

Estimating the Weight of Evidence in Forensic Speaker Verification

Beat Pfister and René Beutler

Speech Processing Group
Computer Engineering and Networks Laboratory
ETH Zurich

{pfister,beutler}@tik.ee.ethz.ch

Abstract

In forensic casework, the application of automatic speaker verification (SV) aims to determine the likelihood ratio of a suspect being vs. being not the speaker of an incriminating speech recording. For that purpose, the likelihood of the anti-speaker has to be estimated from the speech of an adequate number of other speakers. In many cases, speech signals of such an anti-speaker population are not available and it is generally too expensive to make an appropriate collection.

This paper presents a practical procedure of forensic SV which is based on a text-dependent SV system and instead of an anti-speaker population, a special speech database is used to calibrate the valuation scale for an individual case.

1. Introduction

Generally, the application of SV aims to automatically decide if a person is a true user or an impostor. Since even the best SV systems are not able to make this decision 100% correctly, the trade-off between false acceptance rate (FAR) and false rejection rate (FRR) has to be optimized, taking also the application-specific error costs into account.

In forensic application of SV, although the question is basically the same (“Are the speakers of two signals¹ identical?”), a Yes/No answer is not appropriate. For reasons explained e.g. in [1] and [2], the outcome of a forensic investigation aiming to identify a person by his/her voice has to be probabilistic, namely the likelihood ratio (LR) of the two hypotheses H_0 (“same speaker”) and H_1 (“different speakers”).

As shown e.g. in [3], using a SV system in forensic casework is far from being straight forward. Conventional performance measures of SV systems such as ROC (receiver operating characteristics) curves, specifying the system’s behavior over a large number of trials, cannot be used to estimate the LR of an individual case. But in forensic casework it’s the individual case that matters.

¹One of these two signals originates from the unknown perpetrator and is called hereafter “the incriminating signal”. The other one, the so-called “test signal”, has been spoken by a known suspect.

An individual case can be characterized by a set of case-specific factors that have some influence on the numeric output of the SV system. We can distinguish between technical factors (characteristics of the recording and transmission equipment, ambient noise, signal duration, etc.) and speaker properties (e.g. sex, age, mental and health state, language). In order to use SV in forensic casework, we have to take these factors into account. There are basically two ways to achieve this:

- a) We can collect speech signals from a so-called anti-speaker population, while maintaining the above factors according to the individual case, and set the outcome of the SV for the suspect in relation to the outcome for the other speakers.
- b) We can estimate the performance of the SV system for the particular set of factors given by the individual case and compare the error probabilities for the hypotheses H_0 and H_1 . This can be seen as case-specific calibration of the valuation scale.

Although approximately equivalent, these two ways are in practice very different, and their feasibility depends strongly on the type of SV used.

In this paper an approach of type b) is shown. It is not meant as a general solution to the forensic SV problem, but the presented approach addresses a class of cases that are fairly frequent, namely the SV by means of speech signals from telephone tapping.

The outline of the paper is as follows: The next section presents the used SV system and in Section 3 the factors that influence the SV result are discussed. Section 4 shows how the result of the SV can be valued in terms of evidence. An overview of the major steps of our approach to forensic SV is given in Section 5. Finally, in Section 6 some advantages and drawbacks of the procedure are discussed.

2. The used SV system

In forensic casework, we use a text-dependent SV system that compares two identically worded speech signals. This comparison is based on dynamic time warping

sampling window	8 kHz / 8 bit log PCM
preemphasis	37.5 ms Hamming / 15 ms shift
features	-0.9
DTW	14 LPC ceps / cep mean subtracted
	endpoints mismatch: max 75 ms
	slope constraint: 0.5, 2
local distance	Euclidean dist of cepstral vectors
global distance	mean of intensity-weighted local distance

Table 1: Parameters of the text-dependent SV

(DTW) of the two corresponding cepstral contours (details are given in Table 1). The minimum accumulated distance resulting from DTW, normalized to the duration of the utterance, gives some indication, if the two speech signals originate from the same speaker or not.

It goes without saying, that such a distance value d_0 can only be interpreted in comparison to statistically representative sets of self-distances d_s and cross-distances d_c , as shown in Figure 1. Self-distances have been evaluated from speech signal pairs where both signals are known to originate from the same speaker. In contrast to that, cross-distances have been calculated from the signals of different speakers.

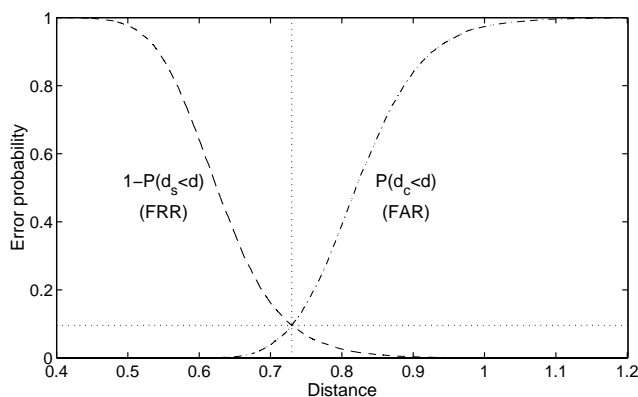


Figure 1: Distributions of some 23'000 self- and cross-distances for about 5 s long speech signals from male speakers; one of the signals has been recorded from a fixed telephone (F), the other one from an arbitrary one (X), i.e. from a fixed or a mobile telephone.

3. Influence of various factors

Speech signals, even if identically worded and spoken by the same speaker, exhibit a high degree of variability. There are various reasons for that, some related to the speaker (e.g. mental conditions and health state), others are external factors such as the recording and transmission conditions. As a consequence, the self- and cross-distances are generally distributed across largely overlapping intervals as can also be seen in Figure 1.

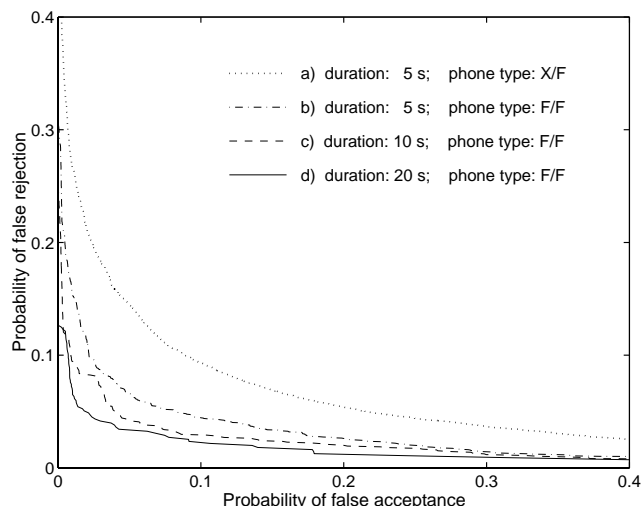


Figure 2: ROC curves of the used SV system for several conditions: a) corresponds to the situation in Figure 1; curve b) results, if both signals have been recorded from a fixed telephone; for curves c) and d) the mean duration of the signals has been increased to 10 and 20 s, resp. (the other factors are equal to b).

By means of ROC curves as shown in Figure 2, it can easily be demonstrated that such factors greatly influence the performance of the SV system. Therefore, the conclusion to be drawn from a SV depends not only on the evaluated distance between the incriminating signal and the test signal, but also on the set of influencing factors that characterize this case.

In order to compensate for the influence of these factors, we want to calibrate the scale that evaluates the output of SV system for the individual case (cf. Section 1), i.e. we estimate the ROC curve (or the distributions of the self- and cross-distances), given the set of influencing factors of the case at hand.

For a text-dependent SV system this can be achieved as follows: We use a database with speech signals that have been labeled in terms of factors such as utterance length, signal quality, telephone type, speaker's sex, etc. From this database, we select pairs of identically worded speech signals that also match the set of factors. Of course we know for each of these pairs if the speakers are identical or not and therefore can split the distances computed from these pairs into self- and cross-distances and determine the corresponding distributions.

It is important to note that the signals within a pair must be identically worded, whereas this is not required between pairs (we want to compute a distance for each pair, but not between pairs). With the assumption that for long enough pairs the evaluated distance is virtually independent of the actual sequence of phonemes, the distance is also independent of the actual wording and even independent of the language. In an individual case, the text-dependent SV system can therefore be calibrated with

speech signals that do neither match the wording nor the language of the incriminating recording.

4. Estimating the weight of evidence

As mentioned in Section 1, the desired outcome of forensic SV is the LR of the hypotheses H_0 : “same speaker” vs. H_1 : “different speakers”. In an individual case, the application of the SV system described in Section 2 results in a single distance value d_0 only. Therefore, we have to estimate the LR for this distance d_0 , given the particular set of influencing factors.

For practical reasons, the logarithm of the LR, i.e. the LLR will be used in the sequel. The LLR in function of the distance d is computed from the probability densities $p(d|H_0)$ and $p(d|H_1)$:

$$LLR(d) = \log \frac{p(d|H_0)}{p(d|H_1)} \quad (1)$$

Densities can be computed from distributions by differentiation, but in the case of continuous distributions only. For experimentally evaluated distributions² as shown in Figure 1, the densities can be estimated by means of a histogram which needs an interval size to be specified. There is a trade-off between smaller intervals which make the density curves more noisy and larger intervals which make the curves smoother but also reduce the resolution. Even for large sets with more than 10'000 distances the resulting densities and the corresponding LLR curve get still very noisy, as can be seen in Figure 3.

In order to circumvent this problem, we have investigated an alternative measure, namely the LER (log error ratio)

$$LER(d) = \log \frac{FRR(d)}{FAR(d)} = \log \frac{1 - P(d_s < d)}{P(d_c < d)} \quad (2)$$

that is directly computed from the distributions. But what's the relation between LLR and LER?

Experiments with continuous densities have shown, that in a limited range around the EER point (equal error rate), i.e. the point where $FRR(d) = FAR(d)$, the curves of $LLR(d)$ and $LER(d)$ are fairly similar. Generally, the absolute value of $LER(d)$ is somewhat higher than the absolute value of $LLR(d)$.

It is currently not possible to give a valid compensation of this small overrating. For the sake of consistency, however, we prefer to rely on the LER, because it seems to be much better to accept a small bias than a large variance on the numeric result of the SV.

²Although the distributions in Figure 1 are extremely smooth, they cannot be differentiated, because the points constituting the curve are equidistant in vertical direction. In horizontal direction neighboring points can be arbitrarily close to each other.

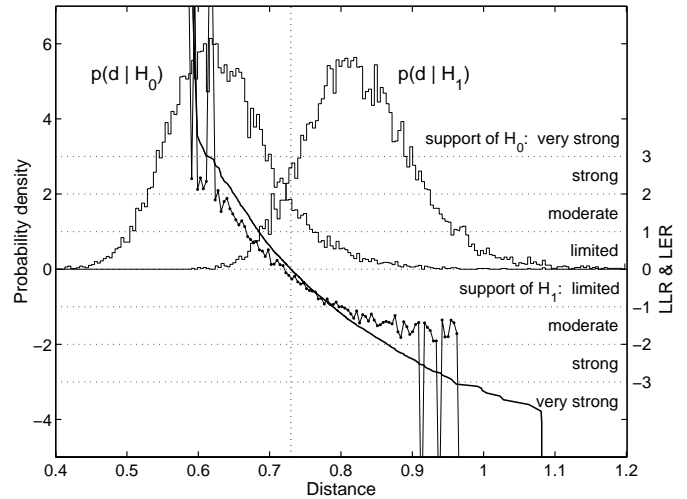


Figure 3: The self- and cross-distances shown in Figure 1 as distributions are represented here as densities $p(d|H_0)$ and $p(d|H_1)$ (stairstep graphs). From these densities the LLR curve³ (thin line with dots) can be computed using equation (1). Similarly the LER curve (strong curve) is evaluated according to equation (2). On the right hand side, the ranges of support according to [2] are set.

5. The practical procedure

In order to apply the SV in an individual case, we proceed roughly as follows (for more details, see [4]):

5.1. Recording test signal from the suspect

For practical reasons, the test signal from the suspect is recorded with telephone-based equipment. Thereby the suspect sits typically at a phone in a police office and repeats the prompted utterances from a remote computer at the investigators office. Thus the computer plays an utterance and in turn records the repetition of the suspect.

Since we use a text-dependent SV system, the test signal and the incriminating signal have to be identically worded. Therefore, the utterances to be prompted have to be selected from the incriminating speech recording. We do not use segments of this recording directly, however. The prompts are preparatively spoken by another person for two reasons: Firstly, the intelligibility of the incriminating signal is often fairly bad and secondly, it is better if the suspect does not hear his/her own voice.

The recording is carefully supervised on both sides in order to achieve signals with identical wording and of optimal quality. In particular each utterance with audible ambient noise, breath noise or amplitude overload has to be repeated.

It has to be emphasized, that this manner of test signal recording, i.e. the suspect has merely to re-

³The LLR curve has been drawn only in a limited range around the EER distance which is indicated by the vertical dotted line; the omitted part is flipping up and down and is therefore completely useless.

peat prompted utterances, is very important, because two problem sources can be omitted: The incriminating recording has not to be transcribed⁴ and the suspect has not to read out.

5.2. Computation of the distance

In the next step, the recorded test signals are manually inspected and auditorily compared with the corresponding segments of the incriminating recording. Every pair of segments exhibiting differences in wording or noise is shorted accordingly. This inspection is not done by the investigator, but by a person with good auditory skills (actually a musician).

Now the text-dependent SV system is used to compute the mean distance d_0 over all speech segment pairs.

5.3. Valuation of the distance

In the last step, we estimate the characteristics of the SV system, i.e. the probability distributions of self- and cross-distances, given the set of influencing factors of the case at hand. For that purpose we use an appropriate database, as has been explained in Section 3. Finally, $LER(d_0)$ is computed using equation (2).

6. Discussion

In an individual case, the appropriateness of the sketched approach to forensic SV depends primarily on finding the correct system characteristics (i.e. ROC curve or distributions of self- and cross-distances) that allows to value the computed distance d_0 between the incriminating signal and the test signal from the suspect.

Given a large database as outlined in Section 3, the only problem is to select the correct factors that characterize the case at hand.⁵ Some of the factors (e.g. speaker's sex or fixed/mobile phone) are discrete and are therefore easy to be figured out for an individual case. Other factors, however, are neither discrete nor easily quantifiable, but clearly affect the SV. Such a factor, a very important one, is the signal quality.

The signal quality includes various aspects such as level of ambient and transmission noise, signal degradation caused by low-quality microphone, amplitude limiting and signal coding (the latter being mainly a problem of mobile phones), breath noise, and last but not least the quality of articulation. We have currently no suitable method to estimate this overall signal quality reliably.⁶

⁴Very often a precise transcription is simply not practical, because of spontaneous conversation, mostly in a foreign language or even in some dialect. Only phoneticians would be able to produce such a (phonetic) transcription, but then the suspect would not be able to read it.

⁵In contrast to other researches (see e.g. [3]), our intention is not to use the most general, but the most specific SV system.

⁶We have subjectively rated and annotated the overall quality of the signals in our database. This information is currently used for our research only.

In forensic casework we handle this issue as follows: Since the recording of the test signal is fully under control of the investigator (cf. Section 5), good quality of the test signal is guaranteed. In contrast to this, the incriminating signal is often of low or even very low quality.

Experiments with our database have shown, that there is a clear dependency between signal quality and the computed distance. Provided the test signal being of good quality, the distance gets higher the more the quality of the "incriminating" signal drops. This was to be expected, because the distance between to similar signals generally is increased by adding noise to one of them.

The consequence of the currently unsolved quality problem for the forensic casework is, that the low quality of the incriminating signal reduces the weight of evidence. In other words: the quality problem takes effect in the sense of "in dubio pro reo".

7. Conclusions

The presented procedure of forensic SV has proved practical and accurate, because the valuation scale for individual cases can be estimated by means of a common database. Therefore, instead of recording speech signals from a number of anti-speakers for each individual case and estimate the LLR, we can use the database to determine the LER of the SV system for the individual case.

Currently, not all individual cases are sufficiently well represented in the database, and consequently not all cases can be handled yet. Nevertheless, the presented procedure has been applied in a considerable number of forensic investigations.

8. Acknowledgement

This work was partly supported by the Swiss National Science Foundation in the framework of NCCR IM2.

9. References

- [1] C. G. G. Aitken. *Statistics and the Evaluation of Evidence for Forensic Scientists*. John Wiley & Sons, Chichester, 1995.
- [2] I. W. Evett. Towards a uniform framework for reporting opinions in forensic science casework. *Science and Justice*, vol. 38:pp. 198–202, 1998.
- [3] J. W. Koolwaaij and L. Boves. On the use of automatic speaker verification systems in forensic casework. In *Audio- and Video-based Biometric Person Authentication*, pages 224–229, Washington, 1999.
- [4] B. Pfister. Personenidentifizierung anhand der Stimme. *Kriminalistik*, 55. Jahrgang, Heft 4, (Hüthig Verlag, Heidelberg), April 2001.