ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan*

ASVspoof consortium
http://www.asvspoof.org/

January 15, 2019

1 Introduction

The ASVspoof 2019 challenge follows on from three special sessions on spoofing and countermeasures for automatic speaker verification held during INTERSPEECH 2013 [1], 2015 [2] and 2017 [3]. While the first edition in 2013 was targeted mainly at increasing awareness of the spoofing problem, the 2015 edition included the first challenge on the topic, accompanied by commonly defined evaluation data, metrics and protocols. The task in ASVspoof 2015 was to design countermeasure solutions capable of discriminating between bona fide (genuine) speech and spoofed speech produced using either text-to-speech (TTS) or voice conversion (VC) systems. The ASVspoof 2017 challenge focused on the design of countermeasures aimed at detecting replay spoofing attacks that could, in principle, be implemented by anyone using common consumer-grade devices.

The ASVspoof 2019 challenge extends the previous challenge in several directions. The 2019 edition is the first to focus on countermeasures for all three major attack types, namely those stemming from TTS, VC and replay spoofing attacks. Advances with regards the 2015 edition include the addition of up-to-date TTS and VC systems that draw upon substantial progress made in both fields during the last four years. Today, well-trained synthetic speech and converted voice is as good as perceptually indistinguishable from bona fide speech. ASVspoof 2019 thus aims to determine whether the advances in TTS and VC technology post a greater threat to automatic speaker verification and the reliability of spoofing countermeasures.

Advances with regards the 2017 edition concern the use of a far more controlled evaluation setup for the assessment of replay spoofing countermeasures. Whereas the 2017 challenge was created from the recordings of real replayed spoofing attacks, the use of an uncontrolled setup made results somewhat difficult to analyse. A controlled setup, in the form of replay attacks simulated using a range of real replay devices and carefully controlled acoustic conditions is adopted in ASVspoof 2019 with the aim of bringing new insights into the replay spoofing problem.

Last but not least, the 2019 edition aligns ASVspoof more closely with the field of automatic speaker verification. Whereas the 2015 and 2017 editions focused on the development and assessment of stand-alone countermeasures, ASVspoof 2019 adopts for the first time a new ASV-centric metric in the form of the tandem decision cost function (t-DCF) [4]. The adoption of the t-DCF does not,

*Document Version 0.4 (January 15, 2019)

however, imply that participation in ASVspoof 2019 requires any expertise in ASV. The cost of entry remains the same as that in the 2015 and 2017 editions; the challenge is still a stand-alone spoofing detection task. By combining spoofing detection scores with ASV scores produced by a fixed system designed by the organisers, adoption of the t-DCF as the primary evaluation metric will ensure that evaluation results and rankings will reflect the impact of spoofing and the performance of spoofing countermeasures upon the reliability of ASV.

This document provides a technical description of the ASVspoof 2019 challenge including details of the two evaluation conditions, namely those involving logical and physical access use-case scenarios. The document also provides details of the evaluation protocols, the new t-DCF metric, a common ASV system, baseline countermeasures, evaluation rules, submission procedures and schedule.

2 Technical objectives

As for previous editions of ASVspoof, the overriding objectives are to promote the development of reliable countermeasures that are able to distinguish between bona fide and spoofed speech. The initiative aims specifically to encourage the design of *generalised* countermeasures, i.e., countermeasures that perform well when faced with spoofing attacks of unpredictable nature. As with both 2015 and 2017 editions, the 2019 evaluation dataset contains training/development and evaluation partitions generated with different technologies (TTS and VC algorithms) and replay scenarios. Successful participation in ASVspoof 2019 therefore depends upon the development of countermeasures that generalise well to unseen forms of spoofing, namely attacks not seen in training/development data.

The specific technical objectives of ASVspoof 2019 are to:

- bring the evaluation of spoofing countermeasure solutions up to date with regards progress in technologies that can be used to implement advanced spoofing attacks;
- to provide new insights in terms of replay attacks and countermeasures, and
- to assess the impact of spoofing and of countermeasures upon the reliability of ASV systems.

Despite the adoption of a new primary metric for ASVspoof 2019, it is stressed that expertise in automatic speaker verification is not a pre-requisite to participation. The task is to design automatic systems capable of discriminating between bona fide and spoofed speech. ASV systems and scores required to support the new metric will be provided by the organisers, as will a comprehensive set of evaluation tools that will be needed in order to assess performance using the new metric. In order to maintain backwards compatibility, in order to observe the correlation between the new metric and the equal error rate (EER) metric used in previous editions, and in order to support applications beyond ASV[1], the EER is nonetheless retained as a secondary metric.

3 Data conditions: logical access

The data used for ASVspoof 2015 included spoofing attacks generated with text-to-speech (TTS) and voice conversion (VC) attacks generated with the state-of-the-art systems at that time. Since

[1]Use of the t-DCF explicitly links the evaluation to ASV applications. Despite this clear focus, the ASVspoof initiative has wider implications and should also attract interest in a more general sense involving fake audio detection, e.g., the fraudulent manipulation of smart-speaker, voice-driven interfaces. Use of an application-independent metric such as the EER is intended to support such alternative interests and use cases.

Table 1: Number of non-overlapping target speakers and number of utterances in training and development sets of the ASVspoof 2019 database. The duration of each utterance is in the order of one to two seconds.

Subset	#speakers		#utterances			
	Male	Female	Logical access		Physical access	
			Bona fide	Spoof	Bona fide	Spoof
Training	8	12	2,580	22,800	5,400	48,600
Development	8	12	2,548	22,296	5,400	24,300

then, considerable progress has been reported by both TTS and VC communities. The quality of well-trained synthetic speech and converted voice produced with today's technology is now perceptually indistinguishable from bona fide speech; the mean-opinion-score of synthetic speech produced with neural-network-based waveform modelling techniques known as WaveNet is comparable to that produced by humans [5]. The best performing system in the Voice Conversion Challenge 2018 [6] also produces converted voice signals with greatly improved naturalness and speaker similarity than the best performing systems in 2015. Since these technologies can be used to project convincing speech signals over the telephone, they pose substantial threats to the reliability of ASV. This scenario is referred to as logical access. The assessment of countermeasures, namely automatic systems that can detect non bona fide, spoofed speech produced with the latest TTS and VC technologies is therefore needed urgently.

The ASVspoof 2019 database for logical access is based upon a standard multi-speaker speech synthesis database called VCTK[2]. Genuine speech is collected from 107 speakers (46 male, 61 female) and with no significant channel or background noise effects. Spoofed speech is generated from the genuine data using a number of different spoofing algorithms. The full dataset is partitioned into three subsets, the first for training, the second for development and the third for evaluation. The number of speakers in the former two subsets is illustrated in Table 1. There is no speaker overlap across the three subsets regarding target speakers used in voice conversion or TTS adaptation.

3.1 Training and development data

The training dataset includes genuine and spoofed speech from 20 speakers (8 male, 12 female). Each spoofed utterance is generated according to one of 2 voice conversion and 4 speech synthesis algorithms. The voice conversion systems include those based on (i) neural-network-based and (ii) transfer-function-based methods. The speech synthesis systems were implemented with (i) waveform concatenation, (ii) neural-network-based parametric speech synthesis using source-filter vocoders and (iii) neural-network-based parametric speech synthesis using Wavenet. They were also built using publicly available toolkits called Merlin[3], CURRENT[4] and MaryTTS[5]. All data in the training set may be used to train spoofing detectors or countermeasures. All of the systems were built using the VCTK corpus.

The development dataset includes both genuine and spoofed speech from a subset of 20 speakers (8 male, 12 female). Spoofed speech is generated according to one of the same spoofing algorithms

[2]http://dx.doi.org/10.7488/ds/1994
[3]https://github.com/CSTR-Edinburgh/merlin
[4]https://github.com/nii-yamagishilab/project-CURRENNT-public
[5]http://mary.dfki.de

used to generate the training dataset. All data in the development dataset may be used for the design and optimisation of spoofing detectors/countermeasures.

Participants should be aware, however, that the spoofing algorithms used to create the development dataset are different from those used to generate the evaluation dataset. They are variants of the spoofing algorithms used to create the development dataset. The aim is therefore to develop a countermeasure which has potential to accurately detect new spoofed data generated with different or unseen spoofing algorithms.

3.2 Evaluation data

The evaluation data includes a set of unseen genuine and spoofed speech collected from multiple speakers. There are around up-to-80k trials including genuine and spoofed speech making the evaluation dataset approximately 4 GB in size. The recording conditions are exactly the same as those for the development dataset. Spoofed data are generated according to diverse unseen spoofing algorithms. However, they are variants of the spoofing algorithms used to generate the development dataset. Being intentionally different, they will give us useful insight into generalised performance of countermeasure models 'in the wild'. This is the same objective as that of ASVspoof 2015.

4 Data conditions: physical access

The physical access condition considers spoofing attacks that are performed at the sensor level. This implies that both bona fide and spoofed signals propagate through a physical space prior to acquisition. Spoofing attacks in this scenario are therefore referred to as replay attacks, whereby a recording of a bona fide access attempt is first captured, presumably surreptitiously, before being replayed to the ASV microphone. The microphone is considered to be a fundamental component of the ASV system. It cannot be bypassed; spoofing attacks cannot be injected into the system post-sensor, as is the case for the logical access scenario. This scenario conforms as much as possible to the ISO definition of presentation attacks [7]. The physical access scenario is relevant not just to ASV, but also to the emerging problem of fake audio detection that is faced in a host of additional applications including voice interaction and authentication with smart objects (e.g. smart-speakers and voice-driven assistants).

The results of ASVspoof 2017, the first ASVspoof edition to focus on replay attack detection, indicated that (i) replay attacks are effective in deceiving ASV systems and (ii) their reliable detection presents a substantial challenge. Nonetheless, progress in the development of countermeasures for replay detection has been rapid, with substantial improvements in performance being reported each year.

In an effort to emulate replay spoofing attacks, the 2017 challenge data was created from the real re-presentation and re-recording of a base corpus [8, 9] in a somewhat uncontrolled setup. This practice, coupled with the use of a base corpora that was captured with varying additive and convolutive noise, made results somewhat difficult to analyse. In seeking to improve upon the 2017 challenge, the 2019 edition is based upon *simulated* and carefully controlled acoustic and replay configurations [10, 11, 12]. The approach used to simulate room acoustics under varying source/receiver positions is that described in [13]. The same approach was applied successfully for data augmentation and x-vector speaker and speech recognition [14, 15]. Acoustic simulation, performed using Roomsimove[6] takes into account source directivity. Finally, replay devices are

[6] http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

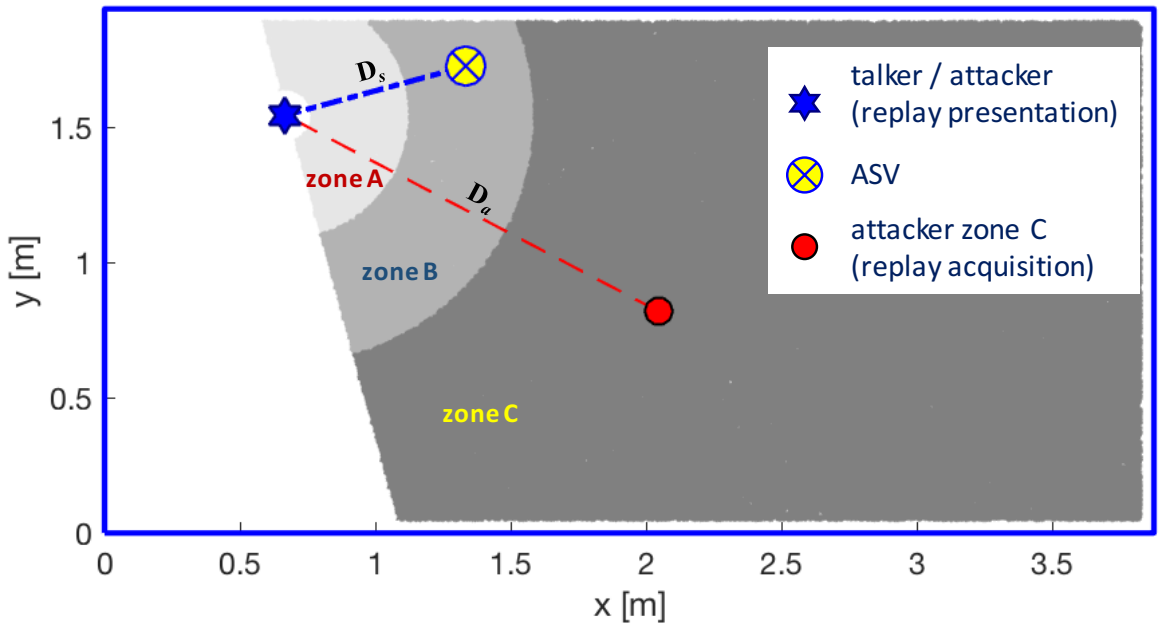

Figure 1: An illustration of the ASVspoof 2019 physical access scenario.

simulated using the generalised polynomial Hammerstein model and the Synchronized Swept Sine tool[7].

Just as is the case for the logical access scenario, the challenge in the physical access scenario involves the design of spoofing countermeasures that can distinguish between bona fide and spoofed ASV trials or access attempts. The physical access scenario is illustrated in Figure 1. The acoustic configuration in which is situated the ASV system (or any other device which is tasked with determining whether a recording is bona fide or spoofed/fake) is defined by the size of the physical space S (7.5m^2 in Figure 1) and the T60 reverberation time R. The height of each room has been set at 2.7m and the height of each source / receiver at 1.1m. The position of the ASV/device microphone is illustrated by the yellow circle in Figure 1. The position of a speaker, from hereon in referred to as a talker (in order to avoid potential confusion with the loudspeaker used to mount replay spoofing attacks), is illustrated by the blue star. Bona fide presentations or access attempts are made by the talker when positioned at a distance D_s from the microphone. The talker is assumed to speak in the direction of the microphone.

The manner by which replay attacks are mounted is also illustrated in Figure 1. A replay spoofing attack is mounted by (i) making a surreptitious recording of a bona fide access attempt and (ii) replaying the recording to the microphone. Whereas recordings are captured at a distance D_a from the talker, their subsequent presentation to the microphone is assumed to be make from the same distance D_s as bona fide access attempts. The recordings are assumed to be made in one of three *zones* illustrated in Figure 1, each representing a different interval of distances D_a from the talker. Recordings captured in zone A are expected to be of higher quality (better signal-to-reverberation ratio) than those made in zones B and C. Further setup details are available in the

[7] https://ant-novak.com/pages/sss/

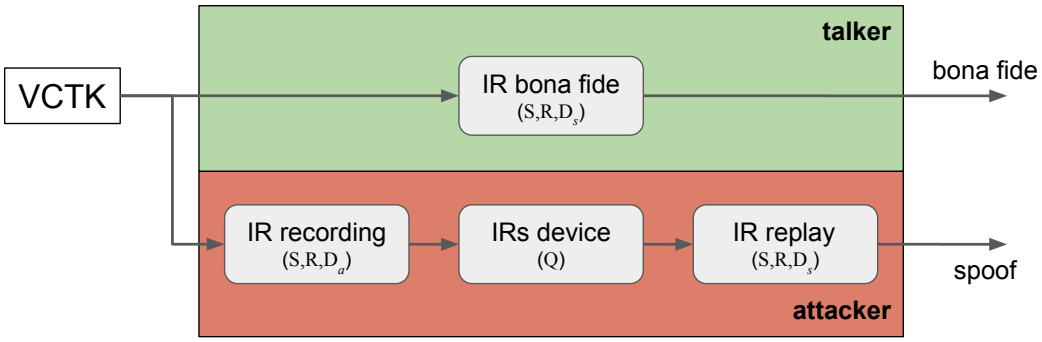

Figure 2: An illustration of the processes based on impulse response (IR) modeling approach, used in simulating the presentation of bona fide and replay spoofing access attempts to the microphone of an ASV system in a given physical space defined by room size S and by reverberation R. Bona fide access attempts are made at a distance D_s from the ASV microphone, whereas surreptitious recordings are made at a distance D_a from the talker before being presented to the ASV microphone. The effects of digital-to-analogue conversion, signal amplification and replay (using a loudspeaker) are all modelled, and are represented with the single, qualitative indicator Q. For simplicity, the attack presentation is performed at the same position of the talker, that is, the IR bona fide and the IR replay are the same.

physical access condition `ASVspoof2019_PA_instructions.txt` file.

Source recordings used to simulate bona fide access attempts are gathered directly from the same VCTK source corpus used for the logical access scenario. Source recording used to simulate replay attack access attempts are gathered from the same corpus, but pre-processed to simulate capture and replay according to the scenario described above. This pre-processing reflects the acoustic effects of propagation within a room comprising floors, ceiling and walls with different absorption coefficients, and hence different levels of reverberation. The simulation also models typical non-linear effects of loudspeakers that might be used to mount replay spoofing attacks The impact of the latter is represented according to a single device quality indicator denoted by Q.

The simulation procedure is illustrated in Figure 2. In brief, bona fide access attempts are characterised by propagation within an acoustic configuration defined by a room of size S with reverberation R and at a distance D_s between the the talker and the microphone. In contrast, replay spoofing access attempts are characterised by propagation in the same acoustic configuration between the talker and an attacker who captures a copy of the bona fide access attempt at a distance D_a from the talker. The recordings are then replayed using a loudspeaker of quality Q from the same distance from the microphone as the talker in a bona fide access attempt. These differences between bona fide and replay spoofing access attempts serve as the cues and opportunities to distinguish between them.

4.1 Training and development data

Training and development data is created according to a total of 27 different acoustic configurations. They comprise an exhaustive combination of 3 room sizes (S), 3 levels of reverberation (R) and 3 speaker-to-ASV microphone distances (D_s). There are 9 different replay configurations, comprising the 3 categories of attacker-to-speaker recording distances (D_a), and 3 categories of loudspeaker

quality (Q).

Table 1 shows the number of speakers and the number of bona fide and replay spoofing access attempts (utterances) in the training and development set. Both bona fide and replay spoofing access attempts in both training and development data partitions were generated according to the same set of randomly selected acoustic and replay configurations.

4.2 Evaluation data

In identical fashion to the logical access scenario, the evaluation set is disjoint in terms of speakers. There are in the order of 135k trials involving either bona fide or replay spoofed access attempts. Evaluation data, some 8.5 GB in size, is generated in the same manner as training and development data, albeit with different, randomly acoustic and replay configurations. Specifically, the set of room sizes, levels of reverberation, speaker-to-ASV microphone distances, attacker-to-speaker recording distances and loudspeaker qualities, while drawn from the same categories, will be different.

Accordingly, while the categories are the same and *known* a priori, the specific impulse responses used to simulate bona fide and replay spoofing access attempts are different or *unknown*. It is expected that reliable performance will only be obtained by countermeasures that generalise well to these conditions, i.e. countermeasures should not be over fitted to the specific acoustic and replay configurations observed in training and development data.

5 Performance measures

In terms of performance assessment, there is a key difference with the two past challenge editions which were focused on stand-alone assessment of spoofing countermeasure (CMs). Improvements in CM technology do not necessarily translate to improved complete system that involves co-operation of CM and ASV. Thus, ASVspoof 2019 will focus on assessment of *tandem* systems whereby a CM serves as a 'gate' to determine whether a given speech input originates from a bona fide (genuine) user, before passing it the main biometric verifier (the ASV system). Both the CM and ASV subsystems will make classification errors. The aim is to assess the performance of the tandem system as a whole taking into account not only the detection errors of both subsystems, but assumed prior frequencies of different types of users and implied monetary (or other similar) loss in the case one of the subsystems makes a mistake.

To this end, ASVspoof 2019 will adopt recently-proposed **tandem detection cost function** (t-DCF) [4] as its primary performance metric. The t-DCF is based on statistical detection theory and involves detailed specification of an envisioned *application* (in statistical terms - defined below). A key feature of t-DCF is the assessment of a tandem system while keeping the two subsystems (CM and ASV) isolated from each other: they can be developed independently of each other. For this reason, ASVspoof 2019 is no different from the two previous challenge editions; the participant will develop standalone CM and submit their detection scores to the organizers who will then evaluate the results. The ASV systems will be developed by the challenge organizers and the t-DCF framework combines the errors of both systems. This new evaluation strategy is intended to not only encourage development of countermeasures that work well when integrated with ASV, but to retain a low entrance barrier to the challenge. **No prior expertise in ASV technology is required. The participant does not need to develop, optimize or execute any ASV system at any stage**. To enable rapid kick-off, the organizers provide t-DCF scoring code and provide pointers to existing CM implementations. Besides t-DCF, the organizers will evaluate equal error rate (EER) of each submitted countermeasure as a secondary metric. This section details the evaluation metrics.

The ASVspoof 2019 challenge task is that of developing a two-class classifier. The positive (*bona fide*) class corresponds to speech segments of natural persons who should be accepted and the negative (*spoof*) class corresponds to attacks that should be rejected. The challenge participants will be provided labeled audio files. The protocol dictates simply a list of trial segments, each corresponding to a randomly named audio file from either bona fide or spoof class. For the development set the organizers provide ground-truth class labels and other metadata while the evaluation set samples will be unlabeled.

Both the t-DCF and EER metrics will be computed from *detection scores*, one score corresponding to a single audio file (evaluation trial). Participants should assign a real-valued, finite numerical value to each trial which reflects the support for two competing hypotheses, namely that the trial is bona fide or spoofed audio. Similar to the former two edition of ASVspoof, **high detection score should indicate bona fide and low score should indicate spoofing attack**. Following the widely-accepted convention in the speaker verification community, we encourage the detection scores to be in the form of *log-likelihood ratios* (LLRs) originating from a statistical detector ($\text{LLR}_t = \log p(\mathcal{X}_t|H_0) - \log p(\mathcal{X}_t|H_1)$. Here $\mathcal{X}_t$ is an audio segment corresponding to t^{th} trial and H_0 and H_1 correspond to, respectively, null hypothesis ($\mathcal{X}_t$ is a bona fide segment) and alternative hypothesis ($\mathcal{X}_t$ is a spoof segment). Nonetheless, LLRs are *not required* but any real-valued scores (such as inner products from a support vector machine) are acceptable; neither the minimum t-DCF nor the EER metric, as defined for the ASVspoof 2019 challenge, require the the scores to be interpreted as LLRs. Please note, however, that *hard binary decisions are not allowed*.

5.1 Primary metric: tandem detection cost function (t-DCF)

The primary metric of ASVspoof 2019 is *minimum tandem detection cost function*, defined below in Eq. (5). Similar to the EER metric (defined in the next Subsection), the minimum t-DCF does not require pre-setting a decision threshold. We explain here the details. The basic form of t-DCF adopted for ASVspoof 2019 takes the form[8]

$$\text{t-DCF}(s) = C_1 P^{\text{cm}}_{\text{miss}}(s) + C_2 P^{\text{cm}}_{\text{fa}}(s), \tag{1}$$

where $P^{\text{cm}}_{\text{miss}}(s)$ and $P^{\text{cm}}_{\text{fa}}(s)$ are, respectively, the *miss rate* and the *false alarm rate* of the CM system at threshold s. They are computed as

$$\begin{aligned} P^{\text{cm}}_{\text{miss}}(s) &= \frac{\#\{\text{bona fide trials with CM score} \leq s\}}{\#\{\text{Total bona fide trials}\}} \\ P^{\text{cm}}_{\text{fa}}(s) &= \frac{\#\{\text{spoof trials with CM score} > s\}}{\#\{\text{Total spoof trials}\}} \end{aligned}$$

The constants C_1 and C_2 are dictated by the t-DCF costs, priors, and the ASV system detection errors:

$$\begin{cases} C_1 = \pi_{\text{tar}} \left(C^{\text{cm}}_{\text{miss}} - C^{\text{asv}}_{\text{miss}} P^{\text{asv}}_{\text{miss}}\right) - \pi_{\text{non}} C^{\text{asv}}_{\text{fa}} P^{\text{asv}}_{\text{fa}} \\ C_2 = C^{\text{cm}}_{\text{fa}} \pi_{\text{spoof}} \left(1 - P^{\text{asv}}_{\text{miss,spoof}}\right) \end{cases}$$

Here, $C^{\text{asv}}_{\text{miss}}$ and $C^{\text{asv}}_{\text{fa}}$ are, respectively, the **costs** of ASV system to miss a target speaker and to accept a non-target (zero-effort impostor). Likewise, for the CM system we assign two costs $C^{\text{cm}}_{\text{miss}}$

[8]The interested reader may easily derive the simplified form (1) from the original formulation in [4] by assuming a fixed ASV system with fixed threshold, and by ignoring an additive constant that is the same for any evaluated CM system for fixed data and fixed ASV system. For the full disclosure, the discarded additive constant is given by $C_0 = \pi_{\text{tar}} C^{\text{asv}}_{\text{miss}} P^{\text{asv}}_{\text{miss}} + \pi_{\text{non}} C^{\text{asv}}_{\text{fa}} P^{\text{asv}}_{\text{fa}}$.

and $C_{\text{fa}}^{\text{cm}}$, respectively, for rejection of a human (bona fide) trial and acceptance of a spoof trial. Additionally, we assert **a priori** probabilities of target (π_{tar}), nontarget (π_{non}) and spoof (π_{spoof}) classes. Note that $\pi_{\text{tar}} + \pi_{\text{non}} + \pi_{\text{spoof}} = 1$. The costs and prior probabilities are fixed in advance to values shown in Table 2. Finally, $P_{\text{miss}}^{\text{asv}}$, $P_{\text{fa}}^{\text{asv}}$ and $P_{\text{miss,spoof}}^{\text{asv}}$ are **detection error rates of a fixed ASV system** at a specific ASV detection threshold (operating point). The former two are the traditional miss rate (proportion of rejected target users) and false alarm rates (proportion of accepted nontarget users). The last one is the miss rate of spoof samples against the ASV system (proportion of spoof samples rejected by the ASV). For further background on t-DCF, please refer to [4].

The raw t-DCF value can be difficult to interpret. Following the standard convention adopted in the NIST speaker recognition evaluation campaigns, it is useful to normalize the cost with that of a non-informative detector — a spoofing countermeasure which either accepts or rejects every input, whichever of these two yields lower t-DCF. Formally, we define the *normalized* t-DCF as

$$\text{t-DCF}_{\text{norm}}(s) = \frac{\text{t-DCF}(s)}{\text{t-DCF}_{\text{default}}}, \tag{2}$$

where $\text{t-DCF}_{\text{default}}$ is a *default* cost defined as $\text{t-DCF}_{\text{def}} = \min\{C_1, C_2\}$. The C_1 and C_2 are obtained from (1) by setting, respectively, $P_{\text{miss}}^{\text{cm}}(s) = 1$ and $P_{\text{fa}}^{\text{cm}}(s) = 0$ (CM threshold $s \to \infty$) and $P_{\text{miss}}^{\text{cm}}(s) = 0$ and $P_{\text{fa}}^{\text{cm}}(s) = 1$ (CM threshold $s \to -\infty$). In the former case, the normalized t-DCF is written as

$$\text{t-DCF}_{\text{norm}}(s) = P_{\text{miss}}^{\text{cm}}(s) + \alpha P_{\text{fa}}^{\text{cm}}(s), \tag{3}$$

where $\alpha = C_2/C_1$. In the latter case, we have

$$\text{t-DCF}_{\text{norm}}(s) = \beta P_{\text{miss}}^{\text{cm}}(s) + P_{\text{fa}}^{\text{cm}}(s), \tag{4}$$

where $\beta = C_1/C_2$. For ASVspoof 2019, we typically have $C_1 > C_2$ so the latter case (4) applies.

The reader familiar with the DCF evaluation framework used in NIST speaker recognition evaluations (SREs) recognizes immediately the similarities. In both SREs and ASVspoof 2019, the performance metric is a weighted sum of the miss and false alarm rates. The weight (either α or β) specifies the importance of one of the two error rates relative to unit weighting of the other one. There is, however, a very important difference: in the NIST SRE campaigns, weightage of the two errors is *known a priori*. In ASVspoof 2019, while all the t-DFC parameters (priors π_{tar}, π_{non}, π_{spoof} and costs $C_{\text{miss}}^{\text{asv}}$, $C_{\text{fa}}^{\text{asv}}$, $C_{\text{miss}}^{\text{cm}}$, $C_{\text{fa}}^{\text{cm}}$) will be fixed (see Table 2), the weighting of miss and false alarm rates of the CM depends *also* on the errors of the ASV system ($P_{\text{miss}}^{\text{asv}}$, $P_{\text{fa}}^{\text{asv}}$, $P_{\text{miss,spoof}}^{\text{asv}}$). The challenge participant cannot interact with the ASV system and will have limited information about it. They *will* be provided labeled ASV scores of the development set to demonstrate t-DCF computation, but will *not* be provided any evaluation set ASV scores. This new evaluation strategy is intended to encourage development of countermeasures that generalize not only across datasets but across different ASV systems. The organizer's ASV system will be scored exactly on the same way on development and evaluation data — but the evaluation set error rates may be different from those of the development set. Besides a common ASV system used for producing challenge rankings, the organizers may include additional ASV systems to address generalization across ASV system. What is common to all the ASV systems is that their detection threshold will be set at the equal error (EER) point, at which $P_{\text{miss}}^{\text{asv}} = P_{\text{fa}}^{\text{asv}}$.

The normalized t-DCF as defined above is a function of the CM threshold s. Similar to the past two challenge editions, ASVspoof 2019 does not focus on threshold setting (calibration). We fix the threshold of each evaluated CM to its optimal value corresponding to perfect calibration. Namely,

Table 2: t-DCF cost function parameters assumed in ASVspoof 2019.

	Priors			**ASV costs**		**CM costs**	
Evaluation condition	π_{tar}	π_{non}	π_{spoof}	$C_{\text{miss}}^{\text{asv}}$	$C_{\text{fa}}^{\text{asv}}$	$C_{\text{miss}}^{\text{cm}}$	$C_{\text{fa}}^{\text{cm}}$
Logical access control	0.9405	0.0095	0.05	1	10	1	10
Physical access control	0.9405	0.0095	0.05	1	10	1	10

challenge rankings will be based on *minimum* normalized t-DCF, defined as

$$\text{t-DCF}_{\text{norm}}^{\text{min}} = \text{t-DCF}_{\text{norm}}(s_*), \tag{5}$$

where $s_* = \arg\min_s \text{t-DCF}_{\text{norm}}(s)$ is the optimal threshold determined from the evaluation set using the ground truth.

Similar to the last ASVspoof2017 evaluation, performance will be computed from detection scores pooled across all trials within the given challenge sub-task (logical or physical attack). That is, within one of these two sub-tasks, all the bona fide trials in the evaluation set will be used to obtain $P_{\text{miss}}^{\text{cm}}(s)$ and all the spoof trials will be used to obtain $P_{\text{fa}}^{\text{cm}}(s)$. This strategy is intended to encourage development of countermeasures that calibrate well across different variants of attacks. That is, to develop countermeasures whose scores are mutually compatible across a wide range of conditions. The two challenge sub-tasks, however, will be scored independent of each other: the organizers will neither pool the logical and physical access scores, nor do they attempt to define a single scalar summary of the evaluation. Participants may register to one, or both, of the two challenge subtasks as detailed below. The two sub-challenges should be considered independent tasks and a participant interested in both tasks is welcome to use completely different technologies to address each task.

The t-DCF parameters used in ASVspoof 2019 are displayed in Table 2. These selected values, same across the two challenge subtasks, are arbitrary but intended to be representative of *user authentication* applications. For instance, a banking domain may involve a very large number of authentication requests, most of which represent genuine target users; therefore, we have asserted high target prior. Priors of both 'undesired' users (nontargets and attackers) are set to low values for the same reason. The costs, in turn, reflect the idea that false acceptances are more detrimental than false rejections. For the reader who is less familiar with the detection cost function approach to performance evaluation, note that the priors do not necessarily reflect the empirical proportions of different classes in a given dataset; that is, one should *not* interpret Table 2 to suggest that ASVspoof 2019 evaluation data contains 98 % of target speaker segments. The prior over the user types is an abstraction that reflects one's prior belief on the prevalence of targets, nontargets and spoofs in a general population, which is different from an evaluation design (as the reader may have already concluded looking at Table 1).

5.2 Secondary metric: equal error rate (EER)

The secondary evaluation metric is *equal error rate* (EER) which was used in the past two editions of ASVspoof. EER corresponds to CM threshold s_{EER} at which the miss and false alarm rates equal each other[9], i.e. $\text{EER} = P_{\text{fa}}^{\text{cm}}(s_{\text{EER}}) = P_{\text{miss}}^{\text{cm}}(s_{\text{EER}})$. Similar to t-DCF, EER will be computed from pooled evaluation trials within the given sub-task.

[9]One may not find such threshold exactly as $P_{\text{fa}}(s)$ and $P_{\text{miss}}(s)$ change in discrete steps. You may use $\theta_{\text{EER}} = \arg\min_\theta |P_{\text{fa}}^{\text{cm}}(s) - P_{\text{miss}}^{\text{cm}}(s)|$ or more advanced methods such as EER based on convex hull (ROCCH-EER)

6 Common ASV system and baseline countermeasures

6.1 Common ASV system

The common ASV system is based on speaker embedding technology based on deep learning techniques. The organizers will provide labeled ASV scores of this system for the development set to demonstrate t-DCF computation. The ASV scores for the evaluation set will not be made available during the evaluation as the final CM evaluation will be done by the organizers. The details of the ASV system will be revealed after the evaluation has been concluded.

6.2 Baseline countermeasures

A Matlab package implementing two baseline systems is available for download from the ASVspoof website `http://www.asvspoof.org`. They are based on two different acoustic frontends, namely linear frequency cesptral coefficients [16] (LFCC) and constant Q cepstral coefficients [17] (CQCC). Both use a GMM binary classifier. Details will be made available through a README file within the package. Table 3 shows bona fide / spoofing detection performance of the baseline systems trained on the ASVspoof 2019 train set and tested on the ASVspoof 2019 development set, for logical and physical access conditions. Performance is measured in terms of minimum t-DCF and EER.

Table 3: t-DCF and EER results for two baseline countermeasures and both logical and physical access scenarios.

	Logical access		**Physical access**	
	$\text{t-DCF}^{\text{min}}_{\text{norm}}$	EER (%)	$\text{t-DCF}^{\text{min}}_{\text{norm}}$	EER (%)
LFCC-GMM	0.0663	2.71	0.2554	11.96
CQCC-GMM	0.0123	0.43	0.1953	9.87

7 Evaluation rules

The submission categories of the challenge, along with the explanations, are provided in Table 4. The main differences from the past two challenge editions are inclusion of two different sub-challenges, omission of 'flexible' submission category, the requirement to submit development scores, and the requirement to submit single sub-system scores (in addition to fusion/ensemble system scores).

For internal experiments, the participants are free to use the development part as they wish. It can be used for optimizing classifier parameters or re-partitioned freely for custom training/dev divisions. For instance, the training and development files can be pooled to form a larger training set or used to composing cross-validation experiments. **For the submission to ASVspoof 2019, however, we require the participants to respect the official protocols defined in Table 5 and submit both development and evaluation set scores.** That is, the submission entries sent to the organizers must correspond to countermeasures trained using the training protocol file (`ASVspoof2019.**.cm.train.trn.txt`) and scored on both the development and evaluation sets (`ASVspoof2019.**.cm.dev.trl.txt` and `ASVspoof2019.**.cm.eval.trl.txt`). If a participant wants to *exclude* some of the training files (*e.g.* to choose more informative training examplars), this *is* allowed, as long as clearly mentioned in the system description. The primary challenge

implemented in the Bosaris toolkit, `https://sites.google.com/site/bosaristoolkit/`. In ASVspoof2019, we use a simplified nearest-neighbor type computation demonstrated in the scoring package.

Table 4: Each participant can submit up to sixteen (16) different scores on the ASVspoof 2019 challenge. A participant can register to either one of the two sub-challenges (LA and PA) of their own choice, or to both For each registered sub-challenge, the participant is required to submit four (4) sets of detection scores to the organizers. They include *Primary* and *Single system* submissions on both development and evaluation parts respecting strictly the common protocols. The exact same CM systems must be scored on the development and evaluation parts without any adaptation. **One, and exactly one, system must be designated as the primary submission, to be used for challenge ranking**. The two different sub-challenges may, however, use two different techniques. The primary system may be an ensemble classifier consisting of multiple subsystems whose output scores are combined, for instance. The 'single system' scores, in turn, should correspond to exactly one of the subsystems in such ensemble. A participant working only on a single CM system should include both primary and single system scores, where the latter is simply a duplicate of the former. Additionally, each participant is allowed to submit two more *contrastive* systems, perhaps for their additional analyses. The organizers will evaluate all the submitted systems, but **the primary system scores on the evaluation set will be the ones used to derive challenge rankings**. Finally, the ASV system scores will be computed internally by the organizers. The participants can neither submit their ASV scores nor interact with the ASV system used in tandem with the submitted CM systems. Finally, the use of any other external data resources is forbidden in any form: all the CM subsystems (and possible internally used ASV systems) must be trained and optimized using *only* the provided dataset. This concerns all the data-driven components of the CM systems.

LOGICAL ACCESS (LA) sub-challenge			
	ASV	Countermeasure (CM) scores	
Submission	scores	Dev	Eval
Primary	—	**Required**	**Required**
Single system	—	**Required**	**Required**
Contrastive1	—	Optional	Optional
Contrastive2	—	Optional	Optional
PHYSICAL ACCESS (PA) sub-challenge			
	ASV	Countermeasure (CM) scores	
Submission	scores	Dev	Eval
Primary	—	**Required**	**Required**
Single system	—	**Required**	**Required**
Contrastive1	—	Optional	Optional
Contrastive2	—	Optional	Optional

ranking evaluation will be based only on the evaluation set scores similar to the past two editions, but the organizers plan to carry out additional analyses of each system across the development and evaluation scores to address generalization in terms of data, ASV system, and possibly score calibration (threshold setting).

To ensure that the submitted CM systems will be evaluated on equal footing in terms of their training data, **the use of any other external datasets/resources is forbidden**. This includes, but is not limited to, the use of any other public or in-house corpora, found speech/spoof/noise samples, externally trained models, feature vectors (or other statistical descriptors) extracted from external data, or externally trained speech activity detectors. All the submitted CMs and all their components must be constructed using data in the training and development portions *only*. The use of data augmentation is permitted, but is limited strictly to the use of official ASVspoof 2019

Table 5: Protocol definitions of ASVspoof 2019. Numbers of utterances are also listed in Table 1.

		Name of protocol file	#utterances (bona fide + spoof)
LA	Training	`ASVspoof2019.LA.cm.train.trn.txt`	25380 (2580 + 22800)
	Development	`ASVspoof2019.LA.cm.dev.trl.txt`	24844 (2548 + 22296)
	Evaluation	`ASVspoof2019.LA.cm.eval.trl.txt`	-
PA	Training	`ASVspoof2019.PA.cm.train.trn.txt`	54000 (5400 + 48600)
	Development	`ASVspoof2019.PA.cm.dev.trl.txt`	29700 (5400 + 24300)
	Evaluation	`ASVspoof2019.PA.cm.eval.trl.txt`	-

training and development data with external, *non-speech* data, *e.g.* impulse responses. The use of any additional, external *speech* data is strictly forbidden; the only source speech data, be it bona fide (genuine) or spoofed (replayed, synthesized or converted), that can be used for either logical access or physical access conditions, is that provided in the ASVspoof 2019 corpus itself. This rule applies also to the use, for any purpose, of VCTK, ASVspoof 2015, or ASVspoof 2017 data. Their use, for any purpose, is also strictly forbidden. Submissions to the logical access condition should not use data corresponding to the physical access condition just as those to the physical access condition should not use data corresponding to the logical access condition.

Any participant who wishes to optimize their spoofing countermeasures jointly with their custom ASV systems are free to use *any* data resources for their internal development purposes, including public and proprietary ones. Please note, however, that you cannot submit any ASV scores to the challenge. The official challenge results will be based on ASV systems developed by the organizers on undisclosed data that will be revealed after completion of the challenge.

The countermeasure scores produced for any one trial must be obtained using *only* the data in that trial segment. The use of data from any other trial segment is strictly prohibited. Therefore, the use of techniques such as normalization over multiple trial segments and the use of trial data for model adaptation is not allowed. Systems must therefore process trial lists segment-by-segment without access to past or future trial segments.

8 Registration and submission of results

8.1 General mailing list

All participants and team members are encouraged to subscribe to the general mailing list. Subscribe by sending an email to:

sympa@asvspoof.org

with 'subscribe asvspoof2019' as the subject line. Successful registrations are confirmed by return email. To post messages to the mailing list itself, emails should be addressed to:

asvspoof2019@asvspoof.org

Subscribers can unsubscribe at any time by following the instructions in the subscription confirmation email.

8.2 Registration and database download

Participants/teams are requested to register for the evaluation. Registration should be performed once only for each participating entity and by sending an email to:

registration@asvspoof.org

with 'ASVspoof 2019 registration' as the subject line. The mail body should include (i) the name of the team, (ii) the name of the contact person, (iii) their country, (iv) their status (academic/non-academic) and (v) the challenge scenario(s) for which they wish to participate (indicative only). Participants may register for one or both challenge scenarios. The organisers aim to validate all registrations within 48 hours. Validated registrations will be confirmed by email.

URIs for the downloading of training/development and evaluation data will be communicated to registered contact persons only. Since download bandwidth is limited and shared among all, participants are kindly requested to download one package at a time and only once per participating team.

8.3 Submission of results

Each participant/team should submit (1) brief system description(s) and (2) up to eight score file(s) per data condition as specified above. Both will be shared among other participants after the evaluation period. The system descriptions should be PDF files detailing the countermeasure approach (features and classifiers etc.) and related technologies. The score file is an ASCII text file. Each line of the score file should contain two entries, separated by white space: the unique trial segment identifier (without the `.flac` extension) and the detection score. An example score file for the physical access is shown below:

```
...
PA_E_10000001 1.571182
PA_E_10000002 -2.735763
PA_E_10000003 -4.084447
PA_E_10000004 77.868048
...
```

Please use the following convention to name the submitted files:

- PDF file of system descriptions: `<team>_<data-condition>_<submission>.pdf`
- Scores on the development set: `<team>_<data-condition>_<submission>_dev.txt`
- Scores on the evaluation set: `<team>_<data-condition>_<submission>_eval.txt`

The field `<data-condition>` can be `LA` for logical access or `PA` for physical access. The field `<submission>` can be `primary`, `single`, `contrastive1`, or `contrastive2`, depending on the submission type listed in Table 4. With all the system descriptions and score file(s) ready, please create a `<team>.tar.gz` or `<team>.zip` package. Here is an example package:

```
MYTEAM.tar.gz
|- MYTEAM_PA_primary.pdf
|- MYTEAM_PA_primary_dev.txt
|- MYTEAM_PA_primary_eval.txt
|- MYTEAM_PA_single.pdf
|- MYTEAM_PA_single_dev.txt
```

```
|- MYTEAM_PA_single_eval.txt
|- MYTEAM_PA_contrastive1.pdf
|- MYTEAM_PA_contrastive1_dev.txt
|- MYTEAM_PA_contrastive1_eval.txt
```

The package should be submitted to a Dropbox account whose link will be announced later on.

9 ASVspoof 2019 special session at Interspeech 2019

On the topic of scientific publications, we will organize a special session on ASVspoof 2019 challenge at Interspeech 2019 in Austria (http://www.interspeech2019.org) and hence we encourage all of you to submit your papers to the conference.

Please note that the review process of papers submitted to the above special session is the same as regular papers. Scientific novelty of your proposed idea should be clearly described and fairly evaluated. Review judgment is independent from the ranking achieved in the challenge.

In case your team decides not to submit any paper to the above conferences, we encourage you to write a technical document and make such document publicly available via your organization's repository or ArXiv. Such document makes the scientific analysis of our challenge results more meaningful.

We look forward to seeing a lot of academic papers written by many of you.

10 Schedule

- Release training and development materials to participants: 19th December 2018
- Participant registration deadline: 8th February 2019
- Release evaluation data to participants: 15th February 2019
- Deadline for participants to submit evaluation scores: 22nd February 2019
- Organisers return results to participants: 15th March 2019
- Interspeech paper submission deadline: 29th March, 2019
- Interspeech 2019 (Graz, Austria): 15th–19th Sept, 2019

11 Glossary

Generally, the terminologies of automatic speaker verification are consistent with that in the NIST speaker recognition evaluation. New terminologies specific to spoofing and countermeasure assessment are listed as follows:

Spoofing attack: An adversary, also named impostor, attempts to deceive an automatic speaker verification system by impersonating another enrolled user in order to manipulate speaker verification results.

Anti-Spoofing: Also known as countermeasure. It is a technique to countering spoofing attacks to secure automatic speaker verification.

Bona fide trial: A trial in which the speech signal is recorded from a live human being without any modification.

Spoof trial: In the case of the physical access, a spoofing trial means a trial in which an authentic human speech signal is first played back through an digital-to-analog conversion process and then re-recorded again through analog-to-digital channel; an example would be using smartphone A to replay an authentic target speaker recording through the loudspeaker of A to the microphone of smartphone B that acts as the end-user terminal of an ASV system. In the case of the logical access, a spoofing trial means a trial in which the original, genuine speech signal is modified automatically in order to manipulate ASV.

12 Acknowledgement

The new ASVspoof 2019 database is the result of more than six months of intensive work including contributions from leading academic and industrial research laboratories. The ASVspoof consortium would like to strongly thank the following 14 organizations and 27 persons who contributed to this database:

- Ms. Avashna Govender and Dr. Srikanth Ronanki from University of Edinburgh, UK,
- Prof. Tomoki Toda and Mr. Yi-Chiao Wu from Nagoya University, Japan, Japan, Mr. Wen-Chin Huang, Mr. Yu-Huai Peng, and Dr. Hsin-Te Hwang from Academia Sinica, Taiwan.
- Prof. Zhen-Hua Ling and Mr. Jing-Xuan Zhang from University of Science and Technology of China, China, and Mr. Yuan Jiang and Ms. Li-Juan Liu from iFlytek Research, China.
- Dr. Ingmar Steiner from Saarland University and DFKI GmbH, Germany and Dr. Sébastien Le Maguer from Adapt centre, Trinity College Dublin, Ireland
- Dr. Kou Tanaka and Dr. Hirokazu Kameoka from NTT Communication Science Laboratories, Japan
- Mr. Kai Onuma, Mr. Koji Mushika, and Mr. Takashi Kaneda from HOYA, Japan
- Dr. Markus Becker, Mr. Fergus Henderson, Dr. Rob Clark from the Google Text-To-Speech team, Google LLC and Dr. Yu Zhang, Dr. Quan Wang from the Google Brain team, and Deepmind, Google LLC
- Prof. Driss Matrouf and Prof. Jean-François Bonastre from LIA, University of Avignon, France
- Mr. Lauri Juvela and Prof. Paavo Alku from Aalto University, Finland

This work was partially supported by JST CREST Grant Number JPMJCR18A6 (VoicePersonae project), Japan and by MEXT KAKENHI Grant Numbers (16H06302, 16K16096, 17H04687, 18H04120, 18H04112, 18KT0051), Japan; The work is also partially supported by research funds received from the French Agence Nationale de la Recherche (ANR) in connection with the bilateral VoicePersonae (with JST CREST in Japan) and RESPECT (with DFG in Germany) collaborative research projects. The is also supported by Academy of Finland (proj. no. 309629 entitled "NOTCH: NOn-cooperaTive speaker CHaracterization"). The authors at the University of Eastern Finland also gratefully acknowledge the use of computational infrastructures at CSC – IT Center for Science, and support of NVIDIA Corporation with the donation of the Titan V GPU used in this research. The work is also partially supported by Region Grand Est, France.

References

[1] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc.*, Lyon, France, August 2013, pp. 925–929.

[2] 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, Annual Conf. of the Int. Speech Comm. Assoc.*, Dresden, Germany, September 2015, pp. 2037–2041.

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

[4] T. Kinnunen, K. 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*, Les Sables d'Olonne, France, June 2018.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[6] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 195–202. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-28

[7] "ISO/IEC 30107. Information technology – biometric presentation attack detection," Standard, 2016.

[8] K. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, M. J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc.*, 2015, pp. 2996–3000.

[9] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, and K. A. Lee, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017.

[10] D. R. Campbell, K. J. Palomäki, and G. Brown, "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching." *Computing and Information Systems Journal, ISSN 1352-9404*, vol. 9, no. 3, 2005.

[11] E. Vincent. (2008) Roomsimove. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

[12] A. Novak, P. Lotton, and L. Simon, "Synchronized swept-sine: Theory, application, and implementation," *J. Audio Eng. Soc*, vol. 63, no. 10, pp. 786–798, 2015. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18042

[13] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[16] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc.*, Dresden, Germany, 2015, pp. 2087–2091.

[17] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech Language*, vol. 45, pp. 516 – 535, 2017.