INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

STC Antispoofing Systems for the ASVspoof2019 Challenge

Galina Lavrentyeva[1,2], *Sergey Novoselov*[2], *Andzhukaev Tseren*[1], *Marina Volkova*[1], *Artem Gorlanov*[1], *Alexandr Kozlov*[1]

[1]STC-innovations Ltd., St.Petersburg, Russia
[2]ITMO University, St.Petersburg, Russia

{lavrentyeva, novoselov, andzhukaev, volkova, gorlanov, kozlov-a}@speechpro.com

Abstract

This paper describes the Speech Technology Center (STC) antispoofing systems submitted to the ASVspoof 2019 challenge. The ASVspoof2019 is the extended version of the previous challenges and includes 2 evaluation conditions: logical access use-case scenario with speech synthesis and voice conversion attack types and physical access use-case scenario with replay attacks. During the challenge we developed anti-spoofing solutions for both scenarios. The proposed systems are implemented using deep learning approach and are based on different types of acoustic features. We enhanced Light CNN architecture previously considered by the authors for replay attacks detection and which performed high spoofing detection quality during the ASVspoof2017 challenge. In particular here we investigate the efficiency of angular margin based softmax activation for training robust deep Light CNN classifier to solve the mentioned-above tasks. Submitted systems achieved EER of 1.86% in logical access scenario and 0.54% in physical access scenario on the evaluation part of the Challenge corpora. High performance obtained for the unknown types of spoofing attacks demonstrates the stability of the offered approach in both evaluation conditions.

Index Terms: spoofing, anti-spoofing, speaker recognition, replay attack, speech synthesis, voice conversion, ASVspoof2019

1. Introduction

Over the past few years, voice biometric technologies have reached impressive performance, which can be confirmed by the results of the NIST Speaker Recognition Evaluation (SRE) Challenges [1]. Automatic Speaker Verification (ASV) systems are already used in security systems of socially significant institutions, in immigration control, forensic laboratories and for identity verification in Internet banking, and other electronic commerce systems.

Alongside the increasing performance and confidence in speaker recognition methods, the privacy level of the information with the necessity to protect it also increases. This leads to higher requirements for the reliability of the biometric systems including their robustness against malicious attacks. Active fraudster attempts to falsify voice characteristics in order to gain unauthorised access referred to as spoofing attacks or presentation attacks (ISO/IEC 30107-1) are the biggest threat for voice biometric systems. The widespread use of ASV systems and new approaches in machine learning has forced the significant quality improvement of these attacks. Many studies show that despite the high performance of the state-of-the-art ASV systems they are still vulnerable to spoofing and the need in reliable spoofing detection methods for ASV systems is apparent.

Automatic Speaker Verification Spoofing and Countermeasures initiative (ASVspoof) has attracted the high interest of the research community to the task of unforeseen spoofing trials detection. It has significantly pushed forward the development of spoofing detection methods by organizing ASVspoof Challenges in 2015 and 2017, that were aimed to develop countermeasures to detect speech synthesis with voice conversion attacks and replay attacks, respectively.

In 2019, the competition was held for the third time and was the extended version of the previous ones [2]. The task was to design the generalised countermeasures in 2 evaluation conditions: logical access use-case scenario with speech synthesis and voice conversion attack types and physical access use-case scenario with replay attacks.

For both scenarios, we proposed several systems based on the enhanced Light CNN architecture, considered by the authors for replay attacks detection in [3] and outperformed other proposed systems during ASVspoof2017 challenge. The proposed systems are based on different types of acoustic features.

This paper explores angular margin based softmax and batch normalization techniques for anti-spoofing systems quality improvements.

Section 2 describes the proposed modifications of the original LCNN-system for spoofing detection from [3] in details. Section 3 contains the overview of all proposed single and submitted systems, while in section 4 the results obtained for these systems on the development and evaluation parts are presented and analysed.

It is worth mentioning that according to the evaluation plan all data used for training and evaluation was modelled using acoustic replay simulation. On the one hand, this helps to carefully control acoustic and replay configurations, but on the other hand, results raise some doubts about the usability of the considered systems for real-case scenarios. According to our experiments performed for spoofing attacks in real and emulated telephone channel [4] systems trained for emulated conditions cannot detect spoofing attacks in real cases.

2. LCNN system modifications

All of the proposed systems for both scenarios were based on the enhanced Light CNN architecture previously used for replay attack detection [3]. The specific characteristic of Light CNN architecture [5] is the usage of the Max-Feature-Map activation (MFM) which is based on Max-Out activation function [6]. Neural network with MFM is capable to choosing features which are essential for task solving. According to impressive results obtained by the authors in [3] for replay attacks, such type of networks can be successfully implemented for anti-spoofing.

http://dx.doi.org/10.21437/Interspeech.2019-1768

2.1. Front-End

We explored several types of acoustic features as input for LCNN, all of them were used in a raw format.

Our experience in spoofing detection confirms that power spectrum contains useful information related to the speech signal and artifacts specific to different spoofing attacks and can be used as informative time-frequency representation for spoofing detection task. We used raw log power magnitude spectrum computed from the signal as features. For this purpose, the spectrum was extracted via:

- constant Q transform (CQT) [7]
- Fast Fourier Transform (FFT)
- Discrete Cosine Transform (DCT)

Additionally, we considered cepstral coefficients from baseline systems, proposed by the organisers of the ASVspoof2019: Linear Frequency Cepstral Coefficients (LFCC) [8] obtained by the use of triangular filters in linear space for local integration of the power spectrum and Constant Q Cepstral Coefficients based on the geometrically spaced filters [7] .We explored efficiency of using simple enegry based Speech Activity Detector (SAD) for solving spoofing detection task for both PA and LA attack types.

2.2. LCNN classifier

In contrast to our LCNN system presented in [3] for replay attacks detection, the proposed systems are used not as high-level features extractor, followed by GMM scoring model. Instead of that LCNN was used here for final score estimation based on the low-level acoustic features.

Additional steps of batch normalization were also used after MaxPooling layers to increase stability and convergence speed during the training process. The detailed architecture is described in Table 1.

2.3. Angular margin based softmax activation

The key difference of the novel LCNN system is angular margin based softmax loss (A-softmax) used for training the described architecture. A-softmax was introduced in [9] and demonstrated an elegant way to obtain well-regularized loss function by forcing learned features to be discriminative on a hypersphere manifold. Thus angular margin softmax loss can be described as:

$$L_{\text{ang}} = \frac{1}{N}\sum_i -\log\left(\frac{e^{\|x_i\|\cos(m\theta_{i,y_i})}}{e^{\|x_i\|\cos(m\theta_{i,y_i})} + \sum_{i\neq y_i} e^{\|x_i\|\cos(m\theta_{i,y_i})}}\right) \quad (1)$$

where N is the number of training samples $\{x_i\}_{i=1}^N$ and their labels $\{y_i\}_{i=1}^N$, θ_{i,y_i} is the angle between x_i and the corresponding column y_i of the fully connected classification layer weights W, and m is an integer that controls the size of an angular margin between classes.

This approach has already used in [10] for high-level speaker embedding extractor. The learned features are constrained to a unit hypersphere. Such regularization technique also addresses the problem of overfitting by separating classes in cosine similarity metric.

We use A-softmax as an effective discriminative objective for training our model.

LCNN weights were initialized using normal Kaiming initialization. And dropout 0.75 was used to reduce overfitting.

Table 1: *LCNN architecture*

Type	Filter / Stride	Output	Params
Conv_1	5×5 / 1×1	$863\times600\times64$	1.6K
MFM_2	–	$864\times600\times32$	–
MaxPool_3	2×2 / 2×2	$431\times300\times32$	–
Conv_4	1×1 / 1×1	$431\times300\times64$	2.1K
MFM_5	–	$431\times300\times32$	–
BatchNorm_6	–	$431\times300\times32$	–
Conv_7	3×3 / 1×1	$431\times300\times96$	27.7K
MFM_8	–	$431\times300\times48$	–
MaxPool_9	2×2 / 2×2	$215\times150\times48$	–
BatchNorm_10	–	$215\times150\times48$	–
Conv_11	1×1 / 1×1	$215\times150\times96$	4.7K
MFM_12	–	$215\times150\times48$	–
BatchNorm_13	–	$215\times150\times48$	–
Conv_14	3×3 / 1×1	$215\times150\times128$	55.4K
MFM_15	–	$215\times150\times64$	–
MaxPool_16	2×2 / 2×2	$107\times75\times64$	–
Conv_17	1×1 / 1×1	$107\times75\times128$	8.3K
MFM_18	–	$107\times75\times64$	–
BatchNorm_19	–	$107\times75\times64$	–
Conv_20	3×3 / 1×1	$107\times75\times64$	36.9K
MFM_21	–	$107\times75\times32$	–
BatchNorm_22	–	$107\times75\times32$	–
Conv_23	1×1 / 1×1	$107\times75\times64$	2.1K
MFM_24	–	$107\times75\times32$	–
BatchNorm_25	–	$107\times75\times32$	–
Conv_26	3×3 / 1×1	$107\times75\times64$	18.5K
MFM_27	–	$107\times75\times32$	–
MaxPool_28	2×2 / 2×2	$53\times37\times32$	–
FC_29	–	160	10.2 MM
MFM_30	–	80	–
BatchNorm_31	–	80	–
FC_32	–	2	64
Total	–	–	10.2MM

3. Experimental setup

3.1. Datasets

All experiments presented further were conducted on ASVspoof 2019 datasets. The detailed description of these datasets can be found in [2]. To train all the systems we used only the train part. The dev part was used for performance validation and weights adjustment for system fusion. The evaluation part includes a set of unseen genuine verification trials and spoofing attacks. These attacks were generated with unknown spoofing algorithms and replay configurations which differ from those in the train and development parts.

3.2. Details of systems implementation

We prepared several single systems for each scenario, based on the features described above and LCNN architecture from 1. For logical access scenario we used the following configurations:

- **LFCC-LCNN**: LFCC were extracted similar to baseline system with 20 ms window length, 512 number of FFT bins and 20 filters.
- **LFCC-CMVN-LCNN**: This system is similar to previous one. The only difference is that LFCC features were

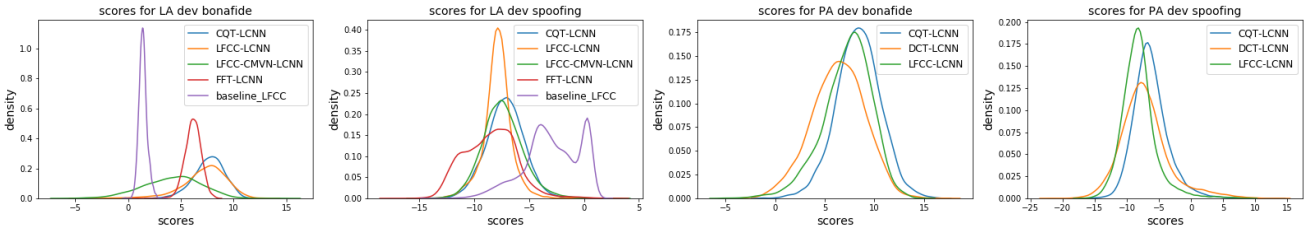

Figure 1: *Scores distributions for LA and PA systems for genuine and spoofing samples respectively.*

normalized by mean and variance.

- **CQT-LCNN**: CQT spectrum was extracted with the use of the default settings from baseline CQCC based system: 96 bins per octave, 1724 window size and 0.0081 step.
- **FFT-LCNN**: FFT spectrum was extracted with 1724 window length and step 0.0081, the Blackman window function was used.

For physical access scenario we used the following configurations:

- **LFCC-LCNN**: similar to LFCC-LCNN for LA scenario
- **CQT-LCNN**: similar to CQT-LCNN for LA scenario
- **DCT-LCNN**: For DCT spectrum the 863 window length and 0.0081 step were used.

Only the first 600 features for each file were used as LCNN input in all single systems. No additionally preprocessing techniques such as speech activity detection or dereverberation was explored in these systems.

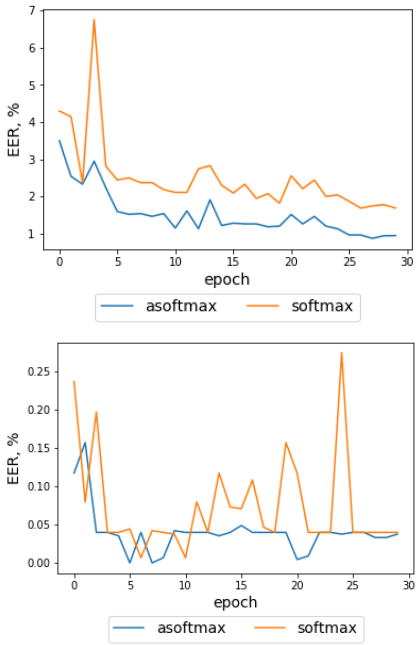

Figure 2: *EER during training process for PA CQT-LCNN system (top) and LA FFT-LCNN system (bottom)*

Table 2: *Performance of baseline systems and their modifications*

	LA		PA	
System	EER	min-tDCF	EER	min-tDCF
LFCC-GMM	3.029	0.078	11.226	0.241
LFCC-CMVN-GMM	6.000	0.153	16.686	0.345
LFCC-VAD-GMM	7.181	0.185	15.503	0.337
CQCC-GMM	0.473	0.014	10.072	0.194
CQCC-CMVN-GMM	3.095	0.086	13.000	0.267
CQCC-VAD-GMM	3.571	0.108	10.144	0.204

3.3. Submission systems

The primary systems submitted to the challenge were the fusion of the single systems on the score level. Fusion of the subsystems scores was done with equal weights. Before fusion scores were normalized by the standard deviation of the genuine class distribution for each single system. The reason for that was the Gaussian distribution of genuine scores in contrast to spoofing scores, (see Figure 1 for LA and PA systems scores distributions).

4. Results and Discussion

The results for our single and fusion systems are presented in terms of Equal Error Rate (EER) and minimum tandem detection cost function (min-tDCF) used as the primary metric in the Challenge.

Results obtained on the development and evaluation sets of the ASVspoof2019 dataset confirm the efficiency of deep learning approaches for the ASV spoofing detection tasks considered in the ASVSpoof2019.

Results of the preliminary investigations of the baseline systems [2], presented in Table 2, demonstrate that the use of SAD leads to the quality reduction in terms of EER and min-tDCF for LFCC and CQCC based systems in both LA and PA scenarios. The possible reason for this is that nonspeech and boundary regions contain discriminative features and distortions specific to various types of spoofing or genuine speech in the opposite. For example, constant energy values in concrete frequency regions, specific to some microphones or recording systems. For this reason, we decided to exclude SAD from the systems we used in the Challenge.

Curious, that cepstral mean and variance (cmvn) features normalisation didn't provide an expected quality improvement on the development set (See Table 2). This behaviour differs from the earlier experiments on ASVspoof2015 [11] and ASVspoof2017 [12] datasets, that did not contain artificially produced data. [13]. Taking into account our experience in spoofing detection in unforeseen conditions we assume that cmvn can increase the robustness of our systems against un-

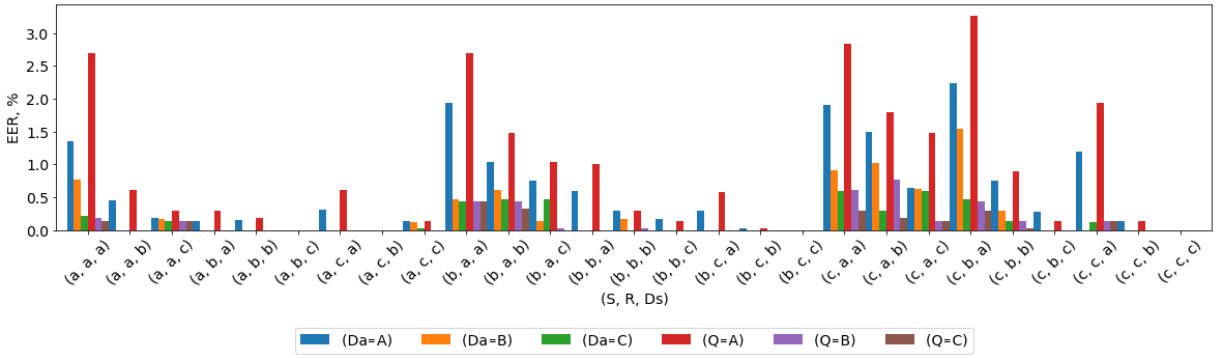

Figure 3: *Performance of the primary PA system pooled by PA spoofing attack types from the evaluation set. D_a relates to distance to a talker at which the replay attack is recorded, Q reletes to loudspeaker quality, S, R, D_s relates to (room size, reverberation and talker to ASV system distance)*

known attacks from the evaluation set. However, according to results, of our single systems on the development and evaluation systems in Table 3, we see the opposite. Such results reinforce our concerns about modelled data and real case mismatch.

Experiment results for deep learning based systems, proposed in the current paper, prove that implementation of angular margin based softmax loss for spoofing detection system training allows to improve system quality and stabilize training process (see Figure 2) for both LA and PA scenarios.

Experiments, conducted on the development part of ASVspoof2019 corpora, confirm that batch normalization and angular margin based softmax activation improve the performance of the original LCNN system for different types of low-level acoustic features in both scenarios (Figure 2).

Table 3 and Table 4 present the performance of all single systems proposed for LA and PA respectively. High performance obtained for the unknown types of spoofing attacks performed on the evaluation part of ASVspoof2019 corpora demonstrates the stability of the offered approach in both evaluation conditions.

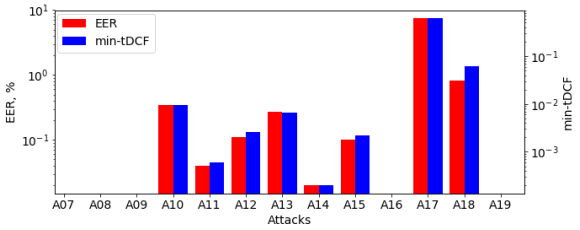

Figure 4: *Performance of the primary LA system pooled by LA spoofing attack types from the evaluation set*

Detailed analysis of our LA final system quality for different types of logical attacks, that are presented in Figure 4 demonstrates that it degrades in case of some unknown types of spoofing attacks (A10-A15, A17-A18) [2]. The most difficult spoofing attack to detect for our system was A17 (voice conversion with waveform filtering) task for our system.

Figure 3 illustrates the analysis of PA detection performance depended on the replay attack configuration: replay device quality, distances to the talker and to ASV system and reverberation characteristics. It can be concluded that replay attack detection performance depends on the replay attacks quality. The most high-quality attacks replay sessions recorded at a small distance to talker with the use of high-quality loudspeaker.

Table 3: *Results for submitted LA systems*

	dev		eval	
System	min-tDCF	EER	min-tDCF	EER
LFCC-LCNN	0.0043	0.157	**0.1000**	5.06
LFCC-CMVN-LCNN	0.0370	1.174	0.1827	7.86
CQT-LCNN	0.0000	0.000	-	-
FFT-LCNN	0.0009	0.040	0.1028	**4.53**
baseline_LFCC	0.069	2.7060	0.2120	8.09
Fusion	0.0000	0.000	**0.0510**	**1.84**

Table 4: *Results for submitted PA systems*

	dev		eval	
System	min-tDCF	EER	min-tDCF	EER
CQT-LCNN	0.0197	0.800	**0.0295**	**1.23**
LFCC-LCNN	0.0320	1.311	0.1053	4.60
DCT-LCNN	0.0732	3.850	0.560	2.06
Fusion	0.0001	0.0154	**0.0122**	**0.54**

5. Conclusion

This paper describes STC systems submitted to the ASVspoof2019 Challenge for LA and PA evaluation conditions. The main difference from the previous ASVspoof challenges is that all data used for training and evaluation was modelled using acoustic replay simulation. In our opinion, this deals with some restrictions from the practical point of view. According to the results obtained on the evaluation part of ASVspoof2019 corpora, the proposed LCNN based systems perform well in both PA and LA cases. Submitted systems achieved EER of 1.86% in LA scenario and 0.54% in PA scenario for unknown types of attacks.

6. Acknowledgements

This work was partially financially supported by the Government of the Russian Federation (Grant 08-08) and by the Foundation NTI (contract 20/18gr) ID 0000000007418QR20002.

7. References

[1] NIST speaker recognition evaluation 2018. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-2018-speaker-recognition-evaluation

[2] ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. [Online]. Available: http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf

[3] 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

[4] G. Lavrentyeva, S. Novoselov, M. Volkova, Y. Matveev, and M. De Marsico, "Phonespoof: A new dataset for spoofing attack detection in telephone channel," in *Proc. ICASSP 2018 (to be published)*, 2018.

[5] X. Wu, R. He, and Z. Sun, "A lightened CNN for deep face representation," *CoRR*, vol. abs/1511.02683, 2015. [Online]. Available: http://arxiv.org/abs/1511.02683

[6] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. C. Courville, and Y. Bengio, "Maxout networks," in *ICML*, 2013.

[7] M. Todisco, H. Delgado, and N. W. D. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey*, 2016.

[8] M. Sahidullah, T. Kinnunen, and C. Hanili, "A comparison of features for synthetic speech detection," 09 2015.

[9] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2017.

[10] S. Novoselov, A. Shulipa, I. Kremnev, A. Kozlov, and V. Shchemelinin, "On deep speaker embeddings for text-independent speaker recognition," 06 2018, pp. 378–385.

[11] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanili, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH*, 2015.

[12] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. W. D. Evans, J. Yamagishi, and K.-A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, 2017.

[13] H. Delgado, M. Todisco, M. Sahidullah, N. W. D. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements," 2018.