SHNU Anti-spoofing Systems for ASVspoof 2019 Challenge

Zhimin Feng, Qiqi Tong, Yanhua Long*, Shuang Wei, Chunxia Yang and Qiaozheng Zhang
Laboratory of Natural Human-Computer Interaction, Shanghai Normal University, Shanghai, China
E-mail: {fzm201612138,toqi877}@163.com, {yanhua,weishuang,chunxiay,zhangqz}@shnu.edu.cn
* Corresponding author: Yanhua Long (yanhua@shnu.edu.cn)

Abstract—**This paper presents an experimental analysis of SHNU anti-spoofing systems for the ASVspoof 2019 challenge. This challenge focused on countermeasures for three major attack types, namely those stemming from the advanced technology of TTS, VC and replay spoofing attacks. According to the type of attacks, the challenge was divided into two independent sub-challenges, the logical access (LA) and physical access (PA). Results of different anti-spoofing technologies on both sub-challenges were reported. Furthermore, the same countermeasures were also evaluated on two previous challenges, the ASVspoof 2015 and 2017. Experiments on cross-databases showed that, it appeared hard to generalize the classifiers trained from ASVspoof 2019 LA and PA databases to the previous challenges. The generalization ability of anti-spoofing technologies to different, new and unknown conditions was still very challenging. In addition, the effectiveness of different acoustic features were also examined and reported. Finally, we investigated the linear and an interfusing score-level fusion methods to individual systems to achieve better performance.**

Index Terms: anti-spoofing, logical access, physical access, LCNN, AFN, ASVspoof

I. Introduction

In the last decades, the automatic speaker verification (ASV) has attracted great research interests. Significant progress has been obtained in this field, and it has been shown to offer promising performance in many real-world applications, such as access control, e-commerce, telephone banking, etc [1], [2]. However, these real applications require the ASV systems to be robust against malicious attacks. Several recent studies have shown the vulnerability of ASV systems to different spoofing attacks without using any countermeasures [3], [4].

In the literature, the main types of spoofing attacks to an ASV system are the impersonation, voice conversion (VC), text-to-speech (TTS), and audio replay [4]. Impersonation generally requires experts to mimic a the voice of target speaker and hence there are limited data and research in the past. With the rapid development of VC and TTS technologies, the well-trained synthetic speech and converted voice is as good as perceptually indistinguishable from bona fide speech. They present a great threat to the ASV systems. Replay uses a pre-recorded speech to spoof the ASV system.

During the recent years, substantial progress has been achieved in the technology of spoofing speech detection, both for the logical access and physical access tasks. To detect the synthetic and voice converted speech, several new features have been proposed and proved to be very effective, such as the Constant Q Cepstral Coefficients (CQCC) [9], the Inverted Mel Frequency Cepstral Coefficients (IMFCC) features [10], the high-dimensional magnitude and phase features [11], the spoofing-vector [12], etc. Moreover, the adaptive weighting and clustering framework [8], combination of the linear and nonlinear classifiers [13] have also shown great improvements. To counteract the audio replay attacks, the best system submitted to the ASVspoof 2017 Challenge is the Light Convolutional Neural Networks (LCNN) anti-spoofing system [14]. The recent proposed Attentive Filtering Network (AFN) [15] has also been proved to be effective. Other researches focused on the new feature extraction and comparison, such as the source and instantaneous frequency and cepstral features (IFCC) [16], the linear frequency residual cepstral features[17], the variable length energy separation algorithm (VESA) based features [18] and different feature analysis[19], etc.

In this paper, we describes the Shanghai Normal University (SHNU) teams effort in participation of both the PA and LA sub-challenges in ASVspoof 2019 Challenge[1]. And moreover, we present an experimental study of different classifiers and features used for detecting both the LA and PA spoofing attacks. We employed the similar LCNN and AFN approaches proposed in [14], [15] as our anti-spoofing classifiers, including the Gaussian Mixture Models (GMMs) used in the official baselines. To assess the generalization ability of classifiers under mismatched data conditions, detail cross-database experiments were performed. In addition, preliminary experiments were performed to examine the behavior of different frequency bands in the acoustic feature of log power magnitude spectrum via Fast Fourier Transform [14]. Furthermore, the general linear and an interfusing nonlinear methods were investigated for system fusion.

The rest of the paper is organized as follows. The ASVspoof 2019 and previous two challenges are briefly introduced in section 2. The overview of single systems and score fusion approaches are described in section 3, followed by experimental results and discussions in section 4. Finally, conclusion and future works are presented in section 5.

[1] http://www.asvspoof.org/

II. Task Description and Baselines

Before the ASVspoof 2019 Challenge, two previous challenges have been held. The ASVspoof 2015 Challenge [5] was designed to find countermeasure solutions to classify the bona fide (genuine) speech and spoofed speech produced using either TTS or VC technologies. The ASVspoof 2017 Challenge [6] was designed to focus on the audio replay attack detection. And the ASVspoof 2019 Challenge [7] extended the previous challenges. It was the first challenge to focus on countermeasures for all the VC, TTS and replay attacks. It was divided into two sub-challenges. The logical access (LA) sub-challenge was designed to focus on countermeasures for the attacks stemming from up-to-date TTS and VC systems. This sub-challenge aimed to determine whether the advances in TTS and VC technology post a greater threat to ASV and the reliability of spoofing countermeasures. While the physical access (PA) sub-challenge focused on the replay spoofing countermeasures. Compared with the replay spoofing created from uncontrolled setup in ASVspoof 2017, the replay attacks in ASVspoof 2019 was simulated using a range of real replay devices and carefully controlled acoustic conditions. This sub-challenge aimed to provide a better assessment of replay spoofing countermeasures, and brought new insights into the replay spoofing problem.

The LA sub-challenge: Compared with the 10 spoofing types in ASVspoof 2015[5], the spoofed utterances of LA sub-challenge in ASVspoof 2019 were the well-trained synthetic speech and converted voice produced with today's technology[7]. They were now perceptually indistinguishable from bona fide speech. The greatly improved naturalness and speaker similarity of these utterances pose substantial threats to the reliability of ASV. This sub-challenge contained training, development and evaluation partitions. The genuine speech was collected from 107 speakers (46 male, 61 female) and with no significant channel or background noise effects. Spoofed speech was generated from the genuine data using a number of different spoofing algorithms. There was no speaker overlap across the three subsets regarding target speakers used in voice conversion or TTS adaptation Each spoofed utterance of training data was generated according to one of 2 voice conversion and 4 speech synthesis algorithms. Spoofed speech of development data was generated according to one of the same spoofing algorithms used to generate the training dataset. The spoofing algorithms used to create the evaluation dataset were variants of the spoofing algorithms used to create the development dataset.

The PA sub-challenge: The speech of replay attacks in ASVspoof 2017 was created from the real re-presentation and re-recording of a base corpus [20] in a somewhat uncontrolled setup. This practice setup made results somewhat difficult to analyze. However, in order to improve the 2017 dataset, the 2019 edition was based upon simulated and carefully controlled acoustic and replay configurations. For the PA sub-challenge, the training and development data was created according to a total of 27 different acoustic configurations. They comprised an exhaustive combination of 3 room sizes, 3 levels of reverberation and 3 speaker-to-ASV microphone distances. There were 9 different replay configurations, comprising the 3 categories of attacker-to-speaker recording distances, and 3 categories of loudspeaker quality. The training, development and evaluation data partitions were generated according to the same set of randomly selected acoustic and replay configurations.

Performance measurements: As a comparison, we also performed experiments on ASVspoof 2015 and 2017 to see the model generalization to unseen conditions. For more details of these tasks, the readers are refer to [5], [6] and the evaluation plan of ASVspoof 2019 Challenge [7]. ASVspoof 2019 was the first time to use a new ASV-centric metric in the form of the tandem decision cost function (t-DCF) [7]. The equal error rate (EER) used in previous challenges was retained as a secondary performance measure. However, in order to support applications beyond ASV and performance comparison between cross ASV challenges, we only used EER to measure our system performances in this study.

TABLE I
Official results of two baseline countermeasures for both LA and PA scenarios of ASVspoof 2019, in EER (%).

Task	Baseline system	Dev set	Eval set
LA	LFCC-GMM CQCC-GMM	2.71 0.43	8.09 9.57
PA	LFCC-GMM CQCC-GMM	11.96 9.87	13.54 11.04

Baselines: Two official baseline systems were provided. They were based on GMM classifier with two different acoustic features, the linear frequency cesptral coefficients [10] (LFCC) and CQCC. Table I showed the performance of these baseline systems trained on the ASVspoof 2019 training set and tested on the development and evaluation sets, for both the logical and physical access conditions. From the EER, it is clear to see that, the LA task is easier than the PA task.

III. Single Systems and Fusion Methods

Two single systems were used in our experiments, one was the system based on the Light Convolutional Neural Networks (LCNN), which was the best system submitted to the ASVspoof 2017 Challenge. The other was the system based on the recently proposed Attentive Filtering Network (AFN). We re-implemented the LCNN and AFN according to the specification described in [14], [15] and the released github repositories[2].

A. LCNN system

As the proposed deep learning framework in [14], we used the same normalized log power magnitude spectrum (logspec) obtained via Fast Fourier Transform (FFT) as the LCNN input

[2] https://github.com/azraelkuan/asvspoof2017, https://github.com/jefflai108/Attentive-Filtering-Network

acoustic features. To obtain an unified time-frequency (T-F) shape of input features. We truncated the normalized FFT spectrograms along the time axis with the size of $864 \times 400 \times 1$ as the input of the first convolution layer of LCNN. During this procedure, short files were extended by repeating their contents if necessary to match the required length.

TABLE II
THE STRUCTURE OF THE LCNN-4 MODEL.

Type	Filter Size/Stride	Output Size
Conv1	$5 \times 5/1, 2$	$864 \times 400 \times 64$
MFM1	-	$864 \times 400 \times 32$
Pool1	$2 \times 2/2$	$432 \times 200 \times 32$
Conv2	$3 \times 3/1, 1$	$432 \times 200 \times 96$
MFM2	-	$432 \times 200 \times 48$
Pool2	$2 \times 2/2$	$216 \times 100 \times 48$
Conv3	$3 \times 3/1, 1$	$216 \times 100 \times 128$
MFM3	-	$216 \times 100 \times 64$
Pool3	$2 \times 2/2$	$108 \times 50 \times 64$
Conv4	$3 \times 3/1, 1$	$108 \times 50 \times 64$
MFM4	-	$108 \times 50 \times 32$
Pool4	$2 \times 2/2$	$54 \times 25 \times 32$
FC1	-	512
MFM FC1	-	256

The LCNN classifier used for the spoofing detection was a reduced CNN architecture with Max-Feature Map activation (MFM). The MFM structure played a feature selector role in the LCNN, because it suppresses a neuron by a competitive relationship rather than the commonly threshold or bias used in Rectified Linear Unit function. Two LCNN architectures were investigated in our experiments, one was the LCNN-9, it was the same as the LCNN architecture of Table 1 in [14], but in our experiments, we set the outputs of FC6 layer (fully connected layer) in the LCNN to 256×2, instead of the 32×2. The other architecture was the LCNN-4, it was a revised version of the Light CNN-4 model used in work [21]. Details are shown in Table II.

B. AFN System

The Attentive Filtering Network (AFN) proposed in [15] was composed of an attention-based filtering (AF) mechanism and a classifier based on the Dilated Residual Network (DRN). The AF enhances feature representations in both the frequency and time domains prior to the DRN, and by including the dilation in convolution, the generalization ability of neural networks can be improved. The AFN input features were the same logspecs for the LCNN system, but with a fixed unified time-frequency map of 257×1091 in our experiments. Other details about the AFN can be found in [15].

C. System Fusion Methods

Our system fusion was performed on score-level in two ways. One was the simplest linear fusion strategy. If the scores of two sub-systems X and Y are L_x and L_y, then final score is $L_{fused} = \alpha \cdot L_x + (1 - \alpha) \cdot L_y$ with $0 \leq \alpha \leq 1$. The other was the method of interfusing the confused region scores (ICRS) proposed in our previous work [22]. Instead of fusing the scores for all the test trials, we combined the scores of sub-systems only at the confused score region using linear weights. This region was estimated through a development test set. In our experiments, the equal score fusion weights were used for both the linear and ICRS fusion. However, we used different confused score regions in ICRS fusion for LA and PA sub-challenges. All of the confused score regions were tuned on the development set, then these confused score regions were directly applied to the evaluation set.

IV. EXPERIMENTS

In this paper, we first focus our efforts on finding the behavior of single systems with the same model architecture, but trained for different anti-spoofing purpose. Then we examined two system fusion methods on both LA and PA sub-challenges. Moreover, we tested both the LCNN and AFN classifiers on three cross-databases, including validation experiments of logspec features with different frequency bands. The official development set was used for model selection and for tuning the combination weights for system fusion.

Since the logspec feature used in the LCNN systems is different from the one we used in AFN system, in this section, we use the “FFT” to represent the logspec feature used in LCNN, and the “AFFT” to represent the logspec feature used in AFN system. “Dev-LA”, “Dev-PA” refers to the development set, and “Eval-LA”, “Eval-PA” refers to the evaluation set for the LA and PA sub-challenges respectively in ASVspoof 2019.

A. Experiments for the LA sub-challenge

TABLE III
RESULTS FOR THE LA SUB-CHALLENGE OF ASVSPOOF 2019, IN EER (%).

System ID	Individual System	Dev-LA	Eval-LA
L0	LFCC-GMM	2.86	8.09
L1	CQCC-GMM	0.43	9.57
L2	FFT-LCNN-4	0.24	25.64
L3	FFT-LCNN-9	0.11	23.21
L4	AFFT-AFN	0.00	15.98
Linear Fusion System		**Dev-LA set**	**Eval-LA set**
L0+L1		0.08	6.36
L1+L3		0.08	7.60
(L0+L1)+L3		0.05	5.82
ICRS Fusion System		**Dev-LA set**	**Eval-LA set**
L0+L1		0.04	6.01
L1+L3 (Primary)		0.05	23.21
(L0+L1)+L3		0.00	18.37

Results for the LA sub-challenge of ASVspoof 2019 are shown in Table III. It is clear that our L2, L3 and L4 system outperformed the L0 and L1 baseline system significantly on the development set. However, the performance on the Eval-LA set was worse than two baselines. It seems that the FFT-LCNN-9 model overfited the training and development datasets. In addition, the AFFT-AFN system obtained zero errors on the Dev-LA set. This indicates that our LCNN-9 model was very sensitive to the training data and difficult to be generalized to new, unseen conditions. The excellent behavior on the Dev-LA set dues to the fact that the spoofed speech in both Dev-LA and training sets were generated from the same spoofing algorithms.

Moreover, we can see that the ICRS score fusion is better than the linear fusion on the Dev-LA set, however, bad results are obtained on the Eval-LA set. This dues to the fact that in the ICRS approach, the confused score region estimated on the Dev-LA set was [0.15, 0.85], then when we applied this score region to perform the ICRS on the Eval-LA set, there was only a few trials (around 100 trials) of the first systems (L0, L1, (L0+L1)) to be selected for score combination with the second systems (L1, L3, L3). However, in the linear fusion, we used equal weights to combined the scores of all the test trials. That's to say, our ICRS was also sensitive to the over-trained speaker models and it was also over-tuned on the development set of LA sub-challenge.

Submission to ASVspoof 2019: primary and single system: our primary system submitted to the LA sub-challenge was the combination of L1 and L3 using ICRS score fusion method. And we submitted the FFT-LCNN-9 (L3) as our single system because L4 system was trained after the submission deadline.

B. Experiments for the PA sub-challenge

TABLE IV
RESULTS FOR THE PA SUB-CHALLENGE OF ASVSPOOF 2019, IN EER (%).

System ID	Individual System	Dev-PA	Eval-PA
P0	FFT-LCNN-4	2.41	3.33
P1	FFT-LCNN-9	2.93	3.79
P2	AFFT-AFN	3.90	3.91
Linear Fusion system		**Dev-PA**	**Eval-PA**
P0+P1(Primary)		2.31	3.01
P1+P2		2.07	2.20
P0+P1+P2		1.65	2.19
ICRS Fusion system		**Dev-PA**	**Eval-PA**
P0+P1		2.16	3.03
P1+P2		1.85	2.17
P0+P1+P2		1.53	2.10

Results for the PA sub-challenge of ASVspoof 2019 are shown in Table IV. It can be seen that the P0, P1 and P2 system achieved much better results than the baseline systems shown in Table I. Such as, the P0 system achieved relative 75% EER reduction than the best CQCC-GMM baseline system. Furthermore, from both the linear and ICRS score fusion, we can see that the complementary information between P1 and P2 is bigger than the one between P1 and P0. It indicates that the LCNN-4 and LCNN-9 model learns almost the same thing with similar network architectures. And consistently, the ICRS score fusion was still a little bit better than the linear fusion on the development and evaluation sets for PA sub-challenge, even the confused score region tuned on the Dev-PA was [e-10, 0.99]. Moreover, different from the observation on LA sub-challenge, the performance gap between the Dev-PA and Eval-PA is very small. This may due to the small variations in acoustic and replay configuration for the training, development and evaluation datasets.

Submission to ASVspoof 2019: primary and single system: Since the AFFT-AFN system was finished after the submission deadline, for the PA sub-challenge, we submitted the FFT-LCNN-9 as the single system and the linear combination of P0 and P1 as our primary system.

C. Cross-database Experiments

Table V shows the results for LA scenarios on the ASVspoof 2015 and LA sub-challenge of ASVspoof 2019. The "ASV15" and "LA19" refer to the training data provided by these two challenges. "Dev-15" and "Eval-15" refer to the development and evaluation set under the required common condition in ASVspoof 2015. We are disappointed to find that the classifiers trained from the ASV15 and LA19 have no generalization ability to cross-database test sets, even they are all under LA scenarios. It indicates that both the LCNN and AFN models were over-trained. However, after combing the ASV15 and LA19 together to train the models, the performance on each self condition test set was still not improved, but the performance of cross-dataset test was significantly improved. This means that the countermeasures are very sensitive to the training data and difficult to be generalized well to unseen conditions.

TABLE V
EER% RESULTS FOR LA SCENARIOS, USING LCNN-9 AND AFN SYSTEMS.

System	Training data	Dev-15	Eval-15	Dev-LA	Eval-LA
FFT-LCNN-9	ASV15	**0.05**	**0.31**	55.82	61.34
	LA19	67.97	70.80	**0.11**	23.21
	LA19+ASV15	**0.08**	**1.92**	**0.19**	**15.24**
AFFT-AFN	ASV15	**0.06**	**4.20**	51.00	44.55
	LA19	46.87	46.06	**0.00**	15.98
	LA19+ASV15	**0.09**	**1.80**	**0.00**	**12.05**

Table VI shows the EERs of cross-database experiments for the PA scenarios in ASVspoof 2017 and PA sub-challenge of ASVspoof 2019. The "ASV17" and "PA19" refer to the training data provided by these two PA challenges. "Dev-17" and "Eval-17" refer to the development and evaluation set under the required common condition in ASVspoof 2017. As in Table V, the similar disappointing cross-database test observations were obtained in Table VI. However, the performance gains obtained from the training data combination are not significant as in Table V, although the EERs on Dev-PA and Eval-PA were reduced from 3.9% to 3.03% and 3.90% to 3.08%. This is because the training utterances of ASV17 was only around 6.3% of the training data in PA19, while the ratio of training data size between ASV15 and LA19 was more balanced.

Table VII shows the EERs for cross-database experiments in PA scenarios, using the FFT-LCNN-9 system. We found

TABLE VI
EER% FOR PA SCENARIOS, USING AFFT- AFN SYSTEM.

Training data	Dev-17	Eval-17	Dev-PA	Eval-PA
ASV17	**8.00**	**12.50**	51.00	45.11
PA19	57.00	60.00	**3.90**	3.91
PA19+ASV17	21.57	27.58	**3.03**	**3.08**

that the "FFT" features with frequency range of 0-4khz, 6k-8khz achieved the best results for the Dev-17 in ASVspoof 2017, and the cross-database test on Dev-17 and Eval-17 using models trained from PA19. However, the performance gains were not consistent when it was generalized to the Eval-17 and Dev-PA test sets. Furthermore, when we compared the 3rd line of Table VI and the last line of Table VII, it seems that the LCNN-9 model was more robust to the cross-database tests than the AFN model.

TABLE VII
EER% FOR PA SCENARIOS, USING FFT-LCNN-9 SYSTEM.

Training data	f_{min}-f_{max}	Dev-17	Eval-17	Dev-PA	Eval-PA
ASV17	0-8khz	**6.80**	**14.50**	45.00	48.23
	0-4khz	15.26	23.67	39.61	46.28
	6k-8khz	9.80	19.80	55.00	56.43
	0-4khz, 6k-8khz	**2.96**	**15.87**	64.20	68.70
PA19	0-8khz	28.00	45.00	**2.93**	**3.79**
	0-4khz, 6khz-8khz	25.26	38.83	**3.87**	**5.11**
PA19+ ASV17	0-8khz	10.26	23.00	4.55	5.78

V. Conclusion

In this paper, we have presented the anti-spoofing approaches used in SHNU system for the ASVspoof 2019 Challenge. Detail experimental results and analysis have been demonstrated, both for LA and PA sub-challenges and cross-database experiments. Our analysis showed that the classifiers based on LCNN and AFN were very vulnerable to unseen conditions. And they were very easily to be over-trained. From the cross-database experiments, we found that the most discriminative information contained in different frequency bands for the PA tasks in the ASVspoof 2017 and 2019 challenges. Moreover, results showed that the ICRS score fusion outperformed the linear fusion significantly, the complementary information between LCNN and AFN was very big. Future works will focus on improving the generalization ability of anti-spoofing countermeasures.

Acknowledgment

This work is funded by the Project 61701306 supported by National Natural Science Foundation of China.

References

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

[2] K.A. Lee, B. Ma and H. Li, "Speaker verification makes its debut in smartphone," *IEEE signal processing society speech and language technical committee newsletter*, 2013.

[3] P.D. Leon, M. Pucher, J. Yamagishi, I. Hernaez and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280-2290, 2012.

[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.

[5] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanili, M. Sahidullah and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in Proc. *INTERSPEECH*, September 6-10, Dresden, Germany, 2015, pp. 2037-2041.

[6] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi and K.A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in Proc. *INTERSPEECH*, August 20-24, Stockholm, Sweden, 2017, pp. 2-6.

[7] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen and K.A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection", *arXiv*, preprint arXiv:1904.05441.

[8] Y. Zhao, R. Togneri and V. Sreeram, "Spoofing Detection Using Adaptive Weighting Framework and Clustering Analysis," in Proc. *INTERSPEECH*, September 2-6, Hyderabad, India, 2018, pp. 626-630.

[9] M. Todisco, H. Delgado and N. Evans "Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification," *Computer, Speech and Language*, vol. 45, pp. 516-535, 2017.

[10] M. Sahidullah, T. Kinnunen and C. Hanilci, "A comparison of features for synthetic speech detection," in Proc. *INTERSPEECH*, September 6-10, Dresden, Germany, 2015, pp. 2087-2091.

[11] X. Xiao, X. Tian, S. Du, H. Xu, E.S. Chng and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in Proc. *INTERSPEECH*, September 6-10, Dresden, Germany, 2015, pp. 2052-2056.

[12] N. Chen, Y. Qian, H. Dinkel, B. Chen and K. Yu, "Robust deep feature for spoofing detection-The SJTU system for ASVspoof 2015 challenge," in Proc. *INTERSPEECH*, September 6-10, Dresden, Germany, 2015, pp. 2097-2101.

[13] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in Proc. *ICASSP*, March 20-25, Shanghai, China, 2016, pp. 5475-5479.

[14] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in Proc. *INTERSPEECH*, August 20-24, Stockholm, Sweden, 2017, pp. 82-86.

[15] C.I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak and S. King, "Attentive Filtering Networks for Audio Replay Attack Detection," in Proc. *ICASSP*, May 12-17, Brighton, United Kingdom, 2019, pp. 6316-6320.

[16] S. Jelil, R.K. Das, S.M. Prasanna, R. Sinha, "Spoof Detection Using Source, Instantaneous Frequency and Cepstral Features," in Proc. *INTERSPEECH*, August 20-24, Stockholm, Sweden, 2017, pp. 22-26.

[17] H. Tak and H.A. Patil, "Novel Linear Frequency Residual Cepstral Features for Replay Attack Detection," in Proc. *INTERSPEECH*, September 2-6, Hyderabad, India, 2018, pp. 726-730.

[18] M.R. Kamble and H.A. Patil, "Novel Variable Length Energy Separation Algorithm Using Instantaneous Amplitude Features for Replay Detection," in Proc. *INTERSPEECH*, September 2-6, Hyderabad, India, 2018, pp. 646-650.

[19] R. Font, J.M. Espn and M.J. Cano, "Experimental analysis of features for replay attack detection-Results on the ASVspoof 2017 Challenge," in Proc. *INTERSPEECH*, August 20-24, Stockholm, Sweden, 2017, pp. 7-11.

[20] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R.G. Hautamaki and et.al., "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in Proc. *ICASSP*, March 5-9, New Orleans, USA, 2017, pp. 5395-5399.

[21] X. Wu, R. He, Z. Sun and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 13, pp. 2884-2896, 2018.

[22] Y. Long, W. Guo and L. Dai, "Interfusing the confused region score of speaker verification systems," in Proc. *ISCSLP*, December 16-19, Kunming, China, 2008, pp.1-4.