ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan*

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

[1]University of Eastern Finland, Finland
[2]Eurecom, France
[3]University of Edinburgh, UK
[4]Institute for Infocomm Research, Singapore

http://www.asvspoof.org/index2017.html
E-mail: info@asvspoof.org

September 19, 2018

1 Introduction

The ASVspoof 2017 challenge follows on from two special sessions on spoofing and countermeasures for automatic speaker verification held during INTERSPEECH 2013 [2] and 2015 [3]. While the first edition in 2013 was targeted mainly at increasing awareness of the spoofing problem, the 2015 edition included a first challenge on the topic, with commonly defined evaluation data, metrics and protocols. The task in ASVspoof 2015 was to discriminate genuine human speech from speech produced using text-to-speech (TTS) and voice conversion (VC) attacks. The challenge was drawn upon state-of-the-art TTS and VC attacks data prepared for the "SAS" corpus [4] by TTS and VC researchers.

ASVspoof 2015 has a number of shortcomings, some which were discussed already in the INTERSPEECH 2015 special session. Firstly, in terms of attacks, ASVspoof 2015 focused exclusively on TTS and VC, excluding **replay attacks**. In contrast to state-of-the-art TTS and VC which requires substantial dedication, replay attacks could, in principle, be implemented by anyone using common consumer devices to playback audio to the ASV system microphone. Secondly, in terms of data quality, the SAS corpus consists of high-bandwidth clean speech without extrinsic condition variation. Thirdly, ASVspoof 2015 was built upon text-*independent* ASV even if text-*dependent* systems are the ones used more frequently in authentication scenarios. Finally, the evaluation metric in ASVspoof 2015, averaged equal error rate, ignores countermeasure calibration across differing conditions or attacks as it uses (implicit) attack-dependent thresholds — which is *not* known *a priori* in reality.

The primary technical goal of ASVspoof 2017 is to assess spoofing attack detection accuracy with 'out in the wild' conditions, thereby advancing research towards generalized spoofing countermea-

*Document Version 1.2 (September 19, 2018) - Please note that the ASVspoof 2017 challenge has been concluded in 2017. To obtain the latest version of ASVspoof 2017 data, please navigate to https://datashare.is.ed.ac.uk/handle/10283/3055. Refer to [1] for details about this latest version of ASVspoof 2017. Please address any future correspondonces to info@asvspoof.org. At the time of updating this document, the next ASVspoof 2019 is planned and will include replay attacks too– please follow http://www.asvspoof.org/ for updates.

sure, in particular to detect replay. In addition, ASVspoof 2017 attempts to better interlink the research efforts from spoofing and text-dependent ASV communities. To this end, ASVspoof 2017 makes an extensive use of the recent text-dependent *RedDots* corpus [5], as well as a *replayed* version of the same data [6].

This document provides a technical description of the ASVspoof 2017 challenge. Similar to ASVspoof 2015, the organizers have done their best to retain the challenge easy and open to both specialists and non-specialists: the challenge task, discrimination of human speech from replay spoofs does not require specific ASV knowledge but can be addressed using generic machine learning or signal processing techniques.

2 Technical objective

The challenge focuses on the development of novel spoofing attack detectors capable of detecting varying replay attacks embedded in both known and unknown conditions. It aims to:

- Answer the fundamental question *whether it is feasible to discriminate a live speech from a replayed one based on 'audio' cues only in the first place.*[1]
- Promote development of replay attack spoofing detectors that discriminate human and replayed speech embedded across varied replay environments, playback devices, and speakers.
- To study the impact of replay attack condition factors to both ASV and replay attack detection accuracy.
- Stimulate the development of countermeasures that require as little re-calibration/re-training to novel replay environments as possible
- Test the feasibility of a crowd-sourced replay attack data collection as a source for a spoofing evaluation.

Expertise in automatic speaker verification is not a pre-requisite to participation in ASVspoof 2017. The task is to discriminate authentic human speech from replay spoofed speech.

3 Data conditions

The challenge will primarily be based on the recent text-dependent *RedDots* corpus [5] and its replayed version [6], the former serving as a source of *genuine* recordings and the latter as a source of *replay spoof* recordings. The latter was collected through a crowd-sourcing exercise in the ongoing H2020-funded OCTAVE project[2], by replaying a subset of the original RedDots corpus utterances through various replay configurations consisting of varied devices, loudspeakers, and recording devices, under a variety of different environments across four European countries[3].

The full dataset is partitioned into three subsets, the first for training, the second for development and the third one for evaluation. The number of speakers in training and development subset with corresponding number of genuine and spoofed utterances is presented in Table 1. The evaluation set is expected to contain no more than 50,000 audio files. It is expected to contain a larger number of speakers than the training and development parts, and a much larger replay-to-genuine ratio in comparison to the development part.

3.1 Training data

The training set includes genuine and spoofed speech from 10 male speakers. Spoofed speech is generated with three different replay configurations in six different sessions. Participants may use all of them to train spoofing detectors or countermeasures. Importantly, the SAME training data is shared across development and evaluation subsets. Thus, the only difference between "develop-

[1] One argument is that high-end loudspeakers, used for implementing replay attacks in silent environments, yield audio that might be indistinguishable from authentic recordings. Besides treating replay detection as a statistical 'channel pattern noise' detection problem, other known techniques involve checking whether an utterance was earlier encountered by the ASV system, using digital audio watermarking techniques, for instance. A challenge design for such approaches would require an entirely different design and will not be addressed in ASVspoof 2017.

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

[3] This sidesteps the issue of *covert (far-field) recording* of the target speakers. ASVspoof 2017 corresponds to a 'stolen voice' scenario where the attacker has an access to the digital copy of authentic recording.

Table 1: Number of speakers and the corresponding number of genuine and spoofed utterances in the training and development sets.

Subset	#speakers	#utterances	
		#genuine	#spoofed
Training	10	1508	1508
Development	8	760	950

ment protocol" and "evaluation protocol" is the set of test files on which one executes their countermeasure.

3.2 Development data

The development subset includes genuine and replayed speech recordings from a total of 8 speakers. The replay spoof samples originate from 10 different replay sessions with different playback and recording devices. The devices used in the dev-part are mostly different from those used in the training set. All of this data may be used for the design and optimization of replay spoofing detectors.

Participants should be aware, however, that the evaluation data may include new speakers, environments, replay-recording device combinations or novel attacks that differ substantially from those in the development part. The aim, therefore, is to develop countermeasures which have the potential to generalise well to new data generated with different replay configuration.

Besides the primary label of the ASVspoof 2017 task (genuine/replay), each audio file in the training and development parts are provided with information of the phrase ID (text content), speaker, recording environment, playback device, and (re)-recording device. The participants are free to use this information as they wish; for instance, developing countermeasures calibrated to certain texts, or developing speaker/session/device normalization techniques.

3.3 Evaluation data

The evaluation data includes a similar mix of genuine and replay spoof speech. Some of the the replay conditions will be exactly the same as those in the training and/or development parts. Majority of the replay attacks will originate from different (unseen) configurations from those in the training and development parts. Being intentionally different, they will provide insight into countermeasure performance 'in the wild', while the inclusion of the known conditions provides reference diagnostic data for analyzing the factors that affect replay attack detection performance. The challenge performance (system ranking) will be based on pooled evaluation data across both the seen and unseen conditions.

4 Performance measures

Similar to the 2015 edition, ASVspoof 2017 challenge concentrates on stand-alone spoofing detection without 'integration' with an ASV system. A standard protocol will be released with both development and evaluation datasets. The protocol dictates simply a list of trial segments, each corresponding to a randomly named audio file of either genuine speech or replayed speech. Participants should assign a real-valued, finite *score* to each trial which reflects the relative strength of two competing hypotheses, namely that the trial is genuine speech or spoofed speech[4]. Similar to ASVspoof 2015, we assume that the positive class represents genuine speech. Therefore, **high detection score should indicate genuine speech whereas low scores should indicate a replay spoofing attack**.

Participants are not required to optimize a decision threshold or submit hard decisions; the primary metric will be the "threshold-free" *equal error rate* (EER). Let $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ denote the false alarm and miss rates at threshold θ:

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

$$P_{\text{miss}}(\theta) = \frac{\#\{\text{human trials with score} \leq \theta\}}{\#\{\text{Total human trials}\}},$$

so that $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ are, respectively, monotonically decreasing and increasing functions of θ. EER corresponds to the threshold θ_{EER} at which

[4]Examples include log-likelihood ratios or support vector machine (SVM) discriminant values.

the two detection error rates are equal[5], i.e. EER = $P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. Systems will be ranked according to the EER obtained on the evaluation set.

Differently from the 2015 edition, where EERs were first computed per spoofing attack and then averaged to form a single scalar summary, in ASVspoof 2017 the summary EER (used for comparing the submitted systems) will be computed from *scores pooled across all the trial segments*. The rationale is to promote development of replay attack countermeasures that produce consistent detection scores across varied spoofing attack configurations.

5 Evaluation rules

The participants are free to use the development part as they wish. It can be used for optimizing classifier parameters or re-partitioned freely for modified training/dev division – the organizers are interested only in your test set scores. Each registered participant can submit up to **six** detection scores for the evaluation set as shown in Table 2, according to usage of training data (**common** or **flexible**) and whether a submission is a **primary** or one of two possible **contrastive** (alternative) submissions. The definitions of training conditions are:

Common: Audio data shared by the challenge organizers, listed collectively in files `ASVspoof2017_train.trn` and `ASVspoof2017_dev.trl` containing in total 1508 + 760 = 2268 genuine and 1508 + 950 = 2458 replay spoof utterances; the training and development lists are provided only as an efficient means of starting work with the data but participants are free to repartition the total 2268 + 2458 = 4726 files for system development as they wish. Participants are kindly requested to describe their use of training and development data in their system descriptions.

Flexible: Any dataset, including possibly non-public ones, that the participants can obtain or generate at their sites. The only exceptions are the source data used for the challenge, RedDots [5] and its replayed derivative data [6], which cannot be used under any circumstances. Participants submitting scores in the flexible training category must provide sufficient details about their training sets: source, number of utterances/speakers, human/spoof balance, etc.

In addition to the required common submission, the participants are allowed to submit three different systems in each training category. **Exactly one submission must be designated as the primary submission which uses only the common training data**. This submission will be used for comparing and ranking different countermeasures. All the score files should be submitted in the format described in Section 6. Submissions must contain **valid detection scores** for the full set of trials.

Similar to the speaker recognition evaluations (SRE) administered by the National Institute for Standards and Technology (NIST) in the US, scores produced for any one trial must be obtained using *only* the data in that trial segment. The use of data from any other trial segment is strictly prohibited. Therefore, the use of techniques such as normalization over multiple trial segments and the use of trial data for model adaptation is not allowed. Systems must therefore process trial lists segment-by-segment without access to past or future trial segments.

6 Submission of results

Each participant/team should submit (1) a brief system description and (2) up to six score file(s) as specified above. Both will be shared among other participants after the evaluation period.

The system description should be a PDF file detailing the countermeasure approach (features and classifiers etc.) and related technologies. The description should list and detail any external data sources used for any purpose.

The score file is a single ASCII text file. Each line of the score file should contain two entries, separated by white space: the unique trial segment

[5] One may not find such threshold exactly as $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ change in discrete steps. You may use $\theta_{\text{EER}} = \arg\min_\theta |P_{\text{fa}}(\theta) - P_{\text{miss}}(\theta)|$ or more advanced methods such as EER based on convex hull (ROCCH-EER) implemented in the Bosaris toolkit, https://sites.google.com/site/bosaristoolkit/

Table 2: Each participant can submit up to six different scores on the evaluation set. Only one set is required, and it must use training only from the set provided by the organizers (see text for details).

Submission	Training condition	
	Common	Flexible
Primary	**required**	Optional
Contrastive1	Optional	Optional
Contrastive2	Optional	Optional

identifier (without the `.wav` extension) and the detection score. An example is shown below:

```
...
E10000001 1.571182
E10000002 -2.735763
E10000003 -4.084447
E10000004 77.868048
...
```

The resulting score file(s) should be submitted by e-mail attachment to:

`asvspoof2017@cs.uef.fi`

with the following subject line:

`ASVspoof17 submission for <participant/team name>`

For the score files, please use the naming convention `<team>_<training-cond>_<submission>` following definitions in Table 2. For instance, `UEF_common_primary.txt`. Score files in excess of 10 MB should be compressed in `.tar.gz` or `.zip` format.

7 Reference replay detector

While the ASVspoof 2015 was a first spoofing challenge to provide a shared data, protocols and metrics, the participants were required to implement their spoofing detectors from scratch. In order to promote reproducible research and to let the participants get quickly started with ASVspoof 2017, the organizers will provide a state-of-the-art reference implementation to detect replay attacks. It is based on *constant-Q transform*, technique extensively used in music information processing and adapted to detect TTS and VC spoofing in [7].

8 Schedule

- Release training and development materials to participants: 23th Dec 2016
- Release evaluation data to participants: 10th Feb 2017
- Deadline for participants to submit evaluation scores: 24th Feb 2017
- Organisers return results to participants: 3rd March 2017
- Interspeech paper submission deadline: 14th March 2017
- Interspeech 2017 (Stockholm, Sweden): August, 2017

9 Glossary

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

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

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

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

Replay spoof trial: A trial in which an authentic human speech signal is first played back through an digital-to-analog conversion process and then re-recorded again through analog-to-digital channel;

an example would be using smartphone A to replay an authentic target speaker recording through the loudspeaker of A to the microphone of smartphone B that acts as the end-user terminal of an ASV system.

References

[1] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 296–303. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-42

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

[3] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proc. Interspeech, Annual Conf. of the Int. Speech Comm. Assoc.*, Dresden, Germany, September 2015, pp. 2037–2041.

[4] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z. 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 Trans. Audio, Speech & Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.

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

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

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