Detection of Replay Attack Based on Normalized Constant Q Cepstral Feature

Yongchao Ye, Lingjie Lao, Diqun Yan*, Lang Lin
Faculty of Electrical Engineering and Computer Science
Ningbo University
Ningbo, China
e-mail: zjnbyyc@foxmail.com, {166001013, yandiqun}@nbu.edu.cn, ll_linlang@163.com

***Abstract*—Since the voice is easy to be recorded and replayed, the replay attack is considered a major threat to the voiceprint authentication system. However, few works have focused on the text-independent detection. We found that there exist differences in spectral features between the original voice and the replayed voice. Hence, constant Q cepstral coefficients which can well describe the spectral features are extracted as acoustic features. Then the cepstral mean and variance normalization is used as the postprocessing method to eliminate the impact of channel noise on detection performance. Finally, the Gaussian Mixture Model determines whether the suspected voice is replayed or not. The experimental results on the ASVspoof 2017 database indicate the proposed algorithm can significantly reduce the equal error rate and keep high robustness when the replayed voice came from various spoofing devices.**

Keywords-replay detection; CQCC; CMVN; GMM

I. INTRODUCTION

Nowadays, identity authentication systems based on biometric technology have become more prevalent because of the advantages over traditional password/ID card-based authentication systems. However, such systems can also have vulnerabilities. One of the most common forms of attack is called spoofing, which the attacker successfully impersonates as another by falsifying data and obtaining the unauthorized right of the authentication system. Most common attacks of voice spoofing include impersonation [1], synthesis [2], conversion [3] and replay [4]. Among them, voice replay is an easy-to-operate spoofing attack without the need of special processing techniques. The attacker recorded the genuine target speaker's voice and replayed by an imposter during authentication. Since no additional operations have been performed on the replayed voice, no modification traces have been left. Meanwhile, due to the availability of high quality recording devices, the replayed voice could have a higher similarity with the original voice than other attacks. The existing voiceprint authentication systems tend to be easily fooled when the pre-recorded voice of the target speaker is replayed [5, 6].

The existed studies on detecting voice replay consist of three categories, methods based on the randomness of voices, methods based on the channel noise and methods based on deep learning. Firstly, Shang et al. [7] proposed a replayed voice detection algorithm based on the randomness of the voices. The algorithm detects the difference between the original voice and the suspected voice in the peak map. Jakub Galka et al. [8] used the positional relationship of each frequency point in the peak map based on the Shang's algorithm. However, these methods are text-dependent. Secondly, distortion is introduced during encoding and decoding. Zhang used the distortion phenomenon of replayed voice [9], and proposed a method for modeling the channel noise based on the Mel's cepstral coefficient of the silent segments. The algorithm works by comparing the channel noises of the suspected voice with the established model. Wang Zhifeng [10] obtained satisfied results by extracting the 6th-order Legendre polynomial coefficients and using the SVM as classifier. In recent years, with the wide application of deep learning, some scholars try to apply this technique to the replayed voice detection. Lin et al. [11] preprocessed features based on electric network frequency (ENF) analysis and used convolutional neural networks (CNN) to distinguish the replayed voice. Since the performance is greatly affected by the ENF signal extraction, this method has a large limitation.

In real scenario, the voiceprint authentication system will be attacked from various spoofing devices. And different devices may cause different effects on the performance of the system. Few studies on detecting voice replayed by various spoofing devices have been reported. By analyzing a large amount of experiments, we found that replaying leaves traces in the high-frequency components compared with the original voice. The Mel filter generally has a high resolution in the low frequency bands but not the high frequency bands. the replay attack affects the high frequency bands most. Therefore, the detection rates are not satisfying using Mel-frequency cepstral coefficients (MFCC). In this work, a new method for detecting replay attack is proposed. We first extract the constant Q cepstral coefficients (CQCC), then cepstral mean and variance normalization (CMVN) is adopted to obtain the acoustic features. Gaussian Mixture Model (GMM) is utilized as the classifier to achieve the replay attack detection. The experimental results on the ASVspoof 2017 database show that the proposed algorithm can effectively reduce the equal error rate (EER) of the replay detection system. Furthermore, the cepstral mean and variance normalization makes the algorithm show good robustness against various replay attack devices.

The remainder of the paper is organized as follows. Section II describes the details of the proposed replay detection algorithm including feature extraction, feature

978-1-7281-1410-1/19/$31.00 ©2019 IEEE

normalization, and Gaussian Mixture Model. Section III gives the experimental results. Finally, the paper is concluded in Section IV.

II. Replay Detection Algorithm Based on Normalized CQCC

In this section, we firstly introduce the extraction procedure of CQCC feature. Then cepstral mean and variance normalization is described. Finally, the Gaussian Mixture Model and the replay detection algorithm are given.

A. Feature Extraction

CQCC is a magnitude-based feature which provides a time-frequency analysis method. Compared to traditional short-term Fourier transform (STFT), CQCC tends to capture more acoustic information at lower frequencies and more time information at higher frequencies. It achieves significantly low equal error rate for voice synthesis and conversion attacks on ASVspoof 2015 challenge [12], which indicates that CQCC can capture the traces of artificial operations more effectively than other traditional features. In this work, we have adopted this feature for the proposed replay detection algorithm.

As seen from Figure 1, CQCC is based on constant Q transform (CQT). CQT was firstly introduced by Boll in 1978 as a perceptually driven time-frequency analysis method [13], which has been further improved over the past few decades, such as [14]. Compared with the traditional short-term Fourier transform, the center frequency of CQT is geometrically distributed, which makes CQT widely concerned in voice signal processing [15].

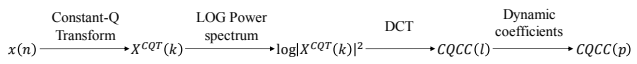

Figure 1. Calculation procedure of CQCC.

CQT can be considered as wavelet transform, the attenuation factor of the filter is fixed, and the center frequency is exponentially increased. The *k*th component of the CQT spectrum of the *n*th frame signal is:

$$X^{CQT}(k,n)=\sum_{j=n-\lfloor N_k/2\rfloor}^{n+\lfloor N_k/2\rfloor} x(j)a_k^*(j-n+N_k/2) \tag{1}$$

where $\lfloor\ \rfloor$ is down rounding. $k=1,2,...,K$, k represents the CQT frequency subscript and K represents the total number of frequency subscripts,

$$K=\left\lceil B\cdot\log 2\cdot\left(\frac{f_{\max}}{f_{\min}}\right)\right\rceil \tag{2}$$

where $\lceil\ \rceil$ is up rounding, B represents the number of subscripts for each octave, which determines the time resolution and frequency resolution. f_{min} and f_{max} represent the highest and lowest center frequency individually. $a^*_k(n)$ is the complex conjugate of the $a_k(n)$, $a_k(n)$ is defined as:

$$a_k(n)=\frac{1}{N_k}\omega\left(\frac{n}{N_k}\right)\exp\left(-i2\pi n\frac{f_k}{f_s}\right) \tag{3}$$

where f_s is sampling frequency. $\omega(t)$ uses Hamming window, when it is outside the range of [0,1], $\omega(t)$=0. The center frequency of *k*th subscript is defined as:

$$f_k=f_{\min}2^{(k-1)/B} \tag{4}$$

Q represents attenuation factor, which is defined as:

$$Q=(2^{1/B}-1)^{-1} \tag{5}$$

The width of N_k for *k*th frequency subscript is:

$$N_k=\left\lceil Q\frac{f_s}{f_k}\right\rceil \tag{6}$$

The selected range of frame H_k is defined as:

$$0<H_k\le\frac{1}{2}N_k \tag{7}$$

Since the ratio of the center frequency to the bandwidth in the CQT is a constant, the distribution of the CQT frequency is geometrically distributed. However, the traditional DCT has to be modified since it is linear. Spline interpolation is a suitable choice which resamples the points of the geometric distribution onto the linear distribution, and then operating the spectrum derived from the linearized CQT to obtain the cepstrum. The detail definition is as follow:

$$CQCC(r)=\sum_{i=0}^{K-1}\log\left|X^{CQT}(i)\right|^2\cos\left[\frac{(2i-1)\pi}{2K}q\right] \tag{8}$$

where $q=0...K-1$, i is the subscript of CQT at the point corresponding to the linear distribution.

B. Cepstral Mean and Variance Normalization

Robustness is an important criterion for evaluating the performance of replay detection algorithms. In actual scenarios, when a replay detector trained on one dataset is applied to voices from a different dataset (for example, with different background/channel noises), generally the performance of the detector degrades due to the mismatch between both datasets. To improve the robustness of the detection algorithm, we introduce cepstral mean and variance normalization (CMVN) to eliminate the deviation caused by channel noise in the cepstral domain and the convolution noise in the time domain, such as channel distortion, corresponding to the additive deviation of the cepstral domain.

Cepstral mean and variance normalization is an effective feature normalization method. It matches mean and variance by converting each training and test samples to zero mean and unit variance [16]. Let x_t be the N-dimensional cepstral

feature vector at time t, and $x_t(i)$ represents the ith component of x_t. $x=\{x_1, x_2, ..., x_t, ..., x_T\}$ represents a voice segment of length T. CMVN first calculates the mean μ and the variance σ^2 using the maximum likelihood estimate for each feature dimension,

$$\mu_{ML}(i)=\frac{1}{T}\sum_{t=1}^{T}x_t(i) \quad 1\le i\le N \tag{9}$$

$$\sigma^2{}_{ML}(i)=\frac{1}{T-1}\sum_{t=1}^{T}(x_t(i)-\mu_{ML}(i))^2 \quad 1\le i\le N \tag{10}$$

then each dimension feature vector is normalized,

$$x_t(i)=\frac{x_t(i)-\mu_{ML}(i)}{\sigma_{ML}(i)} \quad 1\le i\le N, 1\le t\le T \tag{11}$$

C. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is the linear combination of multiple Gaussian distributions, which can well describe the probability density distribution of data. In the GMM-based classification system, the main purpose of model training is to estimate the parameters so that the Gaussian mixture distribution can better match the distribution of feature vectors in the training set. In the training stage, assuming $x=\{x_1, x_2, x_3, ..., x_N\}$ is the CQCC features vector of the training voice segment, it's likelihood for the features can be expressed as:

$$P(x\mid\lambda)=\sum_{i=1}^{N}\omega_i P(x\mid\lambda_i) \tag{12}$$

where $\lambda_i=(\mu_i, \sigma_i^2)$ is the parameters set, ω_i is a weighted value for each category, N is the dimension of the feature vector and $P(x|\lambda)$ is probability density function. All the parameters of GMM are trained via Expectation-Maximization (EM) algorithm. The EM algorithm is one of the most common methods for estimating the maximum likelihood method. It makes parameters converged to the local optimal solution through the iterative process.

In the test stage, the features of the suspected voice are first extracted, the result is determined on the score when matching with the trained GMM models. Assuming that λ_i is the parameter of ith submodel, and x_t is the feature vector of the tth frame of the suspected voice, the probability that the suspected voice comes for the ith targets s_i is as follows:

$$P(x\mid s=s_i)=\frac{1}{T}\sum_{t=1}^{T}P(x_t\mid\lambda_i) \tag{13}$$

Assuming s_0 is the GMM model trained by the original voice samples, and s_1 is the GMM model trained by replayed voice samples. Whether the suspected voice is replayed or original depends on the following formula:

$$\begin{cases} P(x\mid s=s_0)\ge P(x\mid s=s_1) & \text{the voice is original} \\ P(x\mid s=s_0)< P(x\mid s=s_1) & \text{the voice is replayed} \end{cases} \tag{14}$$

D. Proposed Replay Detection Algorithm

The proposed detection algorithm is based on CQCC acoustic features and GMM classifier. CQCC have high resolution in high frequency bands compared with traditional features such as MFCC. The difference between the original voice and the replayed voice are easier to be distinguished. As a result, the original voice samples and replayed voice samples are used to train the GMM model respectively. When the suspected voice segment is put into the classifier, the voice can be distinguished by comparing the probability on each model. Hence, such an algorithm can be used to detect the replayed voice. The proposed algorithm consists of training and testing stages, as shown in Figure 2.

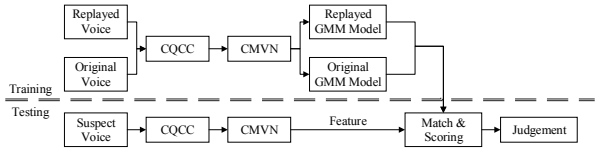

Figure 2. Diagram of proposed replay detection algorithm.

In the training stage, the CQCC features of the original voice and the replayed voice are first extracted based on Formula (8), and then the extracted features are normalized respectively using CMVN as shown in Formula (11). Finally, the GMM models for the replayed and original voices are trained separately using these features.

In the testing stage, the CQCC features of the suspect voice are first extracted and then normalized to obtain the corresponding acoustic feature. Then the feature vector is matched with the GMM models trained by original and replayed voices. Finally, whether the suspected voice is replayed or not depends on the score of the GMM model. Formula (14) gives the probability on GMM model which indicates the voice is more likely to be the original one or the replayed one.

III. Results and Discussion Experimental Results

A. Experiment Setup

In the experiments, ASVspoof 2017 database is adopted to evaluate the performance of algorithm proposed, which was collected in order to foster the development of countermeasures to protect ASV systems from replay spoofing attacks. In this database, there are 4724 voice samples in the training set, including 2267 original voice samples and 2457 replayed voice samples. And the test set consists of 13306 voice samples, including 1298 original voice samples and 12008 replayed voice samples. replayed voice samples in the test set are more complex than those in training set, such as different recording environment and spoofing devices. Each voice is WAV, 16KHz sampling rate, 16-bit quantization and mono.

The experimental settings are as follows: maximum frequency $F_{max}=F_{NYQ}$, where F_{NYQ} is Nyquist frequency and its size is $F_s/2$ (F_s is the sampling ratio). The minimum frequency $F_{min}=F_{max}/2^{oct}$, oct is an adjustable parameter to determine F_{min}, further determine the entire sampling range. The sampling period d is 16. The feature dimension K of CQCC is set to 19 to confirm whether the higher order coefficients contains useful additional information.

B. Experimental Result and Analysis

1) Comparison to traditional features

In Figure 3, we compare the performance of our proposed algorithm and other replay detection algorithms. From Figure 3, we can see that the proposed normalized CQCC obtained the best EER value at 15.96% as compared to other features. Meanwhile, relative improvements of ~34.7% (from ~23.64% to 15.28%) for CQCC and ~54% (from ~37.08% to 17.81) for MFCC in EER are achieved when the proposed CMVN normalization is applied.

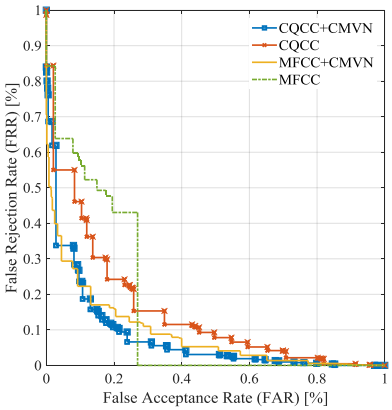

Figure 3. DET curve of proposed algorithm and traditional algorithms.

Generally, various randomness will be involved during GMM training stage, which will lead the model converge in local optimal solutions. The detection accuracy will be fluctuated due to such randomness. Hence, in this paper, we repeat the experiment for 50 times to analyze the fluctuation of EER. The statistical results are shown in Figure 4, where the error bars at the top of the histogram indicate the variance of EER fluctuations in repeated experiments. Firstly, it can be seen that the fluctuation caused by the training randomness will be reduced by increasing the number of GMM kernels, but the effect is limited. Secondly, with the same number of GMM kernel, the proposed normalization can greatly reduce the fluctuation in detecting performance. In addition, since the normalization can eliminate the deviation caused by the channel noise in the cepstral domain and the convolution noise in the time domain, the redundant features are optimized and lower EER values are obtained when different number of GMM kernel is adopted. From Figure 4, the best EER is achieved (from 15.21% to 15.35%) when the number of GMM kernel is 512.

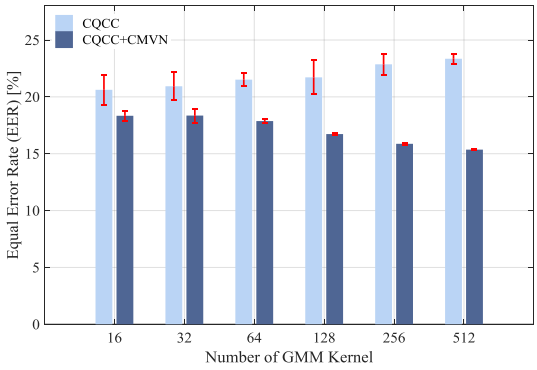

Figure 4. EER fluctuation under diff erent number of GMM kernel.

2) Text-Independent Evaluation

Since text-independent replayed detection is more practical, in this experiment, we evaluate the proposed algorithm on the voice samples with different content respectively. The ASVspoof 2017 database contains 10 English short sentences and the test set is divided into 10 groups, each group has about 1000 replayed voice samples.

TABLE I. EER OF SPEECH WITH DIFFERENT CONTENT.

Group	Number of Samples	EER
S01	1392	13.74
S02	1381	20.46
S03	1366	14.14
S04	1282	13.79
S05	1217	13.88
S06	1449	12.76
S07	1342	15.4
S08	1359	19.34
S09	1229	15.05
S10	1226	12.67

The experimental results are shown in Table I. It can be seen that only two groups of voice samples (S02 and S08) have higher EER values than the average level of 15.28%, and the values of rest groups are lower than the average level. It indicates that the proposed detection algorithm has a stable performance for different voice contents and can be used for text-independent replay detection system.

3) Cross-Devices Evaluation

In real scenarios, the replayed voice against the voiceprint authentication system may be come from different playback devices. Hence, it is necessary to reveal the effect of various spoofing devices on the detection algorithm.

In the paper, we choose Dynaudio BM5A speaker and VIFA M10MD-39-08 speaker as the playback device, and Røde NT2 microphone and Røde smartLav+ microphone as the recording device. For they have relatively more voice samples in the database. The number of samples in the cross-experiment is shown in Table II. In the training set, there are 2267 original voice samples and about 100 replayed voice samples meet the requirements of the cross-experimental device. And in the test set, there are 1298

original voice samples and about 100 replayed voice samples meet the same requirements.

The detection rates of the cross-device experiment are shown in Table II. When exchange both the recording device and playback device, as the two diagonals in table present, the detection rates is a little worse than the non-cross results (from ~96% to ~93%). However, it can be seen that the detection rates are higher than 90% under any circumstance, with the EERs lower than 10%, indicating that, the proposed algorithm achieves outstanding detection performance and has strong robustness.

TABLE II. DETECTION RATE [%] OF THE CROSS-DEVICE EXPERIMENT.

	Recording Device			**Test Set**				
	Playback Device	***Recording Device***	***Number of Samples***	Dynaudio BM5A speaker		VIFA M10MD-39-08 speaker		**EER**
				Røde NT2 microphone	Røde smartLav+ microphone	Røde NT2 microphone	Røde smartLav+ microphone	
				169	116	105	84	
Training Set	Dynaudio BM5A speaker	Røde NT2 microphone	95	92.30	93.71	95.15	94.28	9.04
		Røde smartLav+ microphone	95	91.07	98.16	94.37	94.65	6.28
	VIFA M10MD-39-08 speaker	Røde NT2 microphone	95	91.96	94.34	96.86	95.56	7.19
		Røde smartLav+ microphone	95	92.84	94.13	97.01	96.53	5.74

IV. CONCLUSION

In this paper, an algorithm for replay attack detection is proposed. The basic idea of the proposed algorithm is to utilize the difference between the original and replayed voice in the high frequency band. Constant Q cepstral coefficients are extracted as acoustic feature and CMVN is used as the postprocess on features. Gaussian Mixture Model is adopted as the classifier to detect the replayed voice. The experimental results on ASVspoof 2017 database show that the proposed algorithm achieves better EER performance than other traditional methods and especially achieves much better robustness to various voice contents and spoofing devices. Next, it is an important and necessary task to further improve the performance of the algorithm by using other post-processing methods on extracted features. On the other hand, replayed voice detection algorithm based on deep learning is also one of the directions worth studying.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China, grant numbers (61300055, 61672302); Natural Science Foundation of Zhejiang, grant number LY17F020010; Natural Science Foundation of Ningbo, grant number 2017A610123.

REFERENCES

[1] H. Wu, Y. Wang, and J. Huang, "Identification of Electronic Disguised Voices," IEEE Transactions on Information Forensics and Security, vol. 9, 2014, pp. 489-500.

[2] M. C. Ozbay, A. Khodabakhsh, A. Mohammadi, and C. Demiroglu, "Spoofing attacks to i-vector based voice verification systems using statistical speech synthesis with additive noise and countermeasure," the 24th European Signal Processing Conference (EUSIPCO), 2016.

[3] T. Toda et al., "The Voice Conversion Challenge 2016," in Interspeech 2016, 2016,pp. 1632-1636.

[4] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," Multimedia Tools and Applications, vol. 75, 2016, pp. 5311-5327.

[5] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, and K. A. Lee, "RedDots Replayed: A New Replay Spoofing Attack Corpus for Text-Dependent Speaker Verification Research," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, 2017, pp. 5395-5399.

[6] C. Si et al., "You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones," in 2017 IEEE 37th International Conference on Distributed Computing Systems, 2017, pp. 183-195.

[7] W. Shang and M. Stevenson, "A playback attack detector for speaker verification systems," in 2008 3rd International Symposium on Communications, Control and Signal Processing, 2008, pp. 1144-1149.

[8] J. Gałka, M. Grzywacz, and R. J. S. C. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," vol. 67, 2015, pp. 143-153.

[9] L. Zhang, J. Cao, M. Xu, and Z. Fang. "Prevention of impostors entering speaker recognition systems," Journal of Tsinghua University Science and Technology, no. s1, 2008, pp. 699-703.

[10] Z. F. Wang, G. Wei, and Q. H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in 2011 International Conference on Machine Learning and Cybernetics, 2011, pp. 1708-1713 .

[11] X. Lin, J. Liu, and X. Kang, "Audio Recapture Detection With Convolutional Neural Networks," IEEE Transactions on Multimedia, vol. 18, 2016, pp. 1480-1487,.

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

[13] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in ICASSP '78. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, 1978, pp. 375-378.

[14] H. Delgado et al., "Further Optimisations of Constant Q Cepstral Processing for Integrated Utterance and Text-dependent Speaker Verification," in 2016 IEEE Spoken Language Technology Workshop (SLT), 2017, pp. 179-185.

[15] I. Kaminskyj and T. Czaszejko, "Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using k NNC," Journal of Intelligent Information Systems,vol. 24, 2005, pp. 199-221,.

[16] N. V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using Bayesian framework," in 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013, pp. 156-161.