Making Likelihood Ratios Digestible for Cross-Application Performance Assessment

Andreas Nautsch, *Student Member, IEEE*, Didier Meuwly, Daniel Ramos, Jonas Lindh, and Christoph Busch

***Abstract*—Performance estimation is crucial to the assessment of novel algorithms and systems. In detection error tradeoff (DET) diagrams, discrimination performance is solely assessed targeting one application, where cross-application performance considers risks resulting from decisions, depending on application constraints. For the purpose of interchangeability of research results across different application constraints, we propose to augment DET curves by depicting systems regarding their support of security and convenience levels. Therefore, application policies are aggregated into levels based on verbal likelihood ratio scales, providing an easy to use concept for business-to-business communication to denote operative thresholds. We supply a reference implementation in Python, an exemplary performance assessment on synthetic score distributions, and a fine-tuning scheme for Bayes decision thresholds, when decision policies are bounded rather than fix.**

***Index Terms*—Bayes decision framework, biometric verification, binary decisions, detection error tradeoff (DET), verbal scales.**

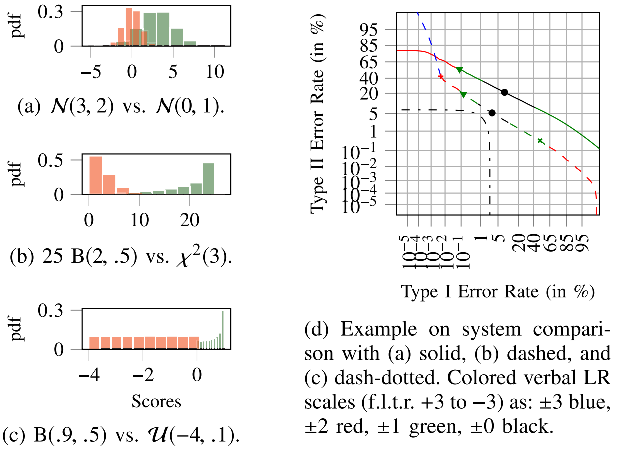
I. Introduction

PERFORMANCE estimation of binary decisions is essential to the research and development of two-class machine learning problems such as biometric verification. Conventionally, biometric verification performance [1], [2] is reported by depicting the tradeoff between type I and type II error rates in form of a detection error tradeoff (DET) plot [3]. Thereby, scalar representations provide easy tractability at distinct operating points, such as the equal-error rate (EER).

Research in biometrics is constrained to specific performance characteristics, such as specified within border control regulations and in mobile banking, whereas forensic evaluation aims at application-independent performance estimates.[1] Biometric research conducted under different optimization constraints is limited in comparability, since constrained measures are reported, which may be irrelevant under different decision policies. For the purpose of improving the ability of interchanging research ideas and results, while accounting for the application domain of a system, we perform the following.

1) Propose an augmentation to DET diagrams based on a verbal security scale, depicting the intercorrelation of error tradeoffs to decision risk by security/convenience levels, utilizing the concept of an angular operating point.
2) Propose a scheme to derive representative operating points per band of verbally alike decision policies, based on which operative thresholds can be adjusted with respect to relative changes in parameters characterizing a decision policy (for optimizing usiness-to-business (B2B) communication).
3) Provide a reference implementation, supporting the BOSARIS toolkit [4] functionality via the SIDEKIT [5].

Exemplary, the proposed DET augmentation is depicted in Fig. 1 on three systems of simulated scores. Considering any binary decision score effectively is a likelihood ratio (LR),



(a) $\mathcal{N}(3, 2)$ vs. $\mathcal{N}(0, 1)$.

(b) 25 B(2, .5) vs. $\chi^2(3)$.

(c) B(.9, .5) vs. $\mathcal{U}(-4, .1)$.

(d) Example on system comparison with (a) solid, (b) dashed, and (c) dash-dotted. Colored verbal LR scales (f.l.t.r. +3 to −3) as: ±3 blue, ±2 red, ±1 green, ±0 black.

Fig. 1. Proposed system comparison on simulated scores (normal $\mathcal{N}$, beta B, chi-square χ^2, uniform $\mathcal{U}$ distributions), with (a – c) depicting probability density functions (pdf) of 5×10^4 mated (green) and 1×10^5 nonmated scores (red). In (d), colors depict verbal scales, i.e., which LRs can be supported by a system, markers denote depending verbal scale centers.

Manuscript received June 21, 2017; revised August 29, 2017; accepted August 30, 2017. Date of publication September 4, 2017; date of current version September 18, 2017. This work was supported in part by the German Federal Ministry of Education and Research as well as in part by the Hessen State Ministry for Higher Education, Research, and the Arts within the Center for Research in Security and Privacy (www.crisp-da.de), and in part by the BioMobile II Project (518/16-30). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philip N. Garner. *(Corresponding author: Andreas Nautsch.)*

A. Nautsch and C. Busch are with the da/sec — Biometrics and Internet Security Research Group, Hochschule Darmstadt, Darmstadt 64295, Germany (e-mail: andreas.nautsch@h-da.de; christoph.busch@h-da.de).

D. Meuwly was with the Universiteit Twente, Enschede 7522NB, The Netherlands, and also with Netherlands Forensic Institute, The Hague 2490AA, The Netherlands (e-mail: d.meuwly@nfi.minvenj.nl).

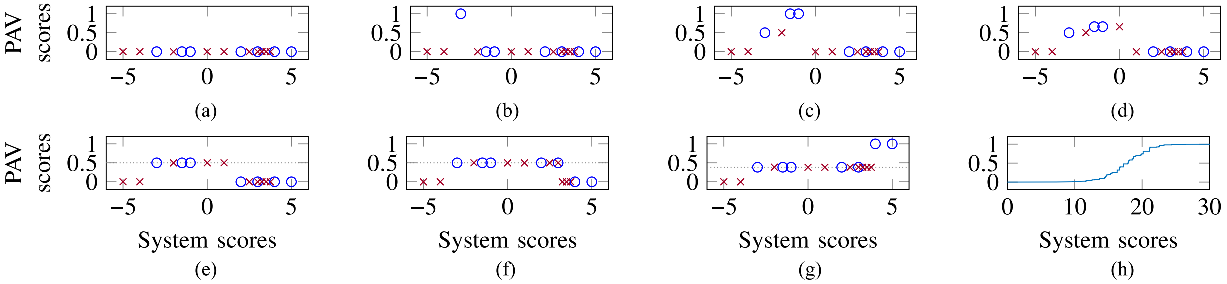
D. Ramos was with the ATVS — Biometric Recognition Group, EPS, Universidad Autónoma de Madrid, Madrid 28049, Spain (e-mail: daniel.ramos@uam.es).

J. Lindh was with the Göteborgs Universitet, Gothenburg 405 30, Sweden (e-mail: jonas@voxalys.se).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2748899

[1] In order to decouple decision policies (province of the trier of fact) from the assignment of the strength of evidence (duty of the forensic practitioner).



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Example on the PAV algorithm on selected iterations over score-aligned class labels, crosses: nonmated scores, circles: mated scores. In (h), PAV is depicted on a score simulation: PAV scores resemble optimal calibration, therefore system scores are grouped by their overall contribution to decision making. (a) Initial. (b) Iteration 2. (c) Iteration 5. (d) Iteration 6. (e) Iteration 7. (f) Iteration 11. (g) final Iteration. (h)χ^2_3 versus $30B_{2,1/2}$ simulation.

LR-based decisions are made utilizing the Bayes risk, and that DET curves are threshold independent, the proposed DET augmentation provides transparency regarding each system's discrimination power. In the DET space, the receiver operating characteristic's convex hull (ROCCH) is depicted, i.e., optimal-calibrated LR scores considering Bayes operating points, where calibration mappings depend on the score distributions of each system. ROCCH curves are afflicted with further discrimination information: The application levels supported by a system, improving the interchangeability of research. In the example, system (c) yields the best error tradeoff, however solely supporting the ±0 level (without the equal tradeoff verbal scale center in the visible area), whereas systems (a,b) are capable of supporting wider ranges of applications.

This letter is organized as follows: Sections II and III depict related work on performance estimation in biometrics, and on likelihood ratio verbal scales introduced in forensic evaluation. In Section IV, we propose augmenting DET curves, visualizing the intercorrelation between error rates and decision risk performance. Discussion and conclusion are provided in Section V.

II. Performance Paradigms in Biometrics

Conventionally, receiver operating curves (ROC) are discussed, e.g., in iris biometrics, whereas y-inverted log-compress ROCs are considered, e.g., in face biometrics. Contrastively, voice biometrics refers to DET plots, which are y-inverted ROCs on Gaussian-scaled axes. Biometric standardization of performance testing reports [1], [2] aims at harmonization and reproducibility of evaluations, utilizing a harmonized vocabulary[2] [6]. The standard requires DET plots, when algorithmic and system performance are reported,[3] i.e., to depict error rates in a quantile-quantile plot utilizing the probit transform [3]. However, the standard spares an assessment of decision risk, e.g., the impact of varying costs associated to error types across applications. In order to examine decision risks of biometric systems under different decision policy constraints, calibration is relevant [7]–[11].

A. Unified Calibration: Pool Adjacent Violators (PAV) Algorithm

The PAV algorithm [12], [13] maps scores of any distribution into probabilities, by conducting an isotonic regression of score-aligned class labels as 0 s and 1 s. Thereby, the PAV algorithm poses optimal calibration for two-class problems, mapping the entire range of score values to a unified score space, cf., Fig. 2. Furthermore, as shown in [4] and [7], PAV relates to the convex hull of the ROC (ROCCH) [14] as well as to the minimum decision cost function[4] (minDCF).

B. Impact of Bayes Decision Theory

In Bayes decision theory, the consequence of a decision outcome is expressed as a cost function. For the purpose of biometric verification, correct outcomes are assigned with zero cost, leaving the type I error cost C_I and the type II error cost C_II to be specified. Given a (evidence) score s, a confirming Bayes decision minimizes the *a posteriori* risk [4], [7]

$$P(\text{nonmated}\,|\,s,\pi)\,C_\mathrm{I} \leq P(\text{mated}\,|\,s,\pi)\,C_\mathrm{II} \qquad (1)$$

where a target prior probability π denotes the (effective) ratio of the two mutually exclusive hypotheses[5] *mated* and *nonmated*. Considering the *Bayes' theorem* and log likelihood ratios (LLRs) as similarity scores s_LLR in (1), Bayes thresholds η are denoted with $\operatorname{logit} x = \log\frac{x}{1-x}$ as[6]

$$\log\frac{C_\mathrm{I}}{C_\mathrm{II}} - \operatorname{logit}\pi = \eta \;\leq\; s_\mathrm{LLR} = \log\frac{P(s\,|\,\text{mated})}{P(s\,|\,\text{nonmated})}. \qquad (2)$$

C. Bayes Operating Points in y-Inverted ROC Space

In this context, the Bayes risk as a DCF is computed for an empirical set of scores S given a specific operating point $(\pi, C_\mathrm{I}, C_\mathrm{II})$ with respect to type I and type II error rates $p_\mathrm{I}(\eta), p_\mathrm{II}(\eta)$ as [7]

$$\mathrm{DCF}(S\,|\,\pi, C_\mathrm{I}, C_\mathrm{II}) = \pi\, C_\mathrm{II}\, p_\mathrm{II}(\eta) + (1-\pi)\, C_\mathrm{I}\, p_\mathrm{I}(\eta) \qquad (3)$$

where [7] further introduces an effective prior mapping the operating point into a scalar representation $\tilde{\pi}$

$$\tilde{\pi} = \frac{\pi\, C_\mathrm{II}}{\pi\, C_\mathrm{II} + (1-\pi)\, C_\mathrm{I}}, \quad \text{with } \eta = -\operatorname{logit}\tilde{\pi}$$

$$\mathrm{DCF}(S\,|\,\tilde{\pi}) = \tilde{\pi}\, p_\mathrm{II}(\eta) + (1-\tilde{\pi})\, p_\mathrm{I}(\eta). \qquad (4)$$

In other words, the Bayes operating point is a linear combination of the type I and type II error rates, and the ROCCH

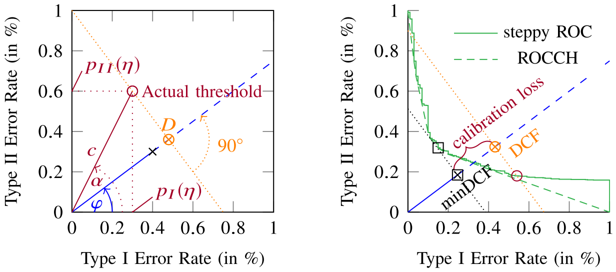
[2] Example: *Match* is a result, while *mated* is a statement, i.e., *of same source*.

[3] In [1], type I / II errors are referred to on algorithm domain as false match rate and false nonmatch rate, and on system domain as false accept rate and false reject rate, incorporating precomparison as well as algorithm errors.

[4] Furthermore, the ROCCH's EER is obtained by $\max(\mathrm{minDCF})$ [4], [7].

[5] For binary decisions, the Bayes decision framework requires the hypotheses to be mutually exclusive, but not to be exhaustive: regarding $\pi\,/\,(1-\pi) = P(\text{mated})\,/\,P(\text{nonmated})$, a value for π can be found, not necessarily equaling $P(\text{mated})$. On full prior uncertainty: $\pi = 0.5$.

[6] PAV LLRs define s_LLR via Bayes' rule as $\operatorname{sigmoid}(s_\mathrm{LLR} + \operatorname{logit}\pi)$ [4], i.e., as the posterior ratio, which is compared to $\frac{C_\mathrm{I}}{C_\mathrm{II}}$ for decision making.



(a) (b)

Fig. 3. Operating points as linear combination in y-inverted ROC space: Deriving DCF values by dropping perpendiculars onto the φ-depending line. (a) Outline of geometric proof with $\overline{OD} = \mathrm{DCF}(S \mid \varphi)$, $\sin\alpha = \frac{p_{\mathrm{II}}(\eta)}{c}$, $\frac{\overline{OD}}{c} = \cos(\alpha - \varphi)$, $\cos\alpha = \frac{p_{\mathrm{I}}(\eta)}{c}$. (b) Deriving minDCF as the ROCCH tangent on an exemplary system with depicted calibration loss to the φ-corresponding DCF value, cf., [15].

visualizes its minimum decision risk for all $\tilde{\pi}$ associated operating points. Thus, before computing similarity scores, an operating point can be denoted in the y-inverted ROC space by a line with $\tilde{\pi}$ depending slope, i.e., we propose to denote the angular operating point φ

$$\tan\varphi = \frac{C_{\mathrm{II}}}{C_{\mathrm{I}}}\,\frac{\pi}{1-\pi} = \frac{\sin\varphi}{\cos\varphi} = e^{-\eta}$$

$$\mathrm{DCF}(S \mid \varphi) = \sin(\varphi)\, p_{\mathrm{II}}(\eta) + \cos(\varphi)\, p_{\mathrm{I}}(\eta)$$

$$\text{note}\quad \varphi = \cot^{-1}(\tilde{\pi}^{-1} - 1), \quad \tilde{\pi} = (1 + \cot(\varphi))^{-1}. \qquad (5)$$

Fig. 3 provides the outline of a geometric proof, cf., [15], a proof via ROCCH and minDCF relations is in [7]. Smaller φ reflect more secure requirements as perpendiculars being tangent to its depending line, i.e., putting emphasis on ROCCHs of high type II errors. On poorly calibrated systems, actual threshold diverges from the threshold of minimum risk, i.e., the calibration loss is represented by the distance between the φ perpendiculars of the actual threshold to the ROCCH tangent. In this letter, we put emphasis on minDCFs, since calibration performance is not assessing discrimination.

III. Verbal Scales: Making LR Scores Digestible

In order to associate a human-interpretable meaning to LRs, verbal LR scales were developed over the past decades in forensics, mapping bands of LR values to verbal interpretation in terms of support for either prosecution or defendant hypotheses. Early verbal scales put emphasis on LRs < 100 [16], until the forensic field considered LRs for DNA. In 2015, the European Network of Forensic Science Institutes (ENFSI) [17] recommended the verbal scale suggested by the association of forensic service providers [18]. Nordgaard *et al.* [19] proposed to interpolate verbal bands concerning two fix points, i.e., $\mathrm{LR} = 100$ and $\mathrm{LR} = 10^6$, such that the increase in the base 10 logarithm of two consecutive interval limits of the verbal bands is proportional to the next band's. In other words, the interpolation is conducted from log-LR perspective, which is more suitable than within the LR-domain due to its linear symmetry regarding the depending scales of conclusion, where base 10 is considered for human-emphasized assessment. Table I compares both verbal scales on LRs ≥ 1, which favor the *mated* hypothesis (i.e., prosecution). Verbal scales for LRs favoring the *non-mated* hypothesis (i.e., defense) with values < 1 are symmetric regarding $\frac{1}{\mathrm{LR}}$.

TABLE I
Verbal Scales for Communicating LR Values, cf., [16], [19], Scales Depicted for LRs ≥ 1

LR	verbal
$\leq 10^1$	weak / limited
$\leq 10^2$	moderate
$\leq 10^3$	moderately strong
$\leq 10^4$	strong
$\leq 10^6$	very strong
$> 10^6$	extremely strong

(a) ENFSI guideline [17, 18].

LR	verbal
≤ 5.625	($\pm$0) neither / nor
≤ 100	(+1) some extent
≤ 5625	(+2) support
$\leq 10^6$	(+3) strong
$> 10^6$	(+4) extremely strong

(b) Scale of conclusion [19] with non-approximated LR values.

IV. Proposed Augmentation to DET Plots

In order to visualize the intercorrelation of error-rates and cross-application discrimination performance, we propose to utilize the verbal scales, e.g., the scale of conclusion. Since at most one minDCF point can lie on the φ depending line due to the ROCCH's convexity, we propose to color-encode levels of security and convenience on the ROCCH.

A. Verbal Bands in Performance Visualization

When associating verbal scales to LLR value of (2), verbal bands are also put in context to Bayes operating points $(\eta, (\pi, C_{\mathrm{I}}, C_{\mathrm{II}}), \tilde{\pi}, \varphi)$, i.e., considering $\mathrm{LLR} = \eta$ leads to favor the *mated* hypothesis at minimal cost advantage. Thus, the limits of verbal bands can be depicted by utilizing (5) in terms of the depending LR limits[7] $\mathrm{LR}_{-4,\ldots,\pm 0,\ldots,+4}$

$$\varphi = \tan^{-1}\left((\mathrm{LR}_{-4,\ldots,\pm 0,\ldots,+4})^{-1} \right). \qquad (6)$$

Fig. 4 depicts verbal scales in y-inverted ROC space, y-inverted, log-compressed ROC space, and in a security-emphasized DET space as well as an example on single system analysis. Levels of decision policy requirements can be depicted, when aggregating applications by verbal scales.

B. Verbal Bands: Representative Operating Points

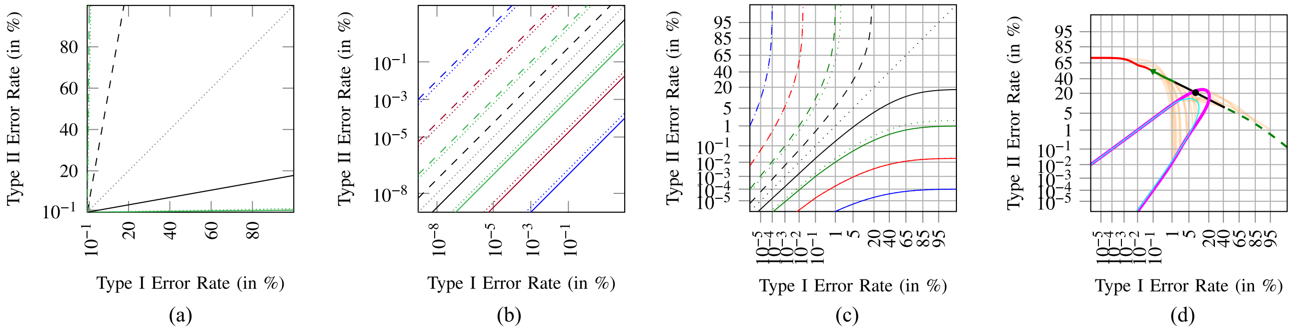
Verbal bands represent a range of operating points, however for the purpose of deriving an application threshold based on DET, one may want to start from an operating point representing a verbal band, and to proceed with a fine-tuning of the decision policy parameters $(\pi, C_{\mathrm{I}}, C_{\mathrm{II}})$. Therefore, we propose to seek the center of gravity of costs depending on verbal bands. Since DCFs are dependent on $\tilde{\pi}$, an application-independent cost measure is necessary, since different DCF setups are compared. Thus, we utilize C_{llr} [20]

$$C_{\mathrm{llr}}(S) = \int_0^1 \mathrm{DCF}(S \mid \tilde{\pi})\, \mathrm{d}\tilde{\pi} \qquad (7)$$

$$= \frac{1}{2\,\log(2)} \left(\sum_{g \in S_G} \frac{\log(1 + e^{-g})}{|S_G|} + \sum_{i \in S_{\mathrm{I}}} \frac{\log(1 + e^{i})}{|S_{\mathrm{I}}|} \right)$$

with the sets of mated scores and nonmated scores S_G, S_{I}, where $S = S_G \cup S_{\mathrm{I}}$. We propose to examine the ratio of S_G, S_{I}

[7] In this example, we refer to the [19] scale of conclusion, since bounds are denoted from the LLR-domain and the amount of bands is more limited to fewer categories, such that B2B decisions become easier to make.



(a) (b) (c) (d)

Fig. 4. Depicting the intercorrelation of error-rates and discrimination performance by φ-depending DCF slopes, solid lines indicating security levels $(+1, +2, +3, +4)$, dashed lines indicating convenience levels $(-1, -2, -3, -4)$. Blue, red, green, and black lines indicating $\pm 4, \pm 3, \pm 2, \pm 1$ levels, respectively. Representative operating points are indicated by dotted lines (EER-line at ± 0 level). From left to right: in y-inverted ROC space (cf., Fig. 3), with log-compression, in the DET space, in an analytic example of the proposed augmentation with simulated scores. Gray grid lines are provided for (c,d), i.e., in DET space. Contrastively to (a – c), (d) depicts the ROCCH with segments colored reflecting minDCF properties of Fig. 3 persisting the line style of (a – c). Markers indicate centers of gravity. In (d), minDCF and DCF values are depicted regarding φ as cyan and pink lines in terms of Fig. 3, respectively. Relations to minDCF points on the ROCCH are indicated by orange lines. Note: the integral of the pink line is closely related to the C_{llr} performance estimate. (a) y-inverted ROC space. (b) y-inv., log-compr. ROC space. (c) DET space. (d) $\mathcal{N}(3, 2)$ versus $\mathcal{N}(0, 1)$ example.

TABLE II
CENTERS OF $C_{\text{llr}}^{\text{ratio}}(\eta)$ GRAVITY ON THE [19] SCALE

Verbal scale	− 3	− 2	− 1	±0	+1	+2	+3
Application		Convenience				Security	
$\eta_{\min}$	− 13.82	− 8.63	− 4.61	− 1.73	1.73	4.61	8.63
η_{center}	− 13.18	− 8.03	− 4.07	0	4.07	8.03	13.18
$\eta_{\max}$	− 8.63	− 4.61	− 1.73	1.73	4.61	8.63	13.82

depending C_{llr} terms, symmetrically emphasizing security and convenience scenarios, i.e., the integral of $C_{\text{llr}}^{\text{ratio}}$

$$C_{\text{llr}}^{\text{ratio}}(\eta) = \frac{\log(1 + e^{a\,\eta})}{\log(1 + e^{-a\,\eta})} \text{ with } a = \text{sign}(\eta). \qquad (8)$$

For $\eta = 0$, i.e., on equal costs under full uncertainty ($C_{\text{I}} = C_{\text{II}} = 1, \pi = 0.5$), the EER line is resembled, cf., Fig. 4. By reaching towards higher levels of security or convenience, the centers visually collapse towards the outer limits of the depending verbal band, cf., Table II and Fig. 4.

Decision policy parameters can be fine-tuned in terms of $\eta \pm \delta$ utilizing (2) regarding threshold offsets, by denoting $C_{\text{II}} = 1$, with δ as

$$\delta = \log \frac{C'_{\text{I}}}{C_{\text{I}}} + \operatorname{logit} \pi + \log\left(\frac{1}{\pi} - \frac{\pi'}{\pi}\right) - \log \frac{\pi'}{\pi}. \qquad (9)$$

In other words, δ is denoted with respect to relative changes in C_{I} as $\frac{C'_{\text{I}}}{C_{\text{I}}}$ and in π as $\frac{\pi'}{\pi}$, where $0 < C'_{\text{I}}$, and $0 < \pi' < 1$.

For the purpose of deriving operating points verbally, vendors, operators, and owners of systems can first discuss on the application type as $-3, \ldots, +3$, second agree on a range of considerable priors,[8] then derive dependent costs, cf., (4), and third, adjust the threshold depending on the representative operating point by utilizing (9), when considering $C_{\text{II}} = 1$, e.g., the C_{I} cost proposed by the representative operating point may vary in a $(-10\%, +15\%)$ band, or need to be downscaled to a distinct C'_{I}. By adjusting thresholds, other verbal bands can be reached, e.g., a threshold of scale $+2$ increasing to scale $+3$.

V. Discussion and Conclusion

The presented work provides a recipe towards cross-application decision risk assessment for biometric researchers, hence DET plots are emphasized. This letter proposes to depict aggregated levels of decision risk on ROCCHs with respect to verbal scales, assuming optimal calibration.[9] Whereas, calibration performance is conventionally depicted by applied probability of error and empirical cross-entropy plots [7], [8].

Using the relationship between PAV and ROCCH, the proposed augmentation to DET plots increases transparency and motivates to reflect resulting PAV groups, e.g., due to a system or postscore binning, regarding minDCF in order to 1) yield ROCCHs not collapsing into a few supporting points, and thus to 2) support a wider range of application requirements. The concept of depicting minDCF or DCF values in the DET space is not new, cf., application-independent evaluation methods [21]. However, we contribute a novel scheme for depicting ranges of minDCFs, which are aggregated by similarity in terms of verbal scales, suitable for comparing a few systems of interest.

As a result of this letter, error tradeoffs can be categorized into levels of security and convenience. A clear distinction between scales aid the reflection of depending changes in decision policy parameters. We depicted the intercorrelation of error rates and cross-application decision risk with respect to minDCF, i.e., solely in terms of estimates for discrimination power. Furthermore, we introduced representative operating points per band of verbal scale, alongside a scheme for verbally conducting the setup of operative thresholds in B2B communication. For forensic purposes, the ENFSI verbal scale may be utilized instead. Finally, we provide a public available reference implementation.[10]

Acknowledgment

The authors would like to thanks to E. Tabassi for discussions.

[8] One may interpret $1 - \pi$ as the prior *"attack probability"* to a system.

[9] As indicated in Fig. 4, calibration performance can be depicted as well.

[10] Online available: https://codeocean.com/algorithm/154591c8-9d3f-47eb-b656-3aff245fd5c1/metadata.

References

[1] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19795-1:2017. Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework*, International Organization for Standardization and International Electrotechnical Committee, Geneva, Switzerland, Mar. 2017.

[2] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC FDIS 30107-3. Information Technology—Biometric Presentation Attack Detection—Part 3: Testing and Reporting*, International Organization for Standardization, Geneva, Switzerland, 2017.

[3] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, 1997, pp. 1895–1898.

[4] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," AGNITIO Research, South Africa, Tech. Rep., Dec. 2011, Accessed on: May 15, 2017. [Online]. Available: https://sites.google.com/site/bosaristoolkit

[5] A. Larcher, K. Lee, and S. Meignier, "An extensible speaker identification SIDEKIT in python," in *Proc. Int. Conf. Audio Speech Signal Process.*, 2016, pp. 5095–5099, Accessed on: 2017-05-15. [Online]. Available: http://lium.univ-lemans.fr/sidekit

[6] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 2382-37:2017 Information Technology—Vocabulary—Part 37: Biometrics*, International Organization for Standardization, Geneva, Switzerland, 2017.

[7] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Dept. Elect. Electron. Eng., University of Stellenbosch, Stellenbosch, South Africa, 2010.

[8] D. Ramos and J. Gonzalez-Rodrigues, "Cross-entropy analysis of the information in forensic speaker recognition," in *Proc. IEEE Odyssey, Speaker Lang. Recognit. Workshop*, 2008.

[9] R. Haraksim, D. Ramos, D. Meuwly, and C. E. H. Berger, "Measuring coherence of computer-assisted likelihood ration methods," *Forensic Sci. J.*, vol. 249, pp. 123–132, Apr. 2015.

[10] D. Ramos, R. Haraksim, and D. Meuwly, "Likelihood ratio data to report the validation of a forensic fingerprint evaluation method," *Data Brief*, vol. 10, pp. 75–92, Feb. 2017.

[11] D. Meuwly, D. Ramos, and R. Haraksim, "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation," *Forensic Sci. Int.*, vol. 276, pp. 142–153, Jul. 2017.

[12] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 694–699.

[13] M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman, "An empirical distribution function for sampling with incomplete information," *Ann. Math. Statist.*, vol. 26, no. 4, pp. 641–647, 1955, Accessed on: May 16, 2017. [Online]. Available: https://projecteuclid.org/euclid.aoms/1177728423

[14] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, pp. 861–874, 2006.

[15] N. Brümmer, "Optimization of the accuracy and calibration of binary and multiclass pattern recognizers, for wide ranges of applications," Feb. 2008, Accessed on: May 17, 2017. [Online]. Available: http://arantxa.ii.uam.es/ jms/seminarios_doctorado/abstracts2007-2008/20070226NBrummer.html

[16] D. Kaye, "The weight of evidence in law, statistics, and forensic science," in *Proc. NIST TC Quantifying Weight Forensic Evidence*, 2016, Accessed on: May 22, 2017. [Online]. Available: https://www.nist.gov/sites/default/files/documents/2016/12/07/03_kaye_16-nist-woe-linear.pdf

[17] S. E. Willis *et al.*, *ENFSI Guideline for Evaluative Reporting in Forensic Science*, European Network of Forensic Science Institutes, Wiesbaden, Germany, Mar. 2015, Accessed on: May 22, 2017. [Online]. Available: http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

[18] Association of Forensic Science Providers, "Standards for the formulation of evaluative forensic science expert opinion," *Sci. Justice*, vol. 49, no. 3, pp. 161–164, Sep. 2009, Accessed on: May 22, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.scijus.2009.07.004

[19] A. Nordgaard, R. Ansell, W. Drotz, and L. Jaeger, "Scale of conclusions for the value of evidence," *Law, Probab. Risk*, vol. 11, no. 1, pp. 1–24, 2012, Accessed on: May 22, 2017. [Online]. Available: https://academic.oup.com/lpr/article-lookup/doi/10.1093/lpr/mgr020

[20] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Comput. Speech Lang.*, vol. 20, no. 2, pp. 230–275, Jul. 2008.

[21] D. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I: Fundamentals, Features, and Methods* (Lecture Notes in Computer Science, vol. 4343). Berlin, Germany: Springer, 2007, pp. 330–353, Accessed on: June 20, 2017.