The distribution of calibrated likelihood-ratios in speaker recognition

David van Leeuwen and Niko Brümmer
d.vanleeuwen@let.ru.nl, nbrummer@agnito.es
Netherlands Forensic Institute / Radboud University Nijmegen, Agnitio Research

15 October 2013\(^1\)

\(^1\)First published at Interspeech 2013
We had these badly-behaving scores depending on utterance duration
We tried to design universal calibration transformations
Question arose: where do calibrated scores hang out?
What is their distribution?

Inspiration for this work

---

Mandasari et al., Interspeech 2011
What is calibration?

Traditionally:

- The capability to set a threshold correctly

Nowadays:

- The ability to give a proper probabilistic statement about identity
  - ...to produce (log) likelihood ratio scores for every comparison
  - ...that lead to optimal Bayes’ decisions
What is calibration?

Traditionally:
- The capability to set a threshold correctly

Nowadays:
- The ability to give a proper probabilistic statement about identity
  - ...to produce (log) likelihood ratio scores for every comparison
  - ...that lead to optimal Bayes’ decisions

Bayes’ decision

\[
\text{Priors} + \text{likelihoods} \rightarrow \text{posteriors} \\
\text{Posteriors} + \text{costs} \rightarrow \text{expected costs} \\
\text{Minimize expected costs} \rightarrow \text{decision}
\]
The forensic motivation of the Likelihood Ratio

Use the log Likelihood Ratio as **weight of evidence** in court

- Using Bayes’s rule, separate contributions
  - Forensic Expert, w.r.t. the material they know about
  - The other evidence / circumstances of the case
to compute the **posterior probability** that suspect is the perpetrator
Use the log Likelihood Ratio as weight of evidence in court

- Using Bayes’s rule, separate contributions
  - Forensic Expert, w.r.t. the material they know about \( E \)
  - The other evidence / circumstances of the case \( I \)

To compute the posterior probability that suspect is the perpetrator \( H_p = \neg H_d \)

- Mathematically,

\[
\frac{P(H_p | E, I)}{P(H_d | E, I)} = \frac{P(E | H_p, I)}{P(E | H_d, I)} \times \frac{P(H_p | I)}{P(H_d | I)}
\]

Judge/jury wants to know given by expert other evidence
A likelihood ratio can be treated like a score
  - All analysis tricks work: ROC, DET, EER, decision cost functions... 
But can we transform a score into a LR?
This is a process known as \textit{calibration}: giving meaning to probabilistic statements
From scores to likelihood ratios

- A likelihood ratio can be treated like a score
  - All analysis tricks work: ROC, DET, EER, decision cost functions...
- But can we transform a score into a LR?
- This is a process known as \textit{calibration}: giving meaning to probabilistic statements

\textbf{problem statement}

But what is the definition of \textit{calibrated} scores / LRs?
Definition of *Calibrated Likelihood Ratios*

Our definition\(^3\)

The LR of the LR is the LR

or, for the mathematically inclined

\[
LR = \frac{P(LR \mid H_p)}{P(LR \mid H_d)}
\]

---

\(^3\)Proof in paper, short version in Mandasari *et al.*, IEEE-TASLP (2013, accepted)
Definition of *Calibrated Likelihood Ratios*

Our definition\(^3\)

The LR of the LR is the LR

or, for the mathematically inclined

\[
LR = \frac{P(LR \mid H_p)}{P(LR \mid H_d)}
\]

which happens to be equivalent to

\[
\log LR = \log \frac{P(\log LR \mid H_p)}{P(\log LR \mid H_d)}
\]

The LLR of the LLR is the LLR

\(^3\)Proof in paper, short version in Mandasari *et al.*, IEEE-TASLP (2013, accepted)
More inspiration: Why are DET curves straight?

- If score distributions are **Gaussian**, then DET curve is straight
  - Slope is ratio of standard-deviations of the score distributions
- If DET is straight, score distributions are not necessarily Gaussian
  - but can be made Gaussian by warping of score axis
For reference: these are the score distributions

- Clearly not Gaussian
- But *still* leading to a straight DET curve
- non-targets: $d(x)$ (different)
- targets: $e(x)$ (equal)
Can Gaussian Scores be Well Calibrated?

Let’s try

- **Gaussian non-targets** \( d(x) = \mathcal{N}(x | \mu_d, \sigma^2_d) \)
- **calibration definition for LLR:**

\[
x = \log \frac{e(x)}{d(x)}
\]

**targets** \( e(x) = e^x d(x) \)

Now use the expression for the normal distribution \( \mathcal{N} \), and see what the targets \( e(x) \) look like

\[
e(x) = e^x d(x) = \frac{1}{\sqrt{2\pi\sigma_d}} e^{x-(x-\mu_d)^2/2\sigma^2_d}
\]
Expanding the exponent for target distribution $e(x)$:

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

Gaussian form

Normalisation constant
Expanding the exponent for target distribution $e(x)$:

$$
- \frac{x^2 - 2\mu_d x + \mu_d^2}{2\sigma_d^2} + \frac{2\sigma_d^2 x}{2\sigma_d^2}
$$

$$
= - \frac{x^2 - 2(\mu_d + \sigma_d^2)x + \mu_d^2}{2\sigma_d^2}
$$

$$
= - \frac{(x - (\mu_d + \sigma_d^2))^2}{2\sigma_d^2} + \frac{2\mu_d \sigma_d^2 + \sigma_d^4}{2\sigma_d^2}
$$

Gaussian form

Normalisation constant

Gaussian form

- if $\mu_e = \mu_d + \sigma_d^2$
- with $\sigma_e = \sigma_d$
- normalization requires $-2\mu_d = \sigma^2$
Conclusions of this little exercise

- Consider non-target distribution $d(x)$ and target score distribution $e(x)$
- Then if $d(x)$ is normally distributed
Conclusions of this little exercise

- Consider non-target distribution \( d(x) \) and target score distribution \( e(x) \)
- Then if \( d(x) \) is normally distributed

...the calibration definition tells us

- \( e(x) \) is normally distributed as well
- Variances are the same for \( d(x) \) and \( e(x) \)
- The means are symmetric around 0,
  \[
  \mu_d = -\mu_e
  \]
- Variance and mean are related
  \[
  \sigma^2 = 2\mu
  \]
Example of well-calibrated scores

- $\text{LR} = 2$
  - density scores around $2$ is $2 \times$ as high for targets (red) as for the non-targets (blue)
Example of well-calibrated scores

- **LR = 2**
  - Density scores around 2 is $2 \times$ as high for targets (red) as for the non-targets (blue)

- **LR = 4**
Example of well-calibrated scores

- \( LR = 2 \)
  - density scores around 2 is \( 2 \times \) as high for targets (red) as for the non-targets (blue)

- \( LR = 10 \)
Some direct consequences

- Well calibrated straight DET curves must be of 45° slope
- Preferred “flat” straight DET curves can’t arise from calibrated scores
  - highly-discriminative systems have flat DET curves,
  - fingerprint, iris, ...

![Graph of ROC at EER = 1%](image)

@ FAR = 10^{-6}

@ FAR = 10^{-3}

Score Distributions at EER = 10% and sigma ratio = 2
All relations are known, now

From this model of scores all other characteristics follow, e.g.,

- Equal Error Rate $E_=$
  - Threshold at 0
  - Integrate the miss error:

$$E_= = \int_{-\infty}^{0} N(x \mid \sigma, \mu) \, dx$$

$$= \Phi\left(-\frac{\mu}{\sigma}\right) = \Phi\left(-\sqrt{\frac{\mu}{2}}\right)$$

- Φ(z) cumulative normal distribution

- Cost of LLR $C_{llr}$

$$C_{llr} = \frac{1}{\log 2} \int_{-\infty}^{\infty} N(x \mid \mu, \sigma) \log(1 + e^{-x}) \, dx$$

- $C_{llr}$ depends only on $E_=$
$C_{llr}$ depends only on $E_{ll}$

Approximate relation:

$$C_{llr} \approx 1 - (2E_{ll} - 1)^2$$
Application: a new way of doing calibration

**Calibration** is the process of fixing scores so that they can be interpreted better as log likelihood ratios

- Traditionally, this is done in speaker recognition by an affine transformation of score $s$

  $$x = as + b$$

- parameters $a$ and $b$ found by logistic regression using a development set of trials
Application: a new way of doing calibration

Calibration is the process of fixing scores so that they can be interpreted better as log likelihood ratios

- Traditionally, this is done in speaker recognition by an affine transformation of score $s$

  $$x = as + b$$

- parameters $a$ and $b$ found by logistic regression using a development set of trials

New calibration method:

Find $a$ and $b$ by constraining the transformed scores to satisfy the Gaussian LLR conditions for $\mu$ and $\sigma$
Raw score means and variances $m_{d,e}$, $s_{d,e}^2$.

- Transformed target mean: $am_e + b = \mu$
- Transformed non-target mean $am_d + b = -\mu$
- Weighted variance $\nu = (1 - \alpha)s_d^2 + \alpha s_e^2$
- Transformed variance $\sigma^2 = a^2 \nu = 2\mu$
Math 101 again

Raw score means and variances $m_{d,e}, s_{d,e}^2$.

- Transformed target mean: $am_e + b = \mu$
- Transformed non-target mean $am_d + b = -\mu$
- Weighted variance $\nu = (1 - \alpha)s_d^2 + \alpha s_e^2$
- Transformed variance $\sigma^2 = a^2 \nu = 2\mu$

...results in solution

- $a = \frac{m_e - m_d}{\nu}$
- $b = -a \frac{m_e + m_d}{2}$

This is a closed-form solution!

Constrained Maximum Likelihood Gaussian: CMLG
First calibration experiment: Miranti’s scores

- RUN i-vector PLDA system
- calibrate on SRE-2008, evaluate using $C_{llr}$ on SRE-2010
- 25 different duration-combinations, to sample range of performances
- Two linear calibration methods
  - Logistic regression
  - This method (CMLG)
Second experiment: Niko’s scores

- Agnitio Research’s SRE-2012 system and scores
- Calibrated using their dev-set
- Evaluated using $C_{primary}$
  - official SRE-2012 metric
  - sensitive to low-FA range
- Contrasting
  - Niko + GD
    Interspeech 2013
  - This method
    CMLG

Comparison of calibration methods

$\log(\alpha) - \log(1 - \alpha)$

logistic regression
CMLG

Comparison of calibration methods

$C_{primary}$

-8 -6 -4 -2 0 2
0.30 0.35 0.40 0.45 0.50 0.55 0.60

logistic regression
CMLG
Conclusions

- We can prove that “the LLR of the LLR is the LLR”
  - ... already in exam questions course Forensic Linguistics...

- Well calibrated Gaussian non-target scores imply
  - Gaussian target scores
  - with same variance
  - and opposite mean
  - and a variance that is equal to the difference in means

- We can use it to find calibration parameters
  - as a closed-form solution
  - that gives same performance as logistic regression, for
    - two different systems
    - two different evaluation data bases
    - two different calibration-sensitive evaluation metrics