PHONESPOOF: A NEW DATASET FOR SPOOFING ATTACK DETECTION IN TELEPHONE CHANNEL

Galina Lavrentyeva[1,2] *Sergey Novoselov*[1,2] *Marina Volkova*[3]
Yuri Matveev[1,2] *Maria De Marsico*[4]

[1]ITMO University, St.Petersburg, Russia, [2]STC-innovations Ltd., St.Petersburg, Russia
[3] STC Ltd., St.Petersburg, Russia, [4] Sapienza University of Rome, Rome, Italy
lavrentyeva@speechpro.com, novoselov@speechpro.com,
volkova@speechpro.com, matveev@mail.ifmo.ru, demarsico@di.unroma1.it

ABSTRACT

The results of spoofing detection systems proposed during ASVspoof Challenges 2015 and 2017 confirmed the perspective in detection of unforseen spoofing trials in microphone channel. However, telephone channel presents much more challenging conditions for spoofing detection, due to limited bandwidth, various coding standards and channel effects. Research on the topic has thus far only made use of program codecs and other telephone channel emulations. Such emulations does not quite match the real telephone spoofing attacks. In order to asses spoofing detection methods in real scenario we present the PHONESPOOF dataset - spoofing data collected through realistic telephone channels. The PHONESPOOF data collection represents most threatening types of spoofing attacks and is publicly available dataset[1]. This work[2] aimed to investigate robustness of the state-of-the-art deep learning based antispoofing systems under telephone spoofing attacks conditions based on the PHONESPOOF data. Moreover newly collected dataset makes it possible to analize language dependency issue for the Anti-Spoofing methods. In the work we also focused on the development of a unified LCNN-based approach for spoofing attack detection. The goal was to train a single system able to detect various types of spoofing attacks in telephone channel. The obtained results approve the effectiveness of such solution.

Index Terms: spoofing, speaker verification, telephone channel

1. INTRODUCTION

Automatic Speaker Verification (ASV) has grown into a reliable, convenient and low-cost approach for person authentication. However, similar to other biometric modalities, it remains vulnerable to spoofing or presentation attacks [1]. During these attacks a fraudster aims to gain the illegal access to the secure information or system by impersonating the enrolled person. This can be done by using voice conversion and speech synthesis technologies like from [2] or by replaying a prerecorded sample [3], [4]. The first competitive evaluation related to voice spoofing detection was Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge in 2015. It aimed to support the investigation of the methods for detecting speech synthesis and voice conversion in microphone channel. Recent researches achieved extremely high spoofing detection performance on the challenge corpus with an equal error rate (EER) close to zero [5]. The second ASVspoof Challenge in 2017 dealt with replay attack detection. Results of the challenge reconfirmed the impressive perspective in detection of unforeseen spoofing trials based on replay techniques produced in microphone channel.

ASVspoof initiative has significantly pushed forward the development of spoofing detection methods for ASV systems in microphone channel. However, ASV systems in telephone channel robust to spoofing attacks are also in high demand, due to the different applications on the mass market that can use it for access control, for example in telephone-banking. Bandwidth, packet-losses, different GSM codecs and other channel distortions, cause much more challenging conditions for the spoofing detection in the telephone channel.

Several papers deal with the impact of additive noise and channel variation on spoofing detection performance [6], [7], and underline its significant degradation in case of additive distortions. The limitations of these investigations are: 1) they only exploit simulated data, while the real conditions can be even worse and vary greatly; 2) speech synthesis and voice conversion detection are considered separately from replay attacks detection, while in real life no prior information is available about the type of faced attack. Additional limitation we payed attention to was language dependency. All experiments performed so far considered genuine and spoofing trials only on english language datasets, but non of them

[1] To obtain the database for non-commercial use, please contact the authors of this paper via the email address given in the paper title.

[2] This work was partially financially supported by the Government of the Russian Federation (Grant 08-08).

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

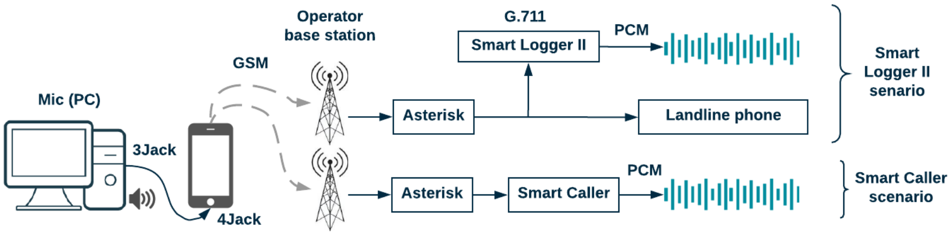

Fig. 1. Recording scheme used for recording the database in the telephone channel

estimate the performance in the cross language scenario and mixed language data.

Our experiments show that systems trained on the simulated data cannot detect spoofing attacks in the real telephone channel. Because of that current research is mainly focused on the database collection that overcomes mentioned limitations. The appropriate database in real telephone channel was collected on the base of ASVspoof2015 corpus, RSR2015 database and other speech synthesis data, that was generated via resources available online. Using this database a single system was designed able to distinguish diverse spoofing attacks in telephone channel. This system was based on LCNN(Light CNN)-approach presented in our previous research in microphone channel [8], that demonstrated high spoofing detection performance on the ASVspoof2017 corpus.

This paper contains the brief description of the collected database in Section 3 and the description of the experiment results in 4. The results obtained by experiments conducted during this investigation, though promising, confirm the need to consider different recording conditions, due to the great number of factors that can influence the channel.

2. EMULATED TELEPHONE CONDITIONS

Having analyzed the problems raised by noise and channel variations described in [7] we started with the same approach of data simulation and applied it to ASVspoof 2015 and 2017 databases. We considered several approaches for channel simulation: down sampling to 8kHz and software codec G.6.10 [9] with 13kbit/s bitrate to emulate lossy speech compression in cellular telephony without package loss (we will refer to both these emulations as Emulation condition 1 - E1) and STC-H219 [10] sound device for recording analogue signals sent through 2 meters telephone cable (Emulation condition 2 - E2). However, experiments conducted with different telephone channel simulations show that systems trained with one type of emulated data demonstrate low quality on the other emulated and real data. Due to this the new database of spoofing attacks PHONESPOOF was collected.

3. REAL TELEPHONE SPOOFING DATA COLLECTION

The real spoofing database was collected by recording over the telephone channel items from the two datasets of ASVspoof challenges. At present these are the most various publicly available spoofing databases in microphone channel in terms of recording conditions and attack types.

In order to cover a great variety of up-to-date high quality spoofing techniques we also collected subsets of the most popular TTS samples, created by cloud services and available libraries: Google [11], Yandex [12], IBM [13], Lyrebird [14], Zamzar [15], Ispeech [16] and STC [17]. When possible, we generated records for English and Russian language. The datasets were recorded in two different scenarios (Figure 1). The first one used the ”Smart Logger II” (Recording scenario 1 - R1) sound system for telephone calls and speech messages registration [18], the second one used the ”Smart Caller” (Recording scenario 2 - R2) system of voice notification via telephone lines [19]. In both cases audio signals with spoofing trials were replayed on the computer and transferred to a mobile phone by 3Jack-4Jack cable.At the same time the mobile phone was calling on the line phone connected to the ”Smart Logger II” system or to one of the channels of ”Smart Caller” system. In both cases these systems recorded the input signal that had gone through all needed channels and codecs, thus obtaining spoofing attack trials in telephone channel. In order to provide recording conditions variability 2 mobile phones (Samsung Galaxy Note II, Xiaomi Redmi 4A) were used with 2 different telecommunications operators. Table 1 presents the amount of collected spoofing trials.

Genuine samples recorded in the described way could not be considered as genuine samples any more and were used as high quality replay attacks. English part of NIST SRE speech dataset was used [20] as a set of real genuine samples. Optionally we used a Russian speech subcorpus RusTelecom to extend the training set. RusTelecom is a Russian speech corpus of telephone data, collected by call-centers. RusTelecom database consists of approximately 4500 sessions. As additional dataset in Russian language we collected 3000 internal calls, recorded according to Recording scenario 1 - CallDB.

Table 1. Total duration of collected spoofing trials in microphone and telephone channels in hours

		Microphone channel	Recorded R1	Recorded R2
ASVspoof2015$_{sp}$	eng	177.1	361.6	856.6
ASVspoof2015$_{g}$	eng	160.2	51.2	-
iSpeech	eng	5.47	-	-
IBM	eng	203.4	-	19.9
	rus	217.0	-	15.1
Zamzar	eng	210.8	-	-
STC	rus	523.1	-	48.4
Yandex	eng	236.3	6.2	53.4
	rus	201.7	3.3	56.8
Google	eng	314.8	6.6	46.0
	rus	240.6	3.3	48.7
Lyrebird	eng	95.9	-	1.64
RSR_phrases	eng	150.7	-	29.9
RedDots2015	eng	142.8	-	30

4. EXPERIMENT RESULTS AND DISCUSSION

The aim of this work was to investigate the possibility to implement a single system able to detect different types of spoofing attacks. In all further experiments for training we used only data recorded with the use of the original training parts of the datasets, and for evaluation the corresponding evaluation part, so that there was no overlap between training and evaluation sets. For TTS datasets, collected during this work, we used different literary text to produce synthesis. For example, to create a training set by Google TTS in english we used "Jane Eyre" novel, while "Harry Potter and the Deathly Hallows" was used to create the evaluation set.

Table 2. Experiment results for CQC system [7], EER(%).

Emulation type	original	8kHz	6.10 codec
ASVspoof2015	2.24	45.46	46.35
ASVspoof2017	49.18	50.00	49.98

As mentioned above, we started our investigation with experiments produced on the emulated data with CQCC based system from [7]. Our results show that such system lacks appropriate quality for replay attacks (Table 2). Therefore we concentrated on deep learning approaches. We considered a system based on LCNN architecture [8] that is illustrated in Figure 2. All experiment results obtained with the use of emulated data confirm that different channel distortions have dramatic impact on the spoofing detection accuracy. This can be seen even for different types of emulation. Table 3 shows results for LCNN system trained on different emulations.

The system trained on the emulated data of ASVspoof 2015, Google and Yandex TTS shows very high quality detection of Ispeech and Zamzar synthesis records in emulated channel. Due to this we concluded that in terms of ability to spoof the ASV system these types of TTS are not critical. In further research we did not consider these types of spoofing in real channel case, assuming to have the same trend. Our current results confirm that it is highly recommended to use data recorded in the same telephone channel that will be used with the implemented system. Results in Table 5 demonstrate the significant improvement in detection of spoofing attacks from ASVspoof2015 on the base recorded in R1 after adding the corresponding subset to the training set.

Table 3. Experiment results, EER(%). $ASVspoof2015_{g}$ and $ASVspoof2015_{sp}$ refer to genuine and spoofing parts of the ASVspoof dataset respectively

Training set	Evaluation set	EER (%)
genuine: - NIST - RusTelecom - $ASVspoof2015_{g}$ E1 **spoof**: - Google (eng + rus) E1 - Yandex (eng + rus) E1 - $ASVspoof2015_{sp}$ E1	**genuine**: - $ASVspoof2015_{g}$ E2 **spoof**: -$ASVspoof2015_{sp}$ E2	10.98
genuine: - $ASVspoof2015_{g}$ E2 **spoof**: - $ASVspoof2015_{sp}$ E2	**genuine**: - $ASVspoof2015_{g}$ E2 **spoof**: -$ASVspoof2015_{sp}$ E2	7.97
	genuine: -NIST -$ASVspoof2015_{sp}$ R1	26.85
genuine: - NIST **spoof**: - $ASVspoof2015_{sp}$ R1	**genuine**: - $ASVspoof2015_{g}$ E2 **spoof**: -$ASVspoof2015_{sp}$ E2	49.90

To estimate the impact of the language dependency we also trained our system with genuine and spoof trials in Russian language. Experiment results show that including Google and Yandex spoofing samples together with genuine subset in the target language to the training set improves detection of these spoofing attacks from 5.52% EER to 0.51% EER (Table 4) on it. Additionally it enhances the performance on the database in the original language used for training. It is highly important to mention that adding samples in a new language only to one class can lead to situation when the system uses language specific features as a spoofing criterion and should be avoided.

The final version of the proposed system was trained with the following datasets: $ASVspoof2015_{sp}$, recorded in both scenarios, Google and Yandex datasets, recorded by Smart Caller, and all available emulations of ASVspoof2015, Google, Yandex, Lyrebird and STC datasets in order to avoid channel overfitting. Results for different types and recording conditions are presented in Table 6. It should be noted that although the EER for each spoofing type are significantly low, the thresholds for each of them differs. Due to this we cannot

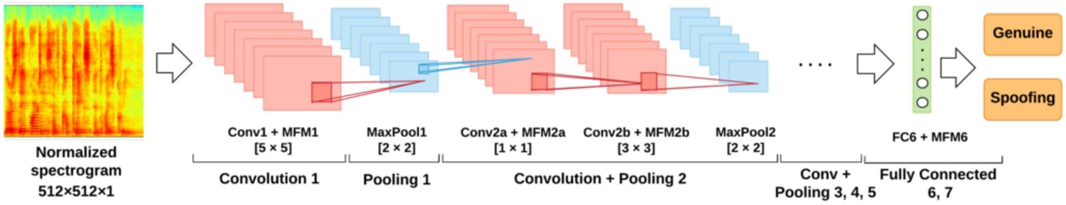

Fig. 2. LCNN architecture

Table 4. Experiment results for different languages

Training set	Evaluation set	EER (%)
genuine: -NIST -$ASVspoof_g$ E1 **spoof**: -$ASVspoof2015_{sp}$ E1 -Google (eng) E1 -Yandex(eng) E1	**genuine**: -RusTelecom **spoof**: -Google(rus) E1 -Yandex(rus) E1	5.52
	genuine: -NIST **spoof**: -Google(eng) E1	0.03
	genuine: -NIST **spoof**: -Yandex(eng) E1	0.4
genuine: -NIST -$ASVspoof2015_g$ E1 - RusTelecom **spoof**: -$ASVspoof2015_{sp}$ E1 - Google (eng + rus) E1 -Yandex(eng + rus) E1	**genuine**: -RusTelecom **spoof**: -Google(rus) E1 -Yandex(rus) E1	0.51
	genuine: -NIST **spoof**: -Google(eng) E1	0.03
	genuine: -NIST **spoof**: -Yandex(eng) E1	0.14

state that our ultimate goal has been achieved. However, this investigation confirms the high efficiency of deep learning approaches for spoofing detection task in telephone channel and determines the agenda for further research.

5. CONCLUSIONS

In the paper we presented PHONESPOOF data collection - audio spoofing attacks data collected through real telephone channels. Based on the collected data we investigated robustness of the state-of-the-art deep learning based antispoofing systems under telephone spoofing attacks conditions. During the investigations we approved that regular telephone channel emulation does not quite match the realistic telephone spoofing attacks scenario which is highly important for the developing of antispoofing systems suitable for real applications. We tested a single unified system, based on the LCNN approach, for the ability to detect different types of spoofing attacks like voice conversion, speech synthesis and replay. We also addressed to the language dependency issue and found out that adding target language to the training set enhance spoofing detection performance on this language. Our experiments conducted on the PHONESPOOF data confirmed effectivness of deep learning frameworks for solving the considered task and highlight several points that should be taken into account in future work.

Table 5. Experiment results for $LCNN$ based system

Training set	Evaluation set	EER (%)
genuine: -NIST -RusTelecom **spoof**: -$ASVspoof2015_{sp}$ R1 -Google R1 (eng + rus) -Yandex R1 (eng + rus)	**genuine**: -CallDB **spoof**: -$ASVspoof2015_{sp}$ R1, R2	25
genuine: -NIST -RusTelecom **spoof**: -$ASVspoof2015_{sp}$ R1, R2 -Google (eng + rus) R1 -Yandex (eng + rus) R1	**genuine**: -CallDB **spoof**: -$ASVspoof2015_{sp}$ R1, R2	1.93
	genuine: -NIST **spoof**: -$ASVspoof2015_{sp}$ R1	2.87
	genuine: -NIST **spoof**: -$ASVspoof2015_{sp}$ R2	0.62
	genuine: -CallDB **spoof**: -$ASVspoof2015_{sp}$ R1	4.96

Table 6. Experiment results for different spoofing types, EER(%)

ASVspoof2015 R1		ASVspoof2015 R2	
TTS	VC	TTS	VC
2.74	3.00	0.97	1.27

Google R2		Yandex R2		IBM R2		Replay R2
Eng	Rus	Eng	Rus	Eng	Rus	
1.88	0.86	0.20	1.49	2.45	3.16	1.77

6. REFERENCES

[1] *ISO/IEC 30107-1:2016(en) Information technology Biometric presentation attack detection Part 1: Framework*, Geneva, Switzerland, 2016.

[2] A. Kaliyev, Y. Matveev, E. Lyakso, and S. Rybin, "Prosodic processing for the automatic synthesis of emotional russian speech," *Proceedings of the 2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies" (ITQMIS)*, 2018.

[3] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA 2010*, 2010, pp. 131–134.

[4] 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, pp. 130 – 153, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639314000788

[5] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "End-to-end convolutional neural network-based voice presentation attack detection," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Oct 2017, pp. 335–341.

[6] C. Hanili, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Communication*, vol. 85, pp. 83 – 97, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316300681

[7] H. Delgado, M. Todisco, N. Evans, M. Sahidullah, W. M. Liu, F. Alegre, T. Kinnunen, and B. Fauve, "Impact of bandwidth and channel variation on presentation attack detection for speaker verification," in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sept 2017, pp. 1–6.

[8] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. Interspeech 2017*, 2017, pp. 82–86. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-360

[9] "RESMG-110610PR2. ETS 300 961 (GSM 06.10) european standard." [Online]. Available: https://portal.etsi.org/webapp/workprogram/Report_WorkItem.asp?WKI_ID=11074

[10] "STC H219 overview." [Online]. Available: http://speechpro.com/product/voice-recording/smartlogger2#tab4

[11] "Google cloud speech API." [Online]. Available: https://cloud.google.com/speech/

[12] "Yandex speech kit." [Online]. Available: https://tech.yandex.ru/speechkit/

[13] "IBM text-to-speech." [Online]. Available: https://www.research.ibm.com/tts/

[14] "Beta version of lyrbird text-to-speech." [Online]. Available: https://lyrebird.ai/

[15] "Zamzar text-to-speech." [Online]. Available: https://www.zamzar.com/

[16] "Text-to-speech API from ispeech." [Online]. Available: http://www.ispeech.org/api/#text-to-speech

[17] P. Chistikov and E. Korolkov, "Data-driven speech parameter generation for russian text-to-speech system," *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference "Dialogue"*, vol. 1, pp. 103–111, 2012.

[18] "Smart logger II: Multi-channel call recording and monitoring system." [Online]. Available: http://speechpro.com/product/voice-recording/smartlogger2

[19] "Smart caller: Automatic 24/7 notification of subscribers." [Online]. Available: http://speechpro.com/product/notification/smartcaller

[20] "NIST speaker recognition evaluation 2016," in *Computational Linguistics and Intellectual Technologies*, 2016. [Online]. Available: https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016