Aural and automatic forensic speaker recognition in mismatched conditions

Anil Alexander[1], Damien Dessimoz[2], Filippo Botti[2] and Andrzej Drygajlo[1]

1 Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
2 Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, Switzerland

ABSTRACT In this article, we compare aural and automatic speaker recognition in the context of forensic analyses, using a Bayesian framework for the interpretation of evidence.We use perceptual tests performed by non-experts and compare their performance with that of an automatic speaker recognition system. These experiments are performed with 90 phonetically untrained subjects. Several forensic cases were simulated, using the Polyphone IPSC-02 database, varying in linguistic content and technical conditions of recording. We estimate the strength of evidence for both humans and the baseline automatic system, calculating likelihood ratios using perceptual scores for humans and log-likelihood scores for the automatic system. A methodology analogous to the Bayesian interpretation in forensic automatic speaker recognition is applied to the perceptual scores given by humans in order to estimate the strength of evidence. 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 fixed telephone, cellular telephone and noisy speech in forensically realistic conditions. The perceptual cues that the human subjects use to perceive differences in voices are studied, along with their importance in different recording conditions. We observe that while automatic speaker recognition shows higher accuracy in matched conditions of training and testing, its performance degrades significantly in mismatched conditions. Aural recognition accuracy is also observed to degrade from matched conditions to mismatched conditions and in mismatched conditions, the baseline automatic systems showed comparable or slightly degraded performance compared to the aural recognition systems. The baseline automatic system with adaptation to noisy conditions showed comparable or better performance than aural recognition. The higher level perceptual cues used by human listeners in order to recognise speakers are discussed. We also discuss the possibility of increasing the accuracy of automatic systems using the perceptual cues that remain robust to mismatched recording conditions.

KEYWORDS Aural speaker recognition, automatic speaker recognition, strength of evidence, mismatched recording conditions

INTRODUCTION

Speaker recognition is the process of recognizing unknown speakers from samples of their voices. Human beings use aural, linguistic and other background knowledge in order to perform this recognition. In the forensic context (often in crimes such as kidnappings, rape, burglary, etc.), it is sometimes required that listeners identify the voice that they heard

ISSN (print): 1748-8885 ISSN (online): 1748-8893

(Kerstholt *et al.* 2003). Speaker recognition in forensic cases has often to be accomplished in difficult conditions where distortions could be introduced either due to the transmission (telephone channels, ambient noise, etc.), system (such as the telephone used, the recording instrument, etc.) or the speaker (disguise, stress, emotion, etc.).

Forensic speaker recognition involves the comparison of recordings of an unknown voice (questioned recording) with one or more recordings of a known voice (voice of the suspected speaker). In a forensic context, the degree of similarity between a questioned recording and a suspected speaker's voice is called the evidence (*E*). In a Bayesian interpretation framework, this similarity is evaluated with respect to two competing hypotheses, i.e. that the questioned recording (trace) has been produced by a suspected speaker, or that it has been produced by any other person. The forensic expert is concerned with the uncertainty regarding the evidence, and not concerned with the guilt or innocence of the suspect. The guilt or innocence of the suspect is the province of the court, which takes into consideration other information and evidence which the expert is not even aware of and makes its decision (Aitken and Taroni 2004). In typical forensic cases, in which recordings are made by the police (anonymous calls and wiretapping), the conditions cannot be controlled. Mismatch between the two recordings owing to differences in the phone handset, in the transmission channel and background noise can affect the estimation of the strength of evidence. Currently, in forensic speaker recognition, aural, automatic and a combination of aural and instrumental methods are used in order to estimate the similarity between voices (Gfroerer 2003, Künzel and Gonzalez-Rodriguez 2003).

Bayesian interpretation methods provide a statistical–probabilistic evaluation, which gives the court an indication of the strength of the evidence (likelihood ratio), given the estimated within-speaker (within-source) variability of the suspected speaker's voice and the between-speakers (between-sources) variability of the questioned recording given a relevant potential population (Drygajlo *et al.* 2003). Recently, there have been suggestions by forensic phoneticians to express the outcome of the aural and instrumental phonetic approaches as a Bayesian likelihood ratio. They have further qualified it as the logically correct way of quantifying the strength of identification evidence, and suggested that it should constitute the conceptual basis for forensic–phonetic comparisons (Rose 2002). Corpus-based methodologies are used in the Bayesian interpretation framework in automatic recognition, using databases to evaluate the within-speaker variability, and between-speakers variability (when it is possible to obtain sufficient recordings of the suspected speaker as well as that of a relevant potential population).

In many cases, only one recording of the suspect is available due to the nature of the investigation, for example, when it is not possible to have additional recordings of the suspect's voice, as it may alert him to the fact that he is being investigated. It is often necessary to perform one-to-one comparisons of the questioned recording and the recordings of the suspect's voice. As a consequence, it is not always possible to evaluate the within-source variability of the suspect with this single recording. However, since this is a recurring problem in forensic speaker recognition, an interpretation framework for evaluating the evidence even in the absence of additional control recordings has been investigated (Botti *et al.* 2004). In the present article, this interpretation framework is considered.

We compare the recognition by laypersons (human subjects without any particular training in phonetics) with a baseline automatic system that is not specially adapted to different recording conditions as well as the same system adapted to noisy conditions. The representation of 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.

MOTIVATION

In forensic casework, the recordings analysed can suffer from mismatch in recording conditions because they are non-contemporaneous, acquired using different recording systems (audiotape, digital recordings, wiretappings, etc.), transmitted through various transmission channels (fixed and cellular telephones), or are recorded in different environmental conditions (clean and noisy). In this work, we analyse the extent to which aural and automatic speaker recognition are adapted to environmental and channel conditions, the relative performance of aural and automatic recognition in conditions of mismatch and ways to adapt for mismatch. We analyse the mismatched recording conditions of the suspect and questioned recordings and their effect on forensic human and automatic speaker recognition. These experiments were done in order to compare and contrast the speaker recognition capabilities of ordinary untrained subjects with those of an automatic speaker recognition system.

The automatic speaker recognition system relies on its feature extraction and statistical modelling algorithm, just as the subjects rely on their experience in extracting features specific to the given speaker and build and memorize a 'model' characterizing the speaker's voice. Thus, the experience and ability of the human brain to extract speaker-specific information and its memory of a particular voice can be compared to the automatic system's feature extraction and statistical modelling algorithms. We extend this analogy further by allowing the subjects to listen to the recordings as many times as they would like, before converging to their memorized model of the identity of the speaker, just as the statistical modelling algorithm is allowed to converge to a statistical model of the identity of the speaker after several iterations.

EXPERIMENTAL FRAMEWORK

In order to simulate forensic case conditions, it was necessary to select a test database from which mock cases could be created. In this study, the Polyphone IPSC-02, which is a forensic speaker recognition database, was chosen. This database was recorded by the Institut de Police Scientifique, Ecole des Sciences Criminelles, University of Lausanne, and the Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL). It contains speech from ten speakers in three different recording conditions and two languages (French and German). The three recording conditions include transmission through a PSTN (Public Switched Telephone Network) and a GSM (Global System for Mobile communications) network, and recordings made using an analogue tape-recorder answering machine as well as a digital recorder.

In our experiments, we have used a subset of five speakers from 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 (SNR) ratio of 10 dB. Five short segments of spontaneous speech (between 10 and 20 seconds) 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 seconds) for the mock suspected speaker. Some studies indicate that using languages different from the listener's mother tongue influence the accuracy of recognition (Yarmey 1995) and hence all the test recordings used were in French (the mother tongue of all the listeners). The mock questioned recordings chosen were simulated forensic cases with undisguised imitations of hoaxes and threats. The test database, from which the mock questioned recording and the mock suspected speaker's voice for each comparison had been selected, consisted of speakers who were of the same geographical region (the French-speaking part of Switzerland), and were all male university students. However, care was taken to ascertain that none of these test speakers were familiar to the 90 subjects who participated in this test.

In phonetic terms, this recognition is naive unfamiliar speaker recognition, where the listeners are tested on their ability to recognize a voice with minimal prior exposure. It has been reported that familiar speaker recognition, where the listener is familiar with the voice of the speaker because they have been exposed to it a number of times, shows significantly lower error rates than unfamiliar naive recognition, often with half the percentage of errors (Rose 2002).

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. Pairs of recordings that simulated a mock case, namely a suspect reference recording and a questioned recording, were presented to each listener using a computer program. Listeners were asked to input their responses directly into the user interface of the program, and none of them had any formal training in phonetics. 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 seven-level score scale was chosen based on studies that suggest that it is difficult for humans to differentiate accurately between more than seven levels of comparison for a certain stimulus (Miller 1956). These scores and their verbal equivalents are presented in Table 1.

The subjects were asked 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 memorization (by which the subjects would remember the voices heard in previous comparisons, and use this knowledge in order to make their decision). All the tests in these experiments were performed with a single computer (the same sound card), using exactly the same set of headphones. At the end of the recognition session, the subjects were also asked to list the factors that they considered important when making their decisions. Each test session corresponded to approximately one hour, and a total of 90 hours was spent performing the aural comparisons.

Automatic speaker-recognition task

In the automatic system, feature extraction was performed using the RASTA-PLP (Hermansky 1994) technique and statistical modelling of these features performed using a Gaussian Mixture Modelling (GMM)-

Table 1 Perceptual scores and their verbal equivalents

Score	Verbal equivalent
1	I am certain that the two speakers are not the same
2	I am almost certain that the two speakers are not the same
3	It is possible that the two speakers are not the same
4	I am unable to decide
5	It is possible that the two speakers are the same
6	I am almost certain that the two speakers are the same
7	I am certain that the two speakers are the same

based classifier with 32 Gaussian mixture components (Reynolds and Rose 1995). These features are a compact representation of the short-term spectral envelope. The zones where speech was not present were removed in a pre-processing phase, in order to avoid including non-speech information in the statistical models of the speakers' voices and in the questioned recordings. 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 in 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. This score is a quantitative measure of similarity between the features of the test recording and the statistical model of the speaker and can take any real value between - ∞ and + ∞ (a continuous scale). These scores were normalized per training utterance to a mean of zero and a unit standard deviation so that the scores obtained across speakers have the same range of values. The total computation time taken to perform feature extraction, training and testing for all the comparisons was approximately three hours.

From the human subjects as well as the automatic speaker-recognition system, a set of scores is obtained as results for the mock cases. The scores from the human subjects are discrete, and range from 1 to 7, while the automatic system returns scores that are on a continuous scale. In order to be able to interpret the similarity scores (*E*) in a Bayesian interpretation framework, it is necessary to compare them with respect to the two competing hypotheses H_0 – the suspected speaker is the source of the questioned recording and H_1 – the speaker at the origin of the questioned recording is not the suspected speaker.

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 questioned recording (Drygajlo *et al.* 2003). 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 test 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 the two recordings have the same source (H_0) and that the two recordings have a different source (H_1). The evaluation of the likelihood ratio of the evidence (*E*), allows us to calculate the degree of support for one hypothesis against the other:

$$LR = \frac{p(E|H_0)}{p(E|H_1)} . \quad (1)$$

Note that in forensic automatic speaker recognition, the expert should be able to evaluate whether the tools he uses perform well with the recordings, whether incompatibilities between the databases that he uses can affect the estimation of the strength of evidence and whether compensations can be performed to reduce the effects of such incompatibilities (Alexander *et al.* 2004b).

Here, it is necessary to calculate the strength of the evidence using the continuous and the discrete scales in automatic and aural recognition, respectively.

The likelihood ratio, as illustrated in Figure 1, is the ratio of the heights of the distributions of scores for hypotheses 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 Bayesian interpretation (Figure 2). Since the scores obtained in this perceptual scale are discrete scores (from 1 to 7) and not continuous as in the case of log-likelihood scores returned by the automatic system, a histogram approximation of the probability density can be calculated. The histograms corresponding to each of the hypotheses are obtained from the perceptual scores corresponding to the mock cases where H_0 and H_1 are true, by calculating the frequency of appearance of these values (from 1 to 7) for all the scores on the perceptual scale. The frequency of appearance is calculated in the following way. Each score (from one to seven) is considered as an interval, and the number of times a certain score is obtained when one of the hypotheses was known to be true is used to calculate the frequency of appearance of that score. The area corresponding to each of the histograms is then normalized so that the total area under each of the histograms is unity by dividing it with the total number of scores per hypothesis. The likelihood ratio would then be the relative heights, on the histogram of the two hypotheses, at the point *E* (shown in Figure 2).

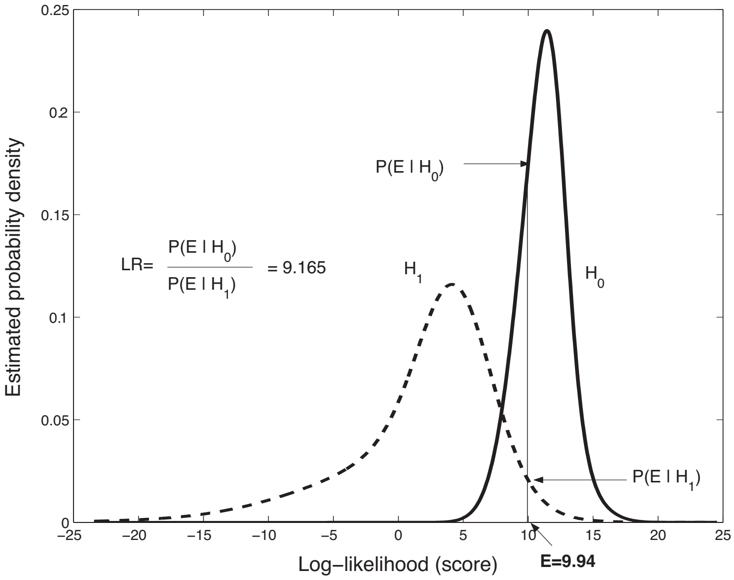

Figure 1 Estimation of likelihood ratio using automatic speaker-recognition scores

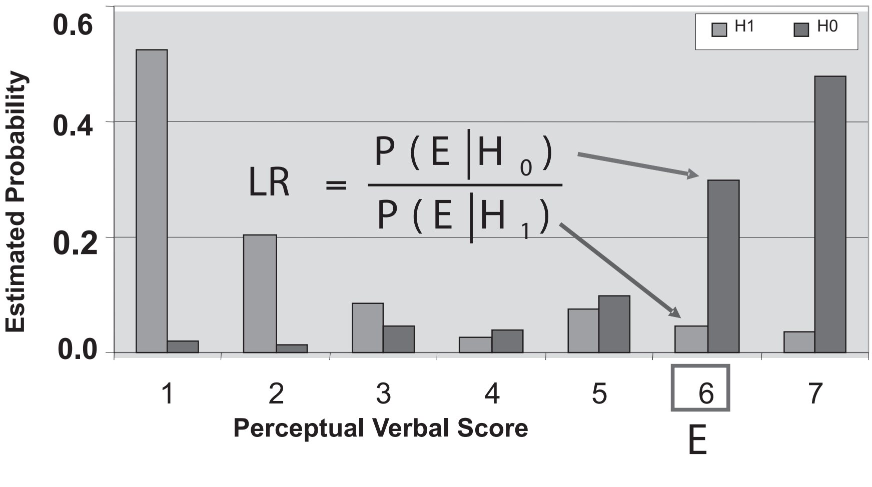

Figure 2 Estimation of likelihood ratio using aural speaker-recognition scores

COMPARING THE STRENGTH OF EVIDENCE IN AURAL AND AUTOMATIC METHODS

The strength of evidence can be evaluated by estimating and comparing the likelihood ratios that are obtained for the evidence E in mock cases, where the hypothesis H_0 is true and when the hypothesis H_1 is true. By creating mock cases which correspond to each of these hypotheses and calculating the *LR*s obtained for each of them, the performance of the speaker-recognition system can be evaluated. In this way, we get two distributions; one for the hypothesis H_0 and the other for the hypothesis H_1. With these two distributions, it is possible to find the significance of a given value of *LR* that we obtain for a case, with respect to each of these distributions.

In order to measure the performance of each of the speaker-recognition methods, the cases described in the previous section are considered, separating them into those 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 cumulative probability distribution plots called Tippett plots, obtained by integration of probability distributions of *LR*s. The Tippett plot represents the proportion of the likelihood ratios greater than a given *LR*, i.e. $P(LR(H_i) > LR)$, for cases corresponding to the hypotheses H_0 and H_1. The separation between the two curves in this representation is an indication of the performance of the system or method in differentiating between voices, with a larger separation implying better performance than a smaller one. This method of representation of performances was proposed by Evett and Buckleton (1996) in the field of interpretation of the forensic DNA analysis. This representation has been named the Tippett plot, referring to the concepts of within-source comparison and between-sources comparison defined by Tippett et al. (1968).

Comparison of aural and automatic recognition in matched and mismatched conditions in terms of Bayesian interpretation of evidence

The experimental results are represented using the Tippett plots $P(LR(H_i) > LR)$ (Figures 3 and 4), which can be used to represent how many cases are above a given value of likelihood ratio with respect to each hypothesis H_0 or H_1, to indicate to the court how strongly a given likelihood ratio can represent either of the hypotheses.

In Figure 3, the aural and automatic likelihood ratios are presented for matched recording 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 between the two curves. The curves of the automatic system are slightly more separated than those of

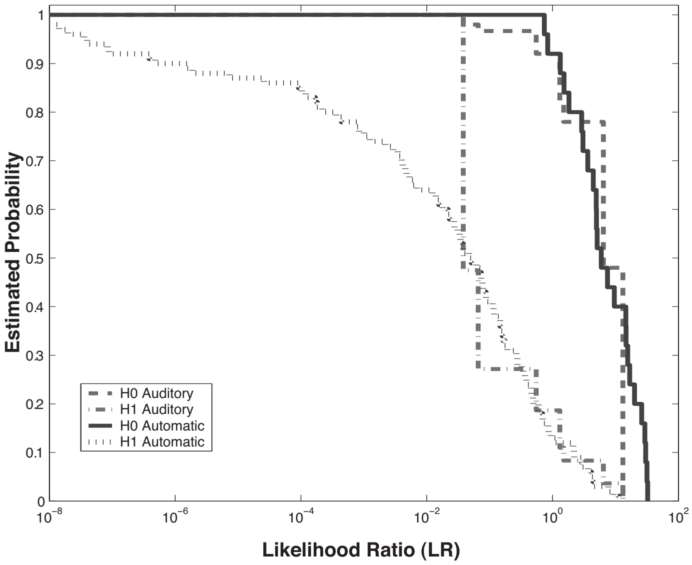

Figure 3 Tippett plot in matched condition (PSTN–PSTN) for aural and automatic recognition

the aural recognition. Here, we can observe, that in matched conditions, the automatic recognition performed better than the aural recognition.

However, using the Tippett plots for both the aural and automatic systems, we observe that in mismatched conditions the likelihood ratios for the H_0 and H_1 hypotheses are not as well separated as in the matched case. In Figure 4, likelihood ratios of aural and automatic recognition are presented for mismatched conditions, i.e. when the suspected speaker's speech was recorded in PSTN transmission conditions and the questioned recording was in noisy-PSTN conditions. In this plot, we observe that both the aural and automatic systems show degraded performance, with the curves of the aural system showing more separation than those of the automatic recognition. This implies, that in mismatched conditions, the aural recognition performed better than the automatic recognition using the baseline system.

In Figure 5, the aural likelihood ratios and the likelihood ratios of the automatic system adapted to mismatched conditions, i.e. when the suspected speaker's speech was in PSTN transmission conditions and the questioned recording in noisy-PSTN conditions are presented. The speech enhancement applied in the automatic system was using an algorithm for spectral subtraction, in order to reduce the effects of the noise (Martin 1994). This method helped boost the performance of the automatic

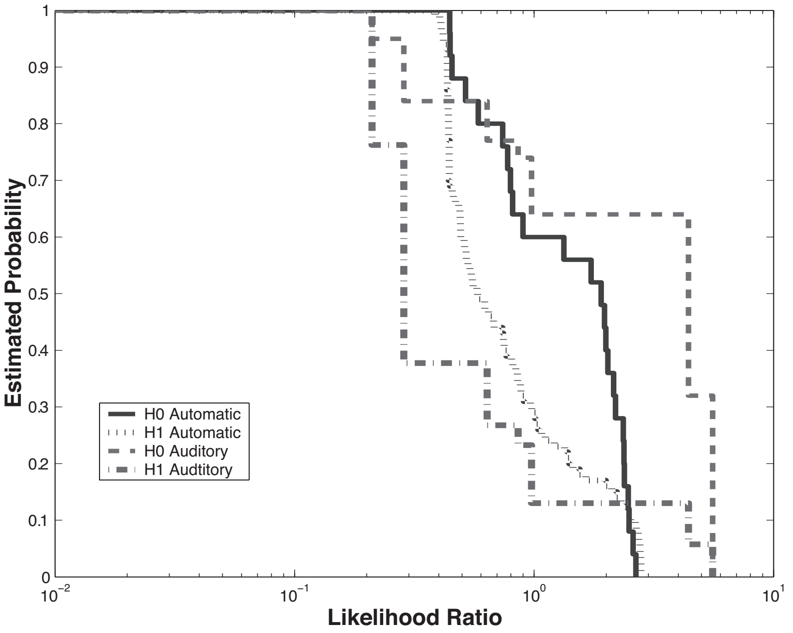

Figure 4 Tippett plot in mismatched conditions (PSTN–Noisy PSTN) for aural and automatic recognition

system when it was applied at the pre-processing phase of the recognition. Here, we observe a better separation between the curves corresponding to the automatic recognition. This performance is similar to the performance of the aural recognition.

Comparison of aural and automatic recognition in matched and mismatched conditions in terms of Bayesian decision theory

In order to compare the performance of automatic speaker verification systems, often, Receiver Operator Curves (ROC) or Detection Error Tradeoff (DET) curves are used (Martin *et al.* 1997). The DET curve plots the relative evolution of False Match Rate (FMR) and False Non Match Rate (FNMR) when using a decision point. The Equal Error Rate (EER) is the point at which the FMR is equal to the FNMR and is used in order to compare the performance of automatic speaker-verification systems. The closer the curve is to the origin (0, 0), the better the performance of the system. The equal error rate can be observed on the DET curve as the point of intersection of the DET curve and the diagonal line between the two axes. Although the forensic speaker-recognition task does not use a threshold, as in the speaker-verification approach, it is informative to compare measures existing in the speaker-verification domain, to mea-

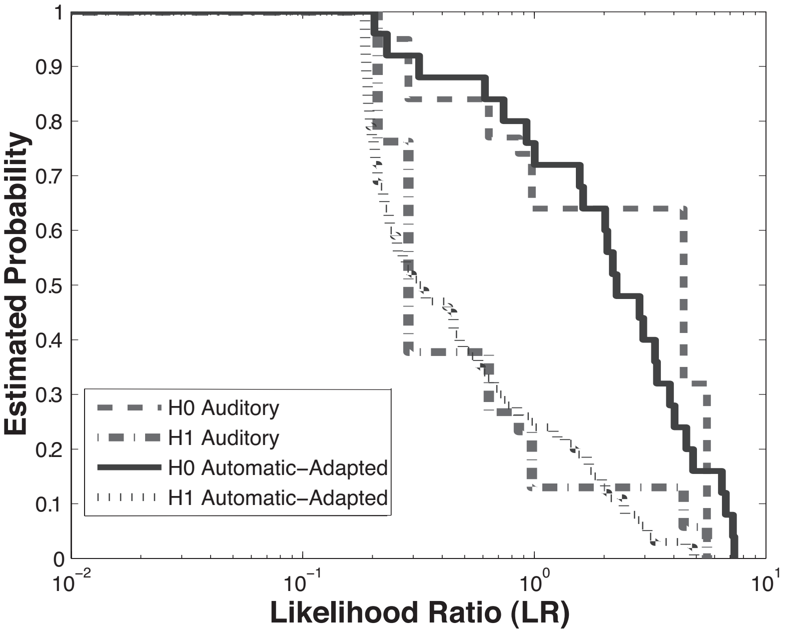

Figure 5 Tippett plot in adapted mismatched conditions (PSTN–Noisy PSTN) for aural and automatic recognition

sures used in forensic speaker recognition in order to ascertain whether the same trends can be observed.

Note that the DET curve for aural recognition is not as smooth as that of the automatic recognition. This is because of the limited number of discrete perceptual verbal scores (from 1 to 7), and not because of the number of comparisons used to plot the DET curves.

Matched conditions
In Figures 6 and 7, we observe the relative performance of aural and automatic recognition in matched conditions (PSTN–PSTN) and (GSM–GSM), respectively. Automatic speaker recognition is seen to outperform aural speaker recognition in matched conditions with EERs as low as 12 per cent and 4 per cent for PSTN–PSTN and GSM–GSM comparisons, respectively.

Mismatched conditions
In Figures 8 and 9, we observe the relative performance of aural and automatic recognition in mismatched conditions, PSTN–GSM and PSTN–PSTN Noisy, respectively. In noisy conditions, human aural recognition is better than the baseline automatic system not adapted to noise. However, in mismatched telephone channel conditions, the automatic

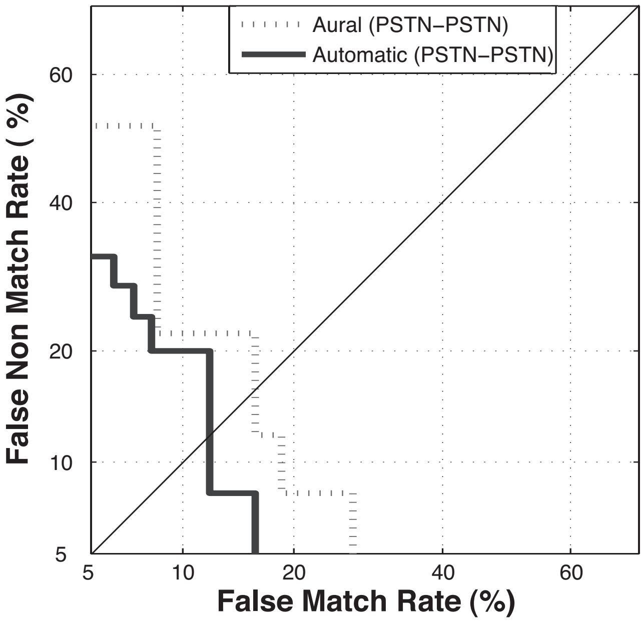

Figure 6 DET plot for comparison between the aural and the automatic recognition (PSTN–PSTN)

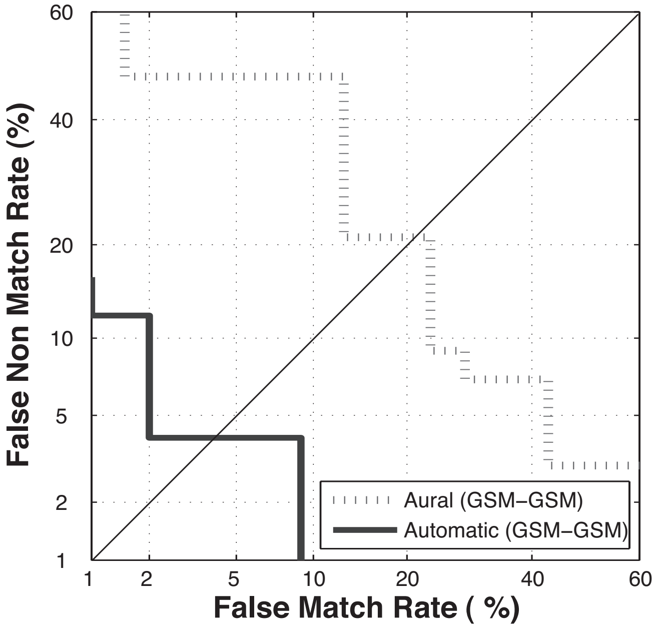

Figure 7 DET plot for comparison between the aural and the automatic recognition (GSM–GSM)

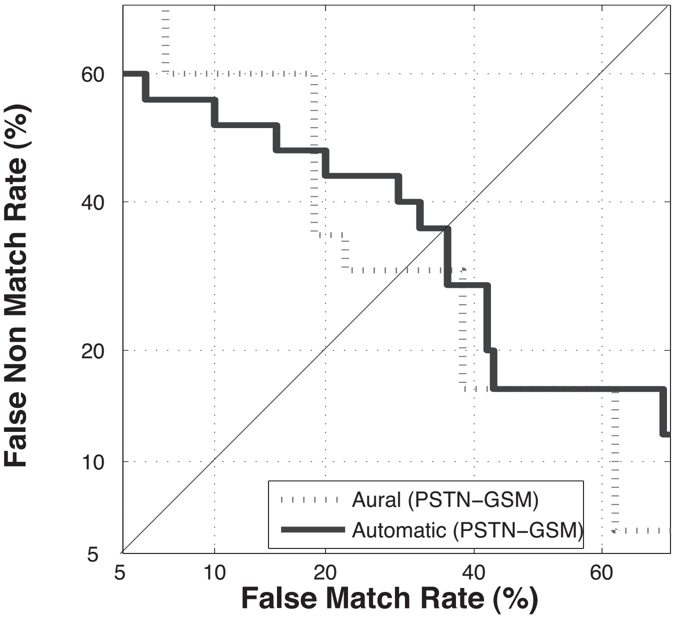

Figure 8 DET plot for comparison between the aural and the automatic recognition (PSTN–GSM)

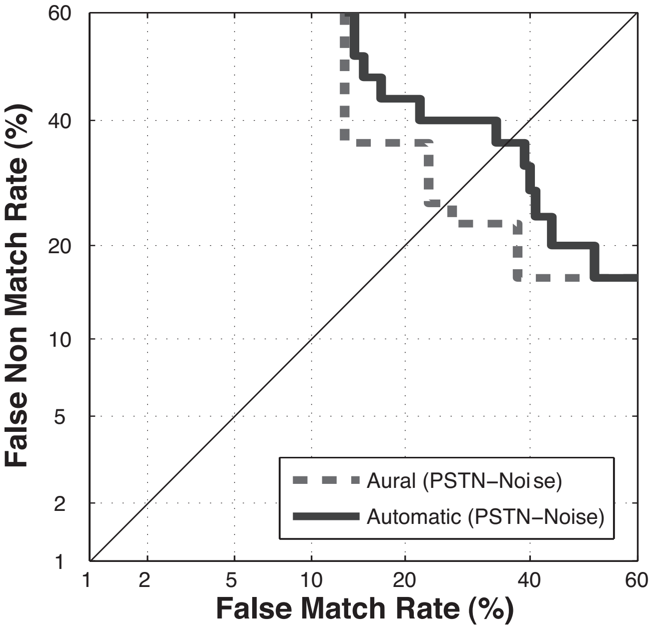

Figure 9 DET plot for comparison between the aural and the automatic recognition (PSTN–Noisy PSTN)

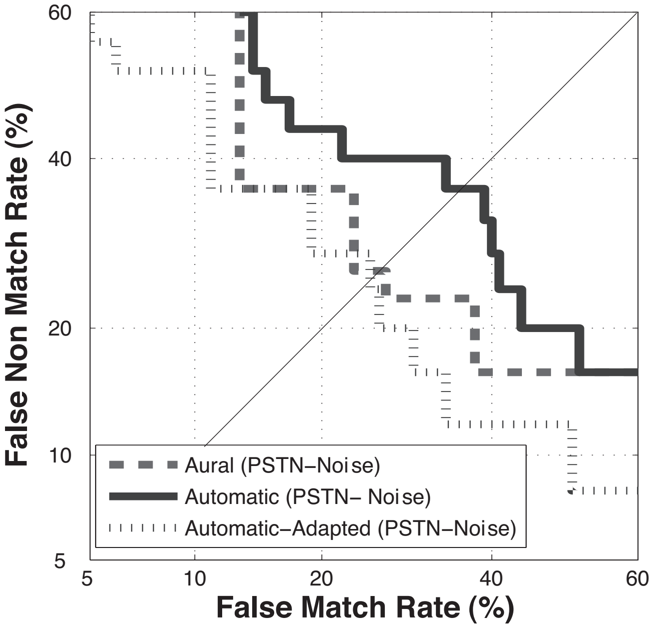

Figure 10 DET plot for comparison between the aural and the automatic recognition (PSTN–Adapted Noisy PSTN)

speaker-recognition system shows accuracies comparable to aural recognition. In general, both aural and automatic recognition perform better in matched conditions than in conditions of mismatch. It is thus important that the baseline automatic system is adapted for the changes in conditions.

In Figure 10, the DET curves after adaptation for noisy conditions are shown. Note that a reduction in the error rates is observed in the DET curves (Figure 10) after this adaptation is performed. This shows a similar trend as in the Tippett plot (Figure 5) where the separation between the curves for automatic recognition was increased after applying adaptation for mismatch.

DISCUSSION

From the previous section we conclude that aural recognition can outperform baseline automatic recognition in mismatched recording conditions. To a large extent, this can be attributed to the human auditory capability to adapt to different conditions, and to use extra information, that is not explicitly modelled in the automatic system. Apart from the speech signal, human recognition relies on other external cues that are difficult to model by automatic speaker-recognition methods.

For instance, when humans recognize other speakers, they include high-level speech features about the identity of the speaker such as mannerisms in their speech, certain anomalous words that they use, their accent, pauses, speaking rate, etc. For example, during telephone calls, although the voice heard at the receiver of the listener may be very different from the original voice of the speaker, most people are able to identify the caller quite easily. Often, listeners are able to make the transition from a rough idea of the identity of the speaker to pinpointing exactly who he or she is simply by paying attention to the high-level data such as the vocabulary chosen, the pronunciation of certain words and using other background information that one may have about the caller. These features are largely independent of the conditions of recording. Since these cues are largely relied upon, normally the drop in accuracy of aural speaker recognition, when there is a change in conditions, remains small. However, for the automatic system, which relies mainly on the low-level speech information, and does not incorporate higher-level information, the changes in conditions may make a large difference in their accuracy.

PERCEPTUAL CUES USED BY LAYPERSONS

Human beings depend largely on perceptual cues in the speech, such as pronunciation, word choice and speech anomalies (Nolan 1983, Schmidt-Nielsen and Crystal 2000) to recognize speakers. There have been studies which attempt to quantify these various aural and perceptual means that laypersons use in order to identify speakers (Voiers 1964).

In our study, we were able to identify the factors important to each of the subjects in order to recognize speakers, i.e. the accent, timbre, intonation, rate of speech, speech defects or anomalies, breathing, loudness, similarity to known voices and their intuition. These criteria were obtained by asking the subjects, at the end of each experiment session, what factors they considered in recognizing the source of the questioned recording. The subjects were allowed to describe the characteristics which they thought they used in order to recognize speakers in their own words, and no prompts were given. These perceived characteristics were compiled and grouped. Since none of the subjects had any training in phonetics, the words were interpreted and categorized by the authors.

In Table 2 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, intonation, timbre, rate of speech and speech anomalies. We have tried to identify these perceptual cues used by the subjects using our interpretation of their responses, in order to understand which of these cues can be incorporated into the automatic system in order to improve its performance.

Table 2 Relative importance of perceptual cues

Criteria	**PSTN–PSTN** (%)	**PSTN–GSM** (%)	**PSTN–Noise** (%)	**GSM–GSM** (%)
Accent	34	31	30	26
Timbre	25	25	22	22
Intonation	16	24	18	21
Rate of speech	9	7	12	12
Speech anomalies (defects)	6	7	8	5
Breathing	5	0	2	2
Volume	3	0	0	0
Imagined physiognomy	3	0	2	2
Similarity to known voices	0	2	0	2
Intuition	0	4	6	9

- *Accents*: Accents can vary widely regionally as well as socially. In this study, a subset of the Polyphone IPSC-02 database was chosen, so that the speakers would have a Swiss French accent, although within this population subtle regional variations exist. All the test speakers had similar social and educational background. In spite of these constraints, the accents were the most important feature that the subjects claimed to use when recognizing the speakers. Accents are difficult to model explicitly in automatic systems, although it is possible to build background statistical models that represent a particular accent or dialect, and use it to normalize the automatic recognition scores.
- *Intonation*: Speakers use intonation to express syntactic change by modifying their pitch (Rose 2002), as well as to signal information about the emotional state of the speaker (e.g. anger, boredom, depression, etc.). Some of the subjects believed they used intonation as a criterion in comparing the recordings belonging to different speakers. Intonation may include differences in pitch level (e.g. high and low voices) and in pitch movement (e.g. monotonous and melodic voices). Pitch has been successfully incorporated in automatic speaker recognition (Arcienega and Drygajlo 2003). It is extremely difficult to incorporate intonation information indicative of the abstract emotional state of the speaker in the automatic system, although the use of suprasegmental features like pitch shows some promise.
- *Rate of speech*: Speakers differ from each other in terms of the rate at which they speak and other speech durational factors (pauses, time taken for a word or phrase, etc.), and this is useful information in ascertaining the identity of a speaker. There have been studies in forensic phonetics in which temporal information about the speech and their use as forensic parameters for speaker recognition were considered

(Hollien 1990, Künzel 1997). In some of our preliminary experiments modelling the normalized rate of speech, we concluded that by normalizing the rate of speech and using information about the transition from silence to speech zones, it is indeed possible to increase the accuracy of automatic recognition. However, it is necessary to have sufficient amounts of recordings to derive these statistics. This is often not the case when we consider forensic cases with short, questioned recordings. In the tests performed, the questioned recordings were too short to derive any meaningful statistics about the rate of speech.

- *Timbre*: Timbre is the subjective correlate of all those sound properties that do not directly influence pitch or loudness. These properties include static and dynamic aspects of the voice quality (laryngeal or supralaryngeal), vowel quality (formant structure) and the temporal evolution of the speech spectral power distribution. The subjects believed that this was a factor they considered in judging the identity of the speaker. Although arguably this information is present in the low-level speech information, the timbre is not explicitly quantified as a parameter in automatic speaker recognition.
- *Speech anomalies*: Because of the rule-based nature of language, speakers of a language can usually discern realizations of speech that deviate from 'normal speech' well (Rose 2002). While certain deviations of speech may be reasonable or acceptable in one language, they may not be acceptable in other languages. These speech- and voice-pathological deviations are often used as clues to the identity of a person in aural recognition. Speech anomalies are difficult to model in an automatic system as these are often linguistic in nature. Baseline text-independent speaker recognition, which is used in the automatic recognition system, does not allow us to take such information into consideration.

Relative importance of the perceptual cues

The relative importance of each of these main characteristics remains very similar across different recording and environmental conditions, implying that human perception of speaker identity mainly depends on characteristics that are robust to conditions. This is in stark contrast to the baseline 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. The human auditory system is able to adapt to the effects of masking by noise and other distortions (Bregman 1990). Consequently, in the automatic system, it is necessary to explicitly adapt the recognition process to each of the conditions.

CONCLUSIONS

Perceptual speaker-recognition 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 suspect and questioned recordings, the automatic systems showed 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. Thus, the baseline automatic speaker-recognition system should be adapted to each of the mismatched conditions in order to increase its accuracy, as was observed with adaptation to noisy conditions. The adapted system shows comparable or better performance than aural recognition in the same conditions. The perceptual cues that human listeners rely upon in order to identify speakers were analysed. The accuracy of automatic systems can be increased using these perceptual cues that remain robust to mismatched conditions.

The scheme of comparison, presented in this article, can also be adapted in order to analyse the performance of the aural perception of persons with phonetic training. It should be stressed that in its current framework, this comparison relies only on the auditory abilities of the listener, and does not take into account any other instrumental techniques. Thus, the future study would essentially be comparing the automatic system with the auditory recognition abilities of the phonetician. It would evaluate the performance of phoneticians and automatic systems, in different conditions, and would indicate methods of combining the aural perceptive approach of trained subjects with that of the automatic system, depending on the conditions of the case.

ACKNOWLEDGEMENTS

The authors would like to thank the 90 subjects for their participation in this study as well as Quentin Rossy for his help in creation of the Polyphone IPSC-02 database which has been used in the experiments in this article. The authors would also like to thank the reviewers and the editors for their valuable comments and suggestions.

NOTE

An earlier, shorter version of this article appeared in *Forensic Science International* (Alexander *et al.* 2004a).

REFERENCES

Aitken, C. and Taroni, F. (2004) *Statistics and the Evaluation of Evidence for Forensic Scientists*, Chichester: Wiley.

Alexander, A., Botti, F., Dessimoz, D. and Drygajlo, A. (2004a) 'The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications', *Forensic Science International*, 146 (Supplement 1): S95–S99.

Alexander, A., Botti, F. and Drygajlo, A. (2004b) 'Handling mismatch in corpus-based forensic speaker recognition', *Proceedings of 2004: A Speaker Odyssey*, Toledo, Spain, 69–74.

Arcienega, M., and Drygajlo, A. (2003) 'A Bayesian network approach for combining pitch and reliable spectral envelope features for robust speaker verification', in J. Kittler and M. S. Nixon (eds), *Proceedings of 4th International Conference on Audio- and Video-Based Biometric Person Authentication*, Guildford: Springer, 78–85.

Botti, F., Alexander, A. and Drygajlo, A. (2004) 'An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data', *Proceedings of 2004: A Speaker* Odyssey, Toledo, Spain, 63–8.

Bregman, A. (1990) *Auditory Scene Analysis*, Cambridge, MA: MIT Press.

Drygajlo, A., Meuwly, D. and Alexander, A. (2003) 'Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition', *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 689–92.

Evett, I. W. and Buckleton, J. S. (1996) 'Statistical analysis of STR data', in Carracedo, A, Brinkmann, B. and Bär, W. (eds) *Advances in Forensic Haemogenetics* Vol. 6, Berlin: Springer Verlag, 79-86.

Gfroerer, S. (2003) 'Auditory instrumental forensic speaker recognition', *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 705–8.

Hermansky, H. (1994) 'RASTA processing of speech', *IEEE Transactions on Speech and Audio Processing*, 2(4): 578–89.

Hollien, H. (1990) *The Acoustics of Crime, The New Science of Forensic Phonetics*, New York: Plenum.

Kerstholt, J., Jansen, E., Amelsvoort, A. van and Broeders, A. (2003) 'Ear-witness line-ups: effects of speech duration, retention interval and acoustic environment on identification accuracy', *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 709–12.

Künzel, H. (1997) 'Some general phonetic and forensic aspects of speaking tempo', *Forensic Linguistics*, 4: 48–83.

Künzel, H. and Gonzalez-Rodriguez, J. (2003) 'Combining automatic and phonetic-acoustic speaker recognition techniques for forensic applications', in *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, 1619–22.

Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M. (1997) 'The DET curve in assessment of detection task performance,' *Proceedings of Eurospeech '97*, Rhodes, Greece, 1895–8.

Martin, R. (1994) 'Spectral subtraction based on minimum statistics,' *EUSIPCO-94*, 1182–5.

Miller, G. A. (1956): 'The magical number seven, plus or minus two: some limits in our capacity for processing information', *Psychological Review*, 63: 81–97.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*, Cambridge: Cambridge University Press.

Reynolds, D. A. and Rose, R. C. (1995) 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE Transactions on Speech and Audio Processing*, 3(1): 72–83.

Rose, P. (2002): *Forensic Speaker Identification*, Forensic Science Series. Taylor & Francis.

Schmidt-Nielsen, A., and Crystal, T. H. (2000) 'Speaker verification by human listeners: experiments comparing human and machine performance using the NIST 1998 speaker evaluation data', *Digital Signal Processing*, 10: 249–66.

Tippett, C. F., Emerson, V., Fereday, M., Lawton, F. and Lampert, S. (1968) 'The evidential value of the comparison of paint flakes from sources other than vehicles', *Journal of the Forensic Science Society*, 8: 61–5.

Voiers, W. D. (1964) 'Perceptual bases of speaker identity', *Journal of the Acoustic Society of America*, 36(6): 1065–73.

Yarmey, A. D. (1995) 'Earwitness speaker identification', *Psychology, Public Policy, and Law*, 1(4): 792–816.