Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation

David A. van Leeuwen[1] *and Jos S. Bouten*[2]

[1]TNO Human Factors,
Soesterberg, the Netherlands.
vanLeeuwen@tm.tno.nl

[2]Netherlands Forensic Institute,
Rijswijk, The Netherlands.
j.bouten@nfi.minjus.nl

Abstract

In this paper we report on the results of the NFI-TNO speaker recognition evaluation held in 2003. The speech material used in this evaluation has been obtained from wire-tapped recordings from real police investigations in the Netherlands. In total six experiments were carried out, one main experiment in Dutch, one experiment in which speech lengths were systematically varied, three language dependence experiments, and one experiment evaluating a proposed forensic procedure for providing evidence in court cases. The lowest equal error rate of all systems was 12.1 % in the condition using 15 seconds test segments and 60 seconds training segments.

Introduction

In 2001, the Netherlands Forensic Institute NFI and the Netherlands Organization for Applied Scientific Research TNO started a joint project, which studied the application of automatic speaker recognition in a forensic context. The intention was to determine the state of the art of text independent automatic speaker recognition systems, and the possibility of using the results of such a system for investigative purposes in police inquiries. The evaluation material consisted of wire-tapped telephone recordings obtained from real police investigations. The project found 14 partners who were willing to participate in the evaluation. Thirteen partners submitted results of which 12 were correct according to the evaluation rules. Eleven agreed to be named in this paper, they are the ELISA consortium, the Israel Police Electronics and Computer Laboratory, LIMSI-CNRS, MEDAV GmbH, MIT Lincoln Laboratory, Roland Auckenthaler, Siemens AG, Universidad Politécnica de Madrid, and the Universities of Patras, Sheffield and West Bohemia. The evaluation was modeled after the yearly evaluations organized by NIST [1], as a one-speaker detection task.

NFI has a lot of experience in the forensic sciences and their speech group treats court cases on issues of authenticity of recordings and of speaker verification. Furthermore, NFI seeks to design and implement new methods and technologies like automatic speaker recognition for police agencies and the Dutch Ministry of Justice. TNO conducts research in a wide range of fields, and their speech group has experience in the evaluation of speech systems (EU sponsored ESPRIT SAM and SQALE projects), and has participated in the NIST evaluation with their own system.

The evaluation was held in the late fall of 2003. Because of the nature of the speech material, partners were obliged to adhere to an agreement that limited storage, usage and distribution of the speech material.

Speech material

The speech material used in the evaluation was taken from real police investigations. This was done in order to obtain field data and to get as close as possible to a real forensic application. It consisted of excerpts from 1809 wire-tapped cellular GSM to GSM telephone conversations recorded over a 23-month period. All speakers were males. The telephone line quality varied between recordings from excellent to moderate (extremes at the lower end were omitted). The telephone handsets used were unknown. The level and nature of background noises of the material varied and included slight room reverberations, music in the background and in some cases background speakers (mostly children playing in the background). The speaking style was spontaneous. Emotions varied between recordings from relaxed (frequent) to stressed (rare). Laughter, shouting and whispering were omitted. The distribution of these parameters among speakers was not homogeneous. The range and distribution of recording dates between speakers varied. The material was edited by NFI in order to select a single conversation side (most recordings contain both sides summed) and to make the material anonymous. Care was taken in editing so that no acoustic artifacts were introduced. Signalling noises in the telephone recordings were removed but speaking pauses were not edited. Meta-information (telephone number, observations, content, etc.) was used to obtain the true identities of the speakers.

In forensic cases the questioned recording is often of limited length; the evaluation test therefore concentrated on smaller test- and training segments than in the NIST evaluation. We also wanted to study the effect of training and test-segment length on the performance of the systems. Given a typical police investigation, and the data that was available for this evaluation, we chose excerpts of conversations for training segments in the range 30–120 seconds. We included conditions where training data was gathered from different conversations. Excerpts from different conversations were not concatenated into a single speech file. This allowed for systems to perform conversation specific normalization on the data. The test samples were collected from a single conversation. Because the recordings were taken from a set of uncontrolled telephone conversations the distribution of speaking time per speaker varied from several seconds to several hours. The requirement to include excerpts of 120 seconds of training material was the most difficult to meet. We found only 40 speakers (Dutch and English) in a set of 129460 recordings.

The evaluation was truly text independent. No transcription of training or test speech was given, although automatic means for determining the contents were allowed. Most of the data was in the Dutch language, a smaller part was in English and in Sranan Tongo and Papiamento. Speakers from different languages were not necessarily distinct persons. Although the language identity was only given for training segments, automatic determination for test segments was allowed.

Evaluation setup

The evaluation was modeled after the yearly NIST evaluations [2]. The task was a one-speaker detection task, where a speaker recognition system was requested to decide whether or not a given test speech sample was uttered by a given target speaker. For all target speakers model training speech material was supplied. With the decision, also a single score was requested, which should be higher when the likelihood that the test speaker is the target speaker is higher. Partners could submit results from more than one system, but had to indicate which was their primary system. In this paper we will only report on the primary systems.

Evaluation Measures

The speaker recognition systems were evaluated in terms of two measures: a detailed Detection Error Tradeoff (DET) curve, indicating the tradeoff between false alarms and misses [3], and a detection cost, based on actual decisions. The Detection Cost Function weights the false alarms (FA) and miss probabilities, according to cost parameters C_{FA} and C_{miss}, and the *a-priori* probability of a target speaker $\Pr(\text{target})$. The detection cost function is

$$\begin{aligned} C_{\mathrm{det}} =& C_{\mathrm{miss}} \Pr(\text{miss} \mid \text{target}) \Pr(\text{target}) \\ &+ C_{\mathrm{FA}} \Pr(\text{FA} \mid \text{non target}) \big(1 - \Pr(\text{target})\big), \end{aligned} \tag{1}$$

as used in NIST evaluations. We have used $C_{\mathrm{FA}} = 10 C_{\mathrm{miss}}$, $\Pr(\text{target}) = 50\,\%$, which leads to only a slightly different operating point from NIST evaluations. The costs are normalized so that the detection cost of a trivial system rejecting all test segments is equal to one, by setting $C_{\mathrm{miss}} = 2$.

Experimental design

With the opportunity to obtain results from so many automatic speaker recognition systems on real forensic speech data, and given the fact that the data could be used for a limited period in time due to its nature, we have tried to make an expermental design allowing as many experiments as reasonably possible on the scores returned by the partners. We have tried to include a structured variation of the following parameters in the design:

Model training speech length. In some forensic cases the amount of speech material can be fairly short. We chose the model lengths at 30, 60 and 120 seconds. The minimum is a balance between the request from forensics to deal with very little training material and the speaker recognition partners who are more experienced with 120 second model training samples from NIST 2001–2003, which is the maximum of this parameter.

Test sample speech length. A piece of speech material available for investigative purposes can be very short, and hence we chose to vary the test speech lengths between 7, 15 and 30 seconds. The range of this parameter is actually not much different from what is customary in the NIST evaluations (typically varying between 3 and 45 seconds).

Number of sessions for model training. With a single session (telephone conversation) the difference between a channel effect and a true speaker characteristic is hard to make. With training material recorded from several sessions this separation should be more easy to make. In order to investigate this effect, we chose model training speech material to be collected from 1, 2, or 4 different telephone conversations.

The experiment was set up around a 'central' condition of 60 seconds of model training material taken out of 1 session, and 15 seconds of test material. In order to obtain good and reliable results, as many trials as possible were defined from the available speech material in these conditions. In prac-

tice, the number of trials was limited by the maximum amount of target trials that could be defined, 521. 9676 non-target trials were added. Although the number of telephone conversations from which the selection of model and test speech could be chosen was very high, the speech length conditions from one conversation limited the number of target speakers to 22. An additional 28 speakers were used for non-target trials (both for test and model speech). This main experiment is numbered 1.

In experiment 2 the three parameters (number of sessions and length for the model training speech, and length for the test speech segment) were varied systematically, allowing selection of virtually any combination in analyzing the results. Again, the source of the speech material limited the number of target trials to 540, a complete design of the 27 conditions with the 20 speakers for which enough material could be found. The trials were augmented with 9360 non-target trials in equally distributed parameter conditions.

Apart from these systematic variations, we included some other experimental conditions:

Spoken language. Because most text independent modeling techniques are also essentially language independent, it is interesting to perform the same test in various languages spoken. Because of the distribution of languages in the source speech material most experiments were set up in Dutch, but also a limited English test was included. This experiment, numbered 3, contained 21 target speakers, two additional non-target test speakers and one additional non-target model. This experiment contained only 126 target trials and 2206 non-target trials.

Cross-language performance. Really interesting is the question how sensitive a text independent speaker recognition system is to difference in language between model training and testing. Such a test can only be performed for bilingual speakers. In our police investigation material we could find a very limited set of speakers speaking Dutch and one other language (mostly English). We therefore included several cross-language trials in the design. One class involved Dutch test segments tried against non-Dutch models (experiment 4) and one involved English test segments tried against Dutch models (experiment 5).

Forensic evidence procedure test. For using a speaker verification system in a forensic context a laborious procedure has been proposed that allows estimation of the likelihood ratio that a piece of evidence in the form of a speech recording supports the odds that a given suspect has produced the evidence [4, 5, 6]. This involves estimation of the within-source variation of the scores of the suspect speaker. In order to accommodate research in this area, a sixth experiment was defined using speech of 10 different speakers, each amounting to 5 separate model training utterances of 60 seconds each, and 6 separate test utterances of 15 seconds each. These test and model training segments were combined to obtain all possible target trials (287 in total, not all of the 300 combinations were possible), and augmented with 2353 non-target trials for each of the target test segments with other non-target models.

Evaluation procedure

The trials of experiments 1 through 6 were all mixed, and organized such that each test segment had to be tested against all relevant models in the experimental trials. The number of models per test segment happened to be a multiple of 11 so that the evaluation test file had a shape identical to the NIST one-speaker detection task. Model training data that needed to be combined from more than one session were identifiable as such by the speech file name. Order and identification of the test segments and model training segments were randomized, and some other measures were taken to hide any production artifacts that could identify the speech files. All speech files were archived in a single `tar` file, encrypted with a separate key for each partner with a symmetric cipher using the GNU Privacy Guard `gpg`. This file was burnt onto a CDROM and sent to each of the 12 partners at the beginning of the evaluation period of four weeks. Unfortunately, due to mail irregularities not all partners could start at the same moment and therefore the deadline for submission of results was postponed for a week. The result for each partner consisted of a score and decision for each trial. Each partner was requested to accompany their results with a short description of the system.

There were several evaluation rules that the partners were obliged to adhere to, such that results between partners can be compared in a meaningful way [7]. The most important one was that for obtaining a result on a trial, only speech material from that test and model segment was allowed to be used, and no other speech material from the evaluation. From the results and descriptions submitted, it could be deduced that on some occasions, results were not obtained according to this rule. More specifically, in all these occasions, the partner apparently had assumed that exactly one model identified a test segment, and hence an identification task was assumed (rather than detection). In the particular case of this evaluation, due to the mix of 6 different experiments with multiple target trials in a single set, this assumption was actually false. One partner had to withdraw his results, but all other cases could be resolved.

Evaluation results and analysis

The evaluation results have been processed and are presented in a standard way introduced in the NIST speaker recognition evaluations [3]. The result scores s from an individual system are separated in true-speaker and impostor trial groups, and the miss and false alarm probabilities are estimated using the cumulative distributions of the trial group scores. The tradeoff between P_{miss} and P_{FA} is then depicted in a parametric plot with a score threshold as implicit parameter. The scales of the graph are warped by the inverse of the cummulative normal distribution function. The graph is known as a DET-plot, and the scales are commonly addressed as 'normal deviate scales' [3]. Mathematically the warping function $Q(x)$ is related to the inverse error function $\text{erf}^{-1}(x)$ as $Q(x) = \sqrt{2}\text{erf}^{-1}(2x-1)$.

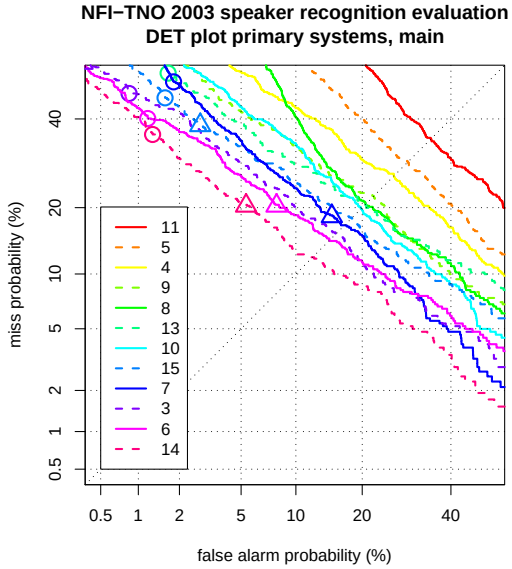

Figure 1. DET plots for all primary systems, for the first experiment. Experimental conditions were Dutch language, 60 sec model training segments obtained from one session, 15 sec test segments. Triangles and circles indicate the actual and minimum detection cost operating points.

Main result: experiment 1

In Figure 1 the results of experiment 1 for the primary system submission of the 12 partners are depicted in a single DET plot. One of the benefits of the normal deviate scale is that tradeoff curves are separated clearly and many curves can be compared in the same figure without cluttering it too much. A system performs better when the curve lies more to the lower left side of the plot (lower error probabilities). The systems are identified anonymously here as a number in the range 3–15. From visual inspection it might be concluded that system 14 is the best performing system.

Apart from scores, the partners were also requested to submit actual decisions for each trial. Making a decision is considered part of the speaker recognition strategy [1] but involves in practice estimating a threshold value based on other speech material than contained in the evaluation set. Since this evaluation contained no development test material on which such threshold could be estimated, the actual detection costs deviate a fair bit from the minimum detection costs. The latter costs are based on a post-evaluation choice of the threshold, minimizing the cost given the evaluation results. In Table 1 the actual and minimum detection costs are tabulated for the twelve systems.

Table 1. Actual and minimum detection costs for the twelve systems that submitted valid results, as well as the equal error rate (EER). Costs are calculated according to equation 1.

System	Actual	Minumum	EER (%)
3	0.582	0.551	15.5
15	0.661	0.613	18.1
14	0.739	0.489	12.1
13	0.742	0.687	20.5
9	0.754	0.744	22.2
10	0.772	0.705	19.8
4	0.959	0.819	26.3
8	0.977	0.969	20.6
6	0.996	0.519	14.4
7	1.669	0.679	16.9
11	2.280	1.004	35.0
5	7.176	0.995	29.5

There is a clear difference between actual and minimum detection cost for most systems, from which we conclude that estimating the threshold for this evaluation data was a difficult task. Systems 8, 9, and 3 are best in choosing the threshold, but it has to be remarked that the cost of system 8 is very close to trivially rejecting all speakers. The actual and minimum detection cost operating points are indicated in Figure 1 as a triangle and a circle, respectively. Due to the large differences in actual and minimum cost, the asymmetric cost balance between false alarms and misses, and the generally difficult task many of the detection operating points lie outside the graph area.

We have included the request for a decision in the evaluation setup because it influences the choice of normalization of scores. For instance, T-normalization of the scores tends to lead to tilting of the DET curve more horizontally around the equal error rate point. We have chosen the cost parameters such that the minimum cost point is virtually the same as in NIST tests so that system developers would have maximum benefit from experience and data obtained in these evaluations. We concentrate

in the remaining of this paper on the DET curve as a whole, and will use the Equal Error Rate point as a one-figure summary describing such a curve. The Equal Error Rate (EER) tends not to be influenced too much by 'DET tilting' optimizations, and can further be considered as representative if the exact purpose or prior probabilities of the application are not known.

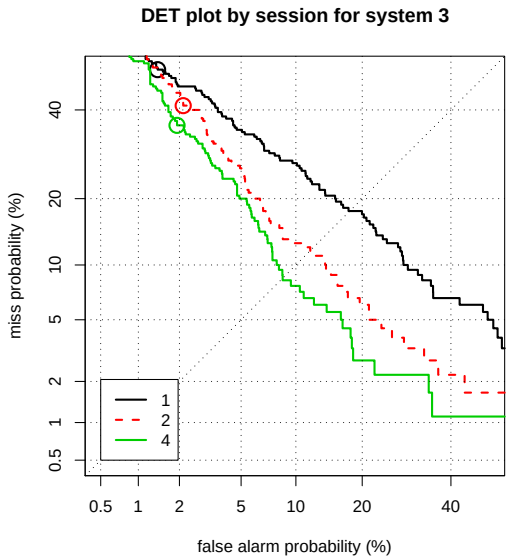

Figure 2. Effect of increasing the number of training sessions (1, 2, 4), while keeping the total training time constant. It should be noted that this result is *not* very typical of all systems, see Figure 3.

Parameter variation: experiment 2

The results of experiment 2 lead to a vast amount of data. For each system a DET plot could be generated for each of the 27 different conditions addressed. One of the problems with such detailed analysis is that the number of trials per DET curve becomes very small and hence the curves are noisy. Visually better curves are obtained by splitting the data for experiment 2 only along one varying parameter, either number of sessions, model training length or test sample length. This way trials that vary over the other two parameters are combined. Still, this leads to nine DET curves per system which still are not easily summarized. We have inspected DET plots such as Figure 2 for all systems and parameter variations. General remarks as a result of this investigation are:

Number of sessions. When the data of experiment 2 is split according to the number of sessions over which the model training data is obtained, we expected curves like those for system 3 in Figure 2, namely a general trend that the more different conversation sessions the model can be based upon, the less dependent the model will be on channel effects [1]. However, this system is quite unique in consistently obtaining lower EER with increasing number of sessions. In Figure 3 the development of EER versus number of training sessions is plotted for all systems, and the only consistent trend appears to be that systems benefit from more than one training session, but in many cases an increase from 2 to 4 sessions does not lower the EER, rather, in some cases the performance degrades. This might be an effect of very short training segments (with 4 sessions, there are trials for which each training session is only 7.5 seconds in length). A generalized linear model fit of the EER with systems and number of training sessions as factors reveals that an increase to 2 sessions lowers the EER on average by 3.6 %, while an increase from 1 to 4 only by 3.2 %.

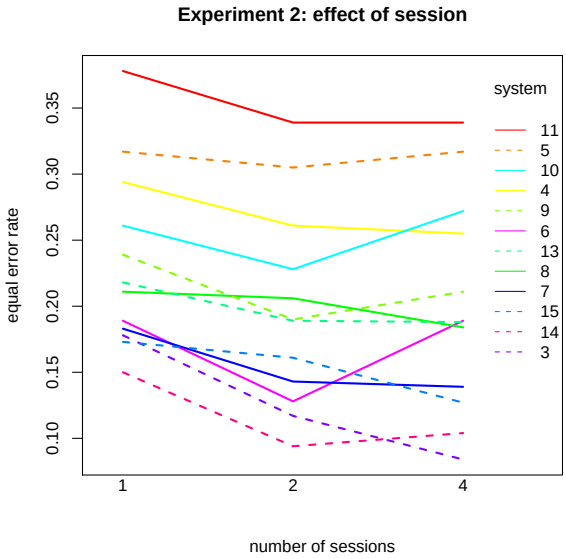

Figure 3. The effect of number of sessions in model training on the EER for all systems. Data points within a system are connected for visibility, using the same colour scheme as in Figure 1.

Model training length. Contrary to the number of sessions, the increase of model training length does lead consistently to better performance across most systems. A graph similar to Figure 3 shows more parallel lines, with the odd exception when the 120 second condition doesn't seem to add much to the performance. A typical example of the DET-curve trend for variation in model training length is given in Figure 4, for system 14. A generalized linear model fit finds a reduction of 1.5 % for each doubling of model training length.

Test utterance length. The influence of the test segment length is, as expected, that the longer the test segments, the better the performance [1]. In a linear model fit the effect is a lowering of 1.3 % of the EER for each doubling of test sample time, on

average. However, the effect is not as clearly visible for all systems as it is in Figure 5 for system 7 or reference [1]. Most of the 2.6 % gain in EER for 30 second samples with respect to 7 second samples is gained in the step to 15 second samples (2.1 %).

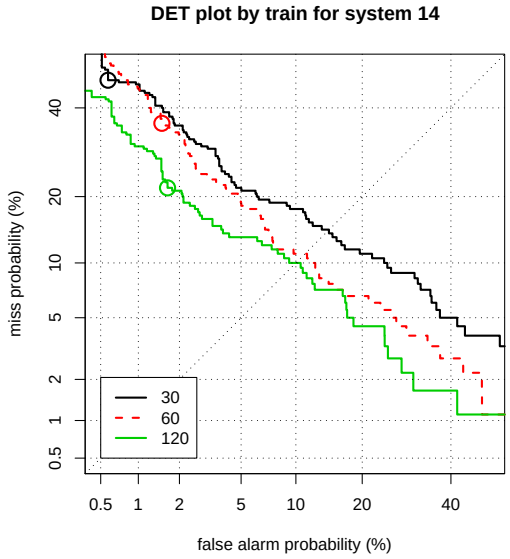

Figure 4. Effect of model training length (30, 60, 120 seconds) on EER, for system 14, the system with the lowest overall EER. This trend is quite general for most systems, with a few exceptions.

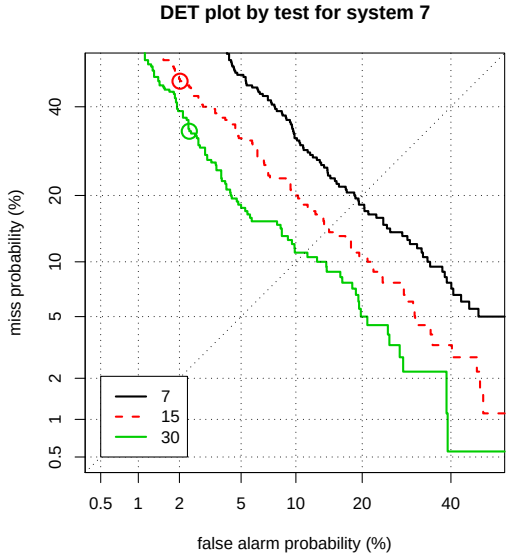

Figure 5. Effect of test sample length (7, 15, 30 seconds) on EER, for system 7, the system for which the effect is clearest.

It is tempting to run an analysis of variance on the complete design of varied parameters. However, one must keep in mind that a one-figure measure such as the EER is not a linear quantity with respect to mixing underlying trial distributions. For instance, the equal error rate based on all trials in experiment 2 for system 14 is 11.7 %, while averages of the equal error rates based on separation according to number of sessions, training length (as in Figure 4) and test length are 11.6 %, 11.8 % and 9.8 %, respectively, for this system. These numbers should then be compared to the more accurate number found in experiment 1 for this system, which is 12.1 %.

Language dependencies: Experiments 3–5

The experiments 3–5 can best be summarized in an interaction plot shown in Figure 6. Here, the EER for all systems is shown for the two main language conditions Dutch (experiments 1–2) and English (experiment 3), and the two cross language conditions: models trained in Dutch tested with another language (experiment 5: xl-dm) and models tested in Dutch trained in another language (experiment 4: xl-dt). The conditions are ordered from left to right in generally increasing EER. On average, changing the evaluation language from Dutch to English increases the EER by 4.3 %. For reference, the difference in average between the two Dutch experiments is only 0.4 %.

Despite the fact that the English test is much smaller in size and hence measurements are less accurate, we are quite surprised by this increase. We may assume that most systems have been developed with English data (e.g., for estimating a universal background GMM model), and that if there would be any language effect, it would be advantageous for English. That this is clearly not the case, might be related to the fact that some of the English speakers are non-native, and that they exhibit less consistency in their use of language and pronunciation, as is often seen in non-native speech research [8].

The cross language conditions give higher EER on average, 6.2 % for Dutch models and 9.0 % for Dutch test segments. Obviously these conditions must suffer from a non-native effect if that exists, for it is unlikely that all speakers in the test have been raised bilingually.

We can ask ourselves what the statistical significance is of the differences between the conditions. It is best to do this for an individual system, and use the underlying statistics that generate the DET curves to compare two DET curves with different conditions. We have not yet performed that analysis, which will require less trivial statistical techniques. At a more global level we may apply a conventional t-test to the average of the EER over all systems. Such a test concentrates on the difference in the data (in this case the language condition) rather than the difference in systems causing the performance of such a system to change. The role of the system is then comparable to that of a test subject in psychologi-

cal experiments. In Table 2 we have tabulated the p-values of a two-sided paired t-test for each of the language conditions.

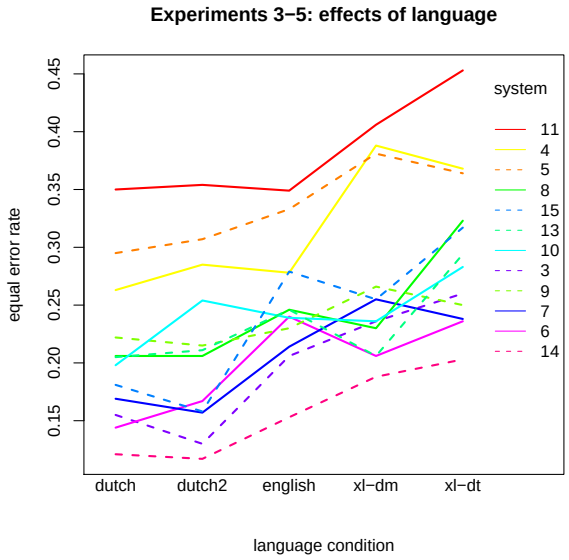

Figure 6. EER in experiments 1–5 for all systems. From left to right are the conditions Dutch from experiments 1 and 2, English, and two cross language condition. The condition 'xl-dt' are trials with Dutch test segments and models trained on other languages, the condition 'xl-dm' has Dutch models and test segments in other languages. Thus, the last two conditions contain target speakers that speak at least two languages.

Table 2. Values of p (the probability of incorrectly concluding that two distributions have a different mean) for the EER averaged over systems, for all pairs of conditions shown in Figure 6.

	dutch2	english	xl-dm	xl-dt
dutch	0.872	0.073	0.048	0.006
dutch2		0.007	0.001	$< 10^{-3}$
english			0.146	$< 10^{-3}$
xl-dm				0.036

The p-values give an indication of which of the conditions really make a difference. We use this wording, because the assumption that the distributions are Gaussian may not be valid and hence a non-parametric test would be more appropriate. A Wilcoxon signed rank test actually shows p-values which are very similar in that the same combinations border the traditional 5 % significance level. Again, it would be better to perform a more detailed test on the individual DET curves, but we were limited in how much we could analyze for this conference.

Proof in court: experiment 6

In a special issue of Speech Communications on speaker recognition, Champod and Meuwly [4] addressed the problems of interpreting miss and false alarm probabilities directly in the sense of evidence in a court trial. They proposed a forensic procedure for providing the judge in a court trial with a likelihood ratio to help him decide whether a suspect is the author (speaker) of a piece of evidence (speech recording) or not, using a speaker recognition system. Several efforts have since been made to evaluate the performance of automatic speaker recognition systems in such a procedure. Two distinct approaches have been taken in order to estimate the within-speaker distribution of scores of a (presumed) suspect. Meuwly and Drygajlo [6] used 6 models for the same suspect speaker, a single 'evidence' test sample, and 35 'control' test samples. The within speaker distribution is not assumed to be normal, and is estimated with the 6×35 trials. The evaluation uses 8 (simulated) suspect speakers. Gonzalez-Rodriguez *et. al.* [5] uses the Ahumada speech database with 103 speakers, but since only two utterances per speaker are available, the standard deviation of each of the speakers' within-speaker distribution is assumed to be the same, and determined by calculating the difference in the two target-test scores for all speakers.

With the availability of real forensic speech data we have taken the opportunity to set up an experiment to do a similar evaluation of systems with respect to their applicability in the procedure for producing evidence in a court trial. For 10 Dutch speakers we could find enough material to select 5 segments of 60 seconds for speaker model training, and 6 segments of 15 seconds for testing. In the procedure, we used one test sample in the function of trace or evidence material, and one model in the function of suspect speaker model. The evidence e is then defined as the score of the trace versus the suspect model. The scores of trials of the other suspect test samples versus the other suspect models give the within-source score distribution p_w. Finally, the trace is tried against a set of non-target models, forming the reference population, for obtaining the distribution of between-source scores p_b. The non-target speakers were selected from the available 50 Dutch speakers in the evaluation.

The ratio of the likelihoods that the evidence is obtained, given the hypothesis that the suspect is the perpetrator rather than somebody else, is called the likelihood ratio L, and is estimated as $L = p_w(e)/p_b(e)$. The evidence material thus determines at which score the probability density functions of the within- and between-source scores are sampled. We determined the probability density function $p_b(s)$ using kernel density estimation of the statistical program 'R,' and that of the p_w by assuming it to be a Gaussian distribution.

By performing this operation for traces with the same identity as the suspect, for each of the 10 speakers in the experimental set, and for each of the available test utterances as trace, we have 60 values for the likelihood ratio L for which we know it is true that the evidence is produced by the subject. We can repeat the same procedure by selecting some of the other speakers' test segment as trace material. For the likelihood ratios found for these traces, we know that the evidence is not produced by the subject. For a given likelihood ratio threshold L_T we can then evaluate in how many cases having $L > L_T$ the suspect has indeed produced the evidence, and in how many cases he has not. A graph of the development of these figures is called a 'Tippet-plot' [9]. In Figure 7 we have made such a Tippet plot for three system, that cover a wide range of performance.

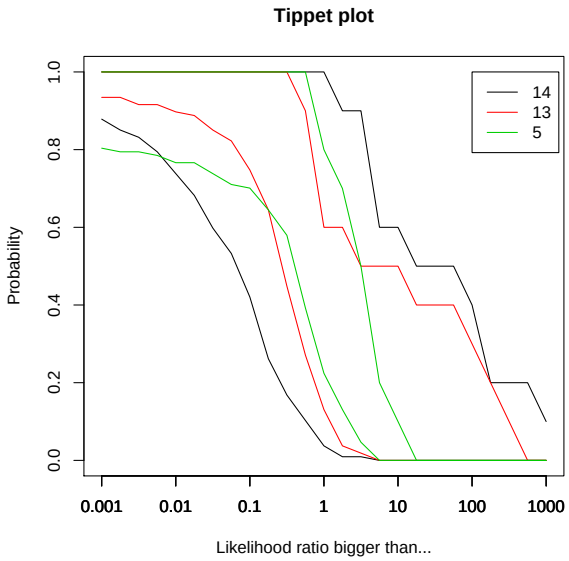

Figure 7. A Tippet plot for three systems. The curves on the upper/right indicate the probability that if the procedure gives a likelihood bigger than the threshold determined by the x-axis value, the evidence is produced by the suspect. The curves on the lower/left give the probability that the evidence is *not* produced by the suspect.

As might be expected we observe that systems with better DET performance also perform better in a Tippet plot, having the two curves further apart. Interesting to note is that the curves have maximum vertical separation around $L_T = 1$.

The precise procedure for producing a Tippet plot depends on various implementation details, such as whether or not to use all available test samples as trace material, and whether or not to use all model training segments as suspect model, which speaker models to use as reference database, etc. In our experience this has a non-negligible influence on the detailed curve of the lines in the Tippet plot. We have the feeling that it has not yet been established what the ideal procedure is, given the tradeoff between number of data points in the Tippet plot and statistically dubious re-use of speech samples in different roles.

Conclusions

We have performed several interesting experiments on the results of 12 different automatic speaker recognition systems obtained from a test with real forensic data. The experimental design allowed for structured analysis of the variation in test conditions, but the interpretation of the results was not always clear. The advantage of being able to work with realistic forensic data has as a downside that experimental conditions and numbers cannot be controlled. The absolute results should therefore be interpreted as a representative, but limited one-shot sample of the state of the art of text independent automatic speaker recognition systems.

References

[1] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. *Speech Communication*, 31:225–254, 2000.

[2] The NIST year 2003 Speaker Recognition Evaluation Plan. `http://www.nist.gov/speech/tests/spk/2003/index.htm`, 2003.

[3] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech '97*, pages 1895–1898, Rhodes, Greece, 1997.

[4] Cristophe Champod and Didier Meuwly. The inference of identity in forensic speaker recognition. *Speech Communication*, 31:193–203, 2000.

[5] J. Gonzalez-Rodriguez, J. Fierrez-Aguillar, and J. Ortega-Garcia. Forensic identification reporting using automatic speaker recognition systems. In *ICASSP*, 2003.

[6] Didier Meuwly and Andrzey Drygajlo. Forensic speaker recognition nased on a Baysian framework and gaussian mixture modelling (GMM). In *2001, a speaker Odyssee*, pages 145–150, June 2001. Crete.

[7] David A. van Leeuwen and Jos S. Bouten. The NFI/TNO forensic speaker recognition evaluation plan. `http://speech.tm.tno.nl/aso/evalplan-2003.pdf`, 2003.

[8] Sander J. van Wijngaarden. Intelligibility of native and non-native Dutch speech. *Speech communication*, 35:103–113, 2001.

[9] C. F. Tippet, V. J. Emerson, M. J. Fereday, F. Lawton, A. Richardson, L. T. Jones, and S. M. Lampers. The evidential value of the comparison of paint flakes from sources other than vehicles. *Journal of the Forensic Science Society*, 8:61–65, 1968.