Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 20 (2006) 128–158
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

NIST and NFI-TNO evaluations of automatic speaker recognition

David A. van Leeuwen [a,*], Alvin F. Martin [b], Mark A. Przybocki [b], Jos S. Bouten [c]

[a] *TNO Human Factors, Postbus 23, 3769 ZG Soesterberg, Utrecht, The Netherlands*
[b] *National Institute of Standards and Technology, Gaithersburg, USA*
[c] *Netherlands Forensic Institute, The Hague, The Netherlands*

Received 1 November 2004; received in revised form 1 June 2005; accepted 18 July 2005
Available online 15 August 2005

Abstract

In the past years, several text-independent speaker recognition evaluation campaigns have taken place. This paper reports on results of the NIST evaluation of 2004 and the NFI-TNO forensic speaker recognition evaluation held in 2003, and reflects on the history of the evaluation campaigns. The effects of speech duration, training handsets, transmission type, and gender mix show expected behaviour on the DET curves. New results on the influence of language show an interesting dependence of the DET curves on the accent of speakers. We also report on a number of statistical analysis techniques that have recently been introduced in the speaker recognition community, as well as a new application of the analysis of deviance analysis. These techniques are used to determine that the two evaluations held in 2003, by NIST and NFI-TNO, are of statistically different difficulty to the speaker recognition systems.

* Corresponding author. Tel.: +31 346 356 235; Fax: +31 346 353 977.
E-mail addresses: david.vanleeuwen@tno.nl (D.A. van Leeuwen), alvin.martin@nist.gov (A.F. Martin), mark.przybocki@nist.gov (M.A. Przybocki), j.bouten@nfi.minjus.nl (J.S. Bouten).

0885-2308/$ - see front matter © 2005 Published by Elsevier Ltd.
doi:10.1016/j.csl.2005.07.001

1. Introduction

Evaluations of text independent speaker recognition systems have been held regularly in the past decade (Przybocki and Martin, 1999; Martin and Przybocki, 2000; Doddington et al., 2000; Martin and Przybocki, 2001; Przybocki and Martin, 2002; Przybocki and Martin, 2004; Van Leeuwen and Bouten, 2004). The evaluations provide the developers of systems an opportunity to assess the quality of their system and inspire them to try out new approaches to the problem of speaker recognition. A leading role in the methodology and focus of the evaluation has been played by NIST and its sponsors. A co-operation with the Linguistic Data Consortium (LDC) has guaranteed regular new challenges with regard to the application domain while the LDC provided a constant quality of the evaluation databases.

Around 2002, two independent efforts resulted in the availability of completely new types of speech database for speaker recognition. The first database was collected by a co-operation between two Dutch parties, the Netherlands Forensic Institute (NFI) and TNO. It consisted of wire-tapped telephone recordings made by the Dutch police forces in police investigations. The second database is the MIXER corpus, collected by LDC, in which a multi dimensional design of controlled recordings of telephone conversations is implemented. Parameters that have proven to be important in earlier speaker recognition evaluations are systematically varied, such that the database now consists of data recorded with several microphones, in five languages, from different handsets and over several transmission lines. Both databases have been used in an evaluation, the former in what has been coined the 'NFI-TNO forensic speaker recognition evaluation' and the latter in the regular NIST evaluation in the year 2004.

The two evaluations differ on many points, such as size, language, design, and collection method. The most important difference is the type of data: On the one hand the NFI-TNO evaluation consists of genuine field data, collected in exactly the same way as it would be used in an application for police investigations, with speech uttered by people suspected of criminal activity, who in no way realized their speech was used for this kind of technology evaluation. The database is uncontrolled, several conditions are unbalanced, and the amount of material useful for a proper evaluation is limited. On the other hand NIST evaluations consist of well-controlled and well-balanced conditions, and vast amounts of speakers and speech. Every subject collected is keenly aware that their conversation is being recorded (although they only know it is for speech research purposes) so in a sense they can be viewed as co-operative subjects. Despite these apparent large differences, it is possible to analyze and compare both evaluations both qualitatively and quantitatively.

Meaningful evaluations are carefully planned. By providing explicit evaluation specifications, common test sets, standard measurements of error, and a forum for participants to openly discuss algorithmic successes and failures, the NIST and NFI-TNO evaluations have provided a means for recording the progress of text-independent speaker recognition performance.

Several relevant papers were presented at Odyssey 2004 The Speaker and Language Recognition Workshop in Toledo, Spain, including a paper on past NIST speaker recognition evaluations (Przybocki and Martin, 2004). The basic results of the NFI-TNO evaluation (Van Leeuwen and Bouten, 2004) and the design of the NIST 2004 evaluation (Przybocki and Martin, 2004) were also presented at Odyssey 2004, but in this paper we have the unique opportunity to present the results of both evaluations together in greater depth where the advance in evaluation methodology and speaker recognition performance will be made apparent.

The layout of this paper is as follows. First a recapitulation is made of the evaluation paradigm, and some notes on statistical analyses are made. Then the results of the NFI-TNO 2003 and NIST 2004 evaluation are presented and various performance factors are analyzed. Finally an attempt is made to compare the results of the NIST 2003 and NFI-TNO 2003 evaluations.

2. Evaluation paradigm

There are many similarities between the various evaluations held, despite the aforementioned differences. We will summarize the more important ingredients of the benchmark evaluations in general, showing the common ground and the specific differences.

Task. The speaker recognition system is evaluated in terms of a *detection task*. The question here is whether or not a given speech segment is uttered by a given speaker. There are several variants of this task defined: the (basic) *one-speaker* detection task, where the speech segment is known to contain only speech from a single speaker, and the *two-speaker* detection task, where both conversation channels are summed and the task is to detect if one of them has the identity of the given speaker. The *extended data* one-speaker detection task is similar to the 'basic' one-speaker detection task except that much larger amounts of data are available to training and test a model. In addition to the speech data, several bits of side-information that can be gathered automatically are made available to the system under test.

Evaluation set-up. The evaluation is carried out at the participating site's premises, for various practical reasons. A site is given a number of speech files containing material for building speaker models (training files) and speech material for testing (test files). The site is required to complete a list of *trials*, each specifying a test segment and a model speaker. A site returns for each trial both a *decision* whether or not the system declares the test speech to be uttered by the model speaker, and a *score*, a real-valued number that increases with the likelihood that the test speech is uttered by the model speaker. An evaluation consists of several thousands of trials.

Performance measure. Performance of a system's ability to detect a speaker is evaluated in terms of a cost function, where the costs for the two types of detection errors are given by the evaluator, as well as the prior probability of a target speaker occurring. This *detection cost* C_{det} (Doddington et al., 2000; Przybocki and Martin, 2004) can be defined as

$$C_{\text{det}} = C_{\text{miss}} P_{\text{miss}} P_{\text{target}} + C_{\text{FA}} P_{\text{FA}} (1 - P_{\text{target}}),$$

where C_{miss} and C_{FA} are, respectively, the costs of a *miss* and a *false alarm*, and P_{target} the detection prior probability. The detection error probabilities P_{miss} and P_{FA} are determined in the evaluation.

The NIST evaluation chose to assume a target poor application scenario (a priori probability of a match set to 1%) with the greater cost assigned to missing such a target ($C_{\text{miss}}/C_{\text{FA}} = 10$). This might be appropriate to searching for speakers of interest in an audio archive. (Note that this did not imply that the actual evaluation trials were as target poor as the supposed application.) For the NFI-TNO we have chosen a prior probability of 50% and $C_{\text{miss}}/C_{\text{FA}} = 0.1$, suggesting that 'a false accusation is worse than a missed perpretator.' This is a little bit misleading, because in a real forensic scenario setting the prior probability is left to the judge,

and the prior chosen here is just to indicate that the speaker detection system does not use such prior information.

The detection cost is normalized to be unity for a system trivially making the same decision irrespective of the test speech segment. The parameters determining the optimal operating point can be combined into a single parameter, the 'effective prior odds' (Bimbot et al., 2000; Brümmer, 2004)

$$\mathcal{O}_{\text{eff}} = \frac{C_{\text{miss}}}{C_{\text{FA}}} \frac{P_{\text{target}}}{1 - P_{\text{target}}}.$$

For NIST evaluations, the choice of cost parameters leads to effective prior odds of 1/9.9, while for the NFI-TNO evaluation these are almost identical, being 1/10. In Table 1, a summary of the evaluation parameters is given. For $\mathcal{O}_{\text{eff}} < 1$, as in these evaluations, the normalized detection cost reduces to

$$C_{\text{det}} = P_{\text{miss}} + P_{\text{FA}}/\mathcal{O}_{\text{eff}}.$$

The primary performance measure of detection is the *actual detection cost*, which is based on the actual decisions of the systems rather than the score information.

Qualitative measure. The detection potential of a system is indicated well by plotting a DET-curve (Martin et al., 1997) showing the Detection Error Trade-off between P_{miss} and P_{FA}. The DET curve is essentially a Receiver Operating Characteristic (ROC) with the axes warped according to the quantile function for the normal distribution (Van Leeuwen and Bouten, 2004). Formally, the DET curve evaluates the quality of the *scores* given by the system to the trials rather than the *decision*. The position of the DET curve is a *post-evaluation* quality measure because an operating point given by P_{miss} or P_{FA} *determines* the score threshold that had to be set in order to obtain these error probabilities. Still, the DET-plots are very useful for investigating differences in systems or conditions. Other post-evaluation measures include the *minimum detection cost* and the *equal error rate* (EER), which are both single valued summaries of the DET-curve.

Conditions. Throughout the years, there have been several different conditions investigated in the evaluations. Most notably, these have been the duration of the training and test speech utterances, the types of handset used, the telephone line type, the coding used, and the language spoken. The largest differences between the various evaluations have been in the choice of conditions. Thus, each evaluation can be said to have had a particular 'focus.'

Evaluation database. Closely related to the evaluation condition in focus is the speech database from which the evaluation is compiled. Almost invariably these databases have been recorded telephone conversations collected by LDC, starting in 1996 with the Switchboard I corpus and extending to the new MIXER corpus in 2004. Exceptions are the Spanish Ahumada database used in 2000 and 2001, the multi-microphone FBI voice database in 2002 and the Dutch

Table 1
The cost parameters for the NIST and NFI-TNO evaluations

Evaluation	C_{miss}	C_{FA}	P_{target}	$1 - P_{\text{target}}$	$\mathcal{O}_{\text{eff}}$
NIST	100	10	0.01	0.99	0.101
NFI-TNO	2	20	0.5	0.5	0.100

The normalization factor has been included in the cost parameters C.

Table 2
A comparison of the conditions of the various NIST and the NFI-TNO evaluation

Evaluation	NIST									NFI-TNO
	1996	1997	1998	1999	2000	2001	2002	2003	2004	
1 speaker detection	•	•	•	•	•	•	•	•	•	•
2 speaker detection				Test	Test	Test	Test	Training and test	training/ and/or test	
1 sp. extended data						Dryrun	•	•	•	
ASR WER						20%	50%	50%	25%	
Database[a]	sw1	sw2p1	sw2p2	sw2p3	sw2p1 + 2	sw3p1	sw3p2	sw3p2	MIXER	NFI
Extended data						sw1	sw2p2 + 3	sw2p2 + 3	MIXER	
Alternative					Ahumada	Ahumada	FBI			
Language/ region	English USA	English Mid-Atlantic	English Mid-West	English South	English/ Spanish	English/ Spanish	English	English	English/ multi-lang	Dutch/ other
Number of speakers	40	~400	~500	233	804	174	330	356	310	50
Line and coding	Land	Land	Land	Land	Land	GSM (primarily) land	CDMA	CDMA	Land cellular cordless	GSM
Training duration conditions[b]	2 m	2 m	2 m	2 m	2 m	2 m	2 m	2 m	10, 30 s	$\frac{1}{2}$, 1, 2 m
Conversation sides							1, 2, 4, 8, 16	4, 8, 16	1, 3, 8, 16	
Test duration conditions	3, 10, 30 s	3,10, 30 s	3, 10, 30 s	5–60 s	5–60 s	5–60 s	5–60 s	5–60 s	10, 30 s	7, 15, 30 s
Conversation sides									1	

ASR WER, automatic speech recognition transcripts, word error rate.
[a] sw*npp*, switchboard release *n* phase *p*, where $n = 3$ means 'cellular.'
[b] m, minutes; s, seconds, a conversation side is about 5 min.

Forensic wire-tapped speech database in NFI-TNO 2003. The type of speech database used can be seen as one of the most important differences between the various evaluations.

Rules. Common to all evaluation is a set of rules to which the sites have to adhere. These are defined in an *evaluation protocol* (see, e.g., NIST year 2004 Speaker Recognition Evaluation Plan; Van Leeuwen and Bouten, 2003), prior to the call for participation in the evaluation. Some of the most important rules are:

- Each trial should be treated independently of all other trials, and only information from the test and training segments referenced in the trial may be used by the system for producing the decision and score. (An allowed exception was the optional 'unsupervised adaptation mode' in the NIST 2004 evaluation).
- Manual interaction with the evaluation data, and listening to the speech, is not allowed.
- Publication of the evaluation results of *other* participating sites is not allowed.

The important characteristics of the past NIST and NFI-TNO evaluations are shown in Table 2. Note, that results in this paper concentrate on evaluations described in the last three columns.

3. Statistics

In order to be able to compare the performance of different systems within an evaluation, or different conditions for one system, or even different evaluations, it is necessary to perform statistical tests that assess the significance of an observed difference. In this section we will discuss the statistical techniques that are commonly used in the speaker recognition community, some of which are used in the remainder of the paper.[1]

3.1. Basic binomial quantities

The dependent variables that are measured in a speaker recognition evaluation are the proportion of trials in error, P_{miss} and P_{FA}, for the target and non-target trials, respectively. Under the assumption that all trials are independent, these error probabilities can be described as a binomial statistic, and hence the variance of the quantities are dependent only on the error probability, $\text{var}(P) = P(1-P)$ (Hays, 1963). This means that the *standard error* of the quantities is given by

$$s_{\text{miss}} = \sqrt{\frac{P_{\text{miss}}(1-P_{\text{miss}})}{N_{\text{tar}}}} \text{ and } s_{\text{FA}} = \sqrt{\frac{P_{\text{FA}}(1-P_{\text{FA}})}{N_{\text{non}}}},$$

where N_{tar} and N_{non} are the number of target trials and non-target trials in the evaluation. The standard error gives an impression of how accurate the determination of P_{miss} and P_{FA} is. If N is large, the normal approximation to the binomial distribution can be used, and the 95%

[1] We have included this section for reference in the speaker recognition community, and to make the paper more self-contained.

confidence interval follows from the quantile function of the normal distribution, evaluated at 2.5% and 97.5%, numerically ± 1.96. Thus, the confidence interval for P is $P \pm 1.96s$. In DET plots, it is customary to plot a box indicating the 95% confidence intervals around the actual C_{det} operating point.

For a typical minimum C_{det}, the value of $P_{\text{miss}} \approx 20 P_{\text{FA}}$, while in evaluations, $N_{\text{non}} \approx 10 N_{\text{tar}}$. This makes standard error of P_{miss} and P_{FA} of the same order of magnitude.

Several statistical tests exist for a comparison of two binomial values. We will use the 'two sample test for equality of proportions.' This test models the difference of the two sample proportions $P_1 = x_1/N_1$ and $P_2 = x_2/N_2$ as a normal distribution with zero mean and a binomial variance based on the common estimate $\hat{P} = (x_1 + x_2)/(N_1 + N_2)$. This procedure is analogous to the classic t test for normally distributed statistics.

Note that strictly speaking the assumption that the trials are independent does not hold: often the same test segment is used in several non-target trials, and the same speaker is usually used for several different test segments.

3.2. Error propagation

The main performance statistic in speaker detection is C_{det}, a linear combination of P_{miss} and P_{FA}. This in itself is not a binomial statistic, but the error in C_{det} can be determined using error propagation. If P_{miss} and P_{FA} are independent measures, then the error in C_{det} is given by

$$s_{\text{det}}^2 = s_{\text{miss}}^2 + s_{\text{FA}}^2 / \mathcal{O}_{\text{eff}}^2.$$

Here, the relative importance $\mathcal{O}_{\text{eff}}^{-1}$ of the false alarms to C_{det} can be appreciated.

The assumption of independence of P_{miss} and P_{FA} holds, because they are determined from independent distributions of trials. This may seem counter-intuitive, because a system generally uses a common threshold that determines both probabilities.

The standard error of C_{det} can be used to test for significant difference between two operating points on different curves, e.g., the minimum C_{det} for two different evaluation conditions.

3.3. DET confidence bandwidth

The confidence intervals around P_{miss} and P_{FA} can be drawn continuously around an entire DET curve by adding the values $\pm 1.96(s_{\text{FA}}, s_{\text{miss}})$ to each point $(P_{\text{FA}}, P_{\text{miss}})$ on the DET curve. Again, this can be used to assess the significance of the difference between two DET curves that have different underlying trial distributions, for instance curves from two different evaluations or different training conditions.

3.4. Comparisons between systems

When two or more systems are evaluated with the same set of trials, we can utilize more powerful methods for comparing the performance than the test of proportions. The basic idea is that individual decisions for each trial can be compared between two systems, rather than the sum of all trials. The McNemar test tabulates the correlation of correct and incorrect decisions between two systems. Thus, for P_{miss}, the target trials can be tabulated as follows:

Trial counts		System B	
		Correct	Incorrect
System A	Correct	N_{cc}	N_{ci}
	Incorrect	N_{ic}	N_{ii}

The test assesses whether or not the number of trials where systems A and B have a different decision, N_{ci} and N_{ic}, are significantly different. We cannot test performance on P_{miss} alone, and so the analysis has to be repeated for P_{FA} and the non-target trials. We can then stipulate that system A is performing significantly better than system B, if $N_{ic} < N_{ci}$ at a $p < 0.05$ significance level for both the analysis of target trials and non-target trials. The test can be used for the actual C_{det}, but also for other operating points such as minimum C_{det} or EER.

An alternative to the McNemar test is the sign test. Here, in order to partition the evaluation into several independent measures, C_{det} is 'decomposed' into 'speaker-specific' $C^{\text{s}}_{\text{det}}$:

$$C^{\text{s}}_{\text{det}} = P^{\text{s}}_{\text{miss}} + P^{\text{s,mod}}_{\text{FA}}/\mathcal{O}_{\text{eff}},$$

where $P^{\text{s,mod}}_{\text{FA}}$ is the false alarm probability over trials where the speaker s is the model speaker. For all speakers with a minimum number of test segments, say 10, $C^{\text{s}}_{\text{det}}$ can be used for a comparison between the systems in a sign test.

3.5. Analysis of variance

An analysis of variance (ANOVA) is capable of testing the effect of several conditions or factors in a single analysis. A standard ANOVA works with normally distributed dependent variables. The two basic parameters P_{miss} and P_{FA} are binomially distributed, and the variance does not have to be estimated from measurements but is determined by P. There exists an *analysis of deviance* which is similar to ANOVA but uses a generalized linear model of the test statistic and exploits the knowledge about the variance.

In an analysis of deviance, the test statistic is transformed by a link function, in our case the *logistic* function

$$\text{logit}(p) = \log\frac{p}{1-p},$$

which has the property that a change in odds ratio has an additive effect on the logit scale. This so-called *logistic regression* analysis forms a generalized linear model with a binomial response distribution and the link function. In a way, it can be seen as the generalization of the test for proportions, just as an ANOVA is a generalization of a t test. The effect of different factors (system, condition) can be analyzed and predictions for other conditions can be made. We will use this in comparing different evaluations.

The logistic regression analysis relies on a binomial statistic. It is therefore not easy to generalize it to C_{det}, which is a linear combination of two binomial statistics. A measure that could be thought of as a binomial statistic is the application-independent metric EER. It is the post-evaluation determined operating point where $P_{\text{miss}} = P_{\text{FA}}$, and we argue that the standard error of EER is determined by the measure with the lesser amount of trials, usually P_{miss}.

4. Designs of the NFI-TNO and NIST 2004 evaluations

In the Odyssey articles (Przybocki and Martin, 2004; Van Leeuwen and Bouten, 2004) the design and data collections paradigm for the two evaluations has been reported on quite elaborately. For completeness, we reproduce the most important issues here.

4.1. NFI-TNO evaluation

Speech material consisted of real field data, collected from recordings made using wire-taps for the purpose of police investigation. Because in Dutch forensic cases, the speech material is often of limited duration, the central durations condition in the evaluation was 60 s for model training and 15 s for test segments. A limitation of working with field data is that only 22 target speakers could be found for which enough material was available. An additional 30 non-target speakers were used in the test segments. All speakers were male and the transmission channel for all recordings was cellular GSM.

The evaluation consisted of several separate experimental conditions, concentrating on different aspects of speaker recognition. The main condition was a general performance evaluation, while other conditions investigated the influence of specific factors such as speech duration and spoken language. All sites participated in all conditions.

The experimental condition in which the effect of speech duration was investigated was set up as an orthogonal design of the variation of three parameters, each sampled with three levels. Each of the 27 conditions thus generated were evaluated using target trials from 20 speakers and augmented with approximately 350 non-target trials (see Section 5.2.1).

4.2. NIST 2004 evaluation

All speech material was taken from the MIXER corpus collection. In total, 310 target speakers occurred in the evaluation, both male and female, and 3426 conversation sides were used for training and 1176 for testing. The factors for speech duration and the one/two-speaker detection task were investigated in a 4×7 design of conditions, see Table 3. Three test and six training conditions were 'one speaker' conditions, where the speech from one side of the telephone conversation was extracted from the recording. One condition in test and training was a 'two speaker' condition,

Table 3
Design of duration and summed channel condition

Test segment condition	Training segment condition						
	10 s	30 s	1 side	3 sides	8 sides	16 sides	3 conv. (2sp)
10 s	10	7	10	4	4	4	3
30 s	6	8	16	4	4	4	3
1 side	7	8	**24**	7	10	6	5
1 conv. (2sp)	3	3	5	3	3	4	6

Numbers indicate how many sites participated in the condition. The primary condition is indicated in bold type. The bottom and right margin are 2-speaker detection task conditions.

where the speech from both conversation sides is summed.[2] Each condition consisted of a full set of evaluation trials. Sites were free to run any of the conditions, but it was compulsory to run the *primary condition* of one conversation side speech for model training, and one conversation side for testing.

All trials in the evaluation had training and test speech segments obtained from different telephone numbers, presumably different handsets. Other factors were included in the design of each experimental condition, such as spoken language and transmission type. It should be noted that the spoken language for all training segments was given.

5. Results and analysis of the NFI-TNO and NIST evaluations

Although the basic results of the NFI-TNO evaluation have been reported in Van Leeuwen and Bouten (2004), we will extend the results with additional statistical analyses here. The results of NIST 2004 have not been published before, and we will integrate the NFI-TNO results and analysis with the NIST results where applicable.

Twelve partners submitted correct system results to NFI-TNO evaluation, 24 sites participated in NIST 2004. The systems are identified anonymously here as a number, there is no correlation between the numbers used in the two evaluations. We will only report on their primary system submission.

The difference between the number of speakers and trials in both evaluations has led to a slightly different statistical analysis. While for NFI-TNO we need an overall system analysis in order to obtain enough statistical power to show effects, the power of the NIST evaluation is generally high enough that systems can be investigated individually.

5.1. Overall results

In Tables 4 and 5, the actual and minimum detection costs are tabulated for all systems, along with the equal error rate. We have indicated the standard error of the measures as well. The tables are ordered according to actual decision point.

In Figs. 1 and 2 the overall results are depicted in a single DET plot. The actual and minimum decision points of Tables 4 and 5 have been indicated as boxes and circles. The fairly large confidence intervals are the result of the relatively low number of target trials.

For many systems, there is a large difference between actual and minimum detection cost, especially for the NFI-TNO evaluation. From this we conclude that estimating the threshold for these evaluations was a difficult task. The reason may be that for both evaluation there was no development test speech material available within the same data collection. 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 actual detection operating points lie outside the graph area.

[2] 'One speaker' is sometimes also referred to as 'one side' or 'four wire' – a term from the analogue recording days – and 'two speaker' is also referred to as 'one conversation' or 'two wire.'

Table 4
Actual and minimum detection costs for the NFI-TNO evaluation, as well as the equal error rate (EER)

System	Actual		Minimum		EER (%)	
	C_{det}	SE	C_{det}	SE		SE
T1	0.582	0.023	0.551	0.024	15.5	0.8
T2	0.661	0.027	0.613	0.025	18.1	0.9
T3	0.739	0.029	0.489	0.024	12.1	0.7
T4	0.742	0.023	0.687	0.025	20.5	0.9
T5	0.754	0.025	0.744	0.025	22.2	0.9
T6	0.772	0.020	0.705	0.023	19.8	0.9
T7	0.959	0.009	0.819	0.022	26.3	1.0
T8	0.977	0.016	0.969	0.015	20.6	0.9
T9	0.996	0.033	0.519	0.024	14.4	0.8
T10	1.669	0.040	0.679	0.026	16.9	0.8
T11	2.280	0.043	1.004	0.003	35.0	1.1
T12	7.176	0.047	0.995	0.003	29.5	1.0

Standard errors (SE) are indicated. The number of trials in this condition were $N_{\text{tar}} = 521$ and $N_{\text{non}} = 9676$. Here, $\mathcal{O}_{\text{eff}}^{-1} = 10$.

Table 5
Actual and minimum detection costs for the 24 systems in the NIST 2004 evaluation, as well as the EER

System	Actual C_{det}	Minimum C_{det}	EER (%)	
				SE
S1	0.423	0.325	7.9	0.6
S2	0.423	0.421	12.1	0.7
S3	0.504	0.478	12.7	0.7
S4	0.524	0.518	11.5	0.7
S5	0.548	0.532	14.6	0.8
S6	0.557	0.512	13.6	0.8
S7	0.564	0.537	15.8	0.8
S8	0.578	0.386	11.1	0.7
S9	0.587	0.553	14.1	0.8
S10	0.604	0.575	13.9	0.8
S11	0.609	0.308	8.3	0.6
S12	0.625	0.544	14.1	0.8
S13	0.630	0.579	17.1	0.8
S14	0.636	0.537	14.9	0.8
S15	0.817	0.627	17.4	0.8
S16	0.932	0.885	25.2	1.0
S17	0.947	0.940	28.0	1.0
S18	1.135	0.910	31.0	1.0
S19	1.215	0.579	15.0	0.8
S20	1.341	0.962	28.0	1.0
S21	2.348	1.000	41.5	1.1
S22	4.645	0.988	39.6	1.1
S23	5.280	0.997	37.3	1.1
S24	9.900	0.643	16.4	0.8

Standard errors (SE) are indicated for the EER. The number of trials in this condition were $N_{\text{tar}} = 568$ and $N_{\text{non}} = 4634$. Here, $\mathcal{O}_{\text{eff}}^{-1} = 9.9$.

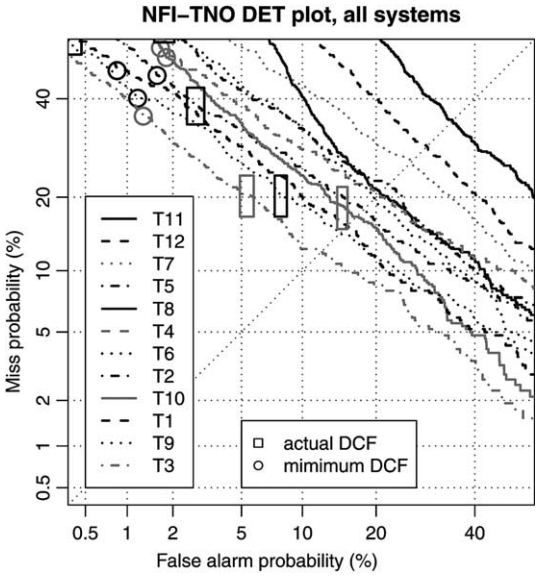

Fig. 1. DET plots for all primary systems, for the NFI-TNO evaluation. Experimental conditions were Dutch language, 60-s model training segments obtained from one session, 15-s test segments. The boxes and circles indicate the actual and minimum detection cost operating points, where the boxes represent the 95% confidence intervals.

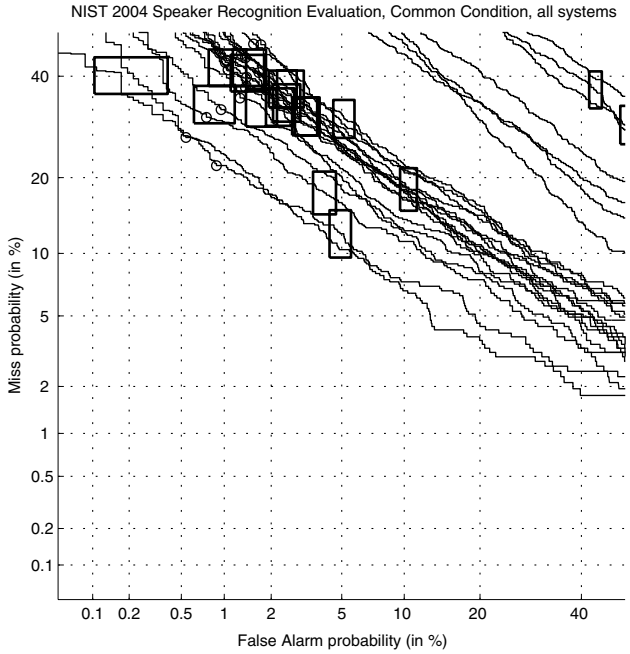

Fig. 2. DET plots for all primary systems, for the NIST 2004 evaluation. Experimental conditions were English language, 1 conversation side model training duration, and an equal test segment duration. The boxes and circles indicate the actual and minimum detection cost operating points, where the boxes represent the 95% confidence intervals.

5.2. Effect of training and test duration

The amount of speech data available, for training or in a test segment, is a key performance factor that has been considered in all of the NIST evaluations. Recent NIST evaluations (2000–2003) have had a separate 'extended data' test involving multiple conversation sides for training and single whole conversations sides as test, while the main evaluation test involved different and smaller amounts of training and test data.

The NFI-TNO evaluation studied the effect of training and test speech duration as a separate part in the evaluation. Here model train speech duration was factored as $t_m \in \{30\text{ s}, 60\text{ s}, 120\text{ s}\}$, and test segment duration was factored as $t_t \in \{7.5\text{ s}, 15\text{ s}, 30\text{ s}\}$. This was carried out in an orthogonal design, together with another 3-level factor 'number of sessions,' meaning conversations from which training data was taken, see Section 5.3.

The NIST 2004 evaluation sought to unify the evaluation with 28 conditions involving seven training conditions and four test conditions in all combinations (see Table 3). One training and one test condition involved summed channel data (see Section 5.6 below). Otherwise there were six training durations and four segment durations. The shorter durations segments were chosen as subsets of those of longer duration, except that fewer speaker models were available for the eight and especially the sixteen conversation sides training condition.

Some differences in approach between the two evaluations are:

- Maximum amount of speech data available is larger for NIST (up to 5 min vs. 30 s for test segments, up to 80 min vs. 2 min for training data).
- The factor 'more conversations' is orthogonal with train duration in the NFI-TNO design, for NIST the two are confounded because, naturally, the very long training duration conditions can only reasonably be formed from multiple conversations.

5.2.1. NFI-TNO effect of duration

Because the design had three orthogonal factors (training duration, test duration and number of training sessions), each with three levels, in total 27 separate conditions were formed. If each of these were to be analyzed separately the number of trials would have become very small for each condition. We have therefore analyzed the effect of each of the three factors separately, where data over the other two factors were pooled. Thus, e.g., for the analysis of the level '120 s' for the factor *training duration*, trials with 1, 2 and 4 sessions training were all used.

In Fig. 3 the effect of model training duration on score performance is shown for one system. The trend for better performance with longer training durations is seen in most systems. Note that, despite the pooling of other factors, the number of trials per analysis was $N_{tar} = 180$ and $N_{non} \approx 3100$. This number is still low compared to NIST evaluations, and hence the standard error is high, so we utilize the power of a joint analysis for all systems here. Using logistic regression we can model the EER by the factor *system* and the linear term $\log(t_m)$, the total model training duration. The choice for a dependence on the logarithm of training duration rather than assuming a linear dependence is motivated by the fact that a linear dependence would be too optimistic: by adding more training material the EER would vanish to zero too quickly.

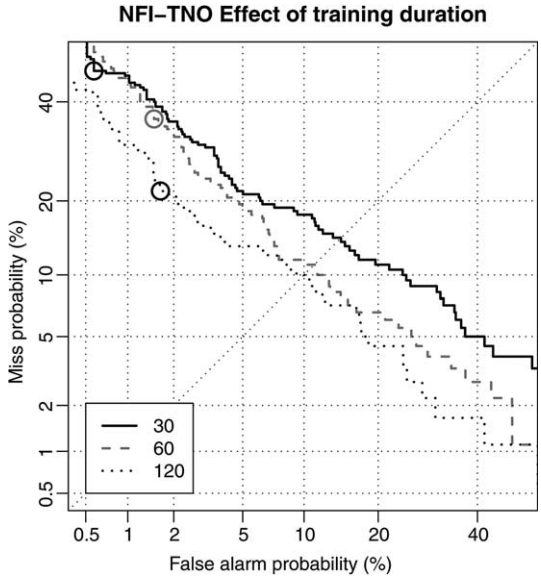

Fig. 3. Effect of training duration (30, 60, 120 s) on EER, for system T3 in NFI-TNO.

Apart from the obviously very significant effect of system, the logarithmic dependence on model training duration has $p = 0.019$. The reduction of logit(EER) with a doubling of t_m is 0.11.

For the factor *test segment duration*, we see a similar but less pronounced effect to training duration. In a logistic regression analysis using factors *system* and $\log(t_t)$, the test utterance duration, the analysis of deviance tables shows $p = 0.09$ for the regression coefficient for $\log(t_t)$, which is -0.075 per doubling of t_t. Apparently, our analysis method is not powerful enough to show an effect at the $p = 0.05$ significance level. Clearly, the low value of $N_{tar} = 120$ limits the power of this test. With twice as many target trials, the effect would have had $p = 0.011$, assuming the same DET curves.

5.2.2. NIST 2004 effect of duration

Fig. 4 shows the variation in the performance DET curve for one system with four training and three test durations. The sixteen sides condition is omitted as the number of trials was limited, and the three sides conditions is omitted to enhance the readability of the chart. The variation is as expected in the sense that longer durations always result in better performance (Doddington et al., 2000).

The training durations for the NIST trials were denominated in conversation sides (16, 8, 3, or 1) or in seconds (30 or 10). Previous NIST evaluation results had suggested that performance results were not very sensitive to small differences in speech durations that were in excess of 15 s or so, and whole conversation sides seemed the most natural units to use for long training durations. Five minutes were used from each conversation, so each side had an average duration of two and a half minutes. The conversations involved two willing adults who did not know each other talking on an assigned topic. Participants were generally polite and desirous of hearing what each to say, so conversations where one speaker strongly dominated the exchange were rare. (Calls were screened to weed out any instances that were not really conversations.) So while there were

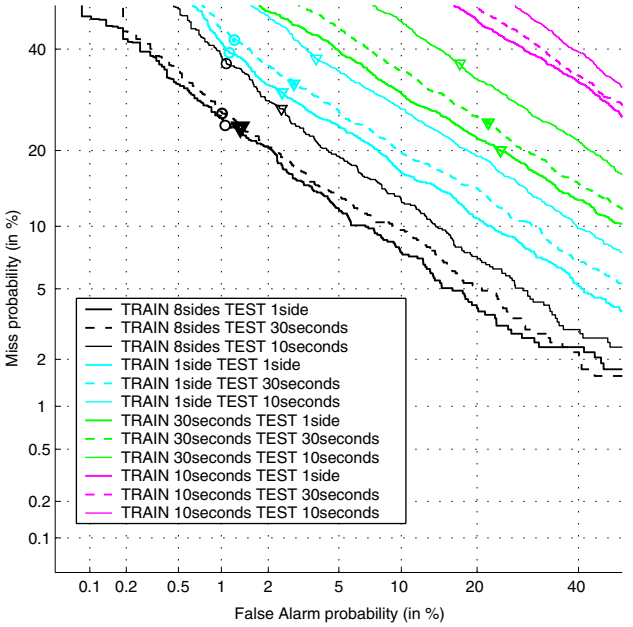

Fig. 4. DET curves for all conditions, formed from the test duration levels 10 s, 30 s, and 1 side, and the training duration levels 10 s, 30 s, 1 side and 8 sides. These are the results for system S9 in the NIST 2004 evaluation.

variations in actual speech duration among the training classes, there was essentially no overlap in total duration among them.

What is most notable, however, is that the training durations have a much greater effect on performance than the test durations. The DET curves shown in the chart separate into groups corresponding to the four training durations. Longer training is always superior, even when 10-s test segments are compared with whole conversation side test segments.

In principle, the speaker detection task is symmetric between the training and test segments. Training speech and test segment speech are both provided, and the system must determine whether the same or different speakers are involved. But comparing, for example, the one side training, 10-s test curve with the 10-s training, one side test curve in Fig. 4 shows very different outcomes. Performance results are quite non-symmetric, with more model training speech giving superior performance results. Note, that this asymmetry is observed for 'traditional' Universal Background Gaussian Mixture Model (UBM/GMM) systems (Reynolds et al., 2000). For one contrastive submission consisting of the recently developed Support Vector Machines (SVM) technique (Campbell, 2002), this asymmetry was not observed.

Looking only at test segment durations, it may be seen that 30-s durations give improved performance over 10-s durations, but that the differences between one conversation side and 30 durations are rather limited. Earlier evaluations had suggested minimal performance advantages to durations in excess of 15 s and up to about a minute. Here we see that even rather long durations (typically a conversation side is about two and a half minutes) result in fairly minimal performance improvement.

It should be noted, however, that most of the participating systems in the evaluations did not run most of the training and test duration conditions. Few systems attempted test conditions

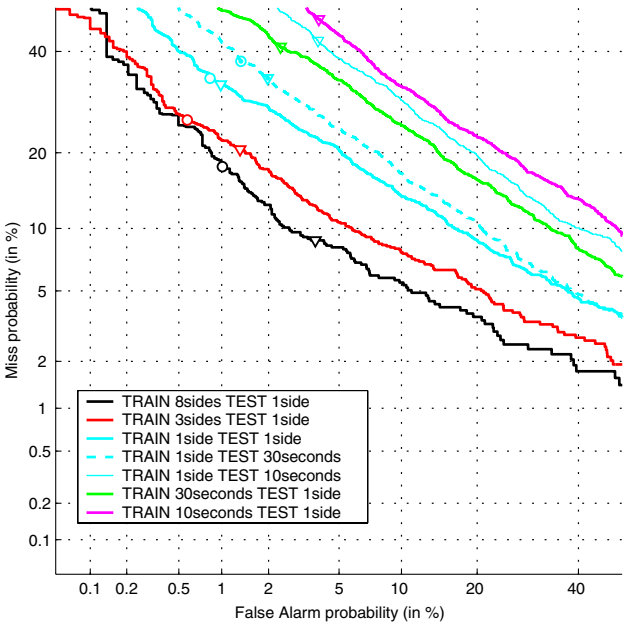

Fig. 5. DET performance as a function of duration condition for system S2.

where the test durations were longer than the training durations (see Table 3). Performance results for another system that attempted some of these conditions are shown in Fig. 5. Here the difference between one side and 30-s test segment durations (with one side training) is greater, but still small. The performance difference between 30-s test and 10-s test is rather larger, however. Comparing the DET curves involving one side and 10-s durations still shows the training duration to have greater effect than the test duration, but the difference between the two is much smaller than for the system in Fig. 4. The reason for this is probably that this GMM-based system S2 used a method choosing the longest duration of the test and training segments for training, and the shortest for testing, for every individual trial.

5.3. Effect of number of training handsets

As seen above, more training data gives better performance. But more variability within the amount of training provided may also be beneficial. Previous NIST evaluations have shown the importance of robustness to handset variability (Doddington et al., 2000). In the 2004 evaluation all target trials involved the use of a different test segment handset from the handset(s) used in training. (The phone number used was taken as indicative of handset distinctness, though different handsets were probably sometimes used with the same number, and the reverse situation is also conceivable.) For NFI-TNO the handset type was unknown. The influence of variability was studied by varying the number of different telephone sessions from which the training material was obtained, either 1, 2, or 4 sessions. We use the word 'session' rather than 'conversation,' in order not to be confused with the meaning of a NIST conversation, where more conversations imply more training data. For a varying number of sessions the total training time stays the same.

5.3.1. NFI-TNO effect of number of sessions

We observed the DET curves for all systems separated for the factor *number of sessions*. Most systems benefited from having training material from more than one session, but only incidentally did we observe a system that showed better performance for 4 session training than for 2. In order to summarize the effect for all systems, we have plotted in Fig. 6 the EER for all systems as a function of the number of training sessions, at a constant training length.

A logistic regression analysis of the EER on the factors *system* and *number of sessions* shows that the factor *system* is very significant, obviously, but the factor *number of sessions* is only significant with $p = 0.038$. A pairwise test of proportions shows that systems only benefit from more than one training session, but that there is no difference going from two to four sessions.

5.3.2. NIST effect of number of training handsets

The 2005 NIST evaluation utilized training data from a single handset for most models, but for the eight conversation side training condition some of the defined models involved two or more different handsets. Most systems showed the expected outcome that having multiple training handsets (all different from the handset used in each trial's test data) produced somewhat improved results, though the degree of improvement was fairly modest. Fig. 7 shows a typical result, comparing the DET curve for single handset training trials with that for multiple handset training trials. Fig. 8 has a similar plot for one system where there is no apparent difference in performance. Why this is so is not clear; the system used an overall approach (GMM-UBM) combining several levels of speech signal information that was not dissimilar to that used by other evaluation participants.

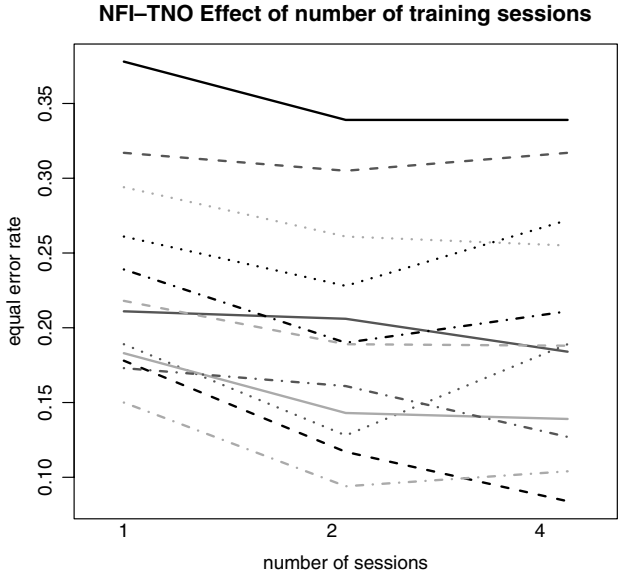

Fig. 6. The effect of number of sessions in model training on the EER for all systems. Data points within a system are connected for visibility, using the same line type scheme as in Fig. 1.

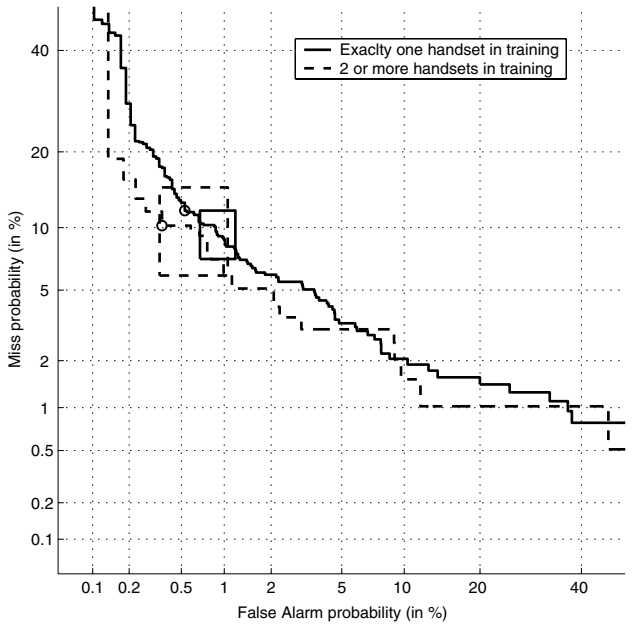

Fig. 7. The effect of handset variation for a system in NIST 2004, which shows typical behaviour.

5.4. Language effects

The NFI-TNO forensic database contained a number of speakers speaking in two different languages. We will call these 'dual language' speakers, rather than 'bilingual,' because these speakers were generally not fluent in both languages. A limited study could be made of the influence of spoken language. The Mixer data provided the first opportunity to look at the effects of language on speaker detection performance in the NIST evaluations. The presence of a good number of dual language speakers allows trials to be segmented in several ways. Several DET plots in this section illustrate different aspects of this for a particular system.

5.4.1. NFI-TNO language effects

In the NFI-TNO design there were the following contrastive language experimental conditions:

English. All training and test segments were spoken in English. Some of the speakers were native speakers, others were not. There were 21 target speakers and only 2 additional speakers that could be used as non-target test speakers.
Cross language, Dutch test segments (xl-dt). In total 9 speakers spoke in Dutch and another language (English, Sranan Tongo or Papiamentu). This test involved trials using Dutch test segments and target models built from non-Dutch speech.
Cross language, Dutch models (xl-dm). For only 5 of the dual language speakers enough model training material in Dutch could be found. This test involved the opposite of the previous test, namely trials combining target models built from Dutch speech with non-Dutch test segments.

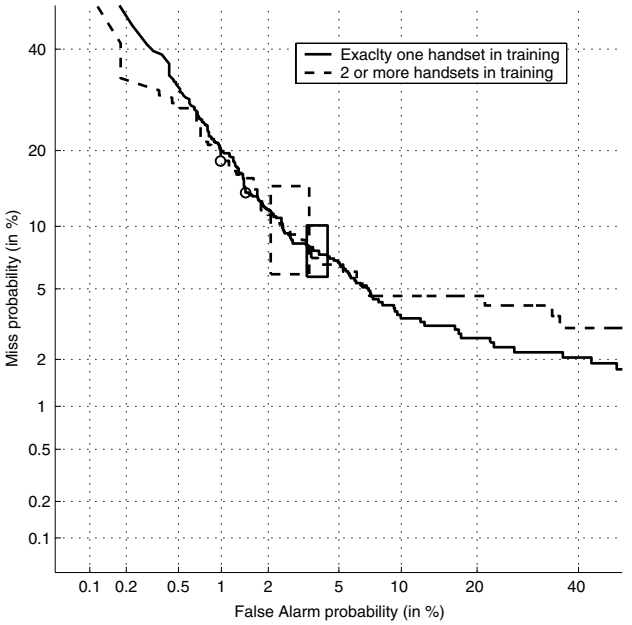

Fig. 8. The effect of handset variation for another system, which shows more robustness to handset variation.

All other parameters were fixed at 1 session training of 60 s, 15-s test segments. Despite the limited number of speakers available, we found some effects of language.

The language conditions can best be summarized in an 'interaction plot' as shown in Fig. 9. Here, the EER for all systems is shown for the two main language conditions Dutch (the overall condition 'Dutch' and the varying duration condition 'Dutch2' from Section 5.2.1), English, and the two cross language conditions: models trained in Dutch tested with another language (xl-dm) and models tested in Dutch trained in another language (xl-dt). The conditions are ordered from left to right in generally increasing EER.

Both the factors *system* and *language* are highly significant in an analysis of deviance on the data shown in Fig. 9. It may be instructive to show the power of the analysis of deviance. In Table 6 the analysis of deviance is reproduced. The table should be interpreted as follows: from all the deviance that the generalized linear model can have (534.76) a large part (405.37) is modeled by 11 parameters for the factor *system*. Of the remaining deviance, 91.39 is modeled by 4 parameters for the factor *language*. The remaining deviance of 38.00 is not modeled. Not shown in the table is that, if we would add the interaction effect of the factors *system* and *language*, it would take 44 more parameters to reduce the deviance to 0. This interaction is highly non-significant, $p = 0.73$.

The question of which language conditions are significantly different in terms of the systems' performance is answered by carrying out a pairwise test of proportions between the language conditions. In Table 7, the corresponding p-values are tabulated. Here, a 'Holm adjustment' to the p-values has been applied to compensate for the many comparisons performed. From the table it follows that only the 'neighbouring conditions' in Fig. 9 are not significantly different, except the 'neighbours' Dutch2-English, which is significant.

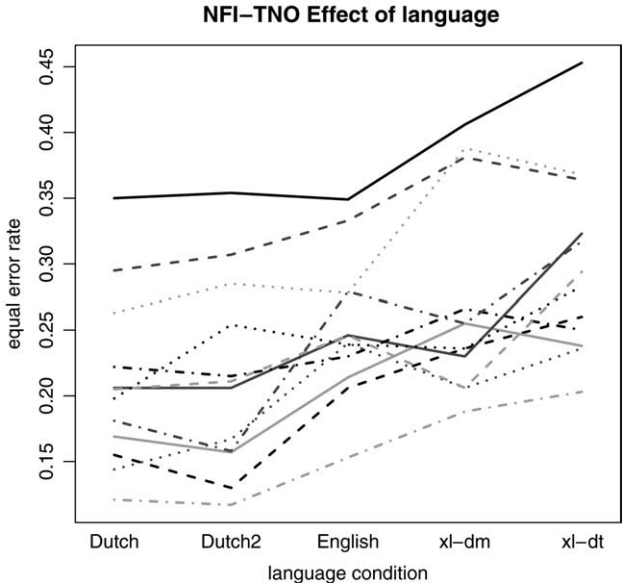

Fig. 9. EER in the several conditions for all systems in NFI-TNO. From left to right are two conditions of Dutch, English, and two cross language conditions. The condition 'xl-dt' are trials with Dutch test segments and models trained on other languages, the condition 'xl-dm' has Dutch models and test segments in other languages. Data points within a system are connected for visibility, using the same line type scheme as in Fig. 1.

Table 6
Analysis of deviance for the effects of *system* and *language*

Analysis of deviance	Df	Deviance	Residual Df	Residual deviance	$P(>\|\chi\|)$
NULL			59	534.76	
System	11	405.37	48	129.39	4.4×10^{-80}
Language	4	91.39	44	38.00	6.7×10^{-19}

The numbers are taken from the analysis program, not all decimal places are relevant. 'Df' means 'degrees of freedom'.

Table 7
Pairwise comparison of the language effect on EER

	Dutch2	English	xl-dm	xl-dt
Dutch	0.57	0.0029	$<10^{-3}$	$<10^{-3}$
Dutch2		0.0089	$<10^{-3}$	$<10^{-3}$
English			0.37	0.012
xl-dm				0.22

5.4.2. NIST language effect

Fig. 10 examines the effect of classifying trials by the language mix of the training and test data. It should be noted in this context that the ASR (automatic speech recognition) transcripts that were made available to all evaluation sites were produced by an English word recognizer, whatever the actual language of the input speech. Restricting to same language trials generally produced slightly better performance than including all trials, as is the case for both systems

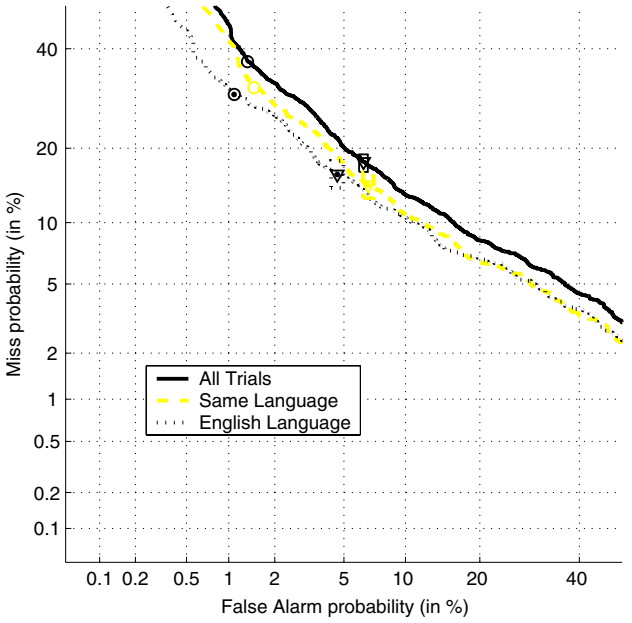

Fig. 10. The effect of language mix in NIST 2004 for system S8. Restricting to same language trials generally aids performance, while restricting to English-only trials is advantageous to systems utilizing lexical information and the English ASR transcripts such as this system.

shown in Figs. 10 and 11. But restricting to all-English trials to benefits slightly systems, like that in Fig. 10 that seek to utilize lexical information provided by the transcripts, while making little difference for systems, like that in Fig. 11 which utilize acoustic information only.

Figs. 12 and 13 consider separately the segmentation of target (same speaker) and non-target (different speaker) trials by language mix. The non-English data has been pooled, as only small differences were found by looking separately at data in the individual languages. In Fig. 12 each curve includes all non-target trials, while in Fig. 13 all target trials are included in each curve. Thus, the actual decision false alarm rates are the same for Fig. 12 curves, and the actual decision miss rates are the same for Fig. 13 curves.

Most notable in Fig. 12 are the superior results when the target trial training and test data are in the same language, especially in a language other than English. Conversely, Fig. 13 shows that the poorest results are obtained for non-target trials where the training and test data are in the same non-English language. This suggests that the system was, to a significant extent, performing language recognition for the non-English data. The numbers of speakers included in the DET curves of Fig. 12 of each ''accent'' are as follows: Arabic 63, Mandarin 58, Russian 68, and Spanish 134.

The use of dual-language speakers also supports another type of analysis illustrated in Fig. 14. In the DET curves, all non-target trials are included, while the target trials all involve only English language data, but are segmented according to the other language, if any, spoken by the target speakers involved. Thus the curves are labeled by the Arabic, Mandarin, Russian, or Spanish 'accent' of their speakers, while an English 'accent' refers to single language speakers. The term 'accent' here suggests that, although the speakers spoke English, their native language is the *other*

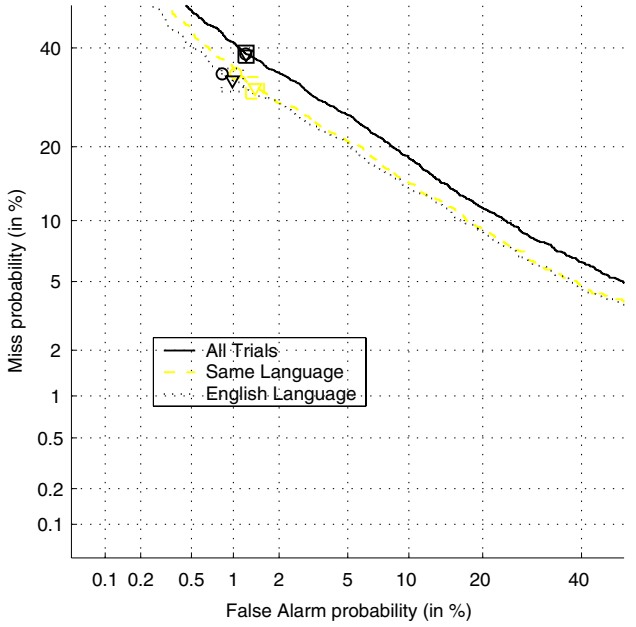

Fig. 11. Restricting trials to English-only is not advantageous to systems relying only on acoustic information, such as here in system S2.

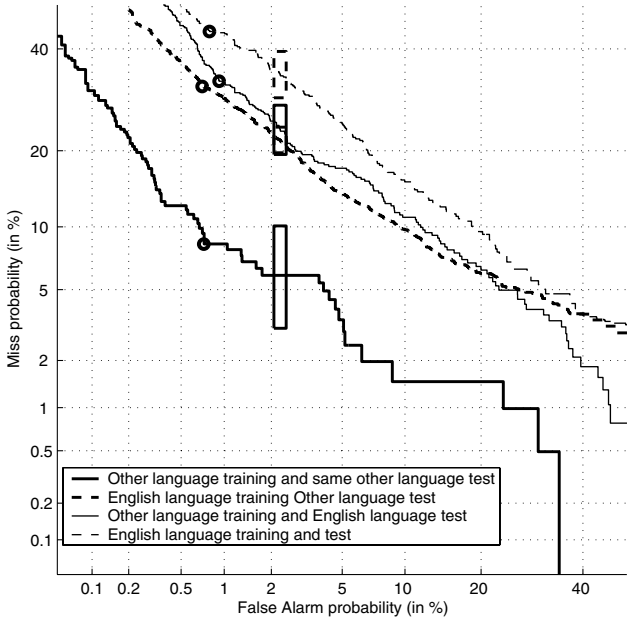

Fig. 12. Separation of target trials by English/other language, for a contrastive submission of site S11.

language that they used in other conversations. What stands out is the superior performance of the system with target trials involving Arabic speakers who are speaking in English. These speakers are apparently well distinguished by the system involved. Other evaluation systems were

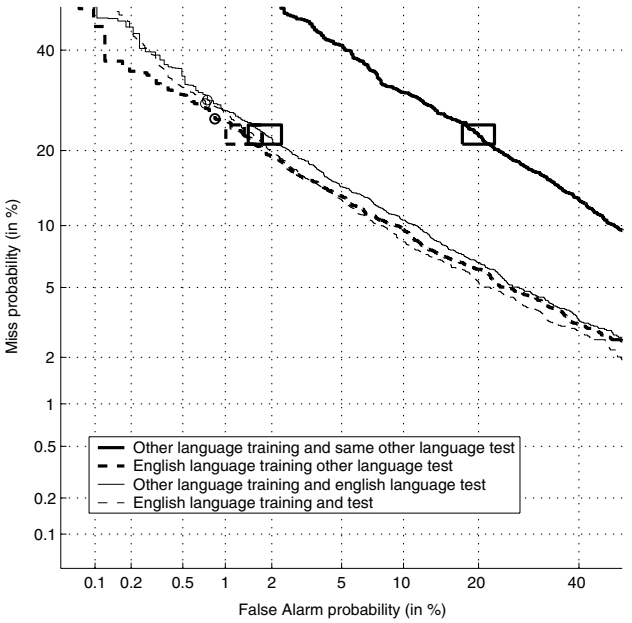

Fig. 13. Separation of non-target trials by English/other language, for the system shown in Fig. 12.

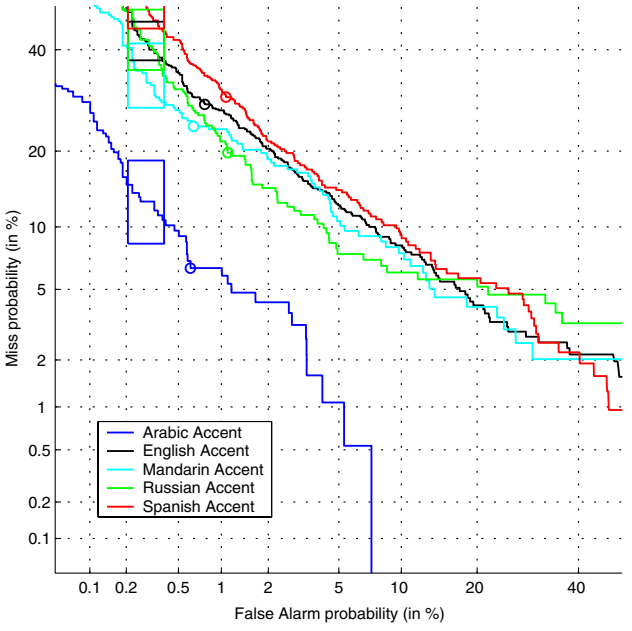

Fig. 14. Separation of target trials by other spoken language (“accent”), for system S1. All trials contained only English speech.

similar in this respect. It would be of interest to examine performance on non-target trials where the training and test speaker both had similar accents, but insufficient trial data was available in the 2004 evaluation for this purpose.

It is to be hoped that these language and accent issues may be further studied in future evaluations.

5.5. Effect of transmission type for NIST 2004

For each call in the Mixer collection each conversant was asked to indicate whether the transmission type of the handset was cordless, cellular, or regular. The last term, intended for ordinary land-line phones, may have been confusing for some users, so the sharpest contrasts appear in comparisons involving either cordless or cellular transmission. The effect of the training and test segment transmission types on performance is likely to be different for target (same-speaker) and non-target (different speaker) trials. Figs. 15 and 16 illustrate the effects of transmission type on performance for one system in target and non-target trials, respectively. It should be noted that all systems implement in their processing various types of normalization to the different channel conditions that are expected to be encountered. The system in question here used RASTA cepstral filtering and a speaker-specific T-norm score normalization based on models which scored most similarly to the given model on a set of impostor utterances. Indeed, most of the evaluation systems used some type of T-norm normalization.

In Fig. 15 the target trials are varied, while all non-target trials are included in each DET curve. Thus the actual decision false alarm rate is fixed across curves. It may be seen that target trials where the training and test transmission types are the same give better performance than those where they are different, with cordless transmission outperforming cellular, as might plausibly be expected. Note again that all trials involve different phone numbers and presumably different handsets. For the mixed trials, better performance is obtained when the training is cordless and

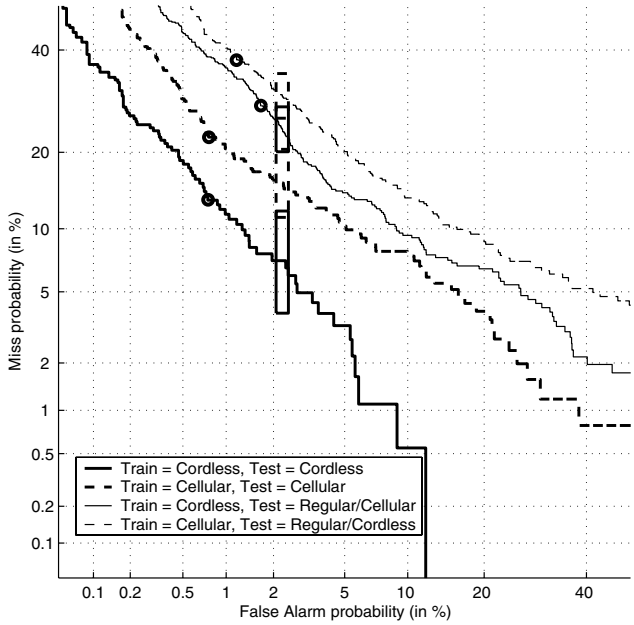

Fig. 15. Separation of target trials by transmission type, for the system shown in Fig. 12.

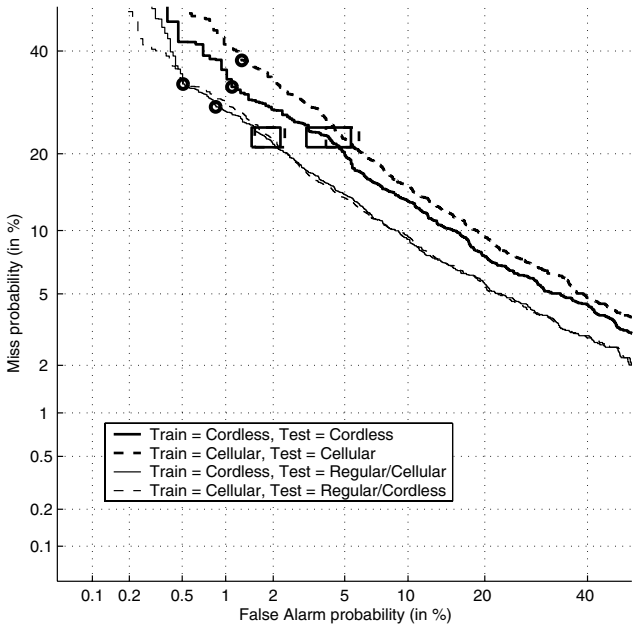

Fig. 16. Separation of non-target trials by transmission type, for the same system as in Fig. 15.

the test is not than when training is cellular and test non-cellular. This is in accord with the notion (see Section 5.2.2) that performance is more sensitive to the training than to the test data.

When the non-target trials are varied using all target trials as in Fig. 16, it is seen, for the given system, that better performance occurs with mixed training and test transmission types. This reverse of the situation for target trials suggests that systems are, to a limited degree, recognizing transmission types in place of voices. Note also that in Fig. 16 performance is better for the matched cordless condition than the matched cellular condition, but there is no difference for the mixed conditions.

5.6. Summed channel data for NIST 2004

Here we examine performance involving the test conditions where either the training or the test data was two-speaker summed channel data. Fig. 17 plots six DET curves involving the three conversation (summed channel), the three side, and the one side training conditions and the one conversation (summed channel) and one side test conditions. All of the trials included involve only English speech and only a single training handset (phone number). In addition, there is a one-to-one correspondence of trials for the six curves, with the speech in one side training or test segments being a subset of that in a corresponding one conversation segment.

It may be observed from the curves in Fig. 17 that for both training and test data, performance is better with single channel data than with an equal amount of summed channel data. Moreover, the performance difference is almost as great when one side training is compared with three conversation training. Having single channel uncontaminated data is the most important factor affecting performance. More surprisingly, perhaps, it may be observed that the three curves

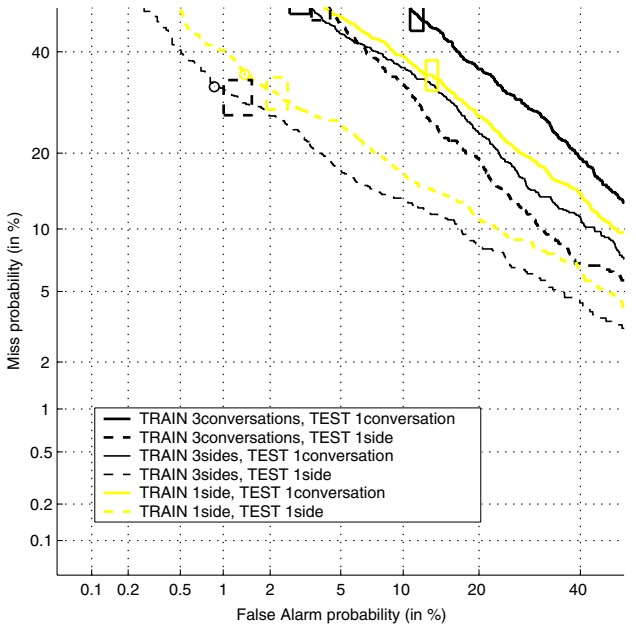

Fig. 17. NIST 2004 DET performance for various summed channel conditions, for system S9.

involving single channel test data all outperform the three involving summed channel test data. In this regard, unlike the situations considered in preceding sections, the nature of the test data has greater effect on performance than that of the training data.

Another issue related to summed channel training and test data is explored by Fig. 18. The summed channel training data for a speaker always involves three conversations with three different other talkers. A non-target (impostor) trial could involve test segment speech by one of these three other speakers in the training. Fig. 18 involves the three conversation training, one conversation test condition with the non-target trials separated into those where one of the test segment speakers is one of these three and those where this is not the case. Fig. 18 shows that for one typical system, performance indeed degrades for the first case.

5.6.1. Gender mix

Performance involving summed channel data in either training or test is also influenced by the gender mix of the summed channel data. Mixed gender speech segments are generally more readily segmented by speaker, avoiding contamination effects. The figures in this section examine this.

Fig. 19 examines the gender mix of test segments for the three conversation training, one conversation test condition. For the system shown, which is typical of most, it is seen that there is a small performance advantage on mixed gender test segments.

For summed channel training, there are three training conversations, so zero, one, two, or all three of these may involve mixed genders. It may be seen in Fig. 20 that there is a considerable performance advantage for the system considered when all training data is mixed gender, helping to avoid contamination of the data actually used for model building. It is also interesting to

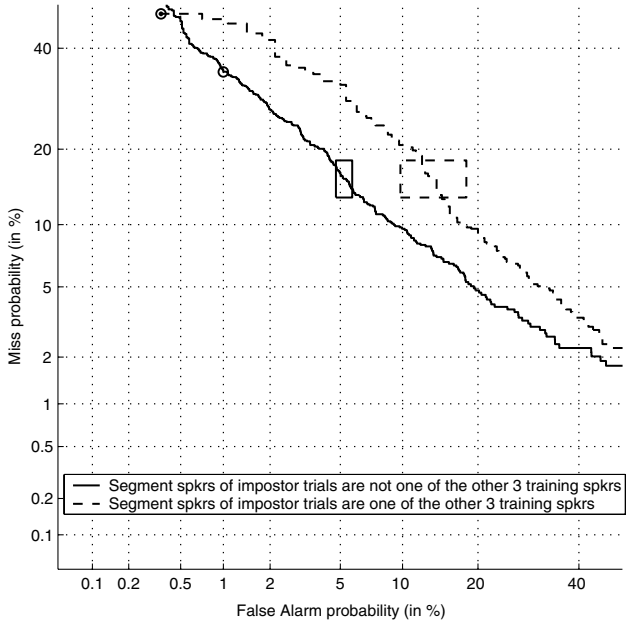

Fig. 18. DET performance for summed channel conditions for system S11, separated for impostor trial speakers as to whether they occurred as conversation partner in one of the three training conversations.

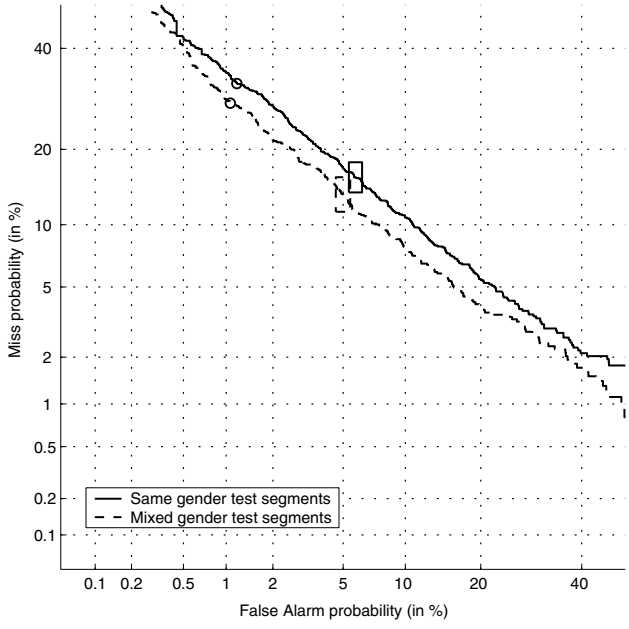

Fig. 19. DET performance of summed channel data for different gender mixes, for system S11. The curves separate the gender mix in test segments.

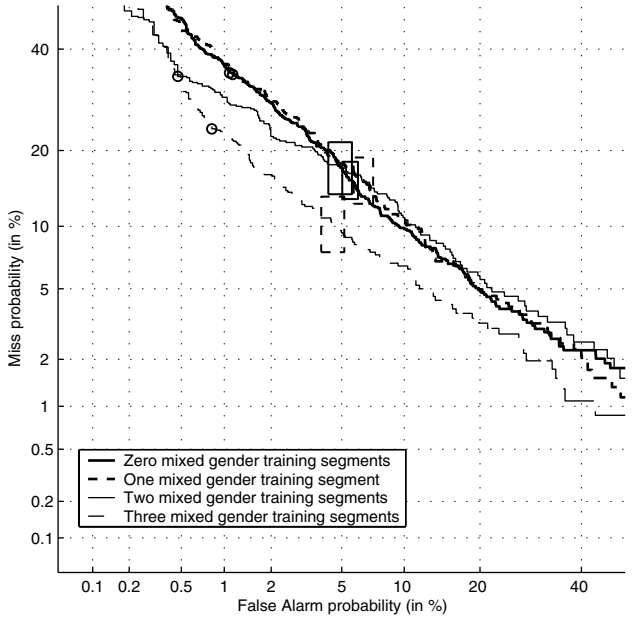

Fig. 20. DET performance of summed channel data for different gender mixes, for system S11. The curves separate the gender mix in training segments.

observe that having two of the training conversations be mixed produces approximately equally good results for low false alarm rate operating points.

The joint influence of training and test gender mix is considered in Fig. 21. In the curves shown here, the training mixed and same gender conditions mean that either all three or none of the three training conversations contain mixed gender speech. It may be seen that, as observed similarly in the preceding section, the test segment condition has considerably greater effect on performance outcomes than the training segment condition.

5.7. Comparison of evaluations

We have seen the effects of several factors on the DET performance, and indicated several statistical techniques for analyzing the significance of effects. We now attempt to address the question: can we measure the difference between *evaluations* themselves? Can we observe that one particular data collection is 'easier' than another? One reason to perform this effort is the observation that from year to year in the NIST evaluations the general performance changes, and it is interesting to separate effects from the change of a system from the effects of the evaluation data set. The method layed out in this section might help in such an analysis.

For this purpose we will try to analyze the difference in performance between both evaluations held in 2003, NIST and NFI-TNO. Four sites participated in both evaluations, which were held within about half a year from each other. We have asked the sites whether their systems changed much between the two evaluations. All reactions were that the changes were minimal, and where there were explicit changes, the expected effect on EER would be very small. For this analysis we take the influence of the individual system's change negligible.

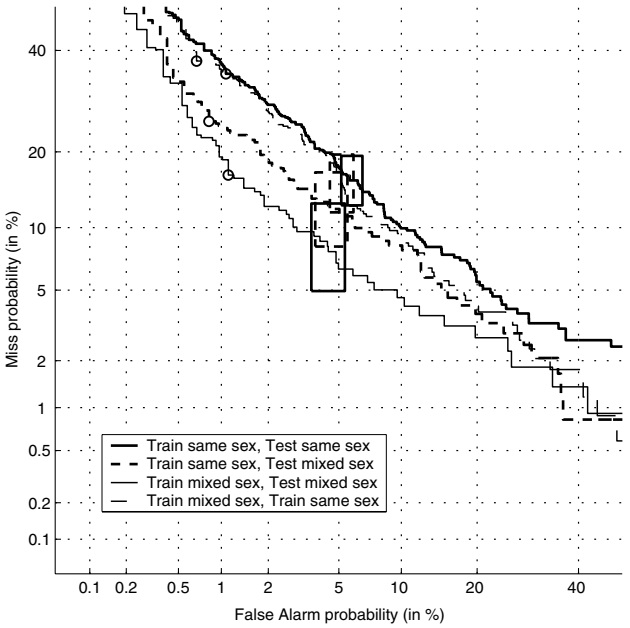

Fig. 21. DET performance of summed channel data for different gender mixes, for system S11. The curves separate the gender mix in both training and test segments.

First, we try to correct for the difference in duration conditions between the two evaluations. We take the representative test duration of NIST 2003 at 30 s, twice that of NFI-TNO. Similarly, the training duration is twice as large. We then correct the EER in NFI-TNO for duration by mapping the EER to the logit domain, and subtracting 0.11 for the doubling in training durations (see Section 5.2.1), and 0.075 for the doubling in test duration. The latter effect could not be proven to be significant at the $p = 0.05$ level, but we correct for the log-linear regression anyway because of the expected effect (see Section 5.2.2 and Doddington et al., 2000). Then the NFI-TNO values can be mapped to the EER domain by the inverse logit function. The comparison of evaluations for the four systems is shown in Fig. 22.

A logistic regression analysis of deviance of the factors *system* and *evaluation* shows that both factors are significant with $p \ll 10^{-3}$. The largest uncertainty in EER is due to N_{tar} in NFI-TNO, and we have not included the possible error introduced by the duration corrections, but we don't think that this will change the significance of the difference in evaluations. We can therefore conclude, that the NFI-TNO task was *harder* than the NIST task held in the same year. Possible explanations are the choice of speakers, the different language used, the speaking style, the signal to noise ratio, or other factors which were not investigated. In summary, we may call the combination of all these unknown factors the effect of the *evaluation*.

We were fortunate that the four systems in this analysis did not change much. In general, however, the comparison of evaluations is confounded with the development of systems. It is hard to attribute the change in performance of a system from one evaluation to the next to either actual

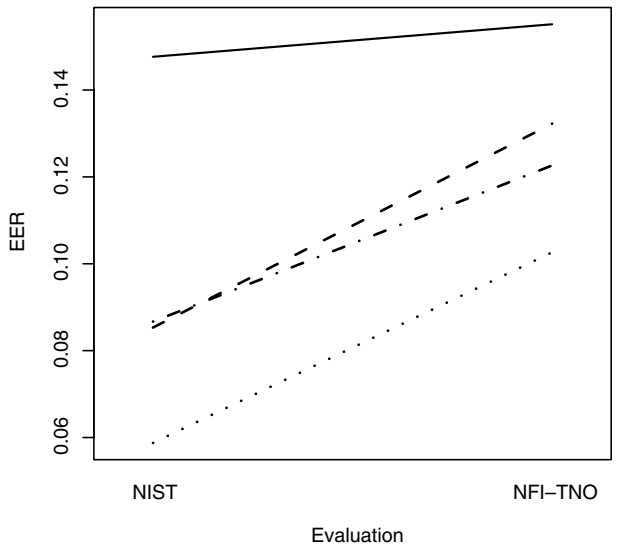

Fig. 22. Comparison of the EER for the four systems that participated in both NIST 2003 and NFI-TNO evaluations. The NFI-TNO EERs have been corrected for the difference in speech duration from NIST.

changes in the system or to a difference in the evaluation. One way to make the different contributions explicit, is to ask sites to not only run their new system on the new evaluation data, but also run the system they have used for the previous evaluation (Doddington, 2004) on the new data. NIST encourages sites to do this in future evaluations.

6. Summary and conclusions

We have given an overview of the evaluation paradigm of the yearly text independent speaker recognition evaluations held by NIST and that of NFI-TNO in 2003. We have presented and analyzed the result of two recent evaluations. We have introduced an analysis of deviance for studying various factors affecting the equal error rate in the NFI-TNO evaluation, and studied various performance factors affecting the DET curve in the NIST 2004 evaluation. Important factors are training segment duration and, to a lesser extent, test segment duration. Longer speech durations make the detection potential of a system better. Being subjected to several handsets in the training material generally makes a system more robust, but some of the better performing systems show no improvement, which suggests that these systems have a proper compensation for handset variability. Language dependence experiments show interesting effects of accents, most clearly indicated by categorizing trials by which *other* language a speaker is able to speak. For the two-speaker detection tasks, the negative effect on the DET performance of the contamination by the other speaker is much larger for test segments than for training segments, which is an interesting contrast to speech segment duration. Finally, we have made an attempt to compare *evaluations* as a whole to each other.

Acknowledgements

We want to thank Roland Auckenthaler, Claude Barras, Todor Ganchev and Doug Reynolds for supplying us with additional results, and Niko Brümmer for the many discussions involving decision theory.

References

Przybocki, M.A., Martin, A., 1999. The 1999 NIST speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking. In: Proceedings of the Eurospeech, pp. 2215–2218.

Martin, A., Przybocki, M., 2000. The NIST 1999 speaker recognition evaluation – an overview. Digital Signal Processing 10, 1–18.

Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A., 2000. The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective. Speech Communication 31, 225–254.

Martin, A.F., Przybocki, M.A., 2001. The NIST speaker recognition evaluations: 1996–2001. In: A Speaker Odyssey – The Speaker Recognition Workshop, pp. 39–42.

Przybocki, M.A., Martin, A.F., 2002. NIST's assessment of text independent speaker recognition performcance. In: The Advent of Biometircs on the Internet, A COST 275 Workshop, pp. 25–32.

Przybocki, M., Martin, A., 2004. NIST speaker recognition evaluation chronicles. In: Proceedings of the Odyssey 2004 Speaker and Language Recognition Workshop, ISCA, pp. 15–22.

Van Leeuwen, D.A., Bouten, J.S., 2004. Results of the 2003 NFI-TNO forensic speaker recognition evaluation, In: Proceedings of the Odyssey 2004 Speaker and Language recognition workshop, ISCA, pp. 75–82.

Bimbot, F., Blomberg, M., Boves, L., Genoud, D., Hutter, H.-P., Jaboulet, C., Koolwaaij, J., Lindberg, J., Pierrot, J.-B., 2000. An overview of the CAVE project research activities in speaker verification. Speech Communication, 155–180.

Brümmer, N., 2004. Application-independent evaluation of speaker detection. In: Proceedings of the Odyssey 2004 Speaker and Language Recognition Workshop, ISCA, pp. 33–40.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. In: Proceedings of the Eurospeech 1997, Rhodes, Greece, pp. 1895–1898.

The NIST year 2004 Speaker Recognition Evaluation Plan. Available from: <http://www.nist.gov/speech/tests/spk/2004/index.htm>.

Van Leeuwen, D.A., Bouten, J.S., 2003. The NFI/TNO forensic speaker recognition evaluation plan. Available from: <http://speech.tm.tno.nl/aso/evalplan-2003.pdf>.

Hays, W.L., 1963. Statistics. Holt, Rinehart and Winston, Inc.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted gaussian mixture models. Digital Signal Processing 10, 19–41.

Campbell, W.M., 2002. Generalized linear discriminant sequence kernels for speaker recognition. In: Proceedings of the ICASSP, pp. 161–164.

Doddington, G., 2004. NIST speaker recognition workshop.