

# Study on Feature Complementarity of Statistics, Energy, and Principal Information for Spoofing Detection

LEIAN LIU<sup>1</sup> AND JICHEN YANG<sup>2</sup> , (Member, IEEE)

<sup>1</sup>School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China

<sup>2</sup>Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117575

Corresponding author: Jichen Yang (nisonyoung@163.com)

This work was supported by the Science and Technology Planning Project of Guangdong Province under Grant (2017A070709012) and higher education teaching reform project of Guangdong province "Reform and practice of the training mode of network engineering talents based on the cooperation of school-school and school-enterprise" (Official document by Department of education of Guangdong province ([2018] no. 180)).

**ABSTRACT** Conventional speaker verification systems become frail or incompetent while facing attack from spoofed speech. Presently many anti-spoofing countermeasures have been studied for automatic speaker verification. It has been known that the salient feature is of a more important role rather than the selection of classifiers in the current research field of spoofing detection. The effectiveness of constant-Q transform (CQT) has been demonstrated for anti-spoofing feature analysis in many research literatures on automatic speaker verification. On the basis of CQT-based information sub-features, i.e. octave-band principal information (OPI), full-band principal information (FPI), short-term spectral statistics information (STSSI) and magnitude-phase energy information (MPEI), three concatenated features are proposed by investigating their information complementarity in this paper, the first one is constant-Q statistics-plus-principal information coefficients (CQSPIC) by combining OPI, FPI and STSSI; the second one is constant-Q energy-plus-principal information coefficients (CQEPIC) by combining OPI, FPI and MPEI and the third one is constant-Q energy-statistics-principal information coefficients (CESPIC) by combining OPI, FPI, MPEI and STSSI. In this paper, we set up deep neural network (DNN) classifiers for evaluation of the proposed features. Experiments show that the proposed features can outperform some commonly used features meanwhile the proposed systems give better or comparable performance comparing with state-of-the-art performance on ASVspoof 2019 logical access and physical access corpus.


**INDEX TERMS** Constant-Q transform, anti-spoofing countermeasure, automatic speaker verification.

## I. INTRODUCTION

Over the past decades, speaker verification technology has gradually matured in commercial products [1], [2]. The speaker verification is designed to determine whether a voice is from a claimant speaker. However, when an imposter forges speech signals to imitate an enrolled speaker, the speaker verification system becomes vulnerable to the spoofing attack. A conventional speaker verification system is relatively insensitive to spoofed speech since the model is trained only for ordinary speaker classification without considering the special situation of the spoofing attack. In order to render the speaker verification system practically viable, an indispensable countermeasure must be developed against the spoof-

ing situation. For this purpose, seeking a feature that is of discrimination between natural speech and spoofed speech becomes a critical factor [3]. There are four main challenging attacks from different sources: synthetic speech [3]–[5], voice converted speech [6]–[8], replay speech [9]–[11] and impersonation [12]. In this paper, we mainly focus on the first three attacks.

For synthetic speech detection, the goal of the feature design is to unveil the artificial characteristics of the speech that is produced using text-to-speech (TTS) method [3], [4], [7], [13]. The artificial characteristics is an essential property of the synthetic speech. On the other hand, long-term average spectra (LTAS) is a salient feature to represent the physical characteristics of the speaker, in particular, the vocal tract resonances [14]. Many features such as linear frequency cepstral coefficient (LFCC), linear prediction cepstral coefficient

The associate editor coordinating the review of this manuscript and approving it for publication was Zhanyu Ma .

(LPCC), perceptual PLPCC, mel frequency cepstral coefficient (MFCC) and inverted MFCC have been investigated for the synthetic speech detection [15]. The source of the synthetic speech is a totally artificial product.

Voice conversion [8] is another way of spoofing attack to speaker verification system. In speaker verification attacks, voice conversion is a process which converts or transforms an impostor's voice to an enrolled speaker's one [16]. The vulnerabilities of traditional speaker verification system may show up when the vocal tract information of target speaker is utilized for spoofing purpose. The speech generated by state-of-the-art voice conversion algorithm is of high-quality imitative effect to fool both human listener and automatic system. Because the converted speech originates from genuine speech, traditional detection may not be so effective to discriminate between converted speech and genuine speech. Most countermeasures adopts training examples to capture specific processing artifacts in order to learn the classification model, which is used to detect similarly treated speech. The weakness of the countermeasures is the constraint of the model to a particular voice conversion algorithm and is unlikely to be generalised to others. The source of the voice conversion is semi-artificial.

In contrast to speech synthesis and voice conversion which requires substantial artificial dedication, replay is just to playback a pre-recorded speech sample [17], [18]. The replay attack is a very low technology spoof and the most easily available for any impostor without prior speech processing knowledge. Replay gives a comprehensive risk to automatic speaker verification technology by using audio playback device. For replay speech detection, the challenging task is to discriminate between authentic human speech and replay speech [19], [20]. Physically replay speech go through both recording and playback channels, it leads to the channel distortion of the speech and thus the replay speech may have compressed spectral component in low and high frequency regions. The source of the replay is totally pure human speech.

The spoofing detection includes frontend feature extraction and backend classification. Countermeasures of spoofing attack are mainly focused on feature exploration. The classification for detection are normally built on traditional classification techniques such as Gaussian mixture model (GMM), support vector machine (SVM), deep neural network (DNN), convolution neural network (CNN), and recurrent neural network (RNN).

There are two clusters of features used for anti-spoofing detection: deep-learning generated feature and handcrafted feature. It is believed that the fusion of the handcrafted and deep-learning generated features can hopefully achieve a breakthrough improvement. The deep-learning feature performs overwhelming advantage on accuracy improvement, however it is based on certain training process and only suitable in the scope of training database circumstances. The handcrafted feature has demonstrated its efficacy in many spoofing detection literatures, it can be categorized into three

categories: long-term spectral statistics based feature [21], phase spectrum based feature [22], [23] and power spectrum based feature. In [24], [25], two types of long-term spectral statistics, i.e. first and second order statistics over the entire utterance in each of discrete Fourier transform (DFT) frequency bin, are concatenated to form a single vector representation of an utterance. In [9], [21] six long-term statistical parameters were introduced, which were obtained from DFT frequency bin and they are the minimal value, maximal value, average value, median, difference between the maximal and minimal values and standard deviation, respectively. The six statistical parameters can be concatenated to form a statistical feature. Typical phase spectrum based features are the cosine normalized phased feature, group delay, instantaneous frequency, and instantaneous frequency cosine coefficients [22], [23]. Although the performance of phase spectrum based feature is worse than traditional power spectrum based features, both features are of certain complementary information. Therefore, phase spectrum based features are often combined with power spectrum based features [26], [27].

There are many variants of the power spectrum based feature such as scattering cepstral coefficients [28], mel-warped overlapped block transformation and speech-signal frequency cepstral coefficients [5], and constant-Q cepstral coefficients (CQCC) [29], [30]. CQCC is the most widely used feature; it was firstly applied in synthetic and voice converted speech detection [29], then used in replay speech detection [31]–[33]. CQCC adopts a constant-Q transform (CQT) for the spectral analysis. The CQT employs geometrically spaced frequency bins. Founded upon the basis of CQT, the CQCC has been reported to achieve effective performance for speech synthesis and voice conversion spoofing detection [29]. Traditional features such as MFCC ignores the resolution variability for different frequency regions, it may bring difficulty to the spoofing detection since the resolution in different frequency region can be critical in revealing the information of the spoofing characteristics [31]. In contrast to the Fourier transform which imposes regular spaced frequency bins and hence leads to variable Q factor, the CQT ensures a constant Q factor across the entire spectrum. This trait allows the CQT to provide higher spectral resolution at lower frequencies while providing a higher temporal resolution at higher frequencies, as a result the distribution of the CQT time-frequency resolution is consistent with human hearing characteristics.

The current work is an extension of our previous work [34]. In this paper, based on CQT, we introduce four sub-features and investigate their information complementarity, so that three concatenated features are their concatenated form. In the concatenated features, each sub-feature is designed to be of complementary information to one another. The first sub-feature is the short-term spectral statistics information (STSSI) that is considered to carry the statistic information at frame level, in which the first- and second-order statistics over different CQT-spectral bins are

obtained. The second sub-feature is the octave-band principal information (OPI), which is to provide the octave principle information, where octave segmentation and discrete cosine transform (DCT) are applied. The third sub-feature is the full-band principal information (FPI), it formulates the full-band principle information from the CQT spectrum. And the fourth sub-feature is the magnitude-phase energy information (MPEI). Finally, the four sub-features are combined to generate their delta and acceleration coefficients as features for spoofing detection. We refer the feature as constant-Q statistics-plus-principal information coefficient (CQSPIC) by combining OPI, FPI and STSSI. In the same way, we refer another feature as constant-Q energy-plus-principal information coefficients (CQEPIC) by combining OPI, FPI and MPEI and the third one as constant-Q energy-statistics-principal information coefficients (CESPIC) by combining OPI, FPI, MPEI and STSSI. In this paper, we adopt DNN as the means for the features evaluation.

The remainder of the paper is organized as follows. The CQT is briefly introduced in Section II. In Section III, we describe in detail the proposed features. Section IV gives the experimental results and corresponding analysis, which is based on ASVspoof 2019 corpus. Finally, Section V concludes the paper.

## II. CONSTANT-Q TRANSFORM

CQT is related to the DFT and closely pertinent to the complex Morlet wavelet transform [35], [36]. Different from DFT, the ratio of center frequency to bandwidth,  $Q$ , is constant. For a discrete time domain signal  $x(n)$ , its CQT  $Y(k, l)$  is defined by:

$$Y(k, l) = \sum_{j=l-\lfloor N_k/2 \rfloor}^{l+\lfloor N_k/2 \rfloor} x(j) a_k^*(j - l - N_k/2) \quad (1)$$

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

$$a_k(n) = \frac{1}{C} \omega\left(\frac{n}{N_k}\right) \exp\left[i(2\pi n \frac{f_k}{f_s} + \phi_k)\right] \quad (2)$$

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

$$C = \sum_{m=\lfloor N_k/2 \rfloor}^{\lfloor N_k/2 \rfloor} \omega\left(\frac{m + N_k/2}{N_k}\right) \quad (3)$$

Since a bin spacing corresponding to the equal temperament is desired, the center frequency  $f_k$  obeys

$$f_k = f_1 2^{\frac{k-1}{B}} \quad (4)$$

where  $f_1$  is the centre frequency of the lowest-frequency bin,  $B$  is the number of bins of per octave.

From the above introduction, we can get bandwidth  $\delta f$ :

$$\begin{aligned} \delta f &= f_{k+1} - f_k \\ &= f_1 2^{\frac{k}{B}} - f_1 2^{\frac{k-1}{B}} \\ &= f_1 2^{\frac{k-1}{B}} \left(2^{\frac{1}{B}} - 1\right) \end{aligned} \quad (5)$$

So we can obtain:

$$Q = \frac{f_k}{\delta f} = \frac{1}{2^{\frac{1}{B}} - 1} \quad (6)$$

From formula (6), we can see that  $Q$  and  $B$  has direct proportional relation, the more  $B$ , the more  $Q$ .

Recently, CQCC was reported to be sensitive to the general form of spoofing attack so that it becomes an effective spoofing countermeasure [29]. There are five modules in CQCC extraction. They are CQT, Power Spectrum, Log, Uniform Resampling and DCT. In particular, the role of each module is briefed as follows. CQT is used to convert speech from the time domain to the frequency domain; Power Spectrum is used to calculate octave power spectrum; Log is used to obtain octave power spectrum in logarithm scale; Uniform Resampling is used to convert logarithm octave power spectrum into linear logarithm power spectrum; and DCT is used to extract principal information by attempting to decorrelate the intermediate coefficients in order to deduce the dimension. Figure 1 depicts the extraction of the CQCC feature.

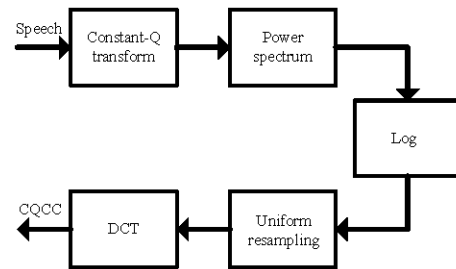


FIGURE 1. Diagram of CQCC extraction.

## III. CONCATENATED FEATURES

CQT that is initially applied in the music processing employs geometrically spaced frequency bins. This ensures a constant  $Q$  factor across the entire spectrum and results in a higher frequency resolution at lower frequencies while providing a higher temporal resolution at higher frequencies. The property of the CQT spectrum distribution coincidentally reflects the human auditory system. This paper investigates the coupling of the CQT with traditional cepstral analysis. In this paper, we aim to seek effective features with different complementary characteristics for spoofing detection on the basis of the advantages of CQT. Consequently, we propose three concatenated features, i.e. **CQSPIC** that includes three characteristics: STSSI, OPI and FPI, **CQEPIC** that includes OPI, FPI and MPEI, and **CESPIC** that includes OPI, FPI, MPEI and STSSI. Next, we will introduce the four sub-features first and then the three concatenated features.

### A. STSSI

In spoofing detection, we face a situation where there is insufficient prior knowledge about the characteristics to distinguish a spoofed speech from genuine speech. It is known that the two kinds of speech signals have two different statistical characteristics. It has been reported that first order statistics is useful to the countermeasures for presentation attacks as natural speech and synthetic speech differ in terms of both spectral statistics and time statistics [25]. On the other hand, the amplitude of the frequency components changes over time due to the non-stationary property of speech signal. Furthermore, natural speech and synthetic speech could be different in terms of such dynamics. Variance of speech spectral amplitude over an utterance for each spectrum is proven to be beneficial to spoofing detection [25]. Indeed, one of the successful approaches to classify natural and synthetic speech signals is the use of dynamic temporal derivative information of short-term spectrum as opposed to static information [37]. Information about the channel characteristics can be modeled through spectral statistics. Mean of cepstral coefficients and the second order spectral statistics have been shown to make the speaker verification system more robust to channel variability, while channel information is desirable for the detection of spoofing attacks [24]. It has been well known that channel information is important for speech synthesis, voice conversion and replay attacks.

In [24], [38], long-term spectral statistics (LTSS) is reported to be effective for spoofing detection in speaker verification system. It is believed that the mean and variance of the spectral amplitude distributed over either a long-term period of certain spectrum or a range of frequencies at a time frame can provide good traits to distinguish the two different kinds of speech signals. However, LTSS is not suitable for small training database due to insufficient feature data generated. In this subsection, on the basis of CQT, we introduce a short-term statistics at frame level for the purpose of solving the small training data issue and build complementary characteristics [34].

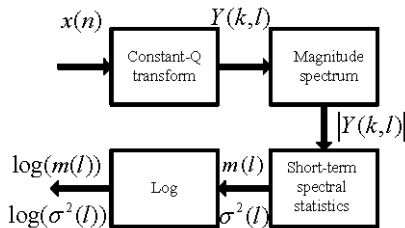


FIGURE 2. Diagram of short-term statistics information extraction.

As mentioned above, there are two short-term statistics, one is first-order statistics (mean) and the other is second-order statistics (variance). There are four modules in STSSI extraction: CQT, magnitude spectrum, short-term statistics and log. The module of CQT is also used to convert speech from the time domain to the frequency domain, magnitude spectrum module is used to calculate magnitude spectrum,

short-term statistics module is to estimate STSSI from magnitude spectrum, and log module is used to obtain mean and variance in logarithmic scale. Figure 2 shows a block diagram of short-term statistics information extraction. To estimate STSSI cross frequency bins at frame-level, one is to estimate the statistics over full frequency-band, the other is to compute the statistics over each individual subband such as the octave-band. To generalize the statistics formula, we give the subband statistics as follows. Supposing  $|Y(k, l)|$  is a frame magnitude spectrum of  $Y(k, l)$ , The mean of the CQT spectral amplitude over subband,  $m_s$ , is defined by

$$m_s(l) = \frac{1}{K_s - K_{s-1}} \sum_{k=K_{s-1}+1}^{K_s} |Y(k, l)| \quad s = 1, \dots, S \quad (7)$$

And the variance of the CQT spectrum amplitude over sub-band is defined by

$$\sigma_s^2(l) = \frac{1}{K_s - K_{s-1}} \sum_{k=K_{s-1}+1}^{K_s} (|Y(k, l)| - m_s(l))^2 \quad (8)$$

where  $\sigma_s^2(l)$  represents variance of  $|Y(k, l)|$ ,  $S$  denotes the number of subbands,  $K_0, \dots, K_S$  is the frequency index of subbands where  $K_0 = 0$  and  $K_S = K$ . Thus, the full-band STSSI becomes the special case of the subband STSSI when  $S = 1$ . In addition,  $l$  represents the time frame index.

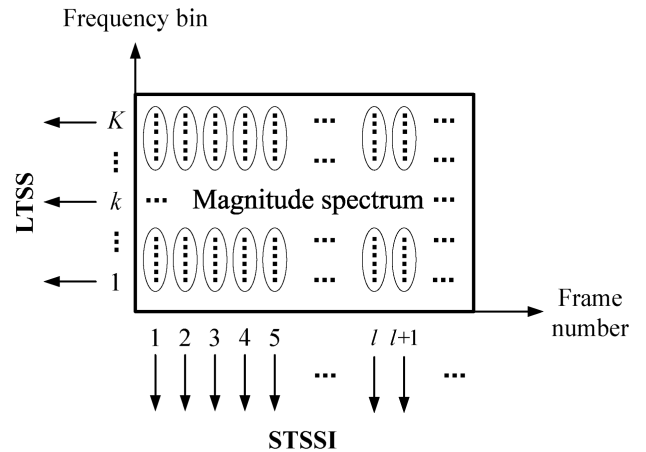


FIGURE 3. Difference between how to estimate LTSS and STSSI from magnitude spectrum.

Figure 3 shows the difference between LTSS and STSSI. From Figure. 3, we can see the difference between LTSS and STSSI. LTSS is obtained cross different frames while STSSI is obtained cross frequency bins. So we can say that LTSS has frequency statistical characteristics while STSSI has temporary-spectral statistical characteristics. Our experiment shows the octave-band statistics of STSSI is not competent with full-band STSSI statistics for spoofing detection. It may be because that there are insufficient frequency bins to approximate the statistics in an octave-band. Subsequently,



in this paper, we only focus on reporting the performance with full-band STSSI statistics.

### B. OPI

An octave is the doubling or halving of a certain frequency. The speech frequency range can be separated into unequal segments called octaves. A band is defined to be an octave in width when the upper band frequency is twice the lower band frequency. In contrast to DFT where frequency region of each frequency bin is equal, the frequency region of different frequency bin in CQT is different. The centre frequency bin of CQT complies with a nonlinear distribution with (4), we have

$$f_{nB+k} = 2^{\frac{nB+k-1}{B}} f_1 = 2^n f_k = 2f_{(n-1)B+k} \quad n = 1, \dots, N \quad (9)$$

where  $N$  denotes the number of octave-bands. So we have  $K = N * B$ . From (9) we can see that  $f_{B+1} = 2f_1$ ,  $f_{2B+1} = 2f_{B+1}, \dots, f_{NB+1} = 2f_{(N-1)B+1}$ . Therefore,  $B$  frequency bins (i.e.  $f_1, f_2, \dots, f_B$ ) between  $f_1$  and  $f_{B+1}$  form the first octave band;  $B$  frequency bins between  $f_{B+1}$  and  $f_{2B+1}$  (i.e.  $f_{B+1}, f_{B+2}, \dots, f_{2B}$ ) form the second octave band;  $\dots$ , and  $B$  frequency bins between (i.e.  $f_{(N-1)B+1}, f_{(N-1)B+2}, \dots, f_{NB}$ ) between  $f_{(N-1)B+1}$  and  $f_{NB+1}$  form the  $N$ -th octave subband. As a result, there are  $B$  frequency bins in each of octave-band with CQT. The higher an octave-band is, the larger frequency region the corresponding frequency bin occupies.

In this subsection, on the basis of CQT, we introduce OPI [34]. In OPI, octave segmentation [39] is applied, and it is followed by a DCT to generate principal information. In particular, OPI includes five modules: CQT, Power spectrum, Log and DCT. The  $p$ -th principal coefficients of the  $n$ -th octave-band is given using discrete cosine transform as follows:

$$X_{np}(l) = \sum_{k=(n-1)B+1}^{nB} \log(|Y(k, l)|^2) \cos \left[ \frac{\pi}{B} \left( k + \frac{1}{2} \right) p \right] \quad p = 1, 2, \dots, P \quad (10)$$

$P$  denotes the number of principal coefficients corresponding to an octave-band, and  $P \leq B$ . Finally, the  $X_{1\{1:P\}}, X_{2\{1:P\}}, \dots, X_{n\{1:P\}}, \dots, X_{N\{1:P\}}$  are concatenated to form a  $N*B$  dimension of OPI vector at the  $l$ -th frame. Figure 4 depicts the procedure of the OPI. In our experiment, we set  $B$  to be 96,  $P$  to be 12, and  $N$  to be 9.

### C. FPI

In this subsection, we introduce FPI as complementary characteristics of the OPI [34]. In CQCC, the frequency bin of CQT domain is resampled by converting geometric space to linear space. Different from the CQCC with linearized log power spectrum resampling, the FPI directly applies DCT on logarithm power spectrum in CQT domain so that the human auditory property can be preserved and thus considered into the feature. In CQT, the property of the higher frequency resolution in low frequency and higher temporal resolution in higher frequency causes the CQT transform domain has some similar trait with human auditory system. MFCC is a good example to adopt the human auditory property by

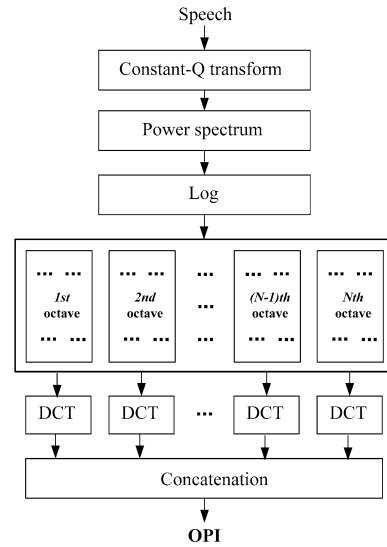


FIGURE 4. Procedure of the OPI extraction.

reorganizing the linear frequency bin into the Mel-frequency domain. In this paper, we consider to directly use the CQT geometrical distribution by applying DCT transformation in the CQT geometrically frequency domain. So that we introduce the FPI. For the FPI feature extraction, there are four modules including CQT, Power Spectrum, Log and DCT. In the FPI, the  $r$ -th principal coefficients are given via DCT as follows:

$$Z_r(l) = \sum_{k=1}^K \log(|Y(k, l)|^2) \cos \left[ \frac{\pi}{K} \left( k + \frac{1}{2} \right) r \right] \quad (11)$$

$r = 1, 2, \dots, R$  where  $R$  is the number of principal coefficients.

Figure 5 shows the block diagrams of the FPI procedure.

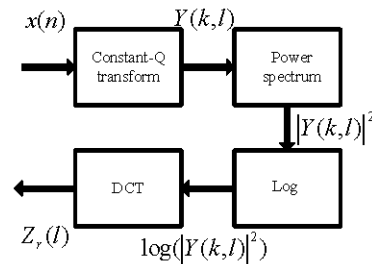


FIGURE 5. Block diagram of FPI extraction.

### D. MPEI

In our previous study [40], the idea of magnitude and phase energy was proposed. In this study, we introduce how to extract MPEI from magnitude and phase spectrum in detail and give its calculation equation step by step. Figure 6 shows how to extract MPEI from magnitude and phase spectrum. From Figure 6, it can be seen that there are four modules in MPEI extraction, which are CQT, magnitude and phase spectrum, magnitude-phase energy and log Magnitude and

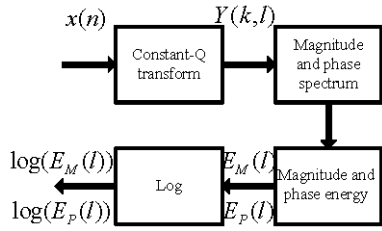


FIGURE 6. Framework of MPEI extraction.

phase spectrum can be obtained at the base of CQT, then magnitude and phase energy can be calculated, finally, MPEI can be obtained by log operation.

Next, we take  $Y(k, l)$  mentioned above as an example to show how to calculate MPEI, in which  $k$  and  $l$  represent frequency bin and time frame index, respectively.

Further,  $Y(k, l)$  can be written as band is given using discrete cosine transform as follows:

$$Y(k, l) = |Y(k, l)| e^{j\theta(k, l)} \quad (12)$$

Suppose  $E_M(l)$  stands for magnitude energy at frame  $l$ , we can obtain

$$E_M(l) = \frac{\sum_{k=1}^K |Y(k, l)|^2}{K} \quad (13)$$

Suppose  $E_P(l)$  stands for magnitude energy at frame  $l$ , we can obtain

$$E_P(l) = \frac{\sum_{k=1}^K |\theta'(k, l)|^2}{K} \quad (14)$$

where

$$\theta'(k, l) = \text{princ}(\theta(k, l)) \quad (15)$$

where  $\text{princ}(\cdot)$  represents the principal value operator, mapping the phase  $\theta(k, l)$  onto  $-\pi \leq \theta(k, l) \leq \pi$ .

On the basis of  $E_M(l)$  and  $E_P(l)$ , supposing  $MPEI(l)$  is MPEI at the frame  $l$ , we can obtain

$$MPEI(l) = [\log(E_M(l)) \log(E_P(l))] \quad (16)$$

### E. CONCATENATION

The proposed CQSPIC is formed by combining the three sub-features: STSSI, OPI and FPI. OPI and FPI are complementary because they represent octave spectral information and full-band spectral information respectively. STSSI represents statistics information, it is of substantial complementarity with both OPI and FPI. The purpose to combining the sub-features is to strengthen the genuine spoofing discriminative ability for the spoofing detection.

Firstly, the static coefficient of the STSSI, OPI and FPI are concatenated; then delta and double-delta of the concatenated coefficients are applied to produce the final CQSPIC. Figure 7 illustrates the block diagram of CQSPIC extraction.

In the same way, we can obtain another two concatenated features, i.e. CQEPIC and CESPIC, which are shown in Figure 8 and Figure 9, respectively.

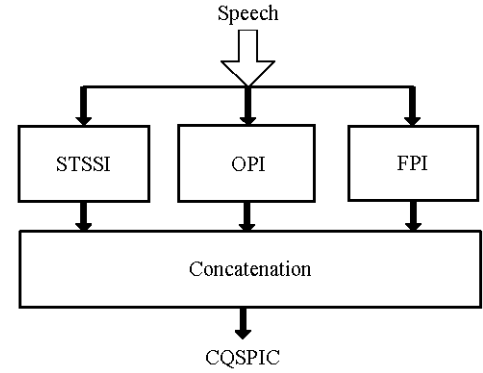


FIGURE 7. Block diagram of the extraction of the proposed CQSPIC.

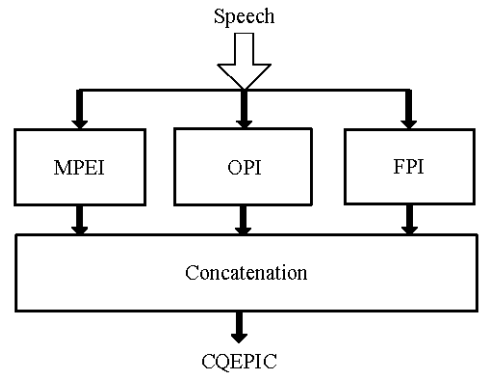


FIGURE 8. Block diagram of the extraction of the proposed CQEPIC.

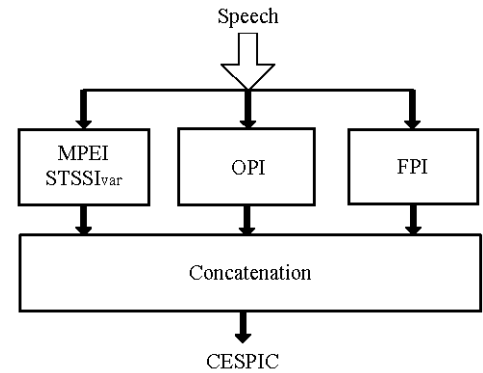


FIGURE 9. Block diagram of the extraction of the proposed CESPIC.

From Figure 8, it can be seen that CQEPIC is formed by combining the three sub-features: MPEI, OPI and FPI. We believe that MPEI is of substantial complementarity with both OPI and FPI. The aim to combine MPEI, OPI and FPI is to enlarge the genuine-spoofing discriminative ability for the spoofing detection.

Different from CQSPIC and CQEPIC, it can be observed that CESPIC is formed by combining the four sub-features: MPEI, STSSIvar, OPI and FPI from Figure 9. Note that only STSSIvar is used here. The reason is that STSSImean has the same information with magnitude energy information

in MPEI. In which, MPEI and STSSIVar can be of substantial complementarity with both OPI and FPI. The objective to combining MPEI, STSSIVar, OPI and FPI is also to enhance the genuine-spoofing discriminative ability for the spoofing detection.

#### F. CLASSIFIER IN SPOOFING DETECTION

The ultimate purpose of the spoofing detection is to distinguish between spoofed speech and genuine speech. Given a speech signal,  $x = [x(0), \dots, x(n), \dots, x(N)]$ , the task of spoofing detection is to determine whether  $x$  is a human natural speech or synthetic/converted/replayed speech. Utterance based spoofing detection actually becomes a classification problem, where two hypothesis  $H_0$  and  $H_1$  is defined below

$H_0$ :  $x$  is a human natural speech;

$H_1$ :  $x$  is a synthetic/converted/replayed speech.

The log-likelihood ratio (LLR) generated by the difference of log-likelihoods of input speech given trained genuine ( $H_0$ ) and spoofing ( $H_1$ ) are used as score of spoofing detection.

$$LLR = \log\{p(x|H_0)\} - \log\{p(x|H_1)\} \quad (17)$$

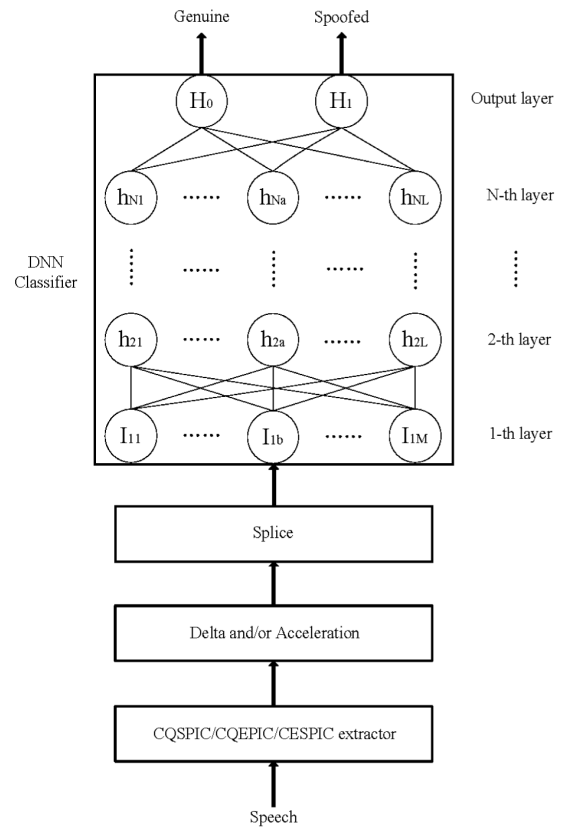
This is done by selecting effective feature in frontend processing and the reliable classifier in backend processing. To verify the effectiveness of the anti-spoofing features, we have to investigate the feature performance on a reliable and effective backend classifier. Presently GMM, SVM, DNN and RNN are popularly used for the spoofing detection.

DNN has been reported to outperform GMM in many classification applications [41], [42]. In this paper, we use DNN to evaluate the performance of the features on which we study for spoofing detection. Figure 10 shows the block diagram of our platform based on the proposed features and DNN classifier to discriminate between genuine and spoofed speech.

We train the DNN using labelled speech and use the trained DNN to discriminate an input utterance between natural and spoofed speech. Figure 10 briefs the CQSPIC spoofing detection system. In particular, the input utterance goes through CQSPIC feature extraction, delta or acceleration usage followed by the splice application where eleven spliced frames centred by the current concatenated to form the input feature, and then DNN classifier is applied, so the posterior probability of the input frame is computed. The final detection result is obtained by averaging the posterior probabilities over all of the frames in an utterance. During training, the output is compared with the target value, and back-propagation algorithm is applied for DNN parameter update. For detection, according to the two output values, the final result (natural or spoofed) can be obtained. The performance of DNN depends on particular feature.

#### IV. PERFORMANCE EVALUATION

In this paper, the experiments for evaluating the anti-spoofing performance of the proposed features are carried out on ASVspoof 2019 logical access and physical access corpus [43]. Tandem detection cost function (t-DCF) [44] and



**FIGURE 10. Our spoofing detection system based on the proposed features and DNN classification.**

equal error rate (EER) are used as the primary and secondary metric to measure the performance.

In CQT computation, all configuration parameters are set to be the same as those in [29]. In the evaluation, we trained DNN models with stochastic gradient descent (SGD) as spoofing detection platform using computational network toolkit (CNTK) [45]. In particular, different DNN models are trained corresponding to different features such as CQSPIC, CQEPIC and CESPIC. In this evaluation, the input layer of the DNN is the feature coefficients of eleven spliced frames centred by the current frame. The feature coefficients of each frame can be the static feature coefficient, or its delta, or acceleration (i.e. double delta), or their combining feature. In this paper, we use S, D and A to represent static, delta and acceleration coefficients respectively. In our experiments, for OPI, we set  $P = 12$ ;  $N = 9$ , as a result, there are 108 dimensions of static OPI. For FPI,  $R$  is set to 30, it means the FPI has 30 dimensions of its principal vector. In addition, the dimension of static STSSI and MPEI is set to 2, respectively. It means that the static dimension of CQSPIC, CQEPIC and CESPIC is 140, 140 and 141, respectively.

#### A. SYNTHETIC SPEECH DETECTION

ASVspoof 2019 logical access corpus contains speech synthesis and voice conversion attacks produced through

logical access. It is constituted by three subsets: training, development and evaluation sets, Table 1 gives the details for the three subsets.

**TABLE 1.** The details of the three subsets in ASVspoof 2019 logical Access corpus.

Subset	#Speakers		# Utterances	
	Male	Female	Genuine	Spoofed
Training	8	12	2,580	22,800
Development	8	12	2,548	22,296
Evaluation	30	37	7,355	63,882

## 1) FEATURE EVALUATION UNDER CONSISTENT PLATFORM

We use 25,380 utterances from the training set of ASVspoof 2019 logical access database to train DNN models, which have two hidden layers with 1024 nodes per layer and one output layer with 2 nodes indicating genuine and spoofed speech, which has shown they can give best performance on the development set comparison with other networks with different layer number and different nodes.

Table 2 gives the experimental results on ASVspoof 2019 logical access development set using different coefficient configurations of CQSPIC, CQEPIC and CESPIC in terms of t-DCF and EER. From Table 2, it can be seen that: (1) In terms of t-DCF, CQSPIC-SD and CQSPIC-SDA can give the best performance on the development set for CQSPIC. (2) For CQEPIC, coefficient configurations A, SD and SDA can give the best performance on the development set. (3) CESPIC-SD gives the best performance on the development set in terms of t-DCF for CESPIC.

**TABLE 2.** Experimental results on ASVspoof 2019 logical Access development set using different coefficient configurations of CQSPIC, CQEPIC and CESPIC in terms of T-DCF and EER (%).

Feature	t-DCF/EER						
	Coefficient configurations						
	S	D	A	SD	SA	DA	SDA
CQSPIC	0.004	0.091	0.076	<b>0.000</b>	0.003	0.038	<b>0.000</b>
	0.16	4.30	2.63	0.04	0.05	1.30	0.04
CQEPIC	0.003	0.046	<b>0.001</b>	<b>0.001</b>	0.081	0.042	<b>0.001</b>
	0.19	1.65	0.08	0.08	2.87	1.37	0.04
CESPIC	0.003	0.050	0.072	<b>0.000</b>	0.001	0.041	0.001
	0.12	1.61	2.40	0.04	0.08	1.41	0.08

However, the coefficient configuration that can obtain the best results on the development set can't obtain the best results on the evaluation set. We adopt ASVspoof 2019 logical access evaluation set for a further comparison. Table 3 shows the experimental result (t-DCF and EER) on ASVspoof 2019 logical access evaluation set using different coefficient configuration of CQSPIC, CQEPIC and CESPIC. It can be seen that the coefficient configuration DA can give the best performance on the evaluation set for the three proposed features in terms of t-DCF and EER.

## 2) PERFORMANCE COMPARISON AMONG DIFFERENT TYPES OF DISCRIMINATIVE INFORMATION

Table 4 shows the experimental results comparison among the proposed features and different types of discriminative

**TABLE 3.** Experimental results on ASVspoof 2019 logical Access evaluation set using different coefficient configurations of CQSPIC, CQEPIC and CESPIC in terms of T-DCF and EER (%).

Feature	t-DCF/EER						
	Coefficient configurations						
	S	D	A	SD	SA	DA	SDA
CQSPIC	0.287	0.184	0.272	0.265	0.307	<b>0.182</b>	0.275
	10.16	8.01	12.57	9.82	11.00	<b>7.64</b>	10.03
CQEPIC	0.288	0.191	0.275	0.271	0.301	<b>0.183</b>	0.262
	10.32	8.13	12.45	9.72	11.01	<b>7.83</b>	9.59
CESPIC	0.183	0.188	0.268	0.266	0.300	<b>0.178</b>	0.282
	7.81	8.24	12.17	9.71	10.90	<b>7.63</b>	10.31

**TABLE 4.** Experimental results on ASVspoof 2019 logical Access evaluation set among the proposed features and different types of discriminative information.

Feature	Information Type	t-DCF	EER
FPI-DA	FPI	0.184	7.91
OPI-DA	OPI	0.188	8.04
(FPI+OPI)-DA	FPI, OPI	0.184	7.82
CQSPIC-DA	FPI, OPI, STSSI	0.182	7.64
CQEPIC-DA	FPI, OPI, MPEI	0.183	7.83
CESPIC-DA	FPI, OPI, STSSI, MPEIvar	0.178	7.63

information on ASVspoof 2019 logical access evaluation set. In which, FPI-DA, OPI-DA and (FPI+OPI)-DA have their respective DNN classifiers and the training method is the same as mentioned above. From Table 4, several conclusions can be obtained: (1) In terms of t-DCF, the performance of FPI+OPI is better than OPI in terms of t-DCF, though FPI and FPI+OPI have the equal t-DCF, FPI+OPI give a little better performance than FPI in term of EER. From the performance comparison between FPI+OPI and OPI, and FPI+OPI and PFI, we can know that OPI and FPI can be complementary with each other. (2) The performance of CQSPIC, CQEPIC and CESPIC is better than FPI+OPI in terms of t-DCF, which means that STSSI, MPEI and MPEI+STSSIvar is helpful for OPI+FPI to detect spoofing. It also confirms that our proposed idea is correct. (3) CESPIC performs a little better than CQSPIC and CQEPIC, the reason is that MPEIvar is useful to capture the artifacts in voice converted speech and synthetic speech. It can be concluded that the four sub-features have very reasonable complementarity each other from the experimental results

**TABLE 5.** Experimental results comparison between DNN and GMM for the proposed features on ASVspoof 2019 logical Access evaluation set.

Feature	Model	t-DCF	EER
CQSPIC-DA	DNN	0.182	7.64
	GMM	0.265	13.20
CQEPIC-DA	DNN	0.183	7.83
	GMM	0.267	13.20
CESPIC-DA	DNN	0.178	7.63
	GMM	0.267	13.43

## 3) COMPARISON WITH GAUSSIAN MIXTURE MODEL

Table 5 shows the experimental results comparison between DNN and GMM for the proposed features on ASVspoof 2019 logical access evaluation set. From Table 5, it can be found



that DNN can give much better performance than GMM on ASVspoof 2019 logical access evaluation set for CQSPIC, CQEPIC and CESPIC. Additionally, it is interesting to find that CQSPIC-DA can perform slightly better than CQEPIC and CESPIC under the model of GMM unlike that CQESPIC can give the best performance under the model of DNN.

#### 4) COMPARISON WITH SOME COMMONLY USED HANDCRAFTED FEATURES

Table 6 gives the experimental results comparison among the proposed features and some commonly used handcrafted features on ASVspoof 2019 logical access evaluation set. In which, every feature has its own DNN classifier. From the above experimental results, we can see that the proposed CQSPIC, CQEPIC and CESPIC greatly outperform the conventional features. The reason may be that the proposed features have several types of discriminative information for spoofing detection.

**TABLE 6.** Experimental results comparison among the proposed features and some commonly used features on ASVspoof 2019 logical Access evaluation set.

Feature	t-DCF	EER
MFCC-DA	0.217	8.50
LFCC-DA	0.310	12.79
IFCC-DA	0.288	11.65
CQCC-DA	0.337	16.64
eCQCC-DA	0.185	7.74
<b>CQSPIC-DA</b>	<b>0.182</b>	<b>7.64</b>
<b>CQEPIC-DA</b>	<b>0.183</b>	<b>7.83</b>
<b>CESPIC-DA</b>	<b>0.178</b>	<b>7.63</b>

#### 5) COMPARISON WITH SOME KNOWN SYSTEMS

Table 7 gives the experimental results comparison among the proposed systems and some known single systems on ASVspoof 2019 logical access evaluation set. In which, MFCC-ResNet, Spec-ResNet and CQCC-ResNet represent ResNet-based end-to-end systems with MFCC, log power spectrum based on DFT and CQCC as the inputs [46], respectively. Dfea\_LPS [47] represents deep feature that is extracted from a trained deep feature extractor with log power spectrum based on DFT as input.

From Table 7, it can be found that the proposed systems perform better than the selected systems. It means that our systems based on proposed features have more discriminative information to capture the artifacts in the evaluation set of ASVspoof 2019 logical access. This also can confirm that the proposed concatenated features are correct, which can work in the field of synthetic speech detection.

### B. REPLAY SPEECH DETECTION

The trait of the replay speech is of the fluctuations of microphone signal caused by transients or changes due to multiple analogue-to-digital conversion and digital-to-analogue conversion. The multiple conversions between analogue and digital signals lead to an apparent artifacts of this kind of

**TABLE 7.** Experimental results comparison among the proposed systems and some known systems on ASVspoof 2019 logical Access evaluation set.

System	t-DCF	EER
CQCC-GMM [43]	0.237	9.57
LFCC-GMM [43]	0.212	8.09
MFCC-ResNet [46]	0.204	9.33
Spec-ResNet [46]	0.274	7.69
CQCC-ResNet [46]	0.217	7.69
(Dfea_LPS)-DNN [47]	0.181	7.57
<b>(CQSPIC-DA)-DNN</b>	<b>0.182</b>	<b>7.64</b>
<b>(CQEPIC-DA)-DNN</b>	<b>0.183</b>	<b>7.83</b>
<b>(CESPIC-DA)-DNN</b>	<b>0.178</b>	<b>7.63</b>

spoofed speech. On the other hand, the multiple channel distortions due to the replaying process causes the frequency distribution different from the genuine speech signal.

In this subsection, CQSPIC, CQEPIC and CESPIC are evaluated on ASVspoof 2019 physical access [40]. Table 8 summarizes ASVspoof 2019 physical access. In which, the corpus also has three subsets: train, development and evaluation sets. According to the training rules aforementioned, a series of classifiers based on DNN are trained, then the trained classifiers are used to evaluate the proposed features on the development set. Unlike the classifiers trained for synthetic and voice converted speech, we found that 4 hidden layers with the node number 1024 can give the best performance, the reason may be that there are 54,000 utterances in ASVspoof 2019 physical access training set while there are only 25,380 utterances in ASVspoof 2019 logical access training set.

**TABLE 8.** The details of the three subsets in ASVspoof 2019 physical Access corpus.

Subset	#Speakers		# Utterances	
	Male	Female	Genuine	Spoofed
Training	8	12	5,400	48,600
Development	8	12	5,400	24,300
Evaluation	30	37	18,089	134,630

#### 1) FEATURES PERFORMANCE EVALUATION WITH ASVspoof 2019 PHYSICAL ACCESS

Table 9 gives the experimental results (t-DCF and EER) on ASVspoof 2019 physical access development set using different feature combinations of CQSPIC, CQEPIC and CESPIC. From Table 9, it can be seen that CQSPIC-SD, CQEPIC-S and CESPIC-S can give the best performance in their own coefficient configurations in terms of t-DCF or EER.

Table 10 shows the experimental results on ASVspoof 2019 physical access evaluation set using different coefficient configuration of CQSPIC, CQEPIC and CESPIC in terms of t-DCF and EER. From the experimental results, we can see that the proposed CQSPIC-SD, CQEPIC-SD and CESPIC-SD greatly outperforms other coefficient configurations of CQSPIC, CQEPIC and CESPIC, respectively. Comparison with Table 9, it can be found that CQEPIC

**TABLE 9.** Experimental results on ASVspoof 2019 physical Access development set using different coefficient configurations of CQSPIC, CQEPIC and CESPIC in terms of t-DCF and EER (%).

Feature	t-DCF/EER						
	Coefficient configurations						
	S	D	A	SD	SA	DA	SDA
CQSPIC	0.054	0.223	0.189	<b>0.051</b>	0.053	0.214	0.055
	2.22	12.70	9.50	2.67	2.72	11.70	2.82
CQEPIC	<b>0.048</b>	0.225	0.195	0.053	0.054	0.216	0.055
	2.41	13.00	10.17	2.74	2.83	12.15	2.85
CESPIC	<b>0.049</b>	0.222	0.188	0.053	0.057	0.215	0.058
	2.29	13.00	9.50	2.73	2.87	12.19	2.39

**TABLE 10.** Experimental results on ASVspoof 2019 physical Access evaluation set using different coefficient configurations of CQSPIC, CQEPIC and CESPIC in terms of t-DCF and EER (%).

Feature	t-DCF/EER						
	Coefficient configurations						
	S	D	A	SD	SA	DA	SDA
CQSPIC	0.206	0.212	0.200	<b>0.168</b>	0.182	0.202	0.162
	8.18	12.44	10.20	<b>7.10</b>	7.69	11.58	7.29
CQEPIC	0.174	0.213	0.202	<b>0.166</b>	0.187	0.205	0.174
	7.68	12.61	10.44	<b>7.44</b>	7.93	11.84	7.53
CESPIC	0.182	0.215	0.198	<b>0.173</b>	0.185	0.204	0.219
	7.81	12.61	9.99	<b>7.35</b>	7.83	11.80	8.31

and CESPIC have different coefficient configurations on the development set and evaluation set.

**TABLE 11.** Experimental results on ASVspoof 2019 physical Access evaluation set among the proposed features and different types of discriminative information.

Feature	Information Type	t-DCF	EER
FPI-SD	FPI	0.252	9.23
OPI-SD	OPI	0.263	9.67
(FPI+OPI)-SD	FPI, OPI	0.188	8.14
CQSPIC-SD	FPI, OPI, STSSI	0.168	7.10
CQEPIC-SD	FPI, OPI, MPEI	0.166	7.40
CESPIC-SD	FPI, OPI, STSSI, MPEIvar	0.173	7.35

## 2) PERFORMANCE COMPARISON AMONG DIFFERENT TYPES OF DISCRIMINATIVE INFORMATION

Table 11 gives the experimental results comparison among the proposed features and different types of discriminative information on ASVspoof 2019 physical access evaluation set. From XI, it can be observed that FPI performs better than OPI; the combination of the two, OPI+FPI, outperforms over individual OPI and FPI separately, this shows the positive complementarity of the two sub-features. It also can be seen that OPI+FPI+STSSI gives better performance than OPI+FPI, and OPI+FPI+MPEI is better than OPI+FPI, and OPI+FPI+MPEI is also better than OPI+FPI, it proves the statement that STSSI and MPEI is useful to discriminate replay speech and genuine speech. However, on the basis of OPI+FPI, the joint MPEI and STSSIvar is worse than MPEI and STSSI on replay spoofing detection. The reason may be that SISSIVar has conflict with MPEI for replay speech detection.

## 3) COMPARISON WITH GAUSSIAN MIXTURE MODEL

Table 12 shows the experimental results comparison between DNN and GMM for the proposed features and on ASVspoof 2019 physical access evaluation set. From Table 12, it can be found that GMM performs better than DNN on ASVspoof 2019 physical access evaluation set for CQSPIC, CQEPIC and CESPIC in terms of t-DCF and EER. In addition, the same as CQEPIC performs better than CQSPIC and CESPIC under the model of DNN, CQEPIC also gives the best performance among the three proposed features under the model of GMM.

**TABLE 12.** Experimental results comparison between DNN and GMM for the proposed features on ASVspoof 2019 physical Access evaluation set.

Feature	Model	t-DCF	EER
CQSPIC-SD	DNN	0.168	7.10
	GMM	0.141	7.07
CQEPIC-SD	DNN	0.166	7.40
	GMM	0.137	6.97
CESPIC-SD	DNN	0.173	7.35
	GMM	0.139	7.10

## 4) COMPARISON WITH SOME COMMONLY USED HANDCRAFTED FEATURES

Table 13 gives the experimental results comparison among the proposed features and some commonly used handcrafted features on ASVspoof 2019 physical access evaluation set. In which, every feature has its own DNN classifier.

**TABLE 13.** Experimental results comparison among the proposed features and some commonly used features on ASVspoof 2019 physical Access evaluation set.

Feature	t-DCF	EER
MFCC-SD	0.452	18.60
LFCC-SD	0.429	18.00
IFCC-SD	0.382	16.80
CQCC-SD	0.426	17.17
eCQCC-SD	0.196	8.30
<b>CQSPIC-SD</b>	<b>0.168</b>	<b>7.10</b>
<b>CQEPIC-SD</b>	<b>0.166</b>	<b>7.40</b>
<b>CESPIC-SD</b>	<b>0.173</b>	<b>7.35</b>

From the experimental results in Table 13, we can see that the proposed CQSPIC, CQEPIC and CESPIC greatly outperform the conventional handcrafted features MFCC and CQCC. Additionally, they also give better performance than eCQCC. The reason may be that the proposed features have several types of discriminative information for spoofing detection.

## 5) COMPARISON WITH SOME KNOWN SYSTEMS

Table 14 shows the experimental results comparison among the proposed systems and some known single systems on ASVspoof 2019 physical access evaluation set. In which, ZTWCC represents zero time windowing cepstral coefficients [48], LFCC-LCNN and DCT-LCNN represent light

**TABLE 14.** Experimental results comparison among the proposed systems and some known systems on ASVspoof 2019 physical Access evaluation set.

System	t-DCF	EER
CQCC-GMM [43]	0.245	11.04
LFCC-GMM [43]	0.302	13.54
ZTWCC-GMM [48]	0.281	12.20
LFCC-LCNN[49]	0.105	4.60
DCT-LCNN [49]	0.560	2.06
Spec-ResNet [46]	0.099	3.81
CQCC-ResNet [46]	0.107	4.43
(CQSPIC-SD)-DNN	<b>0.168</b>	<b>7.40</b>
(CQEPIC-SD)-DNN	<b>0.166</b>	<b>7.39</b>
(CESPIC-SD)-DNN	<b>0.173</b>	<b>7.35</b>
(CQSPIC-SD)-GMM	<b>0.141</b>	<b>7.07</b>
(CQEPIC-SD)-GMM	<b>0.137</b>	<b>6.97</b>
(CESPIC-SD)-GMM	<b>0.139</b>	<b>7.10</b>

CNN-based end-to-end systems with LFCC and DCT spectrum as the inputs [49], respectively; in the same way, spec-ResNet and CQCC-ResNet represent ResNet-based end-to-end systems with the inputs of log power spectrum based on DFT and CQCC [46], respectively.

From Table 14, it can be found that our systems outperform the systems based on handcrafted features such as CQCC, LFCC and ZTWCC. However, our system are a little worse than LFCC-LCNN, Spe-ResNet and CQCC-ResNet. It means that our systems have more discriminative information in replay speech detection. In addition, comparison Table 7 and Table 12, it can be found that our systems perform a little worse than the end-to-end systems such as Spe-ResNet and CQCC-ResNet in replay speech detection while our systems can give better performance in synthetic speech detection.

## V. CONCLUSION

On the basis of the advantages of CQT, we have proposed three concatenated features, CQSPIC, CQEPIC and CESPIC, by extracting information from short-term spectral statistics, magnitude and phase energy, octave-band and full-band for spoofing detection in speaker verification system. The complementarity of the sub-features has been investigated for the different types of spoofing attacks: synthetic speech and replay speech. It is conclusive that the combination of OPI, FPI and STSSI (MPEI, MPEI plus STSSIvar) is effective and useful for spoofing detection. The experimental results show that the proposed concatenated features outperform some commonly used handcrafted features in spoofing detection meanwhile the proposed systems can give comparable performance with some known systems.

## REFERENCES

- [1] S. Furui, "Fifty years of progress in speech and speaker recognition," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2497–2498, Oct. 2004.
- [2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.
- [3] J. Yamagishi, T. Kinnunen, N. Evans, P. D. Leon, and I. Trancoso, "Introduction to the issues on spoofing and countermeasures for automatic speaker verification," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 585–587, Jun. 2017.
- [4] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 768–783, Apr. 2016.
- [5] D. Paul, M. Pal, and G. Saha, "Spectral features for synthetic speech detection," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 605–617, Jun. 2017.
- [6] C. Zhang, S. Ranjan, M. K. Nandwana, Q. Zhang, A. Misra, G. Liu, F. Kelly, and J. H. L. Hansen, "Joint information from nonlinear and linear features for spoofing detection: An i-vector/DNN based approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5035–5038.
- [7] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 588–604, Jun. 2017.
- [8] X. Tian, S. Lee, Z. Wu, E. S. Chng, and H. Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1863–1875, Oct. 2017.
- [9] Q. H. ZhifengWang, Q. He, X. Zhang, H. Luo, and Z. Su, "Playback attack detection based on channel pattern noise," *J. South China Univ. Technol., Natural Sci. Ed.*, vol. 39, no. 10, pp. 7–12, 2011.
- [10] P. Nagarshenth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 97–101.
- [11] W. Shang and M. Stevenson, "A preliminary study of factors affecting the performance of a playback attack detector," in *Proc. Can. Conf. Electr. Comput. Eng.*, May 2008, pp. 459–464.
- [12] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Commun.*, vol. 72, pp. 13–31, Sep. 2015.
- [13] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.
- [14] C. Hanilci, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection Goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Commun.*, vol. 85, pp. 83–97, Dec. 2016.
- [15] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison features for synthetic speech detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 2087–2091.
- [16] Z. Wu and H. Li, "Voice conversion versus speaker verification: An overview," *APSIPA Trans. Signal Inf. Process.*, vol. 3, pp. 1–16, Dec. 2014.
- [17] Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 87–91.
- [18] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification reply spoofing attack: On data augmentation, feature representation, classification and fusion," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 17–21.
- [19] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamaki, and K. A. Lee, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5395–5399.
- [20] T. Kinnunen, H. D. M. Sahidullah, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 2–6.
- [21] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 39, Jul. 2011, pp. 5–12.
- [22] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 22–26.
- [23] T. B. Patel and H. A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detection," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 618–631, Jun. 2017.



- [24] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–6.
- [25] H. Muckenhirn, P. Korshunov, M. Magimai-Doss, and S. Marcel, "Long-term spectral statistics for voice presentation attack detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2098–2111, Nov. 2017.
- [26] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 1700–1703.
- [27] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5475–5479.
- [28] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, "Front-end for antispoofing countermeasures in speaker verification: Scattering spectral decomposition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 632–643, Jun. 2017.
- [29] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Speaker Lang. Recognit. Workshop (ODYSSEY)*, Bilbao, Spain, 2016, pp. 283–290.
- [30] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, Sep. 2017.
- [31] X. Wang, Y. Xiao, and X. Zhu, "Feature selection based on CQCCs for automatic speaker verification spoofing," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Aug. 2017, pp. 32–36.
- [32] M. Withowski, S. Kacprasko, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 27–31.
- [33] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudasher, and V. Shchemelinin, "Audio replay attack detection with deep learning framework," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2017, pp. 82–86.
- [34] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2018, pp. 651–655.
- [35] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 1978, pp. 375–378.
- [36] C. J. Brown, "An efficient algorithm for the calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2698–2701, 1992.
- [37] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2119–2123.
- [38] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: Meta-data analysis and baseline enhancements," in *Proc. Speaker Lang. Recognit. Workshop (ODYSSEY)*, 2018, pp. 296–303.
- [39] J. Yang and L. Liu, "Playback speech detection based on magnitude-phase spectrum," *Electron. Lett.*, vol. 54, no. 14, pp. 901–903, 2018.
- [40] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spectra information for spoofing attack detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2373–2384, Dec. 2019.
- [41] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE Workshop Automat. Speech Recognit. Understand. (ASRU)*, Dec. 2011, pp. 24–29.
- [42] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 2067–2071.
- [43] 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," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1008–1012.
- [44] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and A. D. Reynolds, "t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Speaker Lang. Recognit. Workshop (ODYSSEY)*, 2018, pp. 312–319.
- [45] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, p. 2135.
- [46] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1078–1082.
- [47] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVspoof 2019," in *Proc. IEEE Automat. Speech Recognit. Understand. Workshop (ASRU)*, Singapore, Dec. 2019, pp. 1018–1025.
- [48] K. N. R. K. R. Alluri and A. K. Vuppala, "IIIT-H spoofing countermeasures for automatic speaker verification spoofing and countermeasures challenge 2019," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1043–1047.
- [49] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2019, pp. 1033–1037.



**LEIAN LIU** was born in Xinxian, Henan, China, in October 1979. He received the bachelor's and master's degrees from Zhengzhou University, Zhengzhou, Henan, in 2001 and 2004, respectively, and the Ph.D. degree in circuits and systems from the South China University of Technology, Guangzhou, Guangdong, China, in 2007. He is currently working as an Associate Professor with the School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou. His current research interests mainly include anti-spoofing, speech signal processing, machine learning, RFID technology, network security, and embedded technology.



**JICHEN YANG** (Member, IEEE) received the Ph.D. degree in communication and information system from the South China University of Technology (SCUT), Guangzhou, China, in 2010. He was a Postdoctoral Research Fellow with SCUT, from October 2011 to March 2016. Since April 2016, he has been a Postdoctoral Researcher Fellow with the Department of Human Language Technology, Institute for Infocomm Research (I2R), A\*STAR, Singapore, and then with the Human Language Technology Laboratory, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests mainly include anti-spoofing and forensics.

...