t-EER: Parameter-Free Tandem Evaluation of Countermeasures and Biometric Comparators

Tomi H. Kinnunen, Kong Aik Lee, *Senior Member, IEEE,* Hemlata Tak,
Nicholas Evans, *Member, IEEE,* Andreas Nautsch

Abstract—Presentation attack (spoofing) detection (PAD) typically operates alongside biometric verification to improve reliablity in the face of spoofing attacks. Even though the two sub-systems operate in tandem to solve the single task of reliable biometric verification, they address different detection tasks and are hence typically evaluated separately. Evidence shows that this approach is suboptimal. We introduce a new metric for the joint evaluation of PAD solutions operating in situ with biometric verification. In contrast to the tandem detection cost function proposed recently, the new tandem equal error rate (t-EER) is parameter free. The combination of two classifiers nonetheless leads to a *set* of operating points at which false alarm and miss rates are equal and also dependent upon the prevalence of attacks. We therefore introduce the *concurrent* t-EER, a unique operating point which is invariable to the prevalence of attacks. Using both modality (and even application) agnostic simulated scores, as well as real scores for a voice biometrics application, we demonstrate application of the t-EER to a wide range of biometric system evaluations under attack. The proposed approach is a strong candidate metric for the tandem evaluation of PAD systems and biometric comparators.

Index Terms—Biometrics, presentation attack detection, tandem evaluation, equal error rate, automatic speaker verification

✦

1 INTRODUCTION

B IOMETRIC recognition is nowadays in widespread use across forensic, civilian and consumer domains. Likewise, approaches for the testing and performance reporting of biometric systems [1], [2] under *normal presentation mode*[1] are well established — see e.g. the ISO/IEC 19795 standard [4]. Despite high reliability, biometric systems are unfortunately not infallible outside of the normal presentation mode, e.g. when they are attacked by an adversary. Since the identification of biometric system attack points more than two decades ago [5], the community has been active in addressing vulnerabilities, especially *presentation* or *spoofing attacks,* for all the major biometric modes such as fingerprints [6], face [7], and voice [8], [9]. Well-studied examples of spoofing attacks include printed photographs (face), gummy fingers (fingerprints), and audio replay (voice). Of particular concern are vulnerabilities to *DeepFakes* [10], stemming from rapid developments in deep learning [11], which can be used to implement face swapping and voice cloning.

Having recognized the threat, the ISO/IEC Joint Technical Committee (JTC) 1/SC 37 launched a multi-part 30107

standard series [3] in 2016 to provide a foundation for what became known as *presentation attack detection* (PAD), with a second version being released in 2023 [12]. As defined by the standard, attacks are assumed to be *presented* at the sensor level of the targeted biometric data capture subsystem (such as a fingerprint reader, a camera, or a microphone).[2] PAD has since attracted considerable interest; attacks have potential to interfere with the intended operation of the biometric system and to substantially compromise performance unless prepared for.

PAD is today an established research topic within the biometrics research community and one that is still attracting growing attention. While the practical solutions vary from challenge-response approaches [7] to liveness detection [13], quality assessment [6], and artifact detection [14] specific to each biometric mode, most take the form of binary classifiers whose role is to determine whether or not a given biometric probe is a bona fide sample (real) or a presentation attack (fake).
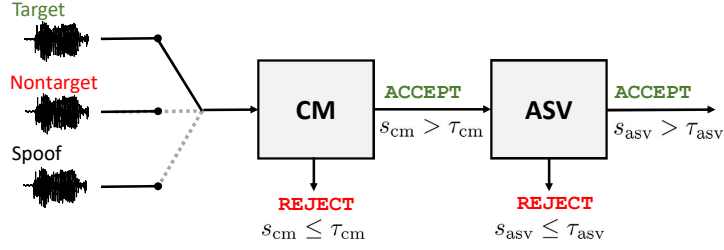
Whatever the approach, the PAD problem has been treated historically as a separate task that is independent of the key biometric comparison task. The latter consists in verifying whether or not a pair of biometric samples in the form of reference (enrolment) and probe (test) samples originate from the same biometric capture subject (individual).

- *T. H. Kinnunen is with the School of Computing, University of Eastern Finland (UEF), Finland, e-mail: tomi.kinnunen@uef.fi.*

- *K. A. Lee is with the Institute for Infocomm Research, A*STAR, Singapore, e-mail: lee_kong_aik@i2r.a-star.edu.sg.*

- *H. Tak and N. Evans are with EURECOM, France, e-mail: {tak,evans}@eurecom.fr.*

- *A. Nautsch is with Avignon Université, France (parts of this work were conceived while he was at EURECOM).*

1. *Normal* or *routine* implies that the system is used in the fashion intended by the system designer [3]. Spoofed trials are considered to be outside of the normal presentation mode.

2. In the authors' view, this definition excludes certain forms of attack that are relevant to specific biometric modes, particularly when the sensor is beyond the control of the biometric system designer. In the particular case of speaker verification technology deployed in many call centers, users choose their *own* biometric capture subsystem — say their smartphone. Thus, besides presentation attacks directed at the microphone, an adversary may also inject attacks *post-sensor* using, e.g. a voice modification system operating somewhere between the microphone and the call center. The ASVspoof initiative [9] differentiates the two cases, termed as *physical attacks* (presented at the sensor) and *logical attacks* (injected post-sensor).

Tandem miss:	Rejection of target (by either subsystem).
Tandem false alarm:	Acceptance (by both subsystems) of either nontarget or spoof.
t-EER path	All pairs $(\tau_{\mathrm{asv}}, \tau_{\mathrm{cm}})$ s.t. Tandem miss rate = Tandem false alarm rate.
Concurrent t-EER:	Unique t-EER path point s.t. target miss = nontarget FA = spoof FA.

(a) A tandem detection system and key definitions.
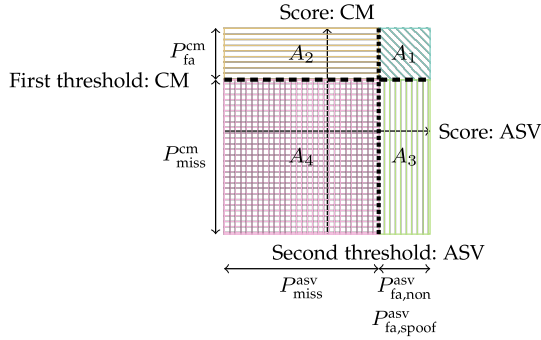
(b) The 2D space of the detection scores $(s_{\mathrm{asv}}, s_{\mathrm{cm}})$ and thresholds $\boldsymbol{\tau} = (\tau_{\mathrm{asv}}, \tau_{\mathrm{cm}})$.
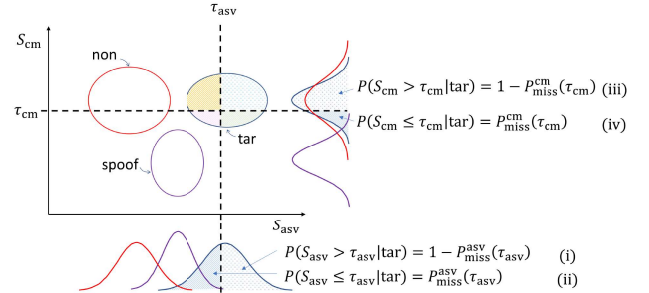
(c) Similar 2D score space as in (b), with computation of per-system error probabilities visualized.

Fig. 1: The *tandem equal error rate* (t-EER) is a metric for the evaluation of biometric systems comprising a cascaded spoofing countermeasure (CM) and comparator such as an automatic speaker verification (ASV) system, illustrated in **(a)**. Extending the conventional EER, the t-EER is found by adjusting the *two* thresholds **(b)** to equalise total miss and false alarm rates. This leads to a *function* with multiple possible t-EERs, summarized using a unique scalar, the *concurrent* t-EER. As illustrated in **(b)**, in the tandem score space, we compose two linear decision boundaries for scores of the two subsystems. In **(c)**, the miss rate of the target class corresponds to the shaded area, which includes trials with both ASV and CM scores fall below the two thresholds, τ_{asv} and τ_{cm}. The marginal distributions shown are those of ASV system (below) and the CM (right).

.

As illustrated in Fig. 1a, practical biometric authentication solutions typically combine separate PAD and biometric comparator sub-systems operating in tandem.

1.1 The Need for Tandem Assessment

Both sub-systems are binary classifiers, each designed to solve a given, different task. Nonetheless, they are combined to solve what is, after all, the *single* task of reliable biometric verification. **We argue that, no matter what the architecture used to combine them, the two sub-systems should be evaluated as a whole**. As noted in [15], however, the PAD subsystem and the biometric comparator are typically evaluated separately.

We suspect there are two main reasons for this. The first is one of convenience which follows the well-known *Maslow's hammer* law, *'if the only tool you have is a hammer, it is tempting to treat everything as if it were a nail'*; since both the PAD subsystem and biometric comparator are binary classifiers, any established performance assessment tool and

figure of merit [2], [16] are readily applicable.[3] While the separate assessment of the two subsystems is useful for analysis purposes, it effectively forces a *ternary* classification task (bonafide target, nontarget, spoofing attack) to be treated as two separate binary tasks [17]. This approach not only *increases* the amount of numbers to be reported (e.g. a pair of EERs corresponding to the PAD subsystem and the biometric comparator) but, worse, provides no guarantees on the performance of the complete biometric verification system as a whole. Previous work [18] confirms that the best-performing PAD subsystem as judged from separate approaches to assessment does not always offer the best protection to the biometric comparator. The separate assessment of the two subsystems is hence sub-optimal.

This leads us to the second suspected reason — even if one might be convinced that the two systems should be

---

3. With the understanding that the positive and negative classes and error measures could be termed differently for clarity; for instance, in biometric verification the proportion of negative class cases that were falsely accepted is known as *false acceptance rate* (FAR), *false match rate* (FMR), or *false alarm rate* depending on the context. In in ISO/IEC 30107 [3], the equivalent property of a PAD subsystem is known as *attack presentation classification error rate* (APCER).

evaluated jointly, there are no commonly agreed guidelines on *how* this should be done. While score- and decision-level combination [19], [20], [21], [22] are popular choices, they can be implemented using various different approaches, with possible tradeoffs [23] in the treatment of bonafide users and spoofing attacks. In fact, even the ISO/IEC 30107 standard leaves the details of the PAD system integration to the general biometric framework open-ended, by noting that the PAD subsystem could be placed e.g. before or after the biometric comparator subsystem, and/or at different levels. The standard explicitly acknowledges that *"Outside the scope are ... overall system-level security or vulnerability assessment"*. While one can expect carefully-tailored solutions for a given biometric mode, the fusion of multiple biometric modes, or the fusion of multiple PAD subsystems [20] to perform well, such solutions tend to be complex and specific to the selected application. **There is clear scope for a common reference approach to the tandem assessment of arbitrary biometric comparator and PAD subsystems.**

1.2 Our Novelty: the Tandem Equal Error Rate (t-EER)

The authors recently proposed a new decision-theoretic framework, namely the *tandem detection cost function* (t-DCF) [18], as a solution to the performance testing of cascaded biometric comparator and PAD subsystems as illustrated in Fig. 1a. The t-DCF was developed as a two-system, three-class generalization of the widely-adopted *detection cost function* (DCF) [16] endorsed by the National Institute of Standards and Technology (NIST). It was adopted as the primary evaluation metric for the two most recent ASVspoof challenge editions [9], and has now been broadly adopted by the voice PAD and biometrics community, even if not embraced by other biometrics communities. One detraction of both the DCF and the t-DCF is that they require the choosing of two sets of parameters — detection costs and class priors — which can be difficult in practice, or feel arbitrary. While the authors generally advocate the adoption of performance measures that make the uncertainty of class prevalence (priors) and implications of decisions (costs) explicit, we recognise the appeal of parameter-free, more easily interpretable metrics. Remaining convinced in the importance of tandem evaluation, we have hence set out to derive an alternative to the t-DCF that is independent of costs and priors.

Our starting point is the traditional *equal error rate* (EER), a popular metric used for the assessment of any ordinary binary classifier. The EER is defined as the error rate at a particular, special operating point at which the false alarm and miss rates are equal. It is also independent of priors, cost parameters and requires no *a priori* setting of detection thresholds [18]. Despite acknowledged shortcomings [3], [4], the EER remains among the most widely-adopted metrics in biometrics research and is used broadly for optimisation and model selection. Since the EER also serves as an upper bound to the Bayes error rate [24], it is often also used as an empirical target in preliminary investigations of new features or model architectures.

We report in this paper an extension to the conventional EER which supports the evaluation of a pair of binary classifiers operating in tandem. The new evaluation methodology is broadly applicable to any such scenario, even those

beyond the study of biometrics. The specific application domain upon which we focus upon in this paper involves the study and evaluation of combined biometric comparators and PAD subsystems. The proposed common reference approach to their tandem evaluation remains agnostic to any specific biometric (e.g. face, iris and fingerprint). To highlight its broad application to both non-biometric and biometric scenarios alike, we report experiments with sets of simulated scores. Nonetheless, in order to demonstrate an application of the new evaluation methodology in a real, practical scenario, we also report experimental work involving voice biometrics.

We summarize the organization, as well as key novelties of the paper, as follows.

- After setting out the detection-theoretic foundations for the evaluation of classical standalone classifiers in Section 2, Section 3 provides a self-contained formulation of the tandem detection task;
- The tandem equal error rate (t-EER), and the extension of a classical operating *point* to an operating *path*, as detailed in Section 4, are new concepts not presented in any prior work. As we will see, the t-EER can be viewed as a *unifying* metric from which the classical per-system (PAD and biometric comparator) EERs are obtained as special cases;
- Section 5 describes a practical and efficient approach to compute the t-EER path. For practitioners, we provide an open-source reference implementation;
- Unlike the classical scalar-valued EER, the new t-EER represents a *function* with the t-EER path as its domain. Section 6 shows how to summarize that path as a unique scalar in an intuitive, meaningful way. The result, coined the *concurrent t-EER*, corresponds to a unique pair of thresholds at which the tandem system miss rate, the false alarm rate for nontargets, and the false alarm rate for spoofing attacks are all equal. The proposed metric hence reflects equal treatment of all three classes.
- Section 7 demonstrates use of the proposed metric in the context of voice biometrics. Finally, conclusions are presented in Section 8.

2 PRELIMINARIES

The proposed metric evaluates a pair of PAD and biometric comparator as illustrated in Fig. 1a. Before detailing the combination of the two systems in Section 3, we first introduce basic notation and describe the evaluation of single detectors as a starting point. For a broader introduction to the evaluation of biometric systems, the interested reader is pointed to [1], [2]. Evaluation of both biometric verification [1], [2] and PAD systems [17] is rooted in detection theory (e.g. [25], [26]).

2.1 Detectors

Let (x, y) denote paired input $x \in \mathcal{X}$ (an element of some signal or feature space $\mathcal{X}$) and the corresponding ground-truth label $y \in \mathcal{Y} \triangleq \{0, 1\}$. A *binary classifier*, or a *detector*,

is a parameterized[4] function $g_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ which makes a prediction, $\hat{y}$, of the label of x. It does so in two steps. First, the detector computes a real-valued *score*, $s(x) \triangleq s \in \mathbb{R}$ with the convention that higher numerical values are associated with greater confidence for x to originate from the positive class ($y = 1$). Typically s is a likelihood ratio, log-likelihood ratio, or class posterior produced by a model (such as a deep neural network). In the second stage, the predicted label $\hat{y}$ (that may differ from the ground-truth label y) is obtained by comparing the score to a preset threshold, τ:

$$\hat{y} = \begin{cases} 1, & \text{if } s > \tau \\ 0, & \text{if } s \leq \tau. \end{cases} \tag{1}$$

In biometric verification, x consists of a paired representation $x = (x_{\mathrm{e}}, x_{\mathrm{t}})$ of enrollment (x_{e}) and test (x_{t}) samples, also known as the *biometric reference* and *biometric probe*, respectively [27]. Depending on the system, these samples could be represented by raw signal/pixel values or feature vectors computed from them. The ground truth label y, in turn, serves to indicate whether or not the enrolment and test samples belong to the same individual (*a target*) or not (*a nontarget*). The ISO/IEC 2382-37 standard [27] refers instead to *biometric mated/non-mated comparison trials*. For the second task considered in this study, PAD, x is (usually) a representation of the test sample alone, whereas y serves to indicate if x is real (*bona fide* human) or a fake (*spoofed*) sample. The corresponding terminologies in the ISO/IEC 30107 standard [3] are *bona fide / attack presentations*.

The approach to evaluation and the metric reported in this paper are agnostic to the semantics of x and the classifier $g_{\boldsymbol{\theta}}$. The only data available to us, as evaluators, are the finite set of detection scores and the ground-truth labels, $\{(s_i, y_i) : i = 1, \ldots, N\}$ for a test set of N samples. From here on we drop the instance subscript i and treat both the scores and their labels as realizations of random variables S and Y, respectively. To this end, we use $F_{S|Y}(\tau|Y) \triangleq P(S \leq \tau|Y)$ to denote the cumulative density function (CDF) of scores for class Y. Here, $P(A|B)$ is the standard notation of conditional probability — the probability of event A given that event B has occured.

2.2 Detection error rates

Any binary classifier is prone to two different types of detection errors — *misses* (false rejections) and *false alarms* (false acceptances). The former take place when the true class of x is 1 but the classifier predicts 0 (rejects x). The latter happen when the true class is 0 but the classifier predicts 1 (accepts x). These two error rates arise naturally from the *total error rate*, $P_{\mathscr{E}}$, defined as the probability that the predicted label differs from the true label. Following the rules of joint and conditional probabilities [28],

$$\begin{aligned} P_{\mathscr{E}} &\triangleq P(\hat{Y} \neq Y) \\ &= P(\hat{Y}{=}0, Y{=}1) + P(\hat{Y}{=}1, Y{=}0) \\ &= P(Y{=}1)P(\hat{Y}{=}0|Y{=}1) + P(Y{=}0)P(\hat{Y}{=}1|Y{=}0) \\ &= \pi P_{\mathrm{miss}} + (1 - \pi)P_{\mathrm{fa}}, \end{aligned} \tag{2}$$

4. Here, $\boldsymbol{\theta} \in \Theta$ denotes model parameters, for instance, the set of weights and biases of a neural network.

where

$$\begin{aligned} P_{\mathrm{miss}} &\triangleq P(\hat{Y} = 0|Y = 1) \\ P_{\mathrm{fa}} &\triangleq P(\hat{Y} = 1|Y = 0) \end{aligned} \tag{3}$$

are the miss and false alarm rates and

$$\pi \triangleq P(Y = 1) \tag{4}$$

denote a *database prior*. As the name suggests, it reflects the relative proportion of positive and negative class samples in the database protocol, known by the evaluator. The database prior is *not* to be confused with a user-supplied *asserted* prior one might have about future data in a predictive context; whereas the database prior is a deterministic property of a database protocol, an asserted prior reflects one's uncertainty of class priors in a future environment. In contrast to the DCF [16] and the t-DCF [18] metrics that require asserted priors, the concurrent t-EER metric proposed in this study has no dependencies, neither database nor asserted priors. Note that if the database contains an equal proportion of trials from the two classes, the database prior is $\pi = \frac{1}{2}$. If the total error rate were to be used as a performance measure, the relative class proportions impose implicit weighting of the two error types given by the convex combination in (2). Such dependency on database class proportions is undesirable, not least because the error rates are noncomparable across datasets with different class proportions.

By combining (1) with (3), the miss rate can be expressed as a function of τ, as

$$\begin{aligned} P_{\mathrm{miss}}(\tau) &= P(\hat{Y} = 0|Y = 1) \\ &= P(S \leq \tau|Y = 1) \\ &= F_{S|Y=1}(\tau|Y = 1). \end{aligned} \tag{5}$$

Similarly, the false alarm rate can be expressed as

$$\begin{aligned} P_{\mathrm{fa}}(\tau) &= P(\hat{Y} = 1|Y = 0) \\ &= P(S > \tau|Y = 0) \\ &= 1 - F_{S|Y=0}(\tau|Y = 0). \end{aligned} \tag{6}$$

A central notion in *receiver operating characteristic* (ROC) [25], [26] and *detection error trade-off* (DET) [29] analysis is to observe the trade off between miss and false alarm rates by varying τ. In particular, the miss and the false alarm rates are, respectively, non-decreasing and non-increasing functions of τ,

$$\begin{aligned} P_{\mathrm{miss}}(\tau) &\geq P_{\mathrm{miss}}(\tau') \quad \text{whenever } \tau \geq \tau' \\ P_{\mathrm{fa}}(\tau) &\leq P_{\mathrm{fa}}(\tau') \quad \text{whenever } \tau \geq \tau', \end{aligned}$$

with the following limits which follow directly from the properties of the CDF:

$$\begin{aligned} \lim_{\tau \to -\infty} P_{\mathrm{miss}}(\tau) &= 0, & \lim_{\tau \to +\infty} P_{\mathrm{miss}}(\tau) &= 1, \\ \lim_{\tau \to -\infty} P_{\mathrm{fa}}(\tau) &= 1, & \lim_{\tau \to +\infty} P_{\mathrm{fa}}(\tau) &= 0. \end{aligned}$$

In the remainder we drop 'lim' and refer to these limits simply by setting $\tau = -\infty$ or $\tau = \infty$. In biometric verification and spoofing detection, τ can be used to control the trade-off between user convenience and security. The 'accept all' setting $\tau = -\infty$ leads to all inputs being accepted, giving $P_{\mathrm{miss}} = 0$ and $P_{\mathrm{fa}} = 1$. Likewise, $\tau = +\infty$ leads to a 'reject all' system, for which $P_{\mathrm{miss}} = 1$ and $P_{\mathrm{miss}} = 0$. While these extreme cases are referred to in the analyses which follow, for any practical detection system, $|\tau| < \infty$.

2.3 Error rate tradeoff and performance summarization

By varying τ, one obtains a ROC/DET curve which displays one error rate as a function of another. ROC curves [25], [30] are obtained by plotting $P_{\mathrm{fa}}(\tau)$ against $1 - P_{\mathrm{miss}}(\tau)$. DET curves [29], used by default in voice biometrics research (and also recommended for PAD evaluation [3]), graph $P_{\mathrm{fa}}(\tau)$ against $P_{\mathrm{miss}}(\tau)$ on a normal deviate designed to highlight better the differences between high-performing systems. The interested reader is pointed to [31, Section 4] for a review and additional approaches to plot error trade-offs.

For performance assessment, a single scalar is preferred over a function. Since no scalar can preserve all the details of the full error tradeoff function, numerous performance indices have been proposed to summarize the aspects relevant in a given domain (for reviews see e.g. [32], [33]). In biometric verification and PAD research papers, the EER (defined below) is among the most popular choice. Others include the *area under the curve* (AUC) [30], the F_1 score and the *cost of log-likelihood ratio* (C_{llr}) [34]. In contrast to these non-parametric indices, *weighted* error metrics include the *detection cost function* (DCF) [16] which measures the average (expected) cost of making an error. Beyond biometrics, readers are referred to [35] for an interesting measure which imposes a weight function (prior distribution) over the detection costs.

For any operational automated biometric verification system, one must set the detection threshold τ in advance to obtain a desirable trade-off between security and convenience as is normally set according to knowledge of the operational environment. This can be done by optimizing a criterion of interest (such as the EER or DCF) using development data or by fixing the threshold according to the analytic form provided by Bayes' minimum-risk threshold. Since the scores produced by a model — even by a probabilistic model — rarely present calibrated probabilities, one may combine this approach with a *score calibration* model [34], [36] trained using development data. Graphical presentations which emphasize this aspect of evaluation — improving the generality of threshold setting or score calibration from development to evaluation data — include the *expected performance curve* (EPC) [37] and the *applied probability of error* (APE) curve [34]. The EPC was further generalized in [17] to an *expected performance and spoofability curve* (EPS) for the evaluation of biometric systems under attack. Since our focus is on a *discriminative* criterion (EER) rather than calibration, threshold selection [5] and score calibration is outside the scope of this paper. As we will see, the generalisation of the EER to the evaluation of two-systems

and three classes is not straightforward. Before doing so, we present the definition of the classical EER.

2.4 Equal error rate (EER)

Noting the monotonicity and contrasting values of miss and false alarm rates at $\pm\infty$, there always exists an interval $\Omega \subset \mathbb{R}$ of thresholds in which the miss and the false alarm rates are equal[6]: $P_{\mathrm{miss}}(\tau) = P_{\mathrm{fa}}(\tau)$, for all $\tau \in \Omega$. Usually this interval collapses to a single point (a scalar). We define the *equal error rate* (EER) as the value $\mathrm{EER} \triangleq P_{\mathrm{miss}}(\tau) = P_{\mathrm{fa}}(\tau)$ where $\tau \in \Omega$ is the interval at which the miss and false alarm rates have the same value. Apart from some special cases, such as as Gaussian score PDFs with known parameters [39], [40], neither the EER nor the EER threshold have closed-form solutions, hence they are determined empirically from the operating point at which the empirical miss and false alarm functions intersect.

The EER is closely related to the total error rate; in fact, substituting $P_{\mathrm{miss}} = P_{\mathrm{fa}} = \mathrm{EER}$ into (2) shows they equal one another at the EER operating point. Further, as detailed in [24], under calibrated probabilities, the EER can be interpreted as a strict upper bound to the Bayes error rate:

$$\max_{\pi} P_{\mathcal{E}} = \mathrm{EER}. \tag{7}$$

Informally, the reader may verify (7) by differentiating (2) w.r.t. π which yields the EER condition $P_{\mathrm{miss}} = P_{\mathrm{fa}}$. Hence, when used as an empirical model selection criterion, the EER can be interpreted as the total error rate with the *worst-possible* prior.

3 TANDEM SYSTEMS

Consider now the case of a combined system comprising *two* binary classifiers operating in tandem (here, PAD and biometric comparator subsystems). Each classifier addresses a different detection task. We want to combine their predictions in a manner which supports their tandem evaluation. As above, each of these two classifiers has two potential prediction outcomes — one positive and one negative. The combined (tandem) system acts as a logical AND gate by producing an overall positive decision if (and only if) both classifiers produce positive predictions. In the remainder of this paper, we adopt special terminology used in the context of *automatic speaker verification* (ASV) — with the understanding that 'ASV' can be replaced by 'fingerprint', 'face', 'iris' verification (or any other biometric mode) without changing the methodology.

3.1 Database Prior

We now consider two sets of labels, $Y_{\mathrm{cm}} = \{\text{bona fide} = 1, \text{spoof} = 0\}$ (label set of a countermeasure) and $Y_{\mathrm{asv}} = \{\text{target} = 1, \text{nontarget} = 0\}$ (label set of an ASV system). We define the set of *tandem labels*, Y_{tdm}, as the cartesian

---

5. For *known* class-conditional feature distributions and class priors, the *optimal* detection score for a detection task is the log-likelihood ratio (LLR) of the null and alternative hypothesis likelihoods, leading to optimal (lowest error rate) decision threshold $\tau_{\mathrm{Bayes}} := -\mathrm{logit}(\pi)$, where π denotes the prior of the target (positive) class and $\mathrm{logit}(\pi) := \log(\pi) - \log(1 - \pi)$. In any practical setting, however, neither the class priors nor the class-conditional data distributions are known, but have to be estimated from data using a model. This results in *miscalibrated* scores that no longer produce optimal decisions derived using τ_{Bayes}. The EER, however, measures the degree of overlap between the score distributions of each class, ignoring completely the scale of the detection scores. Just as the entire ROC or DET curve, it is agnostic to any global order-preserving score transforms (such as the affine map $s \mapsto as + b$ with $a > 0$).

6. Technically, this is guaranteed only when the actual score CDFs are continuous. In practical scenarios, empirical CDFs are constructed from finite sets (typically by error counting) and there may not be an exact threshold where the two are equal. While in this study we rely on simple nearest-neigbor interpolation to determine the EER, the interested reader is pointed to [38] for more advanced methods based on the ROC convex hull [30, Section 6] to determine the EER.

product $Y_{\text{tdm}} = Y_{\text{cm}} \times Y_{\text{asv}}$ which consists of four possible elements. The relative frequency of each of the four cases is given by the joint probability distribution over the two sets of ground-truth labels. Using the shorthand notation $P(Y_{\text{cm}} = y_{\text{cm}}, Y_{\text{asv}} = y_{\text{asv}}) = P(y_{\text{cm}}, y_{\text{asv}})$ to drop random variable names, the joint probability distribution for the four possible tandem labels are given by

$$
\begin{aligned}
P(\text{bona}, \text{tar}) &= P(\text{bona}|\text{tar})P(\text{tar}) \\
P(\text{bona}, \text{non}) &= P(\text{bona}|\text{non})P(\text{non}) \\
P(\text{spoof}, \text{tar}) &= P(\text{tar}|\text{spoof})P(\text{spoof}) \\
P(\text{spoof}, \text{non}) &= P(\text{non}|\text{spoof})P(\text{spoof}),
\end{aligned} \tag{8}
$$

where class names are shortened for brevity. Given the specific tandem tasks of spoofing and speaker detection, we now assert three *necessary* assumptions relevant to the conditional probabilities in the right-hand side of (8):

(A1.) $P(\text{bona}|\text{tar}) = 1$ Target speaker is bona fide.
(A2.) $P(\text{bona}|\text{non}) = 1$ Zero-effort impostor is bona fide.
(A3.) $P(\text{tar}|\text{spoof}) = 1$ A spoofing attack is targeted.
$\Big($**(A4.)** $P(\text{non}|\text{spoof}) = 0$ *(follows from A3)*$\Big)$

The first assumption (A1) means that users who claim their own identity present only bona fide utterances; the second (A2) dictates that non-target or zero-effort impostor presentations contain bona fide speech (not spoofed speech). Finally, (A3) states that spoofed presentations are made to claim only a specific target (speaker) identity. These assumptions reflect a practical verification scenario in which spoofing attacks can be expected. Note that (A4) follows from (A3). Note also that while ISO/IEC 30107 [3] considers both subversive *biometric impostors* (spoofing) and *biometric concealers* (evasion), our focus is on the former.

By substituting assumptions (A1)–(A3) into (8) we arrive at

$$
\begin{aligned}
P(\text{bona}, \text{tar}) &= P(\text{tar}) &&=: \pi_{\text{tar}} \\
P(\text{bona}, \text{non}) &= P(\text{non}) &&=: \pi_{\text{non}} \\
P(\text{spoof}, \text{tar}) &= P(\text{spoof}) &&=: \pi_{\text{spoof}} \\
P(\text{spoof}, \text{non}) &= 0,
\end{aligned} \tag{9}
$$

where π_{tar}, π_{non} and π_{spoof} are nonnegative and sum up to 1. Because of these probabilistic constraints, the database prior $\boldsymbol{\pi} = (\pi_{\text{tar}}, \pi_{\text{non}}, \pi_{\text{spoof}})$ is now uniquely specified by two nonnegative numbers. The above prior definitions align with the notation in [18] though here we make the underlying assumptions explicit. From here on, we use 'target', 'nontarget' and 'spoof' to refer to the tandem classes (bona, tar), (bona, non) and (spoof, tar), respectively. In the ISO/IEC 30107 standard [3], the corresponding terminologies are *bona fide mated* (target), *bona fide non-mated* (nontarget), and *attack* (spoof).

3.2   Decision Rules

The tandem detection system combines information extracted by the two subsystems — CM and ASV — to make the final accept/reject decision. This means that, in order for a particular identity claim to be accepted, the provided test utterance must be declared as bona fide **and** it must match the claimed identity in terms of speaker characteristics.

Depending on the outputs provided by the individual detectors, the combined decision could be obtained, for instance, by `AND`ing the hard decisions by

$$
\begin{aligned}
D &= \text{AND}\big(\mathbb{I}(s_{\text{asv}} > \tau_{\text{asv}}), \mathbb{I}(s_{\text{cm}} > \tau_{\text{cm}})\big) \\
&= \mathbb{I}(s_{\text{asv}} > \tau_{\text{asv}}) \cdot \mathbb{I}(s_{\text{cm}} > \tau_{\text{cm}})
\end{aligned} \tag{10}
$$

where the indicator function $\mathbb{I}(\alpha) = 1$ for a true proposition α (and 0 otherwise), and $D \in \{0, 1\}$ stands for 'decision': $D = 0$ implies a negative prediction (reject) and $D = 1$ a positive prediction (accept). The combined decision could alternatively be produced by passing the detection scores to a back-end classifier followed by thresholding:

$$
D = \mathbb{I}(\texttt{fuse}([s_{\text{asv}}, s_{\text{cm}}]) > \tau_{\text{fuse}}), \tag{11}
$$

where $\texttt{fuse} : \mathbb{R}^2 \to \mathbb{R}$ denotes an arbitrary back-end classifier with an associated single detection threshold, τ_{fuse}.

The theory developed in this study focuses on decision rules of the form (10) where we assume access to two sets of detection scores and where the two detection thresholds τ_{asv} and τ_{cm} can vary concurrently. This leads to an extension of the single-system detection theory where the concept of a scalar threshold is extended to a vector-valued decision boundary in a 2D score space.

While (10) assumes statistical independence of the decisions produced by the two systems (see also [18, Section VIII.A]), we argue that this is reasonable whenever the systems address *different* detection tasks – as is the case here. Moreover, limiting our attention to 2-parameter decision rules specified by two orthogonal hyperplanes (see Fig. 1b) enables tractable analysis (within a reduced space of combination rules). Note that the second, score fusion strategy (11) involves all possible classification rules that map continuous 2D vector onto a scalar. This includes, for instance, affine score combination, Gaussian back-end [22] and arbitrary feedforward DNNs with softmax output.

Another benefit of (10) over the score fusion approach (11) is that we do not need *paired* CM and ASV scores. Due to the statistical independence assumption of ASV and CM decisions, the error rate expressions of the tandem system (detailed in Section 4) involves marginal (subsystem-specific) error rates and their products only. An important practical consequence is that the ASV and CM systems could be produced by two different parties that may not have the capacity or expertise to implement both. This property has been exploited in the organisation of the ASVspoof challenge series [9] for which an ASV system is implemented by the organizers, allowing participants to foucs upon the design of CMs.

3.3   Miss and False Alarm Rates

In summary, we assume three mutually exclusive classes (target, nontarget, spoof), and two possible decisions ($D = 0$ or $D = 1$) produced by the tandem system. Table 1 indicates which cases consitute a classification error.

TABLE 1: Cross-tabulation of tandem system decisions (rows) vs. ground-truth labels (columns) encodes which cases consitute a classification error.

	Target	Nontarget	Spoof
$D = 0$	error	OK	OK
$D = 1$	OK	error	error

Therefore, similarly to (2), the *tandem* total error, $P_{\mathscr{E}}^{\text{tdm}}$, is obtained by summing the probabilities of the three (mutually exclusive) errors:

$$\begin{aligned} P_{\mathscr{E}}^{\text{tdm}} =&P(D = 0, Y_{\text{tdm}} = \text{tar})+ \\ &P(D = 1, Y_{\text{tdm}} = \text{non})+ \\ &P(D = 1, Y_{\text{tdm}} = \text{spoof}) \\ =&P(D = 0|Y_{\text{tdm}} = \text{tar})P(Y_{\text{tdm}} = \text{tar})+ \quad (12) \\ &P(D = 1|Y_{\text{tdm}} = \text{non})P(Y_{\text{tdm}} = \text{non})+ \\ &P(D = 1|Y_{\text{tdm}} = \text{spoof})P(Y_{\text{tdm}} = \text{spoof}) \\ =&\pi_{\text{tar}}P_{\text{miss}}^{\text{tdm}} + \pi_{\text{non}}P_{\text{fa,non}}^{\text{tdm}} + \pi_{\text{spoof}}P_{\text{fa,spoof}}^{\text{tdm}}, \end{aligned}$$

where

$$\begin{aligned} P_{\text{miss}}^{\text{tdm}} &\triangleq P(D = 0|Y_{\text{tdm}} = \text{tar}) \\ P_{\text{fa,non}}^{\text{tdm}} &\triangleq P(D = 1|Y_{\text{tdm}} = \text{non}) \quad (13) \\ P_{\text{fa,spf}}^{\text{tdm}} &\triangleq P(D = 1|Y_{\text{tdm}} = \text{spoof}) \end{aligned}$$

are, respectively, the miss rate, the false alarm rate for nontargets and the false alarm rate for spoofing attacks. Let us rewrite the total error rate in a more comprehensible form by introducing a new variable

$$\rho \triangleq \frac{\pi_{\text{spoof}}}{\pi_{\text{non}} + \pi_{\text{spoof}}}$$

which indicates the proportion of spoofing attacks within the negative class on the evaluation database protocol. We refer to ρ as the *spoof prevalence prior*. We can now rewrite (12) as,

$$\begin{aligned} P_{\mathscr{E}}^{\text{tdm}} &= \pi_{\text{tar}}P_{\text{miss}}^{\text{tdm}} + \pi_{\text{non}}P_{\text{fa,non}}^{\text{tdm}} + \pi_{\text{spoof}}P_{\text{fa,spf}}^{\text{tdm}} \\ &= \pi_{\text{tar}}P_{\text{miss}}^{\text{tdm}} + \frac{\pi_{\text{non}} + \pi_{\text{spoof}}}{\pi_{\text{non}} + \pi_{\text{spoof}}}\left[\pi_{\text{non}}P_{\text{fa,non}}^{\text{tdm}} + \pi_{\text{spoof}}P_{\text{fa,spf}}^{\text{tdm}}\right] \\ &= \pi_{\text{tar}}P_{\text{miss}}^{\text{tdm}} + (\pi_{\text{non}} + \pi_{\text{spoof}})\left[(1 - \rho)P_{\text{fa,non}}^{\text{tdm}} + \rho P_{\text{fa,spf}}^{\text{tdm}}\right] \\ &= \pi_{\text{tar}}P_{\text{miss}}^{\text{tdm}} + (1 - \pi_{\text{tar}})P_{\text{fa},\rho}^{\text{tdm}}, \end{aligned}$$

(14)

where the identity $\pi_{\text{non}} + \pi_{\text{spoof}} = 1 - \pi_{\text{tar}}$ follows from the probabilistic constraints and where

$$P_{\text{fa},\rho}^{\text{tdm}} \triangleq (1 - \rho)P_{\text{fa,non}}^{\text{tdm}} + \rho P_{\text{fa,spf}}^{\text{tdm}} \quad (15)$$

denotes the *total tandem false alarm rate*, a convex combination of the two different types of false alarm rates stemming from the acceptance of either nontarget or spoofed trials. For brevity, in the following we refer to (15) as the tandem false alarm rate.

In summary, we re-expressed the total error rate, $P_{\mathscr{E}}^{\text{tdm}}$, in terms of the three error rates $(P_{\text{miss}}^{\text{tdm}}, P_{\text{fa,non}}^{\text{tdm}}, P_{\text{fa,spf}}^{\text{tdm}})$ and the two priors (π_{tar}, ρ). The first prior π_{tar} expresses the total proportion of the target trials in the database protocol while the latter prior ρ encodes the relative proportion of nontarget and spoofing attack trials corresponding to the negative class. Just as the database prior π determines the relative contribution of miss and false alarm rates to the total error rate in the case of a single classifier (2), *two* database priors are necessary for the tandem scenario to determine the relative contribution of *three* error rates.

4 TANDEM EQUAL ERROR RATE

We are now equipped with all we need to define formally the tandem equal error rate (t-EER). As in the case of a single classifier, we seek a scalar metric which reflects the overall performance of the tandem system. Further, the metric should be independent not only of decision costs, but also the database prior(s) and should not require the setting of detection thresholds in advance.

4.1 t-EER

Just as the classic EER corresponds to an operating point at which the miss and false alarm rates are equal, the **tandem equal error rate** (t-EER) is characterized by the following equation:

$$\boxed{P_{\text{miss}}^{\text{tdm}} = P_{\text{fa},\rho}^{\text{tdm}}} \quad (16)$$

where the miss rate on the left-hand side is again considered a measure of user (in)convenience and the false alarm rate on the right-hand side is a measure of security. The latter combines the two different false alarm rates for nontargets and spoofs. For the time being, the spoof prevalence parameter ρ is considered given (fixed). In Section 6 we describe an approach to estimate a tandem error rate that is, like the conventional EER, independent of the class priors.

It it important to note that while the classic EER is *always* a unique scalar that corresponds *typically* to a unique EER threshold, neither holds for the t-EER. In practice, there will be multiple operating points which satisfy (16), each with different values of $P_{\text{miss}}^{\text{tdm}} = P_{\text{fa},\rho}^{\text{tdm}}$, leading to a **t-EER path** $\Omega(\rho)$ for any given $0 \le \rho \le 1$. Formally,

$$\boxed{\Omega(\rho) \triangleq \{\boldsymbol{\tau} \in \mathbb{R}^2 : P_{\text{miss}}^{\text{tdm}}(\boldsymbol{\tau}) = P_{\text{fa},\rho}^{\text{tdm}}(\boldsymbol{\tau})\}} \quad (17)$$

where $\boldsymbol{\tau} = (\tau_{\text{asv}}, \tau_{\text{cm}})$ denote now a *pair* of subsystem thresholds. While further details will be provided below, the eager reader may already take a sneak peek of Figs. 2, 3, and 4 to see how typical t-EER paths looks like.

4.2 Decision Regions for the AND rule

Thus far we have not assumed any particular form in which CM and ASV subsystems are combined. Let us now invoke the assumption in (10). By letting $D(\boldsymbol{s})$ denote the tandem decision for a particular input score pair $\boldsymbol{s} = (s_{\text{asv}}, s_{\text{cm}})$, the 'accept' and 'reject' decision regions are given by

$$\begin{aligned} \text{ACCEPT} &= \{\boldsymbol{s} \in \mathbb{R}^2 : D(\boldsymbol{s}) = 1\} = A_1 \\ \text{REJECT} &= \{\boldsymbol{s} \in \mathbb{R}^2 : D(\boldsymbol{s}) = 0\} = A_2 \cup A_3 \cup A_4, \end{aligned} \quad (18)$$

as illustrated in Fig. 1b. This classification rule is defined by two orthogonal hyperplanes which partition $\mathbb{R}^2$ into four non-overlapping connected regions labeled A_1 through A_4. A test trial, represented by a pair of scores $\boldsymbol{s}$, is accepted if (and only if) $\boldsymbol{s} \in A_1$. Score vectors $\boldsymbol{s}$ that fall in either one of the other three regions are rejected.

4.3 Tandem Miss and False Alarm rates Under the AND Rule

To proceed algebraically from (16), let us begin from the left-hand side. A miss takes place whenever a positive trial (bona fide target) represented by a pair of scores $\boldsymbol{s} = (s_{\mathrm{asv}}, s_{\mathrm{cm}}) \in \mathbb{R}^2$ falls into the union of A_2, A_3 and A_4. This region could be formed in different ways, such as $(A_2 \cup A_4) \cup A_3$ or $(A_3 \cup A_4) \cup A_2$. Following the former, the miss rate is

$$P_{\mathrm{miss}}^{\mathrm{tdm}} = \underbrace{P(S_{\mathrm{asv}} \leq \tau_{\mathrm{asv}} | \mathrm{tar})}_{P(\mathrm{target} \in A_2 \cup A_4)} + \underbrace{P(S_{\mathrm{asv}} > \tau_{\mathrm{asv}} | \mathrm{tar}) P(S_{\mathrm{cm}} \leq \tau_{\mathrm{cm}} | \mathrm{tar})}_{P(\mathrm{target} \in A_3)}$$

$$= P_{\mathrm{miss}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}) + \left(1 - P_{\mathrm{miss}}^{\mathrm{asv}}(\tau_{\mathrm{asv}})\right) P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}),$$

where $P_{\mathrm{miss}}^{\bullet}(\tau_{\bullet})$ denotes the subsystem-specific miss rates at their respective thresholds. The miss rate consists of two terms, the union of which gives rise to the yellow shaded area in Fig 1c. The first term is given by equation (ii) illustrated to the bottom of Fig 1c. (ii), while the second term is the product of (i) and (iv).

Alternatively, one could also obtain the same shaded area by combining (iv) with the product of (ii) and (iii), as follows

$$P_{\mathrm{miss}}^{\mathrm{tdm}} = \underbrace{P(S_{\mathrm{cm}} \leq \tau_{\mathrm{cm}} | \mathrm{tar})}_{P(\mathrm{target} \in A_3 \cup A_4)} + \underbrace{P(S_{\mathrm{asv}} \leq \tau_{\mathrm{asv}} | \mathrm{tar}) P(S_{\mathrm{cm}} > \tau_{\mathrm{cm}} | \mathrm{tar})}_{P(\mathrm{target} \in A_2)}$$

$$= P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) + P_{\mathrm{miss}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}) \left(1 - P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})\right)$$

It is clear that either approach yields the same result:

$$P_{\mathrm{miss}}^{\mathrm{tdm}} = P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) + P_{\mathrm{miss}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}) - P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) P_{\mathrm{miss}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}). \tag{19}$$

Having determined the expression for tandem miss rate, we proceed with the right-hand side of (16). It consists of the convex sum of the two different types of false alarms in (15). The two different errors take place, respectively, when a nontarget trial or a spoof trial falls into A_1. For the former, we have

$$P_{\mathrm{fa,non}}^{\mathrm{tdm}} = \underbrace{P(S_{\mathrm{cm}} > \tau_{\mathrm{cm}} | \mathrm{bona}) P(S_{\mathrm{asv}} > \tau_{\mathrm{asv}} | \mathrm{non})}_{P(\mathrm{nontarget} \in A_1)}$$

$$= \left(1 - F_{S_{\mathrm{cm}}}(\tau_{\mathrm{cm}} | \mathrm{bona})\right) \left(1 - F_{S_{\mathrm{asv}}}(\tau_{\mathrm{asv}} | \mathrm{non})\right)$$

$$= \left(1 - P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})\right) P_{\mathrm{fa,non}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}).$$

Here, we assert $p(S_{\mathrm{cm}} | \mathrm{bona}) := p(S_{\mathrm{cm}} | \mathrm{tar}) = p(S_{\mathrm{cm}} | \mathrm{non})$. Since both target and nontarget correspond to bona fide (human) data, their corresponding CM score distribution are similar. We point the interested reader to [18, Section III.B] for further discussion. Finally, for spoofing attacks

$$P_{\mathrm{fa,spoof}}^{\mathrm{tdm}} = \underbrace{P(S_{\mathrm{cm}} > \tau_{\mathrm{cm}} | \mathrm{spoof}) P(S_{\mathrm{asv}} > \tau_{\mathrm{asv}} | \mathrm{spoof})}_{P(\mathrm{spoofing\ attack} \in A_1)}$$

$$= \left(1 - F_{S_{\mathrm{cm}}}(\tau_{\mathrm{cm}} | \mathrm{spoof})\right) \left(1 - F_{S_{\mathrm{asv}}}(\tau_{\mathrm{asv}} | \mathrm{spoof})\right)$$

$$= P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) P_{\mathrm{fa,spoof}}^{\mathrm{asv}}(\tau_{\mathrm{asv}})$$

Note that in formulating the tandem miss and false acceptance rates above, we have explicitly assumed that the ASV and CM scores are *conditionally independent* such that $p(s_{\mathrm{asv}}, s_{\mathrm{cm}} | Y_{\mathrm{tdm}}) = p(s_{\mathrm{asv}} | Y_{\mathrm{tdm}}) p(s_{\mathrm{cm}} | Y_{\mathrm{tdm}})$. In biometric verification this is a reasonable assumption which greatly simplifies the computation of the tandem error rates (i.e., the shaded area in Fig. 1c) as the product of the error rates evaluated on the ASV and CM score distributions,

TABLE 2: Values of tandem miss rate, tandem false alarm, tandem spoof false alarm and total false alarm rates at the limit cases.

$\tau = (\tau_{\mathrm{asv}}, \tau_{\mathrm{cm}})$	$P_{\mathrm{miss}}^{\mathrm{tdm}}$	$P_{\mathrm{fa,non}}^{\mathrm{tdm}}$	$P_{\mathrm{fa,spf}}^{\mathrm{tdm}}$	$P_{\mathrm{fa},\rho}^{\mathrm{tdm}}$
$(-\infty, -\infty)$	0	1	1	1
$(-\infty, +\infty)$	1	0	0	0
$(+\infty, -\infty)$	1	0	0	0
$(+\infty, +\infty)$	1	0	0	0

TABLE 3: A unified summary of the commonly-used EERs in speaker verification and spoofing detection (first column) along with the respective positive and negative class definitions, along with the ASV (τ_{asv}) and countermeasure (τ_{cm}) thresholds and spoof prevalence (ρ) of the evaluation database. Here '$*$' indicates the free threshold(s) that are varied to find the operating point(s) at which miss rate equals false alarm rate.

The type of EER	Positive	Negative	$(\tau_{\mathrm{asv}}, \tau_{\mathrm{cm}}, \rho)$
ASV (tar vs non)	target	nontar.	$(*, -\infty, 0)$
ASV (tar vs spf)	target	spoof	$(*, -\infty, 1)$
ASV (tar vs non+spf)	target	nontar. $\cup$ spoof	$(*, -\infty, \rho)$
CM (bona vs spf)	bona fide	spoof	$(-\infty, *, 1)$
t-EER of (ASV,CM)	target	nontar. $\cup$ spoof	$(*, *, \rho)$

respectively. While an ASV system focuses on the detection of voice attributes, the CM system aims to discover artifacts (see Appendix for empirical analysis on correlation of several state-of-the-art ASV and CM systems), thereby justifying the independence assumption; see also [18, Section VIII.A]. Note *class-conditional* in the independence assumption. Without conditioning, the ASV and CM scores tend to exhibit stronger correlation, as seen from typical empirical scores in Fig. 4 as well as from the sketch in Fig. 1c. However, this global ASV-CM score dependency is suppressed once conditioned on the class.

All three tandem error rates are functions of the two thresholds, $\boldsymbol{\tau} = (\tau_{\mathrm{asv}}, \tau_{\mathrm{cm}})$. Let us first examine the error rates at the limits where $\boldsymbol{\tau} \in \{(\pm\infty, \pm\infty)\}$, as listed in Table 2. These values follow directly from the properties of the CDF. The last column also shows the tandem false alarm rate defined in (15). Note that, for these limit cases, its value is independent of ρ.

As for the conventional EER, the total miss and false alarm rates have contrasting 0/1 values for each pair of threshold limits in the first column of Table 2. For instance, a comparison of the first and third rows reveals that, when the CM threshold is set to $-\infty$, the tandem miss rate is 1 for an ASV threshold of $+\infty$ and 0 for an ASV threshold of $-\infty$. The first row corresponds to an accept-all condition, while the remaining rows correspond to a reject-all condition. Whenever either threshold is set to $+\infty$, either $S_{\mathrm{cm}} \leq +\infty$ and/or $S_{\mathrm{asv}} \leq +\infty$ hold. The AND rule then always leads to rejection (by the tandem system), giving miss rate of 1 and false alarm rate(s) of 0.

4.4 t-EER of Accept-All and Reject-All Subsystems

With the above definitions in place, let us first consider a few special systems and database protocols. First, note that the t-EER is *undefined* for all the four cases listed in the rows of Table 2; since the tandem miss rate and the tandem

false alarm rate have opposing values, they cannot be made equal.

Accept-all ASV. Now consider the case when the ASV sub-system accepts everything ($\tau_{\text{asv}} = -\infty$), while $|\tau_{\text{cm}}| < \infty$. In this case the tandem system reduces to a CM-only system handling three classes of trials:

$$P_{\text{miss}}^{\text{tdm}} = P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})$$
$$P_{\text{fa},\rho}^{\text{tdm}} = (1-\rho)(1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})) + \rho P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}})$$

For the special case $\rho = 0$ (i.e. there are no spoofed trials), the tandem false alarm rate is $P_{\text{fa},\rho}^{\text{tdm}} = 1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})$. As per (16), we determine the t-EER by equating the two tandem error rates which gives $P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) = 1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})$, i.e. $P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) = \frac{1}{2}$. That is, the CM threshold should be set so that the CM miss rate is $\frac{1}{2}$, or 50%. This is the case also for the CM false alarm rate, giving a t-EER of $\frac{1}{2}$. For $\rho = 1$ (i.e. there are now no nontarget trials), the t-EER is given by $P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) = P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}})$. The CM threshold τ_{cm} must now be set so that the CM miss and false alarm rates are equal — hence, the t-EER is the familiar CM EER.

Accept-all CM. We now consider a similar treatment, but for an accept-all CM (i.e. $\tau_{\text{cm}} = -\infty$), while $|\tau_{\text{asv}}| < \infty$. Now the tandem system reduces to a ASV-only system:

$$P_{\text{miss}}^{\text{tdm}} = P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}})$$
$$P_{\text{fa},\rho}^{\text{tdm}} = (1-\rho)P_{\text{fa,non}}^{\text{asv}}(\tau_{\text{asv}}) + \rho P_{\text{fa,spf}}^{\text{asv}}(\tau_{\text{asv}}).$$

Let us find again the t-EER by equating the tandem miss and false alarm rates. For the special case $\rho = 0$, the ASV threshold τ_{asv} should be set so that $P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}}) = P_{\text{fa,non}}^{\text{asv}}(\tau_{\text{asv}})$. This is the conventional ASV EER. Similarly, $\rho = 1$ gives the EER of the ASV system, but with no nontarget trials. For the intermediate values $0 < \rho < 1$, the t-EER becomes the EER of an ASV system computed using a mix of nontarget and spoof trials. This EER was referred to as the *spoof-aware speaker verification* (SASV) EER in [41].[7] Thus, *this metric depends on the empirical composition of spoofed and non-target trials (i.e., the spoof prevalence prior) in the test set*, making the SASV EER ill-suited for performance comparisons made across different protocols or databases with a different empirical composition.

A summary of the special case t-EERs is shown in Table 3. The four familiar EERs used in speaker verification and spoofing detection — the EER of the CM and the three different EERs of an ASV system (with nontargets, spoofs and a mix of both) are all special cases of the t-EER with specific subsystems and database assumptions. **Hence, the t-EER is a unifying concept: whenever one computes and reports one of the familiar per-subsystem EERs, it is equivalent to reporting t-EERs with special database protocol or subsystem assumptions**.

5 COMPUTATION OF THE T-EER PATH

We now turn our attention toward the case of non-degenerate subsystems with $-\infty < \tau_{\text{asv}}, \tau_{\text{cm}} < +\infty$ where the practical computation of the t-EER path requires more
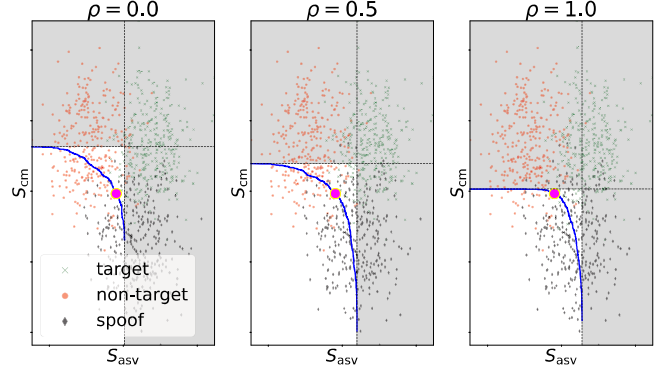

Fig. 2: The t-EER path (blue solid lines) for a given spoofing prevalence prior ρ lives in the constrained region of $\mathbb{R}^2$ (unshaded) defined by two critical thresholds τ_*^{asv} and τ_*^{cm} (dashed black lines). The expressions for the respective critical thresholds are provided by Eq. (20) (ASV) and (22) (CM).

effort. To this end, the t-EER path can be found by considering one of the thresholds fixed (in an outer loop) while sweeping through values of the other (inner loop) in order to determine threshold pairs for which the tandem miss and false alarm rates have the same (or similar[8]) values. Nonetheless, the fixed threshold cannot be set arbitrarily. In the following we provide the technical conditions alongside the visual aid in Fig. 2.

5.1 The t-EER path lives in the lower-left score region

We begin by considering (arbitrarily) a fixed ASV threshold and a variable CM threshold. The following result states that the ASV threshold cannot be selected arbitrarily.

Lemma 5.1. Let $0 \leq \rho \leq 1$ be arbitrary and let $\tau_*^{\text{asv}} \in \mathbb{R}$ be selected so that

$$(1-\rho)P_{\text{fa,non}}^{\text{asv}}(\tau_*^{\text{asv}}) + \rho P_{\text{fa,spoof}}^{\text{asv}}(\tau_*^{\text{asv}}) \geq P_{\text{miss}}^{\text{asv}}(\tau_*^{\text{asv}}) \quad (20)$$

holds (i.e., the total false alarm rate of ASV at τ_*^{asv} is at least as high as the ASV miss rate). Then there exists $\tau_{\text{cm}} \in \mathbb{R}$ so that $(\tau_*^{\text{asv}}, \tau_{\text{cm}}) \in \Omega(\rho)$. Likewise, if (20) does *not* hold, then there is *no* such τ_{cm} for which $(\tau_*^{\text{asv}}, \tau_{\text{cm}}) \in \Omega(\rho)$.

The proof is provided in the Appendix. As illustrated in Fig. 2, the essence of (20) is that the t-EER exists only when the ASV threshold is *sufficiently low*. The CM cannot then be configured to lower the *tandem* miss rate to the same value as the false alarm rate, even by setting $\tau_{\text{cm}} = -\infty$. In particular, note that if $\tau_*^{\text{asv}} = +\infty$ (i.e. 'a reject-all' ASV system), the requirement (20) is

$$(1-\rho) \cdot 0 + \rho \cdot 0 \geq 1 \Leftrightarrow 0 \geq 1, \quad (21)$$

which is identically false. In contrast, for $\tau_*^{\text{asv}} = -\infty$ the condition becomes identically true ($0 \leq 1$). The range of ASV thresholds for which the t-EER exists are those lower than the critical threshold that makes the left and right hand sides of (20) equal.

---

7. To be specific, the particular value of ρ (computed from the evaluation protocol) for the SASV challenge is $\rho \approx 0.65$ — even if not reported in this form.

8. Due to their discrete nature, the tandem miss and false alarm rates may not be precisely identical. Nonetheless, since one is monotonically increasing and the other monotonically decreasing, their profiles will intersect at some operating points defined by a pair of thresholds τ_{cm} and τ_{asv}.
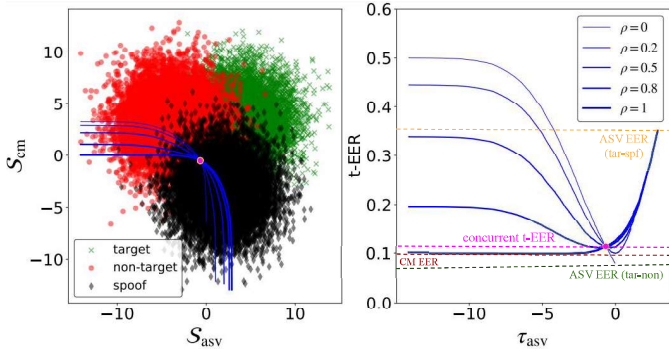
Fig. 3: Demonstration of t-EER paths and t-EER values for different values of ρ on simulated scores drawn from three bivariate Gaussians. The overlaid blue curves on the left-hand side display the t-EER paths (one for each ρ), while the corresponding curves on the right-hand side display the corresponding t-EER *value* along each path. Three familiar special-case EERs are also indicated, along with proposed *concurrent* t-EER (defined in this work).
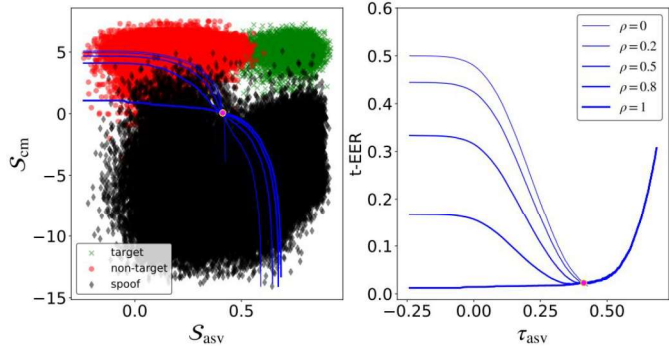

Fig. 4: Demonstration of t-EER paths and t-EER values for different values of ρ on empirical scores from ASV and CM systems (SASV baseline B1). The overlaid blue curves on the left-hand side display the t-EER paths (one for each ρ), while the corresponding curves on the right-hand side display the corresponding t-EER *value* along each path.

The following dual lemma concerns the existence of the t-EER when the CM threshold is fixed, instead of the ASV threshold.

Lemma 5.2. Let $0 \leq \rho \leq 1$ be arbitrary and let $\tau_*^{\mathrm{cm}} \in \mathbb{R}$ be the CM threshold selected so that

$$1 - \rho + \rho P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_*^{\mathrm{cm}}) \geq (2 - \rho) P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_*^{\mathrm{cm}}) \qquad (22)$$

holds. Then there exists $\tau_{\mathrm{asv}} \in \mathbb{R}$ so that $(\tau_{\mathrm{asv}}, \tau_*^{\mathrm{cm}}) \in \Omega(\rho)$.

Proof is similar to that of Lemma 5.1 (omitted). ∎

The result is similar. For the t-EER to exist, the CM threshold should be set low enough in order to ensure a sufficiently high false alarm rate. This allows for the ASV threshold to be set high enough so as to achieve an equal tandem miss rate. Taken together, Eq. (20) and (22) provide useful empirical maximum values for the ASV and CM thresholds that need to be considered in the computation of the t-EER path.

5.2 Construction of the t-EER Path

Up to this point, we have purposefully distanced ourselves from practicalities of the t-EER path construction. Nonetheless, this is an important consideration, particularly in terms of memory consumption. In a practical situation, we have a pair of detection scores, each one generated by one of the two subsystems.

$$\mathcal{S}_{\mathrm{asv}} = \left(s_{\mathrm{asv}}^{(1)}, \ldots, s_{\mathrm{asv}}^{(N)} \right)$$
$$\mathcal{S}_{\mathrm{cm}} = \left(s_{\mathrm{cm}}^{(1)}, \ldots, s_{\mathrm{cm}}^{(M)} \right),$$

along with their respective ground-truth labels $y_{\mathrm{asv}}^{(i)}, i = 1, \ldots, N$ and $y_{\mathrm{cm}}^{(j)}, j = 1, \ldots, M$. Each score set individually takes up linear space in the number of trials and is neglible on modern computer systems. Even if usually not the case, the ASV and CM scores may originate from different sets of test cases, with $N \neq M$. Using the two sets of detection scores one may then construct empirical class-conditional CDFs per subsystem using sorting procedures [30]. As known from conventional single-system ROC analysis (e.g. [25]), the number of distinct pairs of miss and false alarm rate values is the number of unique score values plus one. Hence, the resulting empirical tandem miss and false alarm rates (which consist of all pairs of the two thresholds) can be presented as a 2D array of shape $(N + 1) \times (M + 1)$ which takes quadratic memory for the usual case $N = M$. As typical biometrics and spoofing detection evaluation benchmarks consist of hundreds of thousands to tens of millions of evaluation trials, it is clear that storing the full 2D table becomes prohibitively impractical even for modern computer systems.

Since we are interested only in the t-EER *path*, we do not need to store all these values but, rather, a 1D array of table indices $\mathrm{idx}[i], 1 \leq i \leq N+1$ where $1 \leq \mathrm{idx}[i] \leq M+1$ contains the index of CM threshold corresponding to an ASV threshold indexed by i. Here, each $[i, \mathrm{idx}[i]]$ indices a point on the (now discretely-defined) t-EER path. Depending on whether there are more ASV or CM trials, one may decide to store either the CM threshold indices (one per each ASV threshold that fulfills (20)); or the ASV threshold indices (one per each CM threshold that fulfills (22)). Either case leads to linear, rather than quadratic, memory usage.

6 THE CONCURRENT T-EER: A UNIQUE SCALAR SUMMARY

As noted above, the conventional EER is uniquely defined for a given set of scores. Unfortunately, this is generally *not* the case for the t-EER defined in (16). That is, if $\boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in \Omega(\rho)$ are two distinct points ($\boldsymbol{\tau}_1 \neq \boldsymbol{\tau}_2$) on a t-EER path, then the t-EERs at $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ correspond to *different* values of t-EER $= P_{\mathrm{miss}}^{\mathrm{tdm}} = P_{\mathrm{fa}, \rho}^{\mathrm{tdm}}$. Hence, the t-EER is in fact not a scalar but a *function* with the t-EER path $\Omega(\rho)$ as its domain.

A natural question arises: *which* t-EER value along the t-EER path should we choose as a meaningful summary statistic? Given the optimum (oracle) threshold sentiment of conventional EER, it may seem very reasonable to consider the *minimum*. Unfortunately, since the t-EER path (and the corresponding t-EER values) depends on ρ, that minimum will also depend on ρ. This leads to the same empirical trial

composition dependency problem noted in Subsection 4.4. Our answer to the original question will be *choose the unique t-EER value shared across* all *spoof prevalence priors, at the unique intersection point of all possible paths* $\{\Omega(\rho) : \rho \in [0, 1]\}$.

Before formalizing our proposal, let us consider a simple example. Fig. 3 shows an illustrative example of t-EER analysis for a set of simulated CM and ASV scores sampled from three bivariate normal distributions, one per each of the three classes (target, non-target and spoof).[9] The t-EER paths are derived from the treatment presented above and are reproducible using online code.[10] Profiles are plotted for three different values of the spoofing prevalence prior ρ. The plot to the left shows CM scores (vertical axis) and ASV scores (horizontal axis) for target trials (green), non-target trials (red) and spoofed trials (black). Also shown are super-imposed t-EER paths for each ρ defined by corresponding τ_{cm} and τ_{asv} threshold pairs. The plot to the right in Fig. 3 shows the t-EER as a function of the ASV threshold τ_{asv}.

Recall that the spoof prevalence parameter ρ combines the two tandem false alarm rates in (15) corresponding to the non-target and spoof classes, respectively. The t-EER paths equate the tandem miss rate to the non-target and spoof tandem false alarm rates at different thresholds. The value of ρ determines the prevalence of the spoofed class, with respect to the non-target class, in the tandem false rejection rate. For lower ρ, the nontarget class has a larger weight than the spoofed class. As such, the t-EER is higher for lower τ_{asv} (higher non-target false acceptance rate) but lower for higher τ_{asv} as noted in Fig. 3.

As expected, the profiles show that the value of the t-EER differs along each path for all three values of the spoofing prevalence prior ρ. Of specific interest is the apparent intersection of each path at what appears to be the same values of τ_{cm} and τ_{asv}. At this specific operating point, the value of the t-EER also appears to be identical and is, hence, independent of ρ.

6.1 The concurrent t-EER: where t-EER paths intersect

Up until now, our treatment has concerned t-EER *paths* $\Omega(\rho)$, rather than a single scalar like the conventional EER. Each path characterises a set of operating points that satisfy (16) and for which the t-EER may be different. In further contrast to the conventional EER, the paths are dependent on the database prior, ρ. This dependence is undesirable, not least because it implies that estimates of performance derived from different databases cannot be made using the t-EER; the estimates are dependent on the database or, more specifically, the spoofing prevalence prior for that database.

As shown in Fig. 3, the dependence on the database prior is seen for each point along each t-EER path, except for the particular operating point at which the t-EER paths intersect. At this particular point, the t-EER is independent of ρ just like the conventional EER.

---

9. Simulated scores are produced similarly to [18, Appendix] where the marginals define the ASV and CM score distributions parameterized through their respective EERs. In this simulation, we use arbitrary EER values of 8% (ASV target-vs-nontarget); 35% (ASV target-vs-spoof); and 10% (CM target-vs-spoof), and diagonal covariance matrices.

10. URL to be added upon publication.

Theorem 6.1. (Characterization of t-EER path intersection) Let $0 \leq \rho_1, \rho_2 \leq 1$ be arbitrary but distinct ($\rho_1 \neq \rho_2$) spoof prevalence priors. Then the corresponding t-EER paths $\Omega(\rho_1)$ and $\Omega(\rho_2)$ intersect at $\boldsymbol{\tau}^{\times} \triangleq (\tau_{\mathrm{asv}}^{\times}, \tau_{\mathrm{cm}}^{\times})$, where the two thresholds satisfy

$$\frac{P_{\mathrm{fa,non}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}^{\times})}{P_{\mathrm{fa,spf}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}^{\times})} = \frac{P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}^{\times})}{1 - P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}^{\times})}. \quad (23)$$

Further, the t-EER at $\boldsymbol{\tau}_{\times}$, denoted by $P_{\mathscr{E}}^{\times}$, is

$$P_{\mathscr{E}}^{\times} = P_{\mathrm{fa,spf}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}^{\times}) P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}^{\times}). \quad (24)$$

The proof (see Appendix A) shows that the t-EER paths for different values of ρ have a common intersection point and a common value of t-EER on that point. For simplicity, consider the boundary cases for the spoofing prevalence prior $\rho = 0$ (no spoofing attacks present) and $\rho = 1$ (no non-targets present). Their t-EER at the intersection point is given by

$$\begin{aligned} P_{\mathrm{miss}}^{\mathrm{tdm}}(\boldsymbol{\tau}^{\times}) &= P_{\mathrm{fa,non}}^{\mathrm{tdm}}(\boldsymbol{\tau}^{\times}) & \text{for} \quad \rho = 0 \\ P_{\mathrm{miss}}^{\mathrm{tdm}}(\boldsymbol{\tau}^{\times}) &= P_{\mathrm{fa,spoof}}^{\mathrm{tdm}}(\boldsymbol{\tau}^{\times}) & \text{for} \quad \rho = 1 \end{aligned} \quad (25)$$

since $P_{\mathrm{miss}}^{\mathrm{tdm}}(\boldsymbol{\tau}^{\times})$ is, by definition, the same for all $\rho \in [0, 1]$. From the above equations, it can be seen that, at the point of intersection, the three error rates are *concurrently* equal:

$$\boxed{P_{\mathscr{E}}^{\times} = P_{\mathrm{miss}}^{\mathrm{tdm}}(\boldsymbol{\tau}_{\times}) = P_{\mathrm{fa,non}}^{\mathrm{tdm}}(\boldsymbol{\tau}^{\times}) = P_{\mathrm{fa,spoof}}^{\mathrm{tdm}}(\boldsymbol{\tau}^{\times})} \quad (26)$$

We refer to the value of the t-EER at the point of intersection as the **concurrent t-EER**, which uniquely characterizes the performance of a tandem system with a single parameter-free performance metric. It is worth noting that one could also deduce (26) from (A.3)[Appendix A] using the expression in the parentheses. It is worth noticing that the concurrent relation in (26) indicates a unique point $P_{\mathscr{E}}^{\times}$ on the CM-ASV score space where the target, non-target, and spoofed score distributions intersect with the three error rates being equal. The boundary conditions (where either nontarget or spoof class vanishes) are given by (25).

6.2 The relationship between the concurrent t-EER, the total error rate and the t-DCF

Let us denote the concurrent t-EER value by $P_{\mathscr{E}}^{\times}$. Straightforward substitution of $P_{\mathscr{E}}^{\times}$ in place of $P_{\mathrm{miss}}^{\mathrm{tdm}}$, $P_{\mathrm{fa,non}}^{\mathrm{tdm}}$, and $P_{\mathrm{fa,\rho}}^{\mathrm{tdm}}$ in (12) shows that $P_{\mathscr{E}}^{\mathrm{tdm}} = P_{\mathscr{E}}^{\times}$, i.e. the concurrent t-EER and total tandem error rate equal one another (at $\boldsymbol{\tau}_{\times}$). This is a direct extension of the relationship between the classical total error rate and the EER discussed in Section 2.4.

Another relevant connection can be made to the tandem detection cost function (t-DCF) [18], which has the form

$$\text{t-DCF} = C_{\mathrm{miss}} \pi_{\mathrm{tar}}' P_{\mathrm{miss}}^{\mathrm{tdm}} + C_{\mathrm{fa,non}} \pi_{\mathrm{non}}' P_{\mathrm{fa,non}}^{\mathrm{tdm}} + C_{\mathrm{fa,spf}} \pi_{\mathrm{spf}}' P_{\mathrm{fa,spf}}^{\mathrm{tdm}}, \quad (27)$$

where $C_{(.)} > 0$ are detection costs associated with the three different types of errors and $\pi_{(.)}'$ are *asserted* priors that may differ from the *database priors* introduced in (9). Both the detection costs and asserted priors are constants that one fixes prior to observing the evaluation data. Substituting the concurrent t-EER to (27) gives

$$\text{t-DCF} = P_{\mathscr{E}}^{\times} \underbrace{\left(C_{\mathrm{miss}} \pi_{\mathrm{tar}} + C_{\mathrm{fa,non}} \pi_{\mathrm{tar}} + C_{\mathrm{fa,spf}} \pi_{\mathrm{spf}} \right)}_{\text{Constant}}, \quad (28)$$

TABLE 4: Number of target, non-target and spoof trials in the evaluation partitions of the SASV and ASVspoof 2021 logical access (LA) databases. Due to the nature of the challenge in which participants prepared CM solutions alone, there are different protocols for the training of CMs and the ASV system for ASVspoof 2021 LA database.

SASV			ASVspoof 2021 LA				
			CM trials		ASV trials		
target	non-target	spoof	target	spoof	target	non-target	spoof
38697	33327	63882	14816	133360	13467	543114	133362

i.e. the t-DCF and the t-EER (at $\tau_\times$) are linearly related. Furthermore, by denoting $C_{\min} \triangleq \min\left(C_{\mathrm{miss}}, C_{\mathrm{fa}}, C_{\mathrm{fa,spf}}\right)$ and $C_{\max} \triangleq \max\left(C_{\mathrm{miss}}, C_{\mathrm{fa}}, C_{\mathrm{fa,spf}}\right)$, the convexity of the expression in the parentheses (the asserted priors sum up to 1) can be used to sandwich the t-DCF value between lower and upper bounds:

$$C_{\min} \cdot P_{\mathscr{E}}^{\times} \leq \text{t-DCF} \leq C_{\max} \cdot P_{\mathscr{E}}^{\times}. \tag{29}$$

These results indicate a close connection between the concurrent t-EER and the t-DCF.

7 EXPERIMENTS

To demonstrate use of the new t-EER using empirical scores, we conducted a set of experiments using SASV and ASVspoof databases and baseline systems. Both SASV and ASVspoof 2021 logical access (LA) databases are sourced from the same ASVspoof 2019 database, while the baselines are different.

7.1 Experimental setup

SASV 2022 experiments were performed using the AASIST CM model described in [42] and the ECAPA-TDNN ASV model descibed in [43]. Together they are the B1 baseline described in [44]. Experiments were performed using the publicly available SASV 2022 experimental protocol [45].

We used all four ASVspoof 2021 baseline systems. The ASV subsystem was fixed by the challenge organisers and is a standard *x-vector* [46] system with probabalistic linear discriminant analysis (PLDA) scoring [47]. The CM subsystems use: B1 – constant-Q cepstral coefficients (CQCCs) and a Gaussian mixture model (GMM); B2 – high resolution linear frequency cepstral coefficients (LFCCs) and a GMM; B3 – LFCCs with a light convolutional neural network (LCNN); B4 – a RawNet2 model. Further details of each are available in [48]. Experiments were performed using the ASVspoof 2021 experimental protocols [9], [49].

The number of target, non-target and spoofed trials for the evaluation partition of each database is shown in Table 4. For the ASVspoof 2021 LA challenge, participants prepared CM solutions alone, while the ASV system was designed by the challenge organisers. There are hence different protocols for the evaluation of each. Since we are not concerned with optimisation in this work, all experiments were performed using the pretrained baseline systems with default settings and the respective evaluation partitions of each database.

7.2 Results

Results are illustrated in Table 5. Columns 2-12 show EER results (including concurrent t-EER) for each database and baseline shown in column 1. Column 13 shows minimum tandem detection cost function (min t-DCF) results for comparison. Columns 2-6 show ASV EER results for protocols comprising only: target and non-target trials ($\rho = 0$); target and spoofed trials ($\rho = 1$); target vs. nontarget and spoofed trials for three intermediate spoofing prevalence priors ($0 < \rho < 1$). Column 7 shows the CM EER for a protocol comprising bona fide (target and nontarget) and spoofed trials.

For a given database and ρ, the ASV EERs are the same, no matter what the baseline. They vary between 1.6% and 30.7% for the SASV database and between 7.6% and 38.5% for the ASVspoof database. Unsurprisingly, as the prevalence of spoofed trials increases (increasing ρ), so too does the EER. The dependence of the EER upon the spoofing prevalence prior is undesirable. Experiments involving different values of ρ are, in effect, akin to experiments using different databases and protocols with different spoofing prevalence priors. The differences imply that performance comparisons across databases cannot be made using the EER.

Columns 8-12 show corresponding concurrent t-EER results for the same range of ρ. They are close to being identical for each row, indicating that they are dependent on the system, but not the spoofing prevalence prior (nor any other prior). The independence of the concurrent t-EER to priors implies that it can be used as a more reliable metric for making performance comparisons across different databases.

Plots similar to those in Fig. 3, but for empirically-derived scores for the SASV B1 baseline, are shown in Fig. 4 for the same values of ρ shown in Table 5. The EER for the $\rho = 1$ condition (no nontargets) of 30.7% corresponds to the t-EER value for the highest ASV threshold τ_{asv} (rightmost points in Fig. 4(b)). The EER for the ($\rho = 0$) condition (no spoofs) of 1.6% furthermore corresponds to the lowest point on the t-EER path for $\rho = 0$. The CM EER of 0.7% is shown by the horizontal asymptote of the t-EER path for $\rho = 1$. Finally, the concurrent t-EER of 2.3% is shown by the magenta marker.

While not the goal in this paper, the *ranking* of concurrent t-EER results reflects the same trend as min t-DCF results. In the case of results for the ASVspoof 2021 database, the ranking derived from CM EER results alone (ASV results are in any case invariable across different CM solutions for a given ρ) is different. It might seem at first that B4 is the best among the four baselines (the CM EER is lowest). Tandem results, however, show that B3 might be a better option (the t-EER is lowest). This observation adds further weight to the argument that the selection of CM solutions should not be made from CM results alone, but instead be made from the results of *tandem assessment*. The concurrent t-EER offers one suitable approach.

TABLE 5: Equal error rates (EERs) for SASV and ASVspoof 2021 databases and baseline solutions. In each case, ASV EERs are conventional EERs estimated from a set of positive and negative class trials. Positive class trials are always target trials, whereas negative class trials involve a mix of non-target trials and spoofed trials where the proportions of each are dictated by the spoofing prevalence prior ρ. For each given ρ, the ASV system and hence also the EER is the same for all ASVspoof 2021 experiments. The CM EER is estimated from a mix of positive class (bona fide) trials and negative (spoofed) trials. The concurrent t-EER is estimated from the combination of ASV and CM sub-systems using the same trial sets. Tandem detection cost functions (t-DCF) results are provided for comparison.

System	ASV EER					CM EER	Concurrent t-EER					Min t-DCF
	$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 1$		$\rho = 0$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 1$	
SASV baseline B1	1.63	12.47	20.85	28.15	30.74	0.67	2.28	2.28	2.28	2.28	2.28	0.0660
ASVspoof 2021 LA-B1	7.61	14.56	25.16	34.01	38.49	15.53	14.47	14.47	14.47	14.47	14.47	0.4930
ASVspoof 2021 LA-B2	7.61	14.56	25.16	34.01	38.49	19.97	18.42	18.42	18.42	18.42	18.42	0.5857
ASVspoof 2021 LA-B3	7.61	14.56	25.16	34.01	38.49	9.26	9.37	9.37	9.37	9.37	9.37	0.3445
ASVspoof 2021 LA-B4	7.61	14.56	25.16	34.01	38.49	8.66	10.34	10.34	10.34	10.34	10.34	0.4027

8 DISCUSSION AND CONCLUSIONS

8.1 Summary of the Proposed t-EER Metric

Experimental results reported in this paper corroborate previous findings that the evaluation of spoofing counter-measure solutions should be performed in tandem with biometric recognisers; their evaluation in isolation might result in the selection of a sub-optimal solution. By extending the standard notion of detection theory and the EER, we introduced an approach to tandem evaluation in the form of the tandem equal error rate (t-EER). Its application results in a path of operating points — the t-EER path — along which the tandem false alarm and miss rates are equal. Nonetheless, each operating point along each path corresponds to a different t-EER value and the path itself is dependent upon the prevalence of spoofing attacks in the database. We tackled this problem by introducing *concurrent* t-EER operating point at which all the paths intersect, thereby giving a unique error rate that is prior-independent. Just as the conventional EER corresponds to an operating point at which the errors associated with the two classes are equal ($P_{\text{miss}} = P_{\text{fa}}$), the concurrent t-EER corresponds to an operating point at which the three *tandem* error rates are equal: $P_{\text{miss}}^{\text{tdm}} = P_{\text{fa}}^{\text{tdm}} = P_{\text{fa,spoof}}^{\text{tdm}}$.

Besides the concurrent t-EER point, an intuitive error-rate based performance measure of the tandem system, our work also sheds light upon the various per-system EERs used in the separate evaluation of biometric comparator and PAD subsystems. The t-EER path also reveals the:

- EER of the biometric comparator in the absence of spoofing attacks (an accept-all PAD, and spoofing attack prior $\pi_{\text{spoof}} = 0$);
- EER of the biometric comparator in the absence of non-target trials (an accept-all PAD, spoofing attack prior $\pi_{\text{spoof}} = 1$);
- EER of the biometric comparator with a mix of non-targets and spoofing attacks (an accept-all PAD, and spoofing attack ($0 < \pi_{\text{spoof}} < 1$);
- EER of the PAD subsystem (accept-all comparator).

The t-EER path hence presents a *unified* view of biometric comparator and PAD subsystem performance under a variety of protocols and configurations. The authors advocate for the adoption of the t-EER framework in place of traditional two-class EERs when assessment involves three classes — target, non-target and spoofed trials. We hence discourage the continued use of two-class EER metrics, such as the SASV-EER adopted for the SASV challenge [41]; it is dependent on the database prevalence prior and potentially leaves comparisons of performance made across different datasets essentially meaningless.

8.2 Towards a reference approach to joint evaluation

Though an awareness of vulnerabilities to PAs was founded some two decades ago [5] and though a framework for PAD research and performance assessment has been standardized [3], research into the *joint* operation and evaluation of biometric comparator and PAD subsystems remains in its relative infancy. The usual approach, score fusion, transforms the task of joint *evaluation* into one of *back-end classifier design* with nearly limitless design possibilities and inherent training data dependencies.

Being convinced that simplicity is the key towards defining a reference approach, the authors have advocated a natural two-threshold, three-class extension of single-system ROC analysis to support the tandem evaluation of biometric comparators and PAD subsystems. The main assumption (the same as that in [18]) is the statistical independence of the scores produced by each subsystem. It is justified by noting that PAD and biometric comparators address different classification tasks. We hope the work will inspire further discussion on tandem evaluation through a simple error-rate based approach. By avoiding the need to specify application-specific parameters, we embrace a spirit more in keeping with the ISO standardised approach to the separate evaluation of PAD. The new t-EER metric can be used for the comparison of different solutions and may also be applied when making comparisons between results derived from different databases. Importantly, the metric is agnostic to the biometric mode and requires only scores produced by biometric comparators and PAD subsystems. Hence, the proposed metric has potential to be adapted as a common reference approach to joint evaluation. To this end, the authors provide an open-source Python implementation.[10]

A potential interesting future extension concerns tandem evaluation of *more than two systems*. Biometric systems are known to be vulnerable to many different kinds of

attack vectors (e.g. replay and text-to-speech for speech) that typically benefit from tailored countermeasures combined with the biometric comparator. An even more general situation involves multiple biometric modes (e.g. face and voice), each with potentially multiple biometric comparators and/or PAD subsystems. We are again faced with the key question how the different systems should be combined. Assuming that all the relevant subsystems can be collapsed into a single score (e.g. score fusion of multiple PADs), the t-EER approach remains applicable. More principled tandem assessment of multiple PADs and biometric modes is left as a future work.

8.3 Fundamental Limitations of (t-)EER

The present work should not be viewed as the end but, rather, a first rigorous step towards the unified joint evaluation of biometric comparators and PAD subsystems. The authors are well aware that our starting point — the classical EER — has a number of known limitations, namely:

- the EER threshold is selected by an oracle using the ground-truth class labels of the particular evaluation database, hence the problems of threshold selection or the training of a calibration model is not accounted for;
- an operating point specified by balanced errors (misses and false alarms) will not suit all applications where preference might naturally be given to lowering the miss rate or the false alarm rate.

The first shortcoming is exactly why the community [17], [34], [35], [36], [37] encourages assessment approaches that take threshold selection into consideration from the very beginning. The second accounts for why metrics such as the DCF [16] and t-DCF [18] require the setting of detection costs and asserted priors. Being an extension of the classical EER, the proposed t-EER inherits both of the above shortcomings. So, why did we bother to write this paper?

The main reason is that, **despite the acknowledged drawbacks, the EER remains a popular metric for evaluation and reporting**. As a matter of fact, there is nothing particularly wrong in using the EER as a *discrimination* (class separability) criterion, similar to the popular AUC metric — as long as the above fundamental limitations are understood and acknowledged. The lesser-known interpretation of the EER as an upper bound to the total error rate [24] indeed makes the EER well suited to measurements of class separability. Like the AUC, the EER is useful for quick and easily interpreted analyses and for initial investigations, such as *rejecting* features or models that lack the potential for accurate classification. That the EER is easily visualized on a ROC or DET plots, that it is independent of the class proportions, and that is can be estimated without the need for additional data beyond the evaluation data are all likely reasons for why the EER remains popular.

While calibration remains an important (yet different) problem, the focus in this paper is upon discrimination and a proposal towards a common approach to the joint evaluation of biometric comparators and PAD subsystems built upon foundations that should be familiar to anyone working or having an interest in biometric recognition. In fact,

given the lack of any candidate approaches in the current ISO/IEC PAD standard [3], *any* reasonable metric provides a suitable starting point for the future definition of a common approach to the joint evaluation of biometric comparators operating in tandem with PAD subsystems. Ultimately, we hope the new t-EER serves also as an eventual starting point for the exploration of approaches to the tandem evaluation of calibration in addition to discrimination.

References

[1] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, *Guide to biometrics*. Springer Science & Business Media, 2013.

[2] J. L. Wayman, A. K. Jain, D. Maltoni, and D. Maio, *Biometric systems: Technology, design and performance evaluation*. Springer Science & Business Media, 2005.

[3] "ISO/IEC 30107. Information Technology – Biometric presentation attack detection," International Organization for Standardization, Geneva, Switzerland, Standard, 2016.

[4] "ISO/IEC FDIS 19795-1:2021. Information Technology – Biometric performance testing and reporting – Part 1: Principles and framework," International Organization for Standardization, Geneva, Switzerland, Standard, 2021.

[5] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[6] J. Galbally, J. Fiérrez, and R. Cappelli, "An Introduction to Fingerprint Presentation Attack Detection," in *Handbook of Biometric Anti-Spoofing - Presentation Attack Detection, 2nd Edition*. Springer, 2019, pp. 3–31.

[7] J. Hernandez-Ortega, J. Fiérrez, A. Morales, and J. Galbally, "Introduction to Face Presentation Attack Detection," in *Handbook of Biometric Anti-Spoofing - Presentation Attack Detection, Second Edition*, ser. Advances in Computer Vision and Pattern Recognition. Springer, 2019, pp. 187–206.

[8] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[9] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. H. Kinnunen, M. Todisco, J. Yamagishi, N. W. D. Evans, A. Nautsch, and K.-A. Lee, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *ArXiv*, vol. abs/2210.02437, 2022.

[10] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proc. Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[12] "ISO/IEC 30107-3:2023 information technology — biometric presentation attack detection — part 3: Testing and reporting," International Organization for Standardization, Geneva, Switzerland, Standard, 2023.

[13] M. Micheletto, G. Orrù, R. Casula, D. Yambay, G. L. Marcialis, and S. C. Schuckers, "Review of the Fingerprint Liveness Detection (LivDet) competition series: from 2009 to 2021," *arXiv preprint arXiv:2202.07259*, 2022.

[14] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. International Speech Communication Association (INTERSPEECH)*, 2021, pp. 4259–4263.

[15] I. Chingovska, A. Mohammadi, A. Anjos, and S. Marcel, *Evaluation Methodologies for Biometric Presentation Attack Detection.* Springer, 2019, pp. 457–480. [Online]. Available: https://doi.org/10.1007/978-3-319-92627-8_20

[16] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2, pp. 225–254, 2000.

[17] I. Chingovska, A. R. d. Anjos, and S. Marcel, "Biometrics evaluation under spoofing attacks," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2264–2276, 2014.

[18] T. Kinnunen, H. Delgado, N. W. D. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2195–2210, 2020.

[19] E. Marasco, Y. Ding, and A. Ross, "Combining match scores with liveness values in a fingerprint verification system," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2012, pp. 418–425.

[20] M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z.-H. Tan, "Integrated Spoofing Countermeasures and Automatic Speaker Verification: An Evaluation on ASVspoof 2015," in *Proc. INTERSPEECH*, 2016, pp. 1700–1704.

[21] P. Korshunov and S. Marcel, "Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 695–705, 2017.

[22] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion," in *Proc. INTERSPEECH*, 2018, pp. 77–81.

[23] I. Chingovska, A. Anjos, and S. Marcel, "Anti-spoofing in Action: Joint Operation with a Verification System," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2013, pp. 98–104.

[24] N. Brümmer, L. Ferrer, and A. Swart, "Out of a Hundred Trials, How Many Errors Does Your Speaker Verifier Make?" in *Proc. INTERSPEECH*, 2021, pp. 1059–1063.

[25] W. Krzanowski and D. Hand, *ROC Curves for Continuous Data*, 1st ed., ser. CRC Monographs on Statistics and Applied Probability. Chapman & Hall, 2009.

[26] N. Macmillan and C. Creelman, *Detection theory: A user's guide*, 2nd ed. Lawrence Erlbaum Associates Publishers, 2005.

[27] ISO. (2023) ISO/IEC 2382-37:2022(en) information technology–vocabulary–part 37: Biometrics. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:-37:ed-3:v1:en

[28] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[29] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance." in *Proc. EUROSPEECH*, 1997.

[30] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[31] A. Nautsch, "Speaker recognition in unconstrained environments," Ph.D. dissertation, Technische Universität Darmstadt, 2019.

[32] D. J. Hand, "Assessing the Performance of Classification Methods," *International Statistical Review*, vol. 80, no. 3, pp. 400–414, 2012.

[33] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2020.

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

[35] D. J. Hand and C. Anagnostopoulos, "Notes on the H-measure of classifier performance," *Advances in Data Analysis and Classification*, pp. 1–16, 2022.

[36] L. Ferrer, M. McLaren, and N. Brümmer, "A speaker verification backend with robust performance across conditions," *Computer Speech & Language*, vol. 71, p. 101258, 2022.

[37] S. Bengio and J. Mariéthoz, "The expected performance curve: a new assessment measure for person authentication," in *Proc. the Speaker and Language Recognition Workshop (Speaker Odyssey)*, 2004, pp. 279–284.

[38] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.

[39] N. Poh and S. Bengio, "Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks?" in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2004, pp. V–893.

[40] D. A. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proc. INTERSPEECH*, 2013, pp. 1619–1623.

[41] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-G. Kang, H.-J. Yu, N. Evans, and T. Kinnunen, "SASV challenge 2022: A spoofing aware speaker verification challenge evaluation plan," *arXiv preprint arXiv:2201.10283*, 2022.

[42] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," in *Proc. ICASSP*, 2022.

[43] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. INTERSPEECH*, 2020, pp. 3830–3834.

[44] H. jin Shim, H. Tak, X. Liu, H.-S. Heo, J. weon Jung, J. S. Chung, S.-W. Chung, H.-J. Yu, B.-J. Lee, M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, T. Kinnunen, and N. Evans, "Baseline Systems for the First Spoofing-Aware Speaker Verification Challenge: Score and Embedding Fusion," in *Proc. The Speaker and Language Recognition Workshop (Speaker Odyssey)*, 2022, pp. 330–337.

[45] J.-w. Jung, H. Tak, H.-j. Shim, H.-S. Heo, B.-J. Lee, S.-W. Chung, H.-J. Yu, N. Evans, and T. Kinnunen, "SASV 2022: The First Spoofing-Aware Speaker Verification Challenge," in *Proc. INTERSPEECH*, 2022, pp. 2893–2897.

[46] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 2018, pp. 5329–5333.

[47] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE International conference on computer vision*, 2007, pp. 1–8.

[48] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. The Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[49] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," in *arXiv preprint arXiv:2109.00535*, 2021.

[50] Y. Kwon, H.-S. Heo, B.-J. Lee, and J. S. Chung, "The ins and outs of speaker recognition: lessons from voxsrc 2020," in *Proc. ICASSP*, 2021, pp. 5809–5813.

[51] S. H. Mun, J.-w. Jung, M. H. Han, and N. S. Kim, "Frequency and multi-scale selective kernel attention for speaker verification," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 548–554.

[52] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Proc. INTERSPPECH*, 2022.

[53] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," in *Proc. Speaker Odyssey Workshop*, 2022.

[54] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," *Proc. Speaker Odyssey Workshop*, 2022.

[55] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *Proc. ICASSP). IEEE*, 2021, pp. 6369–6373.

Tomi H. Kinnunen is full professor of speech technology at the University of Eastern Finland (UEF). He received his Ph.D. degree (computer science) from the University of Joensuu in 2005. From 2005 to 2007, he was with the Institute for Infocomm Research (I2R), Singapore. Since 2007, he has been with UEF. His research has been funded by Academy of Finland and he also partnered in H2020-funded OCTAVE project. He chaired the Odyssey workshop in 2014. From 2015 to 2018, he served as an Associate Editor for IEEE/ACM-T-ASLP and from 2016 to 2018 as a Subject Editor in Speech Communication. In 2015 and 2016, he visited the National Institute of Informatics, Japan, for 6 months under a mobility grant from the Academy of Finland. He is one of the co-founders of the ASVspoof challenge, a nonprofit initiative that seeks to evaluate and enhance the security of voice biometric solutions under spoofing attacks.

Nicholas Evans (Member, IEEE) is a Professor at EURECOM, France, where he heads research in Audio Security and Privacy. He received his Ph.D. from the University of Wales Swansea, UK in 2004. He is a co-founder of the community led, ASVspoof, SASV, and VoicePrivacy challenge series. He participated in the EU FP7 Tabula Rasa and EU H2020 OCTAVE projects, both involving antispoofing. Today, his team is leading the EU H2020 TReSPAsS-ETN project, a training initiative in security and privacy for multiple biometric characteristics. He co-edited the second and third editions of the Handbook of Biometric Anti-Spoofing, served previously on the IEEE Speech and Language Technical Committee and serves currently as an associate editor for the IEEE Trans. on Biometrics, Behavior, and Identity Science.

Kong Aik Lee (M'05-SM'16) is currently a Senior Scientist and a Group Leader at the Agency for Science, Technology and Research (A*STAR), Singapore. He was a Senior Principal Researcher at the Data Science Research Laboratories, NEC Corporation, Japan, from 2018 to 2020. He received his Ph.D. degree from Nanyang Technological University, Singapore, in 2006. After which he joined the Institute for Infocomm Research, Singapore, as a Research Scientist and then a Strategic Planning Manager (concurrent appointment). He was the recipient of the Singapore IES Prestigious Engineering Achievement Award 2013 for his contribution to voice biometrics technology, and the Outstanding Service Award by IEEE ICME 2020. He was the Lead Guest Editor for the CSL Special Issue on "Two decades into Speaker Recognition Evaluation - are we there yet?" Currently, he serves as an Editorial Board Member for Elsevier Computer Speech and Language (2016 - present) and was an Associate Editor for IEEE/ACM Transactions on Audio, Speech, and Language Processing (2017 - 2021). He is an elected member of the IEEE Speech and Language Processing Technical Committee and was the General Chair of the Speaker Odyssey 2020 Workshop. His research focuses on the automatic and para-linguistic analysis of speaker characteristics, ranging from speaker recognition, language and accent recognition, diarization, voice biometrics, spoofing, and countermeasure.

Andreas Nautsch is research engineer for the OpenSource toolkit SpeechBrain, working for Avignon Université. He received his doctorate from TU Darmstadt (2019). Andreas served as project editor for ISO/IEC 19794-13:2018 (biometric voice data interchange format), co-initiated the ISCA SIG on Security & Privacy in Speech Communication (SPSC; 2019 to '22), was an Associate Editor for the EURASIP Journal on Audio, Speech, and Music Processing (2020 to '22), co-organised editions of the VoicePrivacy and ASVspoof challenges, and co-lead the 2021 Lorentz workshop on Speech as Personal Identifiable Information.

Hemlata Tak is a PhD candidate at EURECOM, France. She received her M.tech degree in 2018 at DA-IICT, Gandhinagar, India. Her research interests include voice biometrics, audio deepfake detection and anti-spoofing. She co-organised the inaugural edition of the Spoof-Aware Speaker Verification (SASV) Challenge 2022.