Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019

Xingliang Cheng, Mingxing Xu and Thomas Fang Zheng*
Center for Speech and Language Technologies, Research Institute of Information Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
*Corresponding Author E-mail: fzheng@tsinghua.edu.cn

***Abstract*—Automatic Speaker Verification (ASV) technology is vulnerable to various kinds of spoofing attacks, including speech synthesis, voice conversion, and replay. Among them, the replay attack is easy to implement, posing a more severe threat to ASV. The constant-Q cepstrum coefficient (CQCC) feature is effective for detecting the replay attacks, but it only utilizes the magnitude of constant-Q transform (CQT) and discards the phase information. Meanwhile, the commonly used Gaussian mixture model (GMM) cannot model the reverberation present in far-field recordings. In this paper, we incorporate the CQT and modified group delay function (MGD) in order to utilize the phase of CQT. Also, we present a simple 2D-convolution multi-branch network architecture for replay detection, which can model the distortion both in the time and frequency domains. The experiment shows that the proposed CQT-based MGD feature outperforms traditional MGD feature, and performance can be further improved by combining both magnitude-based and phase-based feature. Our best fusion system achieves 0.0096 min-tDCF and 0.39% EER on ASVspoof 2019 Physical Access evaluation set. Comparing with the CQCC-GMM baseline system provided by the organizer, the min-tDCF is relatively reduced by 96.09% and EER is relatively reduced by 96.46%. Our system is submitted to the ASVspoof 2019 Physical Access sub-challenge and won 1st place.**

I. Introduction

Automatic speaker verification (ASV) is a technology that verifys a person's identity through the voice, which is often used in the security-related application. However, the vulnerability to spoofing attack becomes a serious problem [1], [2]. There are four types of spoofing attacks [3]: impersonation, voice conversion, speech synthesis and replay. Impersonation tries to mimic the target speaker voice only by the human itself. Voice conversion, however, converts a talker's voice to mimic the target speaker's voice. Speech synthesis tries to synthesize the target speaker voice by computer directly. The replay attack just needs to replay a pre-recorded speech which is spoken by the target speaker. Among four types of spoofing attacks, the replay attack is accessible and straightforward, because it needs no specialized knowledge or skill and just needs to play back the recording instead of mimicking the target speaker voice. Research shows that the replay attack presents a high risk to the ASV system [2], making it an urgent problem.

Replay detection aims to distinguish whether a speech signal is a replay recording, or the voice spoken by humans directly. There are two standard databases in this field: AVspoof [4] and ASVspoof2017 [5] database. Both of them are collected on real world. Recently, the ASVspoof 2019 [6] physical access sub-challenge provides a database, which simulates the replay attack signal in the computer. One advantage of the simulation is to control variables conveniently. Another advantage is that the simulation can reduce the cost of data collection, providing a shortcut for creating a large-scale database.

In this paper, we describe our system submitted to the ASVspoof 2019 challenge. A novel constant-Q transform (CQT) [7] based modify group delay (MGD) [8] feature is proposed. By replacing the short-time Fourier transform (STFT) used in traditional MGD with CQT, the proposed feature can utilize the phase of CQT. In order to calculate the CQT-based MGD efficiently, we modify the extraction process of MGD. The new extraction process is suitable for calculating the MGD based on the various time-frequency analysis method, including the STFT and CQT. Meanwhile, we present a simple 2D-convolution multi-branch network architecture for replay detection. Our model, named ResNeWt, adopts the same split-transform-merge strategy as the one used in the ResNeXt [9] but replaces the additive aggregate function used in ResNeXt by concatenation. The experiment shows that the CQT-based MGD outperforms the traditional MGD feature, and the performance can be further improved by combining both magnitude-based and phase-based feature. Further analysis reveals that our model can better detect the distortion introduced by the playback device and far-field recording compared with CQCC-GMM baseline model, and the multi-branch architecture can improve the modeling capability while maintaining the complexity at the same time.

The rest of this paper is organized as follows. In Section 2, we review the related works on spoof detection. Section 3 describes the proposed system. The experiment will be described in Section 4, and the result will be discussed in Section 5. In Section 6, we have a conclusion on this paper.

II. Related Work

Feature Engineering Since the microphone and loudspeaker are designed for recording and reproducing the sound as real as possible, the voice after play back should preserve the main information. Meantime, due to the non-ideal characteristics of the physical device, some distortion will be induced. Thus, researchers try to find an effective feature to detect such

distortion. Todisco et al. [7] proposed the constant-Q cepstral coefficient (CQCC) feature, which utilized CQT instead of STFT to convert a voice signal into the frequency domain, and then further transform it to the cepstral domain. The experiment shows that CQCC is generalized well in multiple datasets. Tom et al. [10] use the group delay function in the replay detection task, which not only contains the magnitude information, but also the phase information. Extending previous work, we proposed the CQT-based modified group delay feature, which is a combination of the CQT and modified group delay function. Benefit from the CQT, the proposed feature can have a higher time resolution for higher frequencies, and higher frequency resolution for lower frequencies. Also, the phase based feature is complemented with magnitude feature, so the combination can further boost the performance.

DNN-based Classifier The standard Gaussian mixture model (GMM) is a classic model in this field [3], [11], [12], due to its excellent performance. Recently, researchers have been attempted to use DNN in the replay detection task. Cai et al. [13] utilized ResNet model with spectrogram as the feature, FDNN and BLSTM model with CQCC as the feature to detect the replay attacks. Tom et al. [10] utilized the class activation mapping [14] technology to obtain the implicit attention mechanism presented in ResNet, and further use the attention to mask the group delay feature, then feed the new feature into another ResNet model to make the decision. Lavrentyeva et al. [15] utilized a Light CNN (LCNN) architecture to learn the audio representation from the log normalized power magnitude spectrogram extracted via FFT or CQT, and then use a GMM model to distinguish between genuine and spoof classes using the representation extracted by LCNN. Chen et al. [16] utilized ResNet to learn from CQCC and MFCC feature. ResNeXt [9] is an improved version of ResNet, it could be expected to improve the replay detection accuracy. However, ResNeXt only being applied in the ResNet with more than 50 layers. Due to the limited training data, a complex model does not work well. Thus, we proposed the ResNeWt, which can be applied in 18-layer ResNet to gain accuracy effectively while maintaining the complexity.

III. The Proposed System

In this work, we propose a novel CQT-based modified group delay feature which incorporates the CQT and MGD for a better representation of phase-related information. Also, the ResNeWt model is utilized to detect replay attacks.

A. CQT-based Modified Group Delay Feature

The modified group delay (MGD) function [8] is one of the most commonly used phase feature for speech recognition [17] and converted speech detection [18]. The MGD is derived from the group delay (GD) function, which is defined as the negative derivative of the phase information:

$$\tau(\omega,t) = -\frac{d(\theta(\omega,t))}{d\omega}, \tag{1}$$

where $\theta(\omega,t)$ is the phase spectrogram of signal $x(n)$, n is the index of the sample points, ω and t are the index of frequency bins and frames, respectively. The GD can also be calculated directly from the following formula:

$$\tau(\omega,t) = \frac{X_R(\omega,t)Y_R(\omega,t) + X_I(\omega,t)Y_I(\omega,t)}{|X(\omega,t)|^2}, \tag{2}$$

where $X(\omega,t)$ and $Y(\omega,t)$ are the Fourier transform of the signal $x^t(n)$ and $nx^t(n)$, respectively. The $x^t(n)$ is the signal in frame t. The subscripts R and I denote the real and imaginary parts of Fourier transform.

It should be noted that the GD function will become very spiky when the energy of some frames ($|X(\omega,t)|^2$ in (2)) are close to zero. The MGD function overcomes this by smoothing the spectrum and reduce the dynamic range, defined as:

$$\tau_m(\omega,t) = \left(\frac{\tau'_m(\omega,t)}{|\tau'_m(\omega,t)|}\right)|\tau'_m(\omega,t)|^\alpha, \tag{3}$$

where

$$\tau'_m(\omega,t) = \frac{X_R(\omega,t)Y_R(\omega,t) + X_I(\omega,t)Y_I(\omega,t)}{|S(\omega,t)|^{2\gamma}}. \tag{4}$$

$S(\omega,t)$ is the cepstrally smoothed spectrum of $X(\omega,t)$, γ and α are two parameters which are utilized to reduce the dynamic range, varing from 0 to 1.

The traditional MGD is based on the STFT. Recent studies show that CQT is more powerful than STFT in replay detection task [7]. Thus, we try to incorporate the CQT and MGD to construct a more powerful phase-based feature. However, considering the varying number of samples used in the CQT calculation of each frequency bin, the traditional frame-by-frame extraction is infeasible. To overcome this problem, we proposed a new MGD extraction process. It consists of three steps. Firstly, calculating the spectrogram on the unframed original signal $x(n)$ and $nx(n)$, denoted as:

$$X(\omega,t) = \Phi(x(n)), \tag{5}$$

$$Y(\omega,t) = \Phi(nx(n)), \tag{6}$$

where $\Phi(\bullet)$ can be arbitrary time-frequency analysis method, including the STFT and the CQT. Secondly, calculating the auxiliary spectrogram to compensate the bias between the spectrum calculated on a framed signal $nx^t(n)$, and the spectrum calculated on an unframed original signal $nx(n)$:

$$Y'(\omega,t) = Y(\omega,t) - t \times T \times X(\omega,t), \tag{7}$$

where T is the duration between the beginning of two adjacent frames. Lastly, the MGD is calculated as

$$\tau_m(\omega,t) = \left(\frac{\tau''_m(\omega,t)}{|\tau''_m(\omega,t)|}\right)|\tau''_m(\omega,t)|^\alpha, \tag{8}$$

where

$$\tau''_m(\omega,t) = \frac{X_R(\omega,t)Y'_R(\omega,t) + X_I(\omega,t)Y'_I(\omega,t)}{|S(\omega,t)|^{2\gamma}}. \tag{9}$$

Obviously, the new extraction process is suitable for both traditional MGD and the proposed CQT-based MGD.

B. ResNeWt

ResNet is a popular model in image recognition [19], also has been utilized in the replay detection task [10], [13], [16]. The key point of ResNet is its residual module, as demonstrated in Fig. 1 (a). Based on the ResNet, ResNeXt [9] adopt a splitting, transforming and aggregating strategy to gain accuracy effectively while maintaining the complexity. A two-layer block adopts the same strategy as used in ResNeXt is demonstrated in Fig. 1 (b), however, it equals to trivially a wide, dense module [9]. So the idea of ResNeXt is not suitable for the model using the two-layer building block, for example, the 18-layer ResNet. Thus, we modify the aggregate function used in ResNeXt to construct a new model, named ResNeWt, which can be performed on all types of the ResNet. A basic block of ResNeWt is demonstrated in Fig. 1 (c), which can be defined as:

$$\mathbf{Y} = \mathbf{X} + \underset{i=1}{\overset{D}{\Xi}} f_i(\mathbf{X}), \tag{10}$$

where $\mathbf{Y}$ is the output of the building block, $\mathbf{X}$ is an input tensor, $f_i(\bullet)$ can be an arbitrary function which splits the input first and then transform them, D is the size of the set of transformations to be aggregated, Ξ is the aggregate function that concatenates the tensor along the channel dimension, defined as:

$$\underset{i=1}{\overset{D}{\Xi}} [S_1^{(i)}, ..., S_{C_i}^{(i)}] = [S_1^{(1)}, ..., S_{C_1}^{(1)}, S_1^{(2)}, ..., S_{C_D}^{(D)}], \tag{11}$$

where $[S_1^{(i)}, ..., S_k^{(i)}]$ is a k-channel tensor.

In this work, an 18-layer ResNeWt (ResNeWt18) is constructed by reference to the structure of 18-layer ResNet (ResNet18). The building block of ResNeWt18 has two 3x3 convolutional layers. The first convolutional layer is the same as the one in ResNet; the second 3x3 convolutional layer is split into 32 groups [20]. The number of the channel is doubled compare with ResNet18. To prevent the potentially overfitting problem, we add a dropout layer after the global average layer. The overall structure is described in Table I.

TABLE I
THE OVERALL ARCHITECTURE OF RESNEWT18. THE SHAPE OF A RESIDUAL BLOCK [19] IS INSIDE THE BRACKETS, AND THE NUMBER OF STACKED BLOCKS ON A STAGE IS OUTSIDE THE BRACKETS. "C=32" MEANS THE GROUPED CONVOLUTIONS [20] WITH 32 GROUPS. "2-D FC" MEANS A FULLY CONNECTED LAYER WITH 2 UNITS.

Stage	Output Shape	Detail
conv1	256×128	7×7, 64, stride 2
conv2	128×64	3×3 max pool, stride 2
		$\begin{bmatrix} 3 \times 3,\ 128, \\ 3 \times 3,\ 128,\ C = 32 \end{bmatrix} \times 2$
conv3	64×32	$\begin{bmatrix} 3 \times 3,\ 256, \\ 3 \times 3,\ 256,\ C = 32 \end{bmatrix} \times 2$
conv4	32×16	$\begin{bmatrix} 3 \times 3,\ 512, \\ 3 \times 3,\ 512,\ C = 32 \end{bmatrix} \times 2$
conv5	16×8	$\begin{bmatrix} 3 \times 3,\ 1024, \\ 3 \times 3,\ 1024,\ C = 32 \end{bmatrix} \times 2$
	1×1	global average pool, dropout, 2-d fc, softmax

IV. EXPERIMENTS

A. Dataset

The dataset provided in ASVspoof 2019 [6] physical access sub-challenge was used in this paper. It contains the simulated bona fide and the simulated replay spoofing access attempts. The source signals for performing simulation are from the VCTK[1] corpus. Room acoustics are simulated by Roomsimove toolbox[2] and replay devices are simulated using the generalized polynomial Hammerstein model and the synchronized swept-sine tool[3]. For more details, see [6].

B. Evaluation Metrics

In ASVspoof 2019 challenge, the *minimum normalized tandem detection cost function* (min-tDCF) [6], [21] is used as the primary metric, which can be simply calculated as:

$$\text{t-DCF}_{\text{norm}}^{\text{min}} = \min_s \beta P_{\text{miss}}^{\text{cm}}(s) + P_{\text{fa}}^{\text{cm}}(s), \tag{12}$$

where $P_{\text{miss}}^{\text{cm}}(s)$ and $P_{\text{fa}}^{\text{cm}}(s)$ are, respectively, the *miss rate* and the *false alarm rate* of the countermeasure (CM) system at threshold s, β is a cost which depends on the t-DCF parameters and ASV errors ($\beta \approx 2.0514$ in ASVspoof 2019 physical access development set with the ASV score provided by organizers [6]). The *equal error rate* (EER) [6] is also used as the secondary metric.

C. Experimental Setup

1) Feature: In this work, we utilized the STFT and the CQT to extract the magnitude-based or phase-based time-frequency representation (TFR). They were:

- *Magnitude-based feature*: The traditional log power magnitude spectrogram based on FFT (*Spectrogram*), Mel scale filter banks (*MelFbanks*) and log power magnitude spectrogram based on CQT (*CQTgram*).
- *Phase-based feature*: The tranditional *MGD* feature and the proposed CQT-based MGD (*CQTMGD*) feature.

Spectrogram and MelFbanks were extracted with 50 ms frame length, 32 ms frame shift, 1024 FFT point, Hamming window. Total of 128 Mel filter banks was extracted in MelFbanks. MGD was extracted with 50 ms frame length, 25 ms frame shift, Hamming window, 1024 FFT point, $\alpha = 0.6$, and $\gamma = 0.3$. CQTgram and CQTMGD were extracted with 32 ms frame shift, Hanning window, 11 octaves, and 48 bin per octave. For CQTMGD, we set $\alpha = 0.35$ and $\gamma = 0.3$. All the features were truncated along the time axis to reserved exactly 256 frames. The feature less than 256 frames would be extended by repeating their contents. Finally, for simplicity, all the features are resized to 512×256 by bilinear interpolation.

[1] http://dx.doi.org/10.7488/ds/1994
[2] http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip
[3] https://ant-novak.com/pages/sss/

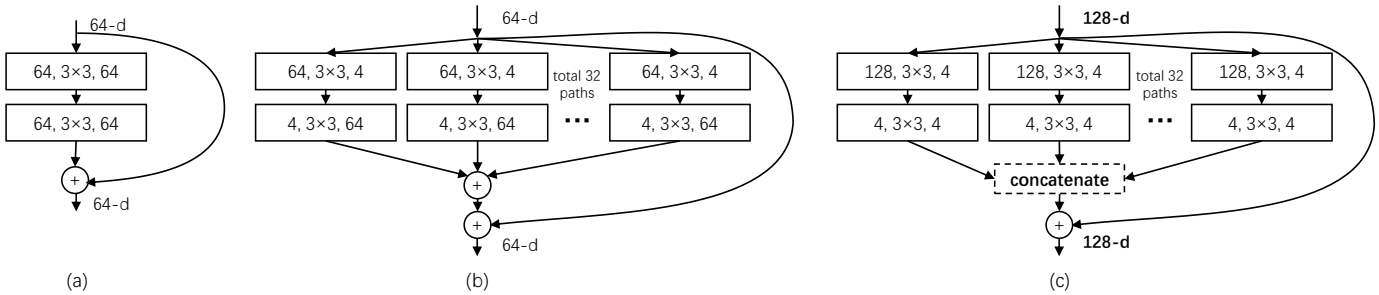

Fig. 1. Demonstration of the building blocks. (a) A block of ResNet [19]. (b) A block adopts the same strategy as used in ResNeXt [9]. (c) A block of ResNeWt. A layer is denoted as (#input channels, kernel size, #output channels).

TABLE II
RESULTS ON ASVSPOOF 2019 PHYSICAL ACCESS CHALLENGE

Description	System	Dev		Eval	
		t-DCF$^{\text{min}}_{\text{norm}}$	EER(%)	t-DCF$^{\text{min}}_{\text{norm}}$	EER(%)
Baseline	LFCC-GMM [6]	0.2554	11.96	0.3017	13.54
	CQCC-GMM [6]	0.1953	9.87	0.2454	11.04
Other Teams	T24 [22]	0.0114	0.44	0.0215	0.77
	T10 [22]	0.0065	0.24	0.0168	0.66
	T44 [22]	**0.0032**	**0.13**	0.0161	0.59
	T45 [22]	0.0054	0.30	**0.0122**	**0.54**
Single System (Magnitude)	Spectrogram	0.0882	3.15	—	—
	MelFbanks (A)	0.0428	1.70	—	—
	CQTgram (B)	0.0110	**0.39**	—	—
	A&B[a](C)	**0.0093**	0.41	**0.0134**	**0.52**
Single System (Phase)	MGD (D)	0.0246	0.97	0.0465	2.15
	CQTMGD (E)	**0.0149**	**0.54**	**0.0250**	**0.94**
Fusion	C+D[b]	0.0061	0.28	—	—
	C+E[b]	0.0072	0.31	—	—
	C+D+E[b]	**0.0049**	**0.20**	**0.0096**	**0.39**

[a] *A&B*: concatenating the feature A and B along the frequency axis[4].
[b] *C+D+E*: fusion by averaging the scores of subsystems C, D, and E.

2) ResNeWt: The ResNeWt18 was optimized by Adam algorithm with $10^{-3.75}$ as learning rate and 16 as batch size. The training process was stopped after 50 epochs. The loss function was the binary cross-entropy between the predictions and targets. The dropout radio was set to 0.5. The output of the "bona fide" node at last full connection layer was obtained as the output score (before softmax).

3) Fusion: A score level fusion was performed to combine the models trained by different features. For the sake of simplicity, the ensemble system averages the output score of all subsystems. A greedy-based strategy was used in selecting subsystems. First, the best system was chosen. Then one system was been selected greedily each time according to the min-tDCF performance of the ensemble system evaluated on the development set. The selection process would not stop until the performance was stable.

V. RESULTS AND DISCUSSION

A. Results on ASVspoof 2019

[4]Specially, We do not resize the shape of the concatenated feature, so it equals to 656×256 (656 = 528(CQTgram) + 128(MelFBank)).

TABLE III
DEFINITION OF ATTACK SOURCE

Factor	Parameter	Level		
		A	B	C
D[a]	distance (cm)	$10 \sim 50$	$50 \sim 100$	> 100
Q[b]	OB[c] (kHZ)	∞	> 10	< 10
	minF[d] (Hz)	0	< 600	> 600
	linearity[e] (dB)	∞	> 100	< 100

[a] *D*: attacker-to-talker distance.
[b] *Q*: replay device quality.
[c] *OB*: occupied bandwith.
[d] *minF*: lower bound of OB.
[e] *linearity*: linear/non-linear OB power difference.

Table II depicts a quantitative comparison of the replay detection systems. Among all single systems, the CQTgram achieve the lowest EER and almost the lowest min-tDCF. The concatenation of CQTgram and MelFbanks slightly improves the min-tDCF, meanwhile, also affects the EER. We attribute it to the variance of the model, thus the main contribution is still from the CQTgram. Also, the CQTgram performs better than Spectrogram or MelFbanks, and the CQTMGD performs better than MGD as well. These phenomena indicate that the CQT are more suitable than FFT in this dataset/task. The fusion of single systems further improves performance, indicate the complementarity between magnitude and phase, and between CQT and FFT. Also, the performance of CQTMGD feature is competitive, and a further improvement achieved when it is fused with other feature, demonstrated the effectiveness of this feature.

Finally, the fusion system and the three subsystems were submitted to ASVspoof 2019 challenge. As we can see in Table II, all systems have a stable performance on both development and evaluation set, indicating a good generalization ability of the model. All the systems perform better than the best baseline system (CQCC-GMM), also better than the systems submitted by other teams.

B. Error Analysis

To better understand the model's ability, let us look at the performance against different attack sources in detail. Two factors have been used in identifying the attack sources:

TABLE IV
REPLAY DETECTION PERFORMANCE UNDER DIFFERENT ATTACK CONDITIONS (EER(%)) ON THE EVALUATION SET

Model	CQCC-GMM			ResNeWt (fusion)		
D \ Q	A	B	C	A	B	C
A	25.28	6.16	2.13	0.86	0.30	0.12
B	21.87	5.26	1.61	0.49	0.33	0.09
C	21.10	4.70	1.79	0.54	0.30	0.09

* See in Table III for the meaning of the symbols.

TABLE V
CONTRIBUTION ANALYSIS ON THE ASVSPOOF 2019 PHYSICAL ACCESS DEVELOPMENT SET[5]

Feature	Model	t-DCF$^{\text{min}}_{\text{norm}}$		EER(%)	
CQCC	GMM [6]	0.1953	—	9.87	—
	ResNet	0.0501	↓*74.4%*	1.98	↓*79.9%*
	ResNeWt	0.0419	↓*78.5%*	1.67	↓*83.1%*
CQTgram	ResNet	0.0124	↓*93.7%*	0.46	↓*95.3%*
	ResNeWt	0.0110	↓*94.4%*	0.39	↓*96.1%*
MGD	ResNet	0.0314	↓*83.9%*	1.28	↓*87.0%*
	ResNeWt	0.0297	↓*84.8%*	1.22	↓*87.6%*
CQTMGD	ResNet	0.0223	↓*88.6%*	0.78	↓*92.1%*
	ResNeWt	0.0180	↓*90.8%*	0.71	↓*92.8%*

- *Recording Distance (D)*: the distance between the talker and the attacker's microphone when the attacker secretly recorded the talker's voice. This factor affects the quality of the recording, specifically, the degree of reverberation.
- *Playback Device Quality (Q)*: the quality of the playback device when the attacker performs the replay attack. This factor is related to the degree of distortion in the frequency domain.

Each factor is categoried into three levels, and the detail information is shown in Table III.

As shown in Table IV, the ResNeWt works well in all conditions, while the GMM model shows a much higher EER when the quality of the playback device is well (Q=A). For the factor of recording distance, the performance is simular between D=B and D=C, but much differently between D=A and D=B. If we let the Q=A, and then observe the performance between D=A and D=B, we will find that the relevant performance drop of the ResNeWt model ($(0.86 - 0.49)/0.86 = 43.02\%$) is much more than the GMM model ($(25.28 - 21.87)/25.28 = 13.49\%$). This indicates that the ResNeWt model has a better ability to detect the distortion introduced by far-field recording. One possible reason for this is that the reverberation present in the far-field recordings will cause time-domain distortion. However, the GMM is a frame-level model, that means it has no ability to model the time-domain distortion. Benefited by the 2D convolution, the ResNeWt can model the distortion both in the time and frequency domains, so it has a better ability to detect the reverberation.

C. Contribution Analysis

In Table V, the performance is increased dramatically when the GMM is replaced by CNN, and a further improvement achieved when the hand-crafted CQCC feature is replaced by the low-level CQTgram. Also, the performance of CQTMGD is better than MGD. This implies that the main contribution of performance improvement comes from CNN's superior modeling capabilities, and the use of low-level feature get better use of the modeling capabilities. Meanwhile, the performance of ResNeWt is consistently better than ResNet, shows that the ResNeWt further improves the modeling capability while maintaining the complexity at the same time.

D. Attention Analysis

To have a better understanding of how the model works, we further visualize the distribution of model attention by class activation mapping (CAM) [14]. According to the binary classification, the evidence that proves the input signal falling to one category in the meantime indicating the absence of the signal in another category. As we only concern about the positive evidence, all the negative value in CAM is set to zero.

Fig. 2 demostrates the visualization of attention distributions. There are two obviously patterns. Firstly, the model concentrates on the low frequencies (the green solid line box in Fig. 2), indicating the importance of the low-frequency band. This could explain why CQT works better than FFT since the frequency resolution of CQT in low frequencies is much higher than FFT, so such low frequencies are hardly distinguished in FFT-based spectrogram. Also, we should be noticed that this phenomenon is different from the conclusion found in the ASVspoof2017 challenge that shows that the high-frequency band has more information [23]. So it may relate to the dataset and need further analysis.

Secondly, the model draws attention to the head and tail of the signal (the white dashed box in Fig. 2), and we found most of the signal has a leading and trailing silence. It indicates that silence contains some efficient information to detect the replay attack. However, it is not we expect to see, because the attacker can easily trim the silence of the speech before playback, then it is impossible to detect anything in the leading and trailing silence.

VI. CONCLUSION

In this paper, we present the replay detection system submitted to ASVspoof 2019 challenge. A novel CQT-based MGD feature is proposed to utilized the phase of CQT. An 18-layer ResNeWt model is utilized to detect the replay attacks. Our models have been evaluated on ASVspoof 2019 physical access challenge dataset and show a significant improvement on the ability to detect the distortion introduced by the playback device and the ability to detect the reverberation introduced by far-field recording, compared with CQCC-GMM baseline

[5]For better comparability, the structure of the ResNet here is the same as Table I except for the C=1. Also, high-dimensional CQTgram feature is not suitable for GMM. The ResNeXt is absent here because the 18-layer ResNeXt does not exist.

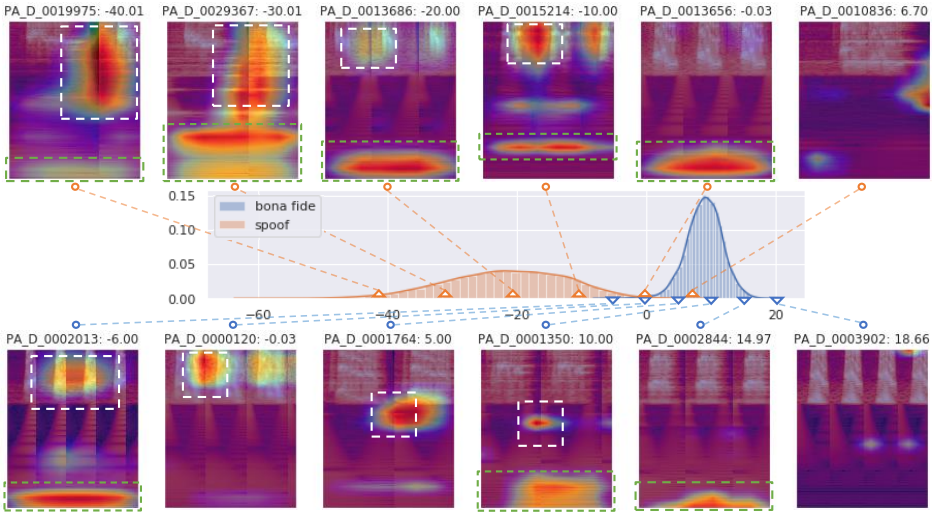

Fig. 2. Activation attention maps for ResNeWt with CQTgram. (**Top / Bottom**: The CAM of spoof / bona fide samples. The foreground heatmap shows the position where the model considers this signal is *not genuine*. The title describes the corresponding audio file name and output score. **Middle**: The score distribution on the development set. Best view in color.)

system. In the future, we will further analyze the method on real-world datasets like ASVspoof2017 and AVspoof dataset.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of China projects under Grant No. 61433018 / 61271389 and the Tsinghua-d-Ear Joint Laboratory on Voiceprint Processing. The authors would like to thank the ASVspoof 2019 challenge organizers.

REFERENCES

[1] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*. IEEE, 2015, pp. 1–6.

[2] M. Singh, J. Mishra, and D. Pati, "Replay attack: Its effect on GMM-UBM based text-independent speaker verification system," in *Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference on*. IEEE, 2016, pp. 619–623.

[3] H. A. Patil and M. R. Kamble, "A survey on replay attack detection for automatic speaker verification (ASV) system," in *APSIPA*. IEEE, 2018, pp. 1047–1053.

[4] S. K. Ergunay, E. el Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *BTAS*. IEEE, 2015, pp. 1–6.

[5] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. W. D. Evans, and M. Todisco, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *J. Sel. Topics Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

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

[7] 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.

[8] Z. Oo, L. Wang, K. Phapatanaburi, M. Iwahashi, S. Nakagawa, and J. Dang, "Phase and reverberation aware DNN for distant-talking speech enhancement," *Multimedia Tools Appl.*, vol. 77, no. 14, pp. 18 865–18 880, 2018.

[9] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[10] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," *INTERSPEECH, Hyderabad, India*, 2018.

[11] L. Li, Y. Chen, D. Wang, and T. F. Zheng, "A study on replay attack and anti-spoofing for automatic speaker verification," *arXiv preprint arXiv:1706.02101*, 2017.

[12] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection," in *Interspeech*. ISCA, 2018, pp. 641–645.

[13] 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 *INTERSPEECH*, 2017, pp. 17–21.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[15] 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.

[16] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "Resnet and model fusion for automatic spoofing detection." in *INTERSPEECH*, 2017, pp. 102–106.

[17] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–68.

[18] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7234–7238.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] 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.

[22] 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*, 2019.

[23] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *INTERSPEECH*, 2017, pp. 27–31.