ASVspoof 2015: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan

Zhizheng Wu[1], Tomi Kinnunen[2], Nicholas Evans[3], and Junichi Yamagishi[1]

[1]University of Edinburgh, UK
[2]University of Eastern Finland, Finland
[3]EURECOM, France
http://www.spoofingchallenge.org

December 19, 2014

1 Introduction

The ASVspoof initiative follows on from the first special session in spoofing and countermeasures for automatic speaker verification (ASV) held during the 2013 edition of INTERSPEECH in Lyon, France [1]. The vision behind that first edition was to promote the consideration of spoofing, to encourage the development of anti-spoofing countermeasures and to gather a community to design, collect and distribute standard databases with standard evaluation protocols and metrics. Such an initiative is deemed vital in order to address weaknesses in research methodologies common to the majority of previous work.

In the past, both spoofing attacks and countermeasures have generally been developed with full knowledge of a particular ASV system used for vulnerability assessments [2]. Similarly, countermeasures have been developed with full knowledge of the spoofing attack which they are designed to detect. This is clearly unrepresentative of the real use case scenario in which neither the specific attack, much less the specific algorithm, can ever be known a priori. It is thus likely that the prior work has as much over-exaggerated the threat of spoofing as it has the performance of countermeasures.

The ASVspoof challenge has been designed to address these shortcomings and to support, for the first time, independent assessments of vulnerabilities to spoofing and of countermeasure performance. While preventing as much as possible the inappropriate use of prior knowledge, the challenge aims to stimulate the development of generalised countermeasures with potential to detect varying and unforeseen spoofing attacks.

The first evaluation, ASVspoof 2015, is being held within the scope of a special session at INTERSPEECH 2015 and with a focus on spoofing detection. Expertise in ASV is therefore not a prerequisite to participation in the 2015 edition. ASVspoof 2015 participants are invited to develop spoofing detection algorithms and to submit results for a freely available standard database and according to the standard protocol and metrics described in this evaluation plan.

2 Technical objective

The objective of ASVspoof 2015 is to stimulate the development of novel, generalised spoofing countermeasures which are able to detect variable spoofing attacks implemented with multiple, different algorithms. It aims to:

- facilitate the development of spoofing countermeasures without the inappropriate use of prior knowledge as regards specific spoofing attacks;

- stimulate the development of generalised countermeasures, and
- provide a level playing field to facilitate the comparison of different spoofing countermeasures on a standard dataset, with standard protocols and metrics.

Whereas future editions will investigate the integration of spoofing countermeasures with ASV, the ASVspoof 2015 challenge focuses on the development of stand-alone spoofing detection techniques. Expertise in ASV is therefore not a prerequisite to participation. The task is to distinguish genuine speech from spoofed speech, i.e. speech manipulated according to some automatic transform or conversion, or artificial speech generated by speech synthesis.

3 Datasets

The evaluation is based upon a standard dataset[1] of both genuine and spoofed speech. Genuine speech is collected from 106 speakers (45 male, 61 female) and with no significant channel or background noise effects. Spoofed speech is generated from the genuine data using a number of different spoofing algorithms. The full dataset is partitioned into three subsets, the first for training, the second for development and the third for evaluation. The number of speakers in each subset is illustrated in Table 1. There is no speaker overlap across the three subsets regarding target speakers used in voice conversion or TTS adaptation

Table 1: Number of non-overlapping target speakers and utterances in the training, development and evaluation datasets.

	#Speakers		#Utterances	
Subset	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	≈ 200000

3.1 Training data

The training dataset includes genuine and spoofed speech from 25 speakers (10 male, 15 female). Each spoofed utterance is generated according to one of three voice conversion and two speech synthesis algorithms. The voice conversion systems include those based on (i) frame-selection, (ii) spectral slope shifting and (iii) a publicly available voice conversion toolkit within the Festvox system[2]. Both speech synthesis systems are implemented with the hidden Markov model-based speech synthesis system (HTS)[3]. All data in the training set may be used to train spoofing detectors or countermeasures.

3.2 Development data

The development dataset 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 generated according to one of the same five spoofing algorithms used to generate the training dataset. All data in the development dataset may be used for the design and optimisation of spoofing detectors/countermeasures. Participants should be aware, however, that the spoofing algorithms used to create the development dataset are a subset of those used to generate the evaluation dataset. The aim is therefore to develop a countermeasure which has potential to generalise well to spoofed data generated with different spoofing algorithms.

3.3 Evaluation data

The evaluation data includes a similar mix of genuine and spoofed speech collected from 46 speakers (20 male, 26 female). There are around 200000 trials including genuine and spoofed speech making the evaluation dataset approximately 20 GB in size. The recording conditions are exactly the same as those for the development dataset. Spoofed data are generated according to diverse spoofing algorithms. They include the same 5 algorithms used to generate the development dataset in addition to others, designated as ‘unknown’ spoofing algorithms. Being intentionally different, they will give

[1] https://wiki.inf.ed.ac.uk/CSTR/SASCorpus

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

[3] http://hts.sp.nitech.ac.jp/

some insight into countermeasure performance 'in the wild', i.e. performance in the face of previously unseen attacks.

4 Performance measures

ASVspoof 2015 focuses on standalone spoofing detection. Released according to the schedule presented in Section 7, both development and evaluation datasets are accompanied with standard protocols. They comprise a list of trials, each corresponding to a randomly named audio file of either genuine or spoofed speech. Participants should assign to each trial a real-valued, finite score which reflects the relative strength of two competing hypotheses, namely that the trial is genuine or spoofed speech[4]. For compatibility with NIST speaker recognition evaluations, we assume that the positive class represents the 'non-hostile' class, i.e. genuine speech. High detection scores are thus assumed to indicate genuine speech whereas low scores are assumed to indicate spoofed speech.

Participants are not required to optimise a decision threshold, and thus neither produce hard decisions; the primary metric for ASVspoof 2015 is the 'threshold-free' *equal error rate* (EER), defined as follows. Let $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ denote the false alarm and miss rates at threshold θ:

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

so that $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ are, respectively, monotonically decreasing and increasing functions of θ. The EER corresponds to the threshold θ_{EER} at which the two detection error rates are equal[5], i.e. EER $= P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. While EERs will be determined independently for each spoofing approach, the average EER for the full evaluation dataset will be used for ranking.

5 Evaluation rules

Participants are free to use the training and development datasets as they wish. They can be used for optimising classifier parameters or re-partitioned for custom protocols – only scores for the evaluation dataset must be produced strictly in accordance to the defined protocol. As illustrated in Table 2, participants may make up to six submissions for the evaluation set according to a **common** or **flexible** use of training data and whether a submission should be considered as a **primary** submission or one of two possible **contrastive** (alternative) submissions. They are defined as follows:

Common: participants shall use only data within the training dataset, namely data referenced in `as_train.trn` which contains 3750 genuine and 12625 spoofed utterances.

Flexible: participants may use **any** dataset, including those which are non-public, that participants can obtain or generate themselves. The only exception to this rules is the original source data used for the challenge, namely VCTK, which cannot be used under any circumstances. Participants who elect to submit scores for this training category must provide full details in their system descriptions of all data used for training: source, number of utterances/speakers, genuine/spoofed balance, etc.

Participants may submit three different systems for each training category (six submissions in total). **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 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

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

[5] It is rarely possible to determine θ_{EER} exactly since $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ change in discrete steps. Participants may optionally use $\theta_{\text{EER}} = \arg\min_{\theta} |P_{\text{fa}}(\theta) - P_{\text{miss}}(\theta)|$ or more advanced methods to determine the EER via the convex hull (ROCCH-EER) approach implemented in the Bosaris toolkit: `https://sites.google.com/site/bosaristoolkit/`

Table 2: Participants may make up to six different submissions. The one required submission must use only the ASVspoof 2015 training dataset.

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

only the data in that trial segment. The use of data from any other trial segment is strictly prohibited. This rule excludes the use of techniques such as normalisation over multiple trial segments and the use of trial data for model adaptation. Systems must therefore process trial lists segment-by-segment without access to past or future trial segments.

6 Submission of results

Participants 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 which details the countermeasure approach (features and classifiers etc.) and related technologies. The description should list any external data sources used for any purpose, be that for background modelling, adaptation or any other datasets of converted voice or synthetic speech, for example.

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

```
...
E10000001 1.571182
E10000002 -2.735763
E10000003 -4.084447
E10000004 2.868048
...
```

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

```
asvspoof2015@spoofingchallenge.org
```

with the following subject line:

```
ASVspoof submission for <participant/team name>
```

Score files should be named according to the following convention `<team>_<training-cond>_<submission>` and according to the definitions in Table 2, for instance, `UEF_common_contrastive1.txt`. Score files in excess of 10 MB should be compressed in `.tar.gz` or `.zip` format. The organisers aim to acknowledge submissions within 24 hours.

7 Schedule

- Release of training and development datasets: 16th Dec. 2014
- Release of evaluation dataset: 6th Feb. 2015
- Deadline for participants to submit evaluation scores: 20th Feb. 2015
- Organisers return results to participants: 27th Feb. 2015
- INTERSPEECH paper submission deadline: 20th Mar. 2015
- Release of meta information (including keys) for evaluation set: 15th June 2015
- Special session at INTERSPEECH, Dreden, Germany: Sept. 2015

8 Glossary

For the most part, terminology is intended to be consistent with that of the NIST speaker recognition evaluations. Additional terminology specific to spoofing and countermeasure assessment are as follows:

Spoofing: an adversary, also referred to as an impostor, who attempts to deceive an ASV system by impersonating another enrolled user in order to manipulate verification results.

Countermeasure: a system employed to detect spoofing attacks and thus to protect ASV systems from such attacks. Also referred to as anti-spoofing.

Training data: a dataset of audio files with known ground-truth which can be used to train or learn systems which can distinguish between genuine and spoofed speech.

Development data: a dataset of audio files with known ground-truth which can be used for the development of spoofing detection algorithms.

Evaluation data: a dataset of audio files with no ground-truth and which must be processed to produce scores.

Genuine trial: a trial in which the speech signal is recorded from a human speaker without modification.

Spoofed trial: a trial in which the original, genuine speech signal is modified automatically in order to manipulate ASV.

9 Acknowledgements

We would like to thank Dr. Daisuke Saito from University of Tokyo, Prof. Tomoki Toda from Nara Institute of Science and Technology, Mr. Ali Khodabakhsh and Dr. Cenk Demiroglu from Ozyegin University, Prof. Zhen-Hua Ling from University of Science and Technology of China, and Dr. Elie Khoury from Idiap Research Institute for their contributions to the spoofing materials used in this evaluation; we also thank Dr. Cemal Hanilci, Dr. Md Sahidullah and Mr. Aleksandr Sizov from University of Eastern Finland for their help with protocol validation.

References

[1] 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.

[2] 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.