A note on performance metrics for Speaker Recognition using multiple conditions in an evaluation

David A. van Leeuwen

9 June 2008

Abstract

In this paper we put forward arguments for pooling different evaluation conditions for calculating speaker recognition system performance measures. We propose a condition-based weighting of trials, and derive expressions for the basic speaker recognition performance measures C_{det}, C_{llr}, as well as the DET curve, from which EER and $C_{\mathrm{det}}^{\mathrm{min}}$ can be computed. We show that trials-based weighting is essential for computing $C_{\mathrm{llr}}^{\mathrm{min}}$ in a pooled condition evaluation. Examples of pooling of conditions are show 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, the importance of so-called auxiliary microphone conditions were stressed, and it was remarked that not many sites participated in this separate evaluation 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 will optimize the EER and cost function for such pooled conditions will not only make systems more robust to these 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 the individual condition.

Condition	NIST	C_{llr}	EER (%)	C_{det}	N_{tar}	N_{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.

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.[1]

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_{llr}. We will show the effects of this new approach using the TNO primary system submission data.

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_{det} and pooling scores for the DET curve ($C_{\text{det}}^{\text{min}}$, EER). In Figure 1 we show the effect of pooling in a DET plot, where the 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, are plotted in 'plot-9' graphs).

Several remarks can be made about the plot. 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 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_{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 sites who have tried to get calibration right over all conditions.

Note, that thus far 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

[1] During the writing of this document, George Doddington pointed out his `DET_tools` perl package, that allows for an analysis pooling DET curve statistics, which is basically the same approach of equalizing the weights of different trial conditions.

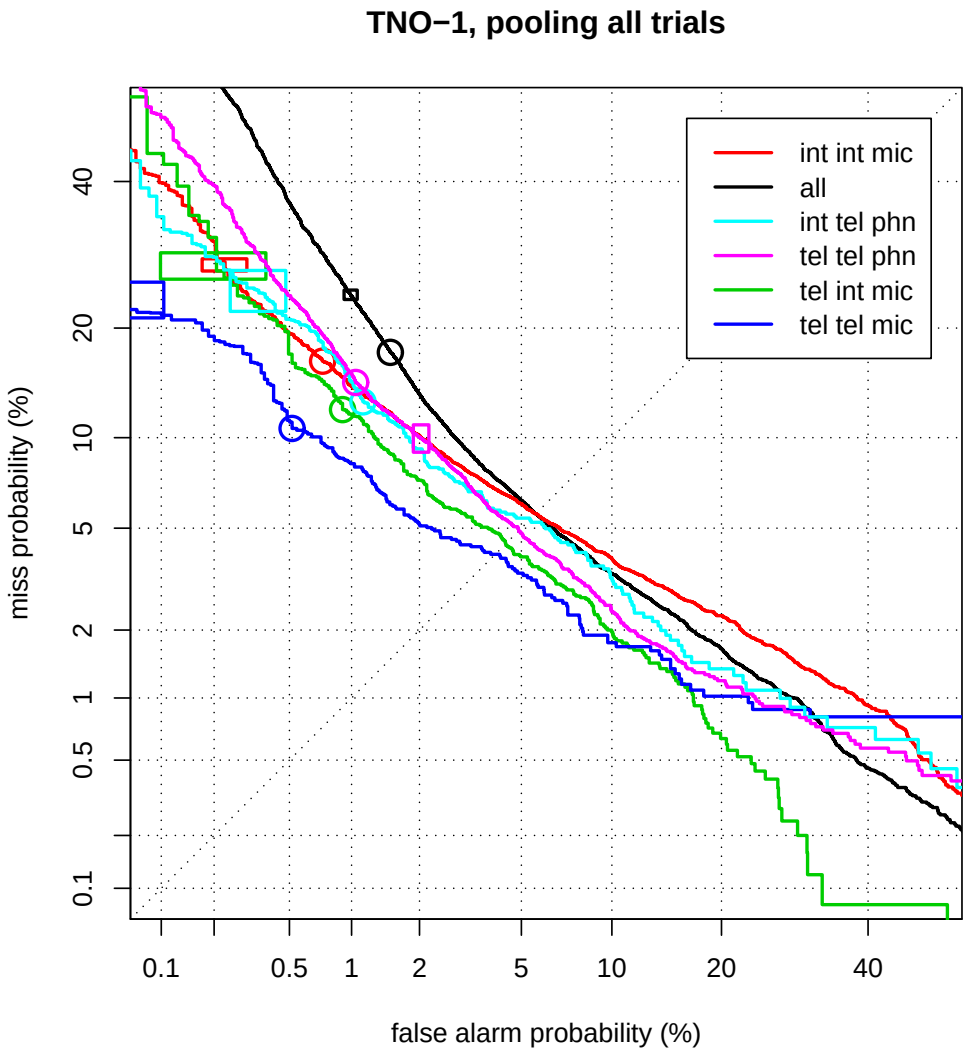

Figure 1: DET curves obtained for TNO-1 in NIST SRE-2008, after pooling all trials in the "short2-short3" core test condition. In colour, DET curves are conditioned on acoustic condition, where 'int' indicates interview style, 'tel' phonecall style, 'phn' recording test segment over phone handset, 'mic' recording over auxiliary microphone. The fist 'int/tel' designates training condition, the second test condition.

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[2], disentangling the evaluation priors from the application priors, we can give the trials in each acoustical condition separate weights. We define the probability of false alarm at a given threshold θ for trials in condition[3] α as

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

Where N_{non}^{α} is the number of non-target trials in condition α, and the sum is the number of false positive trials in this condition, using the unit step function u for counting these trials with score s above the threshold. Similarly, we can defined a conditioned miss probability

$$P_{\text{miss}}^{\alpha}(\theta) = \frac{1}{N_{\text{tar}}^{\alpha}} \sum_{t \in \text{tar},\alpha} u(\theta - s(t)) \tag{2}$$

These formulas are nothing new, they represent the usual estimation of P_{FA} and P_{miss}, but now include the conditioning on α.

The proposal now for computing a performance measure over all trials of an evaluation, is to simply weight the individual conditioned error rates,

$$P_{\text{FA}}(\theta) = \sum_{\alpha} w_{\alpha} P_{\text{FA}}^{\alpha}(\theta), \tag{3}$$

$$P_{\text{miss}}(\theta) = \sum_{\alpha} w_{\alpha} P_{\text{miss}}^{\alpha}(\theta). \tag{4}$$

The weights w_{α} (summing to unity) are the externally defined weights of interest 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. Alternatively, one might prefer to choose $w(\text{tel int mic}) = 0$, to be more in line with the conditions analyzed by NIST.

3.1 Traditional evaluation: C_{det}

From these P_{FA} and P_{miss}, we can go ahead and calculate C_{det} in the usual way. Using hard decision, rather than soft scores, we have

$$P_{\text{FA}}^{\alpha} = N_{\text{non}}^{\alpha}(T)/N_{\text{non}}^{\alpha} \tag{5}$$

$$P_{\text{miss}}^{\alpha} = N_{\text{tar}}^{\alpha}(F)/N_{\text{tar}}^{\alpha} \tag{6}$$

where the numerators count the number of wrong decisions conditioned on α and target/non-target. The actual conditioned miss and false alarm rates can then be averaged as in eqs. 3 and 4, and used in the cost function $C_{\text{det}} = P_{\text{tar}} C_{\text{miss}} P_{\text{miss}} + (1 - P_{\text{tar}}) C_{\text{FA}} P_{\text{FA}}$, where P_{tar}, C_{miss} and C_{FA} are the cost function parameters. Note that C_{det} could also have been obtained as a weighted average over C_{det}^{α} calculated over conditioned parts of the trial set.

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

[3] We removed the adjective 'acoustical' for condition, because one can condition for anything, including sex or even target speaker

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

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_{FA} or P_{miss} by $1/N_{\text{non}}$ or $1/N_{\text{tar}}$, 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

$$\Delta P_{\text{FA}} = \frac{w_\alpha}{N_{\text{non}}^\alpha}, \tag{7}$$

a target trials changes the miss rate by

$$\Delta P_{\text{miss}} \frac{w_\alpha}{N_{\text{tar}}^\alpha}. \tag{8}$$

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

3.3 Application-independent evaluation: C_{llr}

C_{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_{det} for a range of cost parameters for C_{det}. The calculation of C_{llr} is very similar to C_{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_{llr} see [5]. The conditioned version of C_{llr} is expressed as

$$C_{\text{llr}}^\alpha = \frac{1}{2\log 2}\Big(\frac{1}{N_{\text{non}}^\alpha} \sum_{t\in\text{non},\alpha} \log(1+e^s) + \tag{9}$$

$$\frac{1}{N_{\text{tar}}^\alpha} \sum_{t\in\text{tar},\alpha} \log(1+e^{-s})\Big) \tag{10}$$

from which the weighted average

$$C_{\text{llr}} = \sum_\alpha w_\alpha C_{\text{llr}}^\alpha \tag{11}$$

can be calculated.

3.4 $C_{\text{llr}}^{\text{min}}$

For calculating $C_{\text{llr}}^{\text{min}}$, the minimum value of C_{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 $C_{\text{llr}}^{\text{min}}$. Rather, we need a to weight each trial, and use a weighted version of the isotonic regression algorithm. In order to weight trials individually such that the trial set can be treated as a whole, we compute the values

$$\beta_{\text{tar}}^\alpha = w_\alpha \Big/ \frac{N_{\text{tar}}^\alpha}{N_{\text{tar}}}; \qquad \beta_{\text{non}}^\alpha = w_\alpha \Big/ \frac{N_{\text{non}}^\alpha}{N_{\text{non}}}. \tag{12}$$

These β can be interpreted as weights for individual trials to the isotonic regression. They measure the ratio of the desired weight of condition α to the actual proportion of trials in that condition. These weights β hence either boost or diminish the influence of a trial in condition α.

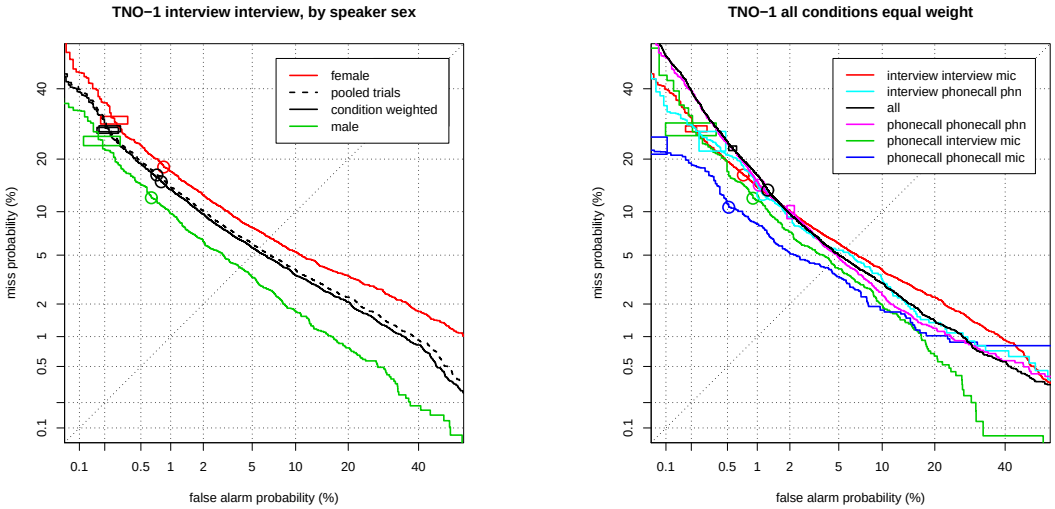

Figure 2: *a.* (left) DET curves obtained for TNO-1 'interview interview' common condition, conditioning on speaker sex. Dashed black is the traditional pooled trial analysis (corresponding to NIST common condition 1 analysis), solid black is the proposed condition-weighted analysis. *b.* (right) DET curves similar to Fig. 1. The difference is that the over-all trials curve (black) has been obtained by weighting the individual curves to equalize their contribution.

3.5 Practical implementation of weighted pooling of conditions

The weights β are useful for more than just the isotonic regression necessary for computing $C_{\text{llr}}^{\text{min}}$. In fact, we can use these weights for individual trials to compute cumulative P_{FA} and P_{miss} for DET plots in the ordinary way, removing the need for conditioned versions as in (1) and averaging afterward. Combining Eqs. 1, 3 and 12 we can derive the false alarm rate at threshold θ as

$$P_{\text{FA}}(\theta) = \frac{1}{N_{\text{non}}} \sum_{t \in \text{non}} \beta_{\text{non}}^{\alpha(t)} u(s(t) - \theta). \tag{13}$$

The advantage of this formulation is that existing infrastructure can be used to produce DET plots, calculate EER and $C_{\text{det}}^{\text{min}}$, after a minor adaptation to the code such that integer counts/steps of 1 are replaced by the trial's weight $\beta_{\text{tar,non}}^{\alpha}$. In our implementation of the speaker recognition performance evaluation tools in the statistical programming language R, we have even used these weighted trials for calculating C_{det} and C_{llr}, see Appendix A for the detailed expressions.

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. 2a 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.

Analysis	C_{llr}	EER (%)	C_{det}	N_{tar}	N_{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.

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.

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, and now using the weighted approach, we obtain the DET plot in Fig. 2b. 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 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 a number of other site's who were willing to share them for this purpose. In Figures 3a and b one can appreciate that the apparent diverse performance seems to be normalized a bit by our equal weighting of the acoustic conditions.[4] Further, notice that the effect of equal weighting is not necessarily lowering the DET curve. For one system, which performed very well in the interview-interview condition, removing the relative weight of this conditions actually raises the overall DET curve a bit.

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 $C_{\text{det}}^{\text{min}}$ operating points. Finally, the computation of condition-weighted $C_{\text{llr}}^{\text{min}}$ 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 would like the sites that provided their system scores for doing this, so that we could show a broader application of trial weighting in Fig. 3.

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

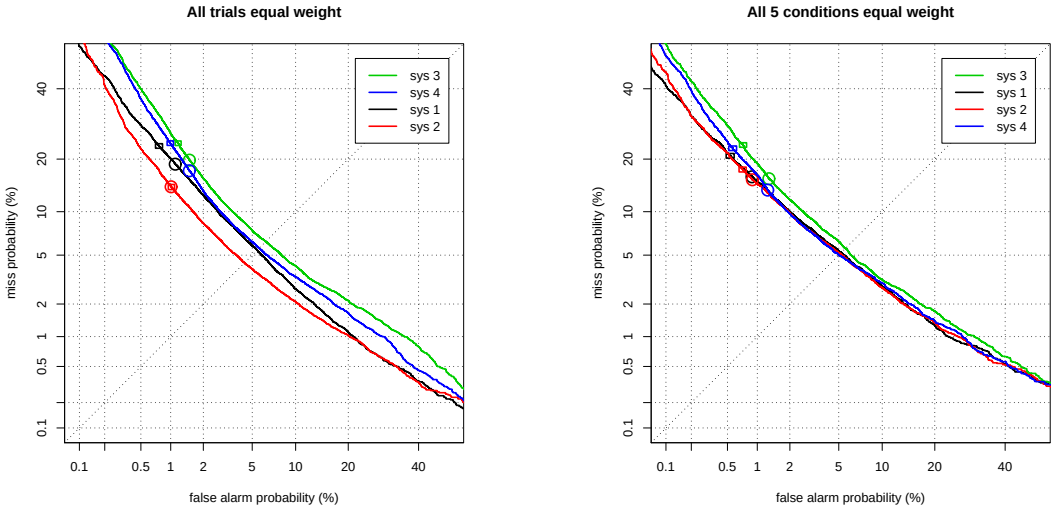

Figure 3: DET curves for 4 systems in SRE-2008, where *a* (left) all trials all pooled, and *(*b) (right) the 5 acoustic conditions are equally weighted.

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.

A Trials-weighted expression for various performance measures.

Here we reproduce, for completeness, how the various performance measures can be computed using the trial weights introduced in Sect. 3.5. The counterpart of $P_{\mathrm{FA}}(\theta)$

$$P_{\mathrm{FA}}(\theta) = \frac{1}{N_{\mathrm{non}}} \sum_{t \in \mathrm{non}} \beta_{\mathrm{non}}^{\alpha(t)} u(s(t) - \theta). \tag{14}$$

is the miss rate at threshold θ

$$P_{\text{miss}}(\theta) = \frac{1}{N_{\text{tar}}} \sum_{t \in \text{tar}} \beta_{\text{tar}}^{\alpha(t)} u(\theta - s(t)). \tag{15}$$

For plotting DET curves the step sizes in Eqs. 7 and 8 become:

$$\Delta P_{\text{FA}} = \beta_{\text{non}}^{\alpha} / N_{\text{non}}; \qquad \Delta P_{\text{miss}} = \beta_{\text{tar}}^{\alpha} / N_{\text{tar}}. \tag{16}$$

From these, EER can be estimate in your favorite way. One approach is to take $P_{\text{miss}}(i)$ or $P_{\text{FA}}(i)$ where $|P_{\text{miss}}(i) - P_{\text{FA}}(i)|$ is minimum, or you can interpolate between the two, or in the case of very ragged DET curves use a convex hull of the ROC curve for interpolation.[5] Above, we have used the index i for the miss and false alarm rates for sorted scores

$$P_{\text{FA}}(i) = 1 - \frac{1}{N_{\text{non}}} \sum_{j=1,\text{non}}^{i} \beta_{\text{non}}^{\alpha(j)}; \qquad P_{\text{miss}}(i) = \frac{1}{N_{\text{tar}}} \sum_{j=1,\text{tar}}^{i} \beta_{\text{tar}}^{\alpha(j)}, \tag{17}$$

where the summation is only over scores from non-target and target trials, respectively. The value for $C_{\text{det}}^{\text{min}}$ can be found quickly as

$$C_{\text{det}}^{\text{min}} = \min_{i} C_{\text{det}}(P_{\text{FA}}(i), P_{\text{miss}}(i)). \tag{18}$$

The actual detection costs $C_{\text{det}}(P_{\text{FA}},P_{\text{miss}})$ is found by summing trial-weighted decisions-in-error,

$$P_{\text{FA}} = \frac{1}{N_{\text{non}}} \sum_{t \in T,\text{non}} \beta_{\text{non}}^{\alpha}; \qquad P_{\text{miss}} = \frac{1}{N_{\text{tar}}} \sum_{t \in F,\text{tar}} \beta_{\text{tar}}^{\alpha}. \tag{19}$$

$$C_{\text{det}}(P_{\text{FA}}, P_{\text{miss}}) = P_{\text{tar}} C_{\text{miss}} P_{\text{miss}} + (1 - P_{\text{tar}}) C_{\text{FA}} P_{\text{FA}} \tag{20}$$

Here the summations run over trials in error, and we have reproduced the definition of C_{det} for cost parameters P_{tar}, C_{miss} and C_{FA}.

Finally, we will give the expression for C_{llr} using weighted trials,

$$C_{\text{llr}} = \frac{1}{2 \log 2} \Big(\frac{1}{N_{\text{non}}} \sum_{t \in \text{non}} \beta_{\text{non}}^{\alpha} \log(1 + e^{s(t)}) + \tag{21}$$

$$\frac{1}{N_{\text{tar}}} \sum_{t \in \text{tar}} \beta_{\text{tar}}^{\alpha} \log(1 + e^{-s(t)}) \Big). \tag{22}$$

This expressions can be appreciated as a 'log-penalty soft version' of C_{det} in Eqs. (19)–(20).

[5] Thanks to Niko Brümmer for pointing this out.