ABSTRACT
Score calibration enables automatic speaker recognizers to make cost-effective accept / reject decisions. Traditional calibration requires supervised data, which is an expensive resource. We propose a 2-component GMM for unsupervised calibration and demonstrate good performance relative to a supervised baseline on NIST SRE’10 and SRE’12. A Bayesian analysis demonstrates that the uncertainty associated with the unsupervised calibration parameter estimates is surprisingly small.

Index Terms—calibration, unsupervised learning, Laplace approximation, automatic speaker recognition

1. INTRODUCTION
Automatic speaker recognizers map trials to scores. A trial has two parts: some speech of a known speaker, and some of an unknown speaker. When the same speaker is present in both parts, we have a target trial. When the speakers differ, we have a non-target trial. The score is a real number, which is expected to be larger (more positive) for target trials and smaller (more negative) for non-target trials. When a speaker recognizer is deployed in a new environment, the behaviour of the scores may change between environments. The ideal calibration of a score, \( s \), would be of the form:

\[
\begin{align*}
  P(s | \text{target, environment}) & \rightarrow \log \frac{P(s | \text{target, environment})}{P(s | \text{non-target, environment})} \\
  \text{but of course, we are not given these score distributions. Our only resource would be some data collected from the new environment. In special cases (usually involving considerable expense), this data can be supervised, such that each trial is labelled as target or non-target. In a more realistic scenario however, all or most of this data would be unsupervised.}
\end{align*}
\]

To date, most works on calibration have made use of supervised data. In this paper, we explore the problem of calibration where our only resource is a large database of completely unsupervised scores.

2. CALIBRATION MODEL
In the supervised setting, score calibration can be viewed as a straight-forward, 2-class pattern recognition problem, for which both generative and discriminative solutions exist [1, 2, 3, 4, 5]. For the unsupervised case, we found the generative approach more convenient. Here we introduce the score model for supervised calibration, followed by a generalization to the unsupervised case.

2.1. Supervised calibration model
For the supervised case, we adopt the simple generative model of [4]. Denoting targets as \( H_1 \) and non-targets as \( H_2 \), we model a score \( s \in \mathbb{R} \), with class-conditional Gaussian distributions:

\[
P(s | H_i, C) = N(s | \mu_i, \sigma^2)
\]

where \( \mu_1, \mu_2 \) are class-conditional means and \( \sigma^2 \) is the common within-class variance. We collectively refer to \( C = (\mu_1, \mu_2, \sigma^2) \) as the calibration parameters. This model gives an affine transformation, from score to log-likelihood-ratio, of the form:

\[
\log R(s | C) = \log \frac{N(s | \mu_1, \sigma^2)}{N(s | \mu_2, \sigma^2)} = \frac{d' \sigma}{\sigma^2} s + \frac{\mu_2 - \mu_1}{2\sigma^2}
\]

where \( d' \), the separation between targets and non-targets [6],

\[
d' = \frac{\mu_1 - \mu_2}{\sigma},
\]

represents accuracy, since the theoretical equal-error-rate is EER = \( \Phi(-\frac{d'}{2\sigma}) \).

\( \Phi \) is the normal cumulative density, given in terms of the error-function as \( \Phi(x) = (1 + \text{erf}(x/\sqrt{2}))/2. \)

2.2. Unsupervised calibration model
In the unsupervised case, we are given a collection of \( T \) training scores, denoted \( S = s_1, \ldots, s_T \), but we are not given the corresponding class labels. Denoting these labels as \( \mathcal{L} = \ell_1, \ldots, \ell_T \), we treat them as hidden variables and our calibration model generalizes to a 2-component Gaussian mixture model (GMM), for which we need an additional mixture-weight parameter, \( \pi_1 \). Letting \( \pi_2 = 1 - \pi_1 \), the GMM likelihood is:

\[
P(S | \mathcal{M}) = \prod_{t=1}^{T} \sum_{i=1}^{2} \pi_i N(s_t | \mu_i, \sigma^2)
\]
where $\mathcal{M} = (\mathcal{C}, \pi_1) = (\mu_1, \mu_2, \sigma^2, \pi_1)$ are the GMM parameters.

Now consider a new test score, $s'$, generated by the same model, and with associated hidden class label $\ell'\in\{H_1, H_2\}$. The conditional dependency structure of all the parameters and variables can be summarized in graphical model [7] notation as:

$$\pi_1 \rightarrow \mathcal{L} \rightarrow S \leftarrow \mathcal{C} \rightarrow s' \leftarrow \ell' \leftarrow \pi'_1 \tag{5}$$

where we have introduced an independent prior, $\pi'_1$, for $\ell'$. The importance of this diagram cannot be overstressed. It is used repeatedly below, to be able to remove irrelevant conditioning terms.2

Our end-goal will be to infer the value of $\ell'$, when $S$ and $s'$ are observed and $\pi'_1$ is given, but where $\pi_1, \mathcal{L}, \mathcal{C}$ are unknown nuisance variables. The result could be given as the posterior $P(\ell'|S, s', \pi'_1)$, or equivalently, as the predictive likelihood-ratio:

$$R(s'|S) = \frac{P(s'|\ell' = H_1, S)}{P(s'|\ell' = H_2, S)} \tag{6}$$

Our end-goal is the calibration of $s'$ via the mapping:

$$s' \rightarrow \log R(s'|S)$$

### 3. Inference

Here we discuss computational strategies to compute $R(s'|S)$. The computation involves summing over all of the hidden labels, $\mathcal{L}$, and integrating out the parameters, $\mathcal{M}$. Unfortunately this cannot be done in closed form. If conjugate priors are used, the parameters can be integrated out in closed form, but this makes the labels dependent, so that summing them out requires an intractable sum over $2^\ell$ terms. Conversely, if you start with the labels, they can be summed out in closed form, using (4), but then the parameters cannot be integrated out in closed form. For this work, we shall follow the latter route, because the parameter space is just 4-dimensional, allowing approximate integration in this space.

The numerator and denominator of (6) are obtained by marginalizing w.r.t. $\mathcal{C}$ and simplifying by (5):

$$R(s'|S) = \langle \frac{P(s'|\mathcal{C}, H_1)}{P(s'|\mathcal{C}, H_2)} \rangle_S \tag{7}$$

where $\langle \cdot \rangle_S$ denotes expectation w.r.t. the parameter posterior $P(\mathcal{C}|S)$. We show below how to derive $P(\mathcal{C}|S)$ from the Laplace approximation for $P(\mathcal{M}|S)$.

Although we shall use (7) in practice, we develop an interesting alternative form below that helps to theoretically illuminate the relationship between the plug-in and predictive LRs.

#### 3.1. Plug-in vs predictive LR

The predictive likelihoods can be expanded by the product rule as:

$$P(s'|S, H_1) = P(s'|\mathcal{C}, H_1) \times \frac{P(\mathcal{C}|S)}{P(\mathcal{C}|S, s', H_1)} \tag{8}$$

where a ratio of unsupervised to semi-supervised posteriors modulates the plug-in likelihood. The numerator is conditioned only on the unsupervised scores, while the denominator is conditioned on one additional supervised score, with assumed label $\ell' = H_1$. Although (8) holds for any value of $\mathcal{C}$ with non-zero posteriors, we find a more convenient form by taking logarithms and the expectation w.r.t. $P(\mathcal{C}|S)$ on both sides:

$$\log P(s'|S, H_1) = \langle \log P(s'|\mathcal{C}, H_1) \rangle_S + D_1(s') \tag{9}$$

where $D_1$ denotes KL-divergence from unsupervised to semi-supervised posterior:

$$D_1(s') = \langle \log \frac{P(\mathcal{C}|S)}{P(\mathcal{C}|S, s', H_1)} \rangle_S \tag{10}$$

Using (9) in (6) gives the predictive log-LR as:

$$\log R(s'|S) = \langle \log R(s'|\mathcal{C}) \rangle_S + D_1(s') - D_2(s') \tag{11}$$

Notice that:

$$\langle \log R(s'|\mathcal{C}) \rangle_S = s'\left(\frac{d^2}{\sigma^2}\right) + \left(\frac{\mu_2^2 - \mu_1^2}{2\sigma^2}\right)_S \tag{12}$$

which remains affine in $s'$, just like (2). Moreover, if $P(\mathcal{C}|S)$ has a sharp, dominant peak, then $\langle \log R(s'|\mathcal{C}) \rangle_S \approx \log R(s'|\mathcal{C})$, where $\mathcal{C}$ is the mode of the dominant peak. Finally, if there are many scores in $S$, then a single additional score $s'$ that is similar to the scores in $S$, will result in small $D_1(s')$, so that $\log R(s'|S) \approx \log R(s'|\mathcal{C})$. Only if $S$ has very few scores, or $s'$ is very far away, will the $D_1(s')$ cause significant non-linearity in $\log R(s'|S)$.

We already know that we have a large collection of unsupervised scores, but it remains to be demonstrated that $P(\mathcal{C}|S)$ has a dominant peak, which we shall do below, via an experimental exploration of the likelihood. We shall also quantify the sharpness of that peak by using the Laplace approximation.

#### 3.2. Laplace approximation

The Laplace approximation (LA) is ideally suited to approximating sharply peaked, low-dimensional posteriors [7, 8]. We have only 4 parameters and the likelihood is sharply peaked because we have lots of data. The only pitfall is that label swapping causes two identical peaks in the likelihood. We kill the unwanted peak by assigning a lag of data. The only pitfall is that label swapping causes two identical peaks in the likelihood. We shall also quantify the sharpness of that peak by using the Laplace approximation.

$$f(\mathcal{M}) = \log P(S, \mathcal{M}) = \log [P(S|\mathcal{M})P(\mathcal{M})] \tag{13}$$

which is computable by (4). Notice $P(\mathcal{M}|S) \propto e^{f(\mathcal{M})}$. Let $\mathcal{M}$ be the dominant mode of $f$ and form a 2nd-order Taylor-series approximation here. The gradient at the mode is zero, but we need the Hessian (2nd derivative matrix), denoted $\Lambda$. This forms a multivariate Gaussian, approximate posterior:

$$\hat{P}(\mathcal{M}|S) = N(\mathcal{M}|\hat{\mathcal{M}}, -\Lambda^{-1}) \tag{14}$$

2Observation at a node with convergent arrows induces dependency between variables linked via this node; when not observed, such nodes block dependency. Conversely, nodes with divergent or aligned arrows induce dependency when not observed; and block dependency when observed. A node is ‘observed’ if it appears to the right of the | in probability notation.

3To see this, use (5) to find: $P(s'_{H_1}|S, \pi'_1) = \frac{\pi'_1}{\pi_1} \times R(s'|S)$.

4By dominant, we mean that the peak contains almost all of the probability mass in $P(\mathcal{C}|S)$. 

---

1700
Since (7) calls for $P(C|S)$, we still have to marginalize 8 (14) w.r.t. $\pi_1$. With the Gaussian approximation this is easy. Let $C = -A^{-1}$ denote the covariance of $P(M|S) = N(M|\hat{M}, C)$, then the corresponding marginal $P(C|S)$ is also multivariate Gaussian, where the elements in $\hat{M}$ and $C$ corresponding to $\pi_1$ have been removed [7]. In summary, we get the 4-dimensional mode and Hessian, invert the Hessian and then discard one dimension.

3.2.1. Model parametrization

The LA is not invariant to parametrization [8]. Moreover, it is obvious that the true posterior for the parameters $\pi_1$ and $\sigma^2$ cannot be Gaussian. But, for a sharply peaked posterior, the parametrization is not that important and the behaviour far from the maximum is almost irrelevant. As long as $f(M)$ is smooth enough so that a 2nd-order approximation is accurate close to the maximum, the posterior peak will be approximately Gaussian. If the likelihood magnitudes are large, then by the time the 2nd-order approximation becomes inaccurate, this inaccuracy becomes irrelevant because of the effect of the exponentiation. The reader is encouraged to consult Wikipedia, 6 where this is graphically illustrated. Our experimental results below are reported for the parametrization $[\mu_1, \mu_2, \log(\sigma^2), \log(\pi_1)]$.

4. EXPERIMENTS

We used two score corpora, DAC and ABC:

The Domain Adaptation Challenge (DAC) 7 has telephone speech from the LDC’s Switchboard and Mixer databases as well as NIST SRE’10 [9]. It has three parts:

Recognizer: Switchboard was used to train the hyperparameters of an i-vector PLDA speaker recognizer [10], which was then used to produce the scores below.

Calibration: About 7 million trials (single enrollment), from pre-SRE’10 Mixer corpora, provided the unsupervised scores, $S$, with target proportion about 4% and (empirical) EER = 2.38%.

Evaluation: About 400 000 trials, composed of pairs of segments from SRE’10, provided the test scores, $s'$, with EER = 5.54%.

The ABC corpus used a different speaker recognizer, applied to a different data set, drawn from the AGNITIO-BUT-CRIM submission to NIST SRE’12 [11, 12]. The mixture of conditions was more diverse than for DAC, having telephone and microphone speech, variable number of enrollment segments, full and truncated test segments and varied noise levels.

Recognizer: Switchboard, Fisher and pre-SRE’12 Mixer was used to train an i-vector PLDA system.

Calibration: About 42 million scores, $S$, from pre-SRE’12 Mixer. The target proportion is about 0.07%, and EER = 2.38%.

Evaluation: About 9 million test scores, $s'$, from SRE’12, with EER = 3.25%.

4.1. Exploration of likelihood

The success of the whole venture depends critically on the behaviour of the GMM likelihood, $P(S|M)$, given by (4). If we are a-priori very uncertain about the proportion of targets in the unsupervised data, and also about the accuracy of the recognizer, it is not at all obvious whether there is enough information in the likelihood 10 to be able to infer calibration parameters with a useful level of accuracy. Moreover, since our inference tools (plug-in and LA) both rely on finding likelihood optima, it is important to know whether the likelihood is plagued by local optima. 9

To learn how the likelihood behaves, we did an exhaustive experimental exploration of the parameter space, $(\mu_1, \mu_2, \sigma^2, \pi_1)$. To facilitate visual representation, we used a 2-dimensional model representation, namely $(d', \log(\frac{\pi_1}{\pi_1^*}))$, which we plotted against log-likelihood, where the remaining two degrees of freedom were optimized. 10 Recall from section 2.1 that $d'$ represents accuracy, while $\pi_1$ represents target proportion. If the likelihood has a single dominant peak for these two critical parameters, then there is hope that the calibration exercise will work.

We made such plots for the calibration parts of DAC and ABC. The results are similar. In figure 1 we show the latter, which we believe is more challenging, because of the smaller target proportion. The log-likelihood is smooth as a function of $d'$, but is bi-modal as a function of $\pi_1$—a warning that initialization for the EM algorithm is important. Although the modes look flat, and similar, in the log-likelihood plot, the normalized 11 likelihood plot of figure 2 reveals that there is just a single, sharp, dominant peak, the location of which is given in table 2.

4.2. Analysis of sharpness

To approximate $P(M|S)$, we use the LA recipe of section 3.2. The mode is found with the EM-algorithm. Complex-step differentiation [14] and the Pearlmutter trick [15] are used for the Hessian. We find the error-bars (posterior standard deviations [13, 8]) for the parameters to be suprisingly small:

<table>
<thead>
<tr>
<th>parameterization</th>
<th>$\mu_1$</th>
<th>$\mu_2$</th>
<th>$\log(\sigma^2)$</th>
<th>$\log(\pi_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC error-bars</td>
<td>0.0226</td>
<td>0.0004</td>
<td>0.0002</td>
<td>0.0071</td>
</tr>
<tr>
<td>DAC error-bars</td>
<td>0.1183</td>
<td>0.0192</td>
<td>0.0006</td>
<td>0.0022</td>
</tr>
</tbody>
</table>

4.3. Calibration experiments

The sharpness of the parameter posterior shows there is no practical difference between the plug-in and predictive likelihood-ratios. We therefore proceed to report our final experimental results for maximum-likelihood plug-in calibration. We estimated the parameters on the calibration parts of DAC (7 million scores, 4% targets) and ABC (42 million scores, 0.07% targets). The performance of these calibrations was tested on the independent evaluation parts of those corpora, in terms of normalized Bayes error-rate, also known as normalized DCF [16]:

$$\text{normDCF}(\pi_1) = \frac{\pi_1 P_{\text{norm}}(\pi_1^*) + (1 - \pi_1) P_{\text{err}}(\pi_1^*)}{\min(\pi_1^*, 1 - \pi_1^*)} \quad (15)$$

8The likelihood function, $P(S|M)$, represents all of the information that our chosen model can extract from $S$.

9This is true even for more sophisticated Bayesian tools, like variational Bayes and Gibbs sampling [7].

10Integrating them out using LA would also be feasible, but we found this unnecessary—when $d'$ and $\pi_1$ are fixed, there remains very little uncertainty about the scale and location. The constrained optimization was done with a bespoke EM algorithm, where the M-step had to make use of numerical optimization.

11Subtract the maximum over the graph and then exponentiate [13].
where $P_{\text{miss}}(\pi')$ and $P_{\text{fa}}(\pi')$ are the empirical miss and false-alarm error-rates obtained when using the log-likelihood-ratios to make Bayes decisions at the theoretical threshold, $-\logit(\pi')$. The denominator is the Bayes error-rate for the default decision that always accepts, or always rejects, depending only on $\pi'$. Smaller values of normDCF are better, while a value of smaller than one shows the recognizer is doing better than the default decision.

Table 1 reports normDCF for 4 different values of $\pi'$. For both databases, we compare the supervised recipe of [4] against the proposed unsupervised recipe. The supervised method used the same data as the unsupervised recipe, except that the labels were supplied. We also report minDCF, which uses an empirically optimized threshold at each operating point, where the optimization makes use of the evaluation labels. Finally, for DAC, we also report results for a supervised calibration trained on the mismatched\textsuperscript{12}, Switchboard data. The high error-rates for this case emphasizes the need for calibration on matched data.

Surprisingly, for the DAC database, the unsupervised method does mostly better than the supervised one. This may be because of errors\textsuperscript{13} in the labels supplied to the supervised method.

In an effort to test whether our method holds up for very low target proportions, we noticed that we could go as far as removing all trials labelled as targets, so that $S$ contained only trials labelled as non-targets. These entries in the table are marked as unsupervised*. The fact that calibration still works in these cases can perhaps also be attributed to labelling errors.

For additional insight, tables 2 and 3 compare the estimates of model and calibration parameters for the supervised and unsupervised cases.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
$\pi'$ & 0.001 & 0.01 & 0.1 & 0.5 \\
\hline
ABC supervised & 0.32 & 0.22 & 0.13 & 0.08 \\
ABC unsupervised & 0.33 & 0.24 & 0.16 & 0.11 \\
ABC unsupervised* & 0.32 & 0.23 & 0.15 & 0.10 \\
ABC minDCF & 0.31 & 0.21 & 0.12 & 0.06 \\
DAC mismatched & 0.73 & 0.34 & 0.35 & 0.21 \\
DAC supervised & 0.63 & 0.44 & 0.28 & 0.13 \\
DAC unsupervised & 0.65 & 0.43 & 0.25 & 0.11 \\
DAC unsupervised* & 0.65 & 0.43 & 0.24 & 0.12 \\
DAC minDCF & 0.63 & 0.42 & 0.24 & 0.11 \\
\hline
\end{tabular}
\caption{Calibration performance in terms of normDCF.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
& $\mu_1$ & $\mu_2$ & $\sigma$ & $d'$ & $\pi_1$ \\
\hline
ABC super & 8.2 & -5.9 & 2.9 & 4.9 & 6.6e-4 \\
ABC unsup & 9.9 & -5.9 & 2.9 & 5.5 & 5.6e-4 \\
ABC unsup* & 9.6 & -5.9 & 2.9 & 5.4 & 5.1e-5 \\
DAC super & 34.0 & -169.3 & 48.4 & 4.2 & 3.9e-2 \\
DAC unsup & 45.9 & -168.7 & 48.8 & 4.4 & 3.4E-2 \\
DAC unsup* & 72.3 & -169.3 & 48.0 & 5.0 & 1.4E-5 \\
\hline
\end{tabular}
\caption{GMM parameter estimates}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
& scale & offset \\
\hline
ABC super & 1.7 & -2.0 \\
ABC unsup & 1.9 & -3.8 \\
ABC unsup* & 1.9 & -3.5 \\
DAC super & 0.087 & 5.9 \\
DAC unsup & 0.090 & 5.5 \\
DAC unsup* & 0.105 & 5.1 \\
\hline
\end{tabular}
\caption{Calibration parameters}
\end{table}

5. CONCLUSION

The outcome of this work held two surprises for us. The first is that unsupervised calibration works at all. The second is that the missing labels contribute surprisingly little uncertainty to the parameter estimates.

For future work on different data, we caution against blind application of the plug-in recipe. We feel that some Bayesian analysis similar to ours should also be done to illuminate the interaction between model and data.

\textsuperscript{12}Switchboard is at least a decade older than Mixer, during which time telephony changed dramatically [17].

\textsuperscript{13}Some Mixer subjects registered multiple times, with different PINs, thereby causing some target trials to be falsely labelled as non-targets.
6. REFERENCES


