ASVspoof: the Automatic Speaker Verification Spoofing and Countermeasures Challenge

Zhizheng Wu, *Member, IEEE,* Junichi Yamagishi *Senior Member, IEEE,*
Tomi Kinnunen, *Member, IEEE,* Cemal Hanilçi, Md Sahidullah, *Member, IEEE,* Aleksandr Sizov,
Nicholas Evans, *Member, IEEE,* Massimiliano Todisco, Héctor Delgado

***Abstract*—Concerns regarding the vulnerability of automatic speaker verification (ASV) technology against spoofing can undermine confidence in its reliability and form a barrier to exploitation. The absence of competitive evaluations and the lack of common datasets has hampered progress in developing effective spoofing countermeasures. This paper describes the ASV Spoofing and Countermeasures (ASVspoof) initiative, which aims to fill this void. Through the provision of a common dataset, protocols, and metrics, ASVspoof promotes a sound research methodology and fosters technological progress. This paper also describes the ASVspoof 2015 dataset, evaluation, and results with detailed analyses. A review of post-evaluation studies conducted using the same dataset illustrates the rapid progress stemming from ASVspoof and outlines the need for further investigation. Priority future research directions are presented in the scope of the next ASVspoof evaluation planned for 2017.**

***Keywords*—Biometric, Automatic Speaker Verification, Spoofing, Presentation Attacks, ASVspoof**

I. Introduction

Automatic speaker verification (ASV) [1] offers a low-cost and flexible biometric solution to person authentication. The reliability of ASV technology has advanced considerably recently and is currently deployed in a growing variety of practical applications such as in call centres, for spoken dialogue systems, and in many mass-market, consumer products.

Unfortunately, and as is the case for any biometric technology, concerns regarding vulnerabilities to *spoofing* [2], also referred to as *presentation attacks* [3], can undermine user confidence; thus, form a barrier to exploitation. By masquerading as another enrolled client, i.e. by mimicking their biometric traits, fraudsters can use spoofing attacks to infiltrate systems or services protected using biometric technology. Acknowledged spoofing attacks with regards to ASV include impersonation, replay, speech synthesis, and voice conversion [4].

In response to the threat of spoofing, researchers have sought to develop effective approaches to anti-spoofing. There are two general directions, i.e. that involving ever-more robust and advanced ASV techniques and that involving dedicated spoofing countermeasures. Advanced ASV techniques are expected to provide greater inherent resilience to spoofed speech, whereas dedicated countermeasures offer the potential for explicit spoofing detection. Perhaps for this reason, the latter has attracted the greatest interest. The literature shows growing momentum behind the development of spoofing countermeasures, a comprehensive survey of which is presented in [5].

All early investigations on developing spoofing countermeasures were conducted with purpose-collected datasets typically generated using a limited number of specially crafted spoofing-attack algorithms. Furthermore, the datasets and countermeasures were usually developed by the same researchers. Such practice was necessary to support early research, as there were no common benchmark datasets. This methodology nonetheless raises three concerns. First, only the use of common datasets can support reproducible research and meaningful comparison of results generated by different research teams. Second, *a priori* knowledge of a spoofing attack does not reflect the practical scenario in which the nature of the spoofing attack can never be known beforehand. Third, the development of countermeasures using only a small number of spoofing attacks or spoofing algorithms may not offer the greatest potential for generalisation of different or unforeseen attacks that will almost certainly be encountered in the wild.

Common datasets and competitive evaluations to support the development of spoofing countermeasures for other biometric modalities have existed for some time. They include the series of LivDet evaluations for fingerprint recognition [8] and a similar initiative for face recognition [9]. As of 2014,

Zhizheng Wu was with the Centre of Speech Technology Research, University of Edinburgh, U.K. He is now with Apple Inc. e-mail: zhizheng.wu@ed.ac.uk.

Junichi Yamagishi is with the Centre of Speech Technology Research, University of Edinburgh, U.K. and with the National Institute of Informatics, Japan. e-mail: jyamagis@nii.ac.jp.

Tomi Kinnunen, Md Sahidullah and Aleksandr Sizov are with the University of Eastern Finland, Finland. e-mail: tkinnu@cs.uef.fi, sahidullahmd@gmail.com, aleksandr.sizov.work@gmail.com

Cemal Hanilçi was with the University of Eastern Finland, Finland. He is now with the Bursa Technical University, Turkey. e-mail: cemal.hanilci@btu.edu.tr

Nicholas Evans, Massimiliano Todisco and Héctor Delgado are with EURECOM, France. e-mail: evans@eurecom.fr, todisco@eurecom.fr, delgado@eurecom.fr

The work presented in this paper was partially supported by EPSRC through Programme Grants EP/I031022/1 (NST) and EP/J002526/1 (CAF), by the Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency (JST) (uDialogue project), by MEXT KAKENHI Grant Numbers (26280066, 26540092, 15H01686, 15K12071, 16H06302), by the Academy of Finland (project no. 253120 and 283256) and by the Scientific and Technological Research Council of Turkey (TUBITAK), project no. 115E916.

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

Manuscript received August 23, 2016; revised January 11, 2017.

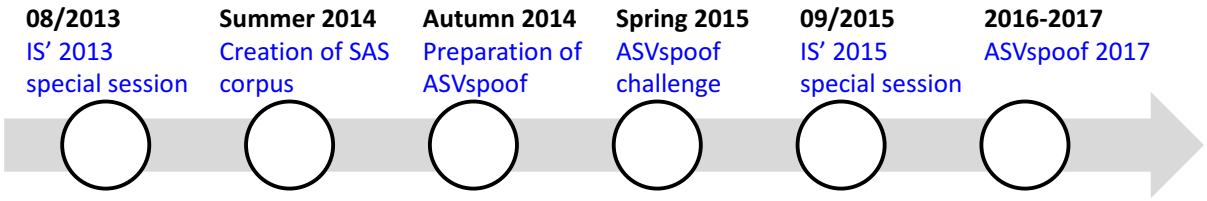


Fig. 1. The ongoing, community driven Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) initiative aims to advance the state-of-the-art in spoofing countermeasures for voice biometrics. Kicked off in 2013 as a special session of Interspeech 2013 and followed up by the creation of the Spoofing and Anti-Spoofing (SAS) corpus of speech synthesis and voice-conversion spoofing data [6], [7], the first ASVspoof challenge was co-organized in 2015. The initiative lead to the submission of 16 countermeasure systems, all benchmarked and ranked using a common database, protocol, and metric. Results were disseminated at a follow-up special session of Interspeech 2015 that attracted wide participation from ASV, speech-synthesis, and voice-conversion communities. Plans for the second edition in 2017 are based on community feedback and lessons learnt from the first edition.

none encouraged the development of spoofing countermeasures for ASV. The need for a common dataset, protocols, and metrics [10] was the principal finding of a special session on spoofing and countermeasures for ASV [4] held at the INTERSPEECH 2013 conference in Lyon, France. The ASV Spoofing and Countermeasures (ASVspoof) initiative was created shortly afterwards. Its goal was to foster ASV spoofing countermeasures by (i) gathering the necessary expertise to produce and to make publicly available a large dataset of genuine and spoofed speech and (ii) organising competitive evaluations to benchmark different countermeasure solutions. The key milestones of the ASVspoof initiative are illustrated in the timeline of Fig. 1.

Stemming from the special session at INTERSPEECH, the creation of a large dataset of genuine and spoofed speech [6], [7] started later in 2013. Following refinements and improvements to the dataset and protocols, the first ASVspoof challenge [11][1] was organised as a special session at INTERSPEECH 2015 held in Dresden, Germany. To lower the cost of entry, thus, maximising participation, ASVspoof 2015 involved a spoofing-detection challenge in isolation from ASV. The challenge was based upon a common database of both genuine speech and spoofed speech generated using ten mainstream speech-synthesis and voice-conversion spoofing-attack algorithms. Participants were tasked with designing spoofing-detection algorithms to distinguish between genuine and spoofed speech in accordance with a common protocol and to submit their detection scores to the organisers for evaluation and ranking. This first edition was successful in attracting wide participation, advancing the state-of-the-art in spoofing countermeasures, and identifying directions for future research.

This paper presents the first ASVspoof evaluation with details of the dataset, protocols, and metrics. Also included are brief descriptions of participants' systems, their respective results, including those obtained from system fusion, with detailed analysis. Results reported in the literature post-evaluation are also described and show the rapid progress in spoofing detection brought about by the ASVspoof initiative. Directions for future research are then discussed before current ideas for subsequent editions of ASVspoof, the next of which is planned for 2017.

[1]http://www.spoofingchallenge.org

II. Database

The ASVspoof 2015 database[2] contains both genuine and spoofed speech data collected from 106 speakers (45 male and 61 female). All genuine speech recordings were collected in the same semi-anechoic chamber having a solid floor. Spoofed speech samples of each speaker were generated artificially using one of ten different, well-known speech-synthesis or voice-conversion spoofing-attack algorithms.

A. Spoofing-attack algorithms

Work prior to the release of the ASVspoof 2015 database generally produced countermeasures using spoofed speech data generated with a single or small number of methods implemented by the same researchers. The focus on specific spoofing-attack algorithms and on the use of *a priori* knowledge is clearly unrepresentative of the real use-case scenario in which it is impossible to know the nature of a spoofing attack.

To reflect more closely a real scenario in which a wide variety of unknown spoofing attacks can be expected, the ASVspoof 2015 database was created using ten state-of-the-art spoofing-attack algorithms. They were contributed by members of the speech-synthesis and voice-conversion communities; therefore, introducing a level of independence from the effort within the ASV community to develop countermeasures. This approach is expected to favour the development of generalised countermeasures capable of detecting previously unseen spoofing attacks.

Each of the ten spoofing-attack algorithms (**S1** to **S10**) are described below.

S1 A simplified frame-selection-based [12], [13] voice-conversion algorithm, in which converted speech is generated from the selection of target speech frames. For computational efficiency, target frames are selected without taking the inter-frame joint cost into account. The latter is determined using the Euclidean distance. Further details can be found in [13].

[2]The ASVspoof database is freely available under a creative commons (CC-BY) license and can be downloaded from http://dx.doi.org/10.7488/ds/298. The ASVspoof 2015 database is a subset of the Spoofing and Anti-Spoofing (SAS) corpus [6], [7] which is also freely available online at http://dx.doi.org/10.7488/ds/252 and also under a creative commons (CC-BY) license.

TABLE I. SUMMARY OF SPOOFING-ATTACK ALGORITHMS IMPLEMENTED IN CHALLENGE DATABASE. S1 TO S5 ARE CONSIDERED KNOWN ATTACKS, EXAMPLES OF WHICH ARE AVAILABLE FOR SYSTEM DEVELOPMENT. S6 TO S10 ARE CONSIDERED UNKNOWN ATTACKS SEEN ONLY IN THE EVALUATION SET. HERE "TRAIN" MEANS THE TRAINING SET, "DEV" MEANS THE DEVELOPMENT SET, AND "EVA" MEANS THE EVALUATION SET.

Subset	Number of trials or utterances			Waveform generation	Spoofing method	Feature representation	Using open source toolkit?
	Train	Dev	Eva				
Genuine	3750	3497	9404	None	None	N.A.	N.A.
S1	2525	9975	18400	STRAIGHT vocoder	Frame-selection voice conversion	Mel-cepstrum, Band aperiodicity, F_0	No
S2	2525	9975	18400	STRAIGHT vocoder	Slope shifting voice conversion	Mel-cepstrum, Band aperiodicity, F_0	No
S3	2525	9975	18400	STRAIGHT vocoder	HMM-based speech synthesis	Mel-cepstrum, Band aperiodicity, F_0	Yes
S4	2525	9975	18400	STRAIGHT vocoder	HMM-based speech synthesis	Mel-cepstrum, Band aperiodicity, F_0	Yes
S5	2525	9975	18400	MLSA vocoder	GMM-based voice conversion	Mel-cepstrum, F_0	Yes
S6	0	0	18400	STRAIGHT vocoder	GMM-based voice conversion	Mel-cepstrum, Band aperiodicity, F_0	No
S7	0	0	18400	STRAIGHT vocoder	GMM-based voice conversion	Line spectrum pair, F_0	No
S8	0	0	18400	STRAIGHT vocoder	Tensor-based voice conversion	Mel-cepstrum, Band aperiodicity, F_0	No
S9	0	0	18400	STRAIGHT vocoder	KPLS-based voice conversion	Mel-cepstrum, Band aperiodicity, F_0	No
S10	0	0	18400	Diphone concatenation	Unit selection-based speech synthesis	Waveform	Yes

S2 One of the simplest voice-conversion algorithms that only adjusts the first Mel-cepstral coefficient (MCC) [14] to shift the slope of the source spectrum towards that of the target. This algorithm produces converted speech with high perceptual quality, but low speaker similarity, a common trade-off in voice conversion.

S3 A speech-synthesis algorithm implemented with the hidden Markov model (HMM)-based speech-synthesis system (HTS) and state-of-the-art speaker-adaptation techniques [15]. Adaptation is carried out using only 20 target speaker utterances.

S4 The same algorithm as S3, but using 40 adaptation utterances. This algorithm is expected to produce synthesised speech with higher perceptual quality and higher speaker similarity than S3.

S5 A voice-conversion algorithm implemented with the publicly available voice-conversion framework within the Festvox toolkit[3]. Default settings are used.

S6 A voice-conversion algorithm based on joint density Gaussian mixture models (GMMs) and maximum likelihood parameter generation considering global variance [16].

S7 A voice-conversion algorithm similar to S6, but using line spectrum pairs rather than MCCs for spectral representation.

S8 A tensor-based approach to voice conversion [17] for which a Japanese database [18] was used to construct the speaker space.

S9 A voice-conversion algorithm that uses a kernel-based partial least square (KPLS) approach to implement a non-linear transformation function [19]. For simplicity, conversion is carried out without the use of dynamic information.

S10 A speech-synthesis algorithm implemented with the open-source MARY text-to-speech system (MaryTTS)[4]. Spoofed speech is generated from the concatenation of diphone waveforms.

To provide spoofed utterances of varying quality, 20 utterances per speaker were used to train S1, S2, S3, S5, S6, S7, S8, and S9, whereas the same 20 utterances were supplemented with an additional 20 utterances for training S4 and S10. The S1, S2, S3, S4, S6, S7, S8, and S9 spoofing-attack algorithms use the same state-of-the-art STRAIGHT vocoder [20] for synthesis, whereas S5 uses the Mel log spectrum approximation (MLSA) vocoder [14] implemented in the Speech Signal Processing Toolkit[5]. Both vocoders are commonly used in speech synthesis and voice conversion. In contrast to the MLSA vocoder, the STRAIGHT vocoder includes spectral smoothing using F_0 adaptive windows and mixed excitation, which additionally uses aperiodicity features. The S10 spoofing-attack algorithm uses a diphone concatenation method to generate speech waveforms. These spoofing-attack algorithms are summarized in Table I. Of note, four of the ten spoofing-attacks are implemented with open-source software, i.e. software available to the public, including fraudsters. More details including protocols used to generate spoofed speech can be found in [6], [7].

B. Data visualization

Figure 2 provides an intuitive visualisation of the acoustic similarities between each spoofing-attack algorithm and genuine speech. This visualisation is based upon i-vector data, the extraction process described in [21]. Briefly, each utterance in the ASVspoof 2015 database was first converted into a 600-dimensional i-vector [22] then projected into a two-dimensional space using an optimized variant [23] of the *t-distributed stochastic neighbour embedding* (t-SNE) method [24]. The i-vector extractor was trained on Mel frequency cepstral coefficient (MFCC) features. Raw i-vectors were then processed using standard post-processing operations: raw i-vectors are first whitened, projected onto a unit sphere [25], and finally treated with *within-class covariance normalization* (WCCN) [26]. The WCCN matrix was trained by treating data corresponding to the 10 spoofing-attack algorithms and

[3]http://www.festvox.org/

[4]http://mary.dfki.de/

[5]http://sp-tk.sourceforge.net/

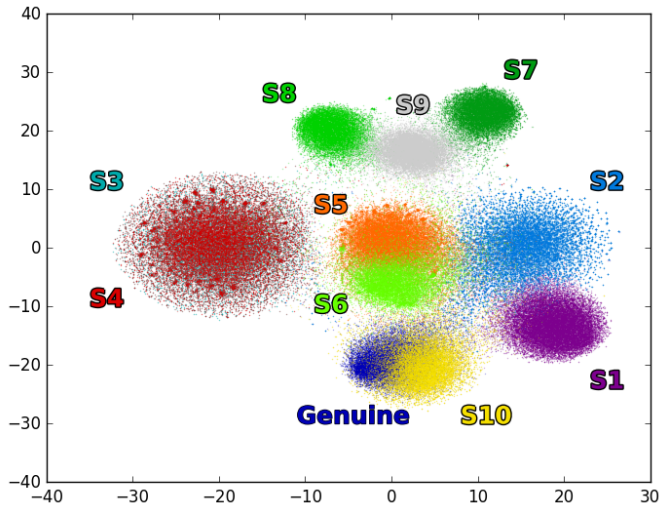


Fig. 2. Visualisation of differences between genuine and spoofed speech in the ASVspoof 2015 database. Two-dimensional data were produced using a t-distributed stochastic neighbour embedding (t-SNE) algorithm applied to high-dimensional i-vector representations of each utterance.

genuine speech as 11 distinct classes. The entire ASVspoof database (a total of 263,151 i-vectors from 11 classes) was used to train the whitening and WCCN matrices. The t-SNE algorithm was applied with a perplexity parameter set to 40.

The visualisation in Figure 2 shows mostly well separated spoofing-attack algorithms. Only S3 and S4 overlap, but this is to be expected; they correspond to the same algorithm trained with different quantities of data. The S10 spoofing-attack algorithm appears to overlap somewhat with genuine speech. With differing similarities to genuine speech and with varying characteristics, the automatic detection of the ten spoofing-attack algorithms is therefore expected to present a significant challenge. Reliable detection then calls for strong but generalised spoofing countermeasures.

C. Impact upon speaker verification

While Figure 2 illustrates the relative differences between genuine and spoofed speech, it does not reflect the impact of each spoofing attack on ASV performance. Accordingly, ASV experiments were conducted to gauge ASV vulnerabilities.

These experiments were conducted with a state-of-the-art, i-vector ASV system [27] with probabilistic linear discriminant analysis (PLDA) [28] based on MFCC features. The universal background model (UBM) [29] was trained with data from the Wall Street Journal (WSJ0, WSJ1 and WSJCAM) [30] and Resource Management (RM) [31] corpora. Models have 512 Gaussian components, whereas the total variability space is of dimension 400. All i-vectors were centred, length-normalized, and whitened. The whitening transformation was estimated from i-vectors in the development set. The Gaussian PLDA model with an eigenspace of dimension 100 was trained using an expectation maximisation (EM) algorithm. The system was implemented using the Microsoft Research (MSR) Identity Toolbox [32].

TABLE II. NUMBER OF NON-OVERLAPPING TARGET SPEAKERS AND NUMBER OF UTTERANCES IN TRAINING, DEVELOPMENT, AND EVALUATION SETS OF THE ASVSPOOF 2015 DATABASE. THE DURATION OF EACH UTTERANCE IS IN THE ORDER OF ONE TO TWO SECONDS.

Subset	Number of speakers		Number of utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

ASV performance is illustrated through detection-error trade-off (DET) profiles in Figures 3 (a) and (b) for female and male speakers, respectively. Separate profiles are illustrated for genuine speech and each of the ten spoofing-attack algorithms (S1-S10). Both plots show that all degrade ASV performance, with the baseline equal error rates (EERs) of 2% increasing to between 3 and 44%. More specifically, we see that S2, which only shifts the slope of the spectrum, increases only slightly the error rates for both male and female speakers. In contrast, other spoofing attacks generated from the state-of-the-art voice-conversion and speech-synthesis systems produce significant increases in error rates.

D. Protocols

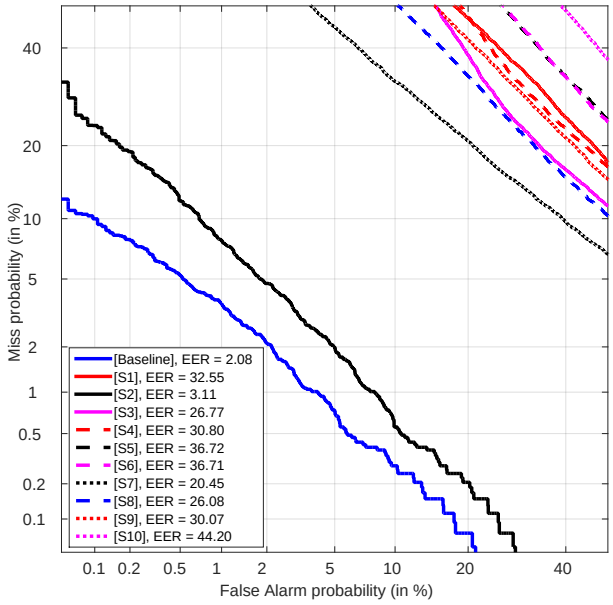
The ASVspoof 2015 database was provided with protocols for three distinct subsets, namely training, development, and evaluation. In addition to genuine speech data, the training and development sets are composed of spoofed data generated from a subset of five spoofing-attack algorithms (S1–S5) referred to as *known* attacks. The evaluation set contains data generated with all ten spoofing-attack algorithms, where the additional algorithms (S6–S10) are referred to as *unknown attacks*. This strategy, whereby there is no training or development data for unknown attacks, was chosen to encourage the development of generalised countermeasures [33]. The number of speakers and trials in each subset is summarized in Table II and described below.

Training set - The training set includes 3750 genuine and 12625 spoofed utterances collected from 25 speakers (10 male, 25 female). Spoofed data are generated using one of the five known attacks (S1–S5). All meta information, including speaker identities and exact spoofing-attack algorithms, is provided in the ground truth. Meta information may be used for system optimisation.

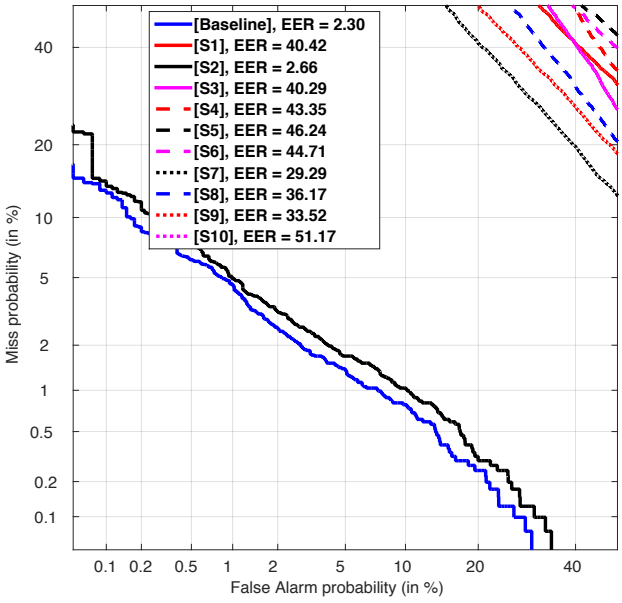
Development set - The development set includes both genuine and spoofed speech from a subset of 35 speakers (15 male, 20 female). There are 3497 genuine and 49875 spoofed trials. Spoofed speech is again generated in accordance with one of the five known attacks (S1–S5). All data in the development set can be used for designing and optimising spoofing countermeasures, for example, to tune classifier hyper-parameters. For the training set, all meta information is again provided and may be used freely.

Evaluation set - The evaluation set is composed of 9404 genuine and 184000 spoofed utterances collected from 46 speakers (20 male, 26 female).

In contrast to training and development sets, spoofed data in the evaluation set are generated in accordance with the



(a) Female speakers



(b) Male speakers

Fig. 3. DET profiles illustrating degradations in ASV performance caused by each of the then spoofing-attack algorithms (S1-S10) in ASVspoof database. Results are shown for an i-vector PLDA ASV system.

full set of ten spoofing-attack algorithms (both known and unknown attacks). This is more representative of the practical application scenario in which there is potential for previously unseen attacks. Spoofing-detection results for the unknown attacks are then expected to shed light on the potential for countermeasures 'in the wild'. No meta information is included in the evaluation set.

Known vs. unknown attacks - (S1–S5) were used as known attacks since they are either easily implemented using standard

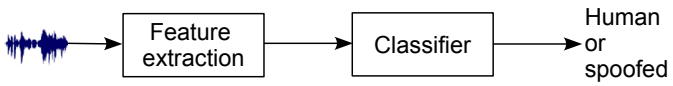


Fig. 4. Simple spoofing-detection framework adhered to by all 16 submissions to ASVspoof 2015.

routines or by off-the-shelf and publicly available, open-source software packages. They are methodologically simple; thus, accessible to non-experts. The S1 and S2 spoofing-attack algorithms are two of the most easily implemented voice-conversion algorithms. The S3, S4, and S5 spoofing-attack algorithms were all implemented using either the HTS[6] or Festvox[7], both of which are publicly available and open source. The remaining unknown attacks (S6–S10) are generally more sophisticated. The S10 spoofing-attack algorithm was also generated using a publicly available open-source toolkit, namely the MaryTTS[8].

III. The ASVspoof 2015 Challenge

This section outlines the ASVspoof 2015 Challenge, which ran from December 2014 to March 2015.

A. Strategy and evaluation rules

Whereas it is the impact of spoofing and countermeasures on ASV performance that is of the greatest interest, ASVspoof 2015 focused exclusively on spoofing detection, that is, detection in isolation from ASV. The approach is illustrated in Figure 4. The focus on simplicity decoupled the evaluation of standalone spoofing detection from the complexities of integrated ASV. This strategy helped maximise participation, which then required no prerequisite expertise in ASV.

The ASVspoof 2015 participants were allowed to submit scores for up to six systems – three for a common training condition and three for a flexible training condition. The common training condition restricts participants to the use of only the ASVspoof 2015 training-set data for countermeasure learning and optimisation. For optional flexible condition submissions, the use of any other database (except VCTK[9]) was permitted. Each participant had to designate one common condition system as their primary submission. Only results for primary submissions were taken into account for system ranking.

B. Submissions

The ASVspoof database was requested by 28 teams from 16 countries, with 16 teams returning primary submissions by the deadline. While 27 additional submissions were also received, this paper focuses on results for primary submissions only. A summary of each of the 16 primary systems is provided in Table III and discussed below.

[6]http://hts.sp.nitech.ac.jp/
[7]http://www.festvox.org/
[8]http://sp-tk.sourceforge.net/
[9]http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

TABLE III. SUMMARY OF 16 PRIMARY SUBMISSIONS TO ASVSPOOF 2015. SYSTEMS ARE RANKED ANONYMOUSLY IN ACCORDANCE WITH AVERAGE EER ON EVALUATION SET, FROM LOWEST (SYSTEM A) TO HIGHEST (SYSTEM P).

System ID	Feature representations	Classifiers	Fusion
A	MFCC, CFCCIF	GMM	Score fusion
B	MFCC, MFPC, CosPhasePC	SVM with i-vectors	i-vector concatenation
C	DNN-based with filterbank output and their deltas as input	Mahalanobis distance on s-vectors	None
D	LMS, RLMS, GD, IF, BPD, MGD, PSP	MLP	Score fusion
E	MFCC, MFCC-CNPCC, PS-MFCC, MGDFCC, MGDCC, WLP-GDCC	GMM	Score fusion
F	NULBP of filterbank features, MFGD, CNPF	SVM, GMM-SVM	Feature and score fusion
G	MFCC, MGDCC, PPP Tandem, Misc features with openSMILE toolkit	i-vector/SVM, Fusion	Feature and score fusion
H	MFCC, TEO, PMVDR, TEO+PMVDR	i-vector with Gaussian back-end	i-vector concatenation
I	MGCC, MGDCC, RPI	GMM	Score fusion
J	Filterbank features, RPS	DNN	Score fusion
K	RPS	GMM	None
L	MLP-based with DCT of raw speech as input	SVM-RBF	None
M	Parametrisation of LP and LTP error	Logistic classifier	None
N	MFCC	i-vector/PLDA	None
O	MFCC	i-vector/SVM	None
P	MFCC	i-vector/PLDA, SVM	None

A System A [34] used two feature parametrisations: 36-dimensional MFCCs and 36-dimensional cochlear filter cepstral coefficients plus instantaneous frequency (CFCCIF). CFCCIF features are based on the combination of an auditory transformation based on CFCCs [35] and instantaneous frequency [36]. A GMM classifier with 128 components was learned separately for each feature set. Finally, the two GMM classifier scores were fused.

B System B [37] exploited three different feature parametrisations: MFCCs, Mel-frequency principal coefficients (MFPCs), and cosine-phase principal coefficients (CosPhasePCs) [38]. MFCCs and MFPCs are spectral magnitude features, while CosPhasePCs use phase information. An i-vector was then computed for each feature set, and the resulting three i-vectors were concatenated into one "super" i-vector that was then passed to a support vector machine (SVM) for scoring.

C System C [39] used deep learning techniques for feature extraction instead of hand-crafted features. Filter bank energies with their deltas were fed into to a deep neural network (DNN). The outputs of the last hidden layer were averaged over all speech frames to produce a new utterance representation referred to as a *spoofing vector* (s-vector). Back-end scoring was carried out using the Mahalanobis distance between s-vectors. Scores were normalized with test normalization.

D System D [40] used multiple feature parametrisation. Magnitude-based features include the log magnitude spectrum (LMS) and residual log magnitude spectrum (RLMS). Phase-based features include group delay (GD), modified group delay (MGD) [41], instantaneous frequency derivative (IF) [42], baseband phase difference (BPD) [43], and pitch synchronous phase (PSP). A multilayer perceptron (MLP) was trained for each feature. All seven MLP-based systems were combined by score averaging.

E System E [44] used several amplitude-, phase-, and linear-prediction-based features: MFCC, product spectrum MFCC (PS-MFCC) [45], MGD with and without energy, weighted linear prediction group delay cepstral coefficients (WLP-GDCCs), and MFCC cosine-normalized phase-based cepstral coefficients (MFCC-CNPCCs) [38]. A GMM back-end with 512 components was used for scoring. Scores of all seven GMMs were fused to obtain the final score.

F System F [46] used three SVM-based approaches. The first used 6844-dimensional histograms of normalised unique local binary patterns (NULBPs) [47], [48] of filterbank features. The other two used 6144-dimensional GMM supervectors computed using modified group delay cepstral coefficients (MGDCCs) and cosine-normalized phase features (CNPFs) [38]. Scores produced from the three systems were further fused with an anti-spoofing supervector-based SVM system, which used all three features described above.

G System G [49] used MFCCs, MGDCCs [38], phonetic level phoneme posterior probability (PPP) tandem features [50] and openSMILE [51] utterance-level feature vectors, which include various speech features and their functionals, e.g. MFCC, line spectral pairs, voicing probabilities, F_0, F_0 envelope, jitter, and shimmer. Utterance level i-vectors were extracted separately with MFCCs, MGDCCs, PPP, and their combinations. For each type of utterance-level feature or i-vector, an SVM with a polynomial kernel was used for scoring. Scores from the five different SVM-based systems were fused.

H Similar to System B, System H [52] used an i-

vector framework. First, two frame-level feature parametrisations were extracted: cepstral-based perceptual minimum variance distortionless response (PMVDR) [53] and speech-production-motivated Teager energy operator (TEO) autocorrelation features [54]. The i-vectors of each feature parametrisation were then extracted. The PMVDR and TEO i-vectors were then concatenated into one vector, which was passed through LDA for dimensionality reduction before a Gaussian back-end was used for classification.

I System I [55] used three types of features: relative phase information (RPI) [56], MFCCs, and MGDCCs [38]. The relative phase feature contains phase information only. All features have a dimensionality of 38. A GMM-based classifier with 256 components was trained for each type of feature, and the scores produced from each sub-system were linearly combined at the score level for classification.

J System J [57] used both magnitude and phase information in the form of linear filter bank energies and relative phase shift (RPS) [58], a phase-based feature for spoofing detection, respectively. For each type of feature, a DNN with two hidden layers was used to carry out classification. Scores from the two DNNs were fused to make the final decision.

K System K [59] used RPS [58]. The RPS values were processed with Mel filters and the discrete cosine transform (DCT). The average value of unwrapped RPS was also included. Finally, 63-dimensional features were obtained after augmenting delta and double delta coefficients. A GMM with 512 components was used for scoring. All signals were downsampled to 8 kHz before feature extraction.

L System L [60] used an MLP network for feature extraction, using 128 DCT coefficients from raw speech signals as input, and producing hidden activations of the third hidden layer as features. Features were modelled with an SVM back-end using a radial basis function (RBF) kernel.

M System M [61] is based on the analysis of linear prediction (LP) error, estimates of which were passed through a long-term predictive coding (LTP) algorithm. Ten different parameters, including the mean LP energy and LTP error [62], were used to construct feature parametrisations, which where scored using a logistic classifier.

N System N used i-vectors as features for spoofing detection. A 200-dimensional i-vector was first derived from MFCCs with respect to a UBM with 512 components. The i-vector was passed through a PLDA model to compute the likelihood score for classification.

O System O also used an i-vector-based framework. It first extracted MFCC features then used a UBM with 512 components to extract a 400-dimensional i-vector. The i-vector was passed through an SVM with an RBF kernel for spoofing detection.

P System P is similar to Systems N and O, but after the i-vector was extracted, a PLDA model was applied to remove channel and speaker-identity effects while keeping spoofing effects. The resultant low-dimensional vector was passed through an SVM with an RBF kernel for detection.

In general, ASVspoof 2015 participants focused more on the development of new feature parametrisations tailored to spoofing detection rather than on the development of new classifiers.

IV. ASVspoof 2015 results and analysis

This section summarises the results of ASVspoof 2015, presents an analysis of the correlation between scores produced from each submitted system, and reports a series of fusion experiments to evaluate performance through system combination.

A. Metrics

To objectively measure and rank the relative performance of different countermeasures submitted to ASVspoof 2015, we adopted the EER, a standard metric in assessing the accuracy of ASV and other biometric systems. The participants were required to assign a single, real-valued *detection score* to each of the evaluation set recordings. We adopted the (arbitrary) convention that higher detection scores indicate greater likelihood to observe genuine human speech, while relatively lower scores indicate greater likelihood of a spoofing attack.

As a binary classification task, any countermeasure system may face two types of errors. A *false alarm* occurs when a countermeasure system incorrectly classifies an actual spoofing attack as a human sample. A *miss*, in turn, occurs when an actual human speech sample is misclassified as a spoofing attack. Given all the detection scores of a particular system, we first compute the empirical false alarm and miss rates, denoted respectively as $P_{\rm fa}(\theta)$ and $P_{\rm miss}(\theta)$ at threshold θ. They are computed as

$$P_{\rm fa}(\theta) = \frac{\#\{\text{spoof trials with score} > \theta\}}{\#\{\text{total spoof trials}\}},$$
$$P_{\rm miss}(\theta) = \frac{\#\{\text{genuine trials with score} \leq \theta\}}{\#\{\text{total genuine trials}\}},$$

so that $P_{\rm fa}(\theta)$ and $P_{\rm miss}(\theta)$ are, respectively, monotonically decreasing and increasing functions of θ. The EER then corresponds to the threshold $\theta_{\rm EER}$ at which the two detection error rates coincide[10], i.e. $\text{EER} = P_{\rm fa}(\theta_{\rm EER}) = P_{\rm miss}(\theta_{\rm EER})$.

[10]It is rarely possible to determine $\theta_{\rm EER}$ exactly since $P_{\rm fa}(\theta)$ and $P_{\rm miss}(\theta)$ change in discrete steps. One can interpolate the values near the EER operating point or use more advanced methods, such as the ROC convex hull (ROCCH-EER) approach implemented in the Bosaris toolkit[11].

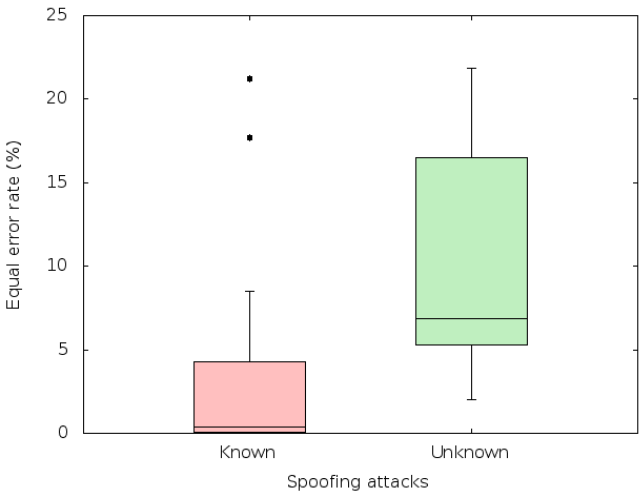


Fig. 5. Tukey boxplots showing a comparison of detection performance for known and unknown spoofing attacks.

The EERs were calculated independently for each of the spoofing-attack algorithms. In the analysis, we used the average EER as the objective measure to be consistent with ASVspoof 2015. The averaged EER is calculated as the mean of individual attack-specific EERs.

B. Results

We first summarise the challenge results for 16 primary submissions used for ranking. The results are presented in Table IV. Figure 5 shows Tukey boxplots of the same results, sub-divided across known and unknown attacks. The left red box shows that known attacks are detected relatively easily; most of the participants (12 out of 16) achieved very low EERs of below 1% for known attacks. The right green box shows that EERs for unknown attacks are considerably higher than those for known attacks. System A [34], which provided the best results for unknown attacks, exhibited an EER of 2.013%, which is still five times higher than with same systems for known attacks.

The results also illustrate the potential of over-fitting countermeasures to known attacks, leading to low performance against unknown attacks absent from the training set. For example, Systems D and I exhibited much lower EERs (0.003 and 0.005%, respectively) than System A (0.408%) for known attacks. However, the same systems produced EERs more than 2 and 3 times that obtained with System A for unknown attacks. One reason unknown attacks produced higher EERs is that countermeasures were not reliable in detecting one of the unknown attacks (S10), the only waveform concatenation approach to speech synthesis used in the creation of the ASVspoof 2015 dataset. As can been seen from Table IV, even the best EER for S10 was as high as 8.49%. In general, the results confirm the importance of developing generalised countermeasures.

TABLE IV. SUMMARY OF PRIMARY SUBMISSION AND ORACLE-FUSION RESULTS FOR THE ASVSPOOF 2015 CHALLENGE.

	Average Equal Error Rates (EERs) [%]				
	Known	Unknown			All
System ID	AVG S1-S5	AVG S6-S9	S10	AVG S6-S10	AVG
A	0.408	0.394	8.490	**2.013**	**1.211**
B	0.008	0.009	19.571	3.922	1.965
C	0.058	0.098	24.601	4.998	2.528
D	**0.003**	0.003	26.142	5.231	2.617
E	0.041	0.085	26.393	5.347	2.694
F	0.358	0.453	28.581	6.078	3.218
G	0.405	0.304	30.021	6.247	3.326
H	0.670	0.042	37.068	6.041	3.355
I	0.005	0.839	32.651	7.447	3.726
J	0.025	0.033	40.708	8.168	4.097
K	0.210	0.195	43.638	8.883	4.547
L	0.412	7.310	35.890	13.026	6.719
M	8.528	17.423	31.574	20.253	14.391
N	7.874	15.580	43.991	21.262	14.568
O	17.723	14.532	41.519	19.929	18.826
P	21.206	15.763	46.102	21.831	21.518
Oracle Fusion	0.000	0.000	5.225	1.045	0.523

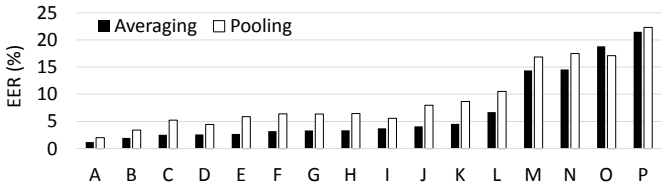


Fig. 6. Comparison of primary submitted systems in terms of EERs computed by averaging (as in ASVspoof evaluation plan) and pooling.

C. Averaged versus pooled EER

To further analyse the primary systems in terms of their *calibration* performance, Fig. 6 contrasts the averaged EERs over all ten spoofing-attack algorithms (as in the evaluation plan) against EERs computed from scores pooled across all spoofing-attack algorithms. A large relative difference between these two EERs would be indicative of mutually incompatible output scores across the different attacks, implying the difficulty in setting attack-independent thresholds in practical spoofing-detector implementations. Such analysis was carried out recently [63, Table 12] for deep spoofing features. We decided to repeat it here for all the submitted primary systems to ASVspoof 2015.

Fig. 6 shows that most systems were not calibrated well across the different attacks; the relative increase from averaged to pooled EERs is 64.9% on average, and notably higher for some top-ranked systems. While the two best systems (A and B) retained their top rank, the order of Systems C and D changed. Similarly, System I appeared to have better calibrated scores compared to Systems E, F, G, and H, exhibiting a higher rank in terms of pooled EERs. Finally, System O was the only one with a pooled EER being lower than the averaged EER.[12]

[12]Unlike the other submitted systems, System O scores were *binary* values (decisions) rather than real values (scores), making the EER not well-defined; hence, causing such an anomaly.

D. Correlation analysis

The previous subsection discussed the performance of each countermeasure proposed by the 16 participants. However, some may be complementary to each other, and their combined performance may be further improved by fusing countermeasure scores. To confirm this possibility, we analysed the correlation between countermeasures using scores for the ASVspoof 2015 evaluation set. Countermeasures that are independent from others are expected to produce a correlation of zero. Countermeasures strongly dependent on others should in contrast produce a correlation nearing unity.

Fig. 7 (a) represents a boxplot of the average absolute correlations across countermeasures for known attacks. Systems A, B, C, I, and N were more correlated, while System L had the lowest correlation to other systems. This might be explained by noting (Table III) that System L directly uses DCT coefficients obtained from raw speech signals as the inputs to the DNN system to automatically learn discriminative features, while all the other systems use traditional hand-crafted features.

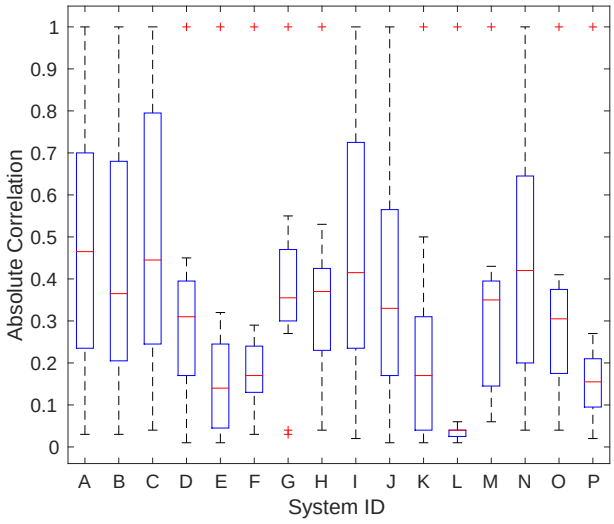
Fig. 7 (b) presents a similar boxplot for unknown attacks. The correlation patterns between the known and unknown attacks are quite different. For the unknown attacks, we see that Systems C, F, and G seemed to behave differently from other systems. This indicates that some of the 16 systems are expected to be complementary to each other and that there is potential to improve performance through score fusion.

E. Countermeasure fusion

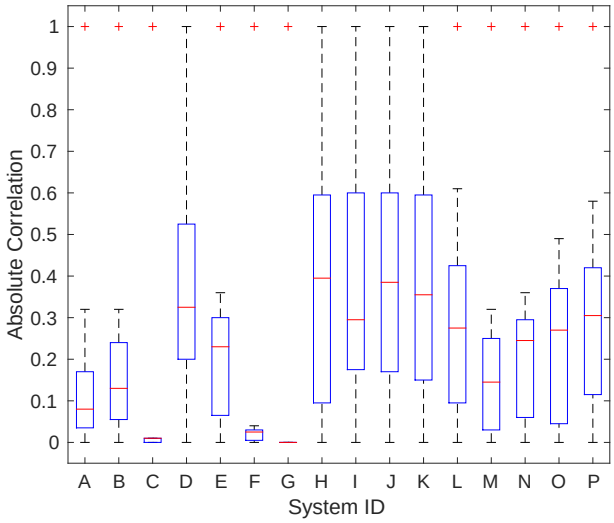
Classifier-fusion techniques are commonly used to improve the performance of modern ASV systems [64], [65], [66] by combinations of many base classifiers that are, in some sense, complementary to each other. The most common approach is linear fusion in the form $s_{\text{fused}} = w_0 + w_1 s_1 + \cdots + w_K s_K$, where s_{fused} denotes the fused score, $\{s_k\}$ are the base classifier scores and $\{w_k\}$ are the fusion weights, w_0 indicating a bias term. The fusion weights are determined by optimizing a logistic loss [64] on a labeled development set of base classifier scores.

Fusion is well-justified in the context of spoofing countermeasures: certain features or classifier architectures might excel in detecting certain attacks, whereas alternative approaches might perform better for different forms of spoofing attacks. As mentioned above, many ASVspoof participants already fused acoustic features, i-vectors, or back-end classifiers internally in their countermeasures. In contrast, we explore *site-wise* fusion to combine all the 16 systems. Since they were developed by different individuals and teams using diverse methods, control parameter settings, and implementations, the set of scores should make an interesting fusion pool.

Since we have access only to the evaluation-set scores produced by each team, we consider an *oracle*-fusion system, where the weights are trained directly on the evaluation-set scores. Nonetheless, this gives us an experimental bound of the best achievable performance of the countermeasure pool. We first normalised the scores of the individual systems to have zero mean and unit variance and then used the BOSARIS



(a) Known attacks



(b) Unknown attacks

Fig. 7. Boxplot of absolute correlations across countermeasures

toolkit[13] to train the fusion weights. Fusion results are presented in the last row of Table IV. As expected, the oracle-fusion system exhibited the lowest error rates for both known and unknown attacks, with average EERs of 0 and 1.045%, respectively. Even though the fusion system can detect all known attacks, it still has difficulty detecting some unknown attacks. In particular, the EERs for S8 and S10 were 0.05 and 5.18%, respectively.

We now analyse the relative contribution (importance) of each countermeasure subsystem in our fusion. We first display the fusion weights in Fig. 8. Unsurprisingly, Systems A and B had the highest weights, as they exhibited considerably lower EERs compared to the other systems. Interestingly, System K also had a large weight, even though it does not rank particularly high nor is the least correlated system with the

[13] Available at: https://sites.google.com/site/bosaristoolkit/

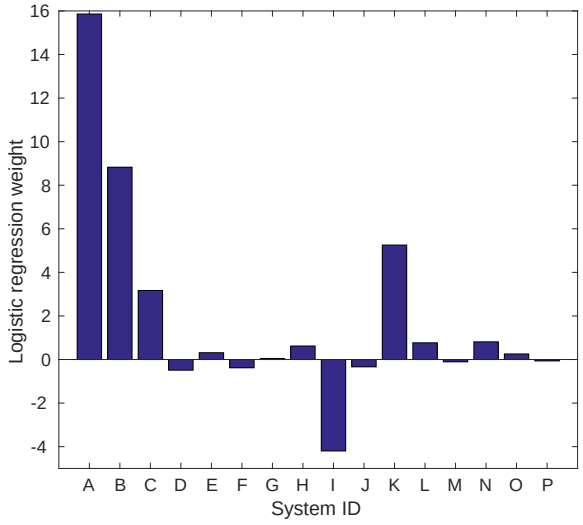


Fig. 8. Bar plot for logistic regression weights in oracle-fusion system

other systems.

Another way to evaluate the importance of a subsystem is to observe a change in the EER of the fusion result when that subsystem is excluded from the fusion pool. To this end, we exclude one system at a time and report the absolute EER difference with respect to a reference EER (obtained by fusing all the systems). Hence, the *higher* the change in EER, the more useful we deem that system to be. Note that we trained all the 16 leave-one-out fusions from scratch.

The results are plotted in Fig. 9 separately for the known and the unknown attacks. Concerning the known attacks, as our reference EER (fusion of all systems) was 0.0%, leave-one-out fusion could only increase or retain the EER. By excluding any of the systems (A, C, F, I, or K), the EERs increased. The greatest EER increase, 0.023%, was obtained by excluding System K. Interestingly, even though System D exhibited the lowest EERs for the known attacks, fusion performance did not change by excluding it. The relative small fusion weight of System D may explain this result. For the unknown attacks, excluding System A, B, C, I, K, or L yields considerably increased EERs. Exclusion of the other systems had little to no impact.

V. Progress and future directions

This section surveys progress in the field of ASV spoofing countermeasures reported in the literature since ASVspoof 2015. While it does not stem from the evaluation itself, the availability of a common database, protocols, and metrics nonetheless facilitates meaningfully comparable research. Included is a review of other related work and research directions for the future.

A. Post-evaluation progress

Reviewed below are post-evaluation studies conducted with the ASVspoof 2015 database. They are categorised in terms of their focus on features and classifiers. The results are discussed afterwards.

*a) **Features**:* An extensive evaluation of three different categories of features is reported in [67]. They include short-term power spectrum features, short-term phase features, and spectral features with long-term processing. Static and dynamic coefficients are also evaluated separately. Dynamic coefficients are found to be more useful than static coefficients. This is reasonable since voice-conversion and speech-synthesis techniques may not properly model the dynamic characteristics of speech.

A comparison of six different feature parametrisations is reported in [68]. They include two magnitude-based features (LMS and RLMS) and 4 phase-based features (IF, BPD, GD, and MGD). Each feature is studied using different vector dimensions including high-dimensional (256) and low-dimensional (23 after application of a Mel filterbank) and a low-frequency range of high-dimensional features (first 128 bins) and high-frequency range of high-dimensional features (last 128 bins). From each combination, dynamic coefficients are also studied (first and second derivatives). High-dimensional configurations and dynamic coefficients are shown to enable better performance. Magnitude- and phase-based features are shown to enable similar levels of performance. The fusion of high-dimensional static parameters and of high-dimensional dynamic coefficients are shown to be the most effective combinations.

The use of i-vector representations for spoofing detection is explored further in [52]. Two base features are used to derive i-vectors: PMVDR and the non-linear speech-production-motivated TEO critical band auto-correlation envelope [69], [70]. A UBM and i-vector extractor are trained separately for each feature using all the training data. Super-i-vectors are then obtained from the concatenation of each individual i-vector, resulting in an i-vector of double dimensionality. Finally, LDA is used to reduce the dimensionality to the original i-vector size.

One study [63] investigates various features based on deep learning techniques for spoofing detection. Fusion of DNN-based features with an LDA classifier and bidirectional long short-term memory (BLSTM)-based features with an SVM classifier show encouraging spoofing-detection performance.

Excitation source-based features, such as fundamental frequency contour and strength of excitation [71], are investigated in [72]. These are fused at score level with MFCCs and CFCCIFs.

Constant Q cepstral coefficients (CQCCs) based on the constant Q transform (CQT) [73] are reported in [74]. The CQT is an alternative time-frequency analysis tool to the short-term Fourier transform that provides variable time and frequency resolution. It provides greater frequency resolution at lower frequencies but greater time resolution at higher frequencies [74]. Multiple combinations of static, delta, and acceleration coefficients are evaluated.

Other recent studies [75], [76], [77], [78] also investigate different features and their combinations. Study [75] compares MFCCs, Mel-warped overlapped block transformation parameters [79], and speech-signal-based frequency cepstral coefficients [80]. Following the hypothesis that speech synthesis and voice conversion do not model high-frequency infor-

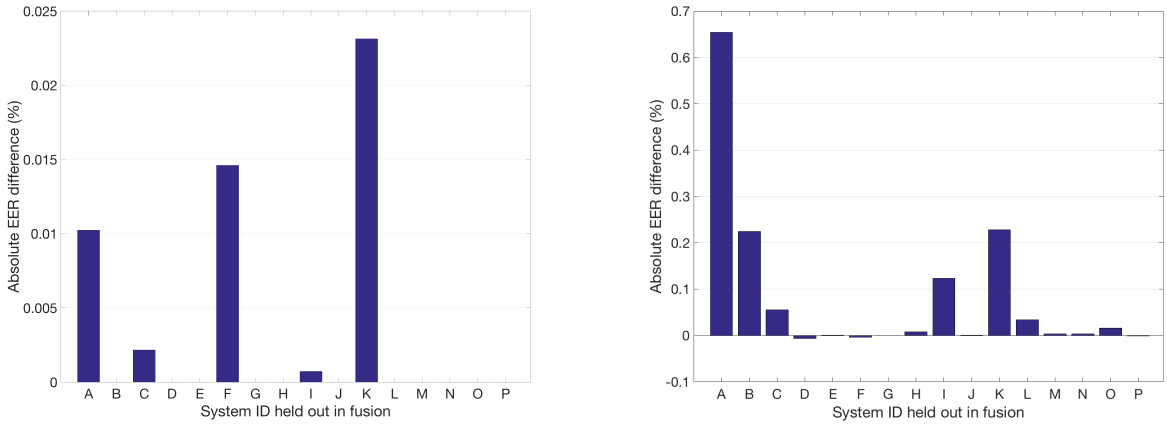


Fig. 9. Absolute equal error rate (EER) difference for both known (left) and unknown (right) attacks using leave-one-out fusion. We leave out one of 16 systems at a time and train a linear-fusion system using other 15 systems. Y-axis shows absolute change in EERs of this fusion relative to fusion of all 16 systems. Hence, higher value in graph indicates that system to be more 'important', as excluding it increases EER.

mation particularly well, an inverted filterbank, which offers greater resolution at higher frequencies, is reported in [75]. Study [76] explores a set of acoustic features: MFCCs, MFCC-CNPCCs, PS-MFCCs, linear-prediction cepstral coefficients (LPCCs), and linear-prediction residual cepstral coefficients, and bottleneck features extracted from each acoustic feature. Study [77] investigates sub-band autoencoder (SBAE)-based features. The connectivity of autoencoder units is restricted in such a manner that each unit in the first hidden layer captures information about a specific band of the speech spectrum. Other studies [78] explore nonlinear prediction-based features for spoofing detection.

b) ***Classifiers****:* As was the case with the ASVspoof 2015 evaluation, the post-evaluation literature is dominated by the use of two general techniques to classification, namely deep learning, e.g. [68], [52], [76] and GMMs, e.g. [67], [72], [74], [75], [76].

A broader comparison of five different classifiers with a common MFCC-based front-end is reported in [81]. The study of a GMM, GMM supervectors, generalized linear discriminant sequence kernel SVMs, and i-vectors found that discriminative classifiers outperform generative ones in the case of the ASVspoof 2015 development set. In contrast, generative classifiers perform better for the evaluation set, especially in the case of unknown attacks. This might suggest that generative classifiers have better potential for generalization.

c) ***Performance****:* Table V summarises spoofing-detection results reported in the post-evaluation studies discussed above. To illustrate progress, the results for the top three performing systems (Systems A, B, and C) of ASVspoof 2015 in addition to brief system descriptions in terms of features and classifiers are also included. In contrast to Table III, the average results for unknown attacks are illustrated separately for S6–S9 and S10 on its own.

For known attacks, many recently reported systems exibit close to 0% EER. However, these systems tend not to generalise well to known attacks, with some exhibiting comparatively poor performance for unknown attacks. Results for S10 remain particularly poor. Other systems, including those that are less reliable regarding known attacks, generalise much better to unknown attacks.

While the performance of the system reported in [74] delivers an EER which is six times higher than that of System B of the ASVspoof evaluation for known attacks, the EER for unknown attacks is less than 0.5% (S6–S10) and only marginally greater than 1% for S10. Overall, only two systems improve on the performance of the best ASVspoof 2015 system [34] for both known and unknown attacks, delivering EERs of 0.9% [67] and 0.3% [74] (S1–S10). Both systems use a simple GMM classifier and a single feature set (no fusion).

B. Investigations from other perspectives

We now discuss a number of different aspects that could affect the reliability of spoofing countermeasures in practical scenarios.

Study [82], [83] evaluates the effect of additive noise on spoofing-detection performance. System performance is found to degrade significantly with increasing noise level. Phase-based features seem to be more robust than magnitude-based features. Similar work is reported in [84], which also studies multi-condition training using both clean and noisy signals in addition to the effectiveness of speech-enhancement methods in improving performance. Although both studies show that error rates are still far above those obtained with clean signals, multi-condition training and speech enhancement significantly improves reliability under noisy conditions.

The effect of mismatch between training and evaluation data is evaluated through a cross-database assessment in [85]. The ASVspoof 2015 database and the similarly named AVspoof database [86] are used alternatively for training/testing. The results indicate that countermeasures that are effective on matched conditions are much less effective in the case of mismatched conditions, another example of poor generalization.

TABLE V. COMPARATIVE PERFORMANCE OF SEVERAL POST-EVALUATION RESULTS (INCLUDING TOP THREE PERFORMING SYSTEMS OF ASVSPOOF 2015). ALL RESULTS EXPRESSED IN TERMS OF EER (%) FOR ASVSPOOF 2015 EVALUATION SET. ERROR RATES ARE SHOWN SEPARATELY FOR KNOWN ATTACKS, UNKNOWN ATTACKS WITH AND WITHOUT S10, S10 ON ITS OWN, AND GLOBAL AVERAGE (S1–S10).

			Known	Unknown			All
System	Features	Classifiers	AVG S1-S5	AVG S6-S9	S10	AVG S6-S10	AVG
[74]	CQCC (acceleration coeffs.)	GMM	0.048	0.312	1.065	0.463	0.255
[67]	Linear Frequency Cepstral Coeffs. (dynamic coeffs.)	GMM	0.110	0.065	8.185	1.689	0.900
[63]	DNN-based deep features, BLSTM-based deep features	LDA, SVM, score fusion	0.000	0.025	10.700	2.160	1.080
A	MFCC, CFCCIF	GMM	0.408	0.394	8.490	2.013	1.211
[72]	MFCC + CFCCIF + 3rd derivative of F0	GMM, score fusion	0.111	0.072	15.300	3.118	1.614
[77]	SBAE+MFCC (static+dynamic)	GMM, score fusion	0.352	0.270	16.520	3.520	1.936
B	MFCC, MFPC, CosPhasePC	SVM with i-vectors	0.008	0.009	19.571	3.922	1.965
[76] (2)	PS-MFCC (dynamic coeffs.)	DNN	1.164	0.798	12.860	3.210	2.187
C	DNN-based with filterbank output and their deltas as input	Mahalanobis distance on s-vectors	0.058	0.098	24.601	4.998	2.528
[68]	LMS, RLMS, IF, BPD, GD, MGD	ANN, score fusion	0.000	0.000	27.790	5.558	2.779
[81]	MFCC (static + dynamic coeffs.)	GMM	0.500	-	-	5.520	3.010
[76] (1)	LPCC (dynamic coeffs.) + bottleneck features	DNN	0.000	0.000	33.000	6.600	3.300
[52]	F-bank feats. + i-vector (TEO+PMVDR)	Fusion of 2 DNN	0.628	0.765	27.940	6.200	3.414
[78]	Long-term prediction+LP-Nonlinear LP + MFCC	GMM, score fusion	0.012	0.010	51.110	10.230	5.121

Other studies also assess the impact of spoofing and more importantly, of countermeasures on ASV. To date, the only study that investigates their integration through experiments with the ASVspoof database is [87]. This evaluation was carried out using a joint ASV+CM protocol. Five independent countermeasures were evaluated and fused in a standalone spoofing detection task. The GMM-UBM and i-vector ASV systems were also evaluated and fused. Finally, both CM and ASV modules were integrated following two different schemes, namely cascade and parallel integrations. The results indicate that countermeasures are effective in reducing false acceptances due to spoofing attacks, particularly in the case of cascaded integration.

Another important aspect relates to the comparison of automatic means to detect spoofing to the performance of humans in detecting the same attacks. This exploratory study can help shed light on the acoustic cues used by the human auditory system for spoofing detection. There are a few interesting comparisons of human performance to that of ASV systems [88], [89], [90], but only one attempt to compare human performance to automatic spoofing detection [91]. This study compares the performance of 100 native English listeners to an automatic approach using a combination of MFCC- and CNFP-based detectors. Both MFCCs and CNPFs include 18 dimensional static features, their deltas and delta-deltas. The results indicate that, on a subset of the ASVspoof 2015 database, automatic detectors outperform human listeners for all spoofing attacks, except S10. This finding would suggest that automatic spoofing-detection algorithms and human listeners use different cues to distinguish spoofed and genuine speech. This supports similar findings previous studies [92], [93] found for voice imitations. This implies that there is a potential acoustic feature that can detect S10, and more recent studies, such as [74], have found features useful for detecting S10 spoofing attacks.

C. Future directions

The ASVspoof 2015 Challenge formed an important first step towards the benchmarking of spoofing countermeasures for ASV. While achieving substantial inroads towards the development of generalised countermeasures in particular, further work is needed to address a number of outstanding questions regarding their practicality. A number of topics for consideration in further work, not just within the scope of ASVspoof, are now discussed.

*a) **Replay spoofing attacks**:* The ASVspoof 2015 Challenge concentrated on voice-conversion and speech-synthesis spoofing attacks. They are, however, not the only forms of spoofing attacks. Others already reported in the literature include impersonation and replay spoofing. The potential severity of impersonation attacks remains somewhat unclear and, like voice conversion and speech synthesis, they require specialized expertise to successfully impersonate another speaker [94]. In contrast, replay attacks require only everyday recording and replaying equipment, which is readily available to the public. There is also ample evidence that replay attacks also pose a threat to ASV reliability [95], [96], [97]. Replay attacks may therefore be the most prolific in practice; thus, they warrant consideration in the context of the ASVspoof initiative.

*b) **Text-dependence**:* the threat of spoofing relates to authentication applications. The need for user convenience thus dictates the use of only relatively short utterances and, in turn, text-dependent ASV. Some form of text constraints can reduce the acoustic mismatch between enrolment and test utterances; therefore, delivering higher recognition performance than might otherwise be achieved in a text-independent mode. While ASVspoof 2015 focused exclusively on text-independent ASV, future editions should also take into account text-dependent conditions that have greater relevance to authentication scenarios.

*c) **Impact on ASV**:* while the threat of each spoofing attack included in the ASVspoof 2015 database was validated using a state-of-the-art i-vector PLDA ASV system, the evaluation focused exclusively on spoofing detection independent of ASV. The evaluation of integrated spoofing countermeasures and ASV would require the joint optimisation of combined classifiers. Furthermore, different strategies can be adopted

for system combination, e.g. via a simple series or parallel score fusion or through the joint modelling of speaker and spoofing characteristics [98]. The evaluation of integrated spoofing countermeasures would thus hamper the meaningful comparison of different countermeasure technologies, where the latter would otherwise be the only difference between two experiments and results. Even so, the impact of spoofing countermeasures on ASV reliability is the primary interest. Future editions of ASVspoof may therefore take into account the combination of spoofing countermeasures with a standard ASV system using a secondary metric.

d) ***Sensor-level spoofing and channel effects****:* the wider biometrics community considers spoofing to be an attack carried out at the sensor level. In terms of ASV, this would imply spoofing attacks at the microphone level. Automatic speaker verification systems are currently deployed in a diverse range of application scenarios including both logical and physical access. Whereas in the latter, the microphone usually forms an integral part of the full ASV system, the microphone may lie beyond the system realm and be uncontrolled in some logical access scenarios, including telephony applications. The user typically uses his/her telephone handset; thus, the microphone is uncontrolled. The microphone can even be bypassed with the coupling of spoofing-attack data directly to the transmission line [99]. Similar to the NIST speaker-recognition evaluations, future editions of ASVspoof may include multiple microphone and channel conditions, perhaps including different bandwidth conditions, to study the effects on countermeasure performance. The consideration of transmission channel (e.g. telephone, mobile, VoIP) and codec variability are important since codecs, such as CELP and GSM, are technically similar to speech-waveform-generation methods used for speech synthesis and voice conversion. Accordingly, the application of speech coding may cause current countermeasures to label genuine speech as spoofed speech. Physical access scenarios may also be addressed in the future through the re-recording of spoofing attacks with a fixed microphone.

e) ***Additive noise and reverberation****:* recent studies [100], [82], [83] indicate that some countermeasures offer little resistance to additive noise, with spoofing-detection performance degrading much more rapidly as a function of falling signal-to-noise ratio than is typical for ASV performance. This suggests either an intrinsic difficulty in the problem or what is possibly the result of countermeasures over-trained to the ASVspoof database, which consists of technically high-quality speech. The impact of reverberation has also been investigated [83]. With many logical and physical access scenarios offering the potential for adverse acoustic conditions, these effects warrant further consideration in the future.

f) ***Passive countermeasures****:* voice conversion and speech synthesis are perhaps the two forms of spoofing attacks that fundamentally necessitate active countermeasures such as all those discussed in this article. They are not the only line of defence, however. Passive countermeasures, such as prompted-text ASV or challenge-response mechanisms, can provide an alternative or added protection, especially from replay spoofing attacks. Currently lacking is an objective and scientific approach to validate their potential. While the latter is self-evident, the evaluation of passive countermeasures is likely to be much more complex than for active countermeasures. Nonetheless, they warrant greater attention in the future.

g) ***Evaluation metrics****:* for simplicity, ASVspoof 2015 used the EER as the sole evaluation metric. The EER is a 'threshold-free' metric in the sense that the detection threshold need not be optimised as part of the evaluation. Instead, it is set with the use of ground-truth knowledge, namely the genuine or spoofed speech labels for the dataset under evaluation. Furthermore, EERs were computed as averages of attack-specific EERs, rather than from scores pooled across all attacks. While the use of an EER metric probably encouraged the focus on novel feature extraction and classifier architectures, reliable score calibration across development and evaluation sets and across different spoofing attacks and acoustic conditions will be important to any real-world deployment. Evaluation metrics that encourage the development of techniques that are easy to calibrate across different conditions, thus, deserve attention.

h) ***Spoofing corpora for the future****:* Speech-synthesis and voice-conversion technologies have advanced significantly recently on account of new deep learning methods. Spoofed speech generated using advanced DNNs (e.g. [101]) or using spoofing-specific strategies, such as generative adversarial networks [102], should be considered in future spoofing corpora. In the ASVspoof2015 database, the STRAIGHT vocoder was used for eight out of the ten spoofing-attack algorithms to generate high-quality speech waveforms. It is obviously important to include not only high-quality vocoders but also other types of vocoders and speech-waveform-generation methods in the future.

A number of such topics will be considered for study in the second edition of ASVspoof being held in 2017. Appropriate standard anti-spoofing systems and/or features may be provided as publicly available baselines. Interested readers are invited to follow progress via the challenge website[14].

VI. Conclusions

We described the ASVspoof initiative and the first common dataset, protocols, and metrics to foster progress in the field. A comprehensive overview of the database was provided and a new visualisation of differences between genuine and spoofed speech illustrated in a new light the challenge in developing effective spoofing countermeasures. While all ten spoofing-attack algorithms used to generate the ASVspoof 2015 dataset were successful in circumventing a state-of-the-art ASV system, the results of 16 different systems submitted to the evaluation show that some are easily detected. Others prove more difficult to detect with evaluation results showing recurrent weaknesses to some attacks. While fusion experiments have the potential to improve performance, vulnerabilities remain.

While illustrating the potential to protect ASV systems from spoofing, the ASVspoof 2015 results demonstrate the need for further investigation. The latter part of the paper overviewed the literature on post-evaluation and showed tremendous advances in detection performance, mostly as a result of focusing

[14] www.spoofingchallenge.org

on the development of new features, resulting in an average detection EER of less than 0.3%.

Even with such rapid progress, however, many research issues need to be addressed. A number of issues already investigated outside the scope of ASVspoof show that solutions to address vulnerabilities to spoofing are much more diverse than those explored within the scope of ASVspoof. Perhaps the most important of these relates to a study of replay attacks, which may be most prolific in practice, and the study of spoofing impact on text-dependent ASV. Finally, while ASVspoof has been largely successful in laying the foundations for further investigation to develop more reliable approaches to spoofing detection, ultimately, it is the effect on ASV itself that is of the greatest interest.

ACKNOWLEDGEMENTS

The authors would like to thank the following for their help in creating the ASVspoof 2015 database: Dr. Daisuke Saito (University of Tokyo), Prof. Tomoki Toda (Nagoya University), Mr. Ali Khodabakhsh (Ozyegin University), Dr. Cenk Demiroglu (Ozyegin University), Dr. Linghui Chen and Prof. Zhen-Hua Ling (University of Science and Technology of China).

The authors would also like to thank the participants of ASVspoof 2015 and those who contributed to the associated special session held at Interspeech 2015.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639309001289

[2] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Syst. J.*, vol. 40, no. 3, pp. 614–634, Mar. 2001. [Online]. Available: http://dx.doi.org/10.1147/sj.403.0614

[3] ISO/IEC 30107, "Information technology – biometric presentation attack detection," *International Organization for Standardization*, 2016.

[4] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech*, 2013.

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, no. 0, pp. 130 – 153, 2015.

[6] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, April 2016.

[7] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[8] L. Ghiani, D. Yambay, V. Mura, S. Tocca, G. L. Marcialis, F. Roli, and S. Schuckers, "LivDet 2013 fingerprint liveness detection competition 2013," in *International Conference on Biometrics (IBC)*, 2013, pp. 1–6.

[9] M. M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, F. Roli, J. Yan, D. Yi, Z. Lei, Z. Zhang, S. Z. Li, W. R. Schwartz, A. Rocha, H. Pedrini, J. Lorenzo-Navarro, M. Castrilln-Santana, J. Mtt, A. Hadid, and M. Pietikinen, "Competition on counter measures to 2-d facial spoofing attacks," in *Biometrics (IJCB), 2011 International Joint Conference on*, Oct 2011, pp. 1–6.

[10] 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.

[11] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech*, 2015.

[12] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[13] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information." in *Proc. Interspeech*, 2013.

[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[15] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[16] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[17] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. Interspeech*, 2011.

[18] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[19] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[20] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[21] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: a comparison," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2057–2061. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2015/i15_2057.html

[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[23] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.

[24] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[25] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2011, pp. 249–252. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2011/i11_0249.html

[26] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2006/i06_1874.html

[27] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 2010, p. 14. [Online]. Available: http://www.isca-speech.org/archive_open/odyssey_2010/od10_014.html

[28] S. Prince, P. Li, Y. Fu, U. Mohammed, and J. Elder, "Probabilistic models for inference about identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, Jan 2012.

[29] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[30] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[31] D. S. Pallet, J. G. Fiscus, and J. S. Garofolo, "DARPA resource management benchmark test results June 1990," in *Proceedings of the workshop on Speech and Natural Language*, Hidden Valley, Pennsylvania, 1990, pp. 298–305.

[32] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, no. 4, 2013.

[33] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *http://www.spoofingchallenge.org/asvSpoof.pdf*, 2014.

[34] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. Interspeech*, 2015.

[35] Q. Li, "An auditory-based transform for audio signal processing," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 181–184.

[36] T. Quariteri, "Discrete-time speech signal processing," 2002.

[37] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," *arXiv preprint arXiv:1507.08074*, 2015.

[38] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Interspeech 2012*, 2012.

[39] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection-the SJTU system for ASVspoof 2015 challenge," in *Proc. Interspeech*, 2015.

[40] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. Interspeech*, 2015.

[41] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578 – 616, 2007.

[42] P. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[43] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[44] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Proc. Interspeech*, 2015.

[45] D. Zhu and K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–125.

[46] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. Interspeech*, 2015.

[47] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[48] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Proc. Interspeech*, 2013.

[49] S. Weng, S. Chen, L. Yu, X. Wu, W. Cai, Z. Liu, and M. Li, "The SYSU system for the interspeech 2015 automatic speaker verification spoofing and countermeasures challenge," *arXiv preprint arXiv:1507.06711*, 2015.

[50] M. Li and W. Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," in *INTERSPEECH*, 2014, pp. 1120–1124.

[51] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[52] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. L. Hansen, "Joint information from nonlinear and linear features for spoofing detection: An i-vector/DNN based approach," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2016, pp. 5035–5039.

[53] U. Yapanel and J. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, pp. 142 – 152, 2008.

[54] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 201–216, 2001.

[55] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. Interspeech*, 2015.

[56] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1085–1095, 2012.

[57] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. Interspeech*, 2015.

[58] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[59] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "The AHOLAB RPS SSD spoofing challenge 2015 submission," in *Proc. Interspeech*, 2015.

[60] A. Godoy, F. Simões, J. A. Stuchi, M. d. A. Angeloni, M. Uliani, and R. Violato, "Using deep learning for detecting spoofing attacks on speech signals," *arXiv preprint arXiv:1508.01746*, 2015.

[61] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error," in *Proc. Interspeech*, 2015.

[62] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE transactions on speech and audio processing*, vol. 10, no. 8, pp. 620–636, 2002.

[63] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Communication*, vol. 85, pp. 43–52, 2016.

[64] S. Pigeon, P. Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237–248, 2000. [Online]. Available: http://dx.doi.org/10.1006/dspr.1999.0358

[65] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. Leeuwen, P. Matějka, P. Schwartz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, September 2007.

[66] V. Hautamäki, T. Kinnunen, F. Sedlak, K. Lee, B. Ma, and H. Li, "Sparse classifier fusion for speaker verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 21, no. 8, pp. 1622–1631, 2013. [Online]. Available: http://dx.doi.org/10.1109/TASL.2013.2256895

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

[68] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2016, pp. 2119–2123.

[69] U. H. Yapanel and J. H. Hansen, "A new perceptually motivated mvdr-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, no. 2, p. 142152, 2008.

[70] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Audio, Speech and Language Processing*, vol. 9, no. 3, p. 201216, 2001.

[71] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 469–472, 2009.

[72] T. B. Patel and H. A. Patil, "Effectiveness of fundamental frequency (F0) and strength of excitation (SOE) for spoofed speech detection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, March 2016, pp. 5105–5109.

[73] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustic Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[74] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, Bilbao, Spain, June 21-24 2016, pp. 283–290.

[75] D. Paul, M. Pal, and G. Saha, "Novel speech features for improved detection of spoofing attacks," in *Proc. Annual IEEE India Conference (INDICON)*, 2016.

[76] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, Bilbao, Spain, June 21-24 2016, pp. 270–276.

[77] M. H. Soni, T. B. Patel, and H. A. Patil, "Novel subband autoencoder features for detection of spoofed speech," *Interspeech 201*, pp. 1820–1824, 2016.

[78] H. N. Bhavsar, T. B. Patel, and H. A. Patil, "Novel nonlinear prediction based features for spoofed speech detection," *Interspeech 2016*, pp. 155–159, 2016.

[79] M. Sahidullah, "Enhancement of speaker recognition performance using block level, relative and temporal information of subband energies," Ph.D. dissertation, Indian Institute of Technology Kharagpur, India, 1 2015.

[80] K. Paliwal, B. Shannon, J. Lyons, and K. Wójcicki, "Speech-signal-based frequency warping," *IEEE Signal Processing Letters*, vol. 16, no. 4, pp. 319–322, 2009.

[81] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison." in *Proc. Interspeech*. ISCA, 2015, pp. 2057–2061.

[82] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection under noisy conditions: a preliminary investigation and an initial database," *arXiv preprint arXiv:1602.02950*, 2016.

[83] X. Tian, Z. Wu, X. Xiao, E. S. Chng, , and H. Li, "An investigation of spoofing speech detection under additive noise and reverberant conditions," in *Proc. Interspeech*. ISCA, 2016.

[84] H. Yu, A. K. Sarkar, D. A. L. Thomsen, Z.-H. Tan, Z. Ma, and J. Guo, "Effect of multi-condition training and speech enhancement methods on spoofing detection," in *Proc. International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, 7 2016.

[85] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Proc. Interspeech*. ISCA, 2016.

[86] S. Kucur Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics: Theory, Applications and Systems*. IEEE, Sep. 2015, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7358783

[87] M. Sahidullah, H. Delgado, M. Todisco, H. Yu, T. Kinnunen, N. Evans, and Z.-H. Tan, "Integrated spoofing countermeasures and automatic speaker verification: an evaluation on ASVspoof 2015," in *Proc. Interspeech*. ISCA, 2016.

[88] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the nist 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, no. 1, pp. 249 – 266, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200499903561

[89] S. J. Wenndt and R. L. Mitchell, "Machine recognition vs human recognition of voices," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4245–4248.

[90] V. Hautamäki, T. Kinnunen, M. Nosratighods, K.-A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *Proc. INTERSPEECH*, 2010.

[91] M. Wester, Z. Wu, and J. Yamagishi, "Human vs machine spoofing detection on wideband and narrowband data," in *Proc. Interspeech*, Dresden, Sep. 2015.

[92] E. Zetterholm, M. Blomberg, and D. Elenius, "A comparison between human perception and a speaker verification system score of a voice imitation," in *Proc. of Tenth Australian Int. Conf. on Speech Science & Technology*, 2004.

[93] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13 – 31, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639315000503

[94] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[95] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*, ser. Lecture Notes in Computer Science, C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul, and M. Fairhurst, Eds. Springer, 2011, pp. 274–285.

[96] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay

spoofing attacks against automatic speaker verification," in *Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2014.

[97] Z. Wu, S. Gao, E. S. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2014.

[98] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Trans. Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, 2015.

[99] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Speaker recognition anti-spoofing," in *Handbook of biometric anti-spoofing*, S. Marcel, S. Z. Li, and M. Nixon, Eds. Springer, 2014.

[100] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Communication*, vol. 85, pp. 83 – 97, 2016.

[101] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[102] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets

Zhizheng Wu was a research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh since May 2014 and joined Apple as a research scientist in the Siri team in May 2016. He received his Ph.D. from Nanyang Technological University, Singapore. During his studies he joined Microsoft Research Asia (2007 - 2009) and the University of Eastern Finland (2012) as a visiting scientist, and received the best paper award at the Asia Pacific Signal and Information Processing Association Annual Submit and Conference (APSIPA ASC) 2012. He co-organised the first Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) at Interspeech 2015, delivered a tutorial on "Spoofing and Anti-Spoofing: A Shared View of Speaker Verification, Speech Synthesis and Voice Conversion" at APSIPA ASC 2015 and co-organised the first Voice Conversion Challenge (VCC 2016) as a special session at Interspeech 2016. He initiated the first open-source deep learning based speech synthesis system, Merlin: https://github.com/CSTR-Edinburgh/merlin.

Junichi Yamagishi (SM'13) is an associate professor of 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 100 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 in 2010, 2013, 2014, and 2016, respectively.

He was one of organizers for special sessions on "Spoofing and Countermeasures for Automatic Speaker Verification" at Interspeech 2013, "ASVspoof evaluation" at Interspeech 2015 and "Voice conversion challenge 2016" at Interspeech 2016. He has been a member of the Speech & Language Technical Committee (SLTC) and an Associate Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing. He is a Lead Guest Editor for the IEEE Journal of Selected Topics in Signal Processing (JSTSP) special issue on Spoofing and Countermeasures for Automatic Speaker Verification.

Tomi Kinnunen received the Ph.D. degree in computer science from the University of Eastern Finland (UEF, formerly Univ. of Joensuu) in 2005. From 2005 to 2007, he was an associate scientist at the Institute for Infocomm Research (I2R) in Singapore. Since 2007, he has been with UEF. In 2010—2012, his research was funded by a post-doctoral grant from Academy of Finland focusing on speaker recognition. He was the PI in a 4-year Academy of Finland project focusing on speaker recognition and a co-PI of another Academy of Finland project focusing on audio-visual spoofing. He chaired *Odyssey 2014: The Speaker and Language Recognition workshop*. He served as an associate editor in *Digital Signal Processing* from 2013 to 2015. He currently serves as an associate editor in *IEEE/ACM Trans. on Audio, Speech and Language Processing* and *Speech Communication*. He is currently a partner in large H2020-funded OCTAVE project focusing on voice biometrics for physical and logical access control. In 2015–2016 he visited 6 months at National Institute of Informatics (NII), Japan, under a mobility grant from Academy of Finland, with focus on voice conversion, speaker verification and spoofing. He holds the honorary title of Docent at Aalto University, Finland, with specialization area in speaker and language recognition. He has authored and co-authored more than 100 peer-reviewed scientific publications in these topics.

Cemal Hanilçi received B.Sc., M.Sc. and Ph.D. degrees from Uludag University in 2005, 2007 and 2013, respectively, all in electronic engineering. From March to December 2011, he was a visiting researcher at the School of Computing, University of Eastern Finland. From 2014 to 2015, he was a post-doctoral researcher at the same school. Currently, he is an assistant professor at Bursa Technical University, Department of Electrical and Electronic Engineering, in Turkey. His research interests include speaker recognition, anti-spoofing and audio forensics.

Md Sahidullah received his Ph.D. degree in the area of speech processing from the Department of Electronics and Electrical Communication Engineering of Indian Institute 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 (with specialization in Embedded System) from West Bengal University of Technology in 2006. In 2007–2008, he was with Cognizant Technology Solutions India PVT Limited. Since 2014, he has been a postdoctoral researcher with the School of Computing, University of Eastern Finland. His research interest includes speaker recognition, voice activity detection and spoofing countermeasures.

Aleksandr Sizov is a Ph.D. student at University of Eastern Finland. He received his Specialist degree in applied mathematics from the Saint-Petersburg State University, Russia, in 2011, and the M.E degree in computer science from the Saint-Petersburg State University of Information Technologies, Mechanics and Optics, Russia, in 2013. Currently, he is supported by ARAP scholarship from A*STAR, Singapore. His research interests include speaker and language recognition, anti-spoofing and machine learning.

Nicholas Evans is an Associate Professor at EURECOM where he heads research in Speech and Audio Processing. In addition to other interests in speaker diarization, speech signal processing and multimodal biometrics, he is studying the threat of spoofing to automatic speaker verification systems and working to develop new spoofing countermeasures. Previously, his work in spoofing was funded by the EU FP7 ICT TABULA RASA project, continuing today through the EU H2020 project OCTAVE. He co-organised the Spoofing and Countermeasures for Automatic Speaker Verification special session at Interspeech in 2013 and the ASVspoof evaluation at Interspeech in 2015. He was Lead Guest Editor for the IEEE Transactions on Information Forensics and Security special issue in Biometrics Spoofing and Countermeasures, Lead Guest Editor for the IEEE SPM special issue on Biometric Security and Privacy and Guest Editor for the IEEE JSTSP special issue on Spoofing and Countermeasures for Automatic Speaker Verification. He currently serves as an Associate Editor of the EURASIP Journal on Audio, Speech and Music Processing, is a member of the IEEE and the Signal Processing Society and is an elected member of the Speech and Language Technical Committee. He was general co-chair of IWAENC 2014 and technical programme co-chair of EUSIPCO 2015.

Massimiliano Todisco is a postdoctoral researcher in the Speech and Audio Processing Research Group at EURECOM. He received his Ph.D. degree in Sensorial and Learning Systems Engineering from the University of Rome Tor Vergata. Before, he received a post-graduate Master degree in Sound Engineering and a MSc degree in Physics from the University of Rome La Sapienza. He has collaborated as researcher and adjunct professor for many years in the Department of Electronic Engineering and for the Master in Sonic Arts at Tor Vergata University. Massimiliano also joined the Fondazione Ugo Bordoni of Rome, where he stayed several years as a researcher in the area of speech processing. His research interests focus on the areas of artificial intelligence, such as machine learning and pattern recognition, circuits and algorithms for signal analysis, processing, and synthesis, particularly with regard to audio signals, speech and music, biosignals and images.

Héctor Delgado received his Ph.D. degree in the area of speech processing from the Department of Telecommunications and Systems Engineering of the Autonomous University of Barcelona (UAB), Spain, in 2015. Before, he received a BS in Computer Science Engineering from the University of Seville, Spain, in 2008, and a MS in Multimedia Technologies from UAB, Spain, in 2009. From 2008 to 2015, he was with the Center for Ambient Intelligence and Accessibility of Catalonia (CAIAC). In 2013, he was with the Laboratoire Informatique dAvignon (LIA), at University of Avignon, France, as a visiting researcher. Currently, he is a postdoctoral researcher in the Speech and Audio Processing Research Group at EURECOM, France. His research interests include speaker recognition and diarization, speaker recognition anti-spoofing, audio segmentation and speech recognition.