The Role of Score Calibration in Speaker Recognition

George Doddington

george.doddington@gmail.com

Abstract

The performance of speaker recognition, as measured by NIST in its periodic evaluation programs of speaker detection technology, is affected by the predictive accuracy of system decisions as well as the inherent discriminating power of the underlying algorithms. This paper analyses the score calibration accuracy of a few systems that exhibited good performance in the NIST 2010 speaker recognition evaluation (SRE10) and proposes a modified cost function that better represents the accuracy of score calibration in the low false alarm decision region, which is the region of interest for intelligence applications.

Index Terms: speaker recognition, speaker detection, score calibration

1. Introduction

NIST has fostered research in speaker recognition through a continuing series of evaluations. [1] In these evaluations the general speaker recognition problem has been framed as a detection task, with performance characterized in terms of miss and false alarm probabilities. While most research effort toward this objective is focused on improving performance generally without regard to particular tasks or applications, the job of setting detection thresholds and making decisions remains an important consideration and a significant technical challenge. This motivates NIST to feature actual decisions in its periodic evaluations, and the accuracy of determining thresholds often is the deciding factor in determining which system among the highest performing systems exhibits the best performance.

One of the alternative performance measures used by NIST in their periodic evaluations is the log likelihood ratio cost function, defined as:

\[ C_{llr} = \frac{1}{2 \ln(2)} \left\{ \frac{N_{target}^{-1} \sum_{target \ trials} \ln(1 + S_i^{-1})}{N_{non-target}^{-1} \sum_{non-target \ trials} \ln(1 + S_i)} \right\} \]

where

\( N_{target} \) is the number of target trials,

\( N_{non-target} \) is the number of non-target trials, and

\( S_i \) is the likelihood ratio for trial \( i \) (i.e., the probability of the observation data given the target divided by the probability of the observation given not the target).

This measure is attributable to Niko Brummer [2] and has an information theoretic interpretation. It is of particular interest in this study because it explicitly penalizes score calibration errors.

While \( C_{llr} \) has the desirable characteristic of representing errors in score calibration, the contribution of these calibration errors are equally weighted across all trials, including trials outside the region of application interest. As a result, the ranking of systems according to \( C_{llr} \) performance is seen to be at odds with the ranking of systems at the decision point specified in the NIST evaluation plan using the conventional primary cost-based performance measure. This defect can be repaired by limiting the range of computation of \( C_{llr} \) as will be shown below.

2. The Data

The data used were SRE10 extended evaluation [1] results submissions from four participating sites. For the purpose of this study, consideration was limited to just condition 2, interview speech collected from multiple different microphones, with the training and the test sample microphones being different. Condition 2 provides the largest number of trials for analysis.

Table 1 SRE10 extended data speaker and trial statistics

<table>
<thead>
<tr>
<th>Condition #</th>
<th>sex</th>
<th># of speakers</th>
<th># of target trials</th>
<th># of non-target trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 - interview speech</td>
<td>female</td>
<td>229</td>
<td>8152</td>
<td>1,573,948</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>196</td>
<td>6932</td>
<td>1,215,586</td>
</tr>
</tbody>
</table>

The four systems studied all provided output scores intended to represent the target/non-target log likelihood ratio. Thus these scores are amenable to evaluation using \( C_{llr} \), and actual decisions can be inferred from these scores that will minimize the detection cost, assuming that the scores are correctly calibrated.

3. Performance Measures

In addition to \( C_{llr} \), three other performance measures will be used:

- NIST’s primary cost measure for SRE10, with the cost based on decisions made using a score threshold that would minimize the cost assuming correctly calibrated scores (i.e., assuming scores that accurately represent the log likelihood ratio of the data),
- NIST’s primary cost measure for SRE10, with the cost based on decisions made using a score threshold that minimizes the cost over the test set, with this threshold and the resulting cost determined \( ex post facto \), and
- The false alarm probability given a threshold determined to yield a miss probability of 10 percent over the test set. This performance measure is the primary performance measure used in LARPA’s BEST program. [3]

The cost function used in SRE10 is a simple combination of miss and false alarm probabilities, weighted by the prior probability of the target:
\[ C_{\text{Det}} = C_{\text{Miss}} \times P_{\text{Target}} \times P_{\text{Miss|Target}} \]
\[ + C_{\text{FalseAlarm}} \times (1-P_{\text{Target}}) \times P_{\text{FalseAlarm|Non-target}} \]

The parameters of this cost function are:
- \( C_{\text{Miss}} \): the cost of a miss,
- \( C_{\text{FalseAlarm}} \): the cost of a false alarm, and
- \( P_{\text{Target}} \): the \textit{a priori} probability that the segment speaker is the target speaker, and

Table 2: SRE10 Speaker Detection Cost Model Parameters

<table>
<thead>
<tr>
<th></th>
<th>( C_{\text{Miss}} )</th>
<th>( C_{\text{FalseAlarm}} )</th>
<th>( P_{\text{Target}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0.001</td>
</tr>
</tbody>
</table>

To improve the intuitive meaning of \( C_{\text{Det}} \), NIST’s primary cost measure, \( C_{\text{Primary}} \), is defined by dividing \( C_{\text{Det}} \) by the best cost that could be obtained without knowledge of the input data:

\[ C_{\text{Primary}} = \frac{C_{\text{Det}}}{C_{\text{Default}}} \]

where \( C_{\text{Default}} = C_{\text{Miss}} \times P_{\text{Target}} \)

Thus

\[ C_{\text{Primary}} = P_{\text{Miss|Target}} + \beta \times P_{\text{FalseAlarm|Non-target}} \]

where \( \beta = \left( \frac{C_{\text{FalseAlarm}}}{C_{\text{Miss}}} \right) \left( \frac{1-P_{\text{Target}}}{P_{\text{Target}}} \right) \)

For optimum performance, decisions are made to minimize the cost. Thus for log likelihood scores less than \( \log(\beta) \) the target hypothesis should be rejected, and for scores greater than \( \log(\beta) \) the hypothesis should be accepted. If the scores are correctly calibrated, then the detection cost, \( C_{\text{Primary}} \) will be minimized.

4. SRE10 Performance Results

First, in order to provide a context which might help to understand the comparison of the various performance measures, the detection error trade-off curves for the four systems being used are shown in figure 1. Looking at figure 1, the situation seems fairly straightforward. Namely, the best performance is achieved by system A, second by system B, third by system C and fourth by system D. Only for system B might there be some uncertainty, because the ranking of system B changes depending on the particular value of \( P_{\text{Miss}} \) (or \( P_{\text{FA}} \)) at which the ranking is computed.

With this as a context, now look at figure 2, which compares the performance of the four systems for each of the four performance measures. The easiest performance measure to correlate with the DET curve is the false alarm probability at 10% \( P_{\text{Miss}} \). This can be read directly off the DET plot. The actual primary cost function DET points and the minimum cost function DET points are given in table 3. The most notable observations in figure 2 are the miscalibrated scores for systems A and D that result in actual detection costs on the test data that are much greater than the minimum costs on the test data. Yet this miscalibration does not appear to be reflected in the \( C_{\text{llr}} \) measures. To explore this in greater depth, figure 3 shows a scatter plot of each system’s log likelihood scores versus an estimate of the actual log likelihood computed for that score.
The estimate of the actual log likelihood is computed by averaging the miss and false alarm statistics over a small region around the log likelihood score. The formula used is:

\[
\text{LLR}_{\text{estimated}}(s) = \log\left(\frac{P_{\text{Miss}}(s + 0.1) - P_{\text{Miss}}(s - 0.1)}{P_{\text{FA}}(s - 0.1) - P_{\text{FA}}(s + 0.1)}\right)
\]

Figure 3 shows that score calibration for systems B and C was fairly good for high scoring trials in the low false alarm region of application importance, while score calibration for systems A and D was not as good in this region. On the other hand, the calibration of scores for systems A and D was fairly good for the low scoring (mostly non-target) trials, whereas the calibration of scores in this region was not as good for system B.

In order to see the effect of score miscalibration on \(C_{\text{lr}}\) more clearly, it would be helpful to compare \(C_{\text{lr}}\) when computed on the actual system output scores with the scores that would be obtained by linearly transforming the scores so as to minimize \(C_{\text{lr}}\). Figure 4 is a scatter plot of the estimated log likelihood ratio as a function of these transformed scores, and figure 5 compares \(C_{\text{lr}}\) for the four systems before and after the linear transformation.

After the linear transformation, the minimized \(C_{\text{lr}}\) scores for the four systems better reflect what we visually observe in the DET curves, which hide any score miscalibration. This is not to suggest that such a linear transformation be performed prior to computing a \(C_{\text{lr}}\) performance measure. Quite to the contrary, this demonstrates that \(C_{\text{lr}}\) embodies calibration errors, which are important in judging the field ability and application readiness of systems. Comparing \(C_{\text{lr}}\) before and after linear transformation is analogous to comparing NIST’s primary cost function using the predetermined threshold versus the threshold that minimizes the cost.

To make \(C_{\text{lr}}\) more relevant to the NIST region of interest, there needs to be a means of focusing \(C_{\text{lr}}\) on this region. One method of achieving this is simply to limit the computation of \(C_{\text{lr}}\) to only those trials for which the score is greater than some threshold chosen to yield a probability of miss that is deemed to be a bound on the region of interest. For NIST’s region of interest, and for the current state of speaker recognition technology, 10% \(P_{\text{Miss}}\) could serve as a reasonable lower bound defining this region of interest. Let us call this limited computation version of \(C_{\text{lr}}\) to be \(C_{\text{lr-M10}}\). Figure 6 compares \(C_{\text{lr}}\) with \(C_{\text{lr-M10}}\).
Note that figure 6 shows that the ranking of system B changes dramatically, from rank 3 under $C_{llr}$ to rank 1 under $C_{llr-M10}$.

The fact that $C_{llr-M10}$ performance for system B is better than the $C_{llr-M10}$ performance for system A, despite the detection error trade-off being everywhere better for system A than for system B, is attributable to the fact that the score calibration for system B is better than that for system A in the region of interest. If the scores are linearly transformed to minimize $C_{llr-M10}$, then the ranking of the systems according to the minimized $C_{llr-M10}$ changes once again, as illustrated in figures 7 and 8.

![Figure 7](image1.png)

Figure 7 Log likelihood ratio estimates as a function of the system output scores linearly transformed to minimize $C_{llr-M10}$.

![Figure 8](image2.png)

Figure 8 A comparison of $C_{llr-M10}$ computed on the original scores with $C_{llr-M10}$ computed on the scores after being linearly transformed to minimize $C_{llr-M10}$.

5. Summary and Cautionary Note

This follow-up study to the SRE10 extended data evaluation shows the role that accurate calibration of log likelihood scores plays, both in NIST’s evaluation and in application readiness. Correct calibration of scores can and often does result in better performance relative to competing systems.

Note that correct calibration of log likelihood ratio scores has been evaluated here in a very narrowly defined context. The population of speakers is well known and relatively well-behaved, and the acoustical and data collection conditions are benign. Very different results may obtain from different or less well-controlled conditions, and accurate calibration of scores under such conditions remains a daunting challenge. Fielding real systems into real applications is, of course, the ultimate challenge.

6. References