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

# Device Feature Extractor for Replay Spoofing Detection

*Chang Huai You*[1], *Jichen Yang*[2*], *Tran Huy Dat*[1]

[1]Institute for Infocomm Research, A*Star, Singapore
[2]Department of Electrical and Computer Engineering, NUS, Singapore
echyou@i2r.a-star.edu.sg, eleyji@nus.edu.sg, hdtran@i2r.a-star.edu.sg

## Abstract

Device feature, which contains the information of both recording channel and playback channel, is the critical trait for replay spoofing detection. So far there have not been any technical reports about the usage of device information in spoofing detection for speaker verification. In this paper, we propose to build a replay device feature (RDF) extractor on the basis of the genuine-replay-pair training database. The RDF extractor is trained in constant-Q transform (CQT) spectrum domain. A bidirectional long short-term memory (BLSTM) is used in the neural network and finally the RDF extractor is formed by applying discrete cosine transform (DCT) to the output vector of the BLSTM. The experimental result on ASVspoof 2017 corpus version 2.0 shows that equal error rate (EER) of replay detection system with the proposed RDF reaches 15.08%. Furthermore, by combining the RDF with constant-Q cepstral coefficients plus log energy (CQCCE), the EER of the detection system can reduce to 8.99%. In addition, the experimental results also show that the RDF has much complementarity with conventional features.

**Index Terms**: replay speech detection, recording device, playback device, speaker verification [1]

## 1. Introduction

The false acceptance rates of state-of-the-art speaker verification systems have been significantly increased with spoofing attacks. There are four typical spoofing attacks, i.e. voice conversion [1] [2], speech synthesis [3], impersonation [4] and replay [5] [6]. Among the four attacks, replay spoofing is the most difficult to be detected [7] and easily applied with very low technical cost [8]. In the case of replay attacks, prerecorded utterance of the target speaker is played to automatic speaker verification (ASV) system to gain unauthorized access. In such a scenario, the speech signal carries the characteristics of the intermediate recording device as well. During a replay attack, the recorded utterance spoken by a target speaker is played back to an automatic speaker verification (ASV) system. Obviously the replay speech signal itself contains the characteristics of the intermediate replay device that includes recording device and playback device.

Replay speech detector usually contains front-end feature extraction and back-end classification. For frond-end feature extraction part, many effective features have been studied for replay speech detection since ASVspoof 2017 challenge [5, 6]. The features used for spoofing detection is categorized into two types: handcrafted features and deep features. While the handcrafted features are obtained by some transformation to convert signal from time domain into variants of frequency domain, the deep features are usually obtained by deep learning of neural network such as deep neural network (DNN), recurrent neural network (RNN), and convolutional neural network (CNN). There have been several works on how to extract deep features for replay spoofing detection. In [9], log power magnitude from CQT and fast Fourier transform (FFT) are used as the input of a light CNN, then the deep feature is generated from the CNN. In [10], both DNN-based frame-level and RNN-based sequence-level are developed to extract respective deep features, moreover both DNN-based and RNN-based deep features as spoofing identity representation are combined to achieve better performance.

In back-end classification part, most models for the spoofing detection are based on Gaussian mixture model (GMM) [12] [13] [14], support vector machine (SVM) [15] [9] [11], residual network (ResNet) [16] and BLSTM [17] [18]. Neural network classifiers are reported to give better performance than GMM [19].

Although device information is important for replay spoofing detection, there have been no any report about how to extract device information. In this paper, we focus on device information study and propose a deep feature extraction method to depict the trait of replay device. We found that the RDF gives good effectiveness on determining the device-involved (replay) speech and non-device-involved (genuine) speech. In particular, we adopt BLSTM [20] to model the replay device where log magnitude spectrum of CQT feature is used as input to BLSTM, the BLSTM is trained by using genuine-playback-pair database from ASVspoof 2017 corpus. Finally the device feature is generated via DCT which is applied to the output of the BLSTM.

The remainder of the paper is organized as follows. Section II briefs the constant-Q transform domain in which we develop the device feature. In Section III, we introduces the channel system of replay device and propose a device feature extraction method. In Section IV, the experimental results and corresponding analysis on ASVspoof 2017 corpus version 2.0 are described. In Section V, conclusions of the work are presented.

## 2. Constant-Q transform

CQT was proposed in [21] [22]. Q is defined as the ratio of center frequency to bandwidth $Q = \frac{f_k}{\delta_f}$ where $f_k$ and $\delta_f$ stand for centre frequency and bandwidth respectively. In CQT, Q is constant for the entire spectrum, which can make CQT have high frequency resolution at low frequency and high temporal resolution at high frequency. In contrast to CQT, discrete Fourier transform (DFT) has a variable Q factors in entire spectrum.

For a discrete time domain signal $x(n)$, its CQT $\mathbf{Y}(k,n)$ is defined as follows:

$$\mathbf{Y}(k,n) = \sum_{u=n-\lfloor \frac{N_k}{2} \rfloor}^{n+\lfloor \frac{N_k}{2} \rfloor} x(u) a_k^*(u - n - \frac{N_k}{2}) \qquad (1)$$

* Corresponding author



http://dx.doi.org/10.21437/Interspeech.2019-2137

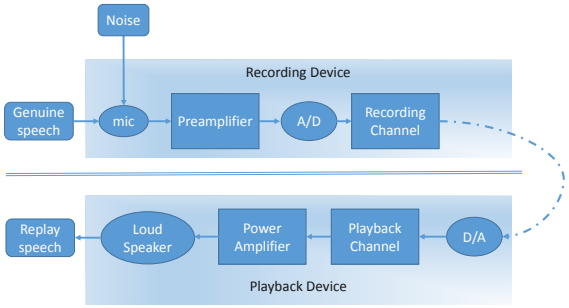


Figure 1: *Diagram of the replay speech generation process.*

where $k = 1, 2, ...K$ is the frequency bin index, $N_k$ are the variable window lengths, $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$, and $\lfloor \bullet \rfloor$ denotes rounding towards negative infinity. The basic functions $a_k(n)$ are complex-valued time-frequency atoms and are defined as follows

$$a_k(n) = \frac{1}{C}\nu(\frac{n}{N_k})exp[i(2\pi n\frac{f_k}{f_s} + \phi_k)] \tag{2}$$

where $f_k$ is the centre frequency of the $k$-th bin. $f_s$ is the sampling rate, and $\nu(t)$ is a window function. $\phi_k$ is a phase offset. C is a scaling factor and

$$C = \sum_{m=-\lfloor \frac{N_k}{2} \rfloor}^{\lfloor \frac{N_k}{2} \rfloor} \nu(\frac{m + \frac{N_k}{2}}{N_k}) \tag{3}$$

where $m$ stands for the sample index in the window.

## 3. Proposed Replay Device Feature Extractor

We start our proposed RDF extraction from the analysis of the device system, then we introduce the training of the neural network, finally the RDF extractor is developed.

### 3.1. Analysis of Replay Device System

This section introduces replay speech generation process and the relationship between genuine speech and replay speech.

The replay device is composed of recording device and playback device. The recording device is used to record genuine speech and the playback device is to play the recorded speech. Recording device contains three modules: preamplifier, filter and analog-to-digital converter (A/D), and playback device contains three modules: digital-to-analog converter (D/A), power amplifier and loudspeaker.

It may be reasonable to assume A/D converter in recording device and D/A converter in playback device are ideal, since the quantization error can be ignorable. Fig.1 shows the process of replay speech generation. The preamplifier in recording device and amplifier in playback device are used to amplify signal linearly.

Let $s_g(t)$ be a genuine speech, $k_r$ the recording preamplifier coefficient, and $h_f(t)$ the impulse response of filter, we have

$$s_r(t) = k_1(s_g(t) + n(t)) * h_f(t) \tag{4}$$

where $n(t)$ denotes the environmental noise in recording environment.

Let $k_p$ be the power amplification factor in playback device, $h_l(t)$ the impulse response of playback channel and loudspeaker. Given input $s_r(t)$, we have the output of the playback speech $s_\rho(t)$ as follows

$$s_\rho(t) = [k_2 s_r(t) * h_p(t)] * h_l(t) \tag{5}$$

Consequently, we have the replay speech as follows

$$\begin{aligned} s_\rho(t) &= k_2[k_1(s_g(t) + n(t)) * h_f(t)] * h_p(t) * h_l(t) \\ &= k_1 k_2 s_g(t) * h_f(t) * h_p(t) * h_l(t) \\ &\quad + k_1 k_2 n(t) * h_f(t) * h_p(t) * h_l(t) \end{aligned} \tag{6}$$

By applying CQT to eq.(6), we have

$$\begin{aligned} S_\rho(jw) &= k_1 k_2 S_g(jw) H_f(jw) H_p(jw) H_l(jw) \\ &\quad + k_1 k_2 N(jw) H_f(jw) H_p(jw) H_l(jw) \end{aligned} \tag{7}$$

where $S_\rho(jw)$, $S_g(jw)$, $H_f(jw)$, $H_l(jw)$, $N(jw)$ are CQTs of $s_\rho(t)$, $s_g(t)$, $h_f(t)$, $h_l(t)$, $n(t)$ respectively.

Ignoring the input noise to microphone, which is the environmental interference, we just consider the replay device affect to the replay speech. We have the replay speech simplified below

$$\tilde{S}_\rho(jw) = S_g(jw) H_\varpi(jw) \tag{8}$$

where $H_\varpi(jw) = k_1 k_2 H_f(jw) H_p(jw) H_l(jw)$. It is clear that $H_\varpi(jw)$ represents the generic replay device system function in any variant of spectrum domain, e.g. CQT domain.

From Eq. (8), $\log|H_\varpi(jw)|$ can be obtained

$$\log|H_\varpi(jw)| = \log|S_\rho(jw)| - \log|S_g(jw)| \tag{9}$$

Eq. (9) describes the relationship among the input genuine speech, replay device system and the replay spoofing speech.

### 3.2. Training of Neural Network

According to the Eq. (9), a system which can extract device information from the pair of input and spoofing speech is proposed here.

For the detection of replay spoofing speech, we consider to use the replay device function $\log|H_\varpi(jw)|$ for the following situation: we train the BLSTM network parameters using the genuine-replay-pair training database. (a) If the input is replay speech, the output is the log-magnitude of device system ($\log|H_\varpi(jw)|$), which can be computed using difference between replay speech and its corresponding genuine speech according to Eq. (9). (b) If the input is genuine speech, the output of the BLSTM is zero to indicate non-device involved. As a result, the neural network system becomes a replay device function generator.

Regression method is used to train the BLSTM here. In other words, square error between the input and output as the training function for the neural network. Fig. 2 shows how to train BLSTM.

From Fig. 2, it can be found that BLSTM training neural network consists of input layer, hidden layer and output layer. As per Eq. (9), both the input ($X_I$) and the output ($X_O$) of BLSTM training are the log magnitude spectrum of CQT (CQLM). In addition, $X_I$ could be the CQLM feature of either replay speech or genuine speech, the regression target $X_O$ is thus defined as follows.

$$X_O = \begin{cases} X_I - X_G & X_I \text{ is the CQLM of replay speech} \\ 0 & X_I \text{ is the CQLM of genuine speech} \end{cases} \tag{10}$$

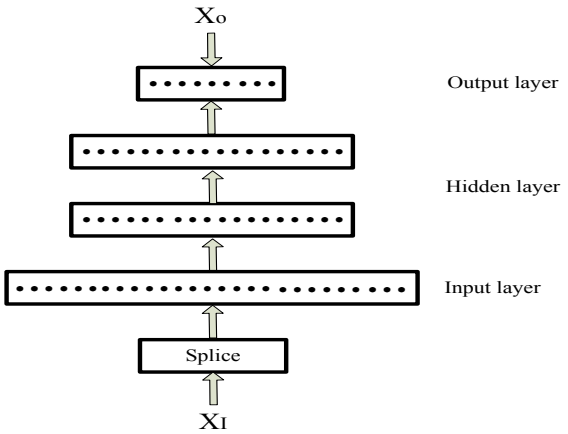


Figure 2: *Schematic diagram of BLSTM training.*

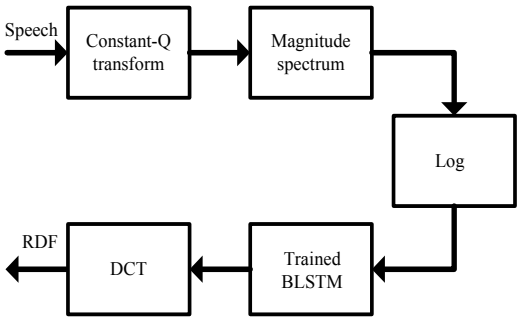


Figure 3: *Schematic diagram of the replay device feature (RDF) extraction.*

where $X_G$ represents the CQLM of the corresponding genuine speech.

### 3.3. RDF Extraction

Fig. 3 shows a block diagram representation for deriving the proposed device feature generator. For a given speech signal, CQT is applied first to convert the signal from the time domain into CQT spectral domain, followed by magnitude and logarithm computation. Finally DCT is applied on the output vector of the BLSTM to generate the RDF.

## 4. Performance Evaluation

We conduct to evaluate the proposed device feature on ASVspoof 2017 corpus version 2.0 (ASVspoof-2017-V2) for replay spoofing detection [23]. ASVspoof-2017-V2 is constituted by three subsets: training data, development data and evaluation data.

The CQT parameters are properly configured as follows: number of frequency bins in each octave is set to 96, number of octaves is set to 9, sampling period is 16, and gamma is chosen as 3.3026. The settings are consistent with those in [24].

1507 genuine-replay-pairs of the speech utterances from ASVspoof 2017 V2 training dataset are used to train BLSTM of the device feature extractor. In our experiment, the static dimension of the device feature set to 863 gives best performance in terms of EER. Many works [25, 26, 14] have shown static feature degrades the performance in replay speech detection, therefore only dynamic features including delta (D), acceleration (A) and delta-acceleration concatenation (DA) are investigated.

To examine the performance of our proposed device feature, we use a series of four-layer DNNs for classification. The input layer of the DNNs are constituted by splicing 11-frames centralized at current frame. The DNNs have two hidden layers with 512 nodes in each layer and its output layer has 2 nodes.

### 4.1. Development of RDF Extractor

We trained DNN model on the development set, where 3,014 genuine/replay utterances from training set are used to train the DNN models for the classification in replay spoofing detection. We show the experimental results on ASVspoof 2017 V2 development dataset using different BLSTM neural network architectures with different dynamic features (i.e. D, A, and DA) for the device feature extractor in Table 1, where ILN, HL1N, HL2N and OLN denote the numbers of input layer nodes, first hidden layer nodes, second hidden layer nodes and output layer nodes respectively. It can be seen: (1) No matter which network structure of BLSTM is, RDF-A gives the best performance on ASVspoof 2017 V2 development set. RDF-D gives the second best performance for the first network structure. While the RDF-DA gives the second best performance for the another three network structures. (2) The RDF-A with network structure 9493:3072:1536:863 performs best on ASVspoof 2017 V2 development set among the four network structures. Thus we use the RDF-A and RDF-DA with network structure 9493:3072:1536:863 to evaluate the RDF performance on ASVspoof 2017 V2 evaluation set.

Table 1: *Experimental results (EER (%)) on ASVspoof 2017 V2 development set using different feature combinations and BLSTM configurations.*

| BLSTM Configuration | Feature Combinations | | |
|---|---|---|---|
| ILN : HL1N : HL2N : OLN | D | A | DA |
| 9493 : 863 : 1536 : 863 | 40.21 | 37.38 | 41.92 |
| 9493 : 2048 : 1536 : 863 | 41.95 | 39.97 | 41.21 |
| 9493 : 3072 : 1536 : 863 | 40.95 | 37.19 | 40.44 |
| 9493 : 4096 : 1536 : 863 | 40.31 | 38.00 | 39.82 |

### 4.2. Experimental Analysis and Performance Evaluation

To evaluate the device feature on the evaluation set, 4,724 utterances from training and development databases are used to train the DNN models for the classification in replay spoofing detection. Our experiment on ASVspoof 2017 V2 evaluation set shows that the EERs using the RDF-A and RDF-DA are 17.00% and 15,08% respectively. It indicates the RDF-DA performs better than the RDF-A. In contrast to the EER result in Table 1, we found an interesting phenomenon, the RDF-A performs better than the RDF-DA on ASVspoof 2017 V2 development set while the RDF-DA performs better than the RDF-A on ASVspoof 2017 V2 evaluation set. The reason behind may be difference between ASVspoof 2017 V2 development set and ASVspoof 2017 V2 evaluation set, such as some types of replay speech only appear in ASVspoof 2017 V2 evaluation set. Consequently, in the remainder of the paper, we only use DA dynamic feature for all features on the evaluation set.

In handcrafted features (e.g. MFCC, CFCC and CQCC), DCT [27] has three functions: de-correlation of the feature dimensions [28], energy concentration and dimension reduction. In this paper, we investigate whether DCT has the same func-

tions in RDF extraction like in handcrafted feature extraction.

We investigate the effect of DCT in the RDF feature. From Table 2, it can be seen that the performance of the RDF with DCT is much better than the RDF without DCT, which means that DCT plays the important role in the RDF, which is the same as the one in tradition handcrafted feature extraction.

Table 2: *Experimental results (EER (%)) on ASVspoof 2017 V2 evaluation set comparison etween RDF-with-DCT and RDF-without-DCT.*

| Feature | DCT | EER |
|---|---|---|
| RDF-with-DCT | Yes | 15.08 |
| RDF-without-DCT | No | 18.41 |

Table 3: *Experimental results (EER (%)) on ASVspoof 2017 V2 evaluation set using the RDF under different static dimension of RDF.*

| Static dimension | EER |
|---|---|
| 13 | 36.25 |
| 20 | 35.59 |
| 30 | 37.06 |
| 256 | 33.80 |
| 512 | 29.41 |
| 863 | 15.08 |

For handcrafted feature like MFCC for speech or speaker recognition, the dimension of static feature is usually selected to be low such as 13, 20 or 30 because the low dimension is able to concentrate most of energy and thus reduce the sensitivity to some interference. Therefore, the low dimension selection gives the best performance since the robustness is strengthened. In order to investigate the relationship between the performance of the RDF and the size of the RDF dimension, not only the low static dimension but also high static dimension are surveyed. From Table 3, it can be seen that low static dimension such as 13, 20 and 30 gives poor performance for the RDF. For this point, there is much difference between handcrafted feature and the device feature. When static dimension equals 863, the RDF performance reaches the best. The reason may be that discriminative information of the RDF hides not only in low dimension but also in high dimension after DCT. In other words, DCT is helpful for de-correlation, but cannot be used for dimension reduction for the device feature;

In this paper, we show the following five conventional features for comparison. They are MFCC, instantaneous frequency cosine coefficient (IFCC) [29], linear frequency cepstral coefficient (LFCC) [30], and CQLM. Since log energy has been found to improve the performance of CQCC [23], CQCC plus log energy (CQCCE) is selected for comparison. The classifiers are setup for the spoofing detection with MFCC, CQLM, CQCC and CQCCE respectively in the exact same training database as that with RDF, where the respective four-layer DNNs are trained. The static dimensions of 13, 20, 20, 30, 31 and 863 are properly selected for MFCC, IFCC, LFCC, CQCC, CQCCE and CQLM respectively.

Table 4 gives the performance comparison between the proposed feature and typical conventional features on ASVspoof 2017 V2 evaluation set. It can be seen that the performance of CQLM, CQCC, CQCCE and RDF are better than MFCC. The reason may be that CQT is a long-term transform and it can provide more frequency details than discrete Fourier transform that is a short-term transform. It also can be seen that though RDF only consists of devices information, its performance is better than the performances of CQLM and CQCC which have multi-aspect information. The performance of CQCCE is much better than the RDF, which means that log energy is very helpful and is of much complementary space with CQCC for replay speech detection.

In Table 4, '$\alpha$ + $\beta$' denotes the two systems '$\alpha$' and '$\beta$' to be fused at the score level. It can be seen that: (1) By score fusion, the EER of RDF + MFCC drops down to 12.50% from its original EER values of 15.08% (RDF) and 23.79% (MFCC). It means that RDF has much complementary space with MFCC. (2) Though RDF is built up from CQLM, the score level fusion of RDF + CQLM improves much from its original respective systems. The reason may be that most of the information in RDF is device and environment information while CQLM has multi-aspect information, and the information in RDF and in CQLM is complementary with each other. (3) RDF also has much complementary information with CQCC and CQCCE, it is surprisingly found that the EER of RDF + CQCCE reaches 8.99%. (4) The fusion of three systems (RDF+CQCCE+MFCC) gives the best performance with the EER of 8.67%.

## 5. Conclusion

Device feature provides a way to depict the characteristics of the replay spoofing speech. Different from existing methods for replay speech detection, this paper addresses the problem of extraction of device information for replay speech detection. We have proposed a method to extract RDF based on CQT, BLSTM and DCT. The experimental results show that the proposed RDF can produce a satisfactory result, which is better than typical conventional features. In addition, we found that DCT only plays the role of de-correlation of the feature but cannot be used to reduce the dimension for the RDF; this is different from traditional handcrafted feature extraction. It has been observed that the proposed RDF has much complementary capability with typical conventional features at score level.

Table 4: *Comparison with some commonly used features on ASVspoof 2017 V2 evaluation set in terms of EER (%).*

| Features | EER | Features | EER |
|---|---|---|---|
| IFCC | 34.61 | RDF+IFCC | 14.83 |
| MFCC | 23.79 | RDF+MFCC | 12.50 |
| CQLM | 16.61 | RDF+CQLM | 10.78 |
| LFCC | 16.60 | RDF+LFCC | 10.12 |
| CQCC | 15.46 | RDF+CQCC | 10.62 |
| RDF | 15.08 | | |
| CQCCE | 10.61 | RDF+CQCCE | 8.99 |
| CQCCE+CQCC | 10.61 | RDF+CQCCE+CQCC | 8.99 |
| CQCCE+CQLM | 10.61 | RDF+CQCCE+CQLM | 8.99 |
| CQCCE+LFCC | 10.56 | RDF+CQCCE+LFCC | 8.98 |
| CQCCE+IFCC | 10.25 | RDF+CQCCE+IFCC | 8.81 |
| CQCCE+MFCC | 9.58 | RDF+CQCCE+MFCC | 8.67 |
| CQCCE+LFCC+CQLM | 10.56 | | |
| CQCCE+LFCC+IFCC | 10.25 | | |
| CQCCE+IFCC+CQLM | 10.24 | | |
| CQCCE+MFCC+CQLM | 9.59 | | |
| CQCCE+MFCC+LFCC | 9.58 | | |
| CQCCE+MFCC+IFCC | 9.02 | | |

# 6. References


[1] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[2] Xiaohai Tian, Siuwa Lee, Zhizheng Wu, Eng Siong Chng, and Haizhou Li, "An example-based approach to frequency warping for voice conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1875, 2017.

[3] Matt Shannon, Heiga Zen, and William Byrne, "Autoregressive models for statical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587–597, 2013.

[4] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne Marial Laukkanen, "Automatic versus human speaker verification:the case of voice mimicrys," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[5] Tomi Kinnunen, Md Sahidullah, Mauro Falcone, Luca Costantini, Rosa Gonzalez Hautamaki, Dennis Thomsen, Achintya Sarkar, Zheng-Hua Tan, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Ville Hautamaki, and Kong Aik Lee, "RedDots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 5395–5399.

[6] Tomi Kinnunen, and Héctor Delgado Md Sahidullah, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2–6.

[7] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Imposture classification for text-dependent speaker verification," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),*, 2014, pp. 739–743.

[8] Zhifeng Wang, Qianhua He, Xueyuan Zhang, Haiyu Luo, and Zhuosheng Su, "Playback attack detection based on channel pattern noise," *Journal of South China University of Technology (Natural Science Edition)*, pp. 1708–1713, 2011.

[9] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudasher, and Vadim Shchemelinin, "Audio replay attack detection with deep learning framework," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 82–86.

[10] Y. M. Qian, N. X. Chen, and K. Yu, "Deep features for automatic spoofing detection," in *Speech Communication*, vol. 85, Dec 2016, pp 43-52

[11] Xianliang Wang, Yanhong Xiao, and Xuan Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 32–36.

[12] Marcin Withowski, Stanislaw Kacprasko, Piotr Zelasko, Konrad Kowlczyk, and Jakub Galka, "Audio replay attack detection using high-frequency features," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 27–31.

[13] Hemant A. Patil, Madhu R. Kamble, Tanvina B. Patel, and Meet Soni, "Novel variable length teager engery separation based on instantaneous frequency features for replay detection," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 12–16.

[14] Jichen Yang, Rohan Kumar Das, and Haizhou Li, "Extended constant-Q cepstral coefficients for detection of spoofing attacks," in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, Hawaii, 2018.

[15] Parav Nagarshenth, Elie Khoury, kailash Patil, and Matt Garland, "Replay attack detection using DNN for channel discrimination," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 97–101.

[16] Mohit Jain Francis Tom and Prasenjit Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 681–685.

[17] Weicheng Cai, Danwei Cai, Wenbo Liu, Gang Li, and Ming Li, "Countermeasures for automatic speaker verification repaly spoofing attack: on data augmentation, feature representation, classification and fusion," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 17–21.

[18] K N R K Raju Alluri, Sivanand Achanta, Sudarsana, Sudarsna Reddy Kadiri, Suryakanth V Gangasheetty, and Jumar Vuppala, "SFF anti-spoofer: Iiit-h submission for automatic speaker verificatio spoofing and countermeasures channlenge 2017," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 107–111.

[19] Frank Seide and Amit Agarwal, "CNTK: Microsoft's open-source deep learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 2135–2135.

[20] Trias Thireou and Martin Reczko, "Bidirectional long short-term networks for predicting the subcellular localization of eukaryotic proteins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 441–446, 2007.

[21] James Youngberg and Steven Boll, "Constant-Q signal analysis and synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1978, pp. 375–378.

[22] Judith C. Brown, "Calculation of a constant Q spectral transform," *Journal of Acoustical Society of America*, vol. 89, 1991.

[23] Héctor Delgado, Massimilliano Todisco, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, KongAik Lee, and Junichi Yamagishi, "ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *In speaker and language recognition workshop (ODYESSEY)*, 2018, pp. 296–303.

[24] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic Speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[25] Jichen Yang and Leian Liu, "Playback speech detection based on magnitude-phase spectrum," *Electronics Letters*, vol. 54, no. 14, pp. 901–903, 2018.

[26] Jichen Yang, Changhuai You, and Qianhua He, "Feature with complementarity of statistics and principal information for spoofing detection," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 651–655.

[27] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Signal Processing*, vol. C-24, 1974.

[28] Qi Li and Yan Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1791–1801, 2011.

[29] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication* vol. 81, pp. 54-71, 2016.

[30] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," *in Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on. IEEE*, pp. 18, 2013.