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

The SJTU Robust Anti-spoofing System for the ASVspoof 2019 Challenge

Yexin Yang†, Hongji Wang†, Heinrich Dinkel, Zhengyang Chen, Shuai Wang, Yanmin Qian, Kai Yu

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

{yangyexin, jijijiang77, richman, feixiang121976, yanminqian, kai.yu}@sjtu.edu.cn,
chenzhengyang117@gmail.com

Abstract

The robustness of an anti-spoofing system is progressively more important in order to develop a reliable speaker verification system. Previous challenges and datasets mainly focus on a specific type of spoofing attacks. The ASVspoof 2019 edition is the first challenge to address two major spoofing types - logical and physical access. This paper presents the SJTU's submitted anti-spoofing system to the ASVspoof 2019 challenge. Log-CQT features are developed in conjunction with multi-layer convolutional neural networks for robust performance across both subtasks. CNNs with gradient linear units (GLU) activations are utilized for spoofing detection. The proposed system shows consistent performance improvement over all types of spoofing attacks. Our primary submissions achieve the 5^{th} and 8^{th} positions for the logical and physical access respectively. Moreover, our contrastive submission to the PA task exhibits better generalization compared to our primary submission, and achieves a comparable performance to the 3^{rd} position of the challenge.
Index Terms: anti-spoofing, spoofing detection, variational autoencoder, convolutional neural network

1. Introduction

As a convenient and reliable method for identity authentication, automatic speaker verification (ASV) [1] has attracted researchers' attention in recent years and gradually become mature, which makes it commercialized such as applications in call centers, security measures, etc. However, the ASV technologies are vulnerable, which makes ASV systems exposed to various spoofing attacks. Therefore, researchers manage to develop effective anti-spoofing systems, also known as presentation attack detection (PAD) systems, to protect ASV systems from malicious spoofing attacks.

At the beginning stage, researches were carried out in diverse datasets using different evaluation metrics, which made the results incomparable. In order to gather a community with standard databases and performance measures, a series of anti-spoofing competitions were born, for example, the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) challenges that serve as special sessions in INTERSPEECH 2013 [2], 2015 [3], 2017 [4] and 2019, respectively. ASVspoof 2013 aimed at raising this serious spoofing problem, but no specific or appropriate solution was proposed. ASVspoof 2015 focused on speech synthesis (SS) and voice conversion (VC), known as logical access condition (LA), while ASVspoof 2017 was designed to develop countermeasures capable of discriminating between bona fide (genuine) audios and replay ones, known as physical access condition (PA). Equal error rate (EER) is the common metric shared by them. ASVspoof 2019 covers both LA and PA but is divided into two separate subtasks.

To enhance the performance of anti-spoofing systems, recent works mainly focus on two approaches. One is to improve the front-end features extracted from audios [5, 6, 7, 8], where GMMs or LightCNN models are usually used as the classifiers. Another approach is to design new deep learning models [9, 10, 11, 12, 13] that learn discriminative representations for this task. Both of these two methods have been shown effective, which suggests that using appropriate front-end features as well as excellent deep learning models are both vital to the spoofing detection.

The rest of the paper is organized as follows, Section 2 briefly introduces the task of ASVspoof 2019 challenge, and Section 3 describes the features we used in the challenge. Section 4 will present the CNN based models and further explore the capabilities of GLU activations. The experiment details and results are given in Section 5. Section 6 concludes the whole paper.

2. Task Description

For better assessment of countermeasures for various spoofing attacks, ASVspoof 2019 challenge comprises two subtasks: logical access (LA) and physical access (PA).

2.1. Logical Access

Logical access (LA) spoofing attacks refer to spoofed speech generated with text-to-speech (TTS) and voice conversion (VC). As the widely use of neural-network-based systems in TTS and VC communities, the quality of generated speech is comparable to human speech, which brings new challenges to the spoofing detection system.

In the ASVspoof 2019 challenge, training data includes spoofed utterances generated according to two voice conversion and four speech synthesis algorithms, while spoofed algorithms in evaluation data are all unseen in the training set. Strong robustness is a requirement for our proposed spoofing detection systems.

2.2. Physical Access

Physical access (PA) spoofing attacks, also known as replay attacks, are performed at the sensor level. Since the somewhat uncontrolled setup in ASVspoof 2017 challenge makes the results difficult to analyze, the acoustic and replay configurations are carefully simulated and controlled in ASVspoof 2019 challenge, thus bringing some new insights into the replay spoofing problem.

†: These authors have contributed equally to this work
Yanmin Qian and Kai Yu are the corresponding authors

http://dx.doi.org/10.21437/Interspeech.2019-2170

The main focus of the PA subtask lies in detecting spoofing speech under different acoustic and replay configurations. Similar to the LA subtask, training and development data are generated from the same, randomly selected acoustic room and distance configuration, while the evaluation data is generated from different ones.

3. Feature Extraction

Here we propose the features used in our work. If not further specified, a normal frame-rate is adopted with 10ms frame shift and 25ms window size was adopted. Librosa [14] was used as our tool of choice for feature extraction.

Log-CQT replaces the standard Fourier transform of an audio signal with the constant-q transform (CQT). The constant q transform is very similar to Fourier transform but has logarithmically spaced center frequencies. In this work 84 dimensional log-CQT features were extracted with a frame shift of 32ms.

Log mel spectrogram (LMS) is a standard feature for ASR and other speech related tasks such as emotion detection [15] and audio event detection [16]. Here, 64 dimensional LMS features were extracted, where the hamming window function was used during pre-processing.

Phase features are extracted in addition to standard magnitude spectrogram features. The frequency spectrum of $\mathbf{X}$ can be decomposed into magnitude ($|X(\omega)|$) and phase ($e^{j\phi\omega}$) of as in Equation (1)

$$X(\omega) = |X(\omega)|e^{j\phi\omega} \tag{1}$$

In this work, we experiment with features extracted from the phase spectrogram ($e^{j\phi\omega}$). Specifically, log-CQT and LMS features are extracted from the phase spectrogram in addition to the traditional magnitude spectrogram .

VAE log-CQT refers to use Variational Autoencoder (VAE) to extract genuine speech specific feature. All bona fide LA log-CQT features are used to train a VAE, which encodes data to 32-dim vectors and then try to reconstruct. Those vectors are our desired features, which are supposed to be meaningful on genuine data and be randomly distributed on spoofing speech.

4. CNN based Spoofing Detection

Convolutional neural network (CNN) based models are used as our classifiers because of their promising performance in [17, 18]. In addition to the heavily investigated models such as ResNet and LightCNN, the use of gated linear unit activation within CNNs for spoofing detection is proposed.

4.1. ResNet

A standard 18-layer ResNet comprised of 8 residual blocks is adopted as one of our single systems. The detailed configuration can be found in Table 1.

4.2. ResNet with i-vector

In order to enhance the generalization capability of our neural network model, i-vector is concatenated to the ResNet embedding layer as an additional feature for joined training. Compared to the naive GMM approach, i-vector is a factor analysis based method which can reduce the impact of spoof-independent factors. The architecture is depicted in Figure 1. In this work, the 400-dim i-vector extracted from log-CQT features is concatenated to a 128-dim ResNet18 embedding.

Table 1: *Detailed Configuration of ResNet model. T denotes the frame number of input utterance and D denotes the feature dimension. Kernel sizes are set to* 3×3.

Layers	Output Size	Channels	Blocks
Conv	$T \times D$	16	-
Res1	$T \times D$	16	2
Res2	$T/2 \times D/2$	32	2
Res3	$T/4 \times D/4$	64	2
Res4	$T/8 \times D/8$	128	2
Average	128	-	-
Linear (embedding)	128	-	-
Output	2	-	-

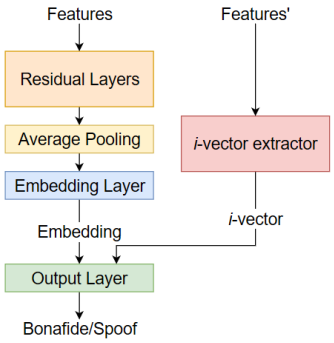

Figure 1: *The proposed ResNet + i-vector architecture. The inputs to the ResNet model and i-vector extractor are features (which are log-CQT + phase and log-CQT in this work, respectively) extracted from the same utterance.*

4.3. LightCNN with multi-task outputs

Following the best system in ASVspoof 2017 challenge [17], a 9-layer LightCNN with max filter map (MFM) activation function is proposed. The general architecture of LightCNN model using multi-task outputs is shown in Table 2. The outputs of FC8_output1 refer to the spoofing labels (1 bona fide node and 1 spoofing node), while the outputs of FC8_output2 are the replay configuration labels (1 bona fide node and 9 replay configuration nodes, seen in Section 5.1). The sum of the outputs in both bona fide nodes is regarded as the detection score.

4.4. Context Gate CNN

In this work we further explore the capabilities of gated linear unit (GLU) activations. This activation function has been used in related tasks such as audio event detection (AED) [19], sound event detection [20], speech recognition[21] as well as natural language processing [22]. GLU can be seen as an alternative to the MFM activation used in the LightCNN. In this work, GLU halves the input tensor over the CNN filter dimension (B and A) and uses one of those filters as weights and applies those weights on the other $f(A, B) = \sigma(A) \times B$ (see Figure 2). Here $\times$ is the Hadamard product of two tensors and σ is the sigmoid activation function.

This activation acts as a context-gate for each filter, which is the reason to denote this network as context gate CNN (CGCNN). A single context gate of our network can be seen in Figure 2. The context gate architecture in this work strictly follows our LCNN approach (see Table 2), however small changes

Table 2: *The LightCNN architecture of LightCNN model using multi-task outputs. The filter size, stride, and pad of Conv1 and MaxPool1 are* $(5 \times 5, 1, 2)$ *and* $(2 \times 2, 2, 0)$*, respectively. Hyper parameters* $C_i (i = 1, 2, 3, 4, 5)$ *is the number of output channels in the i-th layer, which basically control the model size. Along the* T *dimension, Statistics Pooling refers to mean pooling* $(D/32 \times C_5)$ *or mean+std pooling* $(D/32 \times C_5 \times 2)$*.*

Layers	**Output Size**
input	$T \times D \times 1$
Conv1+MFM1+MaxPool1	$T/2 \times D/2 \times C_1$
Block2(C_1,C_2,MFM)	$T/4 \times D/4 \times C_2$
Block3(C_2,C_3,MFM)	$T/8 \times D/8 \times C_3$
Block4(C_3,C_4,MFM)	$T/16 \times D/16 \times C_4$
Block5(C_4,C_5,MFM)	$T/32 \times D/32 \times C_5$
StatisticsPooling6	$D/32 \times C_5(\times 2)$
FC7	128
FC8_output1	2
FC8_output2	10

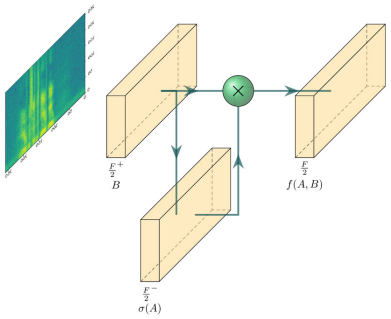

Figure 2: *A single context gate of our proposed context-gate CNN.*

were made: 1) The model only uses Block2 and Block3 with GLUs in order to avoid over-fitting ($C_1 = 48, C_2 = 96, C_3 = 192$). 2) No multi-task training was utilized. 3) Statistics Pooling referred to mean pooling only.

Moreover, for the final system fusion of our LA submission, we also incorporated a bidirectional gated recurrent unit (BGRU) model into the CGCNN model, further referred as CGCRNN. This GRU model was fed abstract features from the CGCNN and predicted posterior probabilities.

5. Experiments

Model training for all experiments was ran for at most 200 epochs using adam optimization where the model producing the lowest cross-entropy loss on the held-out set was chosen for final evaluation. Before training, we split the given train dataset into a 90% training and 10% held-out cross-validation portion in stratified fashion. Since the number of spoofed utterances within the training data set is only a fraction of the bona fide ones, one needs to ascertain that the trained model sees equally many bona fide and spoofed utterances. Therefore we adopt the use of random oversampling the minority class (bona fide) during training.

5.1. Dataset and performance measures

All experiments were conducted on the ASVspoof 2019 dataset respecting the official protocols on training/development divisions. For the LA subtask, 2,580 genuine and 22,800 spoofed speech utterances generated by one of 6 TTS/VC algorithms are used for training. The same spoofing algorithms in training set are used to create the development set, while the algorithms to generate the evaluation dataset are different. For PA task, the training set contains 5,400 genuine speech and 48,600 replay spoofing speech comprising 9 different replay configurations (3 categories of attacker-to-speaker recording distance times 3 categories of loudspeaker quality). The evaluation set for PA task has the same replay spoofing manner as training and development data, with different acoustic configurations. More details of the dataset can be found in ASVspoof 2019 evaluation plan[1].

To evaluate the performance of countermeasure, minimum tandem detection cost function (t-DCF) [23] is adopted as the primary performance metric, while equal error rate (EER) is used as a secondary metric.

5.2. Evaluation on the LA task

The components of our submitted system and their performance on development set is depicted in Table 3. Our single Context Gate CNN system with phase + log-CQT feature reaches 0.034 and 1.09 in min-tDCF and EER, respectively. By fusing all subsystems together, better performance can be achieved, resulting in 0.027 and 0.90 in min-tDCF and EER, respectively. The fusion system is submitted as our primary system.

Table 3: *Performance comparison of the components of our submitted system on development set for LA subtask. "+" mark denotes concatenating features into a multi-channel input.*

Model	**Feature**	**min-tDCF**	**EER**
CGCNN	VAE log-CQT+log-CQT	0.056	1.84
CGCNN	Phase+log-CQT	0.034	1.09
CGCRNN	VAE log-CQT+log-CQT	0.059	1.76
ResNet18	VAE log-CQT+log-CQT	0.040	1.41
ResNet18	Phase+log-CQT	0.051	1.53
ResNet18IVec	Phase+log-CQT	0.087	2.62
Fusion	-	**0.027**	**0.90**

Figure 3 illustrates the detailed results on different spoofing attacks. Although the baseline system (CQCC-GMM) achieves great results on specific spoofing types such as A01 and A02, it fails on most unknown spoofing attacks, potentially indicating an over-fitting problem. In comparison, our proposed system is more robust, resulting in evenly distributed low EERs and min-tDCFs on all spoofing conditions. Table 4 displays the results of the LA subtask. Our proposed system achieves the 5$^{\text{th}}$ position.

5.3. Evaluation on the PA task

OpenSLR26[2], a simulated room impulse response database, is used for data augmentation for the PA task. Specifically, for each genuine speech in the training set, 20 randomly-chosen room impulse response are added. Thus a total number of 108,000 RIR replicas are obtained.

In order to avoid potential over-fitting, 2 different settings of hyper parameters $C_i (i = 1, 2, 3, 4, 5)$ are adopted for the

[1]Refer to `http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf` for details.

[2]Refer to `http://www.openslr.org/26/` for details.

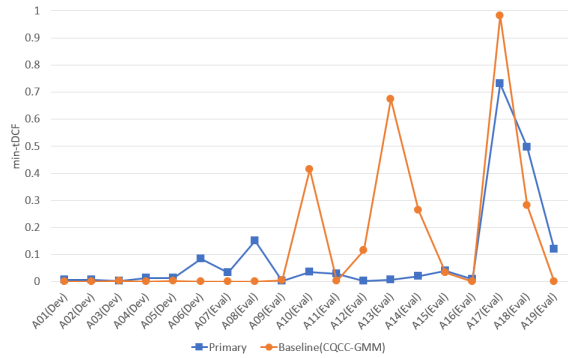

Figure 3: *Performance comparison between the baseline system (red) and our proposed system (blue) on different types of spoofing attacks for the LA subtask. A01 to A06 are known spoofing algorithms seen in the development set, while A07 to A19 are unknown spoofing algorithms in the evaluation set.*

Table 4: *Primary submission results on the evaluation set for LA subtask in ASVspoof 2019 Challenge. The result indicated in bold is our submission.*

ranking	team	min-tDCF	EER
1	T05	0.0069	0.22
2	T45	0.0510	1.86
3	T60	0.0755	2.64
4	T24	0.0953	3.45
5	**T50**	**0.1118**	**3.56**

multi-task LightCNN (LightCNN-MT) models. The larger one (LightCNN-MT-L) uses (48,96,192,128,128), while the smaller one (LightCNN-MT-S) uses (16,32,64,48,48). Furthermore, both mean pooling (denoted as μ) and mean+std (denoted as $\mu\sigma$) pooling are used, leading to 4 different models totally. LMS feature is used as input to our primary system, which is the score fusion of those 4 sub-models shown in Table 5.

Table 5: *Performance of the 4 sub-models, primary as well as the constrastive submission on the development set for the PA subtask. μ indicates the mean pooling while σ refers to the pooled standard deviation.*

Model	Feature	min-tDCF	EER
LightCNN-MT-L-μ	LMS	0.0180	0.59
LightCNN-MT-L-$\mu\sigma$	LMS	0.0189	0.71
LightCNN-MT-S-μ	LMS	0.0235	0.88
LightCNN-MT-S-$\mu\sigma$	LMS	0.0221	0.79
Fusion (above 4)	-	0.0108	0.38
CGCNN	log-CQT	0.0092	0.35
CGCNN (RIR)	log-CQT	0.0078	0.31
Fusion (above 2)	-	**0.0049**	**0.16**

Interestingly, our contrastive submission outperformed our primary submission on the evaluation set. Both of which significantly outperformed the baseline CQCC-GMM model in every replay configuration, shown in Figure 4. The contrastive model is a two way CGCNN fusion using the log-CQT feature - one being trained on the standard PA train set, while the other was trained on the augmented RIR data.

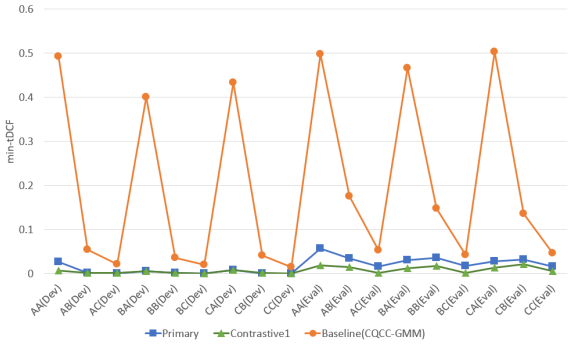

Figure 4: *Performance comparison of the baseline (red), our primary submission (blue), and our contrastive submission (green) for PA subtask.*

Table 6 displays the PA subtask results. Our primary system achieves the 8th position, while our contrastive submission achieves a comparable performance to the 3rd position.

Table 6: *Primary submission results on the evaluation set for PA subtask. The result indicated in bold is our primary submission. The result indicated as * is our submitted contrastive model composed of a two way context-gate CNN fusion.*

ranking	team	min-tDCF	EER
1	T28	0.0096	0.39
2	T45	0.0122	0.54
*	*T50*	*0.0137*	*0.54*
3	T44	0.0161	0.59
4	T10	0.0168	0.66
5	T24	0.0215	0.77
6	T53	0.0219	0.88
7	T17	0.0266	0.96
8	**T50**	**0.0350**	**1.16**

6. Conclusion

In this paper, we investigated multiple CNN based approaches, namely ResNet, LightCNN and most notably CGCNN for the ASVspoof 2019 challenge. Standard LMS as well as log-CQT features were used in conjunction with a newly uncertainty driven VAE model in order to ascertain robustness on development as well as evaluation subsets. Our results show that context-gated CNN networks are viable for both, logical and physical, scenarios. The proposed CGCNN model is shown to be reliable for both tasks. Our submitted system on the LA task, composed of a ResNet and CGCNN fusion, achieves a t-DCF of 0.027 on the development set and the 5th position on the evaluation set. On the other hand, our submission to the PA task, a LightCNN fusion, resulted in a t-DCF of 0.0108 on the development set and the 8th position on the evaluation set. Furthermore, our contrastive submission, a two way CGCNN fusion, outperformed our primary submission, achieving a comparable performance to the 3rd position.

7. Acknowledgement

This work has been supported by the Major Program of National Social Science Foundation of China (No.18ZDA293). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

8. 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] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification." in *Interspeech*, 2013, pp. 925–929.

[3] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "Asvspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 15, p. 3750, 2014.

[4] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "Asvspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 1508, p. 1508, 2017.

[5] H. Sailor, M. Kamble, and H. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," in *Proc. Interspeech*, 2018, pp. 666–670.

[6] B. Wickramasinghe, S. Irtza, E. Ambikairajah, and J. Epps, "Frequency domain linear prediction features for replay spoofing attack detection."

[7] T. Gunendradasan, B. Wickramasinghe, N. P. Le, E. Ambikairajah, and J. Epps, "Detection of replay-spoofing attacks using frequency modulation features."

[8] P. A. Tapkir and H. A. Patil, "Novel empirical mode decomposition cepstral features for replay spoof detection."

[9] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," *CoRR*, vol. abs/1810.13048, 2018.

[10] H.-J. Shim, J.-W. Jung, H.-S. Heo, S.-H. Yoon, and H.-J. Yu, "Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes," in *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 2018, pp. 172–176.

[11] 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)*, March 2017, pp. 4860–4864.

[12] H. Dinkel, Y. Qian, and K. Yu, "Small-footprint convolutional neural network for spoofing detection," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 3086–3091.

[13] Y. Qian, N. Chen, H. Dinkel, and Z. Wu, "Deep feature engineering for noise robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1942–1955, Oct 2017.

[14] B. McFee, M. McVicar, S. Balke, V. Lostanlen, C. Thom, C. Raffel, D. Lee, K. Lee, O. Nieto, F. Zalkow, D. Ellis, E. Battenberg, R. Yamamoto, J. Moore, Z. Wei, R. Bittner, K. Choi, nullmightybofo, P. Friesch, F.-R. Stter, Thassilo, M. Vollrath, S. K. Golu, nehz, S. Waloschek, Seth, R. Naktinis, D. Repetto, C. F. Hawthorne, and C. Carr, "librosa/librosa: 0.6.3," Feb. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2564164

[15] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 03 2017.

[16] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *Transactions on Audio, Speech and Language Processing: Special issue on Sound Scene and Event Analysis*, vol. 25, no. 6, pp. 1291–1303, June 2017.

[17] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." in *Interspeech*, 2017, pp. 82–86.

[18] B. Chettri, S. Mishra, B. L. Sturm, and E. Benetos, "Analysing the predictions of a cnn-based replay spoofing detection system," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 92–97.

[19] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 121–125.

[20] Y. Hou, Q. Kong, J. Wang, and S. Li, "Polyphonic audio tagging with sequentially labelled data using crnn with learnable gated linear units," *arXiv preprint arXiv:1811.07072*, 2018.

[21] X. Chang, Y. Qian, and D. Yu, "Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks," in *Proc. Interspeech 2018*, 2018, pp. 1586–1590. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1547

[22] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 933–941. [Online]. Available: http://proceedings.mlr.press/v70/dauphin17a.html

[23] T. Kinnunen, K. A. 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," *arXiv preprint arXiv:1804.09618*, 2018.