LRPD: LARGE REPLAY PARALLEL DATASET

Ivan Yakovlev *Mikhail Melnikov* *Nikita Bukhal* *Rostislav Makarov*
Alexander Alenin *Nikita Torgashov* *Anton Okhotnikov*

ID R&D Inc., New York, USA

ABSTRACT

The latest research in the field of voice anti-spoofing (VAS) shows that deep neural networks (DNN) outperform classic approaches like GMM in the task of presentation attack detection. However, DNNs require a lot of data to converge, and still lack generalization ability. In order to foster the progress of neural network systems, we introduce a Large Replay Parallel Dataset (LRPD) aimed for a detection of replay attacks. LRPD contains more than 1M utterances collected by 19 recording devices in 17 various environments. We also provide an example training pipeline in PyTorch [1] and a baseline system, that achieves 0.28% Equal Error Rate (EER) on evaluation subset of LRPD and 11.91% EER on publicly available ASVpoof 2017 [2] eval set. These results show that model trained with LRPD dataset has a consistent performance on the fully unknown conditions. Our dataset is free for research purposes and hosted on GDrive[1]. Baseline code and pre-trained models are available at GitHub[2].

Index Terms— Automatic speaker verification, Voice Anti-Spoofing, Physical access, Replay, Dataset

1. INTRODUCTION

Over the last years a lot of work has been done in the voice biometrics field. Recent speaker recognition systems are able to successfully recognize person using the voice over various conditions. Such progress allows to build reliable voice-based solutions for person authentication. However, it is usually assumed that biometrics systems are vulnerable to spoofing attacks, also known as presentation attacks. For the best of our knowledge, there are two main types of spoofing attacks, logical and physical access attacks.

Logical access (LA) attach comprised two different approaches based on speech synthesis systems. Text-to-speech (TTS) systems are used to generate a fully artificial speech based on the specified text, while the voice conversion (VC) systems use a speech of one person as an input and convert it into the speech that resembles the voice of another person.

Physical access (PA) attacks are also known as Replay attacks, and are performed in the following way: the bona fide speech of the target speaker is recorded first, and then it is being presented to the speaker recognition system by a playback using the mobile phone or speaker. In this paper we focus on the problem of physical attacks, due to the ease of implementation and a high difficulty of detection.

There are multiple publicly available datasets for training presentation attack detection systems. The most widespread datasets are related to ASVspoof challenges, that made a huge contribution to the VAS research. The first challenge, ASVspoof 2015 [3], was focused on the Logical access attacks detection. The second challenge, ASVspoof 2017, was focused on the physical access attacks detection. The ASVspoof 2019 [4] challenge aimed to consider both types of presentation attacks. AVspoof [5] is a public audio spoofing database which includes 10 various spoofing threats generated using replay, TTS and VC systems. VoicePA [6] is an extension of the AVspoof database, which includes presentation attacks recorded in different environments with various recording and playback devices. PHONESPOOF [7] is a database that was collected in the telephone channel domain and is used to investigate the robustness of anti-spoofing systems in the telephone channel conditions.

The specific issue associated with replayed databases, is that they tend to become outdated over the time. The main difference between bona fide and replayed speech is a small distortion of source signal, caused by both recording and playback devices, which is likely to be used by anti-spoofing system to discriminate. Thus, even robust replay detection system may be vulnerable to the hardware or software speech preprocessing algorithms of new smartphones, which are being updated at least annually. We believe, that LRPD dataset, containing recent smartphones, with its wide covering of recording environments, is going to be very handful for the further development of replay detection systems.

In Section 2 of this paper we present the LRPD dataset. Section 3 describes experiments conducted on the LRPD and ASVspoof 2017 datasets. And finally, Sections 4 and 5 contain analysis of experiments' results and conclusions on our work.

[1] https://drive.google.com/drive/folders/1lHxQ5tPco5F1N8xv_x7lfDOhU2SEsOU9?usp=sharing

[2] https://github.com/IDRnD/lrpd-paper-code

Table 1. Source datasets

	Train part		Eval part	
Dataset name	Spks	Utts	Spks	Utts
VCTK	-	-	107	1172
LibriSpeech	1252	16724	-	-
MCV	22403	26225	427	427
GLR	250	23377	176	6490
CN-Celeb	1002	6307	-	-
Total files	72633		8089	
Total size, Gb	22.5		2.4	
Total duration, hours	209.5		22.1	

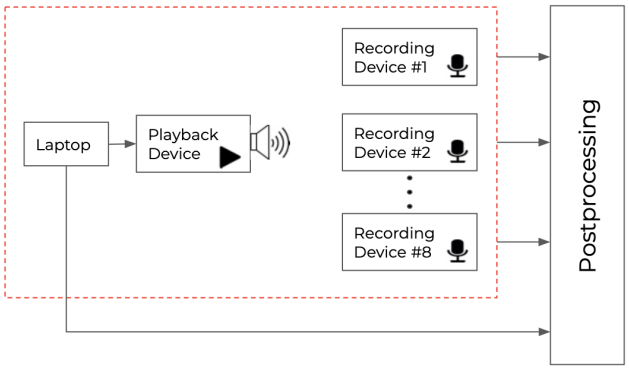

Fig. 1. Schema of a stand

2. DATASET DESCRIPTION

The LRPD corpus was collected and open-sourced to push the boundaries of current research in the field of Replay spoofing attacks detection. The distinctive feature of LRPD dataset is that it contains several copies of replayed audio recorded by different devices at the same time in parallel.

2.1. Audio sources

The LRPD dataset contains both bona fide and replayed types of utterances. Bona fide speech was taken from VCTK [8], LibriSpeech [9], Mozilla Common Voices (MCV) [10], Google Language Resources (GLR) [11], and CN-Celeb [12] datasets. For each dataset we randomly sampled a subset of files to use. The resulting number of speakers and files selected for the replaying is shown in Table 1.

2.2. Collection session

Large replay parallel dataset is made of many recording sessions. One session consists of sampling a subset of data from the source datasets, selecting one playback device, multiple recording devices and pinning one environment.

2.2.1. Session data

For each session we randomly picked the subset of the source data, about 4-8 hours long. All selected files were merged into one big file that was played by a playback device. After finishing recording, we applied a post-processing step to cut the audio back and map the recorded utterances to their original metadata.

2.2.2. Session devices

We used different configurations of playback and recording devices to collect the database. 19 recording and 11 playback devices were used in total (see Table 2 and Table 3). For each session we picked one playback and multiple recording devices. The recording was carried out using standard iOS/Android audio recording API.

2.2.3. Session environment

There are total 17 unique recording environments in LRPD dataset. They include 25 sessions from an anechoic room with a stochastic noise (caused by air conditioning system), and up to 8 sessions for the rest of 16 environments each. These 16 environments represent the rooms and apartments of different sizes, and we also provide 8 labels of known apartments in the dataset. For the data where it was impossible to recover the apartment id we just set the label *unknown* and this is a mix of recordings from the rest 8 environments. For all the environments recording distance varied from 8 to 100 cm and it was pinned for each session and was similar for every device in that session.

2.3. Collection stand

The collection stand was based on a laptop, Wi-Fi router, playback device and multiple recording devices. Playback device was connected to the laptop, and all the devices were connected to a power supply. Stand's schema is shown on Figure 1.

2.4. Post-processing dataset

After collecting the data, we cut audio according to saved playback metadata. To double-check the quality of dataset we applied a cross-correlation function to the original file and its replayed copy.

2.5. Dataset statistics

Aggregated LRPD statistics and distributions across various playback devices (Table 2), recording devices (Table 3),

Table 2. Playback device statistic

	Ratio, %		
Playback	trn aparts	trn office	val aparts
srs xb12	20.95	19.87	20.84
ginzzu gm 877b	18.85	9.05	18.82
lg pj2b	14.53	24.42	14.64
jbl go3	13.28	5.71	13.31
jbl flip4	12.26	11.92	12.37
oklick ok-128	11.08	9.32	10.87
sharp	6.70	0	6.59
defender enjoy s500	2.34	0	2.56
sps 609	0	6.08	0
digma s16	0	5.67	0
jbl clip3	0	7.98	0
Total files	469908	566832	195997
Total size, Gb	95.7	125.6	27.0
Total duration, hours	891.6	1170.6	251.8

source data (Table 4) and recording environments are presented in corresponding tables.

2.6. Dataset partitions

We split LRPD into 3 parts: noised-train (`trn_office`), clean-train (`trn_aparts`) and test (`val_aparts`). Noised train set contains files collected in an anechoic room. Clean train and test sets represent the rest 16 environments from Section 2.2.3.

Metadata is currently included into dataset structure, and for information please check README file in dataset root.

3. EXPERIMENTS

3.1. Description

We conducted experiments for two different tasks:

Task 1. Replay detection: binary classification of spoof / human. The goal of these experiments was to measure impact of adding new LRPD data for replay detection problem.

Task 2. Recording / Playback device classification (Device detection): two multi-label classification problems. Using the models obtained from device detector training we were able to explore embedding space using T-distributed Stochastic Neighbor Embedding (t-SNE) to visualize a distribution of 2-D embeddings on ASVspoof2017 eval subset.

For more information on architectures, data setup, and training hyperparameters please refer to the baseline training pipeline.

Table 3. Record device statistic

	Ratio, %		
Record	trn aparts	trn office	val aparts
huawei matepad pro	13.12	0	13.14
huawei mate 40 pro	13.11	0	13.13
huawei mate 30 pro	13.01	0	13.01
huawei p smart z	8.66	4.59	8.64
honor 10x lite	4.44	4.56	4.46
honor 30 pro+	0	1.27	0
iphone 7	4.19	3.06	4.24
iphone 8	7.85	0	7.8
iphone 11 pro	0	9.18	0
iphone 11 pro max	4.51	9.18	4.54
iphone 12 pro max	0	9.66	0
iphone xr	7.72	8.47	7.62
samsung galaxy a01	0	7.9	0
samsung galaxy a51	0	4.85	0
samsung galaxy m21	6.82	3.12	6.81
samsung galaxy s8+	8.43	10.37	8.4
samsung galaxy s20+	3.9	9.98	3.94
zte blade v2020	4.23	5.2	4.26
sony xperia zx3	0	8.6	0
Total files	469908	566832	195997
Total size, Gb	95.7	125.6	27.0
Total duration, hours	891.6	1170.6	251.8

3.2. Datasets

For the replay detection task, we used different combination of LRPD and ASVspoof 2017:

1. LRPD all train (office + aparts) + ASVspoof 2017 train
2. LRPD all train (office + aparts)
3. ASVspoof 2017 train

For the device detection task, we used LRPD all train (office + aparts) for training and LRPD eval for evaluation.

3.3. Architecture

We have chosen RawNet architecture [13] as a feature extractor model for both tasks, as we found it most suitable for detection of replay attacks. We have slightly changed initial RawNet architecture by replacing Gated recurrent unit (GRU) pooling with Statistical pooling layer and reducing model size using depth multiplier equals to 0.625. For each task we used different Fully Connected (FC) classifier head setup.

3.4. Training setup

For both tasks we used Adam optimizer with the following learning rate (lr) schedule: constant value of lr [$1e^{-3}$, $5e^{-4}$,

Table 4. Source dataset statistic

Source	Ratio, % trn aparts	trn office	val aparts
MCV	36.3	29.4	16.68
GLR	32	30.47	39.43
LibriSpeech	24.54	22.52	0
CN-Celeb	7.15	17.61	0
VCTK	0	0	43.89
Total files	469908	566832	195997
Total size, Gb	95.7	125.6	27.0
Total duration, hours	891.6	1170.6	251.8

Table 5. Results on replay detection task.

Datasets used	EER, % LRPD eval	ASV17 eval	ASV17 dev
LRPD all train	**0.16**	17.18	27.84
ASV17 trn	21.70	13.94	**17.54**
LRPD all train + ASV17 trn	0.28	**11.91**	18.63

$1e^{-4}$, $5e^{-5}$] was dropped down after each 4 epochs, and we trained each model for 16 epochs in total. When training Task 1 models we sampled even number of utterances per class (replay/human) and even number of utterances from each dataset, forming batches of size 64. While for the Task 2 we sampled 4 replayed utterances with the same source utterance and cut them so they maintain aligned resulting in 4 x 32 batch size for Task 2.

Cross-entropy (CE) loss was used for both tasks, except that for Task 2 device detection we summed up CE losses from two simultaneous tasks: playback device classification and recording device classification:

$$\mathcal{L}_{device-detection} = \mathcal{L}_{playback} + \mathcal{L}_{recording}$$

We augmented training data using noises from MUSAN [14], DCASE [15] and DEMAND [16] datasets. One random noise was added on-the-fly to each utterance with 0.5 probability and a random SNR uniformly sampled from 3-15 dB range.

4. RESULTS

4.1. Task 1. Replay detection

The testing results of replay detector on the LRPD eval, ASVspoof17 eval and ASVspoof17 dev datasets are presented in Table 5. Adding LRPD data in training set may results in increasing of VAS system accuracy even on out-of-domain tests, such as ASVspoof17 eval, where EER drops from 13.94% to 11.91%.

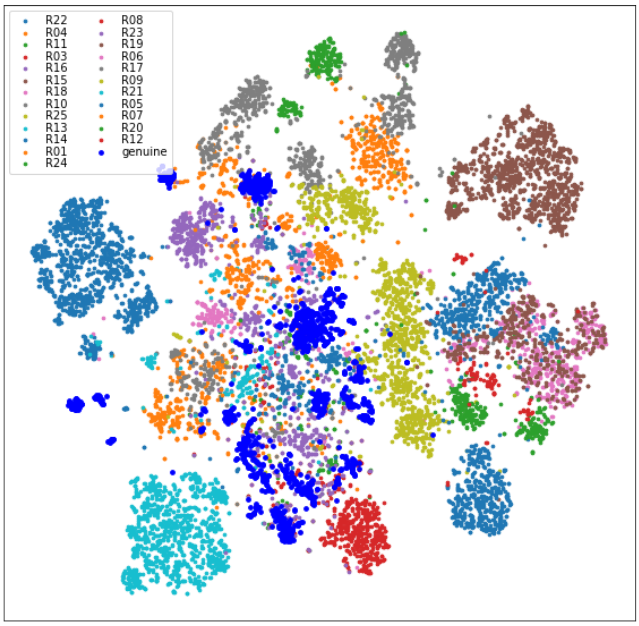

Fig. 2. Device detector embedding space on ASVSpoof 2017 eval. R01-R25 stands for recording device ids in spoof utterances, and genuine class is bonafide.

4.2. Task 2. Device detection

t-SNE Visualisation of the device detector embeddings is presented on the Fig.2. Device detector was trained with LRPD train subset only, and visualized embeddings are extracted from out-of-domain ASVSpoof17 eval subset, for which recording and playback device meta information is given. We got 12.9% EER on ASVSpoof17 eval subset (all with all protocol) for recording/playback pair classification task, using embeddings.

5. CONCLUSIONS

In this paper we presented the LRPD dataset, that was collected to advance future research on replay detection task. Compared with previous open-source datasets, the new corpus is larger, covers up-to-date recording and playback devices and contains more source data variety (speakers, languages). Using evaluations with the proposed data set, we found that the error of the baseline RawNet model drops by relative 15% on target ASVSpoof2017 eval set, when trained with ASVSpoof2017 train and LRPD all train. We hope that proposed data will fuel further research in voice biometrics field by building more robust and protected systems.

6. REFERENCES

[1] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[2] Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection," in *Proc. Interspeech 2017*, 2017, pp. 2–6.

[3] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.

[4] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[5] Serife Kucur Ergünay, Elie Khoury, Alexandros Lazaridis, and Sébastien Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2015, pp. 1–6.

[6] Pavel Korshunov, André R Gonçalves, Ricardo PV Violato, Flávio O Simões, and Sébastien Marcel, "On the use of convolutional neural networks for speech presentation attack detection," in *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. IEEE, 2018, pp. 1–8.

[7] Galina Lavrentyeva, Sergey Novoselov, Marina Volkova, Yuri Matveev, and Maria De Marsico, "Phonespoof: A new dataset for spoofing attack detection in telephone channel," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2572–2576.

[8] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[9] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[10] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[11] Alena Butryna, Shan-Hui Cathy Chu, Isin Demirsahin, Alexander Gutkin, Linne Ha, Fei He, Martin Jansche, Cibu Johny, Anna Katanova, Oddur Kjartansson, et al., "Google crowdsourced speech corpora and related open-source resources for low-resource languages and dialects: An overview," *arXiv preprint arXiv:2010.06778*, 2020.

[12] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang, "Cn-celeb: multi-genre speaker recognition," *arXiv preprint arXiv:2012.12468*, 2020.

[13] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[14] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[15] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.

[16] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," June 2013, Supported by Inria under the Associate Team Program VERSAMUS.