REDDOTS REPLAYED: A NEW REPLAY SPOOFING ATTACK CORPUS FOR TEXT-DEPENDENT SPEAKER VERIFICATION RESEARCH

Tomi Kinnunen[1], *Md Sahidullah*[1], *Mauro Falcone*[2], *Luca Costantini*[2], *Rosa González Hautamäki*[1], *Dennis Thomsen*[3], *Achintya Sarkar*[3], *Zheng-Hua Tan*[3], *Héctor Delgado*[4], *Massimiliano Todisco*[4], *Nicholas Evans*[4], *Ville Hautamäki*[1], *Kong Aik Lee*[5]

[1]University of Eastern Finland, Finland, [2]Fondazione Ugo Bordoni, Italy, [3]Aalborg University, Denmark
[4]EURECOM, France, [5]Institute for Infocomm Research, Singapore

ABSTRACT

This paper describes a new database for the assessment of automatic speaker verification (ASV) vulnerabilities to spoofing attacks. In contrast to other recent data collection efforts, the new database has been designed to support the development of replay spoofing countermeasures tailored towards the protection of text-dependent ASV systems from replay attacks in the face of variable recording and playback conditions. Derived from the re-recording of the original RedDots database, the effort is aligned with that in text-dependent ASV and thus well positioned for future assessments of replay spoofing countermeasures, not just in isolation, but in integration with ASV. The paper describes the database design and re-recording, a protocol and some early spoofing detection results. The new "RedDots Replayed" database is publicly available through a creative commons license.

Index Terms— speaker verification, spoofing, replay

1. INTRODUCTION

Automatic speaker verification (ASV) [1, 2, 3] technology is today exploited in a growing range of real-world user authentication applications. Examples are systems developed by most of today's global technology corporations and a number of large-scale, collaborative projects such as the European Union Horizon 2020 project, OCTAVE[1].

Many of these applications demand not only reliable recognition performance and robustness to environmental and channel variation, but also resilience to circumvention. On this front, recent years have witnessed the emergence of two relatively new, or renewed research directions within the ASV community. The first focuses on *text-dependent* ASV. The second relates to *spoofing countermeasures* [4]. Research in both directions has benefited greatly from community-driven efforts to introduce free and publicly available corpora. These have been essential for the benchmarking of text-dependent ASV systems [5, 6] and spoofing countermeasures [7, 8].

Even if these two research directions have been pursued in relative independence, they are closely intertwined. While text-dependent ASV systems can improve verification reliability beyond what might otherwise be achieved without text constraints, they may be more vulnerable to spoofing through replay attacks. Conversely, spoofing relates to authentication applications which generally demand high usability, short-duration and hence text-dependent ASV.

Previous assessments of ASV and replay attacks have generally involved only a small number of recording and playback conditions, e.g. [9, 10, 11]. As a consequence, countermeasures developed with these databases generally perform well. The reality may be different, however. With the nature of recording and playback conditions being totally unconstrained, the existing databases are probably not representative; ASV systems could be far more vulnerable than past results might suggest. Furthermore, even if the first evaluation of ASV spoofing and countermeasures (ASVspoof) [7] was performed using a range of different spoofing attacks, it lacked a focus on replay attacks and on text-dependent ASV.

A new database is thus needed to support a more meaningful study of replay spoofing and its impact on text-dependent ASV. This is the impetus of the work reported in this paper. It describes the creation of a new database derived from a smale-scale, crowd-sourced re-recording of the text-dependent RedDots database [5]. The new replay corpus represents a significant number of recording and playback conditions while linking research in spoofing to that of the text-dependent ASV community, recent results from which are based upon the RedDots database. The use of the same base corpora for the new replay corpus will thus provide an ideal starting point for further work to assess the impact of replay spoofing and countermeasures on ASV itself, rather than being assessed in isolation.

2. PRIOR REPLAY ATTACK CORPORA

In general, ASV vulnerabilities to replay attacks have received surprisingly little attention in the literature, likely due to a lack of common data. The **AVspoof** corpus[2] is the only publicly available corpus that includes replay attacks [8]. It contains 44 speakers recorded first using two smart-phones and a laptop. Two smart-phones were then used to replay the utterances. The laptop was used both with its built-in speaker and with a high-quality loudspeaker to generate the replayed signals. Playback and recording experiments were conducted in a controlled environment involving varied room acoustics.

Other studies [9, 10, 11] assessed the impact of replay attacks on ASV accuracy using in-house data. The work in [9] compared to the impact of voice conversion and speech synthesis attacks to replay attacks emulated recording and playback impulse responses. In [10], a subset of RSR2015 [6] was used as a source corpus by

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 authors would like to further acknowledge the effort of Sebastiano Trigila (FUB) for his coordination effort of OCTAVE.

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

[2]Not to be confused with the similarly named ASVspoof 2015 corpus [7]

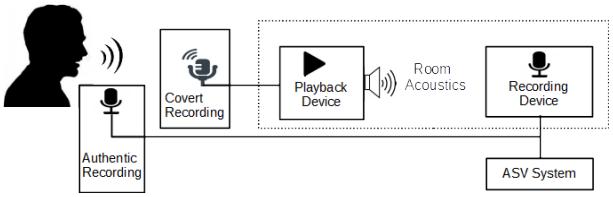


Fig. 1. An illustration of replay spoofing.

playing and re-recording it using a laptop to simulate replay attacks. In [11], a corpus of 175 subjects was collected for different kinds of replay attacks, including telephone and far-field mic recordings.

None of the above work is reproducible since none of the data used is publicly available. These are critical to progress; only through the use of common databases can different research results produced by different teams be meaningfully compared. A new, common database is needed to support future research.

3. REPLAY ATTACKS

This section describes the design and collection of the "RedDots Replayed" database. This was accomplished through a crowd-sourcing approach which captures diverse recording and playback conditions.

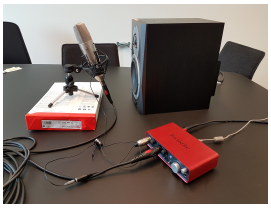
3.1. Definition and resource-constrained collection plan

A replay attack is illustrated in the upper part of Fig. 1. An attacker first acquires a **covert recording** of the target speaker's utterance without his/her consent, then replays it using a **replay device** over some physical space. The replayed sample is captured by the ASV system terminal (**recording device**). In contrast, an authentic target speaker recording illustrated in the lower part of Fig. 1 would also be obtained through some (generally, another) physical space, but captured directly by the ASV system mic.

Creating a corpus to emulate the full scenario of Fig. 1 would ideally require a corpus of multiple simultaneously recorded channels, some of which would be used for representing the covert recording mics, some as target speaker enrolment and some as ASV system test mics. As such data collection is generally tedious, in this study we assume that the attacker has access to the original digital copy of the target speaker recordings. This simplification allows us to re-use any off-the-shelf **source corpus** to study replay.

Attackers in the real-world are not necessarily IT experts but laypersons who share a common-sense understanding of the potential to bypass an ASV system by capturing and replaying someone else's voice using common consumer devices. Thus, we recruited a group of volunteers who were given simple tasks while encouraging them to be creative in emulating replay attacks. The goal was to obtain both unpredictable environments and diverse replay-recording device combinations.

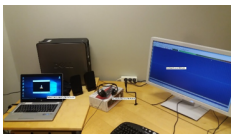
The volunteers, recruited from the ongoing EU H2020 OCTAVE project, were instructed to use their favorite recording/replay audio software in their smart-phones or other devices. To keep the re-recording time feasible, volunteers were provided with long audio files merged from the original utterances to make replay recordings a one-shot task. The long audio files would then be segmented automatically using embedded segment identifiers to obtain the individual utterances.



(a) Recording site 1

(b) Recording site 2

(c) Recording site 3



(d) Recording site 4

Fig. 2. We collect a replayed version of the RedDots [5] corpus through small-scale crowd-sourcing experiment.

3.2. RedDots as a source corpus

We are interested in text-dependent ASV, for which two recent corpora, RSR2015 [6] and RedDots [5], are widely adopted by the community. The RedDots corpus, which contains short phrases recorded using different brands of smartphones was used for the work reported here. It was collected over a time-span of several months to a year and contains subjects from different geographical locations around the globe. The database thus encapsulates diverse channel, session and accent variations. The RedDots database consists of speech files in English, with its Quarter 4 Release having 62 speakers (49 M, 13 F) from 21 countries. The total number of sessions for that release is 572 (473 M and 99 F).

To collect the replay data, we used the Part 01 of the corpus which consists of 10 common short phrases. Since replay attacks are concerned with the playback of target speakers' own voices, we have chosen all the speech samples from speaker-matched trials referred to as target-correct (TC) and target-wrong (TW) in the RedDots evaluation plan [5]. In total, 3498 utterances from 49 male speakers were used. We consider male speakers only due to larger amount of data.

3.3. Replay material preparation and segmentation

The 3498 selected utterances from RedDots corpus were divided into 13 disjoint sets: 3 sets of 250 utterances and one set with the remaining 498 utterances. The utterances were concatenated including interleaved marker tones as embedded segment or utterance identifiers. They signify the beginning of each utterance; corresponding time stamps were used for later segmentation. The marker is a dual-tone multi-frequency (DTMF) tone. The 13 concatenated files of approximately 13 minutes duration were distributed to the volunteers. An additional long file containing all the 3498 samples was provided to two sites. The replayed files were segmented by synchronizing it manually with the corresponding original file and using the recorded time stamps to identify the individual utterances. In total, 130 replayed long files were received from the volunteers.

Table 1. Summary of replay (left) and re-recording (right) devices collected. The recording devices emulate possible ASV system terminal devices on which sensor-level attacks are executed using playback devices.

ID	**P**layback device
P1	ACER "Ferrari ONE" netbook
P2	All-in-one PC speakers
P3	BQ Aquaris M5 smarphone
P4	Beyerdynamic DT 770 PRO headphones with PC
P5	Creative A60 connected to laptop
P6	Dell (SSD) notebook + EdirolUA25 + XXX
P7	Dell laptop with internal speakers
P8	Dynaudio BM5A Speaker connected to laptop
P9	HP Laptop speakers
P10	High-end GENELEC 8020C studio monitorss
P11	MacBook pro internal speakers
P12	PC with Altec lansing Orbit USB iML227 speaker
P13	Samsung GT-I9100
P14	Samsung GT-P6200
P15	VIFA M10MD-39-08 Speakers with laptop

ID	**R**ecording device
R1	AKG C562CM + Marantz PMD670
R2	BQ Aquaris M5 smartphone. Software: Smart voice recorder
R3	Desktop Computer with headset and arecord
R4	H6 Handy Recorder
R5	Logitech C920 connected to Dell (SSD) notebook
R6	Nokia Lumia
R7	Røde NT2 microphone with a laptop
R8	Røde smartlav+ mic with a laptop
R9	Samsung GT-I9100
R10	Samsung GT-P6200
R11	Samsung Galaxy 7s
R12	Samsung Trend 2
R13	Samsung Trend 3
R14	ZoomHD1
R15	iPhone 5c
R16	iphone4

3.4. Data collection sites

The replay recordings were executed in four distinct locations representing four OCTAVE project partners, UEF (Finland), FUB (Italy), AAU (Denmark) and EUR (France). The volunteers were instructed to do at least one recording in a **controlled** condition and at least one in a **variable** condition. The former refers to a silent environment, and the latter to any creative choice by the volunteers. Some set-ups are illustrated in Fig. 2 and a summary of the devices is provided in Table 1. In the following we provide a brief description of each site's approach.

AAU: Controlled recordings where made in a small office room of approximately 6.5m x 3.5m x 3.5m (H) with a large meeting table in the middle. The replay device is P8 from Table 1, while the recording devices are R7, R8 and R11. Uncontrolled recording were made in a student canteen with a large entrance hall. The uncontrolled setup was the same as the controlled setup, except that the replay device was changed to P15.

EURECOM: Controlled recording were made in a silent office. The internal audio card of a desktop PC was used for both playback and recording. The playback device was a pair of Beyerdynamic DT 770 PRO headphones connected to the audio device output. The microphone of a conventional headset device connected to the audio device input was used for audio capture. It was placed immediately between the two headphone speakers. Different mobile and portable devices (BQ Aquaris M5, iPhone 5c, Dell laptop) were used for playback/recording in different environments (silent living room, office room with windows open, bedroom with windows open facing a street producing traffic noise) for uncontrolled recordings.

FUB: Controlled recordings were collected in a silent room of dimensions 6.85m x 3.65m x 2.40m (H). From 500Hz, the isolation level exceeds 40dB. Reverberation time is in the range 0.3-0.5s and the background noise is approximately 2dBSPL. One replay recording came from a notebook controlling an EDIROL AU25 digital board connected to a professional amplified loudspeaker (LEM SoundPressure). It was recorded using a binaural microphone based on two AKG C 562 CM, with professional Marantz PMD670 solid state recorder. Another recording came from the loudspeaker of an ACER ONE (Ferrari) netbook, recorded by a DELL XSP notebook equipped with an external Logitech C920 HD webcam. Uncontrolled recordings were made using a Samsung tablet GT-P6200, Samsung smartphone GT-I9200, iPhone4, and a solid state recorder ZoomHD1. In order to have a stable and reproducible condition for the uncontrolled recording, a special housing-base was developed for the devices, as illustrated in Fig. 2.

UEF: Controlled recordings were made in a silent office and in a silent apartment room. The replay devices were a desktop PC with high quality Genelec Studio speakers and All-in-One PC speakers. The audio was recorded using a Zoom H6 Handy recorder with an omni-directional headset mic (Glottal Enterprises M80). Uncontrolled recordings were collected in a coffee room, office room, and from an open balcony. The office recordings contain additive noises generated from a Nexus 4 smartphone playing bar or small pub noises in the background. The playback devices include a laptop with external Creative A60 speakers, a HP Elite book laptop speakers and a desktop PC with portable Altec lansing Orbit USB iML227 speaker. Recordings used two smartphones: a Nokia Lumia 635 and a Samsung Galaxy Trend 2.

3.5. Analysis of collected data

A total of 130 long files were received within a period of a few days. After segmentation, we extracted 49,432 individual replayed utterances. One of the aims was to collect replay attacks of varied technical quality. To give a sense of how well this was achieved, we measure the signal-to-noise ratios (SNRs) of the collected data, obtained using NIST STNR tool[3]. The SNR histograms of the individual utterances for both controlled and variable conditions, along with the original RedDots files are shown in Fig. 3. The SNR distributions of RedDots and controlled data are similar, though the replay data also contains examples of lower SNRs, as might be expected. Comparing the controlled and variable conditions, the proportion of utterances with high SNR is lower and the mode occurs at lower SNR for the latter, as expected.

4. EXPERIMENTS

To demonstrate corpus usage, we defined pilot protocols both for standalone replay attack detection and ASV. For the former, we use 10 training speakers that are disjoint from the test utterance speakers. Further, any long audio file, whose one or more segments were

[3] http://www.itl.nist.gov/iad/mig//tools/spqa_23sphere25tarZ.htm

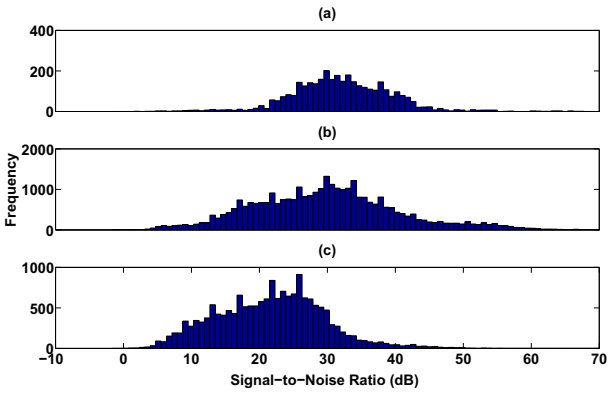


Fig. 3. Histograms of signal-to-noise ratios (SNRs) for speech segments of (a) source RedDots corpus, (b) controlled replay condition and (c) variable replay condition.

Table 2. Protocol for standalone replay attack detection task.

Utterance type	Train	Test
Original	1508	2346
Replayed	9232	16067

used in the training set is excluded from the test side. The training segments include data from the controlled condition only while the test replay data is taken from both controlled and variable conditions. Statistics of the replay detection task are given in Table 2.

The 10 speakers used for countermeasure training are excluded from our ASV protocol. The target models of the remaining 25 speakers are taken from the original RedDots protocol (Part01, Male). Three types of ASV trials are considered: genuine (G), zero-effort spoof (ZS), and replay spoof (RS). Here, G refers to target speaker trials, ZS to original non-target speaker trials (without replay), and RS to collected replay trials. All trials have matched texts. G and ZS trials are similar, respectively, with the target correct (TC) and impostor correct (IC) trials in the original RedDots protocol. The protocol contains a total of 1850 G, 14938 ZS and 52108 RS trials.

Our ASV systems use 19 MFCCs from 20 mel-filters processed through RASTA filter, augmented with deltas and double deltas and normalized via cepstral mean and variance normalization (CMVN) after speech activity detection. We consider both a Gaussian mixture model – universal background model (GMM-UBM) and i-vector back-ends. A gender-dependent UBM of 512 components is trained using all TIMIT male speakers plus 50 male speaker from the RSR2015 background set. A 400-dimensional i-vector extractor is trained using the same data. I-vectors are scored both using cosine similarity and probabilistic LDA, trained using RSR 2015 with an eigenvoice subspace dimensionality of 200. The i-vectors are centered, length normalized and whitened. ASV results in Table 3

Table 3. Speaker verification accuracy (EER, %) using GMM-UBM and two i-vector based systems.

Impostor	GMM-UBM	ivec (cos)	ivec (PLDA)
Zero-effort	2.50	6.64	5.23
Replay	23.18	26.63	24.85

Table 4. Accuracy of two countermeasures (EER, %) on the spoofing protocol, for controlled condition, variable condition and pooled trials.

Feature	Controlled	Variable	All
LFCC 20-DA	5.88	4.43	5.11
CQCC 20-A	2.77	3.50	3.27

indicate that performance degrades considerably under a replay attack scenario. The GMM-UBM system outperforms both i-vector systems, either due to the use of short utterances or non-optimal data selection.

For standalone replay attack detection, we evaluate two countermeasures based on linear frequency cepstral coefficient (LFCC) and constant Q cepstral coefficient (CQCC) [12] features with a common GMM back-end. In each case, two GMMs (genuine speech and replay speech) are trained using all the training data. Table 4 shows the results using the best LFCC configuration reported in [13] and the best CQCC configuration reported in [12]. Neither of these configuration were tuned for replay detection. The LFCC system consists of 20 delta plus double-delta coefficients derived from 20 static coefficients. CQCC features contain only 20 double-deltas which are derived from 20 static coefficients. Countermeasure performance is slightly better for the controlled condition than for the variable condition in the case of CQCC features, while the opposite trend is obtained for LFCC features.

5. CONCLUSIONS

Our study reports the design and crowd-sourced collection of a new database intended to support the development of countermeasures to protect text-dependent automatic speaker verification (ASV) systems from replay spoofing. The crowd-sourcing approach described in this paper enables the collection of replay data with significant recording and playback variation. Efficient data collection was enabled by the reuse of an existing speech corpus. This scenario involves neither far-field covert recordings of the target speaker nor recordings of the original target speakers using the same re-recording devices or environments, but provides a good sampling of different device and environment combinations. The use of the original RedDots database meanwhile aligns the effort to develop replay spoofing countermeasures with that in text-dependent ASV.

Our study also reports early, independent ASV and spoofing detection results. While ASV performance is shown to degrade as a result of replay attacks, the performance of un-optimised spoofing countermeasures gives cause for optimism. Equal error rates lower than those reported in this paper will almost certainly be achieved through further work.

Currently, we use the described RedDots Replayed corpus as our primary data to set-up the ongoing 2017 edition of *automatic speaker verification spoofing and countermeasures challenge* (ASVspoof) [14]. A subset of the data has been made available in December 2016 as a development set for the challenge (containing substantially revised evaluation protocols from those described above). The full corpus will be released after the challenge.

While almost all previous work, including that stemming from the recent, competitive ASVspoof 2015 evaluation has assessed spoofing detection in isolation from ASV, the effect of spoofing on ASV itself is the primary concern. Future research should therefore look towards the joint assessment of spoofing countermeasures and ASV, which is left as a near-future work.

6. REFERENCES

[1] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Comm.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[3] 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.

[4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Comm.*, vol. 66, pp. 130–153, 2015.

[5] 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, M.J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 2996–3000.

[6] 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.

[7] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2037–2041.

[8] S.K. Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Applications and Systems (BTAS)*, Arlington, USA, Sept. 2015, pp. 1–6.

[9] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *BIOSIG 2014 - Proceedings of the 13th International Conference of the Biometrics Special Interest Group, 10.-12. September 2014, Darmstadt, Germany*, 2014, pp. 157–168.

[10] Z. Wu, S. Gao, E.S. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014, Chiang Mai, Thailand, December 9-12, 2014*, 2014, pp. 1–5.

[11] J. Galka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Comm.*, vol. 67, pp. 143–153, 2015.

[12] 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.

[13] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, 2015, Dresden, Germany.

[14] T. Kinnunen, N. Evans, J. Yamagishi, K.A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," http://www.spoofingchallenge.org/.