The following full text is an author’s version which may differ from the publisher’s version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/116267

Please be advised that this information was generated on 2017-08-26 and may be subject to change.
The distribution of calibrated likelihood-ratios in speaker recognition

David A. van Leeuwen¹ and Niko Brümmer²

¹Netherlands Forensic Institute, The Hague and Radboud University Nijmegen, The Netherlands
²AGNITIO Research, Somerset West, South Africa

Abstract

This paper studies properties of the score distributions of calibrated log-likelihood-ratios that are used in automatic speaker recognition. We derive the essential condition for calibration that the log likelihood ratio of the log-likelihood-ratio is the log-likelihood-ratio. We then investigate what the consequence of this condition is to the probability density functions (PDFs) of the log-likelihood-ratio score. We show that if the PDF of the non-target distribution is Gaussian, then the PDF of the target distribution must be Gaussian as well. The means and variances of these two PDFs are interrelated, and determined completely by the discriminability the recognizer characterized by the equal error rate. These relations allow for a new way of computing the offset and scaling parameters for linear calibration, and we derive closed-form expressions for these and show that for modern i-vector systems with PLDA scoring this leads to good calibration, comparable to traditional logistic regression, over a wide range of system performance.

1. Introduction

In recent years, calibration in automatic speaker recognition has received more attention [1][2]. Intuitively, calibration is related to the ability to properly set a threshold in a speaker detection system so as to minimize the expected error [12]. In speaker detection, the task is to decide whether or not two speech signals originate from the same speaker. Because all speaker recognition systems internally work with some scalar score that expresses speaker similarity, a score threshold can control the trade-off between the two types of errors that a system can make [3][4]. Indeed, in the series of NIST Speaker Recognition Evaluations (SRE) the primary evaluation measure has been sensitive to calibration. Until SRE 2010, calibration was assessed in a single operating point, through a single de-

2. Likelihood-ratio idempotence

Here we carefully define the likelihood-ratio (LR) and show that it has the interesting property: the LR of the LR is the LR, which forms a definition of calibration.

The speaker recognition system has as input two speech segments, denoted $X$ and $Y$, which it processes in two steps. We represent the first step as $s = f(X, Y)$. To keep things general, $s$ may represent different kinds of output, e.g., a pair of acoustic feature vector sequences, a pair of i-vectors, or just a single, scalar recognition score. The second step is to compute the likelihood-ratio $r$ as a function of $s$, as:

$$r = \frac{P(s \mid H_1, M)}{P(s \mid H_2, M)}$$

(1)

where $H_1$ is the (target) hypothesis that $X$ and $Y$ originate from the same speaker, $H_2$ the (non-target) hypothesis that they are from two different speakers, and $M$ is a generative probabilistic model for $s$. In current practice, $s$ is always the recognition score, so that $M$ merely models scalar scores—not i-vectors, acoustic feature sequences or speech signals. But our theory...
below is sufficiently general to remain applicable in future to more ambitious models, when \( s \) might have a more complex form. We now assume there is given the hypothesis prior, \( \pi = \text{P}(H_1) \), which allows us to express the hypothesis posterior, via Bayes’ rule as:

\[
\text{P}(H_1 \mid s, M, \pi) = \frac{\pi r}{\pi r + (1 - \pi)}
\]  (2)

This shows that \( r \) is a sufficient statistic: the posterior depends on \( s \) only through \( r \). This allows rewriting the posterior as:

\[
\text{P}(h \mid s, M, \pi) = \text{P}(h \mid r, M', \pi), \quad h \in \{H_1, H_2\}
\]  (3)

where we have introduced \( M' \) to denote \( M \), augmented by asserting \( \{1\} \). Although \( r \) contains all the relevant information that \( M \) can extract from \( s \) to recognize the unknown hypothesis, it must be stressed that \( r \) and \( s \) do not necessarily contain all the relevant information that could have been extracted from the original input \( X, Y \) by some more elaborate model. Now we use the odds form of Bayes’ rule:

\[
\frac{\text{P}(H_1 \mid \rho, M, \pi)}{\text{P}(H_2 \mid \rho, M, \pi)} = \frac{\pi \text{P}(\rho \mid H_1, M)}{1 - \pi \text{P}(\rho \mid H_2, M)}
\]  (4)

where \( \rho \) is a placeholder for \( r \) or \( s \) for \( M \) or \( M' \). Combining this with \( 5 \), we find the desired relationship (the LR of the LR) is the LR \([24]\):

\[
r = \frac{\text{P}(s \mid H_1, M)}{\text{P}(s \mid H_2, M)} = \frac{\text{P}(r \mid H_1, M')}{\text{P}(r \mid H_2, M')}
\]  (5)

If we define \( x \) to be the log-likelihood-ratio (LLR):

\[
x = \log r
\]  (6)

we also find \( \text{the LLR} \) of the LLR is the LLR:

\[
x = \log \frac{\text{P}(x \mid H_1, M'')}{\text{P}(x \mid H_2, M'')}
\]  (7)

where \( M'' \) augments \( M' \) by addition of \( \{6\} \).

### 2.1. Implications

Rewriting \( 5 \) as:

\[
\text{P}(r \mid H_1, M') = r \text{P}(r \mid H_2, M')
\]  (8)

we see that if either of the two distributions is given, then the other distribution is completely determined—they cannot vary independently. Moreover, a further restriction is placed on these distributions: since the LHS must integrate to 1, the expected value of the non-target distribution (the integral of the RHS) must be: \( \langle r \rangle = 1 \). Similarly, for targets: \( \langle \frac{1}{r} \rangle = 1 \). By applying Jensen’s inequality \([25]\), we also find for targets: \( \langle x \rangle \geq 0 \) and for non-targets: \( \langle x \rangle \leq 0 \).

### 2.2. Good and bad calibration

How does \( 5 \) function as a definition of calibration? Since it is an equality, won’t all LRs calculated via \( 1 \) by some model \( M \) just automatically satisfy \( 5 \)? Yes they will, but only if \( M \) and \( M' \) are related as explained above. If we want to independently judge the goodness of the calibration of \( r \), we do not condition the distributions for \( r \) on the recognizer’s model \( M \). Instead, we could empirically observe the target and non-target values of \( r \) as calculated by the recognizer over an independent, supervised database of speaker detection trials. Letting \( O \) denote the empirical observation, we could then say the model \( M \) is well calibrated if:

\[
r = \frac{\text{P}(s \mid H_1, M)}{\text{P}(s \mid H_2, M)} \approx \frac{\text{P}(r \mid H_1, O)}{\text{P}(r \mid H_2, O)}
\]  (9)

Bad calibration is when the LRs given respectively by the recognizer’s \( M \) and empirical observation \( O \), do not agree in this way. This can and does happen, since \( O \) is independent of any development data that was used to determine the form and parameters of \( M \).

It should be noted that \( 9 \) does not give a practical recipe to judge degree of goodness of calibration—it specifies neither how to assign \( \text{P}(r \mid h, O) \), nor how to numerically evaluate the agreement between LHS and RHS. For practical solutions for calibration-sensitive objective functions, see for example \([26]\).

### 3. Gaussian distributed log-likelihood-ratios

Inspired by the fact that DET curves in speaker recognition tend to be straight \([21]\), we explore a Gaussian solution to the LLR distribution constraint \([7]\). Since target and non-target LLR distributions are so tightly coupled, it turns out that if the one is assumed to be Gaussian, then the other must also be. We shall use the shorthand: \( e(x) = P(x \mid H_1, M) \) and \( d(x) = P(x \mid H_2, M''') \). Arbitrarily assuming a Gaussian distribution for non-targets (different-speaker trials):

\[
d(x) = \mathcal{N}(x \mid \mu_d, \sigma_d) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{-\frac{(x - \mu_d)^2}{2\sigma_d^2}}.
\]  (10)

We derive the functional form for target \([2]\) \( e(x) \), when \([7]\) applies:

\[
e(x) = e^x d(x) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{-(x - \mu_d)^2/2\sigma_d^2}.
\]  (11)

We collect the terms in \( x \) in the exponent, which itself can be written like

\[
-\frac{x^2 - 2\mu_d x + \mu_d^2}{2\sigma_d^2} + \frac{2\sigma_d^2 x}{2\sigma_d^2}
\]  (12)

\[
= -\frac{x^2 - 2(\mu_d + \sigma_d^2) x + \mu_d^2}{2\sigma_d^2}
\]  (13)

\[
= -\frac{(x - (\mu_d + \sigma_d^2))^2}{2\sigma_d^2} + \frac{2\mu_d \sigma_d^2 + \sigma_d^4}{2\sigma_d^2}
\]  (14)

The first term is in the familiar form of a Gaussian exponent, the second will result in a constant factor. Gathering terms, and writing

\[
\mu_e = \mu_d + \sigma_d^2,
\]  (15)

the expression for the same-speaker comparison log-likelihood-ratio scores becomes

\[
e(x) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{\sigma_d^2/2 + \mu_d} e^{-(x - \mu_e^2)/2\sigma_d^2}
\]  (16)

\[
e^{\sigma_d^2/2 + \mu_d} \mathcal{N}(x \mid \mu_e, \sigma_d).
\]  (17)

We see that \( e(x) \) is of Gaussian shape, with

\[
\sigma_e = \sigma_d \equiv \sigma.
\]  (18)
Since \( e(x) \) must be a proper PDF, its integral over \( x \) must be unity, from which follows that
\[
e^{\mu^2/2 + \mu x} \int_{-\infty}^{\infty} N(x \mid \mu, \sigma) \, dx = 1 \quad (19)
\]
\[
-2\mu \sigma = \sigma^2. \quad (20)
\]
Finally, with (15) we find
\[
\mu_e = \mu_d + \sigma^2 = -\mu_d \equiv \mu, \quad (21)
\]
This shows that \( d(x) \) and \( e(x) \) are equal variance Gaussians with means symmetric around zero at \( \pm \mu \), and where the variance and mean are related (20)
\[
\sigma^2 = 2\mu. \quad (22)
\]

3.1. Equal Error Rate and \( d' \)

Using the symmetry of the solution, it is clear that the threshold for the equal error rate is at \( x = 0 \). Using the expression for the miss probability, the equal error rate \( E_m \) is
\[
E_m = \int_0^{\infty} N(x \mid \mu, \sigma) \, dx = \int_{-\infty}^{-\mu/\sigma} N(x \mid 0, 1) \equiv \Phi(-\mu/\sigma), \quad (23)
\]
where \( \Phi(x) \) is the cumulative normal distribution.

It is sometimes useful to recognize the parameter \( d' \) from detection theory, which is the difference in means expressed in terms of the standard deviation, here \( d' = 2\mu/\sigma \). With (24) the relation becomes
\[
E_m = \Phi(-\frac{1}{2}d'). \quad (25)
\]
\[
d' = \phi = -2\Phi^{-1}(E_m), \quad (26)
\]
introducing \( \Phi^{-1}(y) \), the inverse of the cumulative normal distribution. The importance of the relations above is that \( \mu \) and \( \sigma \) are determined by the discrimination performance measured by \( E_m \), using (22) and (20)
\[
\mu = \frac{\sigma^2}{2} = 2[\Phi^{-1}(E_m)]^2. \quad (27)
\]

4. A new calibration method

In practice, automatic speaker recognition systems do not deliver scores that can directly be interpreted as a log-likelihood-ratio, even though they are computed as such, for instance in the good old UBM-GMM scoring \( \Sigma \) or the latest i-vector PLDA scoring \( \Sigma \). A practical solution to this is to convert raw scores \( s(X,Y) \) to calibrated log-likelihood-ratios by some transformation function \( x(s) \), usually constrained to be monotonic increasing. There are many ways of doing this. The FoCal \( \Sigma \) and BOSARIS \( \Sigma \) toolkits use logistic regression to discriminatively train linear calibration transformations. Other possibilities include isotonic regression (PAV \( \Sigma \)) and line-up calibration \( \Sigma \) that uses the rank in a line-up of foil speakers. In FoCal or BOSARIS, the score-to-LLR function is affine:
\[
x(s) = as + b \quad (28)
\]
and the parameters \( a \) and \( b \) are found by optimizing cross-entropy, a calibration-sensitive objective function defined on a supervised set of speaker recognition trials.

Here we contrast the popular discriminative logistic regression solution to a new generative, constrained maximum-likelihood (ML) solution. Our constraints follow from assuming (i) Gaussian LLR distributions, and (ii) an affine score-to-LLR transform \( \Sigma \). This implies that (i) the LLR distributions are constrained as derived in Section 3 and (ii) the score distributions are also Gaussians, with equal variances. With no LLR distribution constraints, we would have had 6 free parameters: 2 means, 2 variances and 2 calibration parameters. But we have imposed 3 constraints, equal variances \( \mu \), symmetric means \( \pm \mu \) and common variance \( \sigma \). Setting derivatives to 0, we find the maximum likelihood at the sample means:
\[
m_e = \frac{1}{N_e} \sum_{i \in E} s_i, \quad m_d = \frac{1}{N_d} \sum_{i \in D} s_i \quad (29)
\]
and at a weighted combination of sample variances:
\[
v = \frac{\alpha}{N_e} \sum_{i \in E} (s_i - m_e)^2 + \frac{1 - \alpha}{N_d} \sum_{i \in D} (s_i - m_d)^2 \quad (30)
\]
By (28), the LLR distribution parameters become \( \sigma^2 = \alpha^2 \sigma^2 \), \( \mu_e = \alpha m_e + b \) and \( \mu_d = \alpha m_d + b \). Finally, applying the constraints \( \sigma^2 = \mu_e - \mu_d \) and \( \mu_e = -\mu_d \), we can solve for the calibration parameters:
\[
a = \frac{m_e - m_d}{v}, \quad b = -\frac{m_e + m_d}{2} \quad (31)
\]
We call this recipe constrained, maximum-likelihood, Gaussian (CMLG) calibration. An advantage of CMLG is that it has a closed form, in contrast to the iterative optimization required by logistic regression.

4.1. Experiment

In order to test CMLG we apply it to a number of recognition trials sets. We use a set of trials crafted for duration-dependence experiments \( \Sigma \) from the NIST SRE 2008 and 2010 trial sets, the telephone-telephone “extended” trial lists. We constructed short duration segments of 5, 10, 20, and 40 seconds from both conversation side, were tested in all combinations, leading to 25 different trial list combinations using \( \Sigma \). We contrast CMLG (with \( \alpha = \frac{1}{2} \)) to the traditional logistic regression method. The calibrations are trained on NIST SRE 2008 data \( \Sigma \) and applied to SRE 2010 trials for evaluation \( \Sigma \). We used the 25 different trial list combinations using \( \Sigma \), a cost function that is sensitive to calibration over the whole DET curve \( \Sigma \). We used R’s glm routine for logistic regression.

The results are shown in Fig. 1 where we have plotted the \( \Sigma \) obtained using CMLG calibration versus \( \Sigma \) obtained using logistic regression. The values are highly correlated. For
CMLG, the average $C_{\text{llr}}$ over all 25 conditions is 0.375, for logistic regression it is 0.376. These can be called good, as the mean $C_{\text{llr}}$ is 0.370.

We have also used the NIST SRE12 scores from the ABC-team to study the effect of condition (30) to another calibration sensitive measure $C_{\text{primary}}$, cf. Fig. 2 for details we refer to [26]. The figure shows that with CMLG good calibration results can be obtained for a different system with different data and a different performance measure, if the correct $\alpha$ is chosen.

5. Discussion and Conclusions

We have shown in this paper, that if the different-speaker calibrated log-likelihood-ratio scores from a speaker recognition system follow a Gaussian distribution, then the distribution of the same-speaker scores must also be Gaussian after calibration, with the same variance but opposite mean. Because monotonically increasing score-to-likelihood-ratio functions do not change the DET plot, such equal-variance distributions in the calibrated score domain imply 45° DET-plots in the raw score domain as well—which is neither observed with real data nor desired for applications operating in the low false alarm region.

The logical conclusion then is that real scores, if they are well-calibrated, will not be Gaussian. However, we see that our PLDA system can be calibrated quite well under the Gaussian assumptions, and indeed we have noticed that i-vector PLDA systems tend to have score distributions that appear more Gaussian than earlier technologies, such as i-vector LDA cosine distance scoring, support vector machines or the UBM-GMM likelihood ratio scoring.

The Gaussian solution to the LLR equation (7) is one where both distributions are shaped by the same mathematical function. In signal detection theory, where the distribution represents noise, this seems almost mandatory, but in speaker recognition this is not an obvious assumption. We have experimented with other distributions, e.g., in the likelihood-ratio domain [5].

3We have measured the slope of the DET in the conventional error region 0.1–50 % for the data in the experiment. The mean slope over the 23 conditions is −0.99 with a standard deviation of 0.06, so in fact this data appears to honour the equal variance condition quite well.

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 238803.


7. References


