On the Use of Automatic Speaker Verification Systems in Forensic Casework

Johan Koolwaaij   Lou Boves
A^2RT
Nijmegen University
6500 HD Nijmegen, The Netherlands

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 incriminated because they find it impossible to provide 'hard' estimates of the confidence level of an expert opinion. In this paper it is investigated 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 for several reasons hard estimates of the confidence of an opinion in a specific case cannot be provided by automatic speaker verification either.

1. Introduction

Automatic Speaker Verification (SV) and Forensic Casework have long been considered as essentially unrelated disciplines, because the former was seen as a one alternative forced choice problem, whereas the latter used to be presented as a an open set identification problem. However, [1] has pointed out that many forensic cases boil down to the question 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 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 but also to the forensic phoneticians. In many cases the police is 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 unreliable, 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, where the number of persons who speak the language is small. In these cases a text-independent SV system might be of great help.
In all stages of forensic applications of speaker recognition it is important that one is able to state a confidence interval for conclusions regarding the identity of the voices of a known suspect and an unknown perpetrator. If the statement must be used in a court, a specification of the confidence level is necessary to allow the judge to weigh this piece of evidence. If it is to be used during the police investigation, confidence levels will be used to weigh the evidence in setting priorities for investigating specific suspects. In the harassment case described in this paper, the confidence statement was used to decide on how to proceed with the investigation.
It is well known that forensic phoneticians often have difficulty in making estimates of the confidence level with which they can identify a person by her/his voice. Thus, forensic case workers are interested to know to what extent the use of automatic SV systems could be used to obtain an 'objective' confidence estimate.

In this paper we investigate the implications of using an SV system to estimate the confidence level for an identity statement on the basis of 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 identify 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 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 harassing 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 harassing calls started again, in a whispery voice and from the same classrooms. Now, the obvious question is whether the two sets of harassing calls have been made by the same speaker, and whether this speaker is the same person as the one who read the transcripts. Obviously, this problem can be cast in the form of a one alternative forced choice problem: we can take the test calls for building a voice pattern of a known speaker, and try to answer the question whether all harassing calls have been made by the same person.
In this paper we take this case as the starting point to investigate to the contingencies of applying the procedures and technology developed for Automatic Speaker Verification to forensic cases that can be formulated as speaker verification problems.

2. Evaluation Measures

In principle, one might think that stating confidence levels and intervals for the decision of an automatic SV system should be trivial. For all serious SV products performance figures are available, that specify the proportion of false accepts and false rejects, Equal Error Rates, or ROC or DET curves [2]. Thus, one might expect that a properly built SV system should be able to produce an objective confidence measure on an absolute scale. This 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 characterise the overall behaviour of the system; unfortunately, they do not allow to make inferences about specific individual cases treated by a system. Of course, in forensic work it is only individual cases that matter. A simple way to illustrate why some average measure of the performance of an SV system is not adequate in forensic work is to look at an example. Fig. 1 shows the proportions of false accepts (left hand curve) and false rejects (right hand curve) of our text-independent SV system as a function of thresholds 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 LLR value that corresponds to equal probabilities of false reject and false accept). However, it is

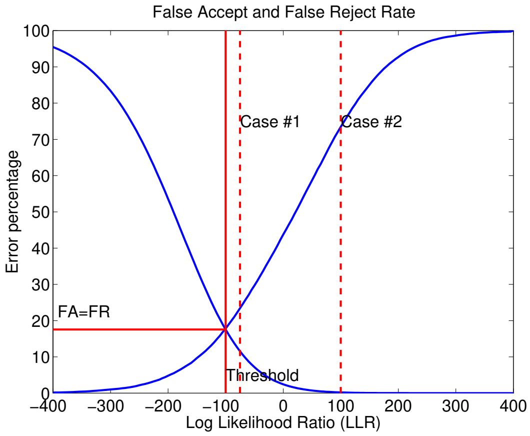

Figure 1: False Reject and False accept rates as a function of the LLR threshold value. The dashed vertical lines represent two individual cases with LLR values greater than the LLR for Equal Error Rate.

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 provided by the system manufacturer).
The example in Fig. 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 normalise the otherwise unscaled raw likelihood values [3]. One might be tempted to assume that likelihood *ratio* scores are measures on a **ratio** scale; unfortunately, in actual practice, LLR's are measures on an **ordinal** scale [4].
There are two reasons why the LLR produced by a speaker verification system must be interpreted as measurements on an ordinal scale:

- The LLR values output by an SV system do not only depend on the characteristics of the test sample(s), but also on the reference models used to normalise the scores. The choice of reference models depends on a large number of design decisions. Some systems use customer dependent cohort models, while other use customer independent world models. In the latter case, a system may choose gender dependent or gender independent world models. All these decisions will affect the LLR score assigned to a test sample. For SV 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, that 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 factors that do 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.

In general terms, the LLR scores depend on the degree of mismatch between the training and testing conditions. For the training not only the speech recordings of the known subject (the suspect, in forensic casework), but also the speech used to create the reference models counts. Ideally, the speech used to train the suspect's model and the speech used to train the reference models should be recorded under the exact same conditions; these conditions should be equal to those prevailing during the recording of the perpetrator's speech. In many real life cases the conditions under which the perpetrator's speech has been recorded may not be known exactly; if more than one recording is available, they may come from different environments, using different microphones and transmission channels.

3. The Speaker Recognition system

In order to be able to build the 'ideal' SV system, in which the influence of the complete set of factors that determine the exact LLR value are explicitly accounted for, one would need to carry out a large number of controlled experiments. Since it is not our intention to even approximate such a system, we will rather demonstrate the impact of a couple of factors on the LLR scores, and therewith on the False Accept and False Reject Rates. The experiments are based on the text-independent speaker verification task in the 1998 NIST Speaker Recognition Evaluation [2]. The speech used in the experiments was taken from the SwitchBoard-2 Phase 1 corpus. Thus, all recordings

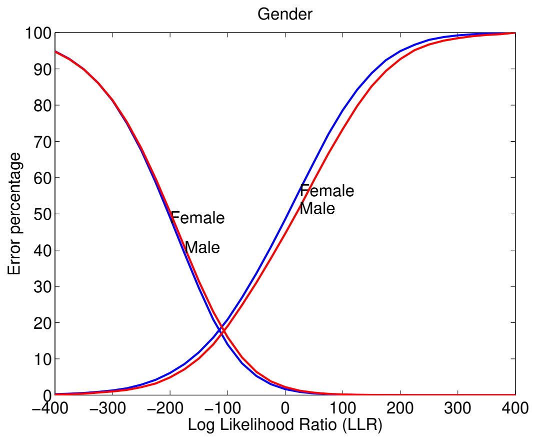

Figure 2: False Reject and False accept rates as a function of the LLR threshold value for female and male subjects.

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, ehm, huhhuh*, etc.). The figures shown in this paper were obtained with the A^2RT system, that appeared to perform reasonably well in the 1998 test campaign.

The speaker recognition system used to generate the FRR and FAR curves is a text-independent SV system. Since the system 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. Parameterisation 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 total feature vector is formed by appending the delta's and delta-delta's of the 13 coefficients, making for a total of 39 features. For each 'client' a single model has been trained, using 2 minutes of speech, recorded in a single session. Ergodic four state HMM models have been trained, with 32 Gaussian mixture densities per state. Reference models with the same topology have been trained using recordings of a large number of speakers, none of whom is among the 'clients'. Separate reference models for male and female speakers were built.

For testing we have used speech samples with a duration of 30 seconds.

3.1. Gender

Gender is among the most obvious factors that one would want to control. From Fig. 2 it is clear that the

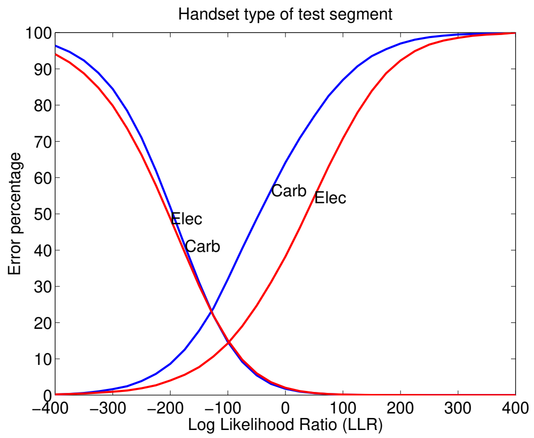

Figure 3: False Reject and False accept rates as a function of the LLR threshold value for carbon button and electret microphone handsets.

FRR and FAR curves for the two sexes are virtually identical. This means that in our SV system the LLR scores are virtually independent of the gender of the 'client'. It should, however, be emphasised that this result cannot be generalised to other SV systems.

3.2. Handset

All experiments with the SwitchBoard data have shown that the type of handset microphone has an enormous impact on the error rates. Fig. 3 shows the FAR and FRR curves for electret and carbon button microphones. From the figure it can be seen that the FAR rates for the two handset types are very close to each other. For the FRR rates, on the other hand, the curves are far apart: for speech samples recorded with a carbon button microphone the FRR is much higher that for samples recorded with an electret microphone. Although the exact EER points will differ between SV systems, it must be expected that the results will generalise to every system. This is due to the significantly larger degree of variability exhibited by carbon button microphones.

4. Bayesian Decision Theory

One (but certainly not the only one) 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. In making accept/reject decisions three important pieces of information must be distinguished and taken into account:

P is the prior probability that the suspect is the searched criminal, without taking the speech evidence into account, but only based on independent evidence or counter-evidence. In forensic cases this kind of information usually comes from the investigating officer. In civil applications of automatic SV information on the prior probability that an identity claim is true can come from the match between previous behaviour of the client and a new transaction that is attempted [5].

LLR value of the speech evidence. This is the output of the verification module of the ASV system while checking the hypothesis that the suspect is the same person as the searched criminal (or the genuine customer in civil applications).

FAR and FRR are the false accept rate and the false reject rate of the ASV system under operating conditions similar to those applying in the case at hand.

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

$$\mathrm{P(error|accept)} = \frac{\text{[1-P] FAR(LLR)}}{\text{[1-P]FAR(LLR) + P[1-FRR(LLR)]}}$$

And given that we reject this hypothesis, the error probability is equal to

$$\mathrm{P(error|reject)} = \frac{\text{P FRR(LLR)}}{\text{[1-P] [1-FAR(LLR)] + P FRR(LLR)}}$$

The posterior error probabilities for the two cases in Fig. 1 are plotted in Fig. 4. In this example the prior probability **P** is (arbitrarily) set equal to 0.75. It is clear that if we accept case #2, the posterior error probability is almost zero. Thus, accepting case #2 only leaves a very minor risk of making the wrong decision. However, if we accept case #1, the risk of making the wrong decision is still approximately 5%.

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 judge in a forensic case, not by the forensic phonetician. If the latter would 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 as unbiased.

Interestingly, in the T-Netix commercial SV system the equivalent of the prior probability **P** is one of the inputs to the function that computes the verification score (in addition to the speech sample under test and -indirectly- a database of anti-speakers that was

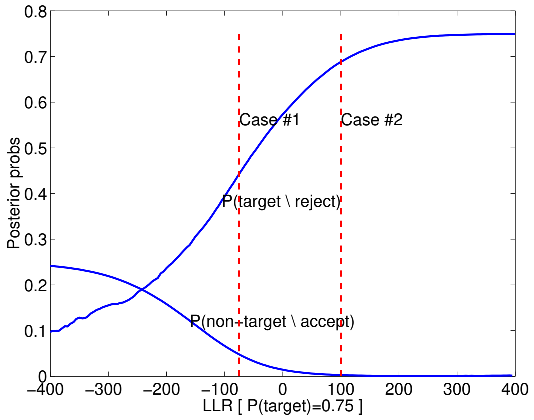

Figure 4: Posterior error probabilities

used during the discriminative training of the model of the claimed speaker) [6]. This should result in a biased score. In the A^2RT SV system the LLR score is not biased by prior probabilities; the application must make an explicit decision as how to combine this LLR score with whatever independent evidence that might be available.

5. The harassment case

We can now revisit the harassment case, introduced above. In some sense it seems to be relatively easy: the recordings of all three sets of harassment calls can be traced to the exact same recording environment. In addition, in both sets of criminal recordings the speaker used a whispery voice. Thus, one would expect that the confidence level for the decision whether or not these sets of calls come from the same speaker should be high. Of course, this intuitive reasoning assumes that the relevant within-speaker variability is not increased significantly because of the non-normal way of speaking. 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 to base a judgement on. The speech recordings came on CD-ROMS in MS-WAVE format (stereo, 44.1 kHz sampling rate and 16 bit per sample). 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 have been treated by some kind of coding mechanism to reduce the number of bytes needed to store messages in the voice mail boxes. This creates one possible mismatch between the data from this case and all other speech data that we had available to build world models. Anyway, since we were not in the position to record large numbers of (male) speakers under the same conditions that applied in the case, we decided to use recordings from the Dutch Polyphone corpus [7] and the Dutch SESP corpus [8] to build the world model; both corpora are recorded over the public switched telephone network. Polyphone comprises only domestic calls, while the better part of SESP consists of international calls. The combination of the two corpora represents a very large range of recording conditions and handsets. This selection is an attempt not to bias the world models to any conceivable specific condition different from the condition prevailing during the recordings in the case. Rather, we try to cover the most general condition possible. From Polyphone we took 439 phonetically rich sentences read by 72 male speakers; from SESP we took 384 spontaneous utterances (answers to questions like *Describe the environment you are calling from*) from 68 male speakers. The recordings from the case were downsampled to 8 kHz to make them compatible with the sampling rate in the speech for the world model.

The total duration of the first harassment calls was 74.4 seconds; the second set of harassment calls had a total duration of 91.5 s. The total duration of the read speech recorded from the suspect was 52.8 s. If we take all speech from each of the three conditions for enrolment, we still have substantially less material than the two minutes used in the NIST evaluation experiments. We decided to use all material to enrol three client models, that will be referred to as CD1 (for the first set of harassment calls), CD2 (for the read material), and CD3 (for the second set of harassment calls). Speaker models and the world model were based on acoustic features consisting of 12 LPC cepstra plus log-energy, and their deltas and delta-deltas, making for feature vectors with 39 components. LPC analysis was performed with 25.6 ms Hamming windows, with 100 frames/second. The models were ergodic HMMs, with 4 states and 4 Gaussian densities per state. The world model was trained first, starting from scratch. The client models were then adapted from the world model.

Impostor distributions were trained for the three client models by matching a large number of Polyphone and SESP recordings, not used to build the world model against each of the models and retaining the resulting LLR score. Subsequently, we obtained LLRs for the speech used to train CD1 matched against CD2 and CD3, for the speech used to build CD2 matched against CD1 and CD3, and for the speech underlying CD3 matched against CD1 and CD2.

Not surprisingly, the match between CD1 and CD3 was very close. Also the matches of CD2 with CD1 and CD3 were very close, especially compared to the LLR values found for the impostor trials. Yet, we still cannot be sure that the extremely unlikely LLR values obtained for the matches between CD1, CD2 and CD3 (unlikely against the impostor distributions that were previously obtained) really imply that the speakers in the two sets of harassment calls must be one and the same person, nor that this person is the suspect who recorded the speech for CD2. It is possible that the large difference between LLR values for our impostor speech and the test speech 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 anti-speakers to train the world model and 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 still is very dangerous to consider it as independent evidence.

5.1. 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 technical mistakes in 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. Therefore, the company dropped the case.

6. Conclusions

In this paper we have analysed the factors that have an impact on the LLR scores produced by automatic speaker recognition systems. It was 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 sufficient data can be provided (to train world models and estimate impostor distributions) that match the case.

References

[1] G. Doddington "Assessment of Speaker Recognition Systems", *Proceedings RLA2C*, Avignon, pp. 60-67, 1998.

[2] M.A. Przybocki & A.F. Martin "NIST Speaker Recognition Evaluation - 1997", *Proceedings RLA2C*, Avignon, pp. 120-123, 1998.

[3] Ch. Lee A "Unified Statistical Hypothesis Testing Approach to Speaker Verification and Verbal Information Verification", *Proceedings COST Workshop Speech Technology in the Public Telephone Network*, Rhodes, pp. 63-72, 1997.

[4] S.S. Stevens *Handbook of experimental psychology,* New York: Wiley, 1951.

[5] L. Boves "Commercial applications of speaker verification: overview and critical success factors", *Proceedings RLA2C*, Avignon, pp. 150-159, 1998.

[6] T-Netix *Speaker Verification Software Reference Manual for Windows NT*, Preliminary Version, Englewood, CO, 1997.

[7] E. den Os, T.I. Boogaart, L. Boves & E. Klabbers "The Dutch Polyphone Corpus", *Proceedings Eurospeech-'95*, Madrid, pp. 829-932, 1995.

[8] J.W. Koolwaaij & L. Boves "On the Independence of digits in connected digit strings", *Proceedings Eurospeech-97*, Rhodes, pp. 2351-2354, 1995.