ISSN 2278-3091

Volume 8, No.4, July – August 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse103842019.pdf

https://doi.org/10.30534/ijatcse/2019/103842019

Fusion of BiLSTM and GMM-UBM Systems for Audio Spoofing Detection

Ivan Rakhmanenko[1], Alexander Shelupanov[2], Evgeny Kostyuchenko[3]

[1]Tomsk State University of Control Systems and Radioelectronics (TUSUR), Tomsk, Russia, ria@keva.tusur.ru

[2]Tomsk State University of Control Systems and Radioelectronics (TUSUR), Tomsk, Russia, saa@keva.tusur.ru

[3]Tomsk State University of Control Systems and Radioelectronics (TUSUR), Tomsk, Russia, key@keva.tusur.ru

ABSTRACT

In this paper, we present our contribution to the ASVspoof 2019 challenge. The main task for this challenge is to find countermeasures that generalize well for different spoofing attacks against automatic speaker verification systems. Some of the approaches used by the authors during participation in the challenge are presented. Described anti-spoofing systems mostly rely on using constant Q cepstral coefficients (CQCC) features and bidirectional long-short term memory (BiLSTM) networks for genuine/spoof audio classification. Fusion of BiLSTM and GMM-UBM system is presented. This approach could give significant improvement to baseline systems results without any data augmentation, especially on physical access (PA) condition. Presented systems give 15.2% min-tDCF relative improvement for logical access (LA) condition and 61.5% min-tDCF relative improvement for PA condition, compared to the best baseline systems results.

Key words : ASVspoof, BiLSTM network, anti-spoofing, playback detection, synthetic speech detection, GMM-UBM.

1. INTRODUCTION

One of the main disadvantages of using automatic speaker verification (ASV) systems is its vulnerability to spoofing attacks. Despite the fact that state of the art ASV systems performs well and are robust to channel variations, different kind of spoofing attacks could significantly reduce ASV system accuracy [1]. Several types of ASV systems used in banking, access control systems, internet of things systems (IOT) [2, 3] could be affected by spoofing attacks [4].

There are several types of spoofing attacks: impersonation, voice conversion (VC), text-to-speech (TTS) and replay attacks. As modern TTS and VC systems made significant progress during last years, the ASVspoof 2019 challenge aims to find relevant countermeasures to such new spoofing methods. Another objective of this challenge is to assess impact of spoofing attacks upon the reliability of ASV systems. In order to have equal conditions for all challenge participants, organizers shared their ASV system scores so there was no need to use existing or create new ASV system. As opposed to the ASVspoof 2017 challenge, this challenge provide data recorded in controlled environment in order to make challenge results easier to analyze.

ASVspoof 2019 adopts for the first time a new ASV-centric metric in the form of the tandem decision cost function (t-DCF) [5]. By combining spoofing detection scores with ASV scores produced by a fixed system designed by the organisers, adoption of the t-DCF as the primary evaluation metric will ensure that evaluation results and rankings will reflect the impact of spoofing and the performance of spoofing countermeasures upon the reliability of ASV [6].

There are two problems in this challenge: synthetic and converted speech recognition, which forms logical access (LA) condition, and replay attack detection, which forms physical access (PA) condition. LA condition includes data processed with VC, TTS and hybrid TTS-VC systems. Part of the spoofed data was generated using voice conversion systems based on neural networks and spectral filtering functions [7]. Another part of spoofed data was processed using the speech synthesis systems that are based on waveform concatenation, waveform filtering, generative adversarial networks, neural-network-based parametric speech synthesis using source-filter vocoders and neural-network-based parametric speech synthesis using Wavenet [8-12].

Spoofing attacks in the PA condition correspond to replay attacks as at first bona fide speech is recorded by supposed intruder. Challenge organizers pre-processed spoofed audio in order to simulate capture and replay according to one of the scenarios. These scenarios include 27 different acoustic configurations, including three room sizes, three levels of reverberation, and three speaker-to-ASV microphone distances.

Most of the spoofing detection systems does not use traditional speech features. Systems of that kind typically use constant Q cepstral coefficient(CQCC) features [13], rectangular frequency cepstral coefficients (RFCC) [14], inverted mel frequency cepstral coefficients (IMFCC) [15] coupled with standard Gaussian Mixture Model (GMM) [16, 17], GMM-UBM, GMM-SVM, i-vectors backend [17]. There are some examples of using deep neural networks (DNN) [18] for spoof attacks detection [19, 20]. In addition, there are systems that includes BiLSTM network as a classifier [21, 22]. In this paper, we describe new methods of boosting BiLSTM networks performance for spoof attacks detection

The results were obtained as part of the implementation of the basic part of the state task of the Russian Federation Education and Science Ministry, project 8.9628.2017/8.9.

using different network training and evaluation approaches. The organisation of the paper is as follows. In Section 2, we present a short description of the proposed methods including features and systems we used. Results of the experimental evaluation and performance analysis are described in Section 3. Conclusions are presented in Section 4.

2. PROPOSED METHODS

In this section, we present components of spoofing countermeasures systems used in ASVspoof 2019 challenge. This includes features, classifiers and fusion details used in submitted systems.

2.1 Features

All of presented systems use constant Q cepstral coefficients (CQCC) as input features [23]. These features were extracted from training, development and test data using default parameters provided by organizers with mean and variance normalization using training set CQCC features distribution. Parameters for CQCC features extraction includes number of bins per octave B = 96, highest frequency to be analyzed fmax = 8000 Hz, lowest frequency to be analyzed fmin ≈ 16 Hz, number of uniform samples in the first octave d = 16, number of cepstral coefficients excluding zeroth coefficient cf = 29. Full feature vector includes 30 static CQCC features as well as 30 delta and 30 double delta coefficients, making total of 90 features.

Some experiments were conducted using other features such as linear frequency cepstral coefficients (LFCC), standard mel frequency cepstral coefficients (MFCC) in combination with voicing probability feature. Last set of features was obtained using OpenSMILE [24] library with avec2013 configuration file that is usually used for emotion recognition.

2.2 Classifiers

A. GMM-UBM system

GMM-UBM classifier is based on the principle of creating universal background model (UBM) which is a GMM trained on the all of the data available. This gives more details to be learned by GMM classifier and allows models for genuine and spoofed data to be more discriminative. Separated genuine and spoofed models are adapted from UBM model using MAP adaptation [25] with means, variance and weights adaptation. All GMM-UBM systems used in this work have 1024-component GMMs with relevance factor *r* = 10. Final score for this system is log-likelihood ratio *LLR* calculated using same approach as for baseline GMM system:

$$LLR = \log p(X \mid H_{genuine}) - \log p(X \mid H_{spoof}) \qquad (1)$$

where is the feature vector of evaluated audio, xi is feature vector of frame i, T is total number of a frames, Hgenuine corresponds to hypothesis that X is genuine audio segment, Hspoof corresponds to hypothesis that X is spoof audio segment. Each log-likelihood is calculated as average likelihood over each frame i:

$$\log p(X \mid H) = \frac{1}{T}\sum_{i=1}^{T} \log p(X_i \mid H) \qquad (2)$$

B. BiLSTM network

Given an input sequence $x = [x_1, x_2, ..., x_T]$ and the hidden vector $h = [h_1, h_2, ..., h_T]$, for a standard recurrent neural networks (RNNs), the output vector $y = [y_1, y_2, ..., y_T]$ can be computed from *t* = 1 to *T* according to the following iterative equations:

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \qquad (3)$$

$$y_t = W_{ht}h_t + b_y \qquad (4)$$

where *H* is the activation function of hidden layer, *W* is the weight matrix, and *b* is the bias vectors [22].

Bidirectional RNNs (BRNNs) were proposed to make full use of the context of feature sequences in both forward and backward directions [26]. Furthermore, an LSTM structure consists of memory blocks was proposed to learn the long-term dependencies [27]. Every block contains self-connected memory cells and three adaptive and multiplicative gate units i.e. input, output and forget gates. These gates can respectively provide write, read, reset operations for the cells. After combining the advantages of BRNN and LSTM, BiLSTM [28], designed as Figure 1, can deal with long-range context in both preceding and succeeding directions.

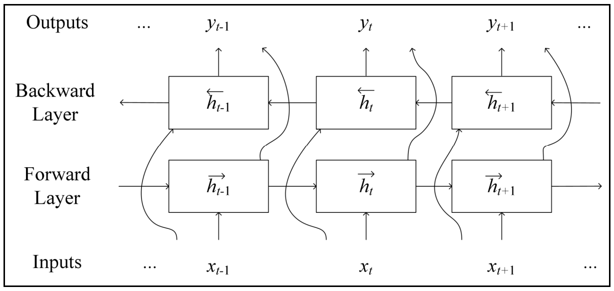

Figure 1: Structure of BiLSTM network module.

Since BiLSTM treats whole feature vector *X* as a sequence, it could find dependencies in time domain, as opposed to GMM classifier, that treats every feature vector in frames separately. This gives us opportunity to combine advantages of this two classifiers with making scores fusion.

BiLSTM network system used in both LA and PA conditions consists of two BiLSTM layers with dropout and a fully connected layer with softmax activation (Figure. 2).

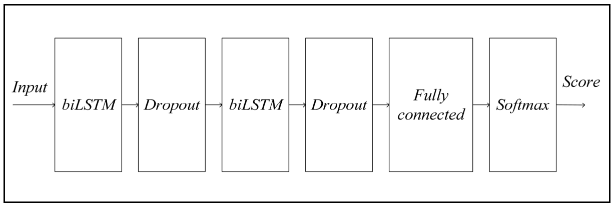

Figure 2: BiLSTM network system architecture.

2.3 Systems description

For both LA and PA conditions different types of GMM-UBM, BiLSTM network systems and scores fusion were used. Here we does not describe other systems structure and parameters because our team did not submit scores for these systems for ASVspoof 2019 Challenge.

2.4. LA condition systems description

There were three systems developed for LA condition: GMM-UBM system, BiLSTM system and system with score fusion from both of them.

A. GMM-UBM system (single system)

First system is GMM-UBM based system. Whole training set was used for UBM training. Three 1024-component GMMs were derived from this UBM: genuine model, spoof model and VC-1 model. VC-1 model was trained separately because bigger part of errors on development dataset was made only regarding this type of voice conversion system.

Final score (log-likelihood ratio) was calculated by scoring all three likelihoods from these GMMs using this algorithm:

```
Score1 = LLKgenuine – LLKspoof;
Score2 = LLKgenuine – LLKVC-1;
if Score1 > Θ
Final_Score = Score1;
else
Final_Score = Score2.
```

Threshold parameter *Θ* was determined using development dataset. This system is considered single because it is fast and uses only three GMM models.

B. BiLSTM system (Contrastive1)

This system consists of two layers of bi-directional long-short term memory units with dropout after each layer, fully connected layer (size = 2) and softmax layer. First layer consists of 100 BiLSTM units, second layer consist of 50 BiLSTM units, p(dropout) = 0.4. All training data was sorted in ascending order in terms of length before training.

Due to disbalance in training data (only 10 % is genuine), training batch with batch size bs = 64 was created with randomly choosing 30 % of spoof sequences and using all genuine speech. For this system training, we used only half of the training sequence skipping every 2-nd feature vector in full sequence.

Final score for this system was calculated using only one trained BiLSTM network, but with three different feature inputs. First input was full test sequence of CQCC vector, which gives output of the BiLSTM system - score1, second – sequence with skipping every 2nd feature vector (score2) and third – sequence with skipping every 2nd and 3rd feature vector (score3). Final score was calculated using this equation:

$$Final\_score = \frac{score_1 + score_3}{2} + score_2 \qquad (5)$$

C. Fusion system (Primary)

Fusion system makes fusion of GMM-UBM and BiLSTM systems scores using this equation:

$$Fusion\_score = score_{GMM-UBM} + \Theta_f * score_{biLSTM} \qquad (6)$$

where Θ_f is fusion weight parameter that was evaluated using development dataset.

2.5. PA condition systems description

Similarly to LA condition, in PA condition three systems were developed: BiLSTM system, GMM-UBM system and system with score fusion from both of them.

A. GMM-UBM system (single system)

Structure of BiLSTM system for PA condition is identical to the system used in LA condition. This system consists of two layers of bi-directional long-short term memory units with dropout after each layer, fully connected layer (size = 2) and softmax layer. First layer consists of 100 BiLSTM units, second layer consist of 50 BiLSTM units, p(dropout) = 0.4. All training data was sorted in ascending order in terms of length before training.

Due to disbalance in training data (only 10 % is genuine), training batch with batch size bs = 64 was created with randomly choosing 30 % of spoof sequences and using all genuine speech. Another aspect of this system is that training batcher used only half of full input training sequence skipping every 2-nd feature vector in sequnce.

Final score for this system was calculated using only one trained BiLSTM network and considered single. All features sequences were cut with skipping every 2nd feature vector as during the training stage. Final score for this system is the direct output from the trained network.

B. BiLSTM system (Contrastive1)

Contrastive1 system is GMM-UBM based system. Whole training set was used for UBM training. 13 GMMs were derived from 1024-component UBM: 3 genuine models (for *a, *b, *c record conditions), spoof model (all spoof data) and 9 for different attack ID models (AA,AB, …). UBM model was also used for scoring and considered as spoof.

Final score (log-likelihood ratio) was calculated by scoring all 14 log-likelihoods from these GMMs, selecting maximum scores of every genuine log-likelihoods and every spoof log-likelihoods using equation:

$$Final\_Score = \max(genuineLLKs) - \max(spoofLLKs) \qquad (7)$$

C. Fusion system (Primary)

Fusion system make fusion of GMM-UBM and BiLSTM systems scores using this equation:

$$Fusion\_score = score_{biLSTM} + \Theta_f * score_{GMM-UBM} \qquad (8)$$

where Θ_f is fusion weight that was calculated using development dataset.

3. EXPERIMENTAL EVALUATION

For experimental evaluation, only the data provided by the challenge organizers was used. This includes using official rules and protocols included in the challenge evaluation plan [6]. Some experimental results are not presented in this section. These results include experiments with convolutional DNN and i-vector systems using CQCC features as a frontend. These systems had very high error rates on development dataset and seemed to be undertrained.

3.1. LA condition results

LA condition systems evaluation results are shown in Table 1. *Baseline1* system is referred to baseline GMM-CQCC system and *Baseline2* system is referred to baseline GMM-LFCC system.

Table 1: LA condition systems evaluation results

System	**Dev**		**Eval**	
	m-tDCF	**EER**	**m-tDCF**	**EER**
Baseline1	0.0123	0.43	0.2366	9.57
Baseline2	0.0663	2.71	0.2116	8.09
GMM-UBM (single)	0.0007	0.04	0.4395	20.08
Primary (fusion)	**0.0002**	**0.03**	0.3856	15.32
BiLSTM (contrastive1)	0.0111	0.47	**0.1793**	**7.78**

As it could be seen from Table 1, our primary submission as well as single submission had highly overfitted on the training and development datasets. This overfitting is a result of assumption that VC systems generate audio with similar features to the VC-1 system. Thus, this assumption was wrong. However, contrastive1 submission (BiLSTM system) for this condition showed 15.2% *t-DCF* relative improvement to *baseline2* system. This gives us conclusion that BiLSTM network system generalises more than baseline GMM system for new VC and TTS spoofing attacks.

Analysing detailed contrastive1 system performance we could see that it recognises well TTS spoofing systems but struggles with VC systems. This gives *min-tDCF* = 0.9992 for waveform filtering VC system (A17) and *min-tDCF* = 0.7269 for vocoder VC system (A18).

Additional systems results for LA condition are shown in Table 2. It includes GMM-UBM system that uses only two GMMs for genuine and spoofed data, quadratic SVM classifier that uses avec2013 features as frontend, BiLSTM network, consisting of only one BiLSTM layer and GMM system that uses 14 MFCC features with delta and double delta coefficients combined with voicing probability (Vp) as a feature vector. We could see that there is small performance boost in *min-tDCF* when using GMM-UBM system instead of simple GMM system.

Table 2: LA condition additional systems development set evaluation results

System	**m-tDCF**	**EER**
Baseline1	0.0123	0.43
Baseline2	0.0663	2.71
GMM-UBM (CQCC)	**0.0085**	**0.43**
SVM (avec2013)	0.0223	0.78
BiLSTM, 1 layer	0.0245	0.79
GMM (MFCC+Vp)	0.0569	2.00

One of the assumptions while developing BiLSTM system was that skipping some of the features in input sequence could significantly affect systems performance. As it could be seen from Table 3, indeed using full input sequence could degrade this system performance. Best results for single input sequence were achieved when skipping every 2nd feature vector in sequence (BiLSTM, 1 skipped in Table 3).

Table 3: BiLSTM systems with different input time steps development set evaluation results

System	**m-tDCF**	**EER**
BiLSTM	0.0231	0.73
BiLSTM, 1 skipped	0.0126	**0.45**
BiLSTM, 2 skipped	0.0232	0.80
BiLSTM(contrastive1)	**0.0111**	0.47

Another approach that gave a little performance boost is using different input time steps for BiLSTM system that was trained on individual input time steps. This approach was used in *contastive1* system, where BiLSTM network was trained on sequence with skipping every 2nd feature vector. Nevertheless, on the evaluation stage this system was also scored with other input time steps sequences (5).

3.2. PA condition results

PA condition systems evaluation results are shown in Table 4. Baseline systems titles are the same as for the LA condition.

Table 4: PA condition systems evaluation results

System	**Dev**		**Eval**	
	m-tDCF	**EER**	**m-tDCF**	**EER**
Baseline1	0.1953	9.87	0.2454	11.04
Baseline2	0.2555	11.96	0.3017	13.54
GMM-UBM (single)	0.1109	4.07	0.1467	5.38
Primary (fusion)	**0.0944**	**3.57**	**0.1309**	**4.87**
BiLSTM (contrastive1)	0.1525	8.22	0.2653	12.32

As in the LA condition, BiLSTM system (single) also uses input features skipping to improve predictive capability of the model. We could see that BiLSTM system alone and as a part of the fusion system gives significant performance boost in comparison with the baseline systems results. Simple weighted scores fusion of the GMM-UBM and BiLSTM systems (8) gives 61.5% *min-tDCF* relative improvement for PA condition.

Analysing detailed *primary* system performance we could see that it does not recognise “BA” attack condition (50-100 cm attacker-to-talker distance, replay device quality is perfect) with environment id “caa” (10-20 m2 room size, 50-200 ms reverberation time, 10-50 cm talker-to-ASV distance), which gives *min-tDCF* = 1. In addition, there is performance loss in “BB” and “CA” attack conditions for “caa” and “cac” environment ids (*min-tDCF* = 0.49).

4. CONCLUSIONS

In this paper, we presented our contribution to the ASVspoof 2019 challenge. Described anti-spoofing systems mostly rely on using constant Q cepstral coefficients (CQCC) features and bidirectional long-short term memory (BiLSTM) networks for genuine/spoof audio classification. Fusion of BiLSTM and GMM-UBM system is presented. BiLSTM network could find dependencies in time domain, as opposed to GMM-UBM classifier that gives us more variability in systems decision methods.

Presented approach gives significant performance improvement compared to the baseline systems results without any data augmentation, especially on physical access (PA) condition. Presented systems give 15.2% *min-tDCF* relative improvement for logical access (LA) condition and 61.5% min-tDCF relative improvement for PA condition, compared to the best baseline systems results.

REFERENCES

1. 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. https://doi.org/10.1016/j.specom.2014.10.005
2. S.V.R.K.Rao, M.Saritha Devi, A.R.Kishore and Praveen Kumar **Wireless sensor Network based Industrial Automation using Internet of Things (IoT)**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, vol. 7, no. 6, pp. 82-86, 2018 https://doi.org/10.30534/ijatcse/2018/01762018
3. J. Sasi Bhanu, J.K.R. Sastry, P. Venkata Sunil Kumar, B. Venkata Sai, K.V. Sowmya **Enhancing performance of IoT networks through high performance computing,** *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, vol. 8, no. 3, pp. 432-442, 2019 https://doi.org/10.30534/ijatcse/2019/17832019
4. B. Usmonov, O. Evsutin, A. Iskhakov, A. Shelupanov, A. Iskhakova, and R. Meshcheryakov. **The cybersecurity in development of IoT embedded technologies**, in *2017 International Conference on Information Science and Communications Technologies (ICISCT)*, 2017, pp. 1–4. https://doi.org/10.1109/ICISCT.2017.8188589
5. T. Kinnunen, K. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds. t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification, in *Proc. Odyssey*, Les Sables d'Olonne, France, June 2018. https://doi.org/10.21437/Odyssey.2018-44
6. M. Todisko, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K.-A. Lee. **ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan**, 2019. [Online]. Availible: http://www.asvspoof.org/
7. D. Matrouf, J.-F. Bonastre, and C. Fredouille. **Effect of speech transformation on impostor acceptance**, in *Proc. ICASSP*, vol. 1. IEEE, 2006, pp. 933–936.
8. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan et al. **Natural tts synthesis by conditioning wavenet on mel spectrogram predictions**, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368
9. A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. **Wavenet: A generative model for raw audio**, *arXiv preprint* arXiv:1609.03499, 2016.
10. D. Griffin and J. Lim. **Signal estimation from modified shorttime Fourier transform**, *IEEE Trans*. ASSP, vol. 32, no. 2, pp. 236–243, 1984. https://doi.org/10.1109/TASSP.1984.1164317
11. K. Tanaka, T. Kaneko, N. Hojo, and H. Kameoka. **Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks**, in *Proc. SLT*, 2018, pp. 632–639. https://doi.org/10.1109/SLT.2018.8639636
12. K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. **Statistical singing voice conversion with direct waveform modification based on the spectrum differential**, in *Proc. Interspeech*, 2014, pp. 2514–2518.
13. M. Todisco, H. Delgado, and N. Evans. **Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification**, *Computer Speech Language*, vol. 45, pp. 516–535, 2017. https://doi.org/10.1016/j.csl.2017.01.001
14. P. Korshunov and S. Marcel. **Cross-database evaluation of audiobased spoofing detection systems**, in *Proc. INTERSPEECH*, 2016, pp. 1705–1709. https://doi.org/10.21437/Interspeech.2016-1326
15. P. Korshunov, S. Marcel, H. Muckenhirn, A. Gonc¸alves, A. S. Mello, R. V. Violato, F. Simoes, M. Neto, M. de Assis Angeloni, J. Stuchi et al. **Overview of BTAS 2016**

speaker anti-spoofing competition, in *Proc. BTAS*, 2016, pp. 1–6.
https://doi.org/10.1109/BTAS.2016.7791200

16. M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka. **Audio Replay Attack Detection Using High-Frequency Features**, in *Proc. INTERSPEECH*, 2017, pp. 27–31.
https://doi.org/10.21437/Interspeech.2017-776
17. Z. Ji, Z. Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao. **Ensemble Learning for Countermeasure of Audio Replay Spoofing Attack in ASVspoof2017**, in *Proc. INTERSPEECH*, 2017, pp. 87–91.
18. I. S. Kipyatkova and A. A. Karpov. **Variants of deep artificial neural networks for speech recognition systems**, *SPIIRAS Proceedings*, vol. 6, no. 49, pp. 80–103, 2016.
https://doi.org/10.15622/sp.49.5
19. P. Nagarsheth, E. Khoury, K. Patil, M. Garland. **Replay Attack Detection Using DNN for Channel Discrimination**, in *Proc. Interspeech*, 2017, pp. 97–101.
https://doi.org/10.21437/Interspeech.2017-1377
20. H. Dinkel, N. Chen, Y. Qian, and K. Yu. **End-to-end spoofing detection with raw waveform CLDNNS**, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4860–4864.
https://doi.org/10.1109/ICASSP.2017.7953080
21. K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala. **SFF Anti-Spoofer: IIIT-H Submission for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2017**, in *Proc. INTERSPEECH*, 2017, pp. 107–111.
https://doi.org/10.21437/Interspeech.2017-676
22. W. Cai, D. Cai, W. Liu, G. Li, and M. Li. **Countermeasures for Automatic Speaker Verification Replay Spoofing Attack: On Data Augmentation, Feature Representation, Classification and Fusion**, in *Proc. INTERSPEECH*, 2017, pp. 17–21.
23. M. Todisco, H. Delgado, and N. Evans. **A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients**, *in Odyssey 2016 - The Speaker and Language Recognition Workshop,* 2016.
24. F. Eyben, F. Weninger, F. Gross, and B. Schuller. **Recent developments in opensmile, the munich open-source multimedia feature extractor**, in *Proc. of the 21st ACM int. conf. on Multimedia*, 2013, pp. 835–838.
https://doi.org/10.1145/2502081.2502224
25. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. **Speaker verification using adapted gaussian mixture models**, *Digital signal processing,* vol. 10, no. 1, pp. 19–41, 2000.
26. M. Schuster and K. K. Paliwal. **Bidirectional recurrent neural networks**, *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
27. F. A. Gers, J. Schmidhuber, and F. Cummins. **Learning to forget: Continual prediction with lstm**, *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
28. A. Graves, N. Jaitly, and A. R. Mohamed. **Hybrid speech recognition with deep bidirectional lstm**, in *Proc. ASRU*, 2013, pp. 273–278.
https://doi.org/10.1109/ASRU.2013.6707742