ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge

Zhizheng Wu\textsuperscript{1}  Tomi Kinnunen\textsuperscript{2}  Nicholas Evans\textsuperscript{3}  Junichi Yamagishi\textsuperscript{1}
Cemal Hanilci\textsuperscript{2}  Md Sahidullah\textsuperscript{2}  Aleksandr Sizov\textsuperscript{2}

\textsuperscript{1}University of Edinburgh, United Kingdom  \textsuperscript{2}University of Eastern Finland, Finland  \textsuperscript{3}EURECOM, France

\texttt{zhizheng.wu@ed.ac.uk, tomi.kinnunen@uef.fi, evans@eurecom.fr, jyamagis@inf.ed.ac.uk}

1. Introduction

Automatic speaker verification (ASV) offers a low-cost and flexible biometric solution to person authentication. While the reliability of ASV systems is now considered sufficient to support mass-market adoption, there are concerns that the technology is vulnerable to spoofing, also referred to as presentation attacks. Spoofing refers to an attack whereby a fraudster attempts to manipulate a biometric system by masquerading as another, enrolled person. Acknowledged vulnerabilities include attacks through impersonation, replay, speech synthesis and voice conversion \cite{1}.

There are two general strategies to protect ASV systems from spoofing: the first involves the continued pursuit of more robust ASV technology in the general sense; the second, more popular approach centres around the development of new spoofing countermeasures. Countermeasures have been reported for replay attacks \cite{2, 3, 4, 5}, speech synthesis \cite{6, 7, 8, 9}, voice conversion \cite{10, 11, 12, 13} and non-speech, artificial signals \cite{14}. For a recent survey, the reader is referred to \cite{1}. While there are currently no alternatives, the use of non-standard databases, protocols and metrics gives rise to two significant problems: (i) a lack of support for comparable and reproducible research, and (ii) countermeasures which lack generalisation.

The focus on highly specific spoofing attacks and the use of non-standard databases often impedes the comparison of different results. For example, much of the work involving voice conversion spoofing attacks was performed with NIST Speaker Recognition Evaluation (SRE) datasets, often with different voice conversion algorithms, protocols and metrics. The Wall Street Journal (WSJ) datasets have been popular in work involving synthetic speech spoofing attacks, but again with a variety of experimental configurations. As a result of database, protocol and metric diversity \cite{15}, comparisons between different experimental results is extremely complicated, if not close to meaningless.

Countermeasures which lack generalisation result from the inappropriate use of prior information in their development. The majority of existing countermeasures are optimised with training data produced using exactly the same spoofing method that is to be detected. This is clearly unrepresentative of the real use case scenario in which it is impossible to know the exact nature of a spoofing attack. At best, research results generated with such methodologies exaggerate countermeasure performance; at worst, they mask the true scale of the problem. Generalised countermeasures \cite{16, 17} are needed to detect previously unseen spoofing attacks, i.e. spoofing attacks in the wild.

The ASVspoof challenge aims to encourage further progress through (i) the collection and distribution of a standard dataset with varying spoofing attacks implemented with multiple, diverse algorithms and (ii) a series of competitive evaluations. Following on from the special session in Spoofing and Countermeasures for Automatic Speaker Verification \cite{18} held during the 2013 edition of INTERSPEECH in Lyon, France, the first ASVspoof challenge \cite{19} will be held during the 2015 edition of INTERSPEECH in Dresden, Germany. The challenge has been designed to support, for the first time, independent assessments of vulnerabilities to spoofing and of countermeasure performance. The initiative provides a level playing field to facilitate the comparison of different spoofing countermeasures on a common dataset, with standard protocols and metrics. While preventing as much as possible the inappropriate use of prior knowledge, the challenge also aims to stimulate the development of generalised countermeasures with potential to detect varying and unforeseen spoofing attacks.

In order to lower the cost of entry and to maximise participation, the first ASVspoof challenge involved only the detection of spoofed speech. By decoupling spoofing detection from ASV, expertise in the latter was not a prerequisite to participation. Participants were invited to develop spoofing detection from ASV, expertise in the latter was not a prerequisite to participation. Participants were invited to develop spoofing detection algorithms and to submit scores for a freely available, standard dataset and protocol. The dataset was generated according to a diverse mix of 10 different speech synthesis and voice con-

\texttt{\url{http://www.spoofingchallenge.org}}
Table 1: Number of non-overlapping target speakers and utterances in the training, development and evaluation sets. The duration of each utterance is in the order of one to two seconds.

<table>
<thead>
<tr>
<th>Subset</th>
<th>#Speakers</th>
<th>#Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Training</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Development</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Evaluation</td>
<td>20</td>
<td>26</td>
</tr>
</tbody>
</table>

version spoofing algorithms. The particular spoofing algorithm involved in any trial was not disclosed during the evaluation. Performance was assessed by the organisers using a standard metric described in the evaluation plan [19].

This paper describes the ASVspoof database, protocol and metrics, all of which are now in the public domain. Also presented is a summary of 16 sets of participant results. Finally, observations and findings are presented with priorities for the future.

2. ASVspoof database and protocols

ASVspoof is based upon a standard database consisting of both genuine and spoofed speech\(^2\). Genuine speech is recorded from 106 human speakers (45 male and 61 female) without any modification, and without significant channel or background noise effects. Spoofed speech is modulated from the original genuine speech data by using a number of speech synthesis (SS) and voice conversion (VC) algorithms. More details and protocols to generate the spoofed speech can be found in [20]. The full dataset is partitioned into three subsets, the first set for training, the second for development and the third for evaluation. The number of speakers and trials in each subset is illustrated in Table 1. There is no speaker overlap across the three subsets.

2.1. Training data

The training set includes 3750 genuine and 12625 spoofed utterances collected from 25 speakers (10 male, 25 female). As illustrated in Table 2, each spoofed utterance is generated by one of the five spoofing algorithms (S1 – S5) as follows:

- **S1** - a simplified frame selection (FS) [21, 22] based voice conversion algorithm, in which the converted speech is generated by selecting target speech frames;
- **S2** - the simplest voice conversion algorithm [23] which adjusts only the first mel-cepstral coefficient (C1) in order to shift the slope of the source spectrum to the target;
- **S3** - a speech synthesis algorithm implemented with the hidden Markov model based speech synthesis system (HTS\(^3\)) using speaker adaptation techniques [24] and only 20 adaptation utterances;
- **S4** - the same algorithm as S3, but using 40 adaptation utterances, and
- **S5** - a voice conversion algorithm implemented with the voice conversion toolkit and with the Festvox system\(^4\).

Table 2: Summary of spoofing algorithms implemented in the challenge database. S1 to S5 are known attacks, examples of which are available for system development. S6 to S10 are unknown attacks seen only in the evaluation set. Dev=Development; Eva=Evaluation.

<table>
<thead>
<tr>
<th>Subset</th>
<th>#trials or #utterances</th>
<th>Vocoder</th>
<th>Spoofing algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Dev</td>
<td>Eva</td>
</tr>
<tr>
<td>Genuine</td>
<td>3750</td>
<td>3497</td>
<td>9404</td>
</tr>
<tr>
<td>S1</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
</tr>
<tr>
<td>S2</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
</tr>
<tr>
<td>S3</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
</tr>
<tr>
<td>S4</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
</tr>
<tr>
<td>S5</td>
<td>2525</td>
<td>9975</td>
<td>18400</td>
</tr>
<tr>
<td>S6</td>
<td>0</td>
<td>0</td>
<td>18400</td>
</tr>
<tr>
<td>S7</td>
<td>0</td>
<td>0</td>
<td>18400</td>
</tr>
<tr>
<td>S8</td>
<td>0</td>
<td>0</td>
<td>18400</td>
</tr>
<tr>
<td>S9</td>
<td>0</td>
<td>0</td>
<td>18400</td>
</tr>
<tr>
<td>S10</td>
<td>0</td>
<td>0</td>
<td>18400</td>
</tr>
</tbody>
</table>

These five algorithms were chosen since they are among the most easily implemented. They are referred to as known attacks, examples of which are available for the training of spoofing detectors. S1 and S2 are two of the most easily implemented VC techniques. S3, S4 and S5 are all implemented with open-source toolkits.

For S1, S2, S3 and S5, 20 utterances were used to train the VC and SS algorithms. These utterances were included in the larger adaptation set used to generate S4. As illustrated in Table 2, S1, S2, S3, and S4, all use the same STRAIGHT vocoder [25] for synthesis, whereas S5 uses an MLSA vocoder [23].

2.2. 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. Spooled 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, for example, to tune classifier hyper-parameters. 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 generalises well to spoofed data generated with different spoofing algorithms.

All meta information, including speaker identities and exact spoofing algorithms was provided to challenge participants for both training and development sets. Participants were allowed to use this information for system optimisation.

2.3. Evaluation data

The evaluation set is comprised of 9404 genuine and 184000 spoofed utterances collected from 46 speakers (20 male, 26 female). Recording conditions for genuine speech are exactly the same as those for the training and development sets. However, spoofed data are generated according to more diverse spoofing algorithms. They include the same five algorithms used to generate the training and development sets and an additional five spoofing algorithms, all referred to as unknown attacks:

\(^2\)http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html
\(^3\)http://hts.sp.nitech.ac.jp/
\(^4\)http://www.festvox.org/
3. Motivation: degraded ASV performance under spoofing attacks

In order to confirm vulnerabilities to spoofing, experiments were conducted using the challenge database and a state-of-the-art Probabilistic Linear Discriminant Analysis (PLDA) [29, 30] ASV system. Five utterances from each target speaker were used as enrolment data. The Wall Street Journal (WSJ0, WSJ1 and WSJCAM) and Resource Management (RM1) databases were used to train the Universal Background Model (UBM) and eigenspaces. More details of the PLDA system can be found in [20].

ASV results for the evaluation set are presented in Table 3. Results are illustrated for the baseline and for the same system when subjected to spoofing (S1-S10). EER=Equal Error Rate.

4. Protocol, metric and results

4.1. Protocol

ASVspoof 2015 focuses on a stand-alone spoofing detection task. The challenge database is accompanied with a standard protocol. It comprises a list of trials, each corresponding to a randomly named audio file. 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. 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.

4.2. Metric

Participants were not required to optimise a decision threshold, and thus neither to produce hard decisions; the primary metric for ASVspoof 2015 is the ‘threshold-free’ EER. For the spoofing detection task the EER is defined as follows. Let \( P_{fa}(\theta) \) and \( P_{miss}(\theta) \) denote the false alarm and miss rates at threshold \( \theta \):

\[
P_{fa}(\theta) = \frac{\#\{\text{spoofed trials with score} > \theta\}}{\#\{\text{total spoofed trials}\}},
\]

\[
P_{miss}(\theta) = \frac{\#\{\text{genuine trials with score} \leq \theta\}}{\#\{\text{total genuine trials}\}},
\]

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 equal i.e. EER = \( P_{fa}(\theta_{EER}) = P_{miss}(\theta_{EER}) \). EERs were estimated using the Bosaris toolkit\(^6\). While the EER was determined independently for each spoofing algorithm, the average EER for the full evaluation dataset was used for ranking submission results.

4.3. Results

Participants were able to submit scores for up to six systems. One of these systems was designated as the primary submission. Spoofing detectors for all primary submissions were trained using only the training data in the ASVspoof 2015 corpus. The dataset was requested by 28 teams from 16 countries; 16 teams returned primary submissions by the deadline. A total of 27 additional submissions were also received. Anonymous results were subsequently returned to each team who were then invited to submit their work to the ASVspoof special session for INTERSPEECH 2015.

This paper summarises the challenge results for primary submissions only. EER results are illustrated in Table 4 in which each line represents the submission of each team. Results are shown independently for known attacks (S1-S5), unknown

---

\(^6\)https://sites.google.com/site/bosaristoolkit/

---

Table 3: PLDA ASV system performance. Results illustrated for the baseline and the same system when subjected to spoofing (S1-S10). EER=Equal Error Rate.

<table>
<thead>
<tr>
<th>Spoofing algorithm</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.42</td>
</tr>
<tr>
<td>S1</td>
<td>32.92</td>
</tr>
<tr>
<td>S2</td>
<td>0.87</td>
</tr>
<tr>
<td>S3</td>
<td>25.42</td>
</tr>
<tr>
<td>S4</td>
<td>28.44</td>
</tr>
<tr>
<td>S5</td>
<td>35.92</td>
</tr>
<tr>
<td>S6</td>
<td>33.76</td>
</tr>
<tr>
<td>S7</td>
<td>29.71</td>
</tr>
<tr>
<td>S8</td>
<td>30.63</td>
</tr>
<tr>
<td>S9</td>
<td>29.50</td>
</tr>
<tr>
<td>S10</td>
<td>45.79</td>
</tr>
<tr>
<td>Average(S1-S10)</td>
<td>29.30</td>
</tr>
</tbody>
</table>
attacks (S6-S10) and the average, ranked according to the latter. Almost all submissions achieved excellent performance for known attacks (for which training data was provided). EERs in the case of unknown attacks are significantly and universally higher. The lowest EER for all attacks is 1.211%, whereas those for known and unknown attacks are 0.003% and 2.013%, respectively. The lowest EER for unknown attacks (2.013%) is 671 times higher than that for known attacks (0.003%).

These results illustrate the potential of over-fitting countermeasures to known attacks which may leave ASV systems prone to unforeseen spoofing attacks. For example, while system D achieves a lower EER than system A in the case of known attacks (0.003% vs 0.408%), the EER for system D is over twice that of system A in the case of unknown attacks (5.231% vs 2.013%). In turn these results thus confirm the importance of developing more generalised countermeasures and also the need for further work and future evaluations.

5. Discussion and future work

Discussed here are some of the limitations of the ASVspoof challenge and priorities for future research. One limitation regards the inclusion of only high-technology speech synthesis and voice conversion spoofing algorithms. While their relative severity is currently uncertain, low-technology replay and impersonation attacks were not considered. Even if these alternative attacks prove to be less severe than speech synthesis and voice conversion, they might well be the most common in practice; their implementation requires no particular expertise, nor equipment. There is thus some cause to include such attacks in future ASVspoof challenges.

The second limitation relates to the focus on STRAIGHT vocoders. Other types of vocoder, such as sinuosoidal vocoders [31] are also popular and their use may have different impacts on spoofing. Accordingly, a greater variety of vocoders and potentially more advanced spoofing algorithms should be considered in future challenges.

The lack of any additive noise or channel effects may also be a limitation. Even if their omission for the first evaluation was a deliberate choice, their effect on spoofing and spoofing detection is currently uncertain. It will thus be important to address additive noise and channel variability in the future.

Future evaluations should also measure the impact of spoofing and detection on ASV. While such work has already been reported, in many cases it considered spoofing attacks implemented with full knowledge of the ASV system. Future evaluations should thus address the integration issue.

It is also stressed that the evaluation was not intended, nor sufficient to compare the relative severity of different spoofing attacks; different levels of effort have been dedicated to developing speech synthesis and voice conversion attacks and different quantities of training data were used in their implementations. A meaningful comparison between the severity of each attack should be a priority for the future. In addition, closer collaboration with the speech synthesis and voice conversion communities should be considered in order that future evaluations include the very best algorithms.

Finally, the focus on text-independent ASV is perhaps not the most representative of authentication applications in which spoofing is relevant. Future evaluations should therefore include an emphasis on text-dependent ASV. The organisers are currently working in this direction.

6. Conclusions

The first automatic speaker verification spoofing and countermeasures challenge (ASVspoof 2015) was highly successful in attracting significant participation. This paper presents the challenge database, organisation, evaluation results, and priorities for future challenges and research.

Even though the best results show an overall average detection EER of less than 1.5%, the EER of unknown attacks is five times greater than that of known attacks. In addition, while some attacks are easily and consistently detected, others (e.g. S10) provoke extremely high error rates nearing 50%. The overall average is not necessarily the best indicator of ASV robustness especially if, as is likely, fraudsters concentrate on the most successful spoofing attacks. Accordingly, the error rates for the most effective spoofing attacks are perhaps more representative than the average error rates. In any case, even if average detection EERs are low, it is the resulting degradation in automatic speaker verification performance which is of the greatest importance; these degradations are often much greater.

The overriding findings from ASVspoof 2015 show the remaining need to develop more generalised spoofing detection algorithms. Generalisation will remain a focus for future evaluations, as will the integration of spoofing detection with automatic speaker verification and text-dependent scenarios.

7. Acknowledgements

We would like to thank Dr. Daiwue 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, and Prof. Zhen-Hua Ling from University of Science and Technology of China for their contributions to the spoofing materials used in the challenge. The work was partially supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology), Academy of Finland (project no. 253120 and 283256) and by the TABULA RASA project funded under the 7th Framework Programme of the European Union (EU) (grant agreement number 257289).
8. References


