IPSC-02 database and ASPIC (an automatic speaker recognition system developed by EPFL and IPS-UNIL in Lausanne, Switzerland) were used in order to test the normalization procedure. Experimental results for three different recording condition scenarios are presented using Tippett plots and the effect of the compensation on the evaluation of the strength of the evidence is discussed.

Keywords: Automatic speaker recognition; Forensic speaker recognition; Normalization techniques; Mismatched recording conditions

1. Introduction

In forensic speaker recognition, the expert has usually a questioned recording and recordings of a suspected speaker. He has to help the court to address the issue as to whether or not the voice in both these recordings comes from the same person. In order to analyse two recordings, an automatic speaker recognition system creates a statistical model of the features of one of them (the suspect's recordings, *R*), and estimates the likelihood of the features of the second recording (the questioned recording, *QR*) by comparing it to this model.

The methodology considered here is the Bayesian interpretation approach applied to forensic automatic speaker recognition presented in [1] and in [3] which requires, apart from the questioned recording (*QR*), the use of three databases: a suspect reference database (*R*), a suspect control (*C*) and a potential population database (*P*).

The *P* database contains the recordings of all the possible voices from the potential population which can be used to evaluate the hypothesis: ''anyone chosen at random from a relevant population could be the source of the questioned recording''.

The *R* database contains recordings of the suspected speaker that are ideally very similar, in terms of recording conditions, to the recordings of speakers of *P* and is used to create the model of the suspected speaker, exactly as it is done with the models of *P*.

The *C* database consists of recordings of the suspect that are as close as possible (in recording conditions and

* Corresponding author.
E-mail addresses: filippo.botti@esc.unil.ch (F. Botti), alexander.anil@epfl.ch (A. Alexander), andrzej.drygajlo@epfl.ch (A. Drygajlo).

0379-0738/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2004.09.032

linguistically) to the *QR* and is used to estimate the within-source variability of the voice of the suspected speaker. However, in forensic casework, it is not always possible to record the suspected speaker in the same recording conditions as the questioned recording (*QR*). On one hand, indeed, it is difficult to know the recording conditions of the *QR* and to reproduce them, while on the other hand, the recordings of the suspect made by the police are often not similar to the *QR* and it is not possible to obtain more recordings, since, for example, the police does not want to alert him to the fact that he is under investigation. Thus, the assumption of compatibility between the *C* database and the *QR* is often not satisfied. In such cases, if the mismatch is not compensated, a bias is introduced in the results.

In this paper, we test an adaptation of the T-norm technique used in the domain of speaker verification, for the Bayesian approach used in forensic automatic speaker recognition.

In this paper, we present: a brief summary of the steps involved in the Bayesian interpretation approach for forensic automatic speaker recognition, a discussion of the bias resulting from the use of mismatched databases, a normalization technique and three scenarios in which the recording conditions of the databases used were different. For each scenario, the obtained results, with and without applying the normalization, are illustrated using Tippett plots.

2. Bayesian interpretation methodology

In forensic speaker recognition, the probability of the result of the comparison between the statistical model of the suspect's speech (obtained from *R*) and the features of the questioned recording (*QR*) is called *E* (evidence) and is evaluated with respect to two hypotheses: H_0—"the suspect is the source of the questioned recording", and H_1—"someone else in the relevant population is the source".

The probability of *E* under the hypothesis H_0 is given by an estimation of the within-source variability of the suspected speaker which corresponds to the distribution of scores obtained by comparing different utterances of the suspected speaker (from the *C* database), with his statistical models (created with the *R* database).

The hypothesis H_1 is represented by an estimation of the between-sources variability of the *QR* with respect to the relevant population, which corresponds to the distribution of the scores obtained by comparing the features of the questioned recording (*QR*) with statistical models of all the speakers in the potential population (*P*).

The strength of the evidence is provided by the likelihood ratio (LR) (i.e., the ratio of support that the evidence (*E*) lends to each of the hypotheses) which is given by the relative probability of observing the score *E* under the distributions defined by each of the hypotheses. This corresponds to the ratio of the density of the within-source and between-sources probability density distributions for the score *E*.

3. Effect of mismatched conditions of the *QR* and the *C* database

As discussed in the previous section, the requirement in the Bayesian interpretation methodology, that the suspect control database (*C*) is in similar conditions to the questioned recording (*QR*) is often not satisfied in real forensic cases. Thus, LRs obtained are biased if the *C* database is not in the same recording conditions as the *QR*, since automatic speaker recognition system depends on the recording conditions.

We define a matched comparison as comparing two recordings in similar conditions (e.g. both from public switched telephone network (PSTN)), and a mismatched comparisons as comparing two recordings in different conditions (e.g. one from PSTN and the other from global system for mobile communications (GSM)).

In Fig. 1, an illustration of such a bias in the estimation of LR is presented. It can be observed that when the H_1 distribution is the result of mismatched comparisons, if H_0 distribution is plotted from the result of matched comparisons, the likelihood ratio estimated for *E* is biased, since the scores range of the two distributions is incompatible. In the example of Fig. 1 the LR would be underestimated if we were using the H_0 distribution obtained in matched conditions while we should employ the H_0 distribution obtained in the same mismatched conditions as for the estimation of H_1 However, in real cases, this estimate of the H_0 distribution in mismatched conditions is often not available (e.g., in absence of control recording of the suspect in same recordings condition of the *QR*). In the following section, we propose to normalize the H_0 matched scores in order reduce the effect of such a bias.

4. Normalization

The normalization performed here is an adaptation of the test-normalization (T-norm) technique used in the domain of speaker verification, for the Bayesian interpretation approach to forensic automatic speaker recognition [2].

The idea behind this normalization is to transform all the scores obtained by comparing a given test utterance with all the models in the potential population and with the suspects' model, to a mean (μ) of zero and a variance (σ) of one. We have observed that, even in conditions of mismatch, the speaker who is the source of the *QR* often obtains a higher score than the other speakers.

The suspect within-source variability scores are normalized with the scores obtained comparing the *C* database with the speakers models of *P*, according to the global distribution $N(0,1)$. The same is performed with the scores obtained comparing *QR* with the models of *P* and of *R*.

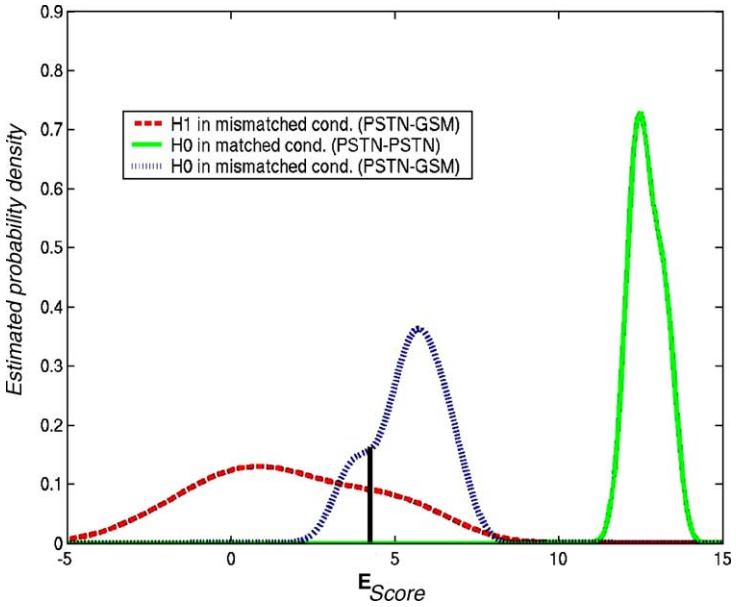

Fig. 1. Representation of the mismatch. The recordings used to create the distribution of H_0 in mismatched conditions are often not available in forensic real cases.

In Fig. 2 it can be observed that the two distributions of H_0 for matched and for mismatched conditions are very similar after the normalization and hence the LR calculated for E, would be less affected.

Although the two distributions were originally in different ranges for matched and mismatched conditions, after normalization they are in a similar range. It is important to note that the respective range of each speaker for the given test recording remains the same after the normalization.

To sum up, the normalization is performed on the scores obtained for each test recording (C and QR databases) compared to all the speaker's models in the case. First,

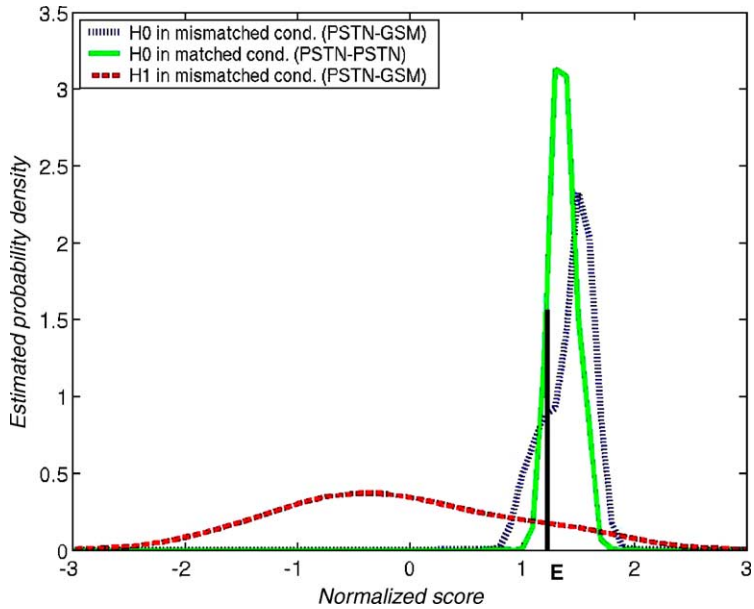

Fig. 2. Effect of the normalization. The two distributions of H_0 after the normalization show similar characteristics.

Table 1
The normalization ($\mu = 0$ and $\sigma = 1$) is performed for each row separately

				Speakers' models (P)		
			R	P_{M1}	P_{M2}	P_{M3}
Test recordings (*t*)		*QR*	*E*	$S_{(QR,\,M1)}$	$S_{(QR,\,M2)}$	$S_{(QR,\,M3)}$
	C database	C_1	$S_{(C1,\,R)}$	$S_{(C1,\,M1)}$	$S_{(C1,\,M2)}$	$S_{(C1,\,M3)}$
		C_2	$S_{(C2,\,R)}$	$S_{(C2,\,M1)}$	$S_{(C2,\,M2)}$	$S_{(C2,\,M3)}$
		C_3	$S_{(C3,\,R)}$	$S_{(C3,\,M1)}$	$S_{(C3,\,M2)}$	$S_{(C3,\,M3)}$

■: Evidence score; ▒: Between-sources variability (H_1); □: Within-source variability (H_0); □: Scores that are used for normalization purpose only.

the scores (S) given by the comparison of the *QR* with each model of P and with the suspect's model (from the R database) are all normalized in order to obtain a mean of zero and standard deviation of one. The same operation is performed with each of the suspect's control recordings (C), which are also compared to all the models of the P database as well as to the suspect's model (R). The scores are normalized separately for each row of Table 1. The proposed normalization is expressed in Eq. (1).

$$S\text{norm}_{(t,\text{M}n)} = \frac{S_{(t,\text{M}n)} - \mu_t}{\sigma_t} \tag{1}$$

where t is the test recording (the *QR* as well as each recording of the C database), M_n is the model of the speaker n (from the P and the R databases), $S_{(t,M\text{n})}$ is the score obtained when the test recording t is compared with the model M_n before the normalization. $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ are respectively the mean and the variance of the scores obtained by comparing the test recording t with all the models from P and R.

5. Experiments

The effect of the normalization has been studied comparing the results obtained with and without normalization for three different scenarios. The recording conditions of the databases used were different in each scenario, in order to obtain a better overview of the effect of the normalization.

ASPIC, a text-independent automatic speaker recognition system, has been used for the experiment. The feature extraction is carried out using 12 RASTA–PLP coefficients [4,5]. The statistical model of the speaker is created using a 32 Gaussian mixture model (GMM) [6,7]. Kernel density estimation is used to calculate the probability densities of distribution of scores for each of the hypotheses.

The database used is a subset of Polyphone–IPSC–02 which contains recordings of 12 male speakers recorded through a PSTN and a GSM phone. All the recordings used are spoken in French. For each speaker, the following recordings were used:

Table 2
Description of the databases recording conditions for each of the scenarios studied

	Scenario 1	Scenario 2	Scenario 3
P database		PSTN (P)	
R database		PSTN (R)	
C database	PSTN (C_a)	GSM (C_b)	PSTN (C_a)
QR	PSTN (T_a)	GSM (T_b)	GSM (T_b)

R database: five PSTN phone recordings of 90–180 s. to train their models.

C database: C_a—10 PSTN phone recordings, C_b—10 GSM recordings.

QR database: T_a—10 PSTN phone recordings; T_b—10 GSM recordings.

For each case constructed, the P database consisted of all the speakers apart from the suspect and the actual speaker considered.

The length of the recordings in C and T databases varies between 10 to 100 s. The mean scores of the comparisons of the five models with the test recording has been considered.

Three scenarios have been created, using the recordings described above. The recording conditions of the databases used in each scenario are explained in Table 2.

Each of the three scenarios was evaluated with and without the normalization, resulting in a total of six sets of cases. Hence, for each set we have 120 cases where the suspected speaker was indeed the source of the questioned recording (H_0 cases) and 1320 cases where the suspected speaker was not the source of the questioned recording (H_1 cases). For each case, a LR is calculated using the Bayesian interpretation methodology.

According to the methodology, in scenario 1 and 2, the conditions are matched and in scenario 3 the conditions are mismatched.

6. Results

The results of the comparisons are represented using a probability distribution plot called Tippett plot. The Tippett plot represents the proportion of the likelihood ratio greater than a given LR, i.e., $P(\text{LR}(H_i) > \text{LR})$, for cases corresponding to the hypotheses H_0 and H_1 [1,8]. Each graphical representation contains two curves, one for the LRs estimated for cases in which H_0 was true (the suspected speaker is the source of the questioned recording) and one for the cases in which H_1 was true (the suspected speaker is not the source of the questioned recording).

The separation between the two curves is indicative of how well the system discriminates between cases corresponding to H_0 and H_1. Figs. 3, 4 and 5 show respectively the Tippett plots for the scenarios 1, 2 and 3 without (not norm) and with the normalization (norm).

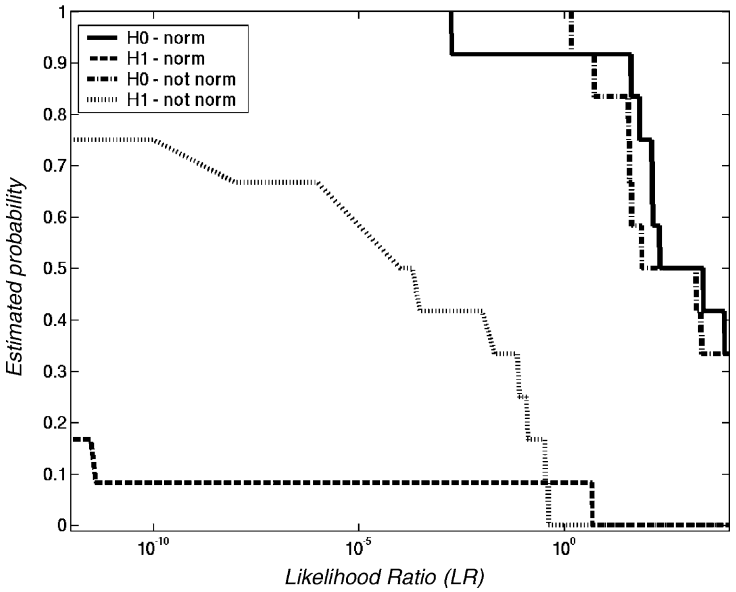

Fig. 3. Tippett plot for Scenario 1.

7. Discussion

A greater separation is observed between the curves corresponding to LR in H_0 cases and H_1 cases when the normalization was applied. In scenario 3, where a mismatch exists between the *C* database and the *QR*, before normalization, all the LRs obtained were less than one, providing support for the hypothesis H_1 even when the suspected speakers were truly the source of the *QR*. In the Tippett plot of this case, without normalization, the two curves show a high degree of overlap since all the LR obtained were very close to zero.

The LRs obtained for H_0 and H_1 cases were better separated when the normalization was applied, even in the cases where there was no mismatch between the *C* database and the *QR*. This can be explained because the rank of the speaker who is the source of the *QR*, with respect to all the other speakers, is in most cases the first, and the

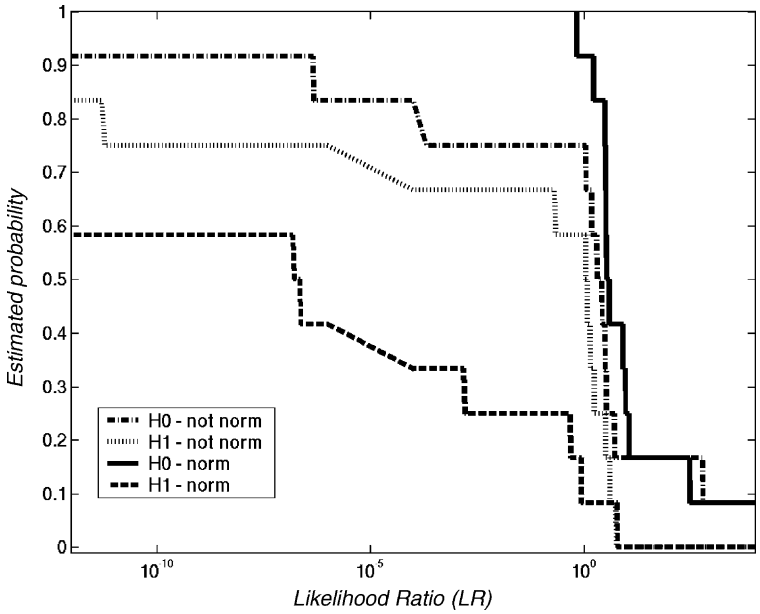

Fig. 4. Tippett plot for Scenario 2.

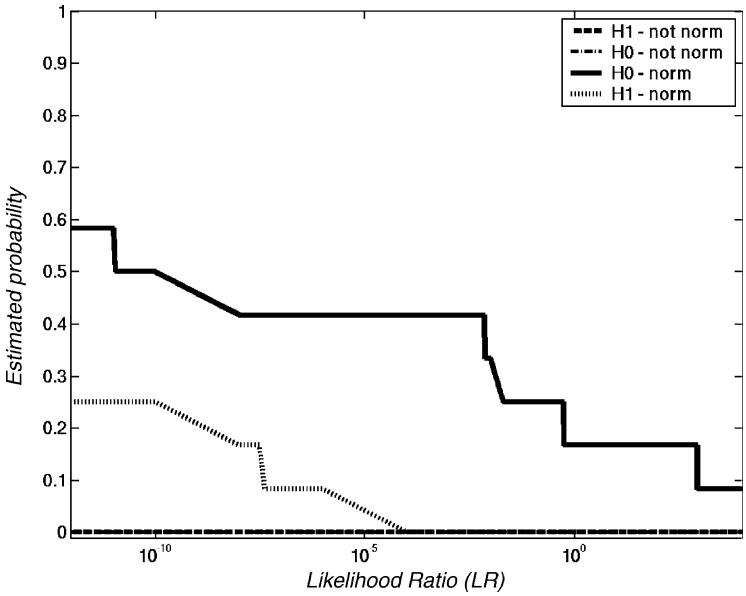

Fig. 5. Tippett plot for Scenario 3.

same is for his control recordings (*C*). However, the H_0 scores of some control recordings for this speaker are lower than H_1 scores of the *QR* (such phenomenon is described in [9]). The normalization helps to overcome this problem, since the ranks are conserved after the normalization and the scores for each test recording (*QR* and control recordings) are normalized with each other.

It is to be noted that the size of the potential population used (*P* database) is small, that the experiments were performed in closed set and were not statistically independent. A larger database and mutually independent cases should be considered. Other normalization techniques (e.g. Z-norm, cohort and unconstrained cohort normalization [2]) can be adapted and tested in this framework.

8. Conclusion

A statistical score normalization to deal with mismatched recording conditions in forensic speaker recognition was presented. In the methodology used for forensic automatic speaker recognition based on the Bayesian approach [1], the similarity of recording conditions between the suspect control database (*C*) and the questioned recording (*QR*) is required. However, in real cases, such compatibility often cannot be satisfied.

In this paper, a statistical normalization on scores based on an adaptation of the T-norm [2] (used in the speaker verification approach) was proposed and tested.

A general improvement in the separation of the likelihood ratio (LR) obtained for H_0 true cases and H_1 true cases has been observed when the normalization was used, for both cases with and without mismatch recording conditions between the *C* database and the *QR*. The effects of the normalization on the estimation of the LRs were also discussed.

References

[1] D. Meuwly, Reconnaissance de locuteurs en science forensiques: l'apport d'une approche automatique, PhD thesis, IPS, University of Lausanne: Lausanne, 2001.
[2] R. Auckenthaler, M.J. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification system, Digital Signal Processing (DSP) 10 (1-3) (2000) 145–150.
[3] D. Meuwly, A. Drygajlo, Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modeling (GMM). A speaker Odyssey, The Speaker Recognition Workshop, Crete, Greece, 2001, pp. 145–150.
[4] H. Hermansky, RASTA processing of speech, IEEE Trans. Speech Audio Process. 4 (2) (1994) 578–589.
[5] H. Hermansky, et al. RASTA-PLP speech analysis technique, IEEE Trans. Speech Audio Process. (San Francisco, USA), 1 (1992) 1121–1124.
[6] D.A. Reynolds, A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification, PhD thesis, Georgia Institute of Technology: Atlanta, USA, 1992.
[7] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. on Speech and Audio Processing 3 (1) (1995) 72–83.
[8] I.W. Evett et al., Statistical analysis of STR (short tandem repeat) data, Advances in Forensic Haemogenetics, Heidelberg, 1996, pp. 79–86.
[9] G. Doddington, Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation, in: International Conference on Spoken Language Processing, Australia, 1998 (Paper 608 on CD-ROM).