INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Out of a hundred trials, how many errors does your speaker verifier make?

Niko Brümmer[1], *Luciana Ferrer*[2], *Albert Swart*[1]

[1]Phonexia, South Africa
[2]Instituto de Investigación en Ciencias de la Computación, CONICET-UBA, Argentina
niko.brummer@gmail.com, lferrer@dc.uba.ar

Abstract

Out of a hundred trials, how many errors does your speaker verifier make? For the user this is an important, practical question, but researchers and vendors typically sidestep it and supply instead the conditional error-rates that are given by the ROC/DET curve. We posit that the user's question is answered by the Bayes error-rate. We present a tutorial to show how to compute the error-rate that results when making Bayes decisions with calibrated likelihood ratios, supplied by the verifier, and an hypothesis prior, supplied by the user. For perfect calibration, the Bayes error-rate is upper bounded by min(EER,P,1-P), where EER is the equal-error-rate and P, 1-P are the prior probabilities of the competing hypotheses. The EER represents the accuracy of the verifier, while min(P,1-P) represents the hardness of the classification problem. We further show how the Bayes error-rate can be computed also for non-perfect calibration and how to generalize from error-rate to expected cost. We offer some criticism of decisions made by direct score thresholding. Finally, we demonstrate by analyzing error-rates of the recently published DCA-PLDA speaker verifier.

Index Terms: speaker recognition, calibration, Bayes decisions

1. Introduction

This is a position paper and tutorial about the decision stage of speaker verification and how to quantify verification accuracy. Our position is:

- Bayes decisions are preferable to ad-hoc alternatives.
- The Bayes error-rate (the accuracy of Bayes decisions) is a good representation of the accuracy of the verifier.

A *speaker verification trial* provides two speech samples that may have been spoken by one-and-the-same speaker (hypothesis H_1), or by two different speakers (H_2). It is required to make an accept/reject decision, where *accept* favours H_1 and *reject* favours H_2. The speech input of the trial is processed by a speaker verifier that outputs a *score*, $s \in \mathbb{R}$, where by convention, more positive s supports H_1, while more negative s supports H_2. For some verifiers, this is all that can be said of the scores and such scores are termed *uncalibrated*. Some verifiers output their scores in (log) likelihood-ratio format, termed *calibrated*. Uncalibrated scores can be post-calibrated by adding a calibration stage to the verifier.

Bayes decisions can be made with either calibrated, or uncalibrated scores as will be explained in this tutorial. The former method provides the user with more flexibility and allows for more informative representation of the verifier accuracy. Decisions *can* also be made by comparing scores to ad-hoc thresholds, obtained e.g. by fixing the false-accept rate. While such methods avoid both likelihood-ratio calibration and specification of prior and costs by the user, we offer some strong criticisms.

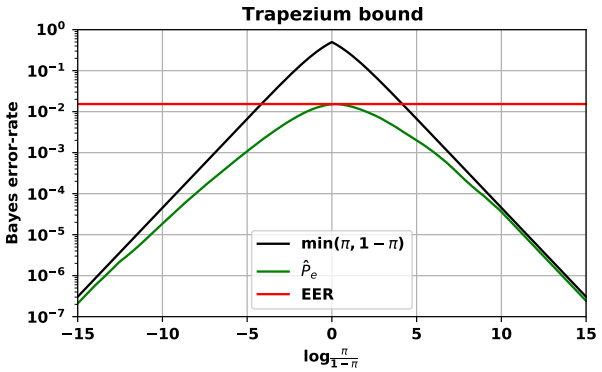

Figure 1: *The trapezium bound:* $\hat{P}_e \leq \min(\pi, 1-\pi, \mathrm{EER})$. *The axes have logit and log scales, to better show small values.*

The tutorial does not discuss how to calibrate scores, but rather how to make decisions given calibrated scores—and how to compute the accuracy of these decisions. We discuss accuracy measures that can be obtained by testing the verifier on a supervised evaluation database. Perhaps the most ubiquitously available accuracy statistic is the equal-error-rate (EER). We highlight a little-known relationship between the EER and the Bayes error-rate: for a well-calibrated verifier, the EER is an approximate upper bound to the Bayes error-rate. More generally, given an ROC/DET curve [1] of the form $P_{\text{miss}}(P_{\text{fa}})$, the Bayes-error-rate for a well-calibrated verifier can be computed as a function of the prior, $P(H_1)$. Most generally, given $P_{\text{miss}}(\theta)$ and $P_{\text{fa}}(\theta)$ as functions of a score threshold, θ, the actual Bayes error-rate can be computed for verifiers that may or may not be well-calibrated.

The theory in several sections below is followed by a demonstration of our methods on the recently published DCA-PLDA speaker verifier [2].

2. Optimal Bayes decisions

We start with a concise summary of some well-known elements of Bayes decision theory [3], applied to binary classification and in particular to speaker verification [4, 5, 6]. In this section, we analyze Bayes decisions in the idealized case of perfect calibration. In the next section, we generalize to realistic, practical calibration.

Consider a verification trial represented as (h, s), where $h \in \{H_1, H_2\}$ is the unknown hypothesis and $s \in \mathbb{R}$ is the verifier score. Given s, we want to make an accept/reject decision. The outcomes (accept, H_1) and (reject, H_2) are considered correct, while the two kinds of classification errors are *miss*: (reject, H_1); and *false-accept*: (accept, H_2). We now suppose that the decision logic has access to the joint distribu-

http://dx.doi.org/10.21437/Interspeech.2021-541

tion:

$$P(h, s) = P(h)P(s \mid h) \tag{1}$$

This implies perfect calibration: the prior, $P(h)$, typically supplied by the user, must equal the actual frequency of H_1 vs H_2 and the likelihoods $P(s \mid H_1)$ and $P(s \mid H_2)$, typically supplied in likelihood-ratio form by the verifier, must equal the actual conditional score distributions. Given these resources, we can compute the hypothesis posterior:

$$P(h \mid s) = \frac{P(s, h)}{P(s, H_1) + P(s, H_2)} \tag{2}$$

If we choose to accept or reject, the probability of error is respectively $P(H_2 \mid s)$ or $P(H_1 \mid s)$. So to minimize the probability of error, make the *Bayes decision*:[1]

$$\text{decision} = \begin{cases} \text{accept} & \text{if } P(H_1 \mid s) \geq P(H_2 \mid s) \\ \text{reject} & \text{if } P(H_1 \mid s) \leq P(H_2 \mid s) \end{cases} \tag{3}$$

This can be rewritten, using basic probability rules, as:

$$\text{decision} = \begin{cases} \text{accept} & \text{if } r \geq \theta_\pi \\ \text{reject} & \text{if } r \leq \theta_\pi \end{cases} \tag{4}$$

where r is the *likelihood-ratio* and θ_π is the *Bayes threshold*:

$$r = \frac{P(s \mid H_1)}{P(s \mid H_2)} \tag{5}$$

$$\theta_\pi = \frac{P(H_2)}{P(H_1)} = \frac{1-\pi}{\pi} \tag{6}$$

where $\pi = P(H_1)$ is short-hand. With this optimal decision rule, the probability for a classification error is:[2]

$$\begin{aligned} \hat{P}_e(\pi) &= P(\text{reject}, H_1) + P(\text{accept}, H_2) \\ &= \pi P(\text{reject} \mid H_1) + (1-\pi)P(\text{accept} \mid H_2) \\ &= \pi P_{\text{miss}}(\theta_\pi) + (1-\pi)P_{\text{fa}}(\theta_\pi) \end{aligned} \tag{7}$$

The above conditional error-rates can be expanded as:

$$P_{\text{miss}}(\theta) = P(\text{reject} \mid H_1, \theta) = \int_0^\theta P(r \mid H_1)\, dr \tag{8}$$

$$P_{\text{fa}}(\theta) = P(\text{accept} \mid H_2, \theta) = \int_\theta^\infty P(r \mid H_2)\, dr \tag{9}$$

The ROC curve can be generated[3] by sweeping π from 0 to 1 and plotting P_{miss} versus P_{fa}. As π is increased, P_{miss} decreases monotonically and P_{fa} increases monotonically.

The full function, $\hat{P}_e(\pi)$ provides a good answer to the question of optimal verifier accuracy, but let's continue to seek a more concise summary via further analysis. If there is no prior uncertainty the error-rate vanishes: $\hat{P}_e(0) = \hat{P}_e(1) = 0$. To understand what happens for $0 < \pi < 1$, we differentiate $\hat{P}_e$ w.r.t. π:

$$\hat{P}'_e = P_{\text{miss}} + \pi P'_{\text{miss}} - P_{\text{fa}} + (1-\pi)P'_{\text{fa}} \tag{10}$$

where, using (6) in (8) and (9):

$$P'_{\text{miss}} = \frac{dP_{\text{miss}}}{d\theta}\frac{d\theta}{d\pi} = -\frac{P(\theta \mid H_1)}{\pi^2} \tag{11}$$

$$P'_{\text{fa}} = \frac{dP_{\text{fa}}}{d\theta}\frac{d\theta}{d\pi} = \frac{P(\theta \mid H_2)}{\pi^2} \tag{12}$$

To complete our derivation, we invoke the *calibration property* [7] of the likelihood-ratio (5):

$$\frac{P(s \mid H_1)}{P(s \mid H_2)} = r = \frac{P(r \mid H_1)}{P(r \mid H_2)} \tag{13}$$

Using this at $r = \theta = \frac{1-\pi}{\pi}$, the 2nd and 4th terms in (10) cancel, to give:

$$\hat{P}'_e(\pi) = P_{\text{miss}}(\theta_\pi) - P_{\text{fa}}(\theta_\pi) \tag{14}$$

After checking that $\hat{P}'' < 0$, we see that $\hat{P}_e$ is concave, with a unique maximum of $\text{EER} = P_{\text{miss}}(\theta_{\pi^*}) = P_{\text{fa}}(\theta_{\pi^*})$, where the derivative vanishes at $\pi = \pi^*$:

$$\begin{aligned} \max_\pi \hat{P}_e &= \pi^* P_{\text{miss}}(\theta_{\pi^*}) + (1-\pi^*)P_{\text{fa}}(\theta_{\pi^*}) \\ &= \pi^* \,\text{EER} + (1-\pi^*)\,\text{EER} = \text{EER} \end{aligned} \tag{15}$$

This upper-bound property of the EER is not well known. It was mentioned in [8] and received detailed treatment in [9], but the derivation above is new and we hope more intuitive. Next, we invoke another upper bound [9]:

$$\hat{P}_e(\pi) \leq \min(\pi, 1-\pi) \tag{16}$$

where the RHS is the probability for error when making a Bayes decision using the prior alone, while the LHS has the advantage of the extra information supplied by the score. The RHS becomes small whenever $\pi \approx 0$, or $\pi \approx 1$, which shows that the problem becomes easier when the prior uncertainty is low.

In summary of the above results, we can now provide a first answer to our question:

Q: Out of a hundred trials, how many errors does your speaker verifier make?

A: The error-rate depends not only on the accuracy of the verifier, but also on the user-supplied prior π. For perfect calibration, the error-rate, $\hat{P}_e$, obeys the *trapezium bound*:

$$\hat{P}_e(\pi) \leq \min(\pi, 1-\pi, \text{EER}) \tag{17}$$

wherein the verifier accuracy is represented by the EER, and the user contribution by $\min(\pi, 1-\pi)$. See figure 1 for an example.

Q: What if calibration is not perfect?

A: Calibration can be measured (see the next section), using the same resources as for the ROC, namely a supervised database of scores. Since calibration can be measured, it can be optimized, so that the actual error-rates can get close to the optimal $\hat{P}_e$.

[1] The decision at equiprobable hypotheses can be chosen arbitrarily.

[2] The $\hat{}$ on $\hat{P}_e$ refers to the optimality given by perfect calibration—later, $\tilde{P}_e$ will be used for the actual error-rate, when perfect calibration cannot be assumed.

[3] The ROC can also be generated from raw scores, s, or from $\log(r)$. For monotonic transforms, $s \to r$ all three ROCs are identical.

3. Actual Bayes decisions

In a real, practical application of a speaker verifier, perfect calibration cannot be assumed, neither for the user-supplied prior, nor for the system-supplied likelihood-ratio. If either of these resources are not perfectly calibrated then the optimal error-rate, $\hat{P}_e$ of (7) becomes a lower bound to the actual error-rate. For bad calibration, the error-rate could become arbitrarily bad, up to a maximum of 1. To see this, consider for example a user that supplies $\pi = 1$, when in actual fact only H_2 trials arrive at the verifier, in which case every decision will be an error.

We shall however dismiss the problem of imperfectly calibrated π, because in this paper, we are interested in quantifying the verifier accuracy and the verifier should not be blamed for user errors. Keep in mind that π cannot be extracted by the verifier from the speech inputs, nor can it be learnt from a typical verifier training database, because π is entirely dependent on the application. In what follows, we shall assume perfect π calibration.

Now, let $P(s \mid h)$ denote the actual conditional score distributions and let $r = \frac{P(s|H_1)}{P(s|H_2)}$ represent the perfectly calibrated likelihood-ratio, which is in practice *not available* to the decision logic. Instead, we assume that the decision logic has access to a *calibration function*, $\tilde{r} = f(s)$, that is designed to give $\tilde{r} \approx r$. (For some systems, the score $s \approx r$ is already well-calibrated, e.g. [10, 2]. For such systems we take f to be the identity function.) Other examples of calibration can be found in [6, 4, 11, 12, 13, 14, 15, 16, 17], and in many other systems submitted to speaker recognition evaluations.

To analyze the actual error-rate that results from using $\tilde{r}$, we use it to replace r in the decision rule (4), while θ_π is still the same theoretically optimal threshold (6). The resulting *actual error-rate* is:[4]

$$\tilde{P}_e(\pi) = \pi \tilde{P}_{\text{miss}}(\theta_\pi) + (1-\pi)\tilde{P}_{\text{fa}}(\theta_\pi) \tag{18}$$

where

$$\tilde{P}_{\text{miss}}(\theta) = \int_0^\theta P(\tilde{r} \mid H_1)\, d\tilde{r} \tag{19}$$

$$\tilde{P}_{\text{fa}}(\theta) = \int_\theta^\infty P(\tilde{r} \mid H_2)\, d\tilde{r} \tag{20}$$

To evaluate $\tilde{P}_e$ in practice, the usual recipe requires the same resource as for the ROC: a supervised database of scores, where $\tilde{r}$ and h are available for every trial. The conditionals, $P(\tilde{r} \mid h)$ are now in empirical form (impulses at the data points) and the integrals reduce to counting the errors that result when thresholding all calibrated scores against a set of thresholds of interest. This can be done[5] efficiently by jointly sorting scores and thresholds and retrieving error-rates from the ranks of the thresholds [4].

Since generally $\tilde{r} \neq r$, neither the calibration property (13), nor the trapezium bound (17) apply to $\tilde{P}_e$. Since $\hat{P}_e$ is optimal and $\tilde{P}_e$ depends on a (hopefully only slightly) suboptimal decision rule, we know that:

$$\tilde{P}_e(\pi) \geq \hat{P}_e(\pi) \tag{21}$$

If the calibration is good however, then $\tilde{P}_e \approx \hat{P}_e$ and the trapezium will still function as *approximate* upper bound to the actual error-rate:

$$\tilde{P}_e(\pi) \approx \hat{P}_e(\pi) \leq \min(\pi, 1-\pi, \text{EER}) \tag{22}$$

For bad calibration however, it is possible for $\tilde{P}_e$ to deteriorate up to 1 for some values of π. See figures 2 and 3, in the section on experiments below, where we show some examples of both good and bad calibration. We conclude with further Q&A:

Q: Which resources are required for good score calibration?

A: The same as for any other machine learning problem: algorithms and data—see the references in this section. Methods vary in their data requirements, from large to small and from supervised to unsupervised.

Q: It seems that not only score calibration, but also the user's π calibration are too hard. Is it not safer and better to directly threshold the scores, with a threshold set at a fixed false-accept rate?

A: Easier, maybe. Safer and better, no. Read on.

4. Generalization to expected cost

Error-rate can be generalized to risk (expected cost), by assigning the respective costs, $C_{\text{miss}}, C_{\text{fa}} > 0$ to miss and false-accept errors. The Bayes decision threshold generalizes to:

$$\theta_\pi^c = \frac{(1-\pi)C_{\text{fa}}}{\pi C_{\text{miss}}} \tag{23}$$

For perfect calibration, $\hat{P}_e$ generalizes to *optimal expected cost*:

$$\hat{C}_e(\pi) = \pi C_{\text{miss}} P_{\text{miss}}(\theta_\pi^c) + (1-\pi)C_{\text{fa}} P_{\text{fa}}(\theta_\pi^c) \tag{24}$$

The trapezium bound becomes [4]:

$$\hat{C}_e(\pi) \leq \min(\pi C_{\text{miss}}, (1-\pi)C_{\text{fa}}, R^*) \tag{25}$$

where R^* is the *equal risk*, obtained at a threshold, θ^*, such that:

$$R^* = C_{\text{miss}} P_{\text{miss}}(\theta^*) = C_{\text{fa}} P_{\text{fa}}(\theta^*) \tag{26}$$

In this sense, depending on the ratio $\frac{C_{\text{miss}}}{C_{\text{fa}}}$, any point on the ROC can act as upper bound to the optimal expected cost. For practical calibration, actual expected cost, $\tilde{C}_e$, can be evaluated in a straight-forward generalization of section 3, see [4].

5. Direct score thresholding

Given a verifier score, s, that is not specifically calibrated to function as likelihood-ratio, consider the decision rule:

$$\text{decision} = \begin{cases} \text{accept} & \text{if } s \geq \bar{\theta} \\ \text{reject} & \text{if } s \leq \bar{\theta} \end{cases} \tag{27}$$

One way to proceed, when given a *calibration database* of supervised scores, is to choose the score threshold, $\bar{\theta} = \bar{\theta}_\pi^c$, to be approximately equivalent to the optimal likelihood-ratio threshold, θ_π^c, by doing:[6]

$$\bar{\theta}_\pi^c = \underset{\theta'}{\text{argmin}}\, \pi C_{\text{miss}} \bar{P}_{\text{miss}}(\theta') + (1-\pi)C_{\text{fa}} \bar{P}_{\text{fa}}(\theta') \tag{28}$$

where $\bar{P}_{\text{miss}}, \bar{P}_{\text{fa}}$ are defined empirically by counting errors when thresholding scores against θ' in the calibration database.

[4] To analyze prior miscalibration: use the true hypothesis frequencies here, but a miscalibrated prior, $\tilde{\pi}$ to compute $\theta_{\tilde{\pi}}$.

[5] See implementations: `github.com/bsxfan/PYLLR`

[6] $P_{\text{miss}}, \tilde{P}_{\text{miss}}, \bar{P}_{\text{miss}}$ are computed from respectively $r, \tilde{r}, s$.

If we assume $\frac{P(s|H_1)}{P(s|H_2)}$ is monotonic rising, then the optimal expected cost can be computed even from *uncalibrated* scores:

$$\hat{C}_e(\pi) = \min_{\theta'} \pi C_{\text{miss}} \bar{P}_{\text{miss}}(\theta') + (1-\pi) C_{\text{fa}} \bar{P}_{\text{fa}}(\theta') \qquad (29)$$

This is equivalent to defining a calibration function, $\tilde{r} = f(s)$, non-parametrically, via the PAV algorithm [11]. Given an independent *test database*, the actual risk at $\bar{\theta} = \bar{\theta}^c_\pi$:

$$\tilde{C}_e = \pi C_{\text{miss}} \bar{P}_{\text{miss}}(\bar{\theta}^c_\pi) + (1-\pi) C_{\text{fa}} \bar{P}_{\text{fa}}(\bar{\theta}^c_\pi) \qquad (30)$$

gives a more realistic representation of the accuracy at this operating point. Note $\hat{P}_e$ and $\tilde{P}_e$ can be obtained from $\hat{C}_e$ and $\tilde{C}_e$ by setting $C_{\text{miss}} = C_{\text{fa}} = 1$.

Another popular solution is to make use of a calibration database containing only H_2 scores (which are often easier to collect than H_1 scores) and to choose $\bar{\theta}$ so that some fixed proportion of H_2 scores, say 1% of them, are above the threshold, thus fixing (at least on this data) the false-accept rate. If H_1 trials are also available, the miss rate at this threshold can be reported as a representation of accuracy.

Does fixing the false-accept rate thus, make this a safe strategy? No, the rate is fixed exactly *only* for the dataset that was used to choose the threshold. There is no guarantee on unseen data—neither for this method, nor for any other calibration method. Moreover, this method does not lend itself to reporting accuracy on an independent test database, because on the test data, *both* the miss and false-accept rates could change.

Surely, this strategy has the advantage that the user does not have to be troubled to provide prior and costs, especially if those could be badly calibrated? No, ignoring prior and costs, does not cause their essential role in making optimal decisions go away [8, 3, 18]. Fixing the false-accept rate can be seen as crude way to take the prior and costs into account. Put simply: a bad choice of false-accept rate can have exactly the same effect as a bad choice of π, C_{miss} or C_{fa}.

In summary, we find no advantages to direct score thresholding, other than its perceived simplicity. At best, it is less flexible than likelihood-ratio calibration, because—unlike the adjustable Bayes threshold—the direct score threshold is fixed. Moreover, the fixed-false-accept method has additional disadvantages: it cannot be tested on independent test data; it obfuscates the role of cost and prior; and in general it does something different from Bayes decisions, so that it does *not* minimize error-rates, nor expected costs.

6. Experimental demonstration

Figure 2 shows plots[7] of the actual Bayes error-rate $\tilde{P}_e$, for an x-vector speaker verifier, with a standard (generative) PLDA scoring backend, followed by an affine log-likelihood-ratio calibration transform, as implemented in [2]. The system was evaluated on three different data sets. On the red and green data sets, the ideal EER bound on $\hat{P}_e$ is somewhat exceeded, but the prior bound, $\min(\pi, 1-\pi)$ is not exceeded. On the blue data set, calibration is very bad, it far exceeds both the EER and the prior bound. The blue data is very clean speech that gives a very low EER (about 0.1%), but this dataset shift effect causes the calibration to fail.

[7] Code is provided at `github.com/bsxfan/PYLLR`.

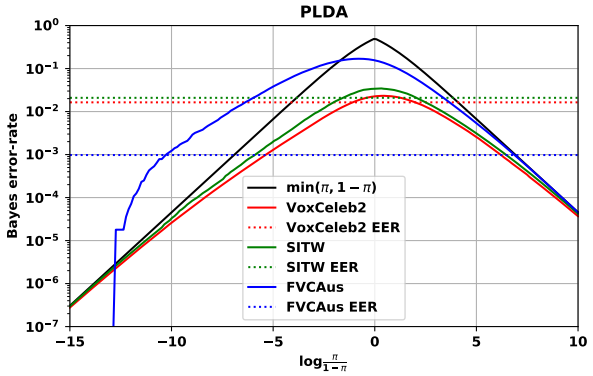

Figure 2: *PLDA: $\tilde{P}_e$ evaluated on three data sets, vs respective trapezium bounds,* $\min(\pi, 1-\pi, \text{EER})$.

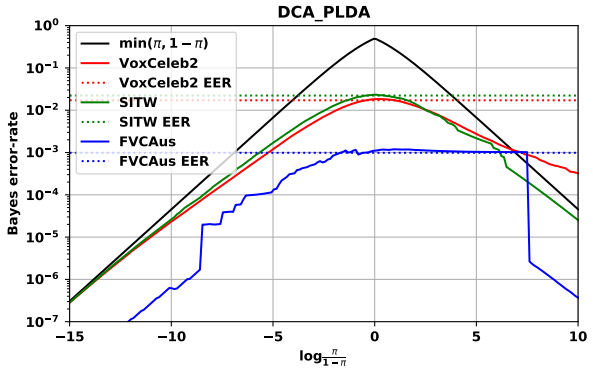

Figure 3: *DCA-PLDA: $\tilde{P}_e$ evaluated on three data sets, vs respective trapezium bounds,* $\min(\pi, 1-\pi, \text{EER})$.

Figure 3 shows a *more complex* backend, DCA-PLDA [2], that was discriminatively trained on a *more diverse* database. It shows much better calibration on all three data sets.[8]

7. Conclusion

Calibrated scores are more versatile than uncalibrated ones. Bayes decisions minimize error-rates and risk. Ad-hoc decisions don't. The Bayes error-rate and risk are more informative to prospective users than P_{miss} and P_{fa}. Calibration and computing Bayes risk both require some hard work, but the benefits that can be thus obtained can be previewed before investing in this work: the *trapezium upper bound* on the Bayes error-rate (risk) can be computed from the EER (ROC) and these can be computed from uncalibrated scores with standard methods. Although this bound applies only to perfectly calibrated scores, it provides an ideal that can be pursued by practical calibration strategies.

8. Acknowledgements

We would like to thank Corne van Biljon of Gendac (Pty) Ltd, for insisting on a simple and direct answer to his question, "Out of a hundred trials, how many errors would a typical speaker verifier make?"

[8] We have preliminary evidence to suggest the violation of the prior bound on the far right may be due to labelling errors in the VoxCeleb 2 test set.

9. References

[1] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *EUROSPEECH*, Rhodes, Greece, Sept 1997, pp. 1895–1898.

[2] L.Ferrer, M. McClaren, and N. Brümmer, "A speaker verification backend with robust performance across conditions," *Submitted to Computer Speech, and Language*, 2021. [Online]. Available: https://arxiv.org/abs/2102.01760

[3] M. H. DeGroot, *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[4] N. Brümmer and E. de Villiers, "The BOSARIS Toolkit: Theory, algorithms and code for surviving the new DCF," 2013. [Online]. Available: https://arxiv.org/abs/1304.2865

[5] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I: Fundamentals, Features, and Methods*, ser. Lecture Notes in Computer Science, C. A. Müller, Ed., vol. 4343. Springer, 2007, pp. 330–353. [Online]. Available: https://doi.org/10.1007/978-3-540-74200-5_19

[6] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[7] D. A. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proceedings INTERSPEECH, Lyon, France*, 2013.

[8] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. John Wiley & Sons, 1994.

[9] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch University, 2010. [Online]. Available: https://scholar.sun.ac.za/handle/10019.1/5139

[10] N. Brümmer, L. Burget, J. Černocky, O. Glembek, F. Grézl, M. Karafíat, D. A. van Leeuwen, P. Matějka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *IEEE TASLP*, vol. 15, no. 7, September 2007.

[11] N. Brümmer, A. Swart, and D. van Leeuwen, "A comparison of linear and non-linear calibrations for speaker recognition," in *Odyssey 2014: The Speaker and Language Recognition Workshop, Joensuu, Finland*, 2014.

[12] N. Brümmer and A. Swart, "Bayesian calibration for forensic evidence reporting," in *INTERSPEECH, Singapore*, 2014.

[13] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," in *ICASSP*, 2014.

[14] S. Cumani and P. Laface, "Tied normal variance–mean mixtures for linear score calibration," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6121–6125.

[15] S. Cumani, "Normal variance–mean mixtures for unsupervised score calibration," in *Interspeech, Graz, Austria*, 2019.

[16] S. Cumani, "On the distribution of speaker verification scores: Generative models for unsupervised calibration," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 547–562, 2021.

[17] L. Ferrer, M. K. Nandwana, M. McLaren, D. Castan, and A. Lawson, "Toward fail-safe speaker recognition: Trial-based calibration with a reject option," *IEEE/ACM Trans. Audio Speech and Language Processing*, vol. 27, Jan. 2019.

[18] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.