Available online at www.sciencedirect.com
SCIENCE DIRECT®
Forensic Science International 146S (2004) S95–S99
Forensic Science International
www.elsevier.com/locate/forsciint

The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications

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

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

Abstract

In this paper, we analyse mismatched technical conditions in training and testing phases of speaker recognition and their effect on forensic human and automatic speaker recognition. We use perceptual tests performed by non-experts and compare their performance with that of a baseline automatic speaker recognition system. The degradation of the accuracy of human recognition in mismatched recording conditions is contrasted with that of the automatic system under similar recording conditions. The conditions considered are of public switched telephone network (PSTN) and global system for mobile communications (GSM) transmission and background noise. The perceptual cues that the human subjects use to perceive differences in voices are studied along with their importance in different conditions. We discuss the possibility of increasing the accuracy of automatic systems using the perceptual cues that remain robust to mismatched conditions. We estimate the strength of evidence for both humans and automatic systems, calculating likelihood ratios using the perceptual scores for humans and the log-likelihood scores for automatic systems.

Keywords: Forensic speaker recognition; Mismatched recording conditions; Automatic and aural recognition

1. Introduction

In a forensic context, the degree of similarity between a questioned recording and a suspected speaker's voice is called the evidence (*E*). Currently, in forensic speaker recognition, aural, automatic and a combination of aural and instrumental methods are used in order to estimate this similarity [1]. In a Bayesian interpretation framework, this similarity is evaluated with respect to two competing hypotheses, i.e., how likely it is that a questioned recording (trace) has been produced by a suspected speaker than by any other person. Conditions typical to forensic cases in which recordings are made by the police (anonymous calls and wiretapping) cannot be controlled and are far from ideal. Mismatch due to differences in the phone handset, in the transmission channel and background noise can affect the estimation of the strength of evidence. In this paper, we compare the recognition of laypersons (phonetically untrained human subjects) with a baseline automatic system that is not specially adapted to different recording conditions. The results of human recognition are adapted into a Bayesian interpretation framework, in order to evaluate the strength of evidence and are then compared with automatic speaker recognition results evaluated in the same framework.

2. Experimental framework

In this study, the Polyphone-IPSC02 database was used. This forensic speaker recognition database contained speech from 10 speakers in three different recording conditions and

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

0379-0738/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2004.09.078

two languages. The recording conditions include transmission through a public switched telephone network (PSTN) and a global system for mobile communications (GSM) network with the recordings performed using an analogue tape recorder answering machine as well as a digital recorder. French and German are the two languages present in this database.

In our experiments, we have used a subset of this database, using speech recorded through a PSTN and GSM network from speakers whose mother tongue is French. For one part of the test, artificially generated white noise was added to the PSTN recordings at a signal to noise ratio of 10 dB. Five short segments of spontaneous speech (between 10 and 20 s) for each speaker in PSTN, GSM and noisy conditions were used for the mock questioned recordings and five longer recordings were used as reference recordings (90 s) for the mock suspected speaker. All the test recordings were in French since it was the mother tongue of all the subjects, and studies had indicated that using languages apart from the speaker's mother tongue would influence the accuracy of recognition [2]. The mock questioned recordings chosen were simulated forensic cases with undisguised imitations of hoaxes and threats.

3. Listening task and test procedure

A total of 90 French speaking subjects were chosen for the experiments, each performing 25 comparisons, with no limitation on the number of times they could listen to a particular recording. None of these speakers had any formal training in phonetics. Care was taken to see that these subjects did not personally know the speakers whose voices appear in the tests. A seven level verbal scale was established, ranging from 'I am certain that the speakers in the two recordings are different' (corresponding to a score of 1) to 'I am certain that the speakers in the two recordings are the same' (corresponding to a score of 7). The option 'I am unable to make any judgement whether the speakers were the same or different' was placed at the middle of the scale (corresponding to a score of 4). The scores 2 and 3 corresponded to 'I am almost sure that the two speakers are different' and 'it is possible that the two speakers are different'. The scores 4 and 5 corresponded to 'it is possible that the two speakers are the same' and 'I am almost sure that the two speakers are the same'. The seven level scale was chosen because of studies that suggest that it is difficult for humans to differentiate accurately for more levels of comparison of a certain stimulus [3]. The subjects were required to listen to the suspect reference recording and the questioned recording, and to estimate their similarity on the verbal scale. The order in which the comparisons were presented was randomized in order to minimize the effect of memorisation, by which the subject would remember the voices heard in previous comparisons, and use it in order to make their decision. All the tests in this analysis were performed with a single computer, using exactly the same set of headphones. At the end of the recognition session, the subjects were also asked to list the factors they considered important when making their decisions. Each test session corresponded to approximately 1 h, with a total of 90 h, spent in order to perform the aural comparisons.

In the automatic system, feature extraction was performed using a RASTA–PLP [4] and statistical modelling of these features using a Gaussian mixture modelling (GMM) based classifier with 32 Gaussian mixture components. The zones where speech was not present had been removed in a pre-processing phase, in order to avoid including non-speech information in the statistical models of the speakers' voices. During the testing phase, features were extracted from the test utterances and compared with the statistical model of the speaker created in the training phase. The likelihood of observing the features of the test recording for the statistical model of the suspected speaker's voice was calculated. The logarithm of the likelihood obtained was used as a score for the comparison of the test utterance and the model. These scores were normalized per training utterance to a mean of zero and a standard deviation of one in order to standardize scores across speakers. The total computation time utilized to perform feature extraction, training and testing for all the comparisons was approximately 3 h.

4. Evaluating the strength of evidence

Both the automatic speaker recognition system and the human subjects give scores for the comparison between the two recordings, which indicate how similar the two recordings are to each other. In forensic automatic speaker recognition the evidence (E) is the degree of similarity, calculated by the automatic speaker recognition system, between the statistical model of the suspect's voice and the features of the trace [5]. In the forensic aural recognition case, however, an estimate of the similarity score (E) between the suspected speaker's voice and the questioned recording is given by a subject on a verbal perceptual scale. In order to evaluate the strength of evidence, the forensic expert has to estimate the likelihood ratio of the evidence given the hypotheses that two recordings have the same source (also known as hypothesis H_0) and that the two recordings have a different source (hypothesis H_1). The likelihood ratio, illustrated in Fig. 1, is the ratio of the heights of the distribution of scores for hypothesis H_0 and H_1 at E.

The likelihood ratio for the perceptual scores given by the test subjects can be calculated in a similar way using the Bayesian interpretation method [5]. Since the scores derived in this perceptual scale are discrete scores (from one to seven) and not continuous as in the case of log-likelihood scores returned by automatic systems, a histogram approximation of the probability density can be performed. The likelihood ratio would then be the relative heights on the histogram of the two hypotheses H_0 and H_1 at the point E (shown in Fig. 2).

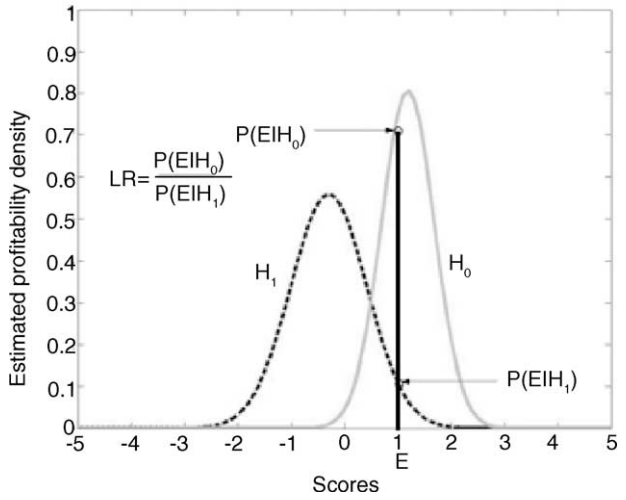

Fig. 1. Estimation of likelihood ratio using automatic speaker recognition scores.

5. Comparing the strength of evidence in aural and automatic methods

In order to measure the performance and reliability of each of the speaker recognition methods, the cases described in the previous section are considered, separating them into cases where 'it was known that the suspected speaker was the source of the questioned recording' and those where 'it was known that the suspected speaker was not the source of the questioned recording'. These results are represented using a probability distribution plot called Tippett plots. The Tippett plot represents the proportion of the likelihood ratios greater than a given L_R, i.e., $P(\mathrm{LR}(\mathrm{H}_i) > \mathrm{LR})$, for cases corresponding to the hypotheses H_0 and H_1. This way of representation of results was proposed by Evett and Buckleton [6] in the field of interpretation of the forensic DNA analysis. This representation has been named the ''Tippett plot'' in [6], referring to the concepts of ''within-source comparison'' and ''between-sources comparison'' defined by Tippett et al. [7].

We observe in the Tippett plots for both the aural and automatic systems, that the likelihood ratios for the H_0 and H_1 hypotheses are well separated. In Fig. 3, the aural and automatic likelihood ratios are presented for matched conditions, i.e., when both the suspected speaker's speech as well as the questioned recording were made in PSTN transmission conditions. In this plot, we observe that both the aural and automatic systems show good separation of the two curves, with the curves of the automatic system slightly more separated than those of the aural recognition.

In Fig. 4, the aural and automatic likelihood ratios are presented for mismatched conditions, i.e., when the suspected speaker's speech was recorded in PSTN conditions and the questioned recordings were made in noisy PSTN transmission conditions. Here, we observe that the separation between the curves corresponding to H_0 and H_1 has decreased in comparison to Fig. 3 (matched conditions). In this plot, we observe, that between the aural and automatic systems, the aural recognition shows slightly better separation than the automatic system.

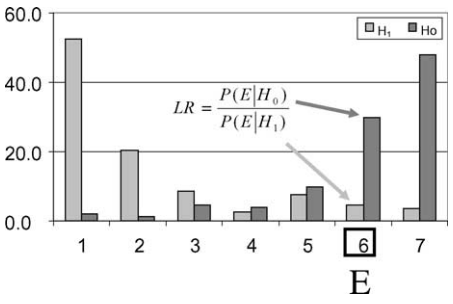

Fig. 2. Estimation of likelihood ratio using aural speaker recognition scores.

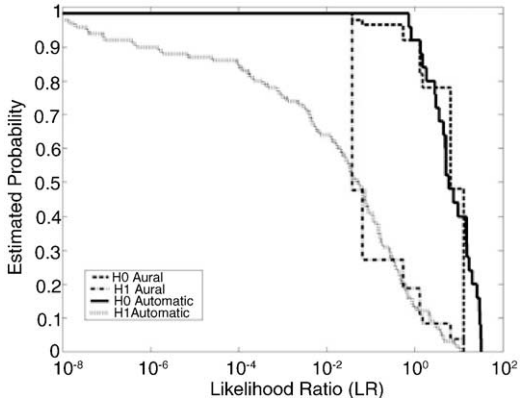

Fig. 3. Tippett plot in matched conditions (PSTN–PSTN) for aural and automatic recognition.

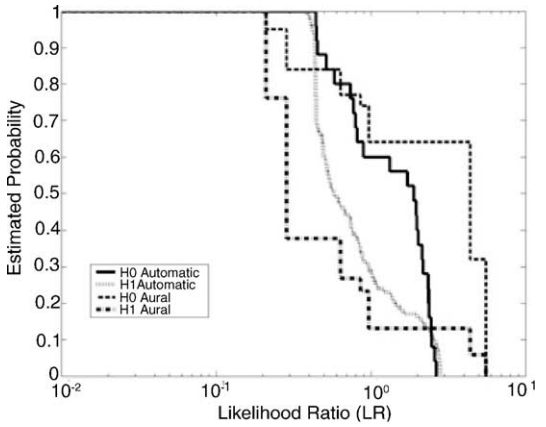

Fig. 4. Tippett plot in mismatched conditions (PSTN–noisy PSTN) for aural and automatic recognition.

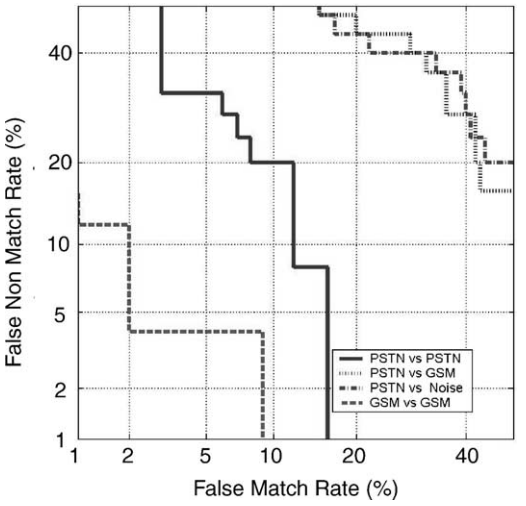

Fig. 6. DET plot for the automatic speaker recognition system.

6. Accuracy of aural and automatic systems in matched and mismatched conditions

In order to compare the performance of automatic speaker verification systems, often, receiver operator curves (ROC) or detection error tradeoff (DET) curves are used. The DET curve plots the relative evolution of false match and false non-match rate when using a decision point. The equal error rate (EER) is the point at which the false acceptance rate is equal to the false rejection rate and is used in order to compare the performance of automatic systems. Although the forensic speaker recognition task does not use a threshold, as in the speaker verification approach, this performance measure is informative.

In Fig. 5, we observe degradation in the accuracy of human recognition in matched conditions (training and testing with PSTN or GSM transmission conditions) compared to mismatched recording conditions (training and testing with PSTN and GSM, PSTN and noisy PSTN speech). Human aural recognition shows good performance in matched conditions with EERs as low as 16% and degraded performance in mismatched conditions with EERs of 30%.

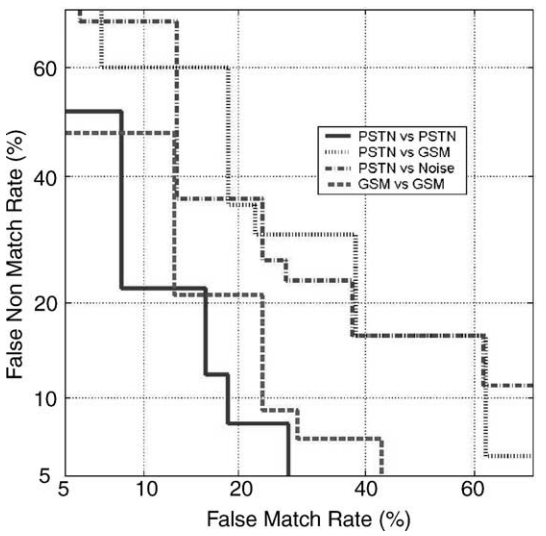

Fig. 5. DET plot for aural speaker recognition.

In Fig. 6, we observe a similar degradation in the accuracy of automatic speaker recognition in matched recording conditions as compared to mismatched recording conditions (training and testing with PSTN and GSM, PSTN and noisy PSTN). Automatic speaker recognition outperforms aural speaker recognition in matched conditions with EERs as low as 4 and 12%. However, in mismatched recording conditions, automatic systems show accuracies comparable to aural recognition when the channel conditions are changed. In noisy conditions, human aural recognition is better than the baseline non-adapted (to noise) automatic system.

7. Perceptual cues used by human subjects

From the previous sections, we can observe that aural recognition is robust to changes in recording conditions. Human beings depend largely on perceptual cues in the speech, such as pronunciation, word choice and laughs [8]. There have been studies which attempt to quantify the various aural and perceptual means that laypersons use in order to identify speakers [9].

In our study, we have been able to identify the factors important to each of the subjects, the accent (or pronunciation), the timbre, intonation, the rate of speech, speech defects or anomalies, breathing, the loudness, similarity to known voices and their intuition. These criteria were

Table 1
Criteria used by the test subjects to identify speakers

No.	Factor	PSTN (%)	GSM (%)	Noise (%)
1	Pronounciation, accent	34	31	30
2	Timbre	25	25	22
3	Intonation	16	24	18
4	Speech rate	9	7	12
5	Defects	6	7	8
6	Breathing	5	0	2
7	Loudness	3	0	0
8	Imagined physiognomy	3	0	2
9	Similarity to known voices	0	2	0
10	Intuitive feeling	0	4	6

obtained by asking the subjects, at the end of each experiment session, what factors they considered in recognizing the questioned recording. In Table 1, we have presented these factors and their relative importance to the subjects in each of the different conditions of recording. The recording conditions have been varied in order to study the differences in the perceptual cues that human beings use to recognize different speakers.

We observe that the main characteristics that humans depend upon, in all the three conditions are mainly the accent, timbre, intonation, rate of speech and speech anomalies. The relative importance of each of these main characteristics is very similar across different conditions implying that human perception of speaker identity mainly depends on characteristics that are robust to differences in conditions. This is in stark contrast to the automatic speaker recognition system which depends heavily on the conditions of recording. Considering these additional factors is of importance in severely degraded conditions, as is often the case in forensic casework.

8. Conclusions

Perceptual tests were performed with laypersons and their performance was compared with that of a baseline automatic speaker recognition system. It was observed that in matched recording conditions of training and testing, the automatic systems showed significantly better performance than the aural recognition systems. In mismatched conditions, however, the baseline automatic systems showed comparable or slightly degraded performance compared to the aural recognition systems. The extent to which mismatch affected the accuracy of human aural recognition in mismatched recording conditions was similar to that of the automatic system under similar recording conditions. The baseline automatic speaker recognition system should be adapted to each of the mismatched conditions in order to increase its accuracy.

References

[1] S. Gfroerer, Auditory instrumental forensic speaker recognition, in: Proceedings of the Eurospeech 2003, Geneva, Switzerland, 2003, pp. 705–708.
[2] A.D. Yarmey, Earwitness speaker identification, Psychol. Public Policy Law 1 (4) (1995) 792–816.
[3] G.A. Miller, The magical number seven, plus or minus two: some limits in our capacity for processing information, Psychol. Rev. 63 (1956) 81–97.
[4] H. Hermansky, RASTA processing of speech, IEEE Trans. Speech Audio Proc. 2 (4) (1994) 578–589.
[5] A. Drygajlo, D. Meuwly, A. Alexander, Statistical methods and bayesian interpretation of evidence in Forensic Automatic Speaker Recognition, in: Proceedings of the Eurospeech 2003, Geneva, Switzerland, 2003, pp. 689–692.
[6] I.W. Evett, J.S. Buckleton, Statistical Analysis of STR Data: Advances in Forensic Haemogenetics, Springer-Verlag, Heidelberg, 1996.
[7] C.F. Tippett, V.J. Emerson, M.J. Fereday, F. Lawton, S.M. Lampert, The evidential value of the comparison of paint flakes from sources other than vehicles, J. Forensic Sci. Soc. 8 (1968) 61–65.
[8] A. Schmidt-Nielsen, T.H. Crystal, Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data, Digit. Signal Process. 10 (2000) 249–266.
[9] W.D. Voiers, Perceptual bases of speaker identity, J. Acoust. Soc. Am. 36 (6) (1964) 1065–1073.