Open access • Journal Article • DOI:10.1109/TBIOM.2021.3059479

ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech — Source link

Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen ...+6 more authors

Institutions: Institut Eurécom, National Institute of Informatics, University of Eastern Finland, Nuance Communications ...+2 more institutions

Published on: 18 Feb 2021

Topics: Spoofing attack, Replay attack, Speaker recognition and Electronic mail

ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech

Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi Kinnunen, Ville Vestman, Massimiliano Todisco, Hector Delgado, Md Sahidullah, Junichi Yamagishi, Kong Aik Lee

ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech

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

Abstract—The ASVspoof initiative was conceived to spearhead research in anti-spoofing for automatic speaker verification (ASV). This paper describes the third in a series of bi-annual challenges: ASVspoof 2019. With the challenge database and protocols being described elsewhere, the focus of this paper is on results and the top performing single and ensemble system submissions from 62 teams, all of which out-perform the two baseline systems, often by a substantial margin. Deeper analyses shows that performance is dominated by specific conditions involving either specific spoofing attacks or specific acoustic environments. While fusion is shown to be particularly effective for the logical access scenario involving speech synthesis and voice conversion attacks, participants largely struggled to apply fusion successfully for the physical access scenario involving simulated replay attacks. This is likely the result of a lack of system complementarity, while oracle fusion experiments show clear potential to improve performance. Furthermore, while results for simulated data are promising, experiments with real replay data show a substantial gap, most likely due to the presence of additive noise in the latter. This finding, among others, leads to a number of ideas for further research and directions for future editions of the ASVspoof challenge.

Index Terms—Spoofing, countermeasures, presentation attack detection, speaker recognition, automatic speaker verification.

1 Introduction

It is well known that automatic speaker verification (ASV) systems are vulnerable to being manipulated by spoofing, also known as presentation attacks [1]. Spoofing attacks can enable a fraudster to gain illegitimate access to resources, services or devices protected by ASV technology. The threat from spoofing can be substantial and unacceptable. Following the first special session on anti-spoofing held in 2013 [2], the effort to develop spoofing countermeasures, auxiliary systems which aim to protect ASV technology by automatically detecting and deflecting spoofing attacks, has been spearheaded by the ASVspoof initiative[1].

ASVspoof 2019 [3], the most recent of three editions and the focus in this paper, was the first to include all three major forms of spoofing attacks involving speech synthesis, voice conversion and replay, in separate logical and physical access scenarios. It also brought several advances with respect to previous editions. First, ASVspoof 2019 aimed to explore whether advances in speech synthesis and voice conversion technologies pose a greater threat to ASV reliability; the latest of these techniques, *e.g.* neural network-based waveform modelling techniques, can produce synthetic and converted speech that is perceptually indistinguishable from bona fide speech. Second, the 2019 edition explored replay attacks using a far more controlled evaluation setup in the form of simulated replay attacks and carefully controlled acoustic conditions. Third, the database is substantially larger than the previous ASVspoof databases and considerably more diverse in terms of attack algorithms. With a comprehensive description of the database available in a published companion paper [4], only a brief description is provided in the current article. The focus here is instead upon challenge results and findings.

Whereas previous editions of ASVspoof utilised the equal error rate (EER) metric to judge performance, the 2019 edition shifted to the ASV-centric tandem detection cost function (t-DCF) metric [5], [6]. While the latter reflects the impact of both spoofing and countermeasures

Manuscript received MMMM DD, YYYY; revised MMMM DD, YYYY; accepted MMMM, DD YYYY. Date of publication MMMM DD, YYYY; date of current version MMMM DD, YYYY. This work was supported by a number of projects and funding sources: VoicePersonae, supported by the French Agence Nationale de la Recherche (ANR) and the Japan Science and Technology Agency (JST) with grant No. JPMJCR18A6; RESPECT, supported by the ANR; the NOTCH project (no. 309629), supported by the Academy of Finland; Region Grand Est, France. (Corresponding author: Andreas Nautsch.)

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

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

T. Kinnunen and V. Vestman are with University of Eastern Finland, Joensuu campus, Länsikatu 15, FI-80110 Joensuu, Finland. E-mail: {tkinnu,ville.vestman}@uef.fi

H. Delgado is with Nuance Communications, C/ Gran Vía 39, 28013 Madrid, Spain. E-mail: hector.delgado@nuance.com.

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

*K. A. Lee is with Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632. E-mail: lee_kong_aik@i2r.a-star.edu.sg*

1. https://www.asvspoof.org

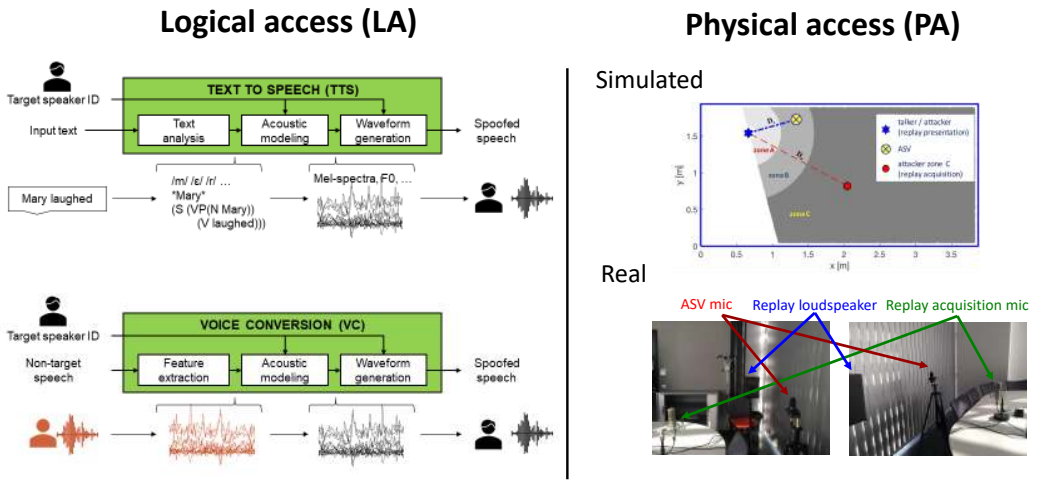


Fig. 1. The ASVspoof 2019 challenge featured four different types of spoofed audio data. The LA scenario contains text-to-speech and voice conversion attacks. In the PA scenario attackers acquire a recording of the target speaker which is then replayed to the ASV system. Both *simulated* and *real* replay attacks are considered. The former refers to simulated acoustic environments/rooms with specified dimensions and controllable reverberation whereas the latter contains actual replay recordings collected at three different sites. Real replay attacks were included in the test set (but excluded from challenge ranking). This paper describes challenge results for all four setups illustrated.

upon ASV performance, participation still calls only for the development and optimisation of *countermeasures*. Strong performance depends upon generalisation, namely countermeasures that perform well in the face of spoofing attacks not seen in training or development data.

Two different baseline systems were provided for the 2019 edition. With data, protocols and metrics being different to those of previous editions, progress is judged in terms of performance relative to the two baseline systems which provide some level of continuity or consistency with previous challenge editions. The article describes the top five single and fused systems for both challenge scenarios, provides insights into the most successful countermeasure (CM) techniques and assesses their impact upon ASV reliability. Finally, the article outlines priorities for the ASVspoof initiative looking to the future, including ideas for the next edition – ASVspoof 2021.

2 Challenge Outline

This section describes the logical and physical access ASVspoof 2019 challenge scenarios (see Fig. 1), the challenge rules, the new t-DCF metric, and baseline CM and ASV systems. Since it is described elsewhere [4], the ASVspoof 2019 database is not described here. Briefly, it is sourced from the *Voice Cloning Toolkit* (VCTK) corpus [7], a multi-speaker, native-English speech database of read sentences recorded in a hemi-anechoic chamber.

2.1 Logical access

Logical access (LA) control implies a scenario in which a remote user seeks access to a system or service protected by ASV. An example is a telephone banking service to which attackers may connect and then send synthetic or converted voice signals directly to the ASV system while bypassing the microphone, *i.e.* by injecting audio into the communication channel post sensor.

The LA subset of the ASVspoof 2019 database was created using a diverse array of 17 text-to-speech (TTS), voice conversion (VC) and hybrid systems. Waveform generation methods vary from waveform concatenation to neural network-based waveform modelling techniques including WaveNet [8]. Acoustic models also vary from Gaussian mixture models to advanced sequence-to-sequence neural networks. Some are constructed using popular open-source toolkits while others are selected based on their superior evaluation results reported in the Voice Conversion Challenge [9] or other literature. Six of these systems are designated as known spoofing algorithms/attacks, with the other 11 being designated as unknown spoofing attacks. Among the 6 known attacks there are 2 VC systems and 4 TTS systems. The 11 unknown attacks comprise 2 VC, 6 TTS and 3 hybrid TTS-VC systems for which VC systems are fed with synthetic speech. Known attacks are used to generate training and development data. Unknown attacks and two of the known attacks are used to generate evaluation data. Attacks are referred to by attack identifiers (AIDs): A1 – A19. Full details of the LA setup, attack groups and analysis are provided in [4].

2.2 Physical access

In the physical access (PA) scenario, spoofing attacks are presented to a fixed microphone which is placed in an environment in which sounds propagate and are reflected from obstacles such as floors and walls. Spoofing attacks in this scenario are referred to as replay attacks and match the ISO definition of *presentation attacks* [10]. The PA scenario, is based upon *simulated* and carefully controlled acoustic and replay configurations [11], [12], [13]. The approach used to simulate room acoustics under varying source/receiver positions is inspired from the approach reported in [14] and based upon an image-source method [15]. Acoustic simula-

TABLE 1
Submission categories for the ASVspoof 2019 challenge for both LA and PA scenarios and primary, single and contrastive submission. Only results for single and primary systems are discussed in this paper.

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

tions are performed using Roomsimove[2], while the replay device effects are simulated using the generalised polynomial Hammerstein model and the Synchronized Swept Sine tool[3].

The ASV system is used within a noise-free acoustic environment defined by: the room size S; the $T60$ reverberation time; the talker-to-ASV[4] distance D_{s}. Each parameter is categorised into three different intervals. The room size S is categorised into: (a) small rooms of size 2-5 m^2; (b) medium rooms of size 5-10 m^2; (c) large rooms of size 10-20 m^2. The $T60$ reverberation time is categorised into: (a) low 50-200 ms; (b) medium 200-600 ms; (c) large 600-1000 ms. The talker-to-ASV distance D_{s} is categorised into: (a) low 10-50 cm; (b) medium 50-100 cm; (c) large 100-150 cm. This results in 27 acoustic configurations denoted by environment identifiers (EIDs) (*aaa*, *aab*, ..., *ccc*).

A replay spoofing attack is mounted through the making of a surreptitious recording of a bona fide access attempt and then the presentation of this recording to the ASV microphone. Attackers acquire recordings of bona fide access attempts when positioned at an attacker-to-talker distance D_{a} from the talker whereas the presentation of recording is made at the talker-to-ASV distance D_{s} using a playback device of quality Q. D_{a} is categorised into three different intervals: (A) low 10-50 cm; (B) medium 50-100 cm; (C) large 100-150 cm. Q is categorised into three quality groups: (A) perfect quality, *i.e.* a Dirac impulse response; (B) high quality; (C) low quality. Their combination results in 9 attack configurations denoted by attack identifiers (AIDs) (*AA*, *AB*, ..., *CC*). Full details of the PA setup are also provided in [4].

2.3 Rules

The submission categories for ASVspoof 2019 are illustrated in Table 1. Participants were permitted to submit up to 4 different score sets (or 8, counting sets for development and evaluation separately) for the LA scenario and an additional 4 for the PA scenario, with the use of different systems being permitted for each. Two of these score sets are required and include *primary* and *single* system scores. Score submissions were required for both the development and evaluation subsets defined in the ASVspoof 2019 protocols. Scores for corresponding development and evaluation subsets were required to be derived using identical CM systems without any adaptation. Ensemble classifiers consisting of multiple sub-systems whose output scores are combined were permitted for primary systems only. Single system scores were required to be one of the sub-systems in the ensemble (normally the single, best performing). While participants were permitted to submit scores for an additional two *contrastive* systems, only results for single and primary systems are presented in this paper.

ASV scores used for scoring and ranking were computed by the organisers using separate ASV protocols. The use of external data resources was forbidden: all systems designed by the participants were required to be trained and optimised using *only* the relevant ASVspoof 2019 data and protocols. The only exception to this rule is the use of data augmentation, but only then using ASVspoof 2019 training and development data with external, *non-speech* data, *e.g.* impulse responses. Use of LA data for PA experiments and vice versa was also forbidden.

Finally, CM 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 segments was strictly prohibited. Therefore, the use of techniques such as normalization over multiple trial segments and the use of trial data for model adaptation was forbidden. Systems must therefore process trial lists segment-by-segment independently without access to past or future trial segments.

2.4 Metrics

While the parameter-free equal error rate (EER) metric is retained as a secondary metric, the primary metric is the *tandem detection cost function* (t-DCF) [5], and the specific **ASV-constrained** variant detailed in [6]. The detection threshold (set to the EER operating point) of the ASV system (designed by the organiser) is fixed, whereas the detection threshold of the CM system (designed by participants) is allowed to vary. Results are reported in the form of **minimum normalized** t-DCF values, defined as

$$\text{min t-DCF} = \min_{\tau_{\text{cm}}} \left\{ \frac{C_0 + C_1 P_{\text{miss}}^{\text{cm}}(\tau_{\text{cm}}) + C_2 P_{\text{fa}}^{\text{cm}}(\tau_{\text{cm}})}{\text{t-DCF}_{\text{default}}} \right\}, \quad (1)$$

where $P_{\text{miss}}(\tau_{\text{cm}})$ and $P_{\text{fa}}(\tau_{\text{cm}})$ are the miss and false alarm rates of the CM at threshold τ_{cm}. Coefficients C_0, C_1 and C_2 [6, Eq. (11)] depend not only on pre-defined target, nontarget and spoofing attack priors and detection costs but *also* on the miss, false alarm and spoof false alarm rates (the ratio of spoofed trials accepted by the ASV to the total number of spoofed trials) of the ASV system.

The denominator $\text{t-DCF}_{\text{default}} = C_0 + \min\{C_1, C_2\}$ is the cost of an uninformative *default* CM that either accepts or rejects every test utterance. Its inclusion ensures that min t-DCF values are in the range between 0 and 1. A value 0 means that both ASV and CM systems are error-free whereas a value 1 means that the CM cannot improve

2. http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

3. https://ant-novak.com/pages/sss/

4. We refer to a talker instead of speaker in order to avoid confusion with the loudspeaker.

upon the default system. Another useful reference value in between these two extremes is the case of an error-free CM (but an imperfect ASV), given by C_0/t-DCF$_{\text{default}}$. This lower bound is referred to as the **ASV floor**.

The above formulation differs slightly from that in the ASVspoof 2019 evaluation plan. Differences include the absence of sub-system-level detection costs and the inclusion of the ASV floor. The numerical scale of the t-DCF values between the formulations differs but the impact upon system rankings is negligible. The scoring toolkits[5,6] have been updated to reflect these changes.

One last, relevant detail concerning the t-DCF is how performance across different attack conditions is aggregated. The straightforward way (used for the ranking of ASVspoof 2019 challenge entries as reported in [3]) is to report performance by pooling CM scores across all attack conditions. As an alternative, we further report **max min t-DCF** across attack conditions in selected cases. Here, 'min' refers again to oracle CM calibration while 'max' refers to the highest per-condition t-DCF. The 'max min' t-DCF, therefore, serves as a reference point for worst-case attacks (see Section 5.2).

2.5 Spoofing countermeasures

Two CM systems were provided to ASVspoof 2019 participants. Baseline **B01** uses constant Q cepstral coefficients (CQCCs) [16], [17] and a bandwidth of 15 Hz to 8 kHz. The number of bins per octave is set to 96 and the re-sampling period set to 16. Static features of 29 coefficients and the zeroth coefficient are augmented with delta and delta-delta coefficients resulting in 90-dimensional features.

Baseline **B02** uses linear frequency cepstral coefficients (LFCCs) [18] and a bandwidth of 30 Hz to 8 kHz. LFCCs are extracted using a 512-point discrete Fourier transform applied to windows of 20 ms with 50% overlap. Static features of 19 coefficients and the zeroth coefficient are augmented with delta and delta-delta coefficients resulting in 60-dimensional features.

Both baselines use a Gaussian mixture model (GMM) back-end binary classifier. Randomly initialised, 512-component models are trained separately using an expectation-maximisation (EM) algorithm and bona fide and spoofed utterances from the ASVspoof 2019 training data. Scores are log-likelihood ratios given bona fide and spoofed models. A Matlab package including both baselines is available for download from the ASVspoof website.[7]

2.6 ASV system

The ASV system was used by the organisers to derive the ASV scores used in computing the t-DCF metric. It utilizes an x-vector [19] embedding extractor network that was pre-trained[8] for the *VoxCeleb recipe* of the Kaldi toolkit [20]. Training was performed using the speech data collected from 7325 speakers contained within the entire VoxCeleb2 corpus [21] and the development portion of the VoxCeleb1 corpus [22]. The network extracts 512-dimensional x-vectors which are fed to a probabilistic linear discriminant analysis (PLDA) [23], [24] back-end (trained separately for LA and PA scenarios) for ASV scoring. PLDA backends were adapted to LA and PA scenarios by using bona fide recordings of CM training data. ASV scores for the development set were provided to participants so that they could calculate t-DCF values and use these for CM optimisation. They were not provided for the evaluation set.

5. https://www.asvspoof.org/resources/tDCF_matlab_v2.zip
6. https://www.asvspoof.org/resources/tDCF_python_v2.zip
7. https://www.asvspoof.org.
8. https://kaldi-asr.org/models/m7

3 Logical access scenario

This section describes submissions to ASVspoof 2019 for the LA scenario and results. Single system submissions are described first, followed by primary system submissions, presenting only the top-5 performing of 48 LA system submissions in each case.

3.1 Single systems

The architectures of the top-5 single systems are illustrated in Fig. 2 (grey blocks). Systems are labelled (left) by the anonymised team identifier (TID) [3]. A short description of each follows:

T45 [25]: A light CNN (LCNN) which operates upon LFCC features extracted from the first 600 frames and with the same frontend configuration as the B2 baseline CM [18]. The LCNN uses an angular-margin-based softmax loss (A-softmax) [26], batch normalization [27] after max pooling and a normal Kaiming initialization [28].

T24 [29]: A ResNet classifier which operates upon linear filterbank (LFB) coefficients (no cepstral analysis). The system extracts embeddings using a modified ResNet-18 [30] architecture in which the kernel size of the input layer is 3×3 and the stride size of all layers is 1×2. Global mean and standard deviation pooling [31] are applied after the last convolutional layer and pooled features go through two fully-connected layers with batch normalization. The output is length-normalized and classified using a single-layer neural network.

T39: A CNN classifier with Mel-spectrogram features. The CNN uses multiple blocks of 1D convolution with ReLU activation, dropout, and subsampling with strided convolution. The CNN output is used as the score without binary classification.

T01: A GMM-UBM classifier with 60-dimensional STFT cepstral coefficients, including static, delta, and delta-delta components. The GMM-UBM uses the same configuration as the two baseline CMs.

T04: Cepstral features with a GMM-UBM classifier. Due to hardware limitations, models are trained using the full set of bona fide utterances but a random selection of only 9,420 spoofed utterances.

3.2 Primary systems

The architectures of the top-5 primary systems are also illustrated in Fig. 2. A short description of each follows:

T05: A fusion of seven sub-systems, six of which derive spectral representations using the DFT, while the seventh uses the discrete cosine transform (DCT). Features are extracted using frame lengths of 256, 160, or 128 samples,

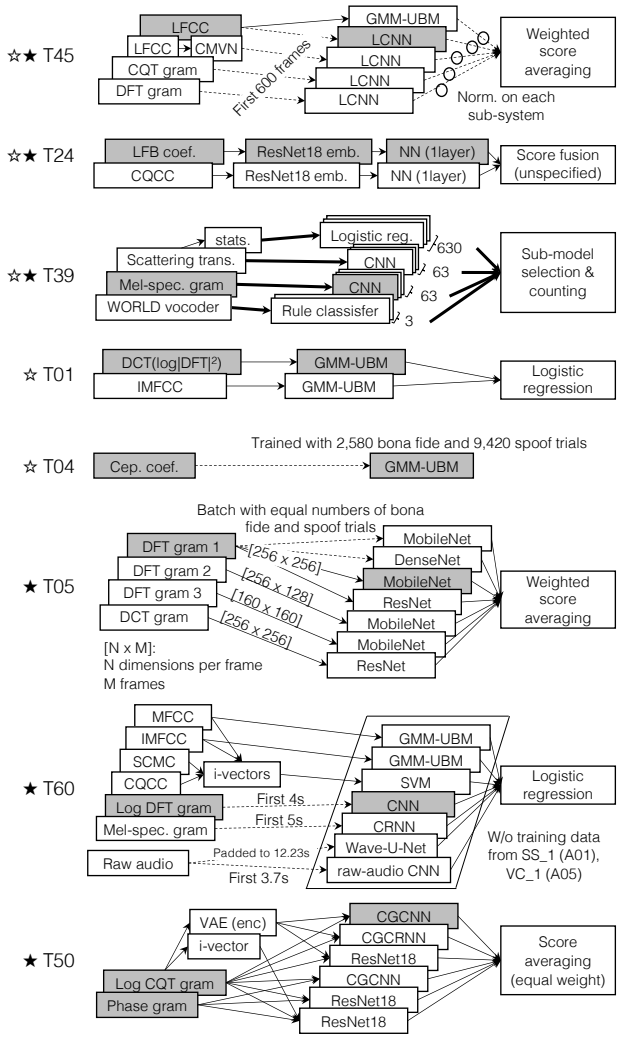


Fig. 2. Illustration of top single (grey blocks) and primary system submissions for the LA scenario. ☆ and ★ denote top-5 single and top-5 primary systems, respectively.

frame overlaps of 100, 60, or 50 samples, and 256 or 160 point DFTs. The input features are sliced in 2D matrices with 256, 160, or 128 columns (frames). All are based upon different neural network architectures: four upon MobileNetV2 [32]; two upon ResNet-50 [30]; one upon DenseNet-121 [33].

T45 [25]: A fusion of five sub-systems, including the LFCC-GMM baseline system (B1) and T45 single system. The remaining three sub-systems use LCNNs, each of which uses different features: LFCCs with CMVN; a log power spectrogram derived from the CQT; log power spectrogram derived from the DFT. All LCNN-based sub-systems use features extracted from the first 600 frames of each file. Sub-system scores are normalized according to the standard deviation of bona fide scores from the same sub-system before being fused using equal weights.

T60 [34]: A fusion of seven sub-systems. Two sub-systems are based upon 128-component GMMs trained with either MFCC or *inverted* MFCC (IMFCC) [35] features appended

TABLE 2
A comparison of top-5 (a) single and (b) primary systems for the LA scenario. Single systems B01 and B02 are the two baselines, whereas *Perfect* refers to the perfect CM (ASV floor for min t-DCF and EER of 0%). Systems are labelled by participating team identifiers (TIDs). Results are presented in terms of the minimum t-DCF (see Section 2.4) and EER metrics. Also illustrated are max min t-DCF results and corresponding attack identifier (AID) for each system (see Section 5).

(a) Single systems

TID	min t-DCF	EER [%]	Max min t-DCF (AID)
T45	0.1562	5.06	0.9905 (A17)
T24	0.1655	4.04	0.8499 (A17)
T39	0.1894	7.01	1.000 (A17)
T01	0.1937	5.97	0.7667 (A17)
T04	0.1939	5.74	0.7837 (A17)
B01	0.2839	9.57	0.9901 (A17)
B02	0.2605	8.09	0.6571 (A17)
Perfect	0.0627	0.0	0.4218 (A17)

(b) Primary systems

TID	min t-DCF	EER [%]	Max min t-DCF (AID)
T05	0.0692	0.22	0.4418 (A17)
T45	0.1104	1.86	0.7778 (A17)
T60	0.1331	2.64	0.8803 (A17)
T24	0.1518	3.45	0.8546 (A17)
T50	0.1671	3.56	0.8471 (A17)

with delta and double-delta coefficients. The third sub-system is based upon the concatenation of 100-dimension i-vectors extracted from MFCC, IMFCC, CQCC [16] features and sub-band centroid magnitude coefficient (SCMC) [36] features, and a support vector machine (SVM) classifier. The fourth sub-system is a CNN classifier operating on mean-variance normalized log DFT grams. The remaining three sub-systems are based upon either mean-variance normalized Mel-scaled spectrograms or raw audio and either convolutional recurrent neural network (CRNN), Wave-U-Net [37] or raw audio CNN classifiers. The NN-based sub-systems process a fixed number of feature frames or audio samples for each trial. Data for two attack conditions were excluded for training and used instead for validation and to stop learning. Scores are combined according to logistic regression fusion.

T24: A fusion of two sub-systems: the T24 single system; a second sub-system using the same ResNet classifier but with CQCC-based features. Scores are derived using single-layer neural networks before fusion (details unspecified).

T50 [38]: A fusion of six sub-systems all based on log-CQT gram features. Features are concatenated with a phase gram or compressed log-CQT gram obtained from a *variational autoencoder* (VAE) trained on bona fide recordings only. Three of the six classifiers use ResNet-18 [30] classifiers, for one of which a standard i-vector is concatenated with the embedding layer of the network to improve generalizability. Two other classifiers use CGCNNs [39]. The last classifier (CGCRNN) incorporates bidirectional gated recurrent units [40]. Scores are combined by equal weight averaging.

3.3 Results

A summary of results for the top-5 single and primary submissions is presented in Tables 2(a) and 2(b) respectively. Results for the two baseline systems appear in the penultimate two rows of Table 2(a) whereas the last row shows performance for a perfect CM, *i.e.* the ASV floor.

In terms of the t-DCF, all of the top-5 single systems outperform both baselines by a substantial margin, with the best T45 system outperforming the B02 baseline by 40% relative. Both systems use LFCC features, whereas T45 uses a LCNN instead of a GMM-UBM classifier. Even so, T01 and T04 systems, both also based upon standard cepstral features and GMM-UBM classifiers, are only slightly behind the better performing, though more complex systems.

Four of the top-5 primary systems perform even better, with the best T05 primary system outperforming the best T45 single system by 56% relative (73% relative to B02). The lowest min t-DCF of 0.0692 (T05 primary) is also only marginally above the ASV floor of 0.0627, showing that the best performing CM gives an expected detection cost that is close to that of a perfect CM. The top-3 primary systems all combine at least 5 sub-systems. All use diverse features, including both cepstral and spectral representations, with at least one DNN-type classifier. Of note also are differences in performance between the same teams' primary and single systems. Whereas the T05 primary system is first placed, the corresponding single system does not feature among the top-5 single systems, implying a substantial improvement through system combination. The first-placed T45 single system, however, is the second-placed primary system and, here, the improvement from combining systems is more modest. The same is observed for T24 primary and single systems.

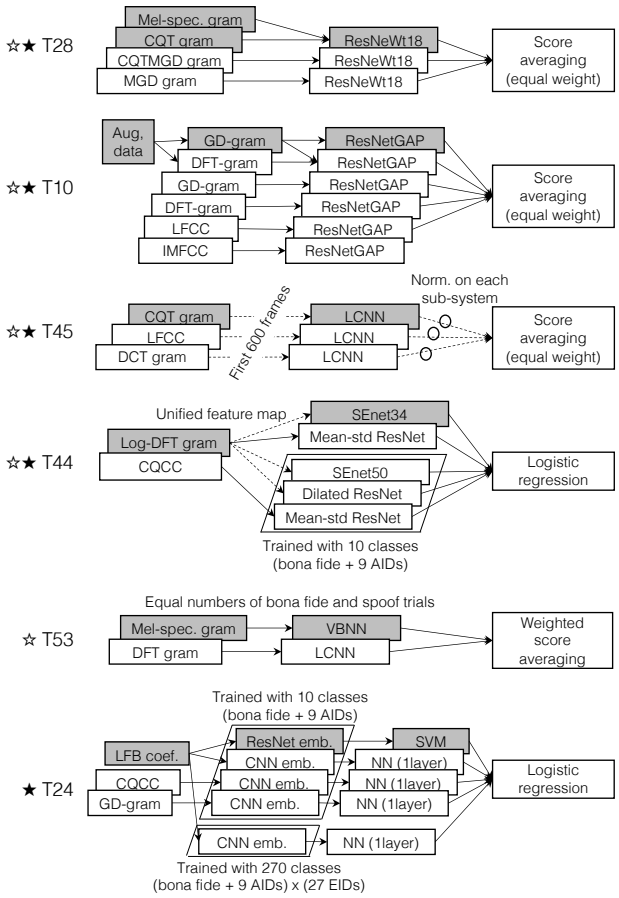


Fig. 3. Illustration of top single (grey blocks) and primary system submissions for the PA scenario. ☆ and ★ denote top-5 single and top-5 primary systems, respectively.

4 Physical access scenario

This section describes submissions to ASVspoof 2019 for the PA scenario and results. It is organised in the same way as for the logical access scenario in Section 3.

4.1 Single systems

The architectures of the top-5 single systems are illustrated in Fig. 3 (grey blocks) in which systems are again labelled with the corresponding TID. A short description of each follows:

T28 [41]: Spectral features based upon the concatenation of Mel-grams and CQT-grams and a ResNetWt18 classifier based upon a modified ResNet18 architecture [42] where the second 3x3 convolution layer is split into 32 groups. A 50% dropout layer is used after pooling and the fully connected layer has a binary output.

T10 [43]: Group delay (GD) grams [44] with cepstral mean and variance normalisation (CMVN) with data augmentation via speed pertubation. The classifier, referred to as ResNetGAP, is based upon a ResNet34 architecture [42] with a global average pooling (GAP) layer that transforms local features into 128-dimensional utterance-level representations which are then fed to a fully connected layer with softmax based cross-entropy loss.

T45 [25]: Log power CQT-grams with a LCNN classifier which uses Kaiming initialization, additional batch normalizations, and angular softmax loss. In identical fashion to the T45 LA system, the PA system operates only upon the first 600 frames from each utterance and fuses scores by averaging.

T44 [45]: Log-DFT grams with a unified feature map and squeeze and excitation network (SEnet34) [46] with a ResNet34 backbone, in which each block aggregates channel-wise statistics (*squeeze* operation) to capture channel-wise dependencies for an adaptive feature recalibration (*excitation* operation) and binary training objective.

T53: Fixed-length log Mel grams (2048 frequency bins) extracted from concatenated or truncated utterances and a variational Bayesian neural network (VBNN) using flipout [47] to decorrelate gradients within mini-batches. Bona fide data is oversampled to make the bona fide and spoofed data balanced [48].

4.2 Primary systems

The architectures of the top-5 primary systems are also illustrated in Fig. 3. A short description of each follows:

T28 [41]: A fusion of three sub-systems, all ResNet variants referred to as ResNeWt18. The first sub-system is the T28

TABLE 3
As for Table. 2 except for the PA scenario. In contrast to the LA scenario, the worst case PA scenario is denoted by both the attack identifier (AID) and the environment identifier (EID). Also illustrated here are min-tDCF results for hidden, real replay data (see Section 5.2) — min t-DCF results for real replay data are computed using C_0, C_1, and C_2 terms derived for simulated replay data.

(a) Single systems

TID	Performance			Hidden track	
	min t-DCF	EER (%)	Max min t-DCF (AID/EID)	min t-DCF	EER (%)
T28	0.1470	0.52	0.2838 (AA/acc)	0.5039	19.68
T10	0.1598	1.08	0.3768 (AA/caa)	0.8826	37.04
T45	0.1610	1.23	0.2809 (AA/acc)	0.7139	25.03
T44	0.1666	1.29	0.2781 ($AA, AC/acc$)	0.7134	41.11
T53	0.1729	1.66	0.2852 (BA/acc)	0.6379	32.64
B01	0.3476	11.04	1.0 ($BA/baa, caa, cac$)	0.3855	12.73
B02	0.3481	13.54	1.0 (BA/caa)	0.6681	29.44
Perfect	0.1354	0.0	0.2781 (AA/acc)	-	-

(b) Primary systems

TID	Performance			Hidden track	
	min t-DCF	EER (%)	Max min t-DCF (AID/EID)	min t-DCF	EER (%)
T28	0.1437	0.39	0.2781 (AA,AC/acc)	0.7160	30.74
T45	0.1460	0.54	0.2803 (AA/acc)	0.6136	20.02
T44	0.1494	0.59	0.2781 (AA,AC/acc)	0.6798	33.66
T10	0.1500	0.66	0.2781 (AA,AC/acc)	0.7987	32.04
T24	0.1540	0.77	0.2781 (AA,AC/acc)	0.9236	31.67

single system and operates upon concatenated Mel and CQT grams. The second operates upon a CQT modified group delay (CQTMGD) gram whereas the third operates directly upon the MGD gram (no CQT). Scores are combined by equal weight averaging.

T45 [25]: A fusion of three sub-systems with different frontends and a common LCNN backend. The first sub-system is the T45 single system operating on CQT grams, while the other two use either LFCC or DCT grams.

T44 [45]: A fusion of five sub-systems with either log-DFT gram or CQCC frontends and either squeeze and excitation network (SEnet) or ResNet based backends. One sub-system is the T44 single system. Two are mean and standard deviation ResNets (Mean-std ReNets) for which the input feature sequences are transformed into a single feature vector through statistics pooling. Other classifiers receive fixed-size 2D feature matrices, referred to as unified feature maps [49]. All are either binary classifiers (*i.e.* bona fide vs. spoof) or multi-class classifiers trained to predict the type of spoofing attack. Scores are combined via logistic regression fusion.

T10 [43]: A fusion of six sub-systems, all ResNet-based architectures with global average pooling (GAP) for utterance level aggregation. Two sub-systems, including the T10 single system, use data augmentation in the form of *speed perturbation* [50] applied to the raw signal. Front-ends include group-delay (GD) gram, DFT gram, LFCCs and IMFCCs. Networks are configured as binary classifiers and trained with cross-entropy loss. Scores coming from the bona fide unit for each sub-system are fused using equal weight score averaging.

T24: A fusion of five sub-systems using either LFB coefficients, CQCCs or GD-gram frontends and either CNN or ResNet backends. Embeddings produced by the ResNet system are length-normalised and classified using a weighted, two-class SVM. Three of the CNN systems and the ResNet system are configured with 10 classes (combination of 9 AIDs and the bona fide class) whereas the other CNN system has 270 output classes (full combination of all EIDs, AIDs and bona fide class). All use statistics pooling to obtain utterance-level representations from frame-level representations. Utterance-level embeddings are computed from the second-to-last fully connected layer in a similar manner to x-vector extraction. Except for the first sub-system, embeddings are processed with a single-layer neural network. Scores are combined with logistic regression.

4.3 Results

A summary of results for the top-5 single and primary submissions is presented in Tables 3(a) and 3(b) respectively, with those for the two baseline systems and the ASV floor appearing in the last three rows of Table 3(a).

Just as is the case for the LA scenario, for the PA scenario all of the top-5 single systems outperform both baselines, again by a substantial margin. In terms of the t-DCF, the best T28 system outperforms baseline B01 by 58% relative. In contrast to the LA scenario, however, all of the top-5 systems use spectral features rather than cepstral features and all also use DNN-type classifiers. Of note also is the small gap in performance between the top-5 systems, and the use of data augmentation by only one of the top-5 systems, but not the top system. The latter is, however, the only single system that uses concatenated Mel-gram and CQT-gram features.

In contrast to single systems, primary systems utilise both spectral and cepstral features, but again with exclusively DNN-type classifiers. It seems, though, that system combination is less beneficial than for the LA scenario; primary system results for the PA scenario are not substantially better than those for single systems. Perhaps unsurprisingly, then, teams with the best single systems are generally those with the best primary systems. The best T28 primary system outperforms the best single system, also from T28, by only 2% relative (59% relative to B01). Lastly, the lowest min t-DCF of 0.1437 (T28 primary) is only marginally above the ASV floor of 0.1354 showing, once again, that the best performing CM gives an expected detection cost that is close to that of a perfect CM.

5 Analysis

This section aims to provide more in-depth analysis of results presented in Sections 3 and 4. We report an analysis of generalisation performance which shows that results can be dominated by detection performance for some so-called worst-case spoofing attacks. Further analysis shows potential to improve upon fusion strategies through the use of more complementary sub-systems.

5.1 Generalisation to unseen attacks

Since its inception, ASVspoof has prioritised strategies to promote the design of generalised CMs that perform reliably

(a) LA (b) PA

Fig. 4. Illustrations of generalisation performance for the top-5 primary systems for the LA scenario (a) and the PA scenario (b), estimated using evaluation set data. For the LA scenario, box plots illustrate performance decomposed across known, varied and unknown attacks. For the scenario, they illustrate performance decomposed across low, medium and high $T60$ reverberation categories. For all plots, the green profiles signify corresponding ASV floors (performance for a perfect CM). The two right-most box plots in each case indicate performance for varied attacks without the worst case AID (LA) and high $T60$ reverberation without the worst case AID/EID (PA) and then for the worst case scenarios on their own (see Section 5.2).

in the face of spoofing attacks not seen in training data. For ASVspoof 2019, the LA evaluation set features TTS and VC spoofing attacks generated with algorithms for which some component (*e.g.* the acoustic model or the waveform generator) is different to those used in generating spoofing attacks in the training and development sets. The situation is different for the PA scenario. While the full set of attack identifier (AID) and environment identifier (EID) *categories* (see last two paragraphs of Section 2.2) are seen in all three data sets, the *specific* AIDs and EIDs in each are different (while the categories are the same, no specific attack or room configuration appears in more than one set).

A view of generalisation performance for the top-5 LA and PA primary system submissions is illustrated in Figures 4(a) and 4(b) respectively. For the LA scenario, the three left-most box plots depict performance in terms of the min t-DCF for: known attacks (attacks that are identical to those seen in training and evaluation data); varied attacks (attacks for which either the acoustic model or waveform generator is identical to those of attacks in the training and development data); wholly unknown attacks (attacks for which both components are unseen). Interestingly, while performance for unknown attacks is not dissimilar to, or even better than that for known attacks, there is substantial variability in performance for varied attacks. This observation is somewhat surprising since, while systems appear to generalise well to unknown attacks, they can fail to detect others that are generated with only variations to known attack algorithms. This can mean either that the unknown attack algorithms produce artefacts that are not dissimilar to those produced with known attacks, or that there is some peculiarity to the varied attacks. The latter implies that knowledge of even some aspects of an attack is of little use in terms of CM design; CMs are over-fitting and there is potential for them to be overcome with perhaps even only slight adjustments to an attack algorithm. Reassuringly, however, as already seen from results in Table 2 and by the green profiles to the base of each box plot in Fig. 4(a) which illustrate the ASV floor, some systems produce min t-DCFs close to that of a perfect CM.

A similar decomposition of results for the PA scenario is illustrated in Fig. 4(b), where the three left-most box plots show performance for low, medium and high $T60$ reverberation categories, the component of the EID which was observed to have the greatest influence on performance. In each case results are pooled across the other AID and EID components, namely the room size S and the talker-to-ASV distance D_{s}. Results show that, as the level of reverberation increases, the min t-DCF increases. However, comparisons of each box plot to corresponding ASV floors show that the degradation is not caused by the CM, the performance of which improves with increasing reverberation; replay attacks propagate twice in the same environment and hence reverberation serves as a cue for replay detection. The degradation is instead caused by the performance of the ASV system; the gap between the min t-DCF and the ASV floor decreases with increasing $T60$ and, for the highest level of reverberation, the min t-DCF is close to the ASV floor. This observation also shows that the effect of high reverberation dominates the influence of the room size and the talker-to-ASV distance.

From the above, it is evident that min t-DCF results are dominated by performance for some worst case attack algorithms (LA) or some worst case environmental influence (PA). Since an adversary could exploit knowledge of such worst case conditions in order to improve their chances of manipulating an ASV system, it is of interest to examine not just the pooled min t-DCF, but also performance in such worst case scenarios.

5.2 Worst case scenario

The worst case or maximum of the minimum (max min) t-DCFs (see Section 2.4) for the top-5 single and primary systems in addition to the baseline systems are shown in Tables 2 and 3. For the LA scenario, the worst case attack identifier (AID) is A17, a VC system that combines a VAE acoustic model with direct waveform modification [51]. While the best individual result for A17 is obtained by the best performing primary system the max min t-DCF is over 6 times higher than the min t-DCF. The lowest max min t-DCF for single systems is that of baseline B02. While this result (0.6571) is not substantially worse than the lowest max min t-DCF for primary systems (0.4418), it suggests that the fusion of different CMs may help to reduce the threat in the worst case scenario. The two, right-most box plots in Fig. 4(a) show performance for varied attacks without attack A17, and then for attack A17 on its own, both for the top-5 performing primary LA systems. A17 is a varied

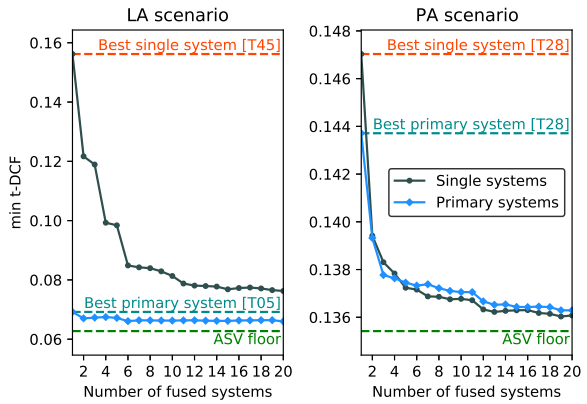


Fig. 5. min t-DCF results for oracle fusion performed with evaluation set scores for the top-20 performing systems for the LA scenario (left) and PA scenario (right). System T08, which returned problematic score distributions, was excluded in compution of results for the LA scenario.

attack and is the single attack that accounts in large part for the differences between the box plots for varied and known/unknown attacks described in Section 5.1.

Performance for the PA scenario is influenced by both the attack (AID) and the environment (EID). Excluding baselines, all but two systems struggle most for acc EIDs with small rooms (a), high $T60$ reverberation times (c) and large talker-to-ASV distances D_s (c), and either the AA or AC AID where recordings are captured in close proximity to the talker. The two, right-most box plots in Fig. 4(b) show performance for high $T60$ reverberation time without the acc EID and then for the acc EID on its own, both for the top-5 performing primary PA systems. Worst case max min t-DCFs are substantially higher than pooled min t-DCFs. Even so, it appears that the greatest influence upon tandem performance in the case of PA is within the system designer's control; the environment in which CM and ASV systems are installed should have low reverberation. Individual system results shown in Table 3 show that a single, rather than a primary system gives almost the best pooled min t-DCF (0.1470 cf. 0.1437). This observation suggests that fusion techniques were not especially successful for the PA scenario. We expand on this finding next.

5.3 Fusion performance

From the treatment of results presented in Sections 3.3 and 4.3, we have seen already that fusion seems more beneficial for the LA scenario than for the PA scenario; the best performing single and pimary LA systems give min t-DCFs of 0.1562 and 0.0692 respectively, whereas the best performing single and primary PA systems give similar min t-DCFs of 0.1470 and 0.1437 respectively.

For the LA scenario, we note that the best performing T45 *single* system still outperforms the fifth-placed T50 *primary* system. The architectures of the top-4 primary systems might then suggest that the benefit from fusion requires substantial investment in front-end feature engineering in addition to the careful selection and optimisation of the classifier ensemble. By way of example, the top-ranked T05 primary LA system uses DFT grams with different time-frequency resolutions, and three different classifiers in the shape of MobileNet, DenseNet, and a large ResNet-50. In addition, two out of the seven sub-systems take into consideration the ratio of bona fide and spoofed samples observed in training.

Even if different front-end combinations are among the top-performing primary PA systems, we do not see the same diversity in the classifier ensembles. We hence sought to investigate whether this lack of diversity could explain why fusion appears to have been less beneficial for the PA scenario. Using logistic regression [52], we conducted oracle fusion experiments for LA and PA evaluation datasets using the scores generated by the top-20 primary and single systems. In each case the number of systems in the fusion was varied between 2 and 20.

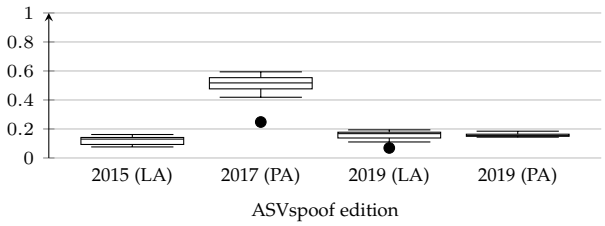
Results are illustrated in Figure. 5. Also illustrated in each plot is the min t-DCF for the best performing primary and single systems in addition to the ASV floor, *i.e.* a prefect CM that makes no errors such that the only remaining errors are made by the ASV system. There are stark differences between the two challenge scenarios. For the LA scenario, the best-performing T05 primary system (left-most, blue point) obtains nearly perfect results (performance equivalent to the ASV floor) and the fusion of multiple primary systems improves only marginally upon performance. While the fusion of multiple single systems (black profile) leads to considerably better performance, even though the fusion of 20 single systems fails to improve upon the best T05 primary system.

As we have seen already, there is little difference between primary and single systems for the PA scenario. In addition, the performance of the best individual primary and single systems is far from that of the ASV floor; there is substantial room for improvement. Furthermore, the fusion of both primary and single systems gives substantially improved performance, to within striking distance of the ASV floor. Fusion of only the best two single systems results in performance that is superior to the best T28 primary system. There was significant scope for participants to improve performance for the PA condition using the fusion of even only a small number of diverse sub-systems; it seems that those used by participants lack complimentarity.

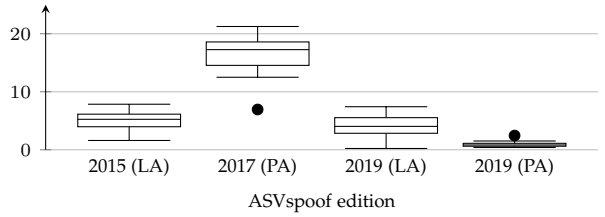
5.4 Progress

Fig. 6 shows box plots of performance for the top-10 systems for the three challenge editions: ASVspoof 2015 (LA), ASVspoof 2017 (PA), and ASVspoof 2019 (LA+PA). Comparisons should be made with caution; each database has different partitioning schemes or protocols and was created with different spoofing attacks. Furthermore, while the ASVspoof 2017 database was created from the re-recording of a source database, the ASVspoof 2019 PA database was created using simulation and, while systems developed for 2015 and 2017 editions were optimised for the EER metric, those developed for 2019 may have been optimised for the new t-DCF metric. Accordingly, Fig. 6 shows results in terms of both EER and min t-DCF.

Results for 2015 and 2019 LA databases shows that progress in anti-spoofing has kept apace with progress in TTS and VC research, including neural network-based waveform modelling techniques including WaveNet [8];



(a) min t-DCF



(b) EER (in %)

Fig. 6. An illustration of performance for top-10 submission to the three ASVspoof challenge editions: 2015, 2017 and 2019. Results are shown in terms of both min-DCF and EER.

EERs and min t-DCFs are similar, despite the use of state-of-the-art neural acoustic and waveform models to generate spoofing attacks in the ASVspoof 2019 database. Results for 2017 and 2019 PA databases seemingly show significant progress, with both EERs and min t-DCFs dropping by substantial margins, though improvements are likely caused by database differences. The 2017 database contains both additive background noise and convolutional channel noise, artefacts stemming from the source database rather than being caused by replay spoofing, whereas the 2019 database contains neither. EER results for the ASVspoof 2019 database are substantially lower than those for any of the other three databases, indicating that results reflect acoustic environment effects upon the ASV system, rather than upon CM systems. While this finding is encouraging differences between 2017 and 2019 PA results show that additive noise might have a considerable impact on performance. These issues are expanded upon next.

6 Results for real replay recordings

Results for simulated replay data were compared to results for real replay data[9] that were concealed within the PA database. This data, results for which were excluded from challenge scoring and ranking, is not described in [4]. Accordingly, a brief description is provided here. Real replay data was recorded in 3 different rooms with two different talker-to-ASV distance categories D_s and in conditions equivalent to two different EID categories: a small meeting room (equivalent EIDs of aaa and aac); a large office (equivalent EIDs of bba and bbc); a small/medium office (equivalent to EIDs of cca and ccc). Recordings were captured using high or low quality capture devices, whereas replay data were recorded using various acquisition devices before presentation to the ASV microphone using various presentation devices. Both recording and presentation devices were of quality equivalent to either B or C categories.

9. https://www.asvspoof.org/database

Data were collected from 26 speakers, each of whom provided 15 utterances selected at random from the same set of phonetically-balanced TIMIT phrases as the VCTK source data. This setup gave 540 bona fide utterances and 2160 replay utterances.

In contrast to simulated data, real replay data contains additive, ambient noise. Differences between simulation and the collection of real replay data also imply that consistent trends between results for the two datasets cannot be expected. The objective of this analysis is to expose consistencies or discrepancies in results derived between simulated and real data in terms of the t-DCF, or to determine whether the use of simulated data leads to the design of CMs that perform well when tested with real data. The two right-most columns of Table 3 show min t-DCF and EER results for the baselines and top-5 single and primary systems. In general, there are substantial differences. Among the top-5 systems considered, the best t-DCF result for real data of 0.3855 is obtained by the B01 baseline. This observation suggests that CMs are over-fitting to simulated data, or that CMs lack robustness to background noise. This possibility seems likely; we observed greater consistency in results for simulated replay data and real replay data recorded in quieter rooms. One other explanation lies in the relation between additive noise and the ASV floor. Results for synthetic data are dominated by the ASV floor, whereas those for real data are dominated by the impact of additive noise. Whatever the reason for these differences, their scale is cause for concern. Some plans to address this issue are outlined in our thoughts for future directions.

7 Future directions

Each ASVspoof challenge raises new research questions and exposes ways in which the challenge can be developed and strengthened. A selection of these is presented here.

Additive noise and channel variability

It is likely that ambient and channel noise will degrade CM performance, thus it will be imperative to study the impact of such nuisance variation in future editions, *e.g.* as in [53]. Even if such practice is generally frowned upon, the *artificial* addition of nuisance variation in a controlled fashion may be appropriate at this stage. LA scenarios generally involve some form of telephony, *e.g.* VoIP. Coding and compression effects are readily simulated to some extent. In contrast, the consideration of *additive* noise is potentially more complex for it influences speech *production*, *e.g.* the Lombard reflex [54]. The simulation of additive noise is then generally undesirable. An appropriate strategy to address these issues in future editions of ASVspoof demands careful reflection.

Quality of TTS/VC training data

For ASVspoof 2019, all TTS and VC systems were trained with data recorded in benign acoustic conditions. This setup is obviously not representative of *in-the-wild* scenarios where an adversary could acquire only relatively noisy training or adaptation data. Future editions of ASVspoof should hence consider TTS and VC attacks generated with more realistic data. Such attacks may be less effective in fooling ASV system, and may also be more easily detectable.

Diversified spoofing attacks

ASVspoof presents an arguably naive view of potential spoofing attacks. Future editions should consider more diversified attacks, *e.g.* impersonation [55], attacks by twins or siblings [56], non-speech [57] or adversarial attacks [58], [59], [60] and attacks that are injected into specific regions of the speech signal rather than the entire utterance. One can also imagine blended attacks whereby, for instance, replay attacks are launched in an LA scenario, or replay attacks in a PA scenario are performed with speech data generated using TTS or VC systems.

Joint CM+ASV score calibration

With the 2019 edition transitioned to an ASV-centric form of assessment with the min t-DCF metric there are now not two, but three decision outcomes: target, non-target (both bona fide) and spoof. The existing approaches to calibrate binary classification scores are then no longer suitable. Future work could hence investigate approaches to joint CM+ASV system optimisation and calibration.

Reproducibility

Anecdotal evidence shows that some ASVspoof results are un-reproducible. While it is not our intention to enforce reproducibility – doing so may deter participation – it is nonetheless something that we wish to promote. One strategy is to adopt the reviewing of system descriptions, either by the organisers or by fellow ASVspoof participants, or the reporting of system descriptions according to a harmonised format. Such a harmonised reporting format should include details of the fusion scheme and weights. This policy, together with a requirement for the submission of scores for each system in an ensemble, would also allow a more fine-grained study of fusion strategies and system complementarity.

Explainability

Explainability is a topic of growing importance in almost any machine learning task and is certainly lacking sufficient attention in the field of anti-spoofing. While the results reported in this paper show promising potential to detect spoofing attacks, we have learned surprisingly little about the artefacts or the cues that distinguish bona fide from spoofed speech. Future work which reveals these cues may be of use to the community in helping to design better CMs.

Standards

The t-DCF metric adopted for ASVspoof 2019 does not meet security assessment standards, in particular the so-called *Common Criteria for Information Technology Security Evaluation* [61], [62], [63]. Rather than quantifying a probabilistic meaning of some attack likeliness, common criteria are based upon a category-based points scheme in order to determine a so-called *attack potential*. This reflects *e.g.* the equipment, expertise and time required to mount the attack and the knowledge required of the system under attack. The rigour in common criteria is that each category of attack potential then demands hierarchically higher *assurance components* in order to meet *protection profiles* that express *security assurance requirements*. In the future, it may prove beneficial to explore the gap between the common criteria and the t-DCF. To bridge this gap pragmatically, we need to determine the attack potentials for ASVspoof, asking ourselves: 1) *How long does it take to mount a given attack?*; 2) *What level of expertise is necessary?*; 3) *What resources (data or computation) are necessary to execute it?*; 4) *What familiarity with the ASV system is needed?* Clearly, providing the answers to these questions is far from being straightforward.

Challenge model

The organisation of ASVspoof has developed into a demanding, major organisational challenge involving the coordination of 6 different organising institutes and 19 different data contributors for the most recent edition. While the organisers enjoy the support of various different national research funding agencies, it is likely that we will need to attract additional industrial, institutional or public funding to support the initiative in the future. To this end, we are liaising with the Security and Privacy in Speech Communications (SPSC), the Speaker and Language Characterisation (SpLC) and Speech Synthesis (SynSig) Special Interest Groups (SIGs) of the International Speech Communication Association (ISCA) regarding a framework with which to support the initiative in the longer term.

With regards the format, the organising team committed to making ASVspoof 2019 the last edition to be run as a special session at INTERSPEECH. With anti-spoofing now featuring among the Editor's Information Classification Scheme (EDICS) and topics of major conferences and leading journals/transactions, it is time for ASVspoof to make way for more genuinely 'special' topics. Accordingly, we will likely transition in the future to a satellite workshop format associated with an existing major event, such as INTERSPEECH.

8 Conclusions

ASVspoof 2019 is the third in the series of anti-spoofing challenges for automatic speaker verification. It was the first to consider both logical and physical access scenarios in a single evaluation and the first to adopt the tandem detection cost function as the default metric and also brought a series of additional advances with respect to the 2015 and 2017 predecessors. With the database and experimental protocols described elsewhere, the current paper describes the challenge results, findings and trends, with a focus on the top-performing systems for each scenario.

Results reported in the paper are encouraging and point to advances in countermeasure performance. For the logical access scenario, reliability seems to have kept apace with the recent, impressive developments in speech synthesis and voice conversion technology, including the latest neural network-based waveform modelling techniques. For the physical access scenario, countermeasure performance is stable across diverse acoustic environments. Like most related fields in recent years, the 2019 edition was marked with a shift towards deep architectures and ensemble systems that brought substantial improvements in

performance, though more so for the logical access scenario than the physical access counterpart. There seems to have been greater diversity among each teams' ensemble systems for the former, while there is evidence that those for the latter suffer from over-fitting. In both cases, however, *tandem* systems exhibit high detection costs under specific conditions: either specific attack algorithms or specific acoustic environments with costs stemming from either countermeasures or automatic speaker verification systems.

Many challenges remain. Particularly for the physical access scenario, though likely also for the logical access scenario, countermeasure performance may degrade in real-world conditions characterised by nuisance variation such as additive noise. Results for real replay data with additive noise show substantial gaps between results for simulated, noise-free data. It is furthermore reasonable to assume that performance will also be degraded by channel variability as well as any mismatch in the training data used to generate spoofing attacks.

Future editions of ASVspoof will also consider greater diversification in spoofing attacks and blended attacks whereby speech synthesis, voice conversion and replay attack strategies are combined. Countermeasure optimisation strategies also demand further attention now that they are assessed in tandem with automatic speaker verification systems. Future editions demand greater efforts to promote reproducibility and explainability, as well as some reflection on the gap between ASVspoof and standards such as the common criteria. Lastly, the paper outlines our plans to adopt a satellite event format, rather than a special session format for the next edition, tentatively, ASVspoof 2021.

ACKNOWLEDGEMENTS

The ASVspoof 2019 organisers thank the following for their invaluable contribution to the LA data collection effort – Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang and Zhen-Hua Ling. We also extend our thanks to the many researchers and teams who submitted scores to the ASVspoof 2019 challenge. Since participants are assured of anonymity, we regret that we cannot acknowledge them here by name.

REFERENCES

[1] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, and K.-A. Lee, *Introduction to Voice Presentation Attack Detection and Recent Advances*. Springer International Publishing, 2019.

[2] N. Evans, J. Yamagishi, and T. Kinnunen, "Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics," *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*, 2013.

[3] 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.

[4] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "ASVspoof 2019: a large-scale public database of synthetic, converted and replayed speech," *Elsevier Computer Speech and Language*, vol. 64, November 2020.

[5] 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*, 2018, pp. 312–319.

[6] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.

[7] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017, http://dx.doi.org/10.7488/ds/1994.

[8] 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.

[9] 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, pp. 195–202.

[10] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 30107-1. Information Technology - Biometric presentation attack detection - Part 1: Framework*, International Organization for Standardization, 2016.

[11] 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*, vol. 9, no. 3, 2005.

[12] E. Vincent, "Roomsimove," 2008, [Online] http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip.

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

[14] A. Janicki, F. Alegre, and N. Evans, "An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks," *Security and Communication Networks*, vol. 9, no. 15, pp. 3030–3044, 2016.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[16] 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.

[17] ——, "Articulation rate filtering of CQCC features for automatic speaker verification," in *Proc. Interspeech*, 2016, pp. 3628–3632.

[18] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, 2015, pp. 2087–2091.

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

[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," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[23] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proc. European Conference on Computer Vision (ECCV)*, A. Leonardis, H. Bischof, and A. Pinz, Eds., 2006, pp. 531–542.

[24] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Intl. Conf. on Computer Vision (ICCV)*, 2007, pp. 1–8.

[25] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. Interspeech*, 2019, pp. 1033–1037.

[26] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 212–220.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Intl. Conf. on Machine Learning (ICML)*, 2015, pp. 448–456.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Intl. Conf. on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[29] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of Audio Deepfake Detection," in *Proc. Odyssey*, 2020, pp. 132–137. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2020-19

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[31] M. Lin, Q. Chen, and S. Yan, "Network in network," *Proc. Intl. Conf. on Learning Representations (ICLR)*, 2014.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[34] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," in *Proc. Interspeech*, 2019, pp. 1018–1022.

[35] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *International Journal of Signal Processing Systems (IJSPS)*, vol. 4, no. 2, pp. 114–122, 2007.

[36] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," in *Proc. Odyssey*, 2010, pp. 34–39.

[37] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. Intl. Society for Music Information Retrieval Conference (ISMIR)*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 334–340.

[38] Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, and K. Yu, "The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge," in *Proc. Interspeech*, 2019, pp. 1038–1042.

[39] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Intl. Conf. on Machine Learning (ICML)*, 2017, pp. 933–941.

[40] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[41] X. Cheng, M. Xu, and T. F. Zheng, "Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 540–545.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[43] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU replay detection system for the ASVspoof 2019 challenge: On data augmentation, feature representation, classification, and fusion," in *Proc. Interspeech*, 2019, pp. 1023–1027.

[44] F. Tom, M. Jain, and P. Dey, "End-toend audio replay attack detection using deep convolutional networks with attention," in *Proc. Interspeech*, 2018, pp. 681–685.

[45] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-Spoofing with Squeeze-Excitation and Residual Networks," in *Proc. Interspeech*, 2019, pp. 1013–1017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1794

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[47] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse, "Flipout: Efficient pseudo-independent weight perturbations on mini-batches," in *Proc. Intl. Conf. on Learning Representations (ICLR)*, 2018.

[48] R. Białobrzeski, M. Kośmider, M. Matuszewski, M. Plata, and A. Rakowski, "Robust Bayesian and Light Neural Networks for Voice Spoofing Detection," in *Proc. Interspeech 2019*, 2019, pp. 1028–1032.

[49] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 6316–6320.

[50] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 171–175.

[51] K. Kobayashi, T. Toda, and S. Nakamura, "Intra-gender statistical singing voice conversion with direct waveform modification using log-spectral differential," *Speech Communication*, vol. 99, pp. 211–220, 2018.

[52] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," [Online] https://sites.google.com/site/bosaristoolkit, AGNITIO Research, South Africa, Tech. Rep., 12 2011.

[53] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, "ReMASC: Realistic Replay Attack Corpus for Voice Controlled Systems," in *Proc. Interspeech 2019*, 2019, pp. 2355–2359. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1541

[54] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Malad. l'Oreille Larynx*, no. 37, pp. 101–119, 1911.

[55] V. Vestman, T. Kinnunen, R. González Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36–54, 2020.

[56] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in anti-spoofing: from the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.

[57] F. Alegre, R. Vipperla, N. Evans, and B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2012, pp. 36–40.

[58] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1962–1966.

[59] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6579–6583.

[60] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *Proc. Interspeech (to appear)*, 2020.

[61] Common Criteria Recognition Arrangement (CCRA), *Common Criteria for Information Technology Security Evaluation — Part 3: Security assurance components*, Common Criteria for Information Technology Security Evaluation (CC) and Common Methodology for Information Technology Security Evaluation (CEM), 2017.

[62] ISO/IEC JTC1 SC27 Security Techniques, *ISO/IEC DIS 19989-1:2019. Information security — Criteria and methodology for security evaluation of biometric systems — Part 1: Framework*, International Organization for Standardization, 2019.

[63] ——, *ISO/IEC DIS 19989-3:2019. Information security — Criteria and methodology for security evaluation of biometric systems — Part 3: Presentation attack detection*, International Organization for Standardization, 2019.

Andreas Nautsch is with the Audio Security and Privacy research group (EURECOM). He received the doctorate from Technische Universität Darmstadt in 2019. From 2014 to 2018, he was with the da/sec Biometrics and Internet-Security research group (Hochschule Darmstadt) 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) respectively in 2012 and 2014. He served as an expert delegate to ISO/IEC and as project editor of the ISO/IEC 19794-13:2018 standard. Andreas serves currently as associate editor of the EURASIP Journal on Audio, Speech, and Music Processing, and is a co-initiator and secretary of the ISCA Special Interest Group on Security & Privacy in Speech Communication.

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.

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 EU 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.

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.

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 is 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.

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.

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.

Kong Aik Lee (M'05-SM'16) is currently a Senior Scientist at Institute for Infocomm Research, A*STAR, Singapore. From 2018 to 2020, he was 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, and the Outstanding Service Award by IEEE ICME 2020. He was the Lead Guest Editor for the CSL SI on "Two decades into Speaker Recognition Evaluation - are we there yet?" Currently, he serves as an Editorial Board Member for Elsevier Computer Speech and Language (2016 - present), and 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, and the General Chair of the Speaker Odyssey 2020 Workshop.