Overall performance metrics for multi-condition Speaker Recognition Evaluations

David A. van Leeuwen

International Computer Science Institute, Berkeley, CA
TNO Human Factors, Soesterberg, The Netherlands,
Radboud University Nijmegen, The Netherlands

Abstract

In this paper we propose a framework for measuring the overall performance of an automatic speaker recognition system using a set of trials of a heterogeneous evaluation such as NIST SRE-2008, which combines several acoustic conditions in one evaluation. We do this by weighting trials of different conditions according to their relative proportion, and we derive expressions for the basic speaker recognition performance measures $C_{\rm det}$, $C_{\rm llr}$, as well as the DET curve, from which EER and $C_{\rm det}^{\rm min}$ can be computed. Examples of pooling of conditions are shown on SRE-2008 data, including speaker sex and microphone type and speaking style.

1. Introduction

One of the recent research focuses in Automatic Speaker Recognition is the challenge to deal with channel variability, or more generally, inter session variability. This direction of focus has led to both the collection of databases containing channel variability and technical approaches to deal with this variability. The MIXER SRE-2004 component can be seen as an exponent of this data collection effort, where all trials in the core test condition were selected to be *different telephone number* trials, assuming different telephone handsets and acoustical environments between train and test segment. Examples of approaches to deal with this variability are (Joint) Factor Analysis (FA) [1], Probabilistic Subspace Adaptation (PSA) [2], Nuisance Attribution Projection (NAP) [3] and Feature Domain channel factor compensation [4], which all are data-driven methods exploiting earlier data collection efforts.

At the SRE-2006 workshop discussion, it was remarked that not many sites participated in the 'auxiliary microphone' condition. It was suggested by the present author to include the various microphone condition trials in the required test condition set of trials of the next SRE, if the community felt that the different microphone conditions are an interesting problem to work on by the community as a whole. NIST has subsequently generalized the inclusion of different microphone conditions in the core test condition to include different speech styles, "interview" and "phone call." NIST included 5 combinations of microphone type and speech style (henceforth called acoustical conditions) in the core test condition trial set "short2-short3" in SRE-2008.

In the evaluation plan it was announced that these acoustical conditions were going to be analyzed strictly separately. Hence, in SRE-2008 the community focused on the problem of session variability in microphone type and speech style, but strictly limiting to per-acoustic-condition analysis, thereby not measuring score consistency across these conditions. However, at TNO, and some other sites, we believe that it an interesting task to get calibration right over all acoustic conditions. This means that a score x for a detection trial should have the same interpretation, regardless of the (analysis) condition it happens to be part of. We believe that developing systems that optimize the EER and cost function for such pooled conditions will not just make systems more robust to these varying conditions and their scores more generally interpretable. This will also, as a side effect, optimize performance of the individual acoustical conditions to some extent, but in a way that is not too focused on that individual condition.

The purpose of this paper is to propose a framework for measuring the overall performance of a system over all trials of an evaluation like SRE-2008 "short2-short3," in a meaningful and sensible way.

We will proceed by starting with a naive approach, identify some of the problems related to this, and then propose a new evaluation scheme that allows for pre-determined weighting the different acoustical conditions in an evaluation. We will show how to compute the basic detection performance parameters, but also treat more advanced measures such as $C_{\rm llr}$. We will show the effects of this new approach using the submitted scores from several of the better performing systems of NIST SRE-2008.

2. Pooling of trials

The simplest approach to measuring the performance over all conditions is to simply pool all trials, meaning pooling decisions for $C_{\rm det}$ and pooling scores for the DET curve ($C_{\rm det}^{\rm min}$, EER). In Figure 1 we show the effect of pooling in a DET plot, where the solid black line at the top represents the DET curve obtained after pooling all 98776 trials of the NIST SRE-2008 "short2-short3" core test condition. Also, in colour, DET plots are made for trials conditioned on the 5 different acoustic conditions for which the evaluation included trials. (Note, that the SRE-2008 evaluation plan does not mention the "phonecall interview (mic)" trials as a common condition. DET curves for this condition, however, have been distributed among participants as 'plot-9' graphs).

Several remarks can be made about the graph. First, note that the TNO systems is not particularly well calibrated: decision points (rectangles) tend to be to the left of the minimum cost points (circles), i.e., (log-likelihood-ratio) scores tend to be too low, there is "under confidence." But more interestingly, one condition is the odd-one-out: "phonecall phonecall (phn)" where scores were over-confident. This is an example of an inconsistent mis-calibration between different acoustic conditions. This leads to an over-all DET curve which lies above the

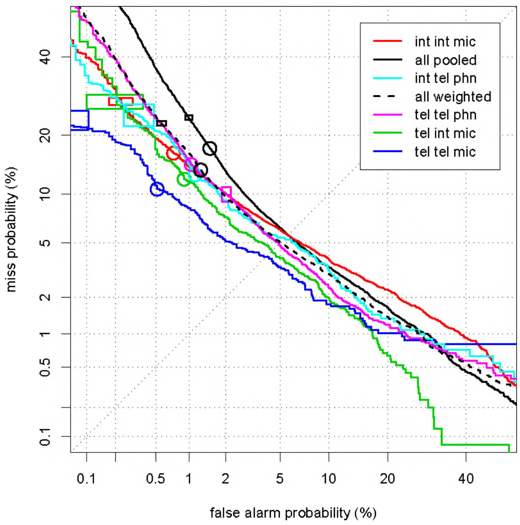


Figure 1: DET curves obtained for TNO-1 in NIST SRE-2008, after pooling all trials in the “short2-short3” core test condition (solid black). In colour, DET curves are conditioned on acoustic condition, where ‘int’ indicates interview style, ‘tel’ phonecall style, ‘phn’ recording test segment over phone handset, ‘mic’ recording over auxiliary microphone. The fist ‘int/tel’ designates training condition, the second test condition. The dashed line is explained in Sect. 4.2

Condition	NIST	$C_{\rm llr}$	EER	$C_{\rm det}$	$N_{\rm tar}$	$N_{\rm non}$
all	-	0.250	5.62	0.0338	20449	78327
int int mic	1	0.238	5.63	0.0301	11540	22641
tel int mic	-	0.241	4.40	0.0297	2500	4850
tel tel mic	5	0.238	4.01	0.0236	1472	6982
int tel phn	4	0.226	5.35	0.0279	1105	10636
tel tel phn	6	0.222	4.90	0.0301	3832	33218

Table 1: Performance summary for TNO-1, pooling all trials. ‘Condition’ is as in Fig. 1. ‘NIST’ indicates equivalent NIST common evaluation condition.

other curves, rather than being in-between. Some performance measures are in Table 1.

There is, however, and important draw-back to this kind of pooling of trials, as was put forward by Doug Reynolds of MIT. If we look at the number of trials per conditions (cf Table 1), we see that these vary widely across condition and target/non-target class. This has an effect on the performance measures.

For instance, the ‘int int mic’ condition—with over half the total number of target trials—completely dominates the $P_{\rm miss}$ behaviour, perhaps most visible at low false alarm rates. Other conditions (such as ‘tel tel mic’) have a very low weight in the overall DET performance. This may be taken as a fact of SRE-2008, but we may want to think of a way of compensating for this, especially for researchers who try to get calibration right over all conditions.

Note, that in all NIST evaluations up to now we have been happily pooling male and female trials, which tend to give different performance, thus forcing system developers to get calibration correct over speaker sex, *even though there are no cross-sex trials*, and systems may actually have separate sub-systems for male and female trials. We believe this pooling is a good thing, but it does lead to sensitivity of the overall performance to the relative amount of trials for female and male. For 2006, the difference was only about 10 %, so the effect was not very large anyway. In the following proposed framework however, we will be able to compensate for this effect as well.

3. Proposed framework for pooling conditions

Just like we weight the trial categories for target and non-targets separately[1], disentangling the evaluation priors from the application priors, we can give the trials in each acoustical condition α separate weights. Let us define the relative weights β for target and non-target trials as

$$\beta^\alpha_{\rm tar} = w_\alpha \Big/ \frac{N^\alpha_{\rm tar}}{N_{\rm tar}}; \qquad \beta^\alpha_{\rm non} = w_\alpha \Big/ \frac{N^\alpha_{\rm non}}{N_{\rm non}}, \tag{1}$$

where $N^\alpha_{\rm tar}/N_{\rm tar}$ (and $N^\alpha_{\rm non}/N_{\rm non}$) are the fraction of target (and non-target) trials belonging to condition α in the evaluation. The weights w_α (summing to unity) are the *desired* weights for conditions α, possibly related to expected usage in an application. These should be specified before any evaluation of interest, but since that has not been done for SRE-2008, we will use $w_\alpha = 1/N_c$, where $N_c = 5$ is the number of conditions.

Using these trial-dependent β, we propose to compute the probability of false alarm at a given threshold θ for a set of trials $\{t\}$ with scores $s(t)$ as

$$P_{\rm FA}(\theta) = \frac{1}{N_{\rm non}} \sum_{t\in{\rm non}} \beta^{\alpha(t)}_{\rm non} u(s(t)-\theta), \tag{2}$$

utilizing the unit step function u. Similarly the miss rate is

$$P_{\rm miss}(\theta) = \frac{1}{N_{\rm tar}} \sum_{t\in{\rm tar}} \beta^{\alpha(t)}_{\rm tar} u(\theta - s(t)). \tag{3}$$

These formulas are nothing new, they represent the usual estimation of $P_{\rm FA}$ and $P_{\rm miss}$, but now include weights conditioned on α. Weighting the trials individually is equivalent to analyzing $P^\alpha_{\rm FA}$ and $P^\alpha_{\rm miss}$ separately for each condition α and taking the weighted average $P_{\rm FA} = \sum_\alpha w_\alpha P^\alpha_{\rm FA}$ and $P_{\rm miss} = \sum_\alpha w_\alpha P^\alpha_{\rm miss}$.

3.1. Traditional evaluation: $C_{\rm det}$

From formulas (2) and (3), we can go ahead and calculate $C_{\rm det}$ in the usual way. Having made decisions at a certain threshold θ, we find the “actual” error rates by summing over trials-in-error

$$P_{\rm FA} = \frac{1}{N_{\rm non}} \sum_{t=T\in{\rm non}} \beta^\alpha_{\rm non}; \qquad P_{\rm miss} = \frac{1}{N_{\rm tar}} \sum_{t=F\in{\rm tar}} \beta^\alpha_{\rm tar}. \tag{4}$$

needed to compute

$$C_{\rm det} = P_{\rm tar} C_{\rm miss} P_{\rm miss} + (1-P_{\rm tar}) C_{\rm FA} P_{\rm FA}, \tag{5}$$

which in its turn is equivalent to analysing conditions α separately and computing a weighted average

$$C_{\rm det} = \sum_\alpha w_\alpha C^\alpha_{\rm det}. \tag{6}$$

[1] through evaluating using a cost function that has externally set target prior and costs for false alarms and misses.

3.2. DET curve, EER and $C_{\rm det}^{\rm min}$

For plotting DET curves, things get slightly more complicated than in the 'pooled trial' case. Normally, each trial in a sorted trial list increases either $P_{\rm FA}$ or $P_{\rm miss}$ by $1/N_{\rm non}$ or $1/N_{\rm tar}$, respectively, but with the condition-weighted probabilities, the step size depends on the condition. A non-target trial in condition α changes the false alarm rate by the amount

$$\Delta P_{\rm FA} = \frac{w_\alpha}{N^\alpha_{\rm non}} = \frac{\beta^\alpha_{\rm non}}{N_{\rm non}}, \tag{7}$$

a target trials changes the miss rate by

$$\Delta P_{\rm miss}\frac{w_\alpha}{N^\alpha_{\rm tar}} = \frac{\beta^\alpha_{\rm tar}}{N_{\rm tar}}. \tag{8}$$

Given these adapted step sizes, we can use the usual cumulative approaches on the sorted scores to compute the DET curve efficiently, and to find post-hoc metrics such as EER and $C_{\rm det}^{\rm min}$.

3.3. Application-independent evaluation: $C_{\rm llr}$

$C_{\rm llr}$ is an evaluation metric proposed by Niko Brümmer that attempts to evaluate the calibration of the scores over more than a single operating point. It can be seen as an integration over $C_{\rm det}$ for a range of cost parameters for $C_{\rm det}$. The calculation of $C_{\rm llr}$ is very similar to $C_{\rm det}$, except that the counting of hard decisions is replaced by a log-error measure of the soft decision score. For further introduction of $C_{\rm llr}$ see [5]. The conditioned version of $C_{\rm llr}$ is expressed as

$$C_{\rm llr} = \frac{1}{2\log 2}\Biggl(\frac{1}{N_{\rm non}}\sum_{t\in{\rm non}}\beta^\alpha_{\rm non}\log(1+e^{s(t)}) + \frac{1}{N_{\rm tar}}\sum_{t\in{\rm tar}}\beta^\alpha_{\rm tar}\log(1+e^{-s(t)})\Biggr). \tag{9}$$

This expressions can be appreciated as a 'log-penalty soft version' of $C_{\rm det}$ in Eqs. (4)–(5). Again, it can also be interpreted as a weighted average over conditions $C_{\rm llr} = \sum_\alpha w_\alpha C^\alpha_{\rm llr}$.

3.4. $C_{\rm llr}^{\rm min}$

For calculating $C_{\rm llr}^{\rm min}$, the minimum value of $C_{\rm llr}$ obtainable by only warping the score scale (i.e., preserving the order of scores), a procedure know as isotonic regression is required, which can be accomplished by, e.g., the Pool Adjacent Violators (PAV) algorithm. Since the warping of the score axis should be performed globally, we cannot perform isotonic regression separately over all conditions and then use a weighted version over the per-condition $C_{\rm llr}^{\rm min}$, as we have shown is possible for $C_{\rm det}$ and $C_{\rm llr}$. Rather, we *have* a to weight each trial as introduced in (2), and use a weighted version of the isotonic regression algorithm.

3.5. Practical implementation of weighted pooling of conditions

By using a weight β for each trial, that is dependent on the condition α and whether it is a target or non-target trial, existing infrastructure can be used to produce DET plots, calculate EER and $C_{\rm det}^{\rm min}$. All that is needed is a minor adaptation to the code such that integer counts/steps of 1 are replaced by the trial's weight $\beta^\alpha_{\rm tar,non}$.

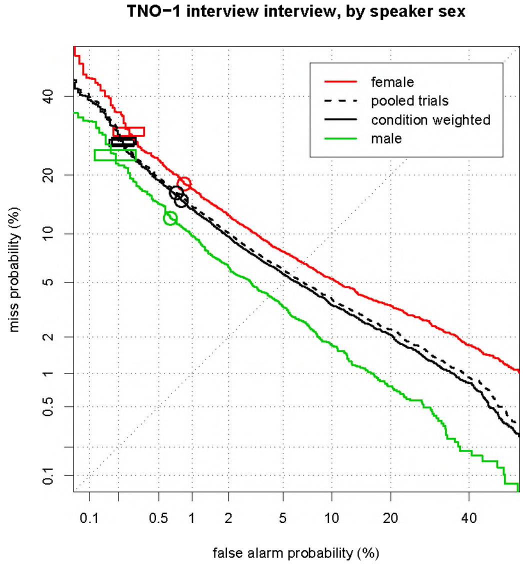


Figure 2: DET curves obtained for TNO-1 'interview interview' common condition, conditioning on speaker sex. Dashed black is the traditional pooled trial analysis (corresponding to NIST common condition 1 analysis), solid black is the proposed condition-weighted analysis.

Analysis	$C_{\rm llr}$	EER (%)	$C_{\rm det}$	$N_{\rm tar}$	$N_{\rm non}$
female	0.277	6.72	0.0328	6639	13137
pooled trials	0.238	5.63	0.0301	11540	22641
condition weighted	0.230	5.41	0.0296	11540	22641
male	0.184	4.04	0.0264	4901	9504

Table 2: Performance figures for the data in Fig. 2a, presented in the same order as the DET curves. The row "pooled trials" corresponds to NIST common condition 1.

4. Application examples of weighted averaging of conditions

4.1. Speaker sex

We will start by a simple example, showing the influence of the slight imbalance of speaker sex trials in traditional analysis. As data we use all interview trials of the TNO-1 submission. In Fig. 2 we have separated the DET curves conditioned on speaker sex, and show traditional (dashed) and condition-weighted analysis. Relevant performance figures are in Table 2.

Apart from the obvious difference in performance between male and female speaker trials, there is the slight effect of the number of female trials on the pooled results, raising error rates w.r.t. condition-weighted analysis. Admittedly, the effect is small.

4.2. Acoustic condition

We will now present the results when we combine all 5 acoustic conditions that occur in the "short2-short3" core condition trial list. The pooled data analysis has been shown earlier in Fig. 1 as the solid black line, and now using the weighted approach, we obtain the dashed black line plotted in the same graph. For comparison, we tabulated the performance metrics for the two approaches in Table 3.

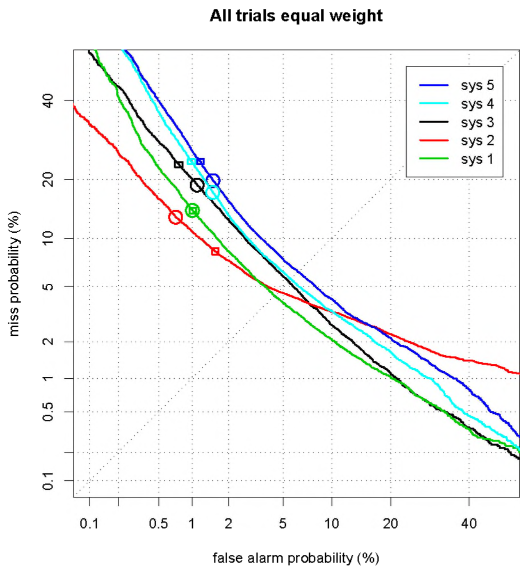
The effect may not seem dramatic, but it changes the position of the DET curve quite a bit for the TNO system, moving

Analysis	C_{llr}	EER (%)	C_{det}	N_{tar}	N_{non}
Pooled	0.250	5.62	0.0338	20449	78327
Weighted	0.233	5.00	0.0283	20449	78327

Table 3: Comparison of performance metrics between the 'naive' pooled trials analysis and the new condition weighted analysis. The data is from the TNO-1 submission, analyzing all trials.

it more towards the middle of the pack. We attribute this to the fact that the 'interview-interview' trials, which this system did not perform extremely well, are less dominant in the weighted condition.

We've applied this condition weighing to the submitted scores of some of the better performing sites who were willing to share them for this purpose. In Figures 3*a* and *b* one can appreciate that the apparent diverse performance seems to be normalized a bit by our equal weighting of the acoustic conditions.[2] Further, notice that the effect of equal weighting is not necessarily lowering the DET curve. For system 1, which performed very well in the interview-interview condition, removing the relative weight of this conditions actually raises the overall DET curve a bit.



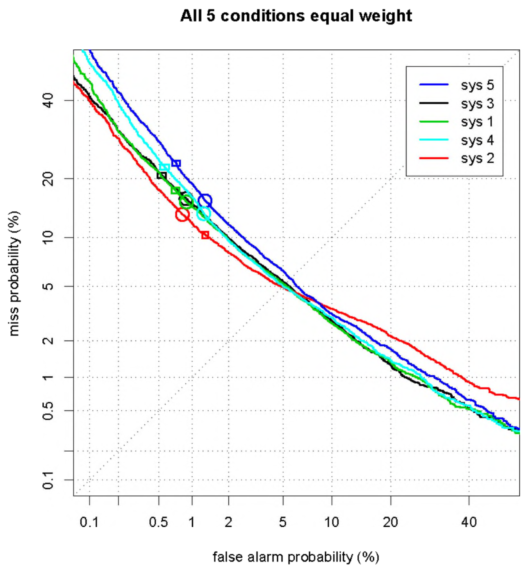


Figure 3: DET curves for five systems in SRE-2008, where (*a*, top) all trials all pooled, and (*b*, right) the 5 acoustic conditions are equally weighted.

5. Conclusions

We argue that both from a detection and calibration point of view, it is an interesting task to develop a speaker recognition system that is robust against different conditions of the train and test data. In order to evaluate such a system, which is a necessary step during the development, a good metric needs to be used. We proposed a metric that simply corrects for the different proportion of trials in the various conditions. By using a trial weighting that reflects the relative proportion of the trial's condition w.r.t. other conditions, we derived expressions for C_{det}, C_{llr} and the cumulative quantities P_{FA} and P_{miss} that govern the DET curve, and EER and $C_{\mathrm{det}}^{\mathrm{min}}$ operating points. Finally, the computation of condition-weighted $C_{\mathrm{llr}}^{\mathrm{min}}$ can be accomplished by using an algorithm for isotonic regression that includes weights. We have made our tools available for computing the various performance metrics. [6]

6. Acknowledgments

We would like to thank George Doddington and Niko Brümmer for stimulating discussions. We are further indebted to BUT, I4U, LPT and SUNSDV that provided their system scores so that we could show a broader application of trial weighting in Fig. 3. This work was supported in part by the European Union 6th FWP project AMIDA, 033812.

7. References

[1] Patrick Kenny and Pierre Dumouchel. Disentangling speaker and channel effects in speaker verification. In *Proc. ICASSP*, pages 37–40, 2004.

[2] Simon Lucey and Tsuhan Chen. Improved speaker verification through probabilistic subspace adaptation. In *Proc. Interspeech*, pages 2021–2024, Geneva, 2003. ISCA.

[3] William Campbell, Douglas Sturim, Douglas Reynolds, and Alex Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. ICASSP*, pages 97–100, Toulouse, 2006. IEEE.

[4] Claudio Vair, Daniele Colibro, Fabio Castaldo, Emanuele Dalmasso, and Pietro Laface. Channel factors compensation in model and feature domain for speaker recognition. In *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.

[5] David A. van Leeuwen and Niko Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In Christian Müller, editor, *Speaker Classification*, volume 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, Heidelberg - New York - Berlin, 2007.

[6] David A. van Leeuwen. SRE-tools, a software package for calculating performance metrics for NIST speaker recognition evaluations. `http://sretools.googlepages.com/`, 2008.

[2]The purpose of this paper is not to compare systems directly, and therefore we have anonymized the entries.