The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection

Tomi Kinnunen¹, Md Sahidullah¹, Héctor Delgado², Massimiliano Todisco², Nicholas Evans², Junichi Yamagishi³,⁴, Kong Aik Lee⁵

¹University of Eastern Finland, Finland – ²EURECOM, France – ³National Institute of Informatics, Japan – ⁴University of Edinburgh, UK – ⁵Institute for Infocomm Research, Singapore

Abstract

The ASVspoof initiative was created to promote the development of countermeasures which aim to protect automatic speaker verification (ASV) from spoofing attacks. The first community-led, common evaluation held in 2015 focused on countermeasures for speech synthesis and voice conversion spoofing attacks. Arguably, however, it is replay attacks which pose the greatest threat. Such attacks involve the replay of recordings collected from enrolled speakers in order to provoke false alarms and can be mounted with greater ease using everyday consumer devices. ASVspoof 2017, the second in the series, hence focused on the development of replay attack countermeasures. This paper describes the database, protocols and initial findings. The evaluation entailed highly heterogeneous acoustic recording and replay conditions which increased the equal error rate (EER) of a baseline ASV system from 1.76% to 31.46%. Submissions were received from 49 research teams, 20 of which improved upon a baseline replay spoofing detector EER of 24.77%, in terms of replay/non-replay discrimination. While largely successful, the evaluation indicates that the quest for countermeasures which are resilient in the face of variable replay attacks remains very much alive.

Index Terms: automatic speaker verification, spoofing, countermeasure, replay attacks, ASVspoof

1. Introduction

Automatic speaker verification (ASV) [1, 2, 3] technology is used in a growing range of applications which require not only robustness to changes in the acoustic environment, but also resilience to intentional circumvention, known as spoofing [4] or, according to ISO/IEC 30107-1:2016 standard, presentation attacks. Among other possible attack vectors, replay attacks are a key concern; they can be performed with ease and the threat they pose to ASV reliability has been confirmed in independent studies [5, 6, 7]. Replay attacks are mounted using recordings of a target speaker’s voice which are replayed to an ASV system in the place of genuine speech. An example is the use of a smart-device to replay a recording of a target speaker’s voice to unlock a smartphone which uses ASV access control.

Spoofing countermeasures have consequently been developed to protect ASV systems from replay attacks. The literature shows three general strategies. Prompted-phrase ASV, e.g. randomised digit sequences [8], and utterance verification [9, 10] offer some protection, although multiple recordings can be remixed to produce a replay attack which matches the prompted phrase. Copy detection [11, 12], also known as audio fingerprinting, can also be used to detect recordings of genuine enrollment utterances or previous access attempts, although this approach calls for the maintenance of a dynamically growing database. This paper concerns a third strategy which aims to detect replay attacks using only the acoustic characteristics of a given utterance. Arguably, this solution has broader utility; this includes any ASV approach/system in addition to any form of replay attack.

The detection of replay attacks using acoustic characterisation is potentially problematic, however. The difficulty relates to the unpredictable variation in the quality of a replay attack. Recordings, perhaps collected surreptitiously, may contain significant additive or convolutional noise. The detection of replay attacks may then boil down to an ambient or channel noise classification problem. In contrast, recordings made with high-quality hardware in benign acoustic environments may be close to indistinguishable from genuine speech signals. At the limit, bit-to-bit digital copies of genuine audio recording, perhaps injected into the input circuitry of the ASV system bypassing the microphone, would be indistinguishable using any method. The question then is, what are the practical limits of replay attack detection?

The search for an answer to this fundamental question is the focus of the ASVspoof 2017 challenge². ASVspoof 2017 follows two special sessions on spoofing and countermeasures for automatic speaker verification at INTERSPEECH 2013 [13] and 2015 [14] which formed the first evaluation, ASVspoof 2015 [15]. The first evaluation promoted the development of generalised countermeasures capable of protecting ASV from diverse text-to-speech (TTS) and voice conversion (VC) spoofing attacks [16]. While the mounting of these attacks may require substantial expertise, replay attacks can be mounted by the layperson using widely available consumer devices for audio recording and replaying. ASVspoof 2017 therefore promoted the development of replay attack countermeasures.

Previous attempts to assess the threat of replay spoofing attacks typically involved a modest number of evaluation conditions, e.g. [5, 6, 7, 17]. Some studies, e.g. [6, 7] report close-to-perfect recognition accuracy, albeit in the case of relatively homogeneous acoustic conditions. Other work [5] suggests that performance may degrade in more practical scenarios where the acoustic conditions can vary greatly. The primary technical goals of ASVspoof 2017 are therefore (i) to assess the practical limitations of replay attack detection and (ii) to promote the development of countermeasures with potential to detect replay attacks.

²http://www.asvspoof.org/
spoofing attacks ‘in the wild’, namely in highly-varying acoustic conditions.

In identical fashion to the 2015 edition, ASVspoof 2017 focuses on standalone spoofing attack detection (here, replay attacks), i.e. spoofing detection in isolation from ASV. However, so that the initiative is at least aligned to ASV research and in contrast to the first edition, ASVspoof 2017 uses the recent text-dependent RedDots [18] data as the base corpus [19]. One additional change to the 2015 edition, made in order to encourage wider participation, is the provision of a baseline spoofing classifier [20]. This strategy appears to have had a positive impact; the organisers received 113 requests for access to the development set, while a total of 49 primary system scores were submitted for the evaluation set.

2. ASVspoof 2017 corpus

The ASVspoof 2017 corpus originates from the RedDots corpus\(^3\) which was collected by volunteers from across the globe (mostly ASV researchers) using Android smartphones. Non-replayed utterances are a subset of the original RedDots recordings whereas replayed recordings are replayed and recaptured versions. The replayed utterances hence correspond to a ‘stolen voice’ scenario where the attacker has access to a digital copy of an original target speaker utterance which is then replayed through transducers of varying quality.

A total of 57% of replay files were collected by four participants of the EU Horizon 2020-funded OCTAVE project\(^4\), (see [19]), while the remaining 43% were collected by other contributors. Replay recordings were collected from the re-playing and re-recording of concatenated RedDots utterances with heterogeneous devices and acoustic environments. Non-replayed evaluation data was supplemented with utterances collected from 7 new speakers.

The ASVspoof 2017 corpus is partitioned into three subsets: training, development and evaluation. Details of each are presented in Table 1. The first two subsets were provided to participants for the design of replay detectors (countermeasures), while re-partitioning of the training and development subsets was permitted. Metadata consisting of replay/non-replay ground-truth labels, in addition to speaker ID, phrase ID, and replay configuration details were provided for the training and development subsets. Only audio data and phrase ID were provided for the evaluation set for which participants were required to submit scores. Results were then determined by the organisers and returned to participants.

All three subsets are disjoint in terms of speakers. They are also somewhat disjoint in terms of data collection sites. The training subset was collected at a single site. The development subset was collected at the same site in addition to two more sites. Finally, the evaluation subset was collected at the same three sites and supplemented with additional data from two new sites. Nonetheless, even data from the same site was collected by different people using different recording and replaying devices and in different acoustic environments. The evaluation subset contains data collected from 161 replay sessions in 110 unique replay configurations\(^5\). Data heterogeneity has proven essential to the development of reliable spoofing countermeasures [21, 22, 23].

2.1. Evaluation conditions

The ASVspoof 2017 corpus comprises six evaluation conditions containing a disjoint set of replay trials (and a shared set of non-replay trials). Since the original RedDots source data and the ASVspoof 2017 replay recordings were collected in diverse conditions, the data exhibits multiple, concurrent variations (e.g. recording device quality, room dimensions, reverberation in addition to the vocal effort in the original recordings). The isolation or marginalisation of such variation is particularly challenging. Hence, the six conditions for the ASVspoof 2017 challenge were defined post-evaluation using a clustering of well-ranked system scores.

To focus on differences in replay configurations rather than the differences relating to individual utterances, clustering was applied to scores averaged across individual replay environments. The clustering process was performed as follows:

1. System scores for all submissions which out-performed the baseline were linearly fused using the Bosaristoolkit\(^6\) to obtain a high-performance ensemble classifier.
2. Fused scores were then averaged across all replay trials corresponding to the same replay session (common replay environment, playback and recording devices).
3. Averaged scores were then clustered using k-means to obtain a non-uniform partitioning of the score axis.
4. Resulting score clusters were then reordered according to increasing average fused score.

This procedure can be applied to cluster results into a number of different replay conditions. Those with a lower average fused score represent replay conditions which are generally easier to detect than replay conditions with a higher average fused score. A clustering into 6 different replay conditions was found empirically to give the most consistent and intuitive results. Condition C1 represents replay trials with significant background noise or channel distortion which are typically detected with ease. Condition C6 represents high-quality replay trials which are comparatively more difficult to detect. A quantitative analysis of the conditions in terms of signal quality measures and error rates is presented in Section 4.2.

2.2. Evaluation metrics

In line with the ASVspoof 2015 challenge, the 2017 edition concentrates on stand-alone spoofing detection without ASV integration. The task requires the assignment to a set of audio files a score which reflects the relative strength of two competing hypotheses, namely that the trial is non-replayed (genuine) or replayed (spoof).

Table 1: Statistics of the ASVspoof 2017 corpus.

<table>
<thead>
<tr>
<th>Subset</th>
<th># Spk</th>
<th># Replay sessions</th>
<th># Replayed Config</th>
<th># Utterances Non-replay</th>
<th># Utterances Replay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>1508</td>
<td>1508</td>
</tr>
<tr>
<td>Devel.</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>760</td>
<td>950</td>
</tr>
<tr>
<td>Eval.</td>
<td>24</td>
<td>161</td>
<td>110</td>
<td>1298</td>
<td>12008</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>177</td>
<td>123</td>
<td>3566</td>
<td>14466</td>
</tr>
</tbody>
</table>

\(^3\)https://sites.google.com/site/thereddotsproject/
\(^4\)https://www.octave-project.eu/
replayed (spoofed) speech. Higher scores are assumed to favor the non-replay/genuine hypothesis. The primary metric is the equal error rate (EER). Let $P_{fa}(\theta)$ and $P_{miss}(\theta)$ be the false alarm and miss rates at threshold $\theta$ defined according to:

$$P_{fa}(\theta) = \frac{\#\{\text{replay trials with score} > \theta\}}{\#\{\text{Total replay trials}\}}$$

$$P_{miss}(\theta) = \frac{\#\{\text{non-replay trials with score} \leq \theta\}}{\#\{\text{Non-replay trials}\}},$$

so that $P_{fa}(\theta)$ and $P_{miss}(\theta)$ are, respectively, monotonically decreasing and increasing functions of $\theta$. The EER corresponds to the threshold $\theta_{EER}$ at which the two detection error rates are (approximately) equal. It is estimated using the convex hull method available in the Bosaris toolkit. In contrast to the ASVspoof 2015 challenge, the EER is computed from scores pooled across all the trial segments instead of condition averaging. The rationale is to promote the development of replay attack detectors yielding scores that are more consistent across variable spoofing conditions; see also [24, Table 12] and [15, Fig. 6].

### 3. Impact of replay to ASV accuracy

The vulnerability of ASV systems to replay spoofing attacks has been confirmed previously by independent teams [6, 7, 25]. This section reports the impact of ASVspoof 2017 replay spoofing attacks on a classical Gaussian mixture model with universal background model (GMM-UBM) [26] ASV system. This has been shown [27] to deliver competitive performance for RedDots enrollment data. The ASV system uses Mel-frequency cepstral coefficient (MFCC) features and a 512-component UBM trained using RSR2015 [28] and TIMIT$^7$ databases. Phrase-dependent target speaker models are created from RedDots enrollment data. The evaluation protocol involves a number of genuine trials and then either zero-effort impostor or replay spoofing attack trials. The number of each are shown in Table 2. Table 3 shows the degradation in ASV performance when zero-effort impostors are replaced with replay spoofing attacks. The baseline EER for speaker discrimination is seen to increase substantially and illustrates the need to develop replay attack countermeasures.

### 4. ASVspoof 2017 challenge results

#### 4.1. Overview

A total of 49 submissions were received. A summary of results for primary systems is illustrated in Table 4, and in the DET plot of Fig. 1, along with the two baseline replay/non-replay detectors$^8$. They use a common Gaussian mixture model (GMM) back-end classifier with constant Q cepstral coefficient (CQCC) features [20] which are based on a perceptually motivated time-frequency transform [29]. Performance is shown for two baseline variants trained using either combined training and development data (B01) or training data alone (B02). The use of pooled data naturally results in better performance. There is substantial variation in EERs with 20 of the 49 submissions achieving better performance than the B01 baseline. This observation indicates the difficulty of the challenge and stresses the importance of avoiding over-fitting. The top-performing S01 system achieves a encouraging EER of 6.73 %. The DET plot in Fig. 1 further indicates that the systems are diverse across all the operating points not limited to the EER region.

#### 4.2. Condition analysis

Table 5 characterizes the six evaluation conditions derived from submission scores using the clustering procedure described in

---

$^7$https://catalog.ldc.upenn.edu/ldc93s1

$^8$http://www.asvspoof.org/data2017/baseline_CMs.zip
Sub-section 2.1. Illustrated are the mean and standard deviation of two standard quality measures. The signal-to-noise ratio (SNR) reflects the level of background noise and is estimated using the NIST STNR tool, while the cepstral distance (CSD) is a two-sided estimate of the distortion between replay utterances and corresponding source recordings [30]. It corresponds to the average Euclidean distance between the two recordings and is estimated from sliding frames of 20ms duration with 10ms overlap and standard cepstral analysis without the DC coefficient $c_0$. Low CSD values characterise high-quality replay attacks, i.e. little distortion.

Table 5 shows no consistent correlation between increasing SNR and difficulty (which increases from C1 to C6). For instance, condition C1 was found to exhibit low background noise but substantial spectral distortion stemming from the use of low quality replay (a netbook) and recording (webcam microphone) devices. Table 5 further indicates that the difficulty of each condition is not entirely correlated with CSD across categories C1-C3, while it is consistent across C4-C6. Replay configurations which introduce greater distortion are in general easier to detect. This is entirely intuitive given that additional noise, reverberation or other distortion induced by low-quality playback or recording devices will inherently distort spectral characteristics.

The last two rows of Table 5 illustrate the general quality of the playback/recording devices which characterise each condition. As expected, replay attacks mounted with low (L) and medium (M) quality devices are more easily detected that those mounted with high (H) quality devices.

4.3. System analysis

Fig. 2 illustrates, independently for each of the 6 conditions and pooled scores, the variation in performance for the top-10 ranked submissions, the B01 baseline system and the best performing S01 submission. Replay detection performance for category C6 is consistently the worst. Performance for conditions C1 and C3 shows a wide variation among systems. Category C2 was the best solved one by all top-10 systems. Conditions C3 to C5 also show a range of performance variation across systems. In general, such variation across conditions and systems’ performance highlights the challenging task of defining the evaluation conditions. System S01 is the best performing for only 1 of the 6 conditions and illustrates consistent and substantial improvements on the baseline.

5. Conclusions

The ASVspoof 2017 challenge was highly successful with more than 100 development data requests and nearly 50 challenge submissions. The second edition of the challenge is new in several respects. Besides new data for what is likely to be the most prolific form of spoofing attack in practice, namely replay, speech signals are collected ‘in the wild’ in a large number of heterogeneous recording conditions. Compared to the first challenge in 2015, the focus is now aligned to a text-dependent ASV scenario where short pass-phrases are used for speaker authentication. This paper summarises the challenge corpus, task, preliminary evaluation results, and categorization of the evaluation data for further analysis.

The average EER of all primary submissions is 25.91% whereas the best single system result shows an average detection EER of 6.73%. The comparison of these results to those from the previous challenge shows that the detection of replay attacks is seemingly more difficult than the detection of speech synthesis and voice conversion spoofing attacks. Countermeasure generalisation also remains an open problem.

Looking to the future, the categorisation of trials according to observed difficulty needs further investigation. To this end, the organisers expect that the challenge data, protocols, keys and evaluation results will be of use to the community in advancing further the state of the art in anti-spoofing in addition to helping ASV researchers to explore new and alternative approaches to protect speaker authentication systems from fraud.

6. Acknowledgements

The authors would like to express their gratitude to all ASVspoof 2017 challenge participants. 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. The work was also funded in part by the Academy of Finland (grant #288558) and by MEXT KAKENHI (Grant Numbers 26280066, 15H01686, 16H06302).
7. References


