On decision making in forensic casework

Johan Koolwaaij and Lou Boves

*Automatic Acoustic Recognition Technologies (*A^{2}RT *), University of Nijmegen*

ABSTRACT In forensic applications of speaker recognition it is necessary to be able to specify a confidence level for a decision that two sets of recordings have been produced by the same speaker (or by different speakers). Forensic phoneticians are sometimes criticized because they find it impossible to provide 'hard' estimates of the confidence level of their expert opinions. This paper investigates to what extent the problem can be solved by deploying automatic speaker verification algorithms, to work alone or to support the work of forensic phoneticians.

It is shown that, although heavily dependent on operating conditions, one of the advantages of automatic systems is that their performance is in fact measurable. We construct a confidence measure which takes into account the past performance of the automatic system, the operating conditions and the probative value of the speech evidence, as well as the non-speech evidence. It is very important to note that such a confidence measure will never lead to a fully automatic procedure, since it still requires human input to weigh the non-speech evidence as well as human explanation of the procedure followed, and, finally, human interpretation. However, when all conditions are met, this procedure is able to (1) provide an interpretative measure in the individual forensic case and (2) join together the strengths of the human interpretation of the non-speech evidence and the automatic interpretation of the speech evidence, so that finally the joint performance of human and machine is better than the performance of one of them in isolation.

KEYWORDS speaker verification, Bayesian theory, decision making

INTRODUCTION

Automatic Speaker Verification (SV) and forensic casework have long been regarded as essentially unrelated disciplines, because the former was seen as a one-alternative forced choice problem, whereas the latter used to be presented as an open set identification problem. However, Broeders (1995) and Doddington (1998) have pointed out that many forensic cases boil down to the question of whether a set of recordings, some of which are definitely from the perpetrator and others from a single suspect, do or do not originate from the same speaker. In other words, many forensic cases can be formulated as a one-alternative forced choice problem.

One broad class of cases where the 'forced choice' paradigm applies, and where automatic SV techniques might prove to be useful in forensic work, is in the processing of telephone taps that are made in the investigation of drug trafficking cases. Very often, the perpetrators are foreigners, who speak a language unknown to the police officers, and also to the

© University of Birmingham Press 1999
1350–1771

Forensic Linguistics 6(2) 1999

forensic phoneticians. In many cases the police are interested in knowing how many different speakers are involved in a given set of telephone taps. Leaving the speaker recognition task to interpreters has been shown to be problematic, if only because of possible links between the interpreters and the criminals. Such links are to be expected if the case is investigated in a small language community. In these cases a text-independent SV system might be of great help.

In forensic applications of speaker recognition it is important that one is able to state a confidence level for conclusions regarding the similarity between the voices of a known suspect and an unknown perpetrator. During police investigations confidence levels will be used to weigh the evidence in setting priorities for investigating specific suspects. Forensic phoneticians often face the reproach that they cannot provide a 'hard' estimate of the confidence attached to a decision in a specific case. Often these reproaches come from people who think that machines, specifically automated speaker recognition systems, are superior to forensic phoneticians, because these machines can provide confidence scores. In reality, this is not true. In this paper we first explain why it is not straightforward to come up with a confidence score for a decision with respect to the (lack of) similarity between two speakers. First, we argue that the concept of 'confidence' in speaker recognition is not easy to define in a statistically tractable way. One obvious operationalization, namely posterior probability, appears to be self-defeating because it requires knowledge about the a priori probability that the decision is correct. We develop the arguments for an automatic speaker recognition system, but similar arguments would apply to the work of forensic phoneticians. Then we proceed to show how potentially uncertain evidence coming from a human or automatic speaker recognition system can be exploited in a Bayesian decision approach. To this end we propose a new measure, based on the false accept and false reject rates observed in the past for an automatic system or a forensic phonetician. We have carried out a large number of simulations to show under which conditions several fallible pieces of information can add up to a single high-confidence judgement. In this way we can show that unreliable systems can still help each other to reach the correct conclusion as long as the evidence they use is independent.

The data used in this paper derive from two sources. The first source of data comes from our involvement in the campaign for evaluating automatic speaker recognition systems organized by the American National Institute for Standards and Technology (NIST). Participants in the NIST campaign have access to very large amounts of speech recorded over the telephone, which can be used to train and test speaker recognition systems. In the research reported here we have also used these data in the simulation experiments.

The second source of data pertains to a specific case that was brought to our attention by a Dutch private investigations bureau. A male person left obscene messages in the voice mail boxes of female employees of a large IT company. The calls could be traced to handsets in in-house classrooms. Three victims identified the same colleague as the likely perpetrator, but the accused person denied all charges, and agreed to collaborate in a test in which he read transcripts of two of the messages. The speech was recorded in one of the classrooms, using the same handset type and the same voice mail system as during the harassment calls. However, while the harassment calls were whispered, probably with the intent to sound 'sexy', the test calls were read with normal voice. Approximately one month after the test recordings the harassment calls started again, in a whispery voice and from the same classrooms. Now, the obvious question is whether the two sets of harassment calls were made by the same speaker, and whether this speaker is the same person as the one who read the transcripts, which is a typical case of a one-alternative forced choice problem.

This paper is organized as follows: first we describe the text-independent speaker verification system that we used in the NIST campaign and in the harassment case. Next, we explain why the usual performance evaluation measures for speaker recognition systems cannot be used as a confidence measure for individual decisions. In the section on Bayesian decision theory we propose posterior probabilities to assess the confidence in an individual accept/reject decision and we then go on to explain their 'operating instructions' as well as the set of factors which should be accounted for when using posterior probabilities. We then discuss the impact of the findings for the harassment case. We finish with a discussion of the way in which (automatic) speaker verification can be of added value in forensic cases, despite potential uncertainties.

AN AUTOMATIC SPEAKER VERIFICATION SYSTEM

In this section we introduce the *A²RT* Automatic Speaker Verification (ASV) system. We will use this system to explain several fundamental aspects of such systems and to demonstrate the impact of a number of operational factors on the scores computed by ASV systems, and therewith on the False Accept Rate (FAR) and False Reject Rate (FRR). The system described here is the text-independent speaker verification system that was built for the 1998 NIST Speaker Recognition Evaluation (Przybocki and Martin 1998). The *A²RT* system appeared to perform reasonably well on the NIST 1998 test data (NIST 1998). Although all results and figures in this paper are based on experiments with the *A²RT* system, the results generalize to all other state-of-the-art ASV systems.

The speech used in the experiments was taken from the Switchboard-2 Phase 1 corpus. Thus, all recordings were made over the US switched

public telephone network, the language used by the speakers was American English, and the speech was conversational (some test samples mainly consist of back channel utterances like *yes*, *erm*, *huhhuh*, etc.).

The *A²RT* system is a text-independent SV system. Since it is intended for use with telephone speech, the signals are sampled with a frequency of 8 kHz. Samples can be either in 8-bit A-law or μ-law format, or in 16-bit linear format. Parameterization is based on 25.6 ms frames, with a 10 ms frame shift. For each frame 12 LPC cepstra and log-energy are computed; the eventual feature vector is formed by appending the deltas and delta-deltas of the thirteen coefficients, making for a total of thirty-nine features. For each client speaker in the NIST data a single model has been trained. These models consist of a mixture of 128 Gaussian distributions; therefore, they are known as Gaussian Mixture Models (GMMs). GMMs do not try to make a kind of phonetic segmentation of the training speech. Thus, there is no interpretable relation between Gaussian distributions in the mixture and specific speech sounds.

All state-of-the-art ASV systems build anti-speaker models in addition to client models. Anti-speaker models can be speaker-specific (in which case they are also referred to as cohort models) or speaker-independent (also called world models). World models, in their turn, can be built for the total population, or for specific sub-populations, like male and female speakers. The anti-models are used to normalize the scores of the ASV system for an individual test sample (Lee 1997). If, due to some distortion caused by the transmission channel or by background noise, a test utterance does not match very well with the model of the true speaker, it is reasonable to expect that this test utterance will also differ from 'speech in general' recorded under undistorted conditions. By relating the likelihood that an unknown test utterance matches with the model of a given speaker to the likelihood that this utterance matches with the world model (or the cohort model) the effect of chance distortions is diminished. Thus, the scores computed by the ASV are Likelihood Ratios. It is customary to present Likelihood Ratios on a logarithmic scale; hence the term Log Likelihood Ratios (LLRs). The LLRs are eventually used to make an accept/reject decision. Anti-speaker models consisting of 128 GMMs have been trained using recordings of a large number of speakers, none of whom is among the 'clients'. Separate anti-speaker models for male and female speakers were built. The anti-speaker model was trained first, starting from scratch. The speaker models were then adapted from the anti-speaker model.

All speech is passed through a silence–speech detector before further processing. For testing, up to 50 000 speech samples with a duration of thirty seconds are available. The proportion of male to female speakers is 1:1, and the proportion of non-target to target trials is 9:1. Obviously, this database contains much more speech from many more speakers than a forensic phonetician could ever hope to process in a realistic experi-

ment. The size of the database allows us to conduct simulation experiments which will provide results with a statistical reliability far beyond what can be obtained in experiments with forensic phoneticians.

EVALUATION MEASURES

Intuitively one might think that stating confidence levels for the decision of an ASV system should be trivial. For all serious ASV products performance figures are available specifying the False Accept Rate (FAR) and the False Reject Rate (FRR), Equal Error Rate (EER), Receiver Operating Characteristic (ROC) curve (Gibbon *et al*. 1997), or Detection Error Tradeoff (DET) curve (Martin *et al*. 1997). Thus, one might expect that a properly built ASV system should be able to produce an objective confidence measure on an absolute scale. If this were true, it would allow the system to be used by virtually every police officer. Unfortunately, the conventional performance measures cannot be used to derive a confidence measure that is appropriate for individual cases. This is because all the measures mentioned above are only valid as averages over large numbers of genuine and impostor attempts. In fact, these measures originate from extensive experiments, in which accept/reject decisions are obtained for large numbers of utterances from true clients and from impostors. The averages computed over all these observations are fundamentally different from individual observations, which can be close to the average, but also far apart. Of course, in forensic work only individual cases matter.

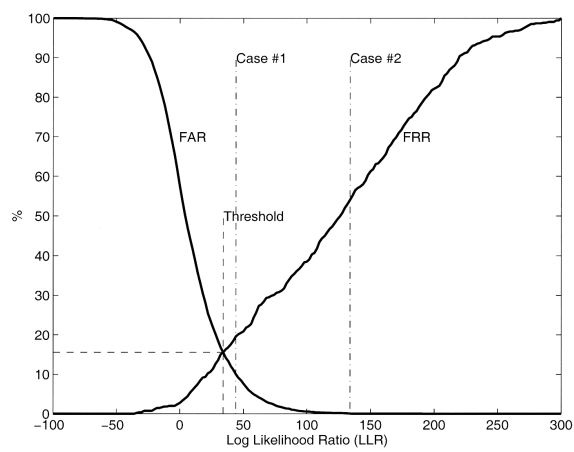

Figure 1 False Accept Rate (FAR) and False Reject Rate (FRR) as a function of the LLR threshold value. The dot-dash vertical lines represent two individual cases with LLR values greater than the equal error rate threshold.

A good way to illustrate why some global measure of the performance of an ASV system is not adequate in forensic work is to look at an example. Figure 1 shows the proportions of false accepts and false rejects of our text-independent SV system as a function of the threshold set in terms of the log-likelihood ratio (LLR) score of test samples. In addition, two individual cases are depicted. Both are in the range of LLR values where the case would probably be accepted as the true speaker (since both are beyond the threshold, which, for this example, is arbitrarily set to the LLR value that corresponds to equal probabilities of false reject and false accept). However, it is obvious that the cases are very different. Even if case #1 has a 'positive' LLR score, it is only marginally so, whereas the LLR score for case #2 makes a false accept very unlikely (but not impossible). Therefore, the confidence that one should attach to an accept/reject decision of this system is certainly different from its EER (or whatever conventional average performance measure is provided by the system manufacturer). If anything, we need a combination of the FAR and FRR (or any other average performance measure) and the LLR assigned to the test utterance, which together give an indication of the confidence of an accept/reject decision. Even for a seemingly mediocre system (with an EER of about 15 per cent) the decision for case #2 seems to be highly reliable.

The example in Figure 1 might suggest that it should be possible to base confidence measures in individual cases on the LLR value proper. This is the more so because the likelihood ratio is introduced to normalize the otherwise unscaled raw likelihood values (Lee 1997). One might be tempted to assume that likelihood ratio scores are measures on a ratio scale; unfortunately, in actual practice, LLRs are measures on an ordinal scale (Stevens 1951).

There are two reasons why the LLRs produced by a speaker verification system must be interpreted as measurements on an ordinal scale:

- The LLR values output by an ASV system not only depend on the characteristics of the test sample(s), but also on the reference models used to normalize the scores. The choice of reference models depends on a large number of design decisions. All these decisions will affect the LLR score assigned to a test sample. For ASV systems that come with built-in world models, it may not always be known in detail what the reference models are. For systems that build cohort models, the client models (and therefore also the LLR scores) will always depend on the cohort database available at the time of enrolment.
- Even if it is known with what kind of speech the reference models have been trained, their actual impact on the LLR score depends on many implementation details, which are often considered as information proprietary to the system manufacturer.

Of course, a laboratory involved in forensic casework could build a custom SV system, so that both the reference models and all relevant implementation details are known. Still, this does not promote the LLR values to the status of scores on a true ratio scale. There remains a long list of additional factors that have an impact on the LLR scores. The confidence interval of an identity statement can only be estimated reliably if the impact of all factors that are relevant in a specific case can be quantified. The factors that intuitively seem to have most impact on the performance of a (human or automatic) speaker recognition system are discussed in the section on operating conditions below.

BAYESIAN DECISION THEORY

In order to be useful in real cases it is necessary to 'transform' the result of forensic speaker recognition expertise or of an ASV system into evidence which can be used in the investigation. To that end, we need an indication of the reliability (or confidence) of the accept/reject decision with reference to the conditions of a particular case. If 'subjective estimates' of forensic phoneticians or the raw numbers produced by an ASV system cannot be used as straightforward confidence measures, we have to look for a better solution.

One (but certainly not the only) way to estimate the confidence to be attached to an accept/reject decision in speaker verification is to compute posterior probabilities in the Bayesian sense. Posterior probabilities have several advantages. First, information about past behaviour of the SV system under similar conditions can be taken into account. In addition, posterior probabilities are transparent measures that can be interpreted by humans. Figure 2 shows graphically how the forensic phonetic information must be used in a specific case. There are three types of information, namely previous experience with the performance of the (human or automatic) speaker recognition system, the speech evidence for this specific case, and independent (non-speech) evidence. (The figure does not show the complexity of the non-speech evidence.) To compute posterior probabilities, and to make accept/reject decisions these three information types must be converted into quantitative measures. These measures are, respectively:

- **FAR and FRR** are the false accept rate curve and the false reject rate curve of the speaker recognition system under operating conditions similar to those applying in the case at hand. If we denote those conditions as *C*, *FAR* and *FRR* can be defined as

$$FAR_C = P(\text{accept} \mid \text{non-target}, C)$$
$$FRR_C = P(\text{reject} \mid \text{target}, C)$$

  Note that FAR_C and FRR_C are not just 'hard decision' values, but functions of a threshold in terms of log likelihood ratio, as depicted

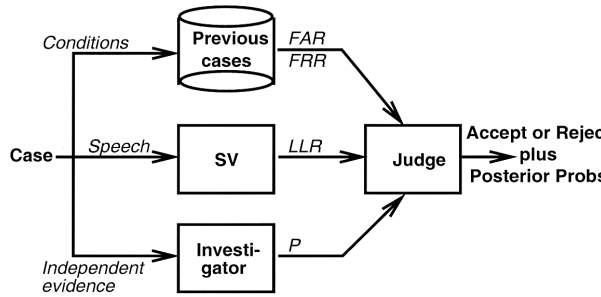

Figure 2 Three important types of input must be taken into account to compute the posterior probabilities

in Figure 1. For example, FAR_C (θ) is the False Accept Rate under condition *C* with threshold θ. For automatic systems FAR and FRR can be computed by means of simulations with pre-recorded speech in a database which reflects the relevant conditions. For human experts meaningful estimates of FAR and FRR are much more difficult to obtain.

- **LLR** value of the speech evidence. This is the output of the scoring module of the ASV system when testing the hypothesis that the suspect is the same person as the wanted criminal (or the genuine customer in civil applications).
- ***P*(target)** is the prior probability that the suspect is the wanted criminal, without taking the speech evidence into account, but based only on independent evidence or counter-evidence. In forensic cases this kind of information is provided by the investigators. In the courtroom the independent evidence is presented by the prosecutor and evaluated by the jury or by the judge (depending on the legal system). In civil applications of ASV independent information on the prior probability that an identity claim is true can come from the match between the previous behaviour of the client and a new transaction that is attempted (Boves 1998). To simplify the equations we define *P(target)* as

$$P(target) = P$$

It is important to note that in this approach both the SV system and the (interpretation of) the independent evidence are fallible. So neither the SV system nor the investigator weighing the independent evidence can claim a 100 per cent confidence in decision making. However, using the scheme of Figure 2 makes it possible to exploit the evidence gathered by

both the SV system (speech evidence) and the investigator (non-speech evidence).

When all three inputs are available, we can compute the posterior error probabilities. Given that we accept the null hypothesis H_0 (the suspect is the same as the criminal), the error probability is equal to (with *P*, *C*, *LLR*, *FAR*, and *FRR* as defined above):

$$P(\text{error} \mid \text{accept}, C) = \frac{(1-P) \cdot FAR_C(LLR)}{(1-P) \cdot FAR_C(LLR) + P \cdot (1 - FRR_C(LLR))}$$

and given that we reject the null hypothesis, the error probability is equal to:

$$P(\text{error} \mid \text{reject}, C) = \frac{P \cdot FRR_C(LLR)}{(1-P) \cdot (1 - FAR_C(LLR)) + P \cdot FRR_C(LLR)}$$

From a forensic phonetics point of view *P* is unknown, since it will be provided by the investigator, based on data which are independent of the actual phonetic research. Therefore, the best thing we can do to get insight into the meaning of the objective performance measures FAR_C(LLR) and FRR_C(LLR) is to take the ratio between the two posterior error probabilities as a function of *P*. We define this ratio R as:

$$R(C, LLR, P) = \frac{P(\text{error} \mid \text{reject}, C)}{P(\text{error} \mid \text{accept}, C)}$$

And it should be noted that R is a function of the triple (*C*, *LLR*, *P*). This measure *R* places *LLR* in its real context, depending on the performance of the SV system in condition *C* and with knowledge of the a priori independent evidence *P* concerning the hypothesis under test. To give an impression of the meaning of *R* an example may help: if $P(\text{error} \mid \text{reject}, C) = 0.8$ and $P(\text{error} \mid \text{accept}, C) = 0.1$ we are 8 times more likely to make an error if we reject than if we accept the charges against the suspect. For the two cases shown in Figure 1 the ratio *R* as a function of *P* is plotted in Figure 3. If we take R=1 as the threshold for accepting H_0, for case #1 the judge or jury can only condemn the accused if, prior to receiving the speech evidence (i.e. based on evidence that is independent of the speech material under investigation), they are more than forty-two per cent sure that the suspect is the perpetrator. However, in case #2 a *P* value as low as three per cent is enough to convict the suspect.

One can interpret the ratio *R* for a given *P* value on a verbal scale, as in Table 1. Two comments must be made on this verbal scale. First, the posterior probabilities – or, in other words, our best approximation of

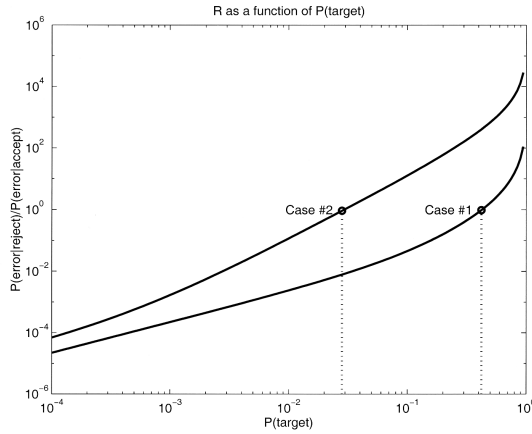

Figure 3 R as a function of P(target) for the two cases shown in Figure 1. Case #1 will be accepted if P(target) > 42% and case #2 will be accepted if P(target) > 3%

Table 1 Verbal scale for the ratio R

Interval	Decision	Confidence
$R<10^{-3}$	reject	very high
$10^{-3}\leq R<10^{-2}$	reject	high
$10^{-2}\leq R<10^{-1}$	reject	moderate
$10^{-1}\leq R<10^{0}$	reject	low
$10^{0}\leq R<10^{1}$	accept	low
$10^{1}\leq R<10^{2}$	accept	moderate
$10^{2}\leq R<10^{3}$	accept	high
$10^{3}\leq R$	accept	very high

the confidence to be attached to the accept or reject decision – are essentially dependent on the prior probability. Therefore, R is really a function of the prior probability *P(target)*. However, the measure R allows us to get a feeling for how big the risk is that the judge makes an erroneous decision when he rejects the hypothesis H_0, given the belief that we have in independent evidence. Second, very big (and very small) values of R can only be obtained if the tails of the target and non-target score distributions can be estimated accurately. This, in turn, requires a very large number of cases, numbers which can only be obtained in simulation experiments on databases with large numbers of test utterances. In most experiments the number of cases will limit the useful range of LLR to the interval (LLR^-, LLR^+), with FRR_C (LLR^-) = α and FAR_C (LLR^+) = α,

and α a small percentile value. For $LLR < LLR^-$ we assume $R = 0$, and for $LLR > LLR^+$ we assume $R = \infty$.

The Bayesian approach to combining prior probabilities and actual scores derived from pieces of evidence (speech samples) requires that *P* and *LLR* are independent. Thus, eventually *P* must be estimated by the investigator in a forensic case, not by the forensic phonetician. If the latter were to bring *P* to bear, the equivalent of the *LLR* score assigned to a set of speech samples *on the basis of the speech only* could no longer be considered unbiased.

THE OPERATING CONDITIONS

To be able to apply automatic speaker verification in real-world forensic cases, we first need to chart the performance of the SV system in several conditions. To that end, we must obtain estimates of $FRR_C(LLR)$ and $FAR_C(LLR)$ in the condition(s) of interest. Each condition is determined by a number of factors which may affect the performance of a (human or automatic) SV system. Research over recent years has addressed a number of factors which affect the performance of virtually any SV system. In the following subsections we discuss the most obvious factors, not necessarily in order of importance. The first three factors are related to the speaker him/herself, the following three to the acoustic background and transmission channel effects, and the last three to the design of the SV system. Most of the results referred to below are based on experiments with Switchboard data. Virtually all are based on experiments with large databases in which the identity of the speakers of all samples was known.

Speaker

Speech is behaviour, and is therefore characterized by both inter-speaker variation and intra-speaker differences (Boves 1998). Speaker verification makes use of the inter-speaker differences, but the intra-speaker variation can mask those inter-speaker differences. Thus, one of the major issues in Automatic Speaker Verification is how to cope with intra-speaker variations. Most state-of-the-art SV systems use statistical speaker modelling (e.g. Hidden Markov Modelling) together with speaker dependent score normalizations, which to some extent deals with the intra-speaker variation problem.

Doddington *et al*. (1998) investigated if there are differences in the recognizability of different speakers based on test data used for the NIST 1998 speaker recognition evaluation. They found that so-called goat speakers (unreliable applicant speakers with a high false reject rate) have the largest performance effect (25 per cent of the most goat-like speakers contributing 75 per cent of the false reject errors). However, the a priori detection of these speakers is still an open question. Moreover, goaty behaviour may be due to a combination of speaker characteristics and the

recording conditions of the training speech. If only one or two recordings are available to enrol speaker models, and all enrolment sessions come from a very noisy environment, one can have little hope that the speaker models are very accurate.

Gender

Gender is among the most obvious factors that one would expect to affect speaker recognition. Indeed, experiments on databases containing speech of both genders invariably show that cross-sex confusions are rare. Also, with few exceptions it has been found that the performance of modern ASV systems is as good for female as it is for male speakers. Figure 4 shows the FRR and FAR curves for the two sexes obtained with the A^2RT system. Clearly, the curves are virtually identical. This means that in our SV system the LLR scores are essentially independent of the gender of the speaker. It should, however, be emphasized that each particular SV system should be checked for gender dependence in performance due to, for example, choices made in the design of that SV system.

Language

Little is known about the impact of language on the performance of SV systems. For example, what happens if the training speech of a speaker is in his native language, but the test speech is in another language? Or can an SV system designed using one language be applied to a second language with the same performance? For text independent speaker identifi-

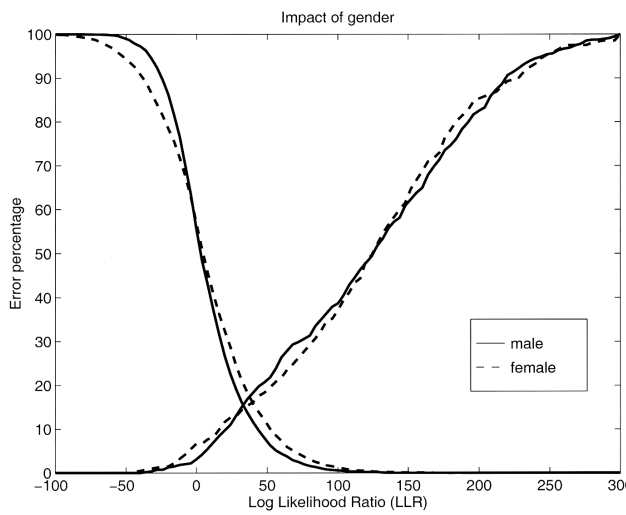

Figure 4 False Reject and False Accept Rates as a function of the LLR threshold value for female and male subjects

cation using telephone speech it has been shown that the impact of language mismatch between training and testing data is minimal, sometimes less than 0.5 per cent (Durou and Jauquet 1998). It should be noted that these results are obtained on the PolyCost database containing European languages only.

Handset

All experiments with the Switchboard data have shown that the type of handset microphone has an enormous impact on error rates. The NIST evaluations show that SV performance on carbon button recordings is significantly worse than on electret recordings. This is due to the significantly larger degree of variability exhibited by carbon button microphones.

Also, when the microphone type is the same (ST) for training and for testing, one might expect better performance results than when the microphone type differs between training and testing recordings (DT). (See Figure 5.) Performance degradation is typically a factor 3 or more when going from the ST to the DT condition (Reynolds 1996). Most state-of-the-art SV systems use some kind of channel normalization technique, like cepstral mean subtraction. Despite such techniques, a performance gap still exists between the ST and the DT condition, because these techniques only remove first-order linear distortions (while carbon button microphones are notorious for their non-linear distortions). The problem with carbon button microphones is likely to disappear, as old fashioned handsets are replaced by modern devices.

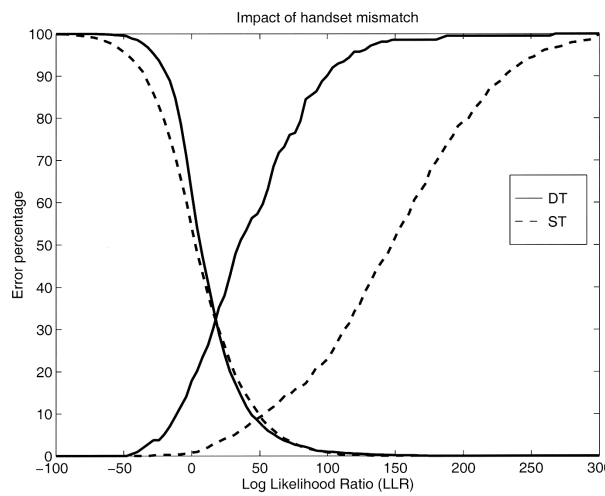

Figure 5 False Reject and false Accept Rates as a function of the LLR threshold value for test segments with the same handset type (ST) or a different handset type (DT) as used during training

Recording platform
Especially in forensic casework the recording platform may be important. Often recordings are delivered on audio cassette or other analogue media, for example answering machine tapes. It is very difficult to predict the impact of analogue recording platforms. The best situation would be to have the same recording platform for training and test speech, which is possible in, for example, phone tap recordings. Traditional analogue telephone tapping devices are increasingly being replaced with advanced digital facilities, with calls being stored in a digital format (Broeders 1995).

Signal quality
Signal quality can vary substantially within a set of test utterances. Moreover, 'signal quality' is a multi-dimensional concept, of which signal-to-noise ratio is just one aspect. Signal quality is determined by the acoustic background, the quality of the transmission channel, and the behaviour of the speaker. If the acoustic background noise level is high, most speakers react by raising their voice level, which affects various characteristics of the speech signal. This is the well known Lombard reflex (Junqua 1996). The ways in which speakers react to distortions in the transmission channel are more difficult to predict. In other situations, the level of the speech signal can be low, for example due to the fact that the utterances mostly contain whispered speech, murmuring, *erm*-sounds or other back-channel utterances. The effects of signal quality on performance are not so evident and consistent over different SV systems as they are for the various handset conditions. Also signal quality is difficult to quantify, except for the well-known signal-to-noise ratio (SNR): for low SNR speaker verification performance drops significantly.

Number of training sessions
It is obvious that speaker models are more powerful if they include a better estimate of intra-speaker variability. Under normal conditions reliable estimates of intra-speaker variability can only be obtained by means of multiple recording sessions, preferably spanning a considerable period of time, and made at different times of the day on different days of the week. In most real applications one must be satisfied with a small number of enrolment sessions.

Recently, attempts have been made to induce intra-speaker variation in a single enrolment session by requesting that the speakers use different speaking styles. As yet, no definitive results of this experiment are available (Karlsson *et al*. 1998).

Amount of test data
Not surprisingly, a longer duration of the test utterance produces better performance. Typically, ten times more testing data halves the equal error

rate of the SV systems participating in the NIST speaker recognition evaluations of 1997 and 1998 (Przybocki and Martin 1998).

Anti-speaker modelling

Anti-speaker modelling (be it world modelling or cohort modelling) has been shown to work well in practice for normalizing the speaker model likelihood given the observations. A study of several kinds of anti-speaker modelling (Rosenberg and Parthasarathy 1996) reveals that cohort modelling performs only marginally better than world modelling, and the performance impact of the number of speakers used for training of the world model or the number of cohort models is limited.

Ideally, the speech used to train the speaker model and the speech used to train the anti-speaker models should be recorded under exactly the same conditions. The better this match is, the better the normalizing effect by the anti-speaker likelihood. For this reason, several SV systems use gender and/or handset dependent world models.

Additional factors

A long list of factors which may or may not affect the performance of SV systems can be made. Some of these factors have been hypothesized in the literature, but not systematically investigated. Other factors are newly introduced, like the effects of source coding in digital cellular telephony. Kuitert and Boves (1997) investigated the effects of GSM coding on the performance of an automatic SV system; they found that the codec *per se* has little impact on the performance.

One factor which has received considerable attention is voice pitch. While pitch can contribute to speaker recognition, it is effectively annihilated in the conventional cepstral coefficients that are used as acoustic features in most modern ASV systems. Moreover, average pitch has been shown to exhibit a relatively large intra-speaker variation (Kraaijeveld 1997).

It is practically impossible to chart the influence of all factors that may affect SV performance. Fortunately, often logical reasoning suffices to convincingly argue that a specific factor can hardly be relevant, because it does not affect the features in a given SV system (e.g. the case of voice pitch with cepstral features). For example, if there exist differences in speech rate between the suspect's recordings and the perpetrator's recordings – which is not unlikely – it can be reasoned that a SV system using HMMs is rather insensitive, because an HMM can adapt its state sequence to changing speech rates.

Prohibitive factors

It should be noted that there are also factors which make the use of (automatic) SV systems impossible. The most extreme cases are those where the perpetrator uses recordings which are digitally altered, for example by applying one of the better voice conversion systems (Genoud

and Chollet 1999). If there is reasonable suspicion that a criminal has used voice conversion software, one should refrain completely from using SV in the investigation or as evidence in court. (But in such a case the type of voice transformation can probably serve as evidence.)

Intentional vocal disguise is another factor that may invalidate speaker recognition, although this is much less obvious. Depending on the disguise technique and its impact on the acoustic features in an SV system, disguise may or may not affect the performance. Most probably, this should be investigated on a case-by-case basis.

As yet there is no technology that is able to reliably detect disguise, or the use of digital voice conversion techniques.

THE HARASSMENT CASE

We can now revisit the harassment case, introduced above. In some senses it seems to be relatively easy: the recordings of all three sets of harassment calls can be traced to exactly the same recording environment. In addition, in both sets of questioned recordings the speaker used a whispery voice. Thus, one would expect that these are optimal performing conditions for an automatic speaker verification system. Of course, this intuitive reasoning assumes that the relevant intra-speaker variability is not increased significantly because of the non-normal way of speaking. In addition, other potentially relevant factors may not make for a simple case. Since the texts spoken in the two sets of recordings differ, we are obliged to use text-independent SV methods, which are known to be less powerful than text-dependent methods.

We had three recordings available on which to base a judgement. The speech recordings came on CD-ROMs in MS-WAVE format (stereo, 44.1 kHz sampling rate and 16-bit per sample). The total duration of the first set of harassment calls was 96.4 seconds; the second set of harassment calls had a total duration of 132.3 seconds. The total duration of the read speech recorded from the suspect was 81.9 seconds. We decided to use all the material to enrol three client models, which will be referred to as:

- PERP1 for the first set of harassment calls (three calls with spontaneous, whispered speech; call durations 36.7, 25.9, and 33.8 seconds),
- PERP2 for the second set of harassment calls (three calls with spontaneous, whispered speech; call durations 52.7, 36.2, and 43.4 seconds),
- SUSP for suspect's calls (read transcripts of the first two calls from PERP1, plus one spontaneously spoken denial of the accusation, all with normal voice; call durations 22.9, 18.5, and 40.5 seconds).

After manually removing the standard answering machine messages (such as 'message for', 'received on', and 'end of message') and automatically

removing silence from the utterances there remained 22.8 seconds of speech in PERP1, 32.0 seconds in PERP2, and 34.7 seconds in SUSP. Because PERP1 and PERP2 are spoken with a whispery voice, the silence–speech detector also removed parts of the whispery speech.

For the speaker verification task we used *A²RT* 's SV system as described above. The first task is to estimate the system's FAR and FRR curves in the appropriate conditions. To this end, we trained the speaker models on only thirty seconds of speech recorded in one session; test utterances were also thirty seconds long, in conformity with the conditions of the case. (Thirty seconds matches more or less with the total duration of the PERP1, PERP2, and SUSP recordings). Since we were not in a position to record large numbers of (male) speakers under the exact same conditions that applied in the case, we used 2642 utterances from an existing corpus: the Switchboard-2 corpus. We used 2347 non-target utterances and 295 target utterances, and a set of 250 target models. All targets were male speakers, speech for training and testing for the true speaker attempts was collected using the same phone number and with the same handset type, as in the case under investigation where the suspect's speech was recorded from the same classroom and with the same handset type as during the harassment calls.

The only mismatch factors are (1) language (Switchboard is English, the case is in Dutch); (2) the recording platform (the detailed specifications of the company voice mail system used to record the three sets of calls were not available to us. Thus, we do not know whether the signals were treated by some kind of coding mechanism to reduce the number of bytes needed to store messages in the voice mail boxes); (3) the Switchboard data base contains spontaneous speech with a normal voice, while the case data is partly spontaneous, partly read speech, partly whispered, and partly with normal voice.

Subsequently, we obtained LLRs for the speech used to train PERP1 matched against SUSP and PERP2, for the speech used to build SUSP matched against PERP1 and PERP2, and for the speech underlying PERP2 matched against PERP1 and SUSP (see Figure 6). Table 2 shows which P(target) suffices to accept H_0 and Table 3 shows the ratio between P(error|reject) and P(error|accept) when P(target) is 50 per cent. Not surprisingly, the match between PERP1 and PERP2 was very close (low P values in Table 2 and high R values in Table 3). In fact, the LLR value of this case falls in the interval (LLR^+; ∞), thus resulting in $R = \infty$. Also, the matches of SUSP with PERP1 and PERP2 were good enough to accept H_0 with high confidence.

The matrices in Tables 2 and 3 are not necessarily symmetrical, because in general a high probability for the test data given a model of the observations on the training data does not necessarily imply an equally high probability for the training data given a model of the observations on the

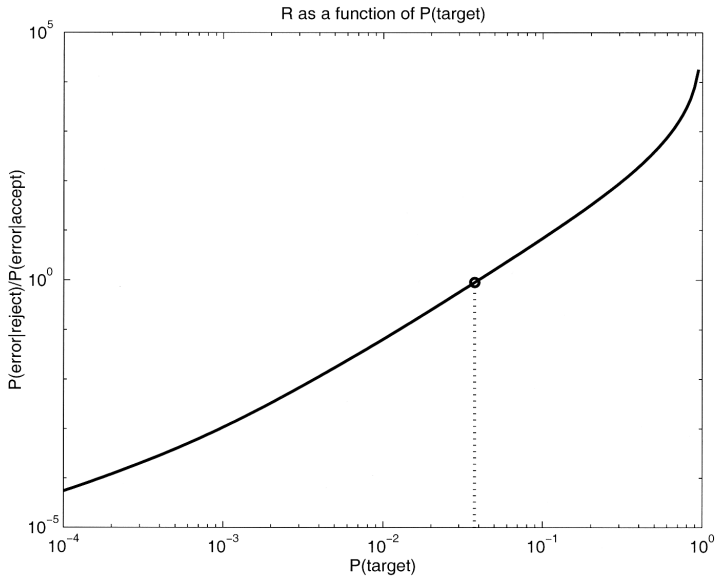

Figure 6 R as a function of P(target) when PERP1 is used as training speech and SUSP is used as test speech. The circle is the operating point where $R=1$

Table 2 P(target) value needed to obtain $R=1$, with fixed LLR and C

P(target)		Trained on		
		PERP1	SUSP	PERP2
Tested on	PERP1	<0.1%	0.2%	<0.1%
	SUSP	3.8%	<0.1%	4.0%
	PERP2	<0.1%	0.2%	<0.1%

Table 3 R value for P(target) = 50%

R		Trained on		
		PERP1	SUSP	PERP2
Tested on	PERP1	∞	$4.0 \cdot 10^2$	∞
	SUSP	$3.3 \cdot 10^2$	∞	$2.9 \cdot 10^2$
	PERP2	∞	0.2%	∞

test data. It is the 'quantifying' effect of data modelling which causes the asymmetry in the distance measures.

However, we should not commit ourselves to the figures in Tables 2 and 3 because we could not solve the three mismatch conditions mentioned above. For example, it is possible that the large difference between our non-target speech and the test speech ($FAR_C(LLR) = 0$) is due to unknown but systematic effects in the recordings of the perpetrator and the suspect. Even if that may be difficult to imagine, based on the limited knowledge that was available to us we cannot completely rule out this possibility. Recall that we were not given any information on the waveform coding employed by the voice mail system used to record all test utterances. Also, all test calls were recorded under the same acoustic conditions, with the same type of handset; both the room acoustics and the handset in the test speech may have been idiosyncratic, thereby adding to the difference between the LLRs computed for (non-matching) impostor trials and the (matching) test trials. Of course, had the case been important enough to warrant the costs, we could have recorded a sufficiently large and varied set of additional speakers to train the LLR distributions of genuine and impostor trials in a matching condition. However, such cases are the exception, rather than the rule.

In the case under analysis we had no information to estimate an independent prior probability of the speakers in the three sets of recordings being the same person. Careful and detailed phonetic analysis yielded a long list of speech features in the three sets of recordings that were very similar, yet sufficiently exceptional to consider them as idiosyncratic. However, even if such phonetic information cannot be brought to bear on the speaker models of our SV system in any direct and explicit way, it is still very dangerous to consider it as independent evidence. Both the phonetic analysis and the speaker models of the SV system are based on the same speech evidence, and as long as it is unclear to what extent they may be exploiting the same information in the speech signal, it is unsafe to regard the outcome of the phonetic analysis and the outcome of the automatic analysis as mutually independent.

The outcome

The eventual decision in a forensic case has very little to say about the confidence and truth of the forensic phonetician's opinion about the identity of the speakers who produced two sets of speech samples. This is so because the final decision may have been based almost completely on other evidence (or in the case of a dismissal on formal grounds, on the way the case was brought before the judge). Yet it is always interesting to know the final verdict. In the case at hand there never was one. The suspect maintained his denial, and the harassment calls stopped after the second set used in this study. (Maybe the threat of hard scientific proof alone was enough to stop the caller.) The company therefore dropped the case.

A SIMULATION EXPERIMENT

We have already made clear that both the investigator and the automatic speaker recognizer are fallible. The main question in this section is how these two can be used to combine the probative values of speech and non-speech evidence together, to achieve more reliable decision making in court. We use our Bayesian decision paradigm to investigate this and choose a subset of the NIST'98 data as experimental data, namely those cases with male speakers, thirty seconds of test data, and two minutes of training data collected in two sessions. Ten thousand utterances are used to compute the FAR and FRR curves of the SV system, and 15 000 utterances are used to do the actual testing.

These test utterances we used in a simulation experiment: for each of the 15 000 test utterances the P(target) value (the prior probability that the suspect is also the perpetrator) is given by an imaginary investigator. One way to create this imaginary investigator with a semi-random output value for P(target) is as follows: in case the suspect is also the perpetrator the P(target) value is a random selection from the uniform distribution over the interval $(1 - p, 1)$, with p a real value between 0.5 and 1.0, and in the opposite case it is a random selection from the uniform distribution over the interval $(0, p)$. So the larger p, the worse the quality of the P(target) value supplied by the investigator. If p varies from 0.5 to 1.0, the total error rate of the imaginary investigator (defined as $TER = (FAR + FRR)/2$) is equal to $2p - 1/2p$, with the accept/reject threshold set to P(target) = 0.5. Then, per test utterance we compute the ratio $R = P(\text{error}|\text{reject}, C)/P(\text{error}|\text{accept}, C)$ as a combination of investigator input and the input of the ASV system, with the semi-random values for P(target), the LLR value of the test utterance and the FAR and FRR curves of the SV system as input, and $R = 1$ serving as accept/reject threshold value. So, finally we obtained 15 000 R values and thus 15 000 accept or reject decisions which take into account both speech and (in this case, simulated) non-speech evidence.

In Figure 7 the total error rate of the combination of investigator and ASV is plotted against the total error rate of the investigator to show what ASV can add to the opinion of an investigator. Two conditions are plotted, one for test segments with the same handset type (ST) as used during training, and one for different handset type (DT). The horizontal lines are the stand-alone performances of the SV system in the two conditions. For example, for the ST condition the ASV TER is 9.6 per cent and if the TER of the investigator is in the range of 3.6 per cent to 40.5 per cent ASV can bring these TERs down to between 3.6 per cent and 9.6 per cent respectively, so giving up to a factor 4.2 reduction in TER. Table 4 shows in which intervals a combination of the investigator and ASV gives a real 'performance gain'. Only if the investigator TER is lower than 3.6 per cent does addition of the SV system (with a TER of 9.6 per cent) start to be 'counterproductive'. However, the better the SV system

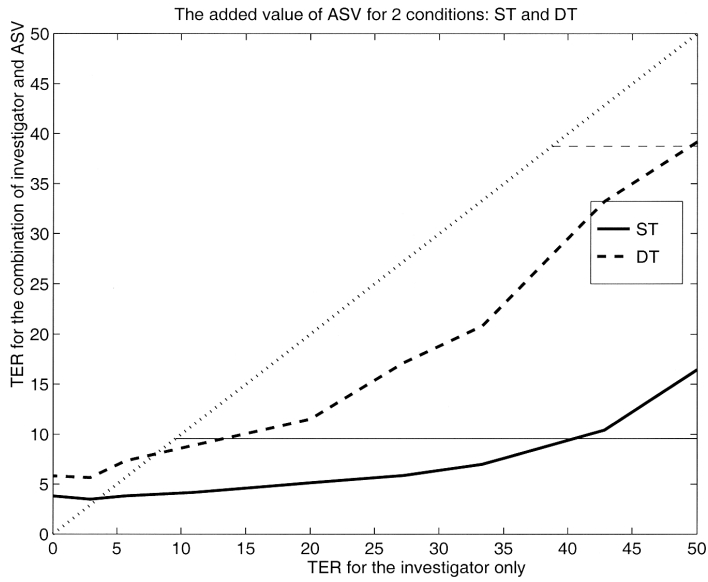

Figure 7 The added value of ASV. The thick lines show the Total Error Rate for the combination of investigator and ASV as function of the Total Error Rate of the investigator only.

Table 4 Break-even points in % for *P*(target) based on Figure 7

Condition	TER of ASV standalone	Jury better than combination	Combination better than both ASV and investigator	ASV better than combination
ST	9.6%	0.0 – 3.6	3.6 – 40.5	40.5 – 50.0
DT	38.7%	0.0 – 8.0	8.0 – 49.5	49.5 – 50.0

performs, the smaller the interval where ASV has a counterproductive effect on the investigator's input and the larger the interval where ASV really helps the judge in taking a decision.

CONCLUSIONS

In this paper we have analysed the factors that have an impact on the log likelihood ratio scores produced by automatic speaker verification sys-

tems. It has been explained why these scores are measurements on an ordinal scale. Therefore, the absolute values of the scores cannot be used as the sole data to attribute a formal confidence value to the decision to accept or reject the test sample as coming from the claimed speaker. Automatic SV systems can only be used in forensic field work to substitute the 'subjective' confidence score attributed to an opinion by a forensic phonetician if three crucial inputs are available: the performance of the automatic system in the appropriate condition must be known (in terms of false accept and false reject rate), the probative value of the speech evidence (in terms of log likelihood ratio), and that of the non-speech evidence (the prior probability that the suspect is also the perpetrator). A combination of these three inputs has led to a measure *R* which can be interpreted intuitively, is robust in different operating conditions, includes self knowledge of the SV system and integration of independent evidence, and contributes to better decision making based on both speech and non-speech evidence in forensic casework.

REFERENCES

Boves, L. (1998) 'Commercial applications of speaker verification: overview and critical success factors', in *Proceedings of RLA2C: La Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques* (Avignon), pp. 150–9.

Broeders, A. P. A. (1995) 'The role of automatic speaker recognition techniques in forensic investigations', in *Proceedings of the XIII International Congress of Phonetic Sciences* (Stockholm), pp. 154–61.

Doddington, G. (1998) 'Speaker recognition evaluation methodology: an overview and perspective', in *Proceedings of RLA2C: La Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques* (Avignon), pp. 60–7.

Doddington, G., Liggett, W., Martin A., Przybocki, M. and Reynolds, D. (1998) 'Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation' in *Proceedings of the International Conference on Spoken Language Processing*, (Sydney), pp. 1351–4.

Durou, G. and Jauquet, F. (1998) 'Cross-language text-independent speaker identification', in *Proceedings of the European Signal Processing Conference* (Rhodes), pp. 1481–4.

Genoud, D. and Chollet, G. (1999) 'Deliberate imposture: a challenge for automatic speaker verification systems', in *Proceedings of the European Conference on Communication and Speech Technology* (Budapest), *Eurospeech*, vol. 5, pp. 1971–4.

Gibbon, D., Moore R. and Winski R. (eds) (1997) *Handbook of Standards and Resources for Spoken Language Systems*, Berlin: Mouton de Gruyter.

Junqua, J.-C. (1996) 'The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex', *Speech Communication*, 20: 13–22.

Karlsson, I., Banziger, T., Dankovicovà, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F. and Schrerer, K. (1998) 'Within-speaker variability due to speaking manners', in *Proceedings of the International Conference on Spoken Language Processing*, (Sydney), pp. 2379–82.

Kraaijeveld, J. (1997) 'Idiosyncrasy in prosody. Speaker and speaker group identification in Dutch using melodic and temporal information', PhD thesis, University of Nijmegen.

Kuitert, M. and Boves, L. (1997) 'Speaker verification with GSM-coded telephone speech', in *Proceedings of the European Conference on Speech Communication and Technology* (Rhodes), pp. 975–8.

Lee, C. (1997) 'Unified statistical hypothesis testing approach to speaker verification and verbal information verification' in *Proceedings COST Workshop Speech Technology in the Public Telephone Network* (Rhodes): pp. 63–72.

Martin, A. *et al*. (1997) 'The DET curve in assessment of detection task performance' in *Proceedings of the European Conference on Speech Communication and Technology* (Rhodes), pp. 1895–8.

NIST (1998) *NIST 1998 Speaker Recognition Evaluation Plan*. http://www.nist.gov/speech/spkrec98.htm

Przybocki, M. and Martin, A. (1998) 'NIST speaker recognition evaluation – 1997' in *Proceedings of RLA2C: La Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques* (Avignon), pp. 120–3.

Reynolds, D. (1996) 'The effects of handset variability on speaker recognition performance experiments on the Switchboard corpus' in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (Atlanta), pp. 113–16.

Rosenberg, A. and Parthasarathy, S. (1996) 'Speaker background models for connected digit password speaker verification' in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (Atlanta), pp. 81–4.

Stevens, S. (1951) *Handbook of Experimental Psychology*, New York: Wiley.