Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals

Tomi Kinnunen, *Member, IEEE,* Héctor Delgado, *Member, IEEE,* Nicholas Evans *Member, IEEE,*
Kong Aik Lee, *Senior Member, IEEE,* Ville Vestman, Andreas Nautsch, *Member, IEEE,*
Massimiliano Todisco, *Member, IEEE,* Xin Wang, *Member, IEEE,* Md Sahidullah *Member, IEEE,*
Junichi Yamagishi, *Senior Member, IEEE,* and Douglas A. Reynolds, *Fellow, IEEE,*

Abstract—Recent years have seen growing efforts to develop spoofing countermeasures (CMs) to protect automatic speaker verification (ASV) systems from being deceived by manipulated or artificial inputs. The reliability of spoofing CMs is typically gauged using the equal error rate (EER) metric. The primitive EER fails to reflect application requirements and the impact of spoofing and CMs upon ASV and its use as a primary metric in traditional ASV research has long been abandoned in favour of risk-based approaches to assessment. This paper presents several new extensions to the tandem detection cost function (t-DCF), a recent risk-based approach to assess the reliability of spoofing CMs deployed in tandem with an ASV system. Extensions include a simplified version of the t-DCF with fewer parameters, an analysis of a special case for a fixed ASV system, simulations which give original insights into its interpretation and new analyses using the ASVspoof 2019 database. It is hoped that adoption of the t-DCF for the CM assessment will help to foster closer collaboration between the anti-spoofing and ASV research communities.

Index Terms—automatic speaker verification, spoofing countermeasures, presentation attack detection, detection cost function.
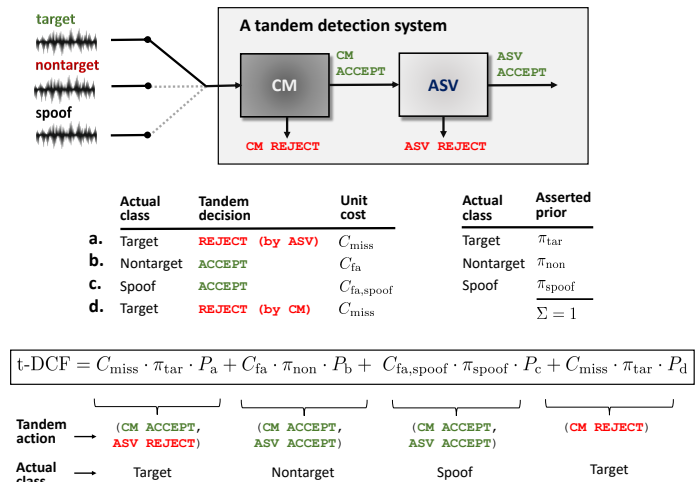
Fig. 1. A tandem system consisting of *automatic speaker verification* (ASV) and *spoofing countermeasure* (CM) modules is evaluated using three types of trials: targets, nontargets and spoofing attacks. *Tandem detection cost function* (t-DCF) is the average cost of erroneous decisions by the tandem system.

I. INTRODUCTION

BINARY classifiers (or detectors) are prone to two different types of errors, *misses* and *false alarms*. For biometric recognition systems such as *automatic speaker verification* (ASV) used for authentication, miss and false alarm rates

T. Kinnunen and V. Vestman are with the School of Computing, University of Eastern Finland, Länsikatu 15, FI-80101 Joensuu, Finland. E-mail: {tomi.kinnunen,ville.vestman}@uef.fi.

H. Delgado is with Nuance Communications Inc., 39 Gran Vía, 28013 Madrid, Spain. E-mail: hector.delgado@nuance.com. Part of his contribution to this work was done during his previous role at EURECOM, France.

N. Evans, A. Nautsch and M. Todisco are with EURECOM, Campus SophiaTech, 450 Route des Chappes, 06410 Biot, France. E-mail: {evans, nautsch, todisco}@eurecom.fr

K.A. Lee is with NEC Corporation, 7-1, Shiba 5-chome Minato-ku, Tokyo 108-8001, Japan. E-mail: k-lee@ax.jp.nec.com

X. Wang and J. Yamagishi are with National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan. E-mail: {wangxin@nii.ac.jp,jyamagis}@nii.ac.jp

M. Sahidullah is with the Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France. E-mail: md.sahidullah@inria.fr

D.A. Reynolds is with MIT Lincoln Laboratory, Massachusetts Institute of Technology, 244 Wood Street, Lexington, MA 02421-6426, E-mail: dar@ll.mit.edu

are proxies for user convenience and security, respectively. User convenience and security are competing requirements. The compromise between them will depend upon the application; whereas an online banking application might call for high security, user convenience might be key to a successful smarthome application. The approach to assessment must hence reflect application considerations. One such approach to assessment, used for the standard ASV evaluation benchmarks run by the *National Institute of Standards and Technology* (NIST) in the US since 1996 [1], is the *detection cost function* (DCF) [2].

The DCF reflects the cost of decisions in a *Bayes risk* sense [3], [4] and was designed for the assessment of ASV systems with a protocol involving a mix of *target* and *non-target* trials. The latter are casual impostors who make no effort to fool the ASV system. This paper concerns the assessment of ASV systems in the face of fake, falsified or spoofed inputs, also referred to as *presentation attacks* [5]. These are specially crafted inputs that are used by a fraudster to deceive an ASV system and hence to provoke false alarms. Just like all other biometrics systems, ASV systems can be vulnerable to spoofing [6]. The usual means to defend against such attacks involves the coupling of ASV systems with spoofing *countermeasures* (CMs) [7], namely sub-systems designed

to distinguish not between target and non-target trials, but between genuine, human or bona fide speech and artificially generated or manipulated inputs.

Since 2013, research in anti-spoofing for ASV has been spearheaded through the community-led ASVspoof initiative. It has produced three, well-supported competitive challenges. The first two editions held in 2015 and 2017 used an *equal error rate* (EER) metric to assess the performance of *isolated* spoofing CMs. Just like ASV systems, CMs are binary classifiers and; they make two types of errors that also have different consequences, depending on the application. The approach to assessment should hence reflect differences in the cost of each type of error. Furthermore, CMs are always used in combination with ASV. Previous work has shown the potential to combine the action of ASV and CM systems in the form of a single, *integrated* system [8], by the back-end *fusion* of independently trained ASV and CM systems [9], [10], or via the *tandem* detection framework illustrated in Fig. 1. Each approach shares the common goal of protecting ASV systems from being deceived by spoofed inputs. Since CMs are never used alone, the use of EER to assess the performance of spoofing CMs is hence questionable. We need better approches to assess CM performance, specifically approaches which reflect application requirements.

In order to preserve trust, ASV systems used for authentication in sensitive applications should have the capacity to defend against spoofing. In this sense, research in ASV and anti-spoofing is inextricably intertwined. Unfortunately, though, the two communities are today somewhat disjointed. One explanation for this situation might stem from the disparity between the metrics used in each field. It is difficult to argue, though, that the two communities should not work together, for they share the same goal to develop ever-more reliable ASV technology. In trying to foster closer collaboration, the work presented in this paper explores how the infrastructure and metrics developed for the assessment of ASV can be adopted for the joint assessment of ASV and spoofing CMs. It presents the *tandem detection cost function* (t-DCF) approach to assess the performance of tandem CM and ASV systems of the form illustrated in Fig. 1.

A preliminary version of this work was presented in [11]. The current work extends it in a number of respects. First, we revisit the t-DCF considering a reduced set of five evaluation parameters (rather than seven). Second, in contrast to the *unconstrained* t-DCF formulation for which both ASV and CM system thresholds can be varied, we present the *ASV-constrained* t-DCF which shows how a CM should be optimised for a given, fixed ASV system with known miss, false alarm and *spoof false alarm* (SFAR) rates. Third, we present a numerical simulation of the t-DCF which sheds light upon its behaviour and helps to interpret results. Fourth, whereas [11] presented results for ASVspoof 2015 and 2017 datasets, the current work presents new results and analysis for the most recent ASVspoof 2019 datasets. Finally, the paper contains several new clarifications and further examples not presented in [11]. The material is intended as a self-contained and accessible introduction for both experts and non-experts. While the paper relates to ASV, it should be of interest to the broader biometrics community, where work in anti-spoofing evaluation remains characterised by the use of *ad-hoc* bases rather than an application-targeted metric.

II. THE TWO SYSTEMS AND THEIR TANDEM COMBINATION

Both *automatic speaker verification* (ASV) and *spoofing countermeasure* (CM) systems are binary classifiers. An ASV trial consists of an enrollment-test utterance pair $(\mathcal{X}_e, \mathcal{X}_t)$ where $\mathcal{X}_e$ is collected at the enrollment stage and $\mathcal{X}_t$ at the verification stage. A pair $(\mathcal{X}_e, \mathcal{X}_t)$ with matched speaker identities is known as a *target* trial, otherwise as a *non-target* trial. The ASV system propositions (hypotheses) are hence given by:

$$\begin{cases} H_0^{\text{asv}} \, (\textbf{target} \text{ hypothesis}) : \; \text{id}(\mathcal{X}_e) = \text{id}(\mathcal{X}_t) \\ H_1^{\text{asv}} \, (\textbf{nontarget} \text{ hypothesis}) : \; \text{id}(\mathcal{X}_e) \neq \text{id}(\mathcal{X}_t), \end{cases}$$

where $\text{id}(\mathcal{X}) \in \mathbb{N} = \{1, 2, \dots\}$ is the unique speaker identity (a categorical variable) of utterance $\mathcal{X}$. The ASV system can encounter spoofed trials too. It is because the ASV system is assumed to have limited (or no) capacity to reject spoofs that dedicated CMs are needed.

The CM operates only upon the test utterance $\mathcal{X}_t$ and aims to verify its authenticity. If it corresponds to genuine speech produced by a human speaker, then the test upon $\mathcal{X}_t$ performed by the CM is referred to as a *bona fide* trial. If it corresponds to non-genuine, manipulated or synthesized speech, then it is referred to as a *spoof* trial. The CM propositions are hence given by:

$$\begin{cases} H_0^{\text{cm}} \, (\textbf{bona fide} \text{ hypothesis}) : \mathcal{X}_t \text{ is bona fide speech} \\ H_1^{\text{cm}} \, (\textbf{spoof} \text{ hypothesis}) : \mathcal{X}_t \text{ is spoofing attack}. \end{cases}$$

The CM system is designed to distinguish bona fide from spoof trials. In the same way that the ASV system has limited capacity to reject spoofing attacks, the CM is assumed to have limited capacity to distinguish target from nontarget trials; both are bona fide. The ASV and CM systems play complementary roles and both are needed to ensure spoofing-robust ASV.

We define the **tandem system** as a cascade of CM and ASV systems, as illustrated in Fig. 1. The CM acts as a *gate* which aims to prevent spoofing attacks from reaching the ASV system. Conventional ASV systems can also be regarded as tandem systems with a dummy 'accept all' CM [11]. Accordingly, the work presented applies also to the analysis of conventional ASV systems. *Internally* the tandem system consists of two subsystems that act together (and whose errors combine). To *end users* the tandem system acts as a single ASV system that either accepts or rejects their identity claim; the tandem system should therefore be viewed as a spoofing-robust ASV system with a CM 'under the hood'. The tandem system can encounter three different types of trials: (i) *target*, (ii) *nontarget* and (iii) *spoof*. It should accept only the target trials. Both nontarget and spoof trials should be rejected.

III. TANDEM DETECTION COST FUNCTION

The *tandem detection cost function* (t-DCF) [11] metric reflects the performance of a combined ASV and CM system

TABLE I
THE INGREDIENTS OF T-DCF. THE SYSTEM ILLUSTRATED IN FIG. 1 CAN FACE THREE TYPES OF TRIALS (FIRST COLUMN): *target*, *nontarget* AND *spoof*, EACH WITH SOME ASSERTED **PRIOR** (SECOND COLUMN). THE SYSTEM EITHER REJECTS OR ACCEPTS THE TRIAL BY TAKING A **TANDEM ACTION** (THIRD COLUMN), WITH NONNEGATIVE **DETECTION COSTS** (FOURTH COLUMN) ASSIGNED TO EACH ERROR (0: NO LOSS). **DETECTION ERROR RATE** OF THE TANDEM SYSTEM (LAST COLUMN) IS OBTAINED BY TREATING THE CM AND ASV SYSTEM DECISIONS INDEPENDENT. THE T-DCF METRIC IS OBTAINED BY MULTIPLYING THE PRIOR (*how often we expect to observe this class?*), COST (*if this error happens, how much does it cost?*) AND THE ERROR RATE (*how many errors were actually observed?*) OF EACH ROW AND SUMMING UP THE TERMS.

Actual class	Class prior	**Tandem action** = (CM action, ASV action)	Detection Cost	Detection error rate
Target	π_{tar}	$\alpha_1 = \textbf{REJECT} = (\text{CM ACCEPT}, \text{ASV REJECT})$	C_{miss}	$(1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})) \times P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}})$
		$\alpha_2 = \textbf{ACCEPT} = (\text{CM ACCEPT}, \text{ASV ACCEPT})$	0	–
		$\alpha_3 = \textbf{REJECT} = (\text{CM REJECT})$	C_{miss}	$P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})$
Nontarget	π_{non}	$\alpha_1 = \textbf{REJECT} = (\text{CM ACCEPT}, \text{ASV REJECT})$	0	–
		$\alpha_2 = \textbf{ACCEPT} = (\text{CM ACCEPT}, \text{ASV ACCEPT})$	C_{fa}	$(1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})) \times P_{\text{fa}}^{\text{asv}}(\tau_{\text{asv}})$
		$\alpha_3 = \textbf{REJECT} = (\text{CM REJECT})$	0	–
Spoof	π_{spoof}	$\alpha_1 = \textbf{REJECT} = (\text{CM ACCEPT}, \text{ASV REJECT})$	0	–
		$\alpha_2 = \textbf{ACCEPT} = (\text{CM ACCEPT}, \text{ASV ACCEPT})$	$C_{\text{fa,spoof}}$	$P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}) \times P_{\text{fa,spoof}}^{\text{asv}}(\tau_{\text{asv}})$
		$\alpha_3 = \textbf{REJECT} = (\text{CM REJECT})$	0	–

(a) Automatic speaker verification (ASV).

(b) Spoofing countermeasure (CM).

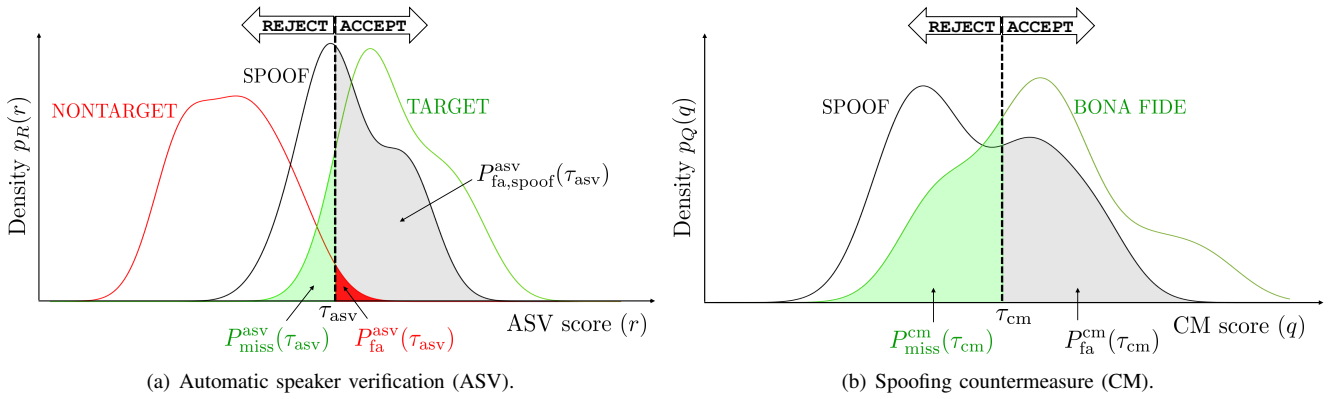Fig. 2. The ASV and CM score density functions $p_R(r)$ and $p_Q(q)$, respectively. Fixing the detection thresholds of ASV (τ_{asv}) and CM (τ_{cm}) yields five different types of error rates: miss and false alarm rates of each system and spoof false alarm rate of the ASV system. The error rates of ASV and CM are multiplied to form tandem error rates (see Table I).

for an assumed operating environment (application). Unlike the EER, the t-DCF is a *parametric* function which requires the specification of application parameters in advance.

A. Detection Costs and Priors (the Application)

Consider a hypothetical 'banking' scenario in which customer authentication is controlled using voice biometrics. Access should be restricted to target users (account owners), while nontarget (zero-effort impostor) and spoofed (dedicated impostor) access attempts should always be denied; the system should be *secure*. Access by customers should always be granted and never denied; they should not be *inconvenienced*. The competing requirements for security and convenience cannot both be satisfied, leading to the potential for detection errors. To each error is associated a monetary loss (*e.g.* loss of funds to fraud, or loss of customers to inconvenience).

In seeking to minimise its costs, the bank will assign a higher penalty to the more costly errors. This is formalized through the specification of *detection costs*, $C(\alpha|\theta) \geq 0$, interpreted as the penalty of taking *action* α (making a decision for a given test trial) when the actual class is θ [3], [4]. Correct decisions are assigned a cost of 0 while erroneous decisions are assigned a positive numerical value,

which signifies the monetary loss to the bank incurred as a result of each type of detection error.

By denoting the class variable by $\theta \in \Theta \equiv \{\theta_{\text{tar}}, \theta_{\text{non}}, \theta_{\text{spoof}}\}$ and the action by $\alpha \in \mathcal{A} \equiv \{\texttt{ACCEPT}, \texttt{REJECT}\}$, we define the following three detection costs:

- $C_{\text{miss}} \equiv C(\texttt{REJECT}|\theta_{\text{tar}})$ – cost of rejecting a target trial;
- $C_{\text{fa}} \equiv C(\texttt{ACCEPT}|\theta_{\text{non}})$ – cost of accepting a nontarget trial;
- $C_{\text{fa,spoof}} \equiv C(\texttt{ACCEPT}|\theta_{\text{spoof}})$ – cost of accepting a spoofed trial,

where the first two correspond to the familiar notations used in NIST *speaker recognition evaluation* (SRE) campaigns [1], [2]. The third cost is specific to the new class of spoofing attacks. It is stressed that actions are *those of the tandem system*. This is different to [11] where costs are specified *per subsystem*, but error rates are those of the tandem system.

The first type of tandem error occurs when either the CM or ASV system rejects a target; the second case occurs when both systems accept a nontarget; the last case occurs when both systems accept a spoofing attack. As displayed in Table I, these four cases cover all the possible errors. The remaining five cases lead to correct tandem decisions and are therefore assigned zero cost. Note the curious cases of nontargets being rejected by the CM system (6th row), and spoofing attacks

rejected by the ASV system (7th row). These trials are rejected by the 'wrong' subsystem but neither incurs loss as the tandem action is correct.

In addition to assigning detection costs for each type of error, one must also take into account the relative occurences of the three classes (target, nontarget, spoof). A relatively expensive error that occurs only infrequently may cost less than a relatively inexpensive error that occurs more frequently. For instance, if the bank expects 99% of authentication requests to originate from bona fide account holders, then the expected monetary loss incurred from bona fide customers (target trials) being denied access to their accounts may override the costs incurred from access being granted erroneously to fraudsters (zero-effort or spoofed trials), who account for only 1% of authentication requests. The assumed commonality of each trial class is encoded in their *prior probabilities*:

- $\pi_{\text{tar}} \equiv P_\Theta(\theta_{\text{tar}})$ – prior probability of target;
- $\pi_{\text{non}} \equiv P_\Theta(\theta_{\text{non}})$ – prior probability of nontarget;
- $\pi_{\text{spoof}} \equiv P_\Theta(\theta_{\text{spoof}})$ – prior probability of spoofing attack,

where $P_\Theta(\theta)$ is a shorthand for $P_\Theta(\Theta = \theta)$, Θ and θ being a random variable and its realization, respectively.

The priors are nonnegative and sum to unity (therefore, fixing any two priors automatically defines the third). The prior is *subjective* — it asserts the belief of the relative frequency of each trial class during the operation of a spoofing-robust ASV (with the actual, empirical class frequencies remaining unknown). The priors of the cost function do not have to (and typically do not) correspond to the empirical trial frequencies in training or evaluation corpora. The costs and priors are set in advance and they remain fixed within a given evaluation or application setting; they might be very different for, *e.g.*, banking, forensics or surveillance applications. The set of evaluation metric parameters are hence given by $\Psi_{\text{t-DCF}} \equiv (\pi_{\text{tar}}, \pi_{\text{spoof}}, C_{\text{miss}}, C_{\text{fa}}, C_{\text{fa,spoof}})$, where the nontarget prior is omitted and obtained from $\pi_{\text{non}} = 1 - \pi_{\text{tar}} - \pi_{\text{spoof}}$.

B. Detection Error Rates of ASV and CM

Let $r = \text{ASV}(\mathcal{X}_{\text{e}}, \mathcal{X}_{\text{t}})$ and $q = \text{CM}(\mathcal{X}_{\text{t}})$ denote ASV and CM scores[1], treated here as realizations of random variables that admit continuous probability density functions $p_R(r)$ and $p_Q(q)$. That is, $p_R(r) \geq 0$, $\int_{-\infty}^{\infty} p_R(r)\,\mathrm{d}r = 1$ (similarly for $p_Q(q)$). The testing of the ASV system with all three types of trials leads to detection scores drawn from the respective class-conditional distributions:

$$p_R(r|\theta_{\text{tar}}) \quad \text{ASV target score distribution;}$$
$$p_R(r|\theta_{\text{non}}) \quad \text{ASV nontarget score distribution;} \quad (1)$$
$$p_R(r|\theta_{\text{spoof}}) \quad \text{ASV spoof score distribution.}$$

Likewise, testing of the CM with trials of bona fide and spoof classes leads to detection scores drawn from the conditional distributions

$$p_Q(q|\theta_{\text{bona}}) \quad \text{CM bona fide score distribution;}$$
$$p_Q(q|\theta_{\text{spoof}}) \quad \text{CM spoof score distribution,} \quad (2)$$

where we introduced θ_{bona} as a realization of a new random variable $\Theta_{\text{bona}} \equiv \{\theta_{\text{tar}}\} \cup \{\theta_{\text{non}}\} \subset \Theta$, a container of any *non-spoof* trials (whether target or nontarget). To obtain the distribution of the bona fide scores, consider the joint distribution of Q and Θ_{bona}, $P(q, \theta_{\text{bona}}) = P(\theta_{\text{bona}})p(q|\theta_{\text{bona}})$, obtained using the *product rule* [12, Eq. (1.11)] (subscripts omitted for brevity). By treating Θ_{bona} as a *latent variable*, the bona fide score distribution is obtained by marginalizing the class variable out, using the *sum rule* [12, Eq. (1.10)]:

$$
\begin{aligned}
p_Q(q|\theta_{\text{bona}}) &= \sum_{\theta \in \Theta_{\text{bona}}} P(q, \theta) = \sum_{\theta \in \Theta_{\text{bona}}} P_{\Theta_{\text{bona}}}(\theta)p_Q(q|\theta) \\
&= P_{\Theta_{\text{bona}}}(\theta_{\text{tar}})p_Q(q|\theta_{\text{tar}}) + P_{\Theta_{\text{bona}}}(\theta_{\text{non}})p_Q(q|\theta_{\text{non}}) \\
&= \tilde{\pi}_{\text{tar}}p_Q(q|\theta_{\text{tar}}) + (1 - \tilde{\pi}_{\text{tar}})p_Q(q|\theta_{\text{non}}),
\end{aligned}
$$
(3)

which is a two-component *mixture distribution* where $\tilde{\pi}_{\text{tar}} \equiv P_{\Theta_{\text{bona}}}(\theta_{\text{tar}})$ is the relative proportion of target trials within the bona fide class (known by the corpus designer, but not necessarily by the evaluee)[2]. While (3) represents the general form of the bona fide score distribution, the target and nontarget score distributions are typically highly overlapped, as speaker-independent CMs are usually not designed to discriminate between them. In the limiting case when the two become indistinguishable, *i.e.* $p_Q(q|\theta_{\text{tar}}) = p_Q(q|\theta_{\text{non}})$, (3) collapses either to the target or nontarget distribution, making *bona fide* an unnecessary relabeling of the original class labels.

As illustrated in Fig. 2, ASV and CM systems convert detection scores r and q into hard binary decisions by comparing their values to preset detection thresholds denoted by τ_{asv} and τ_{cm}, respectively. The ASV system takes an ACCEPT action if and only if $r > \tau_{\text{asv}}$ (otherwise REJECT) whereas the CM takes an ACCEPT action if and only if $q > \tau_{\text{cm}}$ (otherwise REJECT). Score distributions in (1) and (2) combined with fixed decision thresholds τ_{asv} and τ_{cm} yield a set of five error rates illustrated in Fig. 2:

$$
\begin{aligned}
P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}}) &= \int_{-\infty}^{\tau_{\text{asv}}} p_R(r|\theta_{\text{tar}})\,\mathrm{d}r && \text{ASV miss rate;} \\
P_{\text{fa}}^{\text{asv}}(\tau_{\text{asv}}) &= \int_{\tau_{\text{asv}}}^{\infty} p_R(r|\theta_{\text{non}})\,\mathrm{d}r && \text{ASV false alarm rate;} \\
P_{\text{fa,spoof}}^{\text{asv}}(\tau_{\text{asv}}) &\equiv \int_{\tau_{\text{asv}}}^{\infty} p_R(r|\theta_{\text{spoof}})\,\mathrm{d}r && \text{ASV spoof false alarm rate;} \\
P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) &= \int_{-\infty}^{\tau_{\text{cm}}} p_Q(q|\theta_{\text{bona}})\,\mathrm{d}q && \text{CM miss rate;} \\
P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}) &= \int_{\tau_{\text{cm}}}^{\infty} p_Q(q|\theta_{\text{spoof}})\,\mathrm{d}q && \text{CM false alarm rate.}
\end{aligned}
$$
(4)

The first and last two components are the usual miss and false alarm rates of each system, while the third **defines** the *spoof false alarm rate* (SFAR) of the ASV system as the proportion of spoofing attacks accepted by the ASV system[3].

---

[1]Here $\text{ASV}(\cdot, \cdot)$ and $\text{CM}(\cdot)$ denote the two system operations as 'black-boxes' that gives us the detection scores. Usually r and q are the *logarithm of the likelihood ratio* (LLR) of the respective null and alternative hypothesis likelihoods obtained from a parametric model (though the proposed metric does not require a LLR interpretation of scores).

[2]Again, this is not necessarily the same as the proportion dictated by the t-DCF priors, *i.e.* $\pi_{\text{tar}}/(\pi_{\text{tar}} + \pi_{\text{non}})$.

[3]Here spoofs are treated as the negative class (similar to nontargets) but this convention is actually arbitrary; as we did in [11], they could also be defined as a positive class (similar to targets) leading to an equivalent definition of SFAR as the probability of 'ASV does NOT miss a spoof'.

Since the probability density functions are unknown, practical computations must be performed using empirical detection scores $r_i \sim p_R(r)$ and $q_j \sim p_Q(q)$, where i and j index ASV and CM trials respectively. Assuming that the scores produced by each system are independent and identically distributed (i.i.d.) draws from the respective distribution, integrals in (4) may be replaced by summations, illustrated here for the ASV miss rate:

$$P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}}) = \mathbb{E}_{r \sim p_R(r|\theta_{\text{tar}})}\left[\mathbb{I}(r < \tau_{\text{asv}})\right] \approx \frac{1}{N_{\text{tar}}} \sum_{i \in \Lambda_{\text{tar}}} \mathbb{I}(r_i < \tau_{\text{asv}}), \quad (5)$$

where $\mathbb{E}_{z \sim p_Z(z)}[f(z)]$ denotes the expected value of function $f(z)$ w.r.t. distribution $p_Z(z)$, $\mathbb{I}(\cdot)$ is an indicator function that equals 1 for a true proposition and 0 otherwise, Λ_{tar} indices the target trials, and $N_{\text{tar}} = |\Lambda_{\text{tar}}|$ denotes the total number of target trials. The other four error rates are computed similarly. The CM miss rate is computed by pooling CM target and nontarget scores, as both score sets are viewed as samples from the same bona fide score distribution (3).

C. The t-DCF and Its Relation to the NIST DCF

The t-DCF is a measure of the expected (average) cost of all errors made by a tandem system. The following treatment assumes the cascaded setup illustrated in Fig. 1. There are three possible *actions* (tandem decisions), $\alpha_i \in \mathcal{A} = \{\alpha_1, \alpha_2, \alpha_3\}$:

$$\alpha_1 = (\text{CM ACCEPT, ASV REJECT})$$
$$\alpha_2 = (\text{CM ACCEPT, ASV ACCEPT}) \quad (6)$$
$$\alpha_3 = (\text{CM REJECT}),$$

where the last case indicates rejection of a trial at the CM stage. In this case, the ASV action is null or undefined (referred to in [11] as a dummy SLEEP action). As Table I indicates, five of the nine possible (ground truth, action) combinations lead to correct tandem decisions. The remaining four terms all constitute errors:

(a) $(\alpha_1, \theta_{\text{tar}})$: the CM does not miss a bona fide trial, but the ASV system misses a target.

(b) $(\alpha_2, \theta_{\text{non}})$: the CM does not miss a bona fide trial, but the ASV system falsely accepts a nontarget.

(c) $(\alpha_2, \theta_{\text{spoof}})$: both CM and ASV systems falsely accept a spoof.

(d) $(\alpha_3, \theta_{\text{tar}})$: the CM misses a target, which is never processed by the ASV system.

If CM and ASV error probabilities are assumed to be independent (discussed further in Subsection VIII-A), then the probability of each of these four outcomes may be expressed in terms of the error rates in (4) as follows:

$$P_{\text{a}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = (1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})) \times P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}})$$
$$P_{\text{b}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = (1 - P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})) \times P_{\text{fa}}^{\text{asv}}(\tau_{\text{asv}})$$
$$P_{\text{c}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}) \times P_{\text{fa,spoof}}^{\text{asv}}(\tau_{\text{asv}}) \quad (7)$$
$$P_{\text{d}}(\tau_{\text{cm}}, \tau_{\text{asv}}) = P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}).$$

Finally, the t-DCF is obtained by multiplying the class prior, cost and error terms in each non-zero row (rows with nonzero entry in the 4th column) of Table I and summing up the resulting four terms:

Unconstrained t-DCF

$$\text{t-DCF}(\tau_{\text{cm}}, \tau_{\text{asv}}) =$$
$$C_{\text{miss}} \cdot \pi_{\text{tar}} \cdot [P_{\text{a}}(\tau_{\text{cm}}, \tau_{\text{asv}}) + P_{\text{d}}(\tau_{\text{cm}}, \tau_{\text{asv}})]$$
$$+ C_{\text{fa}} \cdot \pi_{\text{non}} \cdot P_{\text{b}}(\tau_{\text{cm}}, \tau_{\text{asv}})$$
$$+ C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}} \cdot P_{\text{c}}(\tau_{\text{cm}}, \tau_{\text{asv}}), \quad (8)$$

---

where the three lines correspond to target, nontarget and spoof related errors, respectively. Eq. (8) is referred to as the 'unconstrained' t-DCF to distinguish this formulation from a special case discussed shortly.

It is readily seen that the NIST DCF [2],

$$\text{DCF}_{\text{NIST}}(\tau_{\text{asv}}) \equiv C_{\text{miss}} \pi_{\text{tar}} P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}}) + C_{\text{fa}}(1 - \pi_{\text{tar}}) P_{\text{fa}}^{\text{asv}}(\tau_{\text{asv}}), \quad (9)$$

is a special case of the t-DCF by assuming (a) the use of a dummy 'accept all' CM ($\tau_{\text{cm}} = -\infty \Rightarrow P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) = 0, P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}) = 1$) **and** (b) there are no spoofing attacks ($\pi_{\text{spoof}} = 0$). In this sense, the NIST DCF could be considered as an *optimistic* t-DCF. Even if the NIST DCF has been instrumental to developments in the ASV field, it may not be well suited to operational environments where there is potential for spoofing attacks.

D. ASV-Constrained t-DCF

In the unconstrained t-DCF (8) both CM and ASV are adjustable, which can make evaluation of the tandem system cumbersome. Here we explore the t-DCF from the perspective of a CM developer who cannot interact with the ASV system (or has no capacity to develop one). The ASVspoof 2019 challenge is representative of such a scenario. The evaluee focuses instead on a special case, **ASV-constrained** t-DCF, where the only information known about the ASV system (a black-box) are the three error rates $P_{\text{miss}}^{\text{asv}} \equiv P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}})$, $P_{\text{fa}}^{\text{asv}} \equiv P_{\text{fa}}^{\text{asv}}(\tau_{\text{asv}})$ and $P_{\text{miss,spoof}}^{\text{asv}} \equiv P_{\text{miss,spoof}}^{\text{asv}}(\tau_{\text{asv}})$ provided by another party (*e.g.*, ASV vendor or challenge organizer).

As the main difference between the two t-DCF variants is whether we can adjust the ASV threshold or not, we use the overloaded notation t-DCF(τ_{cm}) to indicate the ASV-constrained t-DCF. With straightforward manipulation, the t-DCF expression of (8) can then be rewritten as:

ASV-constrained t-DCF

$$\text{t-DCF}(\tau_{\text{cm}}) = C_0 + C_1 P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) + C_2 P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}), \quad (10)$$

---

where C_0, C_1, and C_2 are constants dictated both by the t-DCF parameters and the ASV error rates. They are given by:

$$C_0 = \pi_{\text{tar}} C_{\text{miss}} P_{\text{miss}}^{\text{asv}} + \pi_{\text{non}} C_{\text{fa}} P_{\text{fa}}^{\text{asv}}$$
$$C_1 = \pi_{\text{tar}} C_{\text{miss}} - (\pi_{\text{tar}} C_{\text{miss}} P_{\text{miss}}^{\text{asv}} + \pi_{\text{non}} C_{\text{fa}} P_{\text{fa}}^{\text{asv}}) \quad (11)$$
$$C_2 = \pi_{\text{spoof}} C_{\text{fa,spoof}} P_{\text{fa,spoof}}^{\text{asv}}.$$

We present an analysis of these coefficients in detail below. First, however, we explain the necessity to *normalize* the raw t-DCF values (whether unconstrained or ASV-constrained case).

IV. NORMALIZED AND MINIMUM T-DCF

Both the individual (4) and the tandem (7) error rates take values in $[0, 1]$. This is not the case for the t-DCF in (8), however, which is a linear combination of the tandem errors formed by non-negative but otherwise unconstrained multipliers (the products of costs and priors). The 'raw' t-DCF values can hence be difficult to interpret, especially across different t-DCF parametrizations. Normalization is performed differently depending on whether one focuses on the unconstrained case (both CM and ASV systems are adjustable) or the ASV-constrained case (only the CM system is adjustable). Let us first focus on the unconstrained case.

A. Normalizing the Unconstrained t-DCF

Following the practice adopted in the NIST SREs [1], [2], it is preferable to report the *normalized* t-DCF given by:

$$\text{t-DCF}'(\tau_{\text{cm}}, \tau_{\text{asv}}) = \frac{\text{t-DCF}(\tau_{\text{cm}}, \tau_{\text{asv}})}{\text{t-DCF}_{\text{default}}^{\text{unconstr}}}, \qquad (12)$$

where t-DCF$'$ denotes the normalized cost and $\text{t-DCF}_{\text{default}}^{\text{unconstr}} > 0$ is the t-DCF of a *default* (*reference*) system that yields a fixed decision regardless of input data. The default system either accepts every user, or rejects every user. As an intuitive analogy, the reader may picture a door lock that will either open with any key (inluding that of a burglar), or with no key in the world (including the owner's key). Any useful lock should do better work than either one of these two default options.

In a similar vein, any useful tandem system should yield a lower cost than that of *both* 'accept all' and 'reject all' default systems. The former corresponds to action α_2 in (6). When both thresholds are set to $-\infty$, the two miss rates $P_{\text{miss}}^{\text{asv}}(\tau_{\text{asv}})$ and $P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}})$, and hence also $P_{\text{a}}(\tau_{\text{cm}}, \tau_{\text{asv}})$ and $P_{\text{d}}(\tau_{\text{cm}}, \tau_{\text{asv}})$ in (7) all reduce to zero. $P_{\text{b}}(\tau_{\text{cm}}, \tau_{\text{asv}})$ and $P_{\text{c}}(\tau_{\text{cm}}, \tau_{\text{asv}})$ reduce to one, giving:

$$\begin{aligned} \text{t-DCF}_{\alpha_2} &= \text{t-DCF}(-\infty, -\infty) \\ &= C_{\text{fa}} \cdot \pi_{\text{non}} + C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}}, \end{aligned} \qquad (13)$$

in which there are no target speaker parameters (as there are no misses). In similar fashion and depending on whether each trial is rejected by the ASV system (action α_1) or by the CM (action α_3), the 'reject all' default systems are given by:

$$\begin{aligned} \text{t-DCF}_{\alpha_1} &= \text{t-DCF}(-\infty, \infty) = C_{\text{miss}} \cdot \pi_{\text{tar}} \\ \text{t-DCF}_{\alpha_3} &= \text{t-DCF}(\infty, \tau_{\text{asv}}) = C_{\text{miss}} \cdot \pi_{\text{tar}}, \quad \forall \, \tau_{\text{asv}} \in \mathbb{R} \end{aligned} \qquad (14)$$

which contains neither nontarget nor spoof terms (both types of trials are correctly rejected). The equality $\text{t-DCF}_{\alpha_1} = \text{t-DCF}_{\alpha_3}$, in turn, reinforces the idea that it does not matter whether it was the CM or the ASV which rejected the target — it was rejected, and in both cases the user experiences the same inconvenience.

A useful tandem system should have lower t-DCF than both of the dummy systems (13) and (14). That is, it should yield

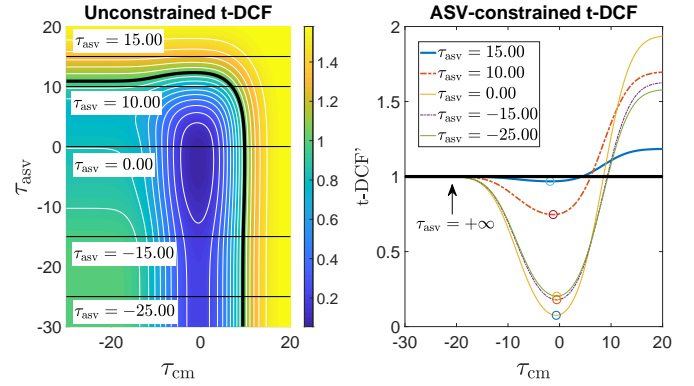| Type of t-DCF	Raw form	Normalized form	Min. value
Unconstrained	(8)	(8)/(15)	0
ASV-constrained	(10)	(10)/(18)	C_0

Fig. 3. Illustration of unconstrained (left) and ASV-constrained t-DCFs (right) using simulated ASV and CM scores. The former involves both thresholds, while the latter considers 'slices' defined by a fixed ASV operating point. Both variants are displayed in their normalized forms (See Table II). The isocontour of t-DCF$' = 1$ is highlighted in each case.

a cost lower than their *minimum*. The default system is hence chosen according to:

$$\begin{aligned} \text{t-DCF}_{\text{default}}^{\text{unconstr}} &= \min \left\{ \text{t-DCF}_{\alpha_1}, \text{t-DCF}_{\alpha_2}, \text{t-DCF}_{\alpha_3} \right\} \\ &= \min \left\{ \text{t-DCF}_{\alpha_1}, \text{t-DCF}_{\alpha_2} \right\} \\ &= \min \left\{ C_{\text{fa}} \cdot \pi_{\text{non}} + C_{\text{fa,spoof}} \cdot \pi_{\text{spoof}}, C_{\text{miss}} \cdot \pi_{\text{tar}} \right\}, \end{aligned} \qquad (15)$$

where the second line follows from (14). Note, however, that the normalized t-DCF obtained by dividing (8) by (15) is *not* an upper bound. With poorly set detection thresholds (alternatively, using Bayes-optimal thresholds but with badly calibrated scores [13]), the normalized cost can exceed 1; it can be higher than that of the default system. Such systems are said to be *badly-calibrated*.

An **optimally calibrated** system provides another useful reference. This *minimum* t-DCF is defined as the minimum cost over all thresholds $(t_{\text{cm}}, t_{\text{asv}}) \in \mathbb{R}^2$,

$$\text{t-DCF}_{\text{min}} = \inf_{(t_{\text{cm}}, t_{\text{asv}})} \text{t-DCF}(t_{\text{cm}}, t_{\text{asv}}), \qquad (16)$$

where the infimum (greatest lower bound) is replaced by min for finite score sets. By definition, $\text{t-DCF}(\tau_{\text{cm}}, \tau_{\text{asv}}) \geq \text{t-DCF}_{\text{min}}$ for any choice of the thresholds (including those of the 'default' tandem system). Thus, the normalized minimum cost t-DCF$'_{\text{min}}$, is upper bounded by unity:

$$\text{t-DCF}'_{\text{min}} = \frac{\text{t-DCF}_{\text{min}}}{\text{t-DCF}_{\text{default}}} \leq \frac{\text{t-DCF}_{\text{min}}}{\text{t-DCF}_{\text{min}}} = 1, \qquad (17)$$

making it a convenient number between 0 and 1. Like the EER metric, the minimum t-DCF uses an oracle threshold determined with use of ground-truth labels (trial key).

B. Normalizing the ASV-Constrained t-DCF

A normalised version of the ASV-constrained t-DCF, $\text{t-DCF}(\tau_{\text{cm}})' = \text{t-DCF}(\tau_{\text{cm}})/\text{t-DCF}_{\text{default}}^{\text{constr}}$, can similarly be defined by selecting an appropriate default cost $\text{t-DCF}_{\text{default}}^{\text{constr}} > 0$. Since the ASV system is now fixed, the default cost is obtained by adjusting the CM threshold only, with either $\tau_{\text{cm}} = -\infty$ (accept all) or $\tau_{\text{cm}} = +\infty$ (reject all) in (10), giving:

$$\begin{aligned} \text{t-DCF}_{\text{default}}^{\text{constr}} &= \min\{C_0 + C_1, C_0 + C_2\} \\ &= C_0 + \min\{C_1, C_2\}. \end{aligned} \quad (18)$$

C. A Summary of the Two t-DCF Variants

A summary of the two different t-DCF versions discussed above is given in Table II. For the unconstrained case, cost scaling is specified by the t-DCF parameters only. For the ASV-constrained case, it is also dependent on the known, fixed ASV error rates. In both cases, however, normalized costs larger than 1 indicate that the system under consideration cannot do better than 'no system'. The 'system' differs between the two cases: for the unconstrained case, it is the (CM, ASV) tandem; for the ASV-constrained case, it is the CM only.

Both t-DCF variants are illustrated in Fig. 3 for simulated scores (see Appendix) with $P_{\text{e}}^{\text{asv}} = 0.01$ (ASV EER=1%), $P_{\text{e}}^{\text{cm}} = 0.02$ (CM EER=2%) and $\xi = 0.85$ (a parameter defined in Eq. (24) to model efficacy of spoofing attacks. The arbitrary value 0.85 used here is a proxy of highly effective spoofing attack, see Fig. 9). The t-DCF parameters are set as described in Section VI (with $\pi_{\text{spoof}} = 0.05$). Even in this idealized simulation, the resulting unconstrained t-DCF has a relatively complicated shape. We observe a valley near the origin $(0, 0)$ in the left panel of Fig. 3. Moving away from this 'sweet spot' yields increased costs. Whenever either threshold is too high/low, we approach the 'accept all' or 'reject all' cases which are generally suboptimal.

Focusing on the ASV-conditional case and comparing the t-DCF values across five arbitrary ASV operating points indicated in Fig. 3, the lowest min t-DCF is obtained for $\tau_{\text{asv}} = 0.00$ (which coincides with the EER operating point in our simulation). We also observe that (a) the t-DCF function flattens with increasing τ_{asv}, and (b) it reaches the value of 1 at one of the infinities (here, at $\tau_{\text{cm}} = -\infty$). These two properties hold for any ASV and CM system, and will be elaborated on below.

V. ANALYSIS OF THE ASV-CONSTRAINED T-DCF

We now provide some intepretation of the coefficients in (11). First, the offset $C_0 \geq 0$ is dubbed the *ASV floor* as it lower bounds (10) and vanishes only for a perfect ASV system ($P_{\text{miss}}^{\text{asv}} = P_{\text{fa}}^{\text{asv}} = 0$). Note that, even though C_0 resembles the NIST DCF (9), they are not the same; unlike for the NIST DCF, $\pi_{\text{tar}} + \pi_{\text{non}} \neq 1$ since some probability mass is assigned to π_{spoof}. C_1 and C_2 reflect the relative importance of the CM miss and false alarm rates in (10).

The coefficients in (11) may seem complicated. Some insight into their influence on the t-DCF can be gained by setting the ASV system to the EER operating point so that $P_{\text{miss}}^{\text{asv}} = P_{\text{fa}}^{\text{asv}} \equiv P_{\text{e}}^{\text{asv}}$, where $P_{\text{e}}^{\text{asv}}$ indicates the EER of the

ASV system. The choice of EER operating point may look arbitrary as it contradicts the t-DCF parameter specifications. Nonetheless, the EER serves as *tight upper bound* on the binary classifier Bayes error-rate [14, p.72]. By fixing the ASV system to the EER operating point, we mimic a miscalibrated ASV system which operates with the *worst possible* target-nontarget discrimination performance.

The coefficients in (11) can now be rewritten as functions of either the EER or the SFAR of the ASV system:

$$\begin{aligned} C_0(P_{\text{e}}^{\text{asv}}) &= \alpha P_{\text{e}}^{\text{asv}} \\ C_1(P_{\text{e}}^{\text{asv}}) &= \beta - \alpha P_{\text{e}}^{\text{asv}} \\ C_2(P_{\text{fa,spoof}}^{\text{asv}}) &= \gamma P_{\text{fa,spoof}}^{\text{asv}}, \end{aligned} \quad (19)$$

where $\alpha = \pi_{\text{tar}}C_{\text{miss}} + \pi_{\text{non}}C_{\text{fa}}$, $\beta = \pi_{\text{tar}}C_{\text{miss}}$, and $\gamma = \pi_{\text{spoof}}C_{\text{fa,spoof}}$ are constants. By substituting (19) to (10) we obtain:

$$\begin{aligned} \text{t-DCF}(\tau_{\text{cm}}) = \alpha P_{\text{e}}^{\text{asv}} &+ (\beta - \alpha P_{\text{e}}^{\text{asv}}) P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) \\ &+ \gamma P_{\text{fa,spoof}}^{\text{asv}} P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}}). \end{aligned} \quad (20)$$

The influence upon the t-DCF, or weight of CM misses (C_1) is a function of the ASV system accuracy (encoded in $P_{\text{e}}^{\text{asv}}$). The weight of CM false alarms (C_2) is a function of the ASV system sensitivity to spoofing attacks (encoded in $P_{\text{fa,spoof}}^{\text{asv}}$). Once the evaluation conditions and the performance of the unprotected ASV system in the same conditions is known, then the CM may be optimised using the t-DCF metric (20) tailored to the specific ASV system and evaluation conditions. Here we are not concerned *how* such optimization (involving generally non-differentiable functions due to hard error counting) should be performed — we are merely stating the objective.

Another way to analyze C_0, C_1 and C_2 is directly in terms of the ASV threshold. This is visualized in Fig. 4 both for simulated, Gaussian-distributed ASV scores (see Appendix) and real x-vector based ASV scores (see Section VI). The data used for drawing the graphs in the two panels are unrelated; the resemblance of simulated and real functions is indicative of their general shape. Also illustrated are the limits of each coefficient as $\tau_{\text{asv}} \to \pm\infty$. As seen, the coefficients are nonlinear functions of the ASV operating point. The '$\mathcal{X}$' shape formed by C_0 and C_1 is explained by the dependence $C_0(\tau_{\text{asv}}) + C_1(\tau_{\text{asv}}) = \pi_{\text{tar}}C_{\text{miss}}$ seen from Eq. (11). We now have the following interpretations:

- C_0 approaches the 'accept all' and 'reject all' dummy systems at $\tau_{\text{asv}} = -\infty$ and $\tau_{\text{asv}} = +\infty$, respectively. What remains in the ASV floor is either the nontarget ($\pi_{\text{non}}C_{\text{fa}}$) or the target ($\pi_{\text{tar}}C_{\text{miss}}$) term.
- C_1 at $\tau_{\text{asv}} = -\infty$ leads to a curious subtractive expression (which can also take negative values), $\pi_{\text{tar}}C_{\text{miss}} - \pi_{\text{non}}C_{\text{fa}}$. How so? The dummy ASV system is set to 'accept all'. The CM nonetheless offers potential to reject some trials. Subtraction of the nonnegative nontarget term *lowers* C_1, which in turn encourages a *higher* CM threshold. For $\tau_{\text{asv}} \to +\infty$, C_1 vanishes. Since the ASV system will always reject target trials, CM behaviour is irrelevant.
- C_2 is a nonincreasing function of τ_{asv}. For the 'accept all' ASV at $\tau_{\text{asv}} \to -\infty$, the CM has the tightest security (highest C_2). Similar to C_1, C_2 vanishes at $\tau_{\text{asv}} \to +\infty$:
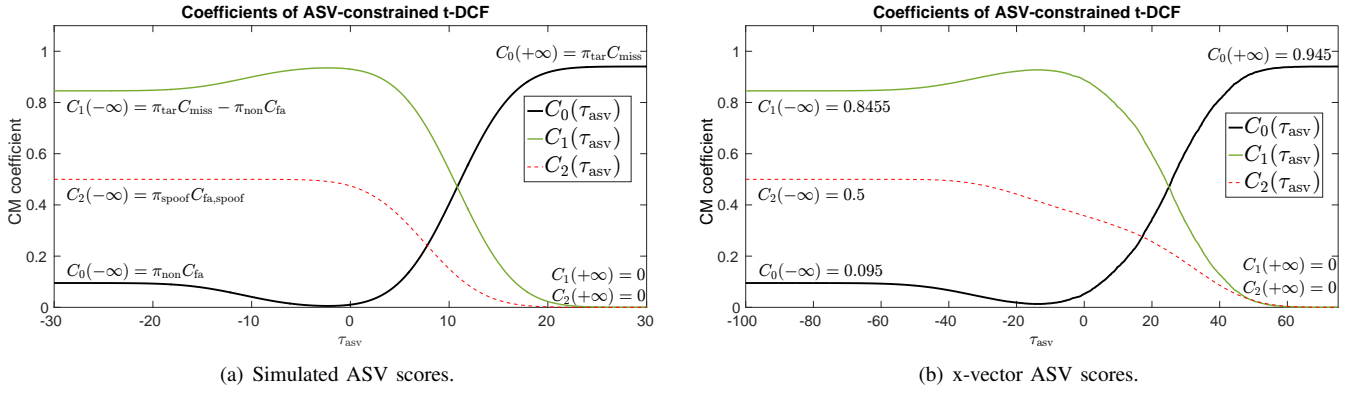
(a) Simulated ASV scores.

(b) x-vector ASV scores.

Fig. 4. Illustration of ASV-constrained t-DCF coefficients (11) for (a) simulated ASV scores ($P_{\mathrm{e}}^{\mathrm{asv}} = 0.01$, $\xi = 0.85$) described in the Appendix, and (b) real x-vector ASV scores computed on the ASVspoof 2019 LA database, with t-DCF parameters described in Section VI (with $\pi_{\mathrm{spoof}} = 0.05$). The indicated special values indicate the coefficient values at $\tau_{\mathrm{asv}} = \pm\infty$. See text for details.

such ASV rejects also the spoofing attacks, so the spoof false alarms of CM does not matter.

Concerning the 'flattening' of the ASV-conditional t-DCF observed in Fig. 3, note that for the 'reject all' case ($\tau_{\mathrm{asv}} \to +\infty$) we have

$$\text{t-DCF}'(\tau_{\mathrm{cm}}) = \frac{C_0 + C_1 P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) + C_2 P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})}{C_0 + \min\{C_1, C_2\}}$$
$$= \frac{C_0 + 0 \cdot P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) + 0 \cdot P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}})}{C_0 + \min\{0, 0\}} = \frac{C_0}{C_0} = 1,$$

regardless of the CM system or its operating point. This simply says there is no way to improve (or for that matter, to degrade) such an ASV system using *any* CM. For the 'accept all' ASV system ($\tau_{\mathrm{asv}} \to -\infty$), the situation is similar. Performance cannot be improved using any CM ($C_1 = C_2 = 0$) if (and only if)

$$(\pi_{\mathrm{tar}} C_{\mathrm{miss}} = \pi_{\mathrm{non}} C_{\mathrm{fa}}) \;\; \text{AND} \;\; (\pi_{\mathrm{spoof}} = 0 \;\; \text{OR} \;\; C_{\mathrm{fa,spoof}} = 0).$$

The first condition above states that overall costs from missed targets and falsely accepted nontargets are the same (there is no preference for either). The second condition states that either spoofing attacks are not anticipated ($\pi_{\mathrm{spoof}} = 0$), or that one does not care about them ($C_{\mathrm{fa,spoof}} = 0$). This is intuitively reasonable. The CM cannot do anything useful to reject spoofs (they do not incur losses), and the potential benefit of CMs helping to reject nontargets will be 'evened out' by equally costly target rejections (that the dummy ASV system would have otherwise accepted). Whenever the t-DCF parameters and the ASV operating point are chosen so that $C_1 \neq 0$ and $C_2 > 0$, there is potential for the CM to improve performance.

Finally, why does the normalized ASV-conditional t-DCF reach the value 1 at one of the infinities (as the right panel of Fig. 3 suggests)? Without loss of generality, suppose that t-DCF$'(+\infty) \neq 1$. Since $P_{\mathrm{miss}}^{\mathrm{cm}}(+\infty) = 1$ and $P_{\mathrm{fa}}^{\mathrm{cm}}(+\infty) = 0$, we have

$$\text{t-DCF}'(+\infty) = \frac{C_0 + C_1 \cdot 1 + C_2 \cdot 0}{C_0 + \min\{C_1, C_2\}} = \frac{C_0 + C_1}{C_0 + \min\{C_1, C_2\}},$$

and since we assumed this expression is $\neq 1$, it follows that $\min\{C_1, C_2\} = C_2$. Therefore, at $\tau_{\mathrm{cm}} = -\infty$, we have

t-DCF$'(-\infty) = (C_0 + C_2)/(C_0 + \min\{C_1, C_2\}) = 1$. Similar argumentation can be made by assuming t-DCF$'(-\infty) \neq 1$, which implies t-DCF$'(+\infty) = 1$. *At either (CM) infinity, the normalized t-DCF equals 1*, as the system collapses to the default system.

VI. EXPERIMENTAL SET-UP

The experimental work aims to assess the tandem operation of ASV and CM systems submitted to the three editions of the *Automatic Speaker Verification Spoofing and Countermeasures* (ASVspoof) challenge. This section defines the cost model parameters, gives an overview of the ASVspoof corpora and the fixed ASV system.

We focus on **authentication** scenarios, to which the problem of spoofing is most relevant. As in [11], we assume a hypothetical banking application where $C_{\mathrm{miss}} = 1$, $C_{\mathrm{fa}} = C_{\mathrm{fa,spoof}} = 10$ and $\pi_{\mathrm{non}} \ll \pi_{\mathrm{tar}}$, and $\pi_{\mathrm{spoof}} \ll \pi_{\mathrm{tar}}$. The parameter of interest is π_{spoof}, which we fix to a small arbitrary value and then obtain $\pi_{\mathrm{tar}} = (1 - \pi_{\mathrm{spoof}}) \times 0.99$ and $\pi_{\mathrm{non}} = (1 - \pi_{\mathrm{spoof}}) \times 0.01$.

The three speech corpora originate from the past ASVspoof challenges. The 2015 edition [15] focused on the detection of synthetic speech and voice conversion, the 2017 edition [16] focused on the detection of replay attacks and the latest 2019 edition [17] focused on the three types of attacks categorized into *logical access* (LA) and *physical access* (PA) scenarios. The data and protocol related details of these corpora are reported elsewhere [15], [16], [17]; the focus here is on aspects relevant to the current evaluation. A summary of trial statistics for the evaluation partitions of the corpora used in this work is presented in Table III.

The ASV system uses *time-delay neural network* (TDNN) based *x-vector* speaker embeddings [18] together with a *probabilistic linear discriminant analysis* (PLDA) [19] backend. The x-vector extractor is a pre-trained[4] neural network model developed with the Kaldi [20] toolkit. It is trained with MFCC features extracted from audio data from 7,325 speakers of the

---

[4] http://kaldi-asr.org/models/m7

TABLE III
NUMBER OF TRIALS IN THE EVALUATION PROTOCOLS FOR ASV
EXPERIMENTS.

| Dataset Name	Target / Non-target / Spoof
ASVspoof 2015	4053 / 77007 / 80000
ASVspoof 2017	1106 / 18624 / 10878
ASVspoof 2019 (LA)	5370 / 33327 / 63882
ASVspoof 2019 (PA)	12960 / 123930 / 116640

Fig. 5. Comparison of the unconstrained and ASV-constrained minimum t-DCF for the top three systems of the ASVspoof 2019 LA challenge ($\pi_{\text{spoof}} = 0.05$). For t-DCF both ASV and CM thresholds are chosen to minimize the cost, while for ASV-constrained t-DCF, the ASV is set to its EER point.

VoxCeleb1 and VoxCeleb2 corpora [21].[5] Further related to network parameters and data preparation are available in [18]. The original Kaldi recipe was modified to include PLDA adaptation using in-domain data. Full details of the ASV system can be found in [22]. PLDA adaptation is applied separately for the ASVspoof 2015, 2017, and 2019 (LA and PA) datasets using in-domain data.

VII. RESULTS

First, the unconstrained and ASV-constrained variants are compared empirically. Then, the ASV-constrained t-DCF is used to assess the performance of submitted CM systems together with a common ASV system. The ASV threshold τ_{asv} is fixed to the EER operating point, while the CM threshold is set to the minimum point of t-DCF. The final experiment addresses the choice of thresholds. All the presented t-DCF values are in their normalised form (see Table II).

A. Unconstrained vs. ASV-Constrained t-DCF

Fig. 5 illustrates a comparison of unconstrained and ASV-constrained t-DCF formulations for the three top-performing systems of the ASVspoof 2019 LA scenario. The ASV-constrained values are systematically higher, as expected: while the unconstrained t-DCF does not assume a pre-defined ASV threshold, the ASV-constrained t-DCF does. The minimum t-DCF for the ASV-constrained is lower bounded by the cost of the ASV system at the EER operating point (which is suboptimal), the unconstrained t-DCF allows both ASV and CM thresholds to be varied jointly, yielding lower t-DCF values. Another difference is in the default cost used to obtain the normalised t-DCF: while the default t-DCF for the unconstrained formulation does not depend on ASV error rates — see Eqs. (13), (14) and (15) — the one for the ASV-constrained t-DCF does; see Eqs. (11) and (18).

B. ASV-Constrained t-DCF of ASVspoof Submissions

Fig. 6 illustrates the ASV-constrained, minimum normalised t-DCF of the ten top-performing submission of each ASVspoof challenge (2015, 2017, 2019 LA and 2019 PA). Submissions are sorted by increasing t-DCF. The upper and lower figures were computed using $\pi_{\text{spoof}} = 0.01$ and $\pi_{\text{spoof}} = 0.05$, respectively. For reference purposes, the green line shows the t-DCF for a *perfect* CM. It corresponds to the ASV floor C_0 as defined in Eq. (11). Another reference, shown by the red

[5]For further details, check the VoxCeleb Kaldi recipe at https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2
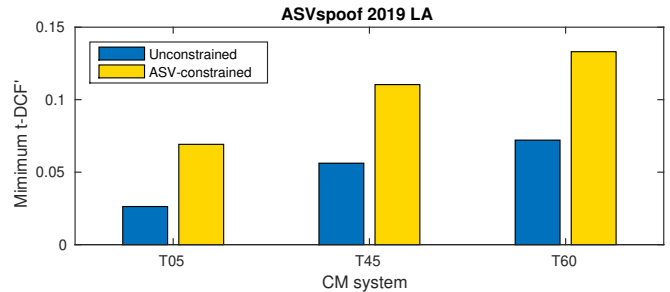
dashed line, corresponds to the default CM that either accepts or rejects all trials, whichever produces a lower cost.

Our first two immediate observations are that CMs are beneficial (all values are below 'no CM') and that the improvements are often substantial. Second, none of the CMs reaches the ASV floor, suggesting potential for future improvements in the CM technology. We see overall *higher* t-DCF values obtained with *lower* spoofing prior $\pi_{\text{spoof}} = 0.01$. This may seem counterintuitive at first but the operation of an imperfect CM will also produce target speaker misses, which increases the overall cost. In the extreme case when no spoofing attacks are expected, one should not use *any* CM. In contrast, when spoofing attacks are likely to occur, CMs are helpful in decreasing the cost by rejecting spoofing attempts (relative to not having any CM).

Note that the CM rank may change when π_{spoof} varies. Differences in the ranks are notable for the ASVspoof 2017 database. This is explained by noting that the target metric for the ASVspoof 2017 edition was the CM EER, which is generally not the optimal operating point for the t-DCF parameters used here. A given system (e.g. S02) can perform better than others at the EER point, but worse in other areas of the DET curve.

Fig. 7 can give some insights into the observations made from Fig. 6. Here, the ASV-constrained normalised t-DCF curve is shown for submission T45 of the ASVspoof 2019 LA challenge, when varying the CM threshold τ^{cm}, for different values of π_{spoof}: 0.05, 0.10, 0.15 and 0.20. We see that the (minimum) t-DCF decreases with an increasing value of π_{spoof}. The optima are reached for different values of τ_{cm} depending on π_{spoof}, as expected. For the lowest spoof prior $\pi_{\text{spoof}} = 0.001$, minimum t-DCF is only slightly below 1, indicating that the CM cannot improve the performance much; even though T45 performs well, the spoofing attack is simply too rare for it to make a substantial difference.

Thus far, we have focused on scores pooled from all attacks, even if their effectiveness (in terms of fooling ASV) varies. It is therefore useful to diagnose attack-specific, empirical C_2 functions, similar to those in Fig. 4 (note that C_0 and C_1 depend on target and nontarget trials only). The attack-specific C_2 graphs are shown in Fig. 8 along with the corresponding ASV EERs. We observe, first, that different attacks produce similarly-shaped but differently located graphs along the τ_{asv}
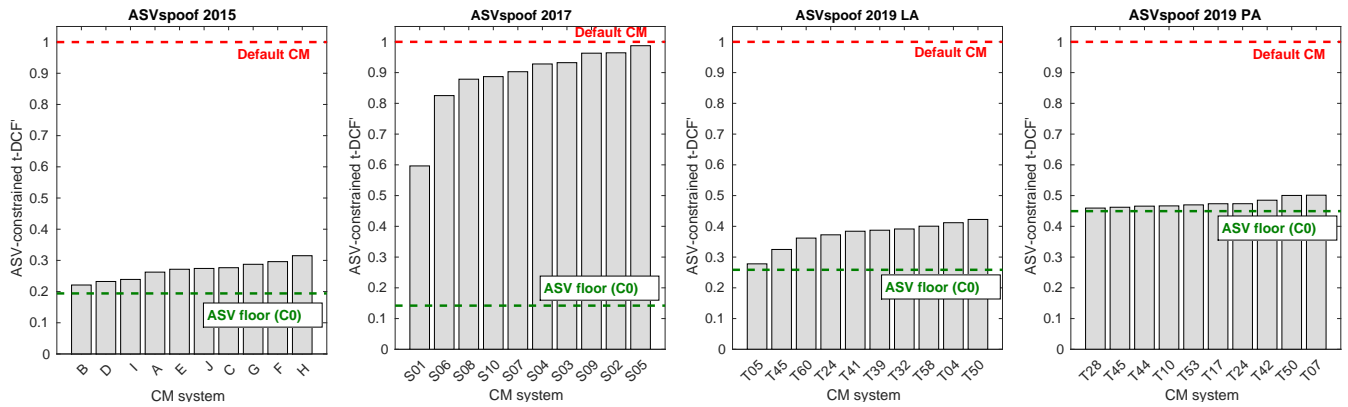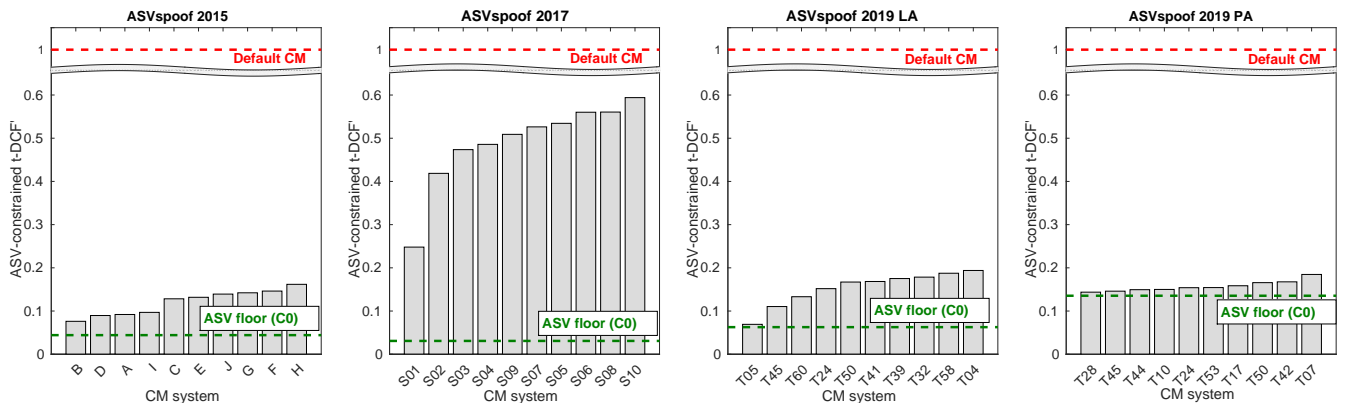
(a) $\pi_{\text{spoof}} = 0.01$.

(b) $\pi_{\text{spoof}} = 0.05$.

Fig. 6. ASV-constrained minimum normalised t-DCF of top-10 systems of ASVspoof 2015, 2017, 2019 LA and 2019 PA challenges for (a) $\pi_{\text{spoof}} = 0.01$ and (b) $\pi_{\text{spoof}} = 0.05$. "ASV floor" indicates the cost of the tandem system if the CM system was perfect (equal to C_0). "Default CM" indicates the cost of the tandem system with a dummy CM that either accepts or rejects all the trials (whichever yields a lower cost).

axis. For fixed τ_{asv}, the more effective attack (*i.e.*, higher EER) gives larger C_2. Second, there is substantial variation of C_2, especially on the LA condition. For the PA condition, both the C_2 graphs and the EERs vary less across attacks. Unlike the LA attacks generated by a large number of researchers and teams, the PA attacks were generated through a common simulation model with a few control parameters only [22], which may explain the more homogenous behavior. The interested reader may refer to [22], [17] for further details on ASVspoof 2019 attack generation and their impact upon ASV.

C. Empirical Threshold Selection Using t-DCF

Until this point, we considered an arbitrary τ_{asv} (set at the EER operating point) along with an optimum τ_{cm}. These were the choices in the ASVspoof 2019 challenge though the former is not aligned with the t-DCF specification. Further, we have considered oracle calibration only where both thresholds are set on the evaluation data. Thus, in our final experiment we demonstrate use of the t-DCF to guide selection of both thresholds (on development and evaluation data).

In line with the ASV-constrained approach, we consider a particular scenario where the ASV and CM systems developers agree upon a specific t-DCF parametrization (specified by the
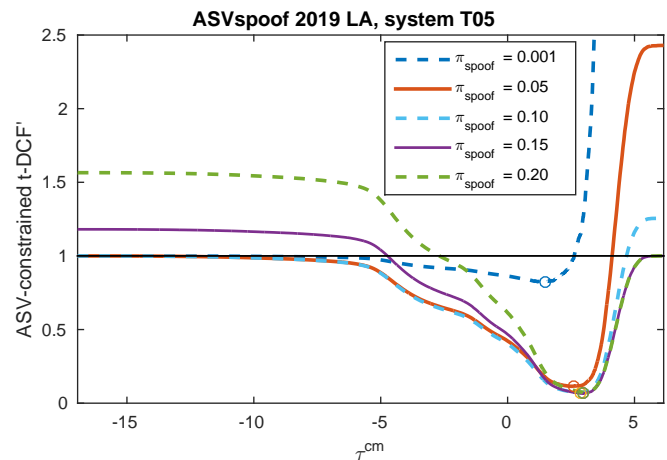

Fig. 7. Normalised ASV-constrained t-DCF w.r.t. the CM threshold for system T45 on ASVspoof 2019 LA.

bank) but optimise their respective systems separately, while sharing ASV error rates only. Using common development data, they proceed as follows:

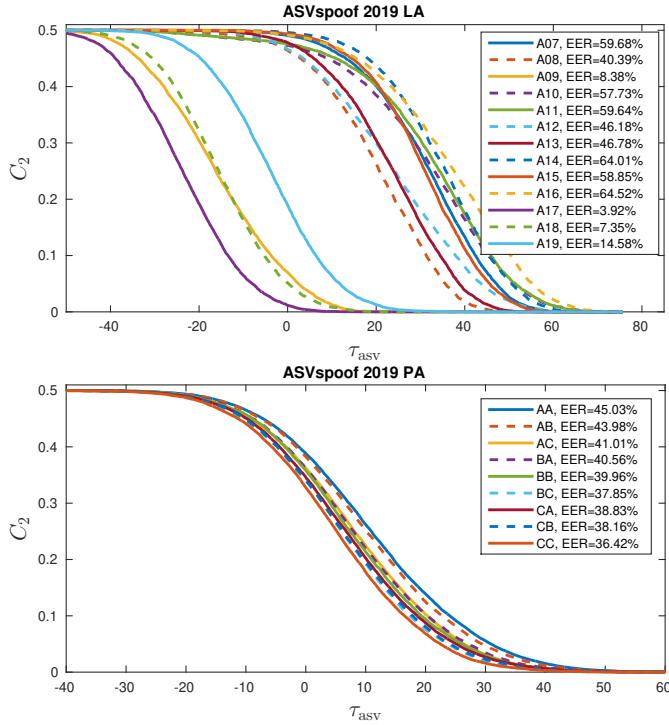1) Since the ASV system is not designed to reject spoofing attacks and hence by using target/non-target trials only,

Fig. 8. C_2 over the ASV threshold τ_{asv} per spoofing attack type, on the ASVspoof 2019 logical access (upper panel) and physical access (lower panel) scenarios. The indicated EERs are those of the x-vector based ASV system.

TABLE IV
ASV-CONSTRAINED t-DCF OBTAINED USING DIFFERENT ASV AND CM THRESHOLDS. THE FIRST TWO LINES INDICATE THE EMPIRICAL CRITERION (AND DATA) TO SET τ_{ASV} AND τ_{CM}, RESPECTIVELY.

Fix $\tau_{asv} \rightarrow$	EER (eval)	EER (dev)	min C_0 (dev)	min C_0 (dev)
Fix $\tau_{cm} \rightarrow$	min t-DCF (eval)	min t-DCF (eval)	min t-DCF (eval)	min t-DCF (dev)
System		min t-DCF		actual t-DCF
ASVspoof 2019 LA				
ASV floor	0.0627	0.0860	0.0304	0.0304
T05	0.0692	0.0923	0.0367	0.3260
T45	0.1104	0.1327	0.0785	0.4989
ASVspoof 2019 PA				
ASV floor	0.1354	0.1389	0.0628	0.0628
T28	0.1437	0.1472	0.0715	0.0718
T45	0.1460	0.1495	0.0740	0.0769

the ASV system developer optimizes τ_{asv} by minimizing the ASV floor, i.e.:

$$\tau_{asv}^* = \arg\min_{\tau_{asv}} C_0(\tau_{asv})$$
$$= \arg\min_{\tau_{asv}} \left\{ \pi_{tar} C_{miss} P_{miss}^{asv}(\tau_{asv}) + \pi_{non} C_{fa} P_{fa}^{asv}(\tau_{asv}) \right\}. \quad (21)$$

2) Using the shared ASV error rates, bonafide (target/non-target) and spoofed trials, the CM developer determines C_0, C_1, and C_2 coefficients in (11) and sets the CM to operate at the empirical minimum of the ASV-constrained t-DCF in (10):

$$\tau_{cm}^* = \arg\min_{\tau_{cm}} \text{t-DCF}(\tau_{cm}). \quad (22)$$

We contrast the above approach with the EER-based ASV threshold selection (as used in previous experiments) in Ta-

ble IV which shows results for the evaluation partitions of the ASVspoof 2019 datasets for LA and PA tasks. The first two columns corresponds to the EER-based optimization of τ_{asv} on evaluation and development data, respectively. The last two columns correspond to choosing the ASV threshold using (21) on development data.

Concerning τ_{cm}, in turn, the first three columns correspond to oracle calibration of the CM (minimum t-DCF on the evaluation set). The last column corresponds to the **actual** t-DCF where both τ_{asv} and τ_{cm} are set on the development set. Results are shown for the top-2 systems for the ASVspoof 2019 LA and PA scenarios. Similar to Fig. 6, the lower bound (ASV floor) is also shown for reference purposes.

Upon comparison of results in the first two columns we see that, for the LA scenario, the ASV EER threshold set on development data is suboptimal compared to the ASV EER threshold tuned on evaluation data; there are differences between results in each column. In contrast, for the PA scenario, results are similar, no matter if the threshold is tuned on the development data or the evaluation data. Results in column 3 indicate that the ASV threshold set to minimize C_0 on the development data leads to substantially lower t-DCF values than in columns 1 and 2. This is expected since the EER operating point represents a miscalibrated ASV system that is generally not intended as a minimizer of the t-DCF.

Finally, results for the LA scenario in column 4 indicate that a CM threshold tuned on the development set does not generalize well to the evaluation set; except for the perfect CM (ASV floor), results in columns 4 are substantially worse than those in column 3. In contrast, the difference is comparatively small for the PA scenario. This might be due to the same reasons noted in Fig. 8 — the diversity in the spoofing attacks for the LA scenario is greater than that for the PA scenario. Data for the latter was generated with a common simulation procedure, leading to more homogeneuos attacks.

These results demonstrate the potential of t-DCF as an empirical threshold selection criterion. The above procedure is intended as a demonstration that follows the format of the ASVspoof 2019 challenge, though there are a number of open questions that we discuss below.

VIII. DISCUSSION AND FUTURE WORK

Before concluding, we discuss here the assumed independence in the t-DCF metric, and outline a number of open research problems exposed in this work.

A. The ASV—CM Independence Assumption

The formulation of the tandem error rates in Eqs. (7) is based upon the assumption that ASV and CM system error rates are independent. While this may seem somewhat questionable, there are a number of reasons that support such a restrictive assumption. They relate to the specific ASVspoof scenario and the fundamental differences between ASV and CM systems:

1) as discussed in Section II, ASV and CM systems address different detection tasks and hence the two detectors provide complementary views, even to the same data;

2) in a speaker-independent CM setting, the two systems are trained using disjoint speakers;

3) ASV systems provide scores for each (target speaker, test) pair whereas speaker-independent CM systems, in turn, use one anti-spoofing model to score all test utterances;

4) the ASV and the CM systems could be developed by different researchers, teams or companies and can hence be based on different ideas, methods, software implementations, control parameters, and training data choices;

5) the two systems typically use different features (e.g. CQCC vs. MFCC features) and classifier back-ends.

Thus, *before* application of the tandem system to evaluation data, ASV and CM scores can be treated as being independent, as can the respective nontarget/target and bonafide/spoof score distributions. *After* the same tandem system is applied to evaluation data, however, ASV and CM cannot necessarily be treated as being independent — there will be some *conditional* dependence by virtue of both systems being executed on the same data. We nonetheless assert in (7) that ASV and CM scores are conditionally *in*dependent.

The primary reason for our conditional independence assumption is practical: it allows the ASV and CM error rates be computed by different parties (and from different data). We have deemed this as a necessity given the practical focus of the ASVspoof challenge series. Nonetheless, if the joint distribution of ASV and CM scores *is* available (for instance, when the same person develops both systems), it may be useful to assess the impact of statistical dependency on the estimated detection error rates. Independence assumptions are sometimes difficult to avoid, e.g. the well known NIST SREs [1] which assume statistically independent trials, yet reuse many times data from the same target/non-target speakers [23]. A deeper study of the in/dependence issue is kept for future work.

B. The Need for a Tandem Calibration Model

Aside from experiments reported in Section VII-C, the issue of optimally calibrated detectors is largely overlooked in this paper. In doing so, we have sidestepped the important but very real problem of threshold setting. The default practice in ASV research, with widespread acceptance by the ASV research community, is to fix τ_{asv} to the Bayes minimum-risk operating point [3], with the assumption that detection scores are well-calibrated log-likelihood ratios (LLRs) [13]. The Bayes threshold is given analytically by the DCF parameters as $\tau_{\text{asv}} = \log[(C_{\text{fa}}/C_{\text{miss}}) \cdot (1 - \pi_{\text{tar}})/\pi_{\text{tar}}]$. Since arbitrary detectors may yield poorly calibrated LLRs it is customary to apply a *calibration transform* in the score domain [24], [25], [26]. Among other benefits, this allows calibrated scores to be used readily with different DCF parameters. Furthermore, calibration allows principled decomposition of *discrimination loss* (how bad the system is in terms of classification; at an ideal threshold) and *calibration loss* (how badly off the threshold is from that ideal threshold) [13], [27]. The apparent benefits of calibration may cause the reader to wonder why we did not apply a tandem calibration model in the case of the t-DCF.

The main reason for having avoided calibration is that the manner by which it should be applied in the tandem setting is far from being a simple extension of calibration in the case of the DCF. Under the tandem framework, we have not only one additional *system* (CM) but also one additional *class* (spoof). At this point, the authors have no analytical expressions for Bayes-optimal ASV and CM thresholds. In addition, the unconstrained and ASV-constrained cases may require different treatments. The ASV-constrained case yields a cost function *for particular spoofing attacks* whose impact upon the ASV system is known; in reality, we do not know that impact in advance. In summary, how calibration transforms should be defined for tandem systems remains an open question and one that deserves attention in future work.

IX. CONCLUSIONS

The intention of the authors has been to provide a self-contained tutorial on the tandem detection cost (t-DCF) framework that generalizes the standard DCF. Extending upon [11], our special focus has been on a *constrained* t-DCF formulation, where the biometric system (here, ASV) is essentially treated as a black-box. The constrained cost serves as a guide for the optimization of a countermeasure for a given biometric system. A variant of the constrained t-DCF was put to its first stress test in the latest ASVspoof 2019 challenge [22] and we anticipate it remaining as the primary metric in future editions of the challenge.

The question of how the performance of *any* binary classifier is to be assessed is much more subtle than it appears on the surface; it took some considerable time for the DCF framework to be absorbed as an integral part of ASV system development — and even longer to migrate from *ad-hoc* 'threshold optimization' recipes to calibrated log-likelihood ratios [24], [27], [14], [26], [25]. We hope that the current study serves to reduce the risk of similarly slow adoption of application-directed metrics within the anti-spoofing community. Since the necessity for tandem systems (consisting of two subsystems) add to the complexity of the assessment issue, the authors have purposefully left out a number of related topics, such as calibration; we focused on *minimum* t-DCF (with oracle threshold). We plan to address calibration in our future work.

The authors note that a variety of different, *adhoc* metrics remain popular in the assessment of biometric systems (beyond the voice trait). Presentation attack detection is a relatively recent, but growing and evolving area of research, and this state of rapid development may go some way to explain the lack of application-directed metrics in use today. What is clear, however, is that next-generation biometric systems *must* be prepared for the possibility of spoofing (whether it be ever-improving DeepFake video and synthetic speech quality, or potential fraud in high-stakes applications including border control and forensics). To help prepare for a future where biometrics may no longer be trusted, we need meaningful metrics both for performance assessment and optimization. One benefit of the proposed t-DCF framework lies in its *generality*. While, on account of the authors' research interests, voice biometrics has been the running example, the t-DCF itself requires nothing beyond the detection scores (or hard decisions) of the biometric recognizer and the presentation

attack detector. The specification of cost parameters is left to the domain expert.

APPENDIX: GAUSSIAN SCORE SIMULATOR

In [24], relations between Gaussian score distributions, well-calibrated log-likelihood ratios (LLRs) and the EER were drawn to derive a supervised score calibration recipe under the assumption of Gaussian nontarget/target scores. In this special case, the means of non/target distributions are symmetric and also relate to their variance. Such constrained score distributions can be parameterized by a single number, such as the EER. In the following, we outline a simple score simulator models for target, nontarget, and spoof classes.

Our score simulator models all the class-conditional score distributions in (1) and (2) as Gaussians,

$$p_R(r|\theta_\bullet) = \mathcal{N}(r|\mu_\bullet^{\mathrm{asv}}, (\sigma_\bullet^2)^{\mathrm{asv}}), \quad \bullet \in \{\mathrm{tar, non, spoof}\}$$
$$p_Q(q|\theta_\circ) = \mathcal{N}(q|\mu_\circ^{\mathrm{cm}}, (\sigma_\circ^2)^{\mathrm{cm}}), \quad \circ \in \{\mathrm{bona, spoof}\}, \quad (23)$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes the univariate normal density with mean μ and variance σ^2. For the CM, the target and nontarget distributions are assumed to be the same — the CM cannot discriminate between target and nontarget classes.

As there are three classes within the ASV system and two classes within the CM system, (23) requires the specification of $(3+2) \times 2 = 10$ parameters. A convenient means to reduce the number is through the EERs of each system — target vs. nontarget EER for ASV and bonafide vs. spoof EER for CM. The EER can be expressed analytically in terms of Gaussian parameters [28], [24]. Even if the tail behavior of empirical ASV scores obtained from typical back-ends [19] differs from that of Gaussians [26], [25], supervised calibration using constrained Gaussians [24] leads to near-optimum calibration near the minimum cost operating point [26], [25].

Let us begin with the ASV system with $\mathcal{N}(r|\mu_{\mathrm{tar}}^{\mathrm{asv}}, (\sigma_{\mathrm{tar}}^2)^{\mathrm{asv}})$ and $\mathcal{N}(r|\mu_{\mathrm{non}}^{\mathrm{asv}}, (\sigma_{\mathrm{non}}^2)^{\mathrm{asv}})$ as the target and nontarget score distributions, respectively. The analytic EER is given by $P_{\mathrm{e}} = 1 - \Phi(F^{\mathrm{asv}})$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $F^{\mathrm{asv}} = (\mu_{\mathrm{tar}}^{\mathrm{asv}} - \mu_{\mathrm{non}}^{\mathrm{asv}})/(\sigma_{\mathrm{tar}}^{\mathrm{asv}} + \sigma_{\mathrm{non}}^{\mathrm{asv}})$. For completely overlapped distributions with equal means, one obtains the chance level $P_{\mathrm{e}}^{\mathrm{asv}} = \frac{1}{2}$. The four parameters collapse to a scalar F^{asv}, which uniquely specifies the EER (as Φ is bijective). We *tie*

the means and variances so that we have only one degree of freedom, specified by the EER, from which we determine the four parameters. To this end, we adopt the approach of [24] with shared variance $\sigma^{\mathrm{asv}} \equiv \sigma_{\mathrm{tar}}^{\mathrm{asv}} = \sigma_{\mathrm{non}}^{\mathrm{asv}}$, symmetric means ($\mu^{\mathrm{asv}} \equiv \mu_{\mathrm{tar}}^{\mathrm{asv}} = -\mu_{\mathrm{non}}^{\mathrm{asv}}$) and mean and variance being related by $\sigma^{\mathrm{asv}} = \sqrt{2\mu^{\mathrm{asv}}}$. To sum up, $p_R(r|\theta_{\mathrm{tar}}) = \mathcal{N}(r|\mu^{\mathrm{asv}}, 2\mu^{\mathrm{asv}})$, $p_R(r|\theta_{\mathrm{non}}) = \mathcal{N}(r| - \mu^{\mathrm{asv}}, 2\mu^{\mathrm{asv}})$.

After having determined the target and nontarget distributions from a given EER, $P_{\mathrm{e}}^{\mathrm{asv}}$, we proceed by defining the ASV spoof score distribution as:

$$p_R(r|\theta_{\mathrm{spoof}}) = \mathcal{N}\Big(r|\mu^{\mathrm{asv}}(2\xi - 1), 2\mu^{\mathrm{asv}}\Big), \quad (24)$$

characterized by an additional parameter $\xi \in \mathbb{R}$ that we dub as the *spoofing factor*. It is illustrated in Fig. 9, with the following interpretation:

- $\xi = 1$ implies spoof mean equal to target mean (attack indistinguishable from the target speaker).
- $\xi = 0$ implies spoof mean equal to nontarget mean, *i.e.* zero-effort spoofing attack.
- $\xi > 1$ implies spoof mean higher than target mean.
- $\xi < 0$ implies spoof mean less than nontarget mean.

The typical case is $0 \ll \xi < 1$, *i.e.* attacks that produce substantially higher scores than nontargets but do not quite reach the target scores due to modeling imperfections, difficulties in gathering spoofing attack training data, or other reasons.

The CM bona fide score distributions are specified in the same way as for the ASV system: given a desired bonafide-to-spoof EER, $P_{\mathrm{e}}^{\mathrm{cm}}$, we find the distributions of bona fide and spoof classes as $\mathcal{N}(\mu^{\mathrm{cm}}, 2\mu^{\mathrm{cm}})$ and $\mathcal{N}(-\mu^{\mathrm{cm}}, 2\mu^{\mathrm{cm}})$, respectively, following the same parameter constraints noted above.

To state our simulator assumptions in an alternative way, the ASV and CM score random variables are assumed to be statistically independent (for practical reasons; see Section VIII-A). The joint probability density functions of target, nontarget and spoof classes are then the product of their marginal distributions, which can be represented as bi-variate Gaussians with a diagonal covariance matrix. For the target distribution, for instance, we have

$$p_{R,Q}(r, q|\theta_{\mathrm{tar}}) = \mathcal{N}(r|\mu_{\mathrm{tar}}^{\mathrm{asv}}, (\sigma_{\mathrm{tar}}^2)^{\mathrm{asv}})\mathcal{N}(q|\mu_{\mathrm{bona}}^{\mathrm{cm}}, (\sigma_{\mathrm{bona}}^2)^{\mathrm{cm}})$$
$$= \mathcal{N}(\boldsymbol{\mu}_{\mathrm{tar}}, \boldsymbol{\Sigma}_{\mathrm{tar}}),$$

with

$$\boldsymbol{\mu}_{\mathrm{tar}} = \begin{bmatrix} \mu^{\mathrm{asv}} \\ \mu^{\mathrm{cm}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathrm{tar}} = \begin{bmatrix} 2\mu^{\mathrm{asv}} & 0 \\ 0 & 2\mu^{\mathrm{cm}} \end{bmatrix}.$$

Similarly, for the nontarget and spoof class we have

$$\boldsymbol{\mu}_{\mathrm{non}} = \begin{bmatrix} -\mu^{\mathrm{asv}} \\ \mu^{\mathrm{cm}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathrm{non}} = \begin{bmatrix} 2\mu^{\mathrm{asv}} & 0 \\ 0 & 2\mu^{\mathrm{cm}} \end{bmatrix}$$

$$\boldsymbol{\mu}_{\mathrm{spoof}} = \begin{bmatrix} \mu^{\mathrm{asv}}(2\xi - 1) \\ -\mu^{\mathrm{cm}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathrm{spoof}} = \begin{bmatrix} 2\mu^{\mathrm{asv}} & 0 \\ 0 & 2\mu^{\mathrm{cm}} \end{bmatrix}.$$

In summary, the three control parameters of our score simulator that define the above distributions, are:

1) Target-to-nontarget EER of ASV, $P_{\mathrm{e}}^{\mathrm{asv}}$, as a model of the discrimination performance of ASV;
2) ASV spoofing factor, $\xi \in \mathbb{R}$, as a model of how effective the spoofing attacks are in fooling the ASV;
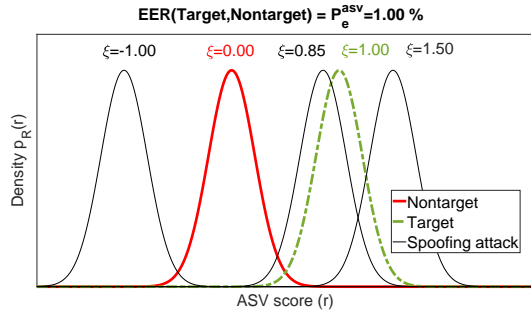
Fig. 9. Simulated, Gaussian-distributed ASV score distributions with varied spoofing factor, ξ. Nontarget and target classes are special cases with $\xi = 0$ and $\xi = 1$, respectively.

3) Bonafide-to-spoof EER of CM, P_e^{cm}, as a model of CM discrimination performance.

Under the above Gaussian model, the detection error rates in (4) are given by,

$$
\begin{aligned}
P_{\mathrm{miss}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}) &= \Phi\left(\frac{\tau_{\mathrm{asv}} - \mu^{\mathrm{asv}}}{\sqrt{2\mu^{\mathrm{asv}}}}\right) \\
P_{\mathrm{fa}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}) &= 1 - \Phi\left(\frac{\tau_{\mathrm{asv}} + \mu^{\mathrm{asv}}}{\sqrt{2\mu^{\mathrm{asv}}}}\right) \\
P_{\mathrm{fa,spoof}}^{\mathrm{asv}}(\tau_{\mathrm{asv}}) &= 1 - \Phi\left(\frac{\tau_{\mathrm{asv}} - \mu^{\mathrm{asv}}(2\xi - 1)}{\sqrt{2\mu^{\mathrm{asv}}}}\right) \quad (25) \\
P_{\mathrm{miss}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) &= \Phi\left(\frac{\tau_{\mathrm{cm}} - \mu^{\mathrm{cm}}}{\sqrt{2\mu^{\mathrm{cm}}}}\right) \\
P_{\mathrm{fa}}^{\mathrm{cm}}(\tau_{\mathrm{cm}}) &= 1 - \Phi\left(\frac{\tau_{\mathrm{cm}} + \mu^{\mathrm{cm}}}{\sqrt{2\mu^{\mathrm{cm}}}}\right),
\end{aligned}
$$

where $\Phi(\tau)$ denotes the cumulative distribution function (CDF) of the standard normal distribution. The standardization operator $\tau \mapsto (\tau - \mu)/\sigma$ yields the CDF of a nonstandard normal distribution with mean μ and variance σ^2.

REFERENCES

[1] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the National Institute of Standards and Technology," *Computer Speech & Language*, vol. 60, 2020. [Online]. Available: https://doi.org/10.1016/j.csl.2019.101032

[2] 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. 2-3, pp. 225–254, 2000. [Online]. Available: https://doi.org/10.1016/S0167-6393(99)00080-1

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[4] E. T. Jaynes, *Probability theory: The logic of science*. Cambridge: Cambridge University Press, 2003.

[5] ISO/IEC 30107-1:2016, "Information technology — Biometric presentation attack detection — Part 1: Framework," https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-1:ed-1:v1:en, 2016, [Online; accessed 22-February-2018].

[6] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech*, 2013, pp. 925–929.

[7] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. EUROSPEECH*, 2001, pp. 759–762. [Online]. Available: http://www.isca-speech.org/archive/eurospeech_2001/e01_0759.html

[8] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Trans. Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015. [Online]. Available: https://doi.org/10.1109/TIFS.2015.2407362

[9] M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z. Tan, "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015," in *Proc. Interspeech*, 2016, pp. 1700–1704. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-1280

[10] 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2289

[11] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey*, 2018, pp. 312–319. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-44

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

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

[14] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch University, 2010.

[15] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015, pp. 2037–2041. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_2037.html

[16] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, 2017, pp. 2–6.

[17] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech*, 2019, pp. 1008–1012.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[19] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[22] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230820300474

[23] J. C. Wu, A. F. Martin, C. S. Greenberg, and R. N. Kacker, "The impact of data dependence on speaker recognition evaluation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 5–18, 2017.

[24] D. A. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proc. Interspeech*, 2013, pp. 1619–1623. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2013/i13_1619.html

[25] S. Cumani and P. Laface, "Tied normal variance–mean mixtures for linear score calibration," in *Proc. ICASSP*, May 2019, pp. 6121–6125.

[26] S. Cumani, "Normal variance-mean mixtures for unsupervised score calibration," in *Proc. Interspeech*, 2019, pp. 401–405.

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

[28] N. Poh and S. Bengio, "Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks?" in *Proc. ICASSP*, 2004, pp. 893–896. [Online]. Available: https://doi.org/10.1109/ICASSP.2004.1327255
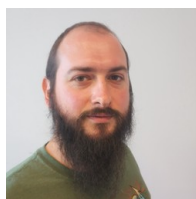
Kong Aik Lee (M'05-SM'16) is currently a Senior Principal Researcher at the Biometrics Research Laboratories, NEC Corp., Japan. He received his Ph.D. degree from Nanyang Technological University, Singapore, in 2006. From 2006 to 2018, he was a Scientist at the Human Language Technology department, I^2R, A*STAR, Singapore, where he led the speaker recognition group. He was the recipient of Singapore IES Prestigious Engineering Achievement Award 2013 for his contribution to voice biometrics technology. He serves as an Editorial Board Member for Elsevier Computer Speech and Language (2016 - present), and an Associate Editor for IEEE/ACM Transactions on Audio, Speech and Language Processing (2017 - present). He is an elected member of IEEE Speech and Language Technical Committee. He chairs thte Speaker Odyssey 2020 Workshop.

Tomi H. Kinnunen is an Associate Professor at the University of Eastern Finland. He received his Ph.D. degree in computer science from the University of Joensuu in 2005. From 2005 to 2007, he was an Associate Scientist at the Institute for Infocomm Research (I2R), Singapore. Since 2007, he has been with UEF. From 2010-2012, he was funded by a postdoctoral grant from the Academy of Finland. He has been a PI or co-PI in three other large Academy of Finland-funded projects and a partner in the H2020-funded OCTAVE project. He chaired the *Odyssey* workshop in 2014. From 2015 to 2018, he served as an Associate Editor for IEEE/ACM Trans. on Audio, Speech and Language Processing 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, with a focus on voice conversion and spoofing. Since 2017, he has been Associate Professor at UEF, where he leads the Computational Speech Group. He is one of the cofounders of the ASVspoof challenge, a nonprofit initiative that seeks to evaluate and improve the security of voice biometric solutions under spoofing attacks.
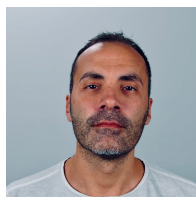
Ville Vestman is an Early Stage Researcher at the University of Eastern Finland (UEF). He received his M.S. degree in mathematics from UEF in 2013. Since 2015, his research work at UEF has been focused on speech technology and, more specifically, on speaker recognition. He is one of the co-organizers of the ASVspoof 2019 challenge.

Héctor Delgado received his Ph.D. degree in Telecommunication and System Engineering from the Autonomous University of Barcelona (UAB), Spain, in 2015. From 2015 to 2019 he was with the Speech and Audio Processing Research Group at EURECOM (France). Since 2019 he is a Senior Research Scientist at Nuance Communications Inc. He serves as an associate editor for the EURASIP Journal on Audio, Speech, and Music Processing. He is a co-organiser of the ASVspoof challenge since its 2017 edition. His research interests include signal processing and machine learning applied to speaker recognition and diarization, speaker recognition anti-spoofing and audio segmentation.

Andreas Nautsch is with the Audio Security and Privacy research group (EURECOM). He received the doctorate from Technische Universität Darmstadt in 2019, where he was with the biometrics group within the German National Research Center for Applied Cybersecurity. He received B.Sc. and M.Sc. degrees from Hochschule Darmstadt (dual studies with atip GmbH) in 2012 and 2014, respectively. He served as an expert delegate to ISO/IEC and as project editor of the ISO/IEC 19794-13:2018 standard. Andreas is a co-initiator and secretary of the ISCA Special Interest Group on Security & Privacy in Speech Communication.

Nicholas Evans is a Professor at EURECOM, France, where he heads research in Audio Security and Privacy. He is a co-founder of the community-led, ASVspoof Challenge series and has lead or co-lead a number of special issues and sessions with an anti-spooing theme. He participated in the EU FP7 Tabula Rasa and H2020 OCTAVE projects, both involving anti-spoofing. Today, his team is leading the EU H2020 TReSPAsS-ETN project, a training initiative in security and privacy for multiple biometric traits. He co-edited the second edition of the Handbook of Biometric Anti-Spoofing, served previously on the IEEE Speech and Language Technical Committee and serves currently as an asscociate editor for the IEEE Trans. on Biometrics, Behavior, and Identity Science.

Massimiliano Todisco is an Assistant Professor within the Digital Security Department at EURECOM, France. He received his Ph.D. degree in Sensorial and Learning Systems Engineering from the University of Rome Tor Vergata in 2012. Currently, he is serving as principal investigator and coordinator for TReSPAsS-ETN, a H2020 Marie Skłodowska-Curie Innovative Training Network (ITN) and RESPECT, a PRCI project funded by the French ANR and the German DFG. He co-organises the ASVspoof challenge series, which are community-led challenges which promote the development of countermeasures to protect automatic speaker verification (ASV) from the threat of spoofing. He is the inventor of constant Q cepstral coefficients (CQCC), the most commonly used anti-spoofing features for speaker verification and first author of the highest-cited technical contribution in the field in the last three years. He has more than 90 publications. His current interests are in developing end-to-end architectures for speech processing and speaker recognition, fake audio detection and anti-spoofing, and the development of privacy preservation algorithms for speech signals based on encryption solutions that support computation upon signals, templates and models in the encrypted domain.

Xin Wang (S'16 - M'18) is a project researcher at National Institute of Informatics, Japan. He received the Ph.D. degree from SOKENDAI, Japan, in 2018. Before that, he received M.S. and B.E degrees from University of Science and Technology of China and University of Electronic Science and Technology of China in 2015 and 2012, respectively. His research interests include statistical speech synthesis and machine learning.

Douglas A. Reynolds (S'85–M'86–SM'98–F'10) is a senior member of the technical staff at MIT Lincoln Laboratory where he provides technical oversight of the projects in speaker and language recognition and speech-content-based information retrieval. Dr. Reynolds received his PhD from the Georgia Institute of Technology in 1992 with a dissertation on applying Gaussian Mixture Models (GMMs) to automatic speaker recognition. His current research is focused on application of speech technology to real-world scenarios and domain adaptation of speech systems. Dr. Reynolds is a Fellow of the IEEE, recipient of the 2017 MIT Lincoln Laboratory Technical Excellence Award, and a founding member of the Odyssey Speaker Recognition Workshop series.

Md Sahidullah (S'09, M'15) received his Ph.D. degree in the area of speech processing from the Department of Electronics & Electrical Communication Engineering, Indian Institute of Technology Kharagpur in 2015. Prior to that he obtained the Bachelors of Engineering degree in Electronics and Communication Engineering from Vidyasagar University in 2004 and the Masters of Engineering degree in Computer Science and Engineering from West Bengal University of Technology in 2006. In 2014-2017, he was a postdoctoral researcher with the School of Computing, University of Eastern Finland. In January 2018, he joined MULTISPEECH team, Inria, France as a post-doctoral researcher where he currently holds a starting research position. His research interest includes robust speaker recognition and spoofing countermeasures. He is also part of the organizing team of two Automatic Speaker Verification Spoofing and Countermeasures Challenges: ASVspoof 2017 and ASVspoof 2019. Presently, he is also serving as Associate Editor for the IET Signal Processing and Circuits, Systems, and Signal Processing.

Junichi Yamagishi (SM'13) is a professor at National Institute of Informatics in Japan. He is also a senior research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK. He was awarded a Ph.D. by Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. Since 2006, he has authored and co-authored over 250 refereed papers in international journals and conferences. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, and the Young Scientists' Prize from the Minister of Education, Science and Technology, the JSPS prize, the Docomo mobile science award in 2010, 2013, 2014, 2016, and 2018, respectively. He served previously as co-organizer for the bi-annual ASVspoof special sessions at INTERSPEECH 2013-9, the bi-annual Voice conversion challenge at INTERSPEECH 2016 and Odyssey 2018, an organizing committee member for the 10th ISCA Speech Synthesis Workshop 2019 and a technical program committee member for IEEE ASRU 2019. He also served as a member of the IEEE Speech and Language Technical Committee, as an Associate Editor of the IEEE/ACM TASLP and a Lead Guest Editor for the IEEE JSTSP SI on Spoofing and Countermeasures for Automatic Speaker Verification. He is currently a PI of JST-CREST and ANR supported VoicePersonae project. He also serves as a chairperson of ISCA SynSIG and as a Senior Area Editor of the IEEE/ACM TASLP.