ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements

Héctor Delgado, Massimiliano Todisco, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Junichi Yamagishi

ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements

Héctor Delgado[1], *Massimiliano Todisco*[1], *Md Sahidullah*[2], *Nicholas Evans*[1], *Tomi Kinnunen*[3], *Kong Aik Lee*[4], *Junichi Yamagishi*[5,6]

[1]Department of Digital Security, EURECOM, France
[2]MULTISPEECH, Inria, France
[3]School of Computing, University of Eastern Finland, Finland
[4]Data Science Research Laboratories, NEC Corporation, Japan
[5]Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan
[6]Centre of Speech Technology Research, University of Edinburgh, U.K.

{delgado,evans,todisco}@eurecom.fr, sahidullahmd@gmail.com,
tkinnu@cs.uef.fi, k-lee@ax.jp.nec.com, jyamagis@nii.ac.jp

Abstract

The now-acknowledged vulnerabilities of automatic speaker verification (ASV) technology to spoofing attacks have spawned interests to develop so-called spoofing countermeasures. By providing common databases, protocols and metrics for their assessment, the ASVspoof initiative was born to spearhead research in this area. The first competitive ASVspoof challenge held in 2015 focused on the assessment of countermeasures to protect ASV technology from voice conversion and speech synthesis spoofing attacks. The second challenge switched focus to the consideration of replay spoofing attacks and countermeasures. This paper describes Version 2.0 of the ASVspoof 2017 database which was released to correct data anomalies detected post-evaluation. The paper contains as-yet unpublished meta-data which describes recording and playback devices and acoustic environments. These support the analysis of replay detection performance and limits. Also described are new results for the official ASVspoof baseline system which is based upon a constant Q cesptral coefficient frontend and a Gaussian mixture model backend. Reported are enhancements to the baseline system in the form of log-energy coefficients and cepstral mean and variance normalisation in addition to an alternative i-vector backend. The best results correspond to a 48% relative reduction in equal error rate when compared to the original baseline system.

1. Introduction

Automatic speaker verification (ASV) [1, 2] has matured considerably over the last few decades. As an efficient, convenient, low-cost and reliable solution to person authentication, ASV technology is finding its way into a growing array of commercial products and services. The research community has also reacted to concerns regarding the vulnerability of ASV technology to spoofing [3], also know as presentation attacks [4], namely attempts by fraudsters to interfere with the normal operation of an ASV system using specially crafted speech signals. Vulnerabilities to spoofing are clearly unacceptable; if not addressed, they stand to dent confidence and jeopardize the commercial exploitation of ASV technology.

So-called spoofing countermeasures, also known as presentation attack detection (PAD) systems, are typically auxiliary systems which operating along side an ASV system in order to detect and deflect spoofing attacks. The Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) challenge series was born in order to spearhead their development. Two such challenges have been held to date. The first focused on developing countermeasures to defend against converted voice and synthetic speech spoofing attacks whereas the second focused on replay attacks. Both challenges were accompanied by a standard database, protocol and metric meaning that, for the first time, the performance of competing countermeasure solutions could be meaningfully compared. The databases for both challenges have been released into the public domain[1,2].

Subsequent to the release of the ASVspoof 2017 database, the organisers became aware of a number of data anomalies. These were patched in a second version of the database released in February 2018[3]. This paper describes Version 2 of the ASVspoof 2017 database in addition to a set of newly released meta-data, which describes the playback and recording devices in addition to the different acoustic environments used in its collection. The paper presents an in-depth analysis of replay detection performance using the new meta-data.

Also new to this paper are a number of enhancements to the original baseline system based upon a constant Q cepstral coefficient (CQCC) [5, 6] frontend and Gaussian mixture model backend. The paper compares results obtained for both Version 1.0 and Version 2.0 of the ASVspoof 2017 database and both the original and enhanced baseline systems. Enhancements comprise log-energy coefficients, cepstral mean and variance normalisation and an alternative backend i-vector classifier.

The paper is organised as follows. Section 2 describes the original ASVspoof 2017 database and the differences introduced in Version 2.0. A description of the replay data collection procedures and meta-data analysis is presented in Section 3. Baseline enhancements are presented in Section 4. Results and conclusions are presented in Sections 5 and 6 respectively.

[1]http://dx.doi.org/10.7488/ds/298
[2]http://dx.doi.org/10.7488/ds/2105
[3]http://dx.doi.org/10.7488/ds/2301

Table 1: Updated statistics of the ASVspoof 2017 2.0 corpus.

Subset	# Spk	# Replay sessions	# Replay Config	#Utterances	
				Bona fide	Replay
Training	10	6	3	1507	1507
Devel.	8	10	10	760	950
Eval.	24	161	57	1298	12008
Total	42	177	61	3565	14465

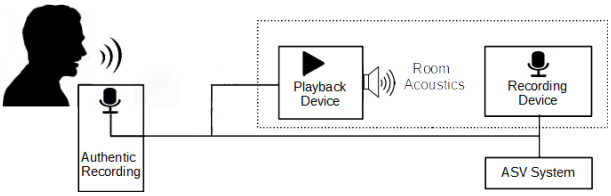

Figure 1: An illustration of the replay spoofing scenario. Figure adapted from [8].

2. ASVspoof 2017 database

The ASVspoof 2017 database was collected in order to foster the development of countermeasures to protect ASV systems from replay spoofing attacks. This section provides a brief overview of the data collection process and data partitions. These are the same as for the original Version 1.0 database that was used for the 2017 challenge. Also described here are changes introduced in Version 2.0.

2.1. Data collection

The ASVspoof 2017 corpus is a collection of *bona fide* and *spoofed* utterances. Bona fide utterances are a sub-set of the *RedDots* corpus[4] [7] collected and released previously in support of research in text-dependent ASV. The RedDots database was collected by volunteers using Android smart phones. Utterances correspond to one of the ten different, fixed pass-phrases (see [8] for more details). Spoofed utterances are the result of replaying and recording bona fide utterances using a variety of heterogeneous devices and acoustic environments. The later are intended to simulate replay spoofing attacks [9, 10]. A total of 57% of replay utterances were collected by four participants of the EU Horizon 2020-funded OCTAVE project[5] (see [8]) while the remaining 43% were collected by other contributors.

2.2. Data partitions

The ASVspoof 2017 corpus is partitioned into three subsets: **training**, **development** and **evaluation**. A summary of their composition is presented in Table 1. Data corresponds to 177 different replay sessions and 61 distinct replay configurations. The number of replay configurations which comprise the evaluation subset is considerably larger than that which comprise the training and development subsets.

Meta-data consisting of bona fide/spoofed (non-replay/replay) ground-truth labels, in addition to speaker IDs, phrase IDs, and replay configuration details were provided to ASVspoof 2017 challenge participants for the training and development subsets only. While corresponding details for the evaluation set was withheld originally, it has since been released publicly with Version 2.0 of the ASVspoof 2017 data. It supports the deeper analysis of countermeasure performance (e.g. per acoustic environment, playback device or recording device).

2.3. Database update

Post-evaluation, the organisers became aware[6] of a number of data anomalies that have potential to influence results and findings [11]. These mostly involve periods of silence, or zero-valued samples that are present in the original RedDots data [7]. While being small in number and while being of little consequence to the use of RedDots data for ASV research, these differences can impact on the assessment of replay detection performance. The same zero-valued samples are distorted to non-zero values as a result of playback, propagation in an acoustic environment and re-recording and such stark differences (zero/non-zero values) are easily detected. While the number of affected trials and duration of zero-valued samples was not substantial, these anomalies may yet bear influence on replay detection results especially, e.g., for approaches which exploit some form of temporal attention mechanism. Accordingly, these data anomalies have been corrected through the release of the ASVspoof 2017 Version 2.0 database. All results presented in this paper relate to this updated version.

3. Replay meta-data

This section introduces the meta-data that accompanies the release of ASVspoof 2017 Version 2.0. Described here is the general approach to replay simulation and the different playback, recording devices and acoustic environments used in database collection. Their different combinations lead to 61 distinct replay configurations.

3.1. Replay simulation

Practical replay spoofing attacks are assumed to be implemented according to the diagram illustrated in Fig. 1. Replay attacks comprise recordings of a bona fide access attempt that are subsequently replayed or represented to the microphone of the ASV system.

As described in [8, 12], the ASVspoof 2017 evaluation considered a worst case scenario where attackers have a digital copy of bona fide recordings, and specifically, a test utterance of the target speaker. These are then replayed and recaptured according to the scenario illustrated in Fig. 1. Accordingly, the acoustic impacts upon an utterance encompass those of a playback device, a recording device and an acoustic environment through which sound propagates. A combination of these is referred to a replay configuration (RC).

With the objective of measuring the *limits* of replay attack detection, the ASVspoof 2017 database was designed to contain a diverse range of replay configurations including conditions for which the detection of replay attacks should be relatively straightforward, to those for which detection should be considerably challenging.

[4] https://sites.google.com/site/thereddotsproject/

[5] https://www.octave-project.eu/

[6] We kindly acknowledge Ricardo P. V. Violato (CPqD lab, Brazil), Pavel Korshunov (IDIAP, Switzerland), and Bhusan Chettri (QMUL, UK), and their colleagues for bringing the problem to our attention.

Table 2: Acoustic environments used in the collection of the ASVspoof 2017 database. Replay recordings made in high noise conditions (noisier than bona fide speech) are assumed to present a low threat (green) to ASV; they should be detected by replay spoofing countermeasures with relative ease Recordings made in low noise conditions are assumed to pose a graver threat (red); they are like to be difficult to detect. Other recordings are made in medium noise conditions (yellow).

ID	Environment	ID	Environment
E01	Anechoic room	**E14**	Office 02
E02	Balcony 01	**E15**	Office 03
E03	Balcony 02	**E16**	Office 04
E04	Home 07	**E17**	Office 05
E05	Home 08	**E18**	Office 06
E06	Cantine	**E19**	Office 07
E07	Home 01	**E20**	Office 08
E08	Home 02	**E21**	Office 09
E09	Home 03	**E22**	Office 10
E10	Home 04	**E23**	Studio
E11	Home 05	**E24**	Analog wire 01
E12	Home 06	**E25**	Analog wire 02
E13	Office 01	**E26**	Analog wire 03

It is stressed that the categorisations of playback and recording devices represented below reflect qualitative indicators (e.g. based on device size) rather than scientific or quantitative measures of quality. They should be interpreted accordingly.

3.2. Acoustic environments

The acoustic environment is the physical space in which original speech data is replayed and re-recorded. The ASVspoof 2017 database was collected in a total of 26 different environments, each listed in Table 2 (Environment) and denoted E01-E26. Variations between them include types and levels of additive ambient and convolutive reverberation noise. The level of noise is assumed to be inversely correlated with the threat to ASV posed by replay attacks recorded in each environment. Colour codes in Table 2 serve as an indicator of the replay threat: low (green), medium (yellow) and high (red).

The two *balcony* and one *cantine* conditions are high ambient noise environments, e.g. traffic noise or background chatter, where corresponding replay recordings are assumed to pose little threat to ASV (high noise should be detected with relative ease). The eight different *home* conditions include recordings made in living rooms and bedrooms characterised by medium ambient noise levels. The ten *office* conditions also contain medium ambient noise, generally produced by air conditioning systems, but also exhibit reverberation.

There are four low noise conditions. *Anechoic room* recordings exhibit very low additive noise and reverberation whereas *studio* recordings contain similarly low levels of ambient noise, but also a degree of reverberation. *Analogue wire* conditions simulate recordings made without physical sound propagation, but with electrical propagation from a playback device directly to an ASV system. Replay recordings made in these condition are assumed to present the stiffest challenge to replay attack detectors.

Table 3: As for Table 2 except for playback devices. Recordings made with high quality playback devices are assumed to be the most difficult to detect (red). Those collected with lower quality devices are assumed to be detected with relative ease (green). Other devices are assumed to be of medium quality (yellow).

ID	Playback device
P01	All-in-one PC speakers
P02	Creative A60 speakers
P03	Genelec 8020C studio monitor
P04	Genelec 8020C studio monitor (2 speakers)
P05	Beyerdynamic DT 770 PRO headphones
P06	Dell laptop internal speakers
P07	Dynaudio BM5A speaker
P08	HP Laptop internal speakers
P09	VIFA M10MD-39-08 speaker
P10	ACER netbook internal speakers
P11	BQ Aquaris M5 smartphone
P12	Logitech low quality speakers
P13	Desktop PC line output
P14	Labtec LCS-1050 speakers
P15	Edirol MA-15D studio monitor
P16	Lenovo Ideatab S6000-H tablet
P17	Logitech S120 multimedia speakers
P18	MacBook pro internal speakers
P19	Altec lansing Orbit USB iML227 portable speaker
P20	Samsung GT-I9100 smartphone
P21	Samsung GT-P6200 tablet
P22	Behringer Truth B2030A studio monitor
P23	Focusrite Scarlett 2i2 audio interface line output
P24	Focusrite Scarlett 2i4 audio interface line output
P25	Genelec 6010A studio monitor
P26	AKG K242HD Headset

3.3. Playback devices

There are 26 playback devices, denoted P01-P26, as listed in Table 3. Replay samples collected from consumer-grade portable replay devices, e.g. smart phones, laptops and tablets equipped with in-built, generally small loudspeakers, are assumed to be of lower quality and should be detected with relative ease (green). Replay samples collected using consumer devices with larger loudspeakers, e.g. desktop PCs with external loudspeakers, are assumed to be of medium quality (yellow); they are assumed to introduce less acoustic distortion than smaller, in-built loudspeakers. Professional audio equipment such as active studio monitors and studio headphones are assumed to result in replay samples of comparatively high quality; they are assumed to pose the gravest threat (red). Also included among high quality devices are audio interfaces or analogue outputs used in the collection of recordings in *analogue wire* environments.

3.4. Recording devices

There are 25 recording devices. They are illustrated in Table 4, denoted R01-R25. Replay samples collected from mobile or battery powered devices with in-built, miniature microphones, e.g. smart phones and laptops are assumed to be of lower quality. Sample collected from such devices are assumed to be detected with ease (green). Samples collected from studio-quality condenser microphones or hand-held recorders are assumed to be of higher quality. Such recording are assumed to present a graver threat (red). Also included among high-quality devices are audio interfaces or analogue inputs used in the collection of recordings in *analogue wire* attacks. Powered desktop devices such as headsets or webcams, are assume to give recordings of medium quality (yellow).

Table 4: As for Tables 2 and 3 except for recording devices. Recordings made with high quality devices are assumed to be the most difficult to detect (red). Those collected with lower quality devices are assumed to be detected with relative ease (green). Other devices are assumed to be of medium quality (yellow).

ID	Recording device
R01	Zoom H6 handy recorder
R02	BQ Aquaris M5 smartphone
R03	Low-quality headset
R04	Nokia Lumia 635 smartphone
R05	Røde NT2 microphone
R06	Røde smartLav+ microphone
R07	Samsung Galaxy S7 smartphone
R08	Desktop PC microphone input
R09	Zoom H6 recorder with Behringer ECM8000 mic.
R10	Zoom H6 recorder with MSH-6 microphone
R11	Zoom H6 recorder. with XY microphone
R12	iPhone 5c smartphone
R13	iPhone 7 plus smartphone
R14	iPhone 4 smartphone
R15	Logitech C920 webcam
R16	miniDSP UMIK-1 microphone
R17	Samsung Galaxy Trend 2 smartphone
R18	Samsung GT-I9100 smartphone
R19	Samsung GT-P6200 tablet
R20	Samsung Trend 2 smartphone
R21	AKG C3000 microphone
R22	SE electronic 2200a microphone
R23	Focusrite Scarlett 2i2 interface line input
R24	Focusrite Scarlett 2i4 interface line input
R25	Zoom HD1 handy recorder

3.5. Replay configurations

The ASVspoof 2017 database contains recordings collected with diverse replay configurations (RCs). According to the schematic illustrated in Fig. 1, each RC comprises one playback device, one acoustic environment and one recording device. More formally, a **RC** is defined as a triplet $\text{RC}_c = (\text{E}_i, \text{P}_j, \text{R}_k)$ where c enumerates all unique triplets (i, j, k), i indexing environments, j playback devices and k recording devices. Only a subset (61) of the $26 \times 26 \times 25 = 16,900$ possible RC combinations are represented; not all devices/environments were available to all the crowd-sourcing data collectors.

In order to aid analysis, the number of distinct RCs reported previously [12] was reduced by the grouping together of overlapping configurations. Version 2 of the ASVspoof 2017 database contains 61 distinct RCs as indicated in Table 1.

As a precursor to further analysis, the set of replay recordings were ranked in terms of the *threat* they present to ASV. The reasoning for such analysis is to compare subsequently the correlation between the supposed threat and spoofing detection performance. Replay attacks which pose the least threat to ASV are assumed to be of poor quality (high noise or distortion) as compared to bona fide speech. It is a reasonable assumption that these replay attacks should also be detected by a spoofing countermeasure with relative ease. In contrast, high quality replay recordings are expected to pose a greater threat to ASV and also be more challenging to detect since they are potentially more similar to bona fide speech.

A standard Gaussian mixture model with universal background model (GMM-UBM) ASV system was used for ranking [13]. It uses a Mel-frequency cepstral coefficient (MFCC) front-end and a 512-component UBM trained using RSR2015 [14] and TIMIT[7] databases. Phrase-dependent target speaker models are created from RedDots enrolment data. Under a spoofing-free scenario, this system achieves an equal error rate (EER) of 1.8% on the evaluation set. The EER for each RC is computed using all replay segments recorded with the given RC and the corresponding bona fide counterparts.

Fig. 2 (blue solid profile) shows a sorted rank of the EER obtained for each of the 57 RCs corresponding to the evaluation set. Without exception, replay attacks collected in all RCs succeed in spoofing the ASV system to some extent; all EERs are above those for the spoofing-free scenario. Also evident from results presented in Fig. 2 is a substantial variation in EER across different RCs (from 2.0% to 48.0%).

Table 5 shows a list of the same 57 distinct RCs of the evaluation set sorted according to the EER ranking in Fig. 2. In addition to the number of evaluation segments corresponding to each RC, also illustrated in Table 5 is the number of distinct playback devices (P), acoustic environments (E) and recording devices (R) used for their collection. Colours reflect the same qualitative indicators of Tables 2-4. Higher quality replay attacks (higher-numbered RCs) that are assumed to present a greater challenge to replay spoofing countermeasures are indicated in red (high quality devices and benign acoustic environments). Lower quality attacks (lower quality devices and harsh acoustic environments) that are assumed to be detected with relative ease are indicated in green.

Noisy acoustic environments appear exclusively to the left in Table 5. With only one exception, RCs with high quality playback *and* recording devices appear to the right. As might be expected, analogue wire attacks (RCs 55-57) are among the most harmful to ASV, producing near random decisions (Fig. 2), and will certainly be among the most difficult to detect. Of course the general trend is more complex. The three components that comprise the RC do not have the same influence on ASV threat and are unlikely to have the same influence of detection difficulty, e.g. the quality of a playback or recording device is of little influence when the acoustic environment is harsh. As a consequence, strict and consistent trends are difficult to observe in practice.

4. Baseline countermeasure enhancements

New to the 2017 edition of the ASVspoof evaluation was the introduction of a baseline countermeasure. It is based upon a constant Q cepstral coefficient (CQCC) frontend [5, 6] and standard Gaussian mixture model (GMM) backend classifier. This section describes the original ASVspoof 2017 baseline configuration and enhancements which deliver improved spoofing detection performance for replay attacks, specifically log-energy coefficients and cepstral mean and variance normalisation, in addition to an alternative i-vector backend classifier.

4.1. CQCC features

CQCC features are derived using the constant Q transform (CQT) [15, 16], a perceptually motivated time-frequency analysis tool and alternative to the short-term Fourier transform (STFT). Whereas the STFT operates with a fixed spectro-temporal resolution, that of the CQT is variable, with a higher frequency resolution at lower frequencies and a higher temporal resolution at higher frequencies. Just like conventional Mel-frequency cepstral coefficients, CQCC extraction is performed

[7] https://catalog.ldc.upenn.edu/ldc93s1

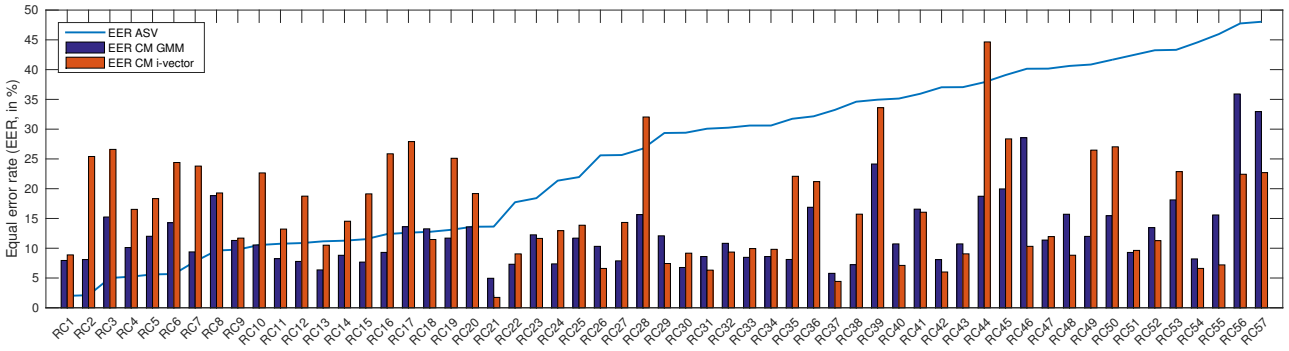

Figure 2: Impact of different replay configurations (RCs) upon ASV performance (solid blue profile) measured in terms of equal error rate (%) when a set of zero-effort impostor trials are replaced with replay spoofing trials generated with each RC. Also illustrated is the detection performance of GMM (blue bars) and i-vector (red bars) spoofing countermeasures (CMs) for the same RCs, also expressed in terms of equal error rate (%).

Table 5: A list of replay configurations sorted according to the ASV EER ranking in Figure 2. Colour codes reflect the supposed threat to ASV (Tables 2 to 4). E denotes acoustic environment, P denotes playback device, R denotes recording device. Numbers reflect distinct replay sessions and number of segments.

ID	E	P	R	#seg.	ID	E	P	R	#seg.	ID	E	P	R	#seg.
RC01	18	18	12	55	RC20	7	21	14	275	RC39	22	3	4	183
RC02	8	20	19	67	RC21	18	11	12	22	RC40	16	7	6	116
RC03	8	21	25	122	RC22	12	16	11	183	RC41	13	14	9	182
RC04	8	20	25	102	RC23	20	10	15	1138	RC42	11	26	16	153
RC05	8	20	14	114	RC24	17	12	17	179	RC43	6	9	6	84
RC06	8	21	14	108	RC25	9	18	12	42	RC44	13	14	10	179
RC07	8	21	18	120	RC26	17	12	10	184	RC45	10	25	16	346
RC08	2	21	25	98	RC27	6	9	7	96	RC46	18	22	21	181
RC09	7	20	25	244	RC28	21	3	1	240	RC47	23	15	13	342
RC10	2	20	25	82	RC29	18	5	3	454	RC48	19	22	22	1200
RC11	7	20	19	272	RC30	15	19	20	74	RC49	14	3	17	180
RC12	2	20	19	75	RC31	16	7	7	145	RC50	1	15	13	748
RC13	3	8	20	113	RC32	17	12	9	180	RC51	16	7	5	169
RC14	7	21	18	279	RC33	13	14	4	183	RC52	10	26	16	181
RC15	2	21	18	150	RC34	17	12	4	181	RC53	14	3	4	181
RC16	2	21	14	116	RC35	13	14	17	178	RC54	6	9	5	105
RC17	7	21	25	266	RC36	22	4	17	181	RC55	26	24	24	178
RC18	7	20	14	265	RC37	15	19	4	48	RC56	25	13	8	182
RC19	2	20	14	120	RC38	12	17	11	184	RC57	24	23	23	183

with a filterbank, where the Q factor is a measure of the selectivity of each filter, defined as the ratio between the centre frequency of the filter and its bandwidth.

Cepstral analysis cannot be applied directly to the CQT since frequency bins are on a different scale to those of the basis functions of the discrete cosine transform; they are respectively geometrically and linearly spaced. This problem is solved by converting geometric space to linear space via re-sampling [6] before otherwise conventional cepstral analysis is applied.

Designed initially for the detection of voice conversion and speech synthesis spoofing attacks, the CQCC frontend was applied to the detection of replay spoofing attacks with only minor modifications.The CQT is applied with a maximum frequency of $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 8kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \simeq$ 15Hz (with 9 being the number of octaves). The number of bins per octave B is set to 96. Re-sampling is applied with a sampling period of 16 bins in the first octave. Resulting feature vectors are of dimension 19, excluding the C_0 coefficient (cf. 29 coefficients + C_0 for the original system). Full details of the CQCC extraction procedure are reported in [5, 6].

4.2. Log-energy coefficients

While frame-level energy coefficients are used extensively in a multitude of different speech-related tasks [17, 18, 19], they have not been applied extensively to spoofing detection problem in the context of the ASVspoof challenge. Even if the absolute energy of both bona fide and replayed speech signals may reflect more the gain of a recording device, rather than acting as an strict indicator of spoofing, it is hypothesised here that non-linear changes to energy *dynamics* may serve as an indication of replay. Accordingly, the use of log-energy parameters for spoofing detection has been explored.

Although the log-energy can be calculated from the time-domain signal, log-energy coefficients are here calculated in the CQT domain in order to preserve the time resolution of CQCC coefficients. Parseval's theorem states that the total energy computed in the time domain is equal to the total energy computed in the frequency domain. In the frequency domain, the log-energy is defined as the logarithm of the summation of the spectral components, normalised by the sample size. It can be calculated according to:

$$\log E(n) = \log \sum_{k=1}^{K} \left| X^{CQ}(k,n) \right|^2 - \log(K) \qquad (1)$$

where n is the frame index, $k = 1, 2, ..., K$ is the frequency bin index and where $X^{CQ}(k, n)$ is the frame-blocked CQT [5, 6].

4.3. Cepstral mean and variance normalisation

Cepstral mean and variance normalization (CMVN) [20] is an efficient normalisation technique used to remove nuisance channel effects which might otherwise degrade the performance of a variety of different approaches to automatic speech and speaker recognition. Other researchers have reported the benefit of using CMVN for spoofing detection, e.g. [21, 22, 23, 24], all of which relate to replay detection within the scope of ASVspoof 2017.

The application of CMVN to replay spoofing detection may at first seem counter-intuitive. The playback and recording of speech in different acoustic environments using different devices is akin to the accumulation of additional channel effects. CMVN, which aims to attenuate channel effects, may then be to the detriment of replay detection. This assumption may only hold, however, if bona fide speech were to be captured across a common, consistent channel. This is not the case for the ASVspoof 2017 source data, namely the RedDots database, which was captured using heterogeneous devices and channels [7].

Accordingly, CMVN may help to align both bona fide and replayed speech distributions to a common scale, and hence force spoofing detection to discriminate between the two according to influences other than those caused by channel differences (which are present in both). The use of CMVN is also expected to improve the reliability of spoofing detection across the diverse variation in replay attacks which characterises the ASVspoof 2017 database.

4.4. Backend classifier

The original baseline system uses a GMM backend classifier. We have also investigated the performance of an alternative i-vector approach [18]. Both are described below.

The GMM backend is the same as the original baseline. It uses models of 512 components. Different models are learned for bona fide and spoofed speech with an expectation-maximisation (EM) algorithm with random initialisation. Classifier scores for a given test utterance are computed as the log-likelihood ratio $\Lambda(X) = \log L(X|\theta_n) - \log L(X|\theta_s)$, where X is a sequence of CQCC feature vectors, L denotes the likelihood function, and θ_n and θ_s represent the GMMs for bona fide and spoofed speech, respectively.

The alternative i-vector backend [18] uses a universal background model of 64 components which is trained on the ASVspoof 2017 training partition. The total variability space matrix T has 100 factors and is learned using the same data. The i-vectors are mean normalised, whitened and treated with within-class covariance normalization (WCCN). The i-vector training data are averaged before length-normalization. Single i-vectors are learned for bona fide and spoofed speech classes. The proposed i-vector backend also uses the same CQCC frontend as the GMM-based classifier.

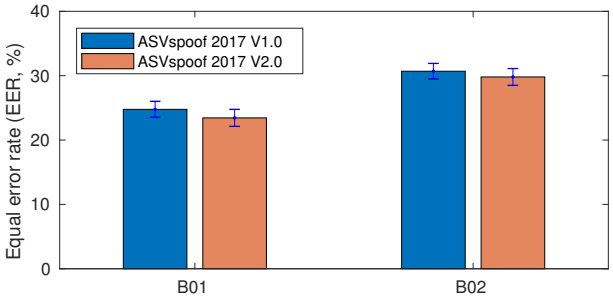

Figure 3: Comparison of replay detection performance in terms of equal error rate (with 95% confidence intervals) for the original baseline system for the evaluation set for versions 1.0 and 2.0 of the ASVspoof 2017 database.

5. Results and analysis

Several sets of experimental results are reported. The first shows differences in performance for original baseline systems for versions 1.0 and 2.0 of the ASVspoof 2017 database. The second set of experiments assess improvements to performance delivered by baseline enhancements. Last, we analyse replay spoofing detection nuances according to different replay configurations.

5.1. ASVspoof 2017 Version 2.0

Fig. 3 illustrates replay detection performance in terms of EER for versions 1.0 and 2.0 of the ASVspoof 2017 database and for original baseline systems B01 and B02 [12]. Also included are 95% confidence intervals calculated using the methodology explained in [25], and using the specific configuration described in [26]. Differences between systems B01 and B02 lie only in the use of training (T) or pooled training and development data (T+D); the systems are otherwise the same. Differences in performance are minor: 25% to 23% EER for B01 and 31% to 30% for B02. Confidence intervals are highly overlapped, suggesting that the variations in performance are not significant. While the same may not hold for alternative approaches to spoofing detection, these results show that in the case of the particular baseline system, performance is reasonably stable across the two database versions.

5.2. Baseline enhancements

Presented here are experiments which assess the benefit of baseline enhancements presented in Section 4, namely the use of log-energy, CMVN and the alternative i-vector backend. Results are presented in Table 6 for different training and testing configurations involving some combination of the default training (T), development (D) and evaluation (E) partitions. Such exhaustive experiments were performed in order to assess the stability of findings across different training and testing conditions.

Results without CMVN appear to the left of Table 6. Those with CMVN appear to the right. Results obtained with 19th order CQCCs and static, delta and acceleration coefficients are denoted by 19-SDA. Those obtained with appended log-energy are denoted by 19E-SDA. Without CMVN normalisation, the use of log-energy decreases performance. Also, the i-vector backend outperforms the GMM backend.

With CMVN, however, the picture is quite different. Results to the right of Table 6 show that the use of log-energy leads to better performance. There is a substantial improvement

Table 6: Replay detection performance in terms of EER for the ASVspoof 2017 Version 2.0 database, several training and testing configurations, different frontend with and without log-energy coefficients and CMVN normalisation and for GMM and i-vector backends.

	training on	T	D	T	D	T+D	T	D	T	D	T+D
	testing on	D	T	E			D	T	E		
	Feat config.	no normalisation					CMVN				
GMM	19-SDA	11.69	1.36	30.79	25.33	23.97	13.31	8.49	19.74	16.89	15.33
	19E-SDA	10.37	1.37	34.95	26.3	29.31	**9.06**	**5.64**	**13.74**	**14.77**	**12.24**
i-vector	19-SDA	**4.43**	**1.23**	**17.82**	**18.81**	**18.60**	11.61	8.74	16.61	15.08	15.63
	19E-SDA	5.11	1.54	21.47	16.25	21.10	10.52	7.27	14.76	14.37	12.93

Table 7: Replay detection performance in terms of equal error rate (EER, %) for different qualities of environments and playback and recording devices.

	Low	**Medium**	**High**
E: acoustic environment	16.68	18.73	21.86
P: playback device	16.64	16.44	18.37
R: recording device	10.80	15.69	17.77

in terms of EER for both GMM and i-vector systems. These observations support the hypotheses outlined in Sections 4.2 and 4.3. CMVN normalises the log-energy distributions to have zero-mean and unit-variance, but retains utterance-level energy dynamics which serve as an indicator of spoofing. With the application of CMVN, the best performance is achieved using 19E-SDA CQCC features and a GMM backend. Compared to the previous baseline score of 24.0% the best enhanced baseline result of 12.2% corresponds to a relative improvement of almost 50%.

Also of interest here is the stability of results and trends across the different training and testing configurations. Whereas results without CMVN show substantial variation for different training configurations, each containing different numbers of RCs, those with CMVN are more stable. These findings appear to confirm the hypothesis that CMVN helps to improve generalisation in spoofing detection.

5.3. Meta-data analysis

Fig. 2 also shows an analysis of performance per RC for the GMM and i-vector backend classifiers. GMM system uses the CQCC 19E-SDA configuration with CMVN, while the i-vector system uses the CQCC 19-SDA configuration. The first interpretation of these results is that there is little correlation between the supposed difficulty of detection, as indicated by the EER profile described in Section 3.5, and replay detection performance. This finding reflects the combined, complex effects of the playback device, the acoustic environment and the recording device. As for CM performance, the GMM system outperforms the i-vector system for most RCs.

Table 7 illustrates a decomposition of results in terms of the qualitative indicators displayed in Tables 2 to 4. Results show the impact of a single element of the RC in terms of EER for all bona fide trials versus all pooled replay trials corresponding to the given qualitative category. While not strictly true, variation from other elements of the RC are at least somewhat marginalised. These results show more consistent trends. The impact of the acoustic environment dominates the effect of the playback and recording devices; the trend for playback devices are inconsistent and show little variation, whereas EERs for recording device variation are universally lower but show a consistent trend.

The observations would suggest that, perhaps unsurprisingly, ambient and reverberation noise are reliable indicators of replay spoofing attacks. Replay recordings made in benign acoustic conditions are more difficult to detect; the highest EER in Table 7 is for high quality acoustic environments. They also suggest that the quality of the recording device plays a more significant role in the difficulty of replay detection than the quality of the playback device. These observations would seem to corroborate the trends shown in Table 7: benign acoustic conditions and high quality recording devices are positioned largely to the right (red colours).

Needless today, however, coupled with the results illustrated in Fig. 2, the physical effects upon a speech signal of playback and recording are complex. An understanding of how the variation in replay configurations impacts on different detection systems, and hence exactly what information these systems are using, remains an issue requiring further investigation.

6. Conclusions

This paper describes the differences between Versions 1.0 and 2.0 of the ASVspoof 2017 database of bona fide and replay spoofing attack recordings. New to this paper is (i) a set of meta-data which describes the diversity of replay configurations used for database collection and (ii) enhancements to the official baseline countermeasure. The latter include the addition of log-energy coefficients and cepstral mean and variance normalisation (CMVN). Despite being somewhat counter-intuitive in terms of replay detection, these enhancements bring a relative reduction to the spoofing detection equal error rate of almost 50%.

Meta-data analysis confirms that the effects of replay spoofing are complex and difficult to interpret. Results suggest that the effect of the acoustic environment is the most influential upon the performance of replay spoofing detection and show that that of the playback device is less important. Different approaches to replay spoofing detection also show substantial variation in performance across different replay configurations, suggesting that they are using different cues to detect replay.

These findings show that the problem of replay spoofing detection remains poorly understood and might suggest that future studies should consider evaluation under controlled conditions. These may allow the influence of each component in a replay attack to be isolated and studied more reliably. Findings from such studies may then help to design more reliable replay spoofing countermeasures. The findings from the study reported in this paper will feed into the roadmap for future ASVspoof evaluations.

7. 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, January 2010.

[2] J.H.L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Proc. Mag.*, vol. 32, no. 6, pp. 74–99, 2015.

[3] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. INTERSPEECH*, Lyon, France, 2013.

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

[5] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Odysssey*, Bilbao, Spain, 2016, pp. 283–290.

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

[7] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, Md. J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proc. INTERSPEECH*, 2015, pp. 2996–3000.

[8] T. Kinnunen, Md Sahidullah, M. Falcone, L. Costantini, R. González Hautamäki, D. Thomsen, A. Sarkar, Z.H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamäki, and K. A. Lee, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. ICASSP*, New Orleans, USA, 2017.

[9] 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. APSIPA*, 2014, pp. 1–5.

[10] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. 13th International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2014, pp. 157–168.

[11] B. Chettri and B. L. Sturm, "A deeper look at Gaussian mixture model based anti-spoofing systems," in *Proc. ICASSP*, 2018.

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

[13] H. Delgado, M. Todisco, M. Sahidullah, A.K. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan, "Further optimisations of constant Q cepstral processing for integrated utterance verification and text-dependent speaker verification," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 179–185.

[14] A. Larcher, K.A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Comm.*, vol. 60, pp. 56–77, 2014.

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

[16] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Semantic Audio*. 2014, Audio Engineering Society.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., "The htk book," *Cambridge university engineering department*, vol. 3, pp. 175, 2002.

[18] N. Dehak, P.J. 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.

[19] R. Saeidi with $>$ 30 coauthors, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. INTERSPEECH*, 2013.

[20] R Auckenthaler, M Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digit. Signal Process.*, vol. 10, no. 1, pp. 42–54, Jan. 2000.

[21] R. Font, J. M. Espn, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge," in *Proc. INTERSPEECH,*, 2017, pp. 7–11.

[22] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. INTERSPEECH*, 2017, pp. 82–86.

[23] Z. Ji, Z.Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in asvspoof2017," in *Proc. INTERSPEECH*, 2017, pp. 87–91.

[24] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using dnn for channel discrimination," in *Proc. INTERSPEECH*, 2017, pp. 97–101.

[25] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. Odyssey*, 2004, number EPFL-CONF-83049.

[26] A. Sholokhov, M. Sahidullah, and T. Kinnunen, "Semi-supervised speech activity detection with an application to automatic speaker verification," *Computer Speech & Language*, vol. 47, pp. 132 – 156, 2018.