Applied Acoustics 157 (2020) 107017

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Improving anti-spoofing with octave spectrum and short-term spectral statistics information

Jichen Yang [1], Rohan Kumar Das [1]

Department of Electrical and Computer Engineering, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 2 March 2019
Received in revised form 19 July 2019
Accepted 24 August 2019

Keywords:
CQT spectrum
Anti-spoofing countermeasure
Octave power spectrum
Short-term spectral statistics

ABSTRACT

The long-term window based features have been found to be effective for spoofing attack detection. One such important countermeasure is constant-Q cepstral coefficients (CQCC) that is derived from constant-Q transform. During its extraction, the octave power spectrum is converted to the linear power spectrum by performing uniform resampling. However, the information from the octave power spectrum is different from that carried by the linear power spectrum. We believe that the octave power spectrum can offer complementary information to the linear power spectrum for spoofing attack detection. In this regard, we propose to combine the coefficients generated using both linear and octave power spectrum. The combined feature is referred to as extended CQCC (eCQCC), which is hypothesized to have better discriminative information for detection of spoofing attacks. In addition, we use the short-term spectral statistics information (STSSI) along with eCQCC feature to form another novel feature representation referred to as eCQCC-STSSI to have improved anti-spoofing countermeasure. We perform the studies with the proposed features for both synthetic and playback attacks using ASVspoof 2015 and ASVspoof 2017 version 2.0 corpus, respectively. The studies reveal that eCQCC outperforms the conventional CQCC feature as well most of the known systems showing importance of octave spectrum information. Further, the hybrid feature eCQCC-STSSI improve the performance of eCQCC feature due to the STSSI information combined with it.

1. Introduction

The recent works in the field of automatic speaker verification (ASV) have shown feasibility for practical systems. With this the detection of spoofing attacks has become a critical issue for successful speaker verification deployments. There are mainly four types spoofing attacks in ASV. They are text-to-speech synthesis (TTS) [1–4], voice conversion [5–7], replay [8–13] and impersonation [14,15]. In order to make ASV systems practically viable, there is a need to detect such kind of attacks. To effectively detect spoofing voice, it is very important to seek the features that can discriminate natural speech from the spoofed speech [16]. For synthetic speech detection, the goal is to seek the artifacts in spoofed speech, which is generated by TTS or voice converted speech [17,16]. While for playback speech detection, the goal is to seek the device and environment information in playback speech, which gets added to the genuine speech in the process of playback speech generation because of environmental effect and the usage of playback and recording devices [10,11,18]. The impersonation attacks are less vulnerable as they are based on the behavioral nature of the speakers. In this paper, we study ways to protect ASV systems from synthetic speech generated using voice conversion or TTS methods and playback speech.

Similar to most of the speech processing systems, a spoofing detection system usually contains a front-end feature extractor and a back-end classifier. In which, feature extractor plays the role of extracting effective signal representation and classifier plays the role of binary class detection for identifying spoofing attacks. Many countermeasures have been proposed along both these directions for protecting ASV systems from spoofed speech. Further, the work [19,20] suggests that more efforts must be used in designing countermeasures from feature rather than complex and advanced classifiers. Therefore, in this paper, we focus on feature level exploration. A new feature that can capture improved discriminative information between natural and spoofed speech from that of the existing features is explored. Next, we provide a brief survey on the past works on the different feature countermeasures.

[1] Both the authors have equal contribution and are joint first authors.
E-mail addresses: eleyji@nus.edu.sg (J. Yang), rohankd@nus.edu.sg (R.K. Das)

https://doi.org/10.1016/j.apacoust.2019.107017
0003-682X/

1.1. Related works

The feature-level countermeasures for spoof detection can be classified into two categories: hand-crafted design features and deep features that are obtained by learning using deep neural network (DNN) based models. The studies have shown that hand-crafted design features can perform better than deep features in synthetic speech detection [21,22] unlike for playback speech detection [23–25]. The reason for this may the variations for synthetic speech may be difficult to capture by deep features.

From the view of magnitude spectrum and phase spectrum, the hand-crafted features designed may be grouped into three categories: the first one is based on power (magnitude) spectrum [26–32], for instance, mel-frequency cepstral coefficients (MFCC) [33], constant-Q cepstral coefficients (CQCC) [19,20] and log magnitude spectrum [34]. The second category is based on phase spectrum [34–36], for instance, modified group delay and instantaneous phase [35,36]. Then, the third one is based on magnitude-phase spectrum, for instance, constant-Q magnitude-phase octave coefficients (CMPOC) [37]. For spoofing attack detection, it has been found that the performance of phase spectrum based feature is worse than traditional power spectrum based features [37]. Therefore, the phase spectrum based features are often combined with power spectrum based features to improve the performance to utilize the complementary information [34,35]. In addition, it has been also found that the features based on magnitude-phase spectrum perform better than features based on magnitude spectrum and feature based on phase spectrum [37].

From the view of transforming a signal from the time domain into frequency domain, the hand-crafted design features may be mainly classified into three categories: the first is based on discrete Fourier transform (DFT), for instance, MFCC, inverted MFCC (IMFCC), mel-warped overlapped block transformation (MOBT) [26], inverted MOBT (IMOBT) [26]. The second category is based on auditory transform, that includes features like cochlear filter cepatral coefficients instantaneous coefficients (CFCC-IF) [38]. The third is based on constant-Q transform (CQT), for instance, CQCC [19,20], CMPOC [37] and constant-Q spectral-plus-statistic information coefficients (CQSPIC) [39]. It is found that the features based on CQT (such as CQCC) performs much better than the feature based on DFT (such as MFCC) and the feature based on auditory transform (CFCC-IF) for spoofing attack detection. The reason is that CQT is a long-term window transform, while DFT is a short-tern window transform, and the features based on CQT provide more detailed artifacts than those based on DFT.

The literature on spoofing attacks detection shows that CQCC is the most widely used feature. It has shown effectiveness for spoofing attack detections for different kinds of attacks as investigated in [19,20,40–42]. In addition, CQCC is used to as baseline system for various studies [43,44]. This is due to the fact that it can seek some artifacts in synthetic speech detection and also capture some devices and environment information in playback speech detection.

1.2. Contribution of the work

Traditional features are mostly extracted based on linear power spectrum. The CQCC as studied in [19,20], uses uniform resampling to convert the octave power spectrum into linear power spectrum, then applies discrete cosine transform (DCT) on linear power spectrum to obtain CQCC. The rationale behind this is that DCT cannot be applied on octave power spectrum directly as every frequency bin has different bandwidth. However, we do believe that DCT can be applied over octave power spectrum to de-correlate among the feature dimensions. We note that the octave power spectrum and linear power spectrum have different characteristic. The octave power spectrum can reflect some characteristic of human auditory system, for instance, higher frequency resolution at low frequency and higher temporal resolution at high frequency. However, linear power spectrum doesn't have this characteristic. In linear power spectrum, every frequency bin has the equal frequency bandwidth. We can say that octave power spectrum and linear power spectrum can offer complementary information. In other words, we believe that the information obtained from octave power spectrum can provide complementary information for the information obtained from linear power spectrum.

The current work is an extension of our previous work [45]. In this paper, in order to have a better discriminative characteristics for spoofing attack detection, the information from linear power spectrum and octave power spectrum are used together. A novel feature is proposed, which is extracted not only from linear power spectrum but also using octave power spectrum. We refer to this proposed feature as extended constant-Q cepstral coefficients (eCQCC). We note that the feature obtained from the linear power spectrum is well known CQCC and we can refer the feature obtained from the octave power spectrum as constant-Q coefficients (CQC) in this work. The proposed eCQCC feature is obtained as a result of concatenation of CQCC and CQC feature.

Additionally, short-term spectral statistics information (STSSI) was proposed in our previous work [39]. It was found that it can help to detect spoofing attacks. Though eCQCC and STSSI are extracted from CQT spectrum, they represent different spectral information. The eCQCC represents spectral principal information, whereas, STSSI represents spectral statistics information. We hypothesize that eCQCC and STSSI carry complementary information and their combined knowledge can improve the detection of spoofing attacks. We therefore, combine them for the proposal of a hybrid feature referred to as eCQCC-STSSI. The proposed feature is investigated for both synthetic and replay speech based attacks using ASVspoof 2015 and ASVspoof 2017 version 2.0 database, respectively. In our studies, DNN based classifiers are used in the back-end as DNNs not only have a classifier function but also have a feature learning ability [46].

The remainder of the paper is organized as follows. Section 2 introduces the feature eCQCC-STSSI in detail. Section 3 and 4 report the experimental results and their analyses for synthetic and replay speech detection, respectively. Finally, Section 5 provides a discussion and Section 6 concludes the paper.

2. Feature extraction based on CQT spectrum

In this section, we provide the details for extracting eCQCC-STSSI feature. The eCQCC-STSSI feature is formed by eCQCC and STSSI, where eCQCC is extracted from linear along with octave power spectrum and STSSI is extracted from octave magnitude spectrum.

Fig. 1 shows the block diagram for extracting eCQCC-STSSI feature. From Fig. 1, we observe that eCQCC-STSSI consists of three sub-features, which are CQCC, CQC and STSSI. Among these, eCQCC is formed by CQCC and CQC as discussed. For CQCC feature extraction, there are six modules. These are CQT, magnitude spectrum, square, Log, uniform resampling and DCT. Compared to CQCC extraction, the CQC extraction differs by only the module of uniform resampling, where it is removed. In case of STSSI extraction, there are four modules that include CQT, magnitude spectrum, short-term spectral statistics and Log.

The extraction of proposed feature eCQCC-STSSI involves CQT, which is used to transform speech from the time domain into the frequency domain. The magnitude spectrum is then used to compute the octave magnitude spectrum value on the basis of CQT by using a module of square. The module of Log is considered to

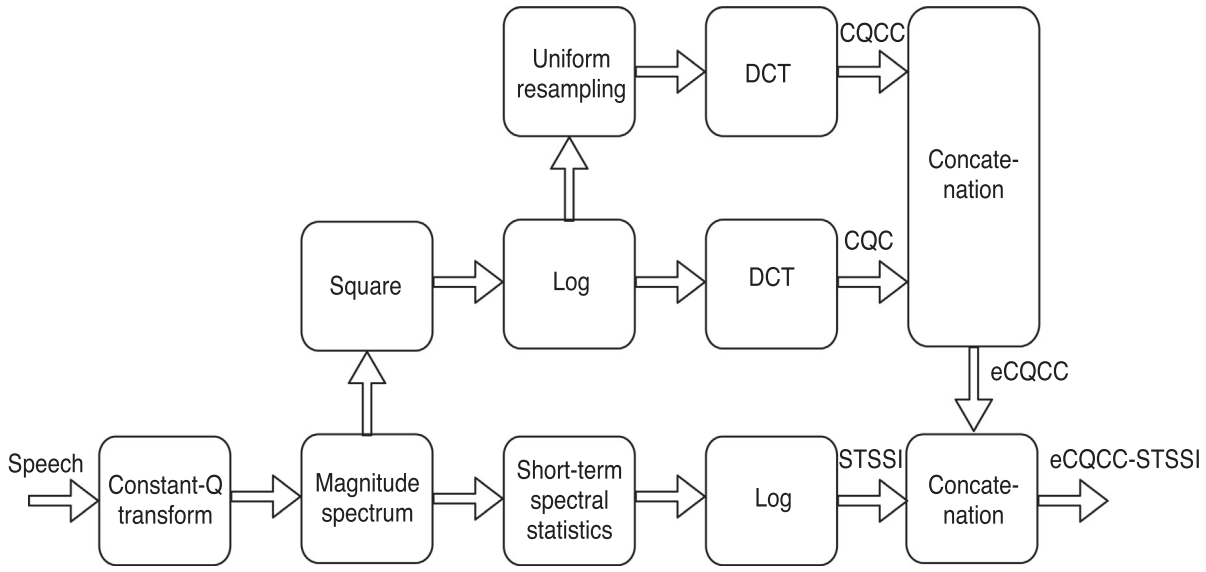


Fig. 1. Schematic diagram of eCQCC-STSSI extraction.

obtain logarithm octave power spectrum, followed by uniform resampling to convert logarithm octave power spectrum into linear power spectrum in logarithm scale. Then DCT is used to decorrelate the feature dimensions and concentrate energy of logarithm octave power spectrum and logarithm linear power spectrum, respectively. Finally, the two DCT outputs are concatenated to form the eCQCC feature vectors. For STSSI extraction, the module of short-term spectral statistics is used to obtain spectral statistics value on the basis of octave magnitude spectrum and then it is passed through Log module to have it in log-scale. Next, we discuss the various modules for extracting eCQCC and eCQCC-STSSI in detail.

2.1. Extended constant-Q cepstral coefficients

The eCQCC feature consists of two sub-feature, which are CQCC and CQC. For its extraction, the module of CQT, uniform resampling, DCT and concatenation are very important and they are will be introduced in detail.

2.1.1. Constant-Q transform

CQT is proposed in [47,48]. It is different from DFT as the ratio of center frequency to bandwidth is constant in CQT. As a result, CQT has a higher frequency resolution in low frequency and higher temporal resolution for higher frequency.

For a discrete time domain signal $x(n)$, its CQT, $Y(k,p)$ is defined as:

$$Y(k,p) = \sum_{j=n-\frac{N_k}{2}}^{n+\frac{N_k}{2}} x(j) a_k^*\left(j - n - \frac{N_k}{2}\right) \tag{1}$$

where $k = 1, 2, \ldots, K$ is the frequency bin index, $p = 1, 2, \ldots, P$ represents frame index and P is total frame number, N_k are the variable window lengths, $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$.

The basic functions of $a_k(n)$ are complex-valued time-frequency atoms and are defined by

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

where f_k is the center frequency of bin f_k, f_s is the sampling rate, and $v(t)$ is a window function (e.g. Hanning window) and ϕ_k is a phase offset. The scaling factor C is computed as

$$C = \sum_{m=-\frac{N_k}{2}}^{\frac{N_k}{2}} v\left(\frac{m + \frac{N_k}{2}}{N_k}\right) \tag{3}$$

In addition, a bin spacing corresponding to the equal temperament is desired in CQT, the center frequency (consider f_k) of k^{th} frequency bin obeys the following

$$f_k = f_1 2^{\frac{k-1}{B}} \tag{4}$$

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

In this way, we can obtain the frequency region (consider δ_f) of k^{th} frequency bin in the following way

$$\begin{aligned} \delta_f &= f_{k+1} - f_k \\ &= f_1 2^{\frac{k}{B}} - f_1 2^{\frac{k-1}{B}} \\ &= f_1 2^{\frac{k-1}{B}} (2^{\frac{1}{B}} - 1) \end{aligned} \tag{5}$$

From Eq. (5), we can observe that each frequency bin corresponds to a different frequency range in the CQT. As k increases its bandwidth also increases. This is different from the DFT, where all the frequency bins have the same bandwidth.

2.1.2. Uniform resampling

Uniform resampling is used to convert logarithm octave power spectrum into logarithm linear power spectrum, its more details can be found in [19,20]. For $Y(k,p)$, its logarithm octave power spectrum is $\log|Y(k,p)|^2$, in which $\log(\cdot)$ represents logarithm operation. In addition, we consider that logarithm linear power spectrum of $\log|Y(k,p)|^2$ is $\log|Y(l,p)|^2$.

2.1.3. Discrete cosine transform

DCT is used to de-correlate the feature dimensions and concentrate energy of logarithm octave power spectrum and logarithm linear power spectrum, respectively. We also can take $Y(k,p)$ as an example. After DCT is employed on $\log|Y(k,p)|^2$ and $\log|Y(l,p)|^2$, we obtain the coefficients as

$$C_O(0,p) = \frac{1}{\sqrt{N_O}} \sum_{k=1}^{N_O} \log|Y(k,p)|^2 \tag{6}$$

$$C_O(z,p) = \sqrt{\frac{2}{N_O}} \sum_{k=1}^{N_O} \log|Y(k,p)|^2 \cos\left\{\frac{(2k-1)z\pi}{2N_O}\right\} \tag{7}$$

$$C_L(0,p) = \frac{1}{\sqrt{N_L}} \sum_{l=1}^{N_L} \log|Y(l,p)|^2 \tag{8}$$

$$C_L(z,p) = \sqrt{\frac{2}{N_L}} \sum_{l=l}^{N_L} \log|Y(l,p)|^2 \cos\left\{\frac{(2l-1)z\pi}{2N_L}\right\} \tag{9}$$

where $C_O(0,p)$ and $C_O(z,p)$ represent 0^{th} and z^{th} order coefficients of *p-th* frame obtained from octave spectrum, and $C_L(z,p)$ represent 0^{th} and z^{th} order coefficients of *p-th* frame obtained for linear spectrum, respectively. Among which, $p = 1,2,\ldots,P, z$ is a positive integer and ranges from 1 to *Z*-1, where *Z* is the number of coefficients selected as feature vector dimension. N_O and N_L are the dimensions of $\log|Y(k,p)|^2$ and $\log|Y(l,p)|^2$, respectively. In addition, l represents linear frequency bin number, $l = 1,2,\ldots,N_l$.

2.1.4. Concatenation

Finally, we concatenate the information from logarithm octave power spectrum and logarithm linear power spectrum together to form eCQCC features. For $x(n)$, we can obtain its eCQCC feature, say $eCQCC_x$, in the following way

$$eCQCC_x = [\, C_O(0,p) \quad C_O(z,p) \quad C_L(0,p) \quad C_L(z,p) \,] \tag{10}$$

where $p = 1,2,\ldots,P$ and z ranges from 1 to *Z*-1.

2.2. eCQCC-STSSI extraction

In this subsection, we describe the details of STSSI and followed by extraction of the proposed eCQCC-STSSI feature.

2.2.1. STSSI extraction

There are two STSSI parameters that can be obtained from every frame of magnitude spectrum, which are mean and variance [39]. Considering $m(p)$ and $\sigma^2(p)$ are the mean and variance of $|Y(k,p)|$, respectively. The STSSI can be computed as follows

$$m(p) = \frac{1}{K} \sum_{k=1}^{K} |Y(k,p)| \tag{11}$$

$$\sigma^2(p) = \frac{1}{K} \sum_{k=1}^{K} (|Y(k,p)| - m(p))^2 \tag{12}$$

where $p = 1,2,\ldots,P$.

At the base of $m(p)$ and $\sigma^2(p)$, considering $STSSI_x$ is STSSI values of $|Y(k,p)|$, we can obtain:

$$STSSI_x = \left[\, \log(m(p)) \quad \log(\sigma^2(p)) \,\right] \tag{13}$$

where $p = 1,2,\ldots,P$.

2.2.2. eCQCC-STSSI extraction

At the base of eCQCC and STSSI, the eCQCC-STSSI feature can be obtained by concatenating both eCQCC and STSSI. For a signal $x(n)$, considering $(eCQCC - STSSI)_x$ is its eCQCC-STSSI feature, we have:

$$(eCQCC - STSSI)_x = \begin{bmatrix} C_O(0,p) & C_O(z,p) & C_L(0,p) \\ C_L(z,p) & \log(m(p)) & \log(\sigma^2(p)) \end{bmatrix} \tag{14}$$

where $p = 1,2,\ldots,P$ and z ranges from 1 to *Z*-1.

3. Synthetic speech detection studies

In this section, eCQCC and eCQCC-STSSI are used to study synthetic speech detection on ASVspoof 2015 database. We describe the experimental setup and report the results next.

3.1. Database

The ASVspoof 2015 corpus is constituted by three subsets: training set, development set and evaluation set, each part consists of natural and spoofed speech. The spoofed speech is generated from original genuine speech with different voice conversion and speech synthesis algorithms. There are 10 spoofing-attack algorithms (referred as S1 to S10) to generate the spoofed utterances, among which, S1, S2, S5-S9 are obtained using voice conversion algorithms and S3, S4 and S10 are obtained using speech synthesis algorithms, more details can be found in [49]. In addition, all the three subsets contain spoofing type S1 to S5, which are denoted as known attacks, whereas S6 to S10 only appear in the evaluation subset and are referred as unknown attacks. ASVspoof 2015 corpus is often used for synthetic speech detection based studies. Table 1 summarizes the composition of the database.

3.2. Experiment setup

According to the ASVspoof 2015 challenge protocol, there are 3750 genuine utterances and 12,625 spoofed utterances from the training set that are used to train respective models. Development data can be used to tune the model parameters. Equal error rate (EER) for individual condition and average equal rate (AEER) across all the conditions are used as evaluation metrics for this database.

In CQT, all parameters are set according to [19,20], which are the number of bins per octave set to 96, the number of octaves set to 9, the sampling period set to 16 and the gamma set to 3.3026. In speaker recognition and speech recognition, 13 and 20 are often selected as the feature static dimension number (SDN). In addition, a high number, for example, 30, can be used to investigate whether higher order coefficients contain additional useful information [19,20]. Thus, *Z* is set as 13, 20 and 30 in our work. In other words, 13, 20 and 30 dimensional feature vectors are obtained from linear power spectrum and octave power spectrum, respectively. We have used the equal dimensions from both the power spectra that results in 26, 40 and 60 as the feature SDN in case of eCQCC.

We use static (S), delta (D) and acceleration (A) coefficients in feature configuration of eCQCC. The computational network Toolkit (CNTK) [50] is used to train DNN, which is used as classifier in our experiment. In the experiments, a series of 6-layer DNN classifier is trained, which has 4 hidden layer with 512 nodes at every layer along with output layer with 2 nodes and the input node is constituted by a 11-frame context window of the input feature vector. The hidden layers training is used by sigmoid network and cross-entropy with softmax is used as training criterion. The input data is normalized by using mean and variance normalization. In DNN training, stochastic gradient descent is used. There are 25 epochs for every DNN training, in which the first epoch has a learning rate of 0.8, 3.2 for the next 14 epochs and then 0.08 for the rest epochs. The first epoch has a minibatch size of 256 and the rest epochs have a minibatch size of 1024. In addition, a momentum value is set as 0.9.

3.3. Studies with eCQCC feature

Table 2 shows the experimental results on the development set of ASVspoof 2015 database using different feature configurations

Table 1
Summary of ASVspoof 2015 database.

Subset	# Speakers		# Utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3,750	12,625
Development	15	20	3,497	49,875
Evaluation	20	26	9,404	184,000

of eCQCC features under different SDNs. We have several observations from Table 2: (1) When SDN equals 26, eCQCC-A, eCQCC-SA, eCQCC-SDA and eCQCC-SD perform much better than eCQCC-D, eCQCC-DA and eCQCC-S according to AEER. (2) When SDN equals 40, except eCQCC-D, the rest six feature configurations of eCQCC can capture the artifacts well in ASVspoof 2015 development set. (3) When SDN equals 60, except eCQCC-S and eCQCC-SD, the rest five feature configurations of eCQCC can capture all the artifacts. (4) Finally, eCQCC-A, eCQCC-SA and eCQCC-SDA consistently outperform others, suggesting that they are more reliable representation. Therefore, we have decided to use eCQCC-A, eCQCC-SA and eCQCC-SDA as the feature of eCQCC for evaluation set on ASVspoof 2015.

We now study the performance of eCQCC features on the ASVspoof 2015 evaluation set. Table 3 shows the experiments under different SDNs using different feature configurations of eCQCC on ASVspoof 2015 evaluation set. We can observe the following: (1) Irrespective of SDN, the dynamic features (eCQCC-D, eCQCC-A and eCQCC-DA) perform better than the static features (eCQCC-S, eCQCC-SD, eCQCC-SA and eCQCC-SDA) on ASVspoof 2015 evaluation set. (2) Among dynamic features of eCQCC, eCQCC-A performs better than eCQCC-D and eCQCC-DA on ASVspoof 2015 evaluation set for all kinds of SDNs. (3) eCQCC-A provides the best performance on ASVspoof 2015 evaluation set when SDN equals 26. An AEER of 0.035% is obtained, which suggests that eCQCC-A well characterizes the artifacts in ASVspoof 2015 evaluation set. (4) In addition, we can observe that the higher order coefficients doesn't lead to better performance, which suggests that the discriminative information in ASVspoof 2015 evaluation set mainly locates in around the low order coefficients.

3.4. Studies with proposed eCQCC-STSSI feature

We then study the proposed eCQCC-STSSI feature for spoofing attack detection. Table 4 shows the experimental results on the ASVspoof 2015 development set using different feature configurations of eCQCC-STSSI feature. We note that the SDNs of eCQCC, STSSI and eCQCC-STSSI are 26, 2 and 28, respectively, where eCQCC-STSSI is a result of combination of eCQCC and STSSI. We observe from Table 4 that feature configurations of SD, SA and SDA can capture all the artifacts in ASVspoof 2015 development set, followed by feature configuration A, and feature configurations S, D and DA perform the worst.

Table 5 shows the different feature configurations of eCQCC-STSSI on ASVspoof 2015 evaluation set. Among which, the static feature dimensions of eCQCC and STSSI are set as 26 and 2, respectively. We can observe the following from Table 5: (1) Dynamic feature configurations (D, A and DA) can perform better than static features configurations (S, SD, SA and SDA) on ASVspoof 2015 evaluation set. However, the static feature configurations SD, SA and SDA can perform better than dynamic feature configurations D, A and DA on ASVspoof 2015 development set. This may be due to the nature of the data mismatch between development and evaluation set. We note that there are five unknown attacks, which only appear in evaluation set. (2) Feature configuration A gives the best performance among the dynamic feature configurations (D, A and DA), which indicates that there is more discriminative information in feature configuration A than in feature configurations D and DA. (3) In addition, we find the average EER reduces from 0.035% from 0.032% on comparing the performance of eCQCC-A in Table 3 and (eCQCC-STSSI)-A in Table 5. This shows that the STSSI with eCQCC can help to capture complementary artifacts, which results in improvement.

3.5. Comparison with features based on different spectrum

We now compare the proposed feature with various features based on different spectra for synthetic speech detection. From Fig. 1, it is observed that CQC and CQCC are extracted from octave power spectrum and linear power spectrum, respectively. Further, the eCQCC is extracted from both octave and linear power spectra.

Table 2
Performance in EER (%) using eCQCC features under different feature configuration (FC) on development set of ASVspoof 2015 database.

SDN	FC	S1	S2	S3	S4	S5	AEER
26	S	0.000	0.000	0.000	0.000	0.010	0.002
	D	0.000	0.000	0.000	0.000	0.010	0.002
	A	0.000	0.000	0.000	0.000	0.000	0.000
	SD	0.000	0.000	0.000	0.000	0.000	0.000
	SA	0.000	0.000	0.000	0.000	0.000	0.000
	DA	0.000	0.000	0.000	0.000	0.028	0.006
	SDA	0.000	0.000	0.000	0.000	0.000	0.000
40	S	0.000	0.000	0.000	0.000	0.007	0.002
	D	0.000	0.000	0.000	0.000	0.000	0.000
	A	0.000	0.000	0.000	0.000	0.000	0.000
	SD	0.000	0.000	0.000	0.000	0.000	0.000
	SA	0.000	0.000	0.000	0.000	0.000	0.000
	DA	0.000	0.000	0.000	0.000	0.000	0.000
	SDA	0.000	0.000	0.000	0.000	0.000	0.000
60	S	0.000	0.000	0.000	0.000	0.007	0.002
	D	0.000	0.000	0.000	0.000	0.000	0.000
	A	0.000	0.000	0.000	0.000	0.000	0.000
	SD	0.000	0.000	0.000	0.000	0.007	0.002
	SA	0.000	0.000	0.000	0.000	0.000	0.000
	DA	0.000	0.000	0.000	0.000	0.000	0.000
	SDA	0.000	0.000	0.000	0.000	0.000	0.000

Table 3
Experiment results on ASVspoof 2015 evaluation set using eCQCC under different SDNs and different feature configurations (FC) in terms of EER (%).

SDN	FC	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	AEER
	S	0.000	0.019	0.000	0.000	0.042	0.035	0.015	0.000	0.008	9.529	0.965
	D	0.000	0.008	0.000	0.000	0.049	0.016	0.005	0.139	0.004	0.938	0.116
	A	**0.000**	**0.007**	**0.000**	**0.000**	**0.005**	**0.005**	**0.000**	**0.004**	**0.000**	**0.300**	**0.035**
26	SD	0.000	0.000	0.000	0.000	0.015	0.011	0.000	0.000	0.000	5.119	0.514
	SA	0.000	0.000	0.000	0.000	0.021	0.013	0.000	0.000	0.000	4.740	0.477
	DA	0.000	0.008	0.000	0.000	0.045	0.016	0.005	0.084	0.000	0.797	0.096
	SDA	0.000	0.000	0.000	0.000	0.022	0.014	0.000	0.000	0.000	4.025	0.406
	S	0.000	0.032	0.000	0.000	0.028	0.041	0.030	0.000	0.015	7.738	0.689
	D	0.000	0.000	0.000	0.000	0.008	0.004	0.000	0.014	0.000	1.688	0.171
	A	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.037**	**0.000**	**0.791**	**0.083**
40	SD	0.000	0.004	0.000	0.000	0.020	0.016	0.004	0.000	0.000	3.743	0.379
	SA	0.000	0.005	0.000	0.000	0.020	0.019	0.000	0.000	0.000	3.750	0.380
	DA	0.000	0.000	0.000	0.000	0.007	0.000	0.000	0.007	0.000	1.043	0.106
	SDA	0.000	0.000	0.000	0.000	0.016	0.015	0.000	0.000	0.000	3.190	0.322
	S	0.000	0.037	0.000	0.000	0.040	0.044	0.035	0.000	0.027	5.557	0.575
	D	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.011	0.000	1.010	0.102
	A	**0.000**	**0.006**	**0.000**	**0.000**	**0.005**	**0.005**	**0.005**	**0.083**	**0.004**	**0.756**	**0.087**
60	SD	0.000	0.005	0.000	0.000	0.021	0.023	0.005	0.000	0.000	3.198	0.325
	SA	0.000	0.009	0.000	0.000	0.024	0.019	0.009	0.000	0.004	3.146	0.321
	DA	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007	0.000	0.938	0.095
	SDA	0.000	0.089	0.000	0.000	0.021	0.027	0.015	0.000	0.004	2.799	0.287

Table 4
Performance in EER (%) using different feature configurations (FC) of eCQCC-STSSI feature on ASVspoof 2015 development set.

FC	S1	S2	S3	S4	S5	Avg.
S	0.000	0.000	0.000	0.000	0.024	0.005
D	0.000	0.000	0.000	0.000	0.027	0.005
A	0.000	0.000	0.000	0.000	0.017	0.003
SD	0.000	0.000	0.000	0.000	0.000	0.000
SA	0.000	0.000	0.000	0.000	0.000	0.000
DA	0.000	0.000	0.000	0.000	0.027	0.005
SDA	0.000	0.000	0.000	0.000	0.000	0.000

Table 5
Performance in EER (%) using different feature configurations (FC) of eCQCC-STSSI feature on ASVspoof 2015 evaluation set.

Type	S	D	A	SD	SA	DA	SDA
S1	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
S2	0.018	0.004	**0.000**	0.000	0.010	0.004	0.000
S3	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
S4	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
S5	0.040	0.032	**0.005**	0.022	0.024	0.025	0.024
S6	0.029	0.015	**0.005**	0.017	0.021	0.009	0.018
S7	0.014	0.000	**0.000**	0.000	0.008	0.000	0.000
S8	0.000	0.076	**0.015**	0.000	0.000	0.045	0.000
S9	0.009	0.000	**0.000**	0.000	0.007	0.000	0.000
S10	7.828	0.829	**0.292**	5.511	4.711	0.637	4.550
Avg.	0.794	0.096	**0.032**	0.555	0.478	0.072	0.459

Table 6
Performance comparison in AEER (%) among CQC-A, CQCC-A, eCQCC-A and (eCQCC-STSSI)-A on ASVspoof 2015 evaluation set.

Feature	Spectrum	AEER
CQC-A	Octave power	0.517
CQCC-A	Linear power	0.112
eCQCC-A	Octave and linear power	0.035
(eCQCC-STSSI)-A	Octave magnitude, octave and linear power	0.032

Table 7
Comparison with some known systems on synthetic speech detection on ASVspoof 2015 evaluation set in terms of AEER(%).

Feature	Classifier	AEER
SpeCNN [22]	RNN	1.860
Dfea [21]	LDA	1.600
Dfea [21]	SVM	1.400
MFCC [33]	GMM	2.120
CQCC-A [19]	GMM	0.260
SCMC [33]	GMM	0.940
CAF [51]	GMM	0.050
CQSPIC-A [39]	DNN	0.038
CQCC-A	DNN	0.112
eCQCC-A	DNN	0.035
(eCQCC-STSSI)-A	DNN	0.032

We note that the proposed feature eCQCC-STSSI is extracted from octave power spectrum, linear power spectrum and octave magnitude spectrum. Table 6 shows a comparison among CQC-A, CQCC-A, eCQCC-A and (eCQCC-STSSI)-A. The SDNs of CQC, CQCC, eCQCC and eCQCC-STSSI are considered as 13, 13, 26 and 28, respectively. In addition, every feature has its own DNN classifier trained for ASVspoof 2015 evaluation set. Their training methods are the same as eCQCC DNN classifiers described earlier for ASVspoof 2015 evaluation set.

From Table 6, the following can be found out: (1) eCQCC-A performs better than CQC-A and CQCC-A on ASVspoof 2015 evaluation set in terms of AEER. The performance with eCQCC-A improves by 92.3% when compared to CQC-A, which indicates that the linear power spectrum has complementary information from the octave power spectrum for synthetic speech detection. (2) With respect to CQCC-A, the AEER of eCQCC-A reduces by 64%, which proves the additional information carried by both the power spectra. (3) (eCQCC-STSSI)-A performs a little better than eCQCC-A on ASVspoof 2015 evaluation set, the average EER can reduce 8.57%, which indicate that STSSI can help eCQCC to improve the detection performance. This confirms our hypothesis to combine eCQCC and STSSI to form eCQCC-STSSI in order to have improved detection of spoofing attacks.

3.6. Comparison with some known systems

In this subsection, we compare the proposed feature with some of the well known systems. Table 7 shows the comparison of eCQCC and eCQCC-STSSI with some known systems on synthetic

Table 8
Summary of ASVspoof 2017 V2 database.

Subset	# Speakers	# Utterances	# Genuine	# Spoofed
Training	10	3,014	1,507	1,507
Development	8	1,710	760	950
Evaluation	24	13,306	1,298	12,008

speech detection on ASVspoof 2015 evaluation set. We note that speCNN represents deep feature obtained from convolutional neural network (CNN) by the input of spectrum [22], Dfea represents deep feature obtained from recurrent neural network (RNN) by the input of Fbank or MFCC [21], CAF represents using score-level fusion from CQCC, all-pole group delay function and fundamental frequency variation [51]. Further, SCMC represents sub-band spectral centroid magnitude coefficients, LDA represents linear discriminant analysis and SVM represents support vector machine.

From Table 7, we can observe the following: (1) Deep feature (SpeCNN and Dfea) based systems perform worse than many hand-crafted features for synthetic speech detection. The reason may be that it may be difficult to capture the variants in synthetic speech that discriminates it from genuine speech by using deep features. (2) The eCQCC based system performs better than systems based on some commonly used features such as MFCC, CQCC and SCMC on synthetic speech detection. (3) (eCQCC-STSSI)-A based system provides the best performance among the considered systems. Among these systems, CQSPIC-A, eCQCC-A and (eCQCC-STSSI)-A are based on DNN as classifier. Thus, we can conclude that eCQCC-STSSI and eCQCC can perform better in synthetic speech detection according to their performance on ASVspoof 2015 evaluation set.

4. Playback speech detection studies

In this section, the studies related to replay attacks using eCQCC and eCQCC-STSSI features are reported on ASVspoof 2017 Version 2.0 database (ASVspoof 2017 V2) for playback speech detection. The details are mentioned in the following subsections.

4.1. Database

The ASVspoof 2017 V2 corpus was collected using 26 playback devices and 25 recording devices in 26 different environments [11,18]. It was originally released for the ASVspoof 2017 challenge [11]. However, the organizers found some zero-value samples and silence regions in ASVspoof 2017 corpus that can affect the result of playback detection. In 2018, the organizers updated ASVspoof 2017 by removing those zero-value samples and silence regions, and named the corrected version as ASVspoof 2017 V2 [18]. This database is constituted by three subsets: training, development and evaluation set. Table 8 summarizes the composition of the ASVspoof 2017 V2 database.

4.2. Experiment setup

According to ASVspoof 2017 challenge protocol, the performances are to be reported on two sets, namely, development and evaluation set. The results on the development set can be used for tuning the performance of the evaluation set. Additionally, EER is used as the primary evaluation metric. The experimental setup for replay attack based studies follows the same that is considered for synthetic speech detection. In addition, a series of 4 layers DNN with 2 hidden layers DNN classifier is trained using ASVspoof 2017 V2 training set, in which the training algorithm is the same as that of DNNs training for synthetic speech detection described in previous section.

4.3. Studies with eCQCC feature

Table 9 shows the results using different feature configurations of eCQCC features on ASVspoof 2017 V2 development set. We can have the following observations: (1) For different SDNs, dynamic features (eCQCC-A, eCQCC-D and eCQCC-DA) perform better than static features (eCQCC-S, eCQCC-SD, eCQCC-SA and eCQCC-SDA) on ASVspoof 2017 development set. The reason may be that there is much discriminative information in dynamic features in playback speech detection. (2) For all SDN setups, eCQCC-A always gives the best performance followed by eCQCC-DA. (3) When SDN equals 60, the EER of eCQCC-A reaches minimum. (4) In conclusion, when SDN equals 60, eCQCC-A and eCQCC-DA can be used as features to evaluate ASVspoof 2017 V2 evaluation set.

Table 10 reports the results using different feature configurations of eCQCC on ASVspoof 2017 V2 evaluation set when SDN equals 60. From Table 10, it can be observed that the performance of eCQCC-DA is much better than eCQCC-A, unlike the trend of results obtained on the development set. This may be due to the fact that the recording and playback devices along with the environments used are very different for evaluation set than that used in development set. In addition, it can be observed that the dynamic feature configurations (D, A and DA) perform better than static feature configurations (S, SD, SA and SDA) on ASVspoof 2017 evaluation set. This indicates that there is more discriminative information in dynamic feature configurations than their static counterparts.

4.4. Studies with eCQCC-STSSI feature

Table 11 reports the results using different feature configurations of eCQCC-STSSI feature on ASVspoof 2017 V2 development set. We note that the SDNs of eCQCC, STSSI, eCQCC-STSSI are 60, 2 and 62, respectively. From Table 11, it can be observed that: (1) The dynamic features (D, A and DA) can perform better than the static features (S, SD, SA and SDA) on ASVspoof 2017 V2 development set. Thus, there is more discriminative information in dynamic features than in static features. (2) The feature configuration A performs best among dynamic features followed by DA on ASVspoof 2017 development set. Therefore, (eCQCC-STSSI)-A and (eCQCC-STSSI)-DA can be used to evaluate on ASVspoof 2017 V2 evaluation set.

Table 12 shows the results using different feature configurations of eCQCC-STSSI on ASVspoof 2017 V2 evaluation set. The SDN of eCQCC-STSSI is 62 as discussed earlier. From Table 12, we find that the dynamic feature configurations (D, A and DA) also perform better than static feature configurations (S, SD, SA and SDA) on ASVspoof 2017 V2 evaluation set. The same conclusion have been obtained from ASVspoof 2017 V2 development set using different feature configurations of eCQCC-STSSI. However, comparison of Table 11 with Table 12 shows an interesting trend. The (eCQCC-STSSI)-A feature performs better than (eCQCC-STSSI)-DA on development set, while (eCQCC-STSSI)-DA can give better performance (eCQCC-STSSI)-A on evaluation set of ASVspoof 2017 V2 corpus. We have obtained the same trend from the performance of eCQCC-A and eCQCC-DA. This may be due to the fact that some of the environments, recording and playback devices only appear in the evaluation set.

4.5. Comparison with features based on different spectrum

In this subsection, the performance of CQC-DA, CQCC-DA, eCQCC-DA and (eCQCC-STSSI)-DA is compared to observe their

Table 9
Performance in EER (%) using different feature configurations of eCQCC under different SDN on ASVspoof 2017 V2 development set.

SDN	Feature Configurations						
	D	A	DA	S	SD	SA	SDA
26	18.53	14.91	16.35	30.72	30.97	29.80	30.88
40	17.23	13.90	17.59	34.89	35.43	35.53	35.87
60	17.26	13.43	13.97	36.59	36.15	36.16	36.85

Table 10
Performance in EER (%) using different feature configurations (FC) of eCQCC on ASVspoof 2017 V2 evaluation set.

FC	S	D	A	SD	SA	DA	SDA
EER	24.27	13.87	16.94	25.06	23.93	**13.38**	27.96

Table 11
Performance in EER (%) using different feature configurations (FC) of eCQCC-STSSI feature on ASVspoof 2017 V2 development set.

FC	S	D	A	SD	SA	DA	SDA
EER	36.66	14.00	**8.94**	37.07	36.67	**10.46**	36.74

Table 12
Performance in EER (%) using different feature configurations (FC) of eCQCC-STSSI feature on ASVspoof 2017 V2 evaluation set.

FC	S	D	A	SD	SA	DA	SDA
EER	25.16	10.55	12.99	23.31	24.36	**10.07**	27.07

importance for playback speech detection. Table 13 shows the comparison for this study with the mentioned feature on ASVspoof 2017 V2 evaluation set. We note that the SDN of CQC, CQCC, eCQCC and eCQCC-STSSI are 30, 30, 60 and 62, respectively. Further, every feature has its individual DNN classifier for ASVspoof 2017 V2 evaluation set. Their training methods are the same as that of eCQCC based DNN classifier for ASVspoof 2017 V2 evaluation set.

From Table 13, we can note the following: (1) The eCQCC-DA feature performs better than CQC-DA and CQCC-DA on ASVspoof 2017 V2 evaluation set in terms of EER. The performance with eCQCC-DA improves by 34% when compared to CQC-DA features. (2) On comparing with CQCC-DA performance, EER of eCQCC-DA reduces by 20%. This indicates the complementary nature of information being carried by linear and octave power spectra. Thus, this confirms our hypothesis for replay attack based spoofing attack detection similar to the case of synthetic speech detection. (3) Finally, (eCQCC-STSSI)-DA can provide the best performance among the four features, where, STSSI reduces the EER of eCQCC by 24.74%, to be precise from 13.38% to 10.07%. This depicts that STSSI can help eCQCC to seek additional artifacts for playback speech detection.

4.6. Comparison with some known systems

Table 14 reports the comparison with some known systems on playback speech detection on ASVspoof 2017 evaluation set. In which, DSia represents deep Siamese obtained from two CNNs [25], FFTL represents log power magnitude based on FFT as the input for light CNN and CNN to learn deep feature [23] and ResNet presents residual neural network [41]. Further, CQCC-E represents combination CQCC and log energy and CQNSC represents constant-Q normalization segmentation coefficients.

From Table 14, we can observe the following:

- The deep feature (DSia and FFTL) based systems can perform better than general hand-crafted features on playback speech detection. However, they depend on heavily training data and therefore if evaluation data does not match to the training data conditions, the performance falls.
- The eCQCC based system performs better than systems based on some commonly used features such as MFCC, CQCC and SCMC on playback speech detection.
- The eCQCC performs a little worse than CQSPIC in playback speech detection. The reason may be that eCQCC is derived from full frequency band unlike CQSPIC, which is extracted using full frequency bands, subbands and STSSI. However, if STSSI is used to combined with eCQCC to form eCQCC-STSSI, the performance of eCQCC-STSSI is much better than CQSPIC. Thus, eCQCC-STSSI can perform better than CQSPIC in playback speech detection

Table 13
Performance comparison in EER (%) among CQC-DA, CQCC-DA, eCQCC-DA and (eCQCC-STSSI)-DA features on ASVspoof 2017 V2 evaluation set.

Feature	Spectrum	EER
CQC-DA	Octave power	18.73
CQCC-DA	Linear power	15.46
eCQCC-DA	Octave and linear power	13.38
(eCQCC-STSSI)-DA	Octave magnitude, octave and linear power	10.07

Table 14
Performance comparison in EER (%) of proposed feature with some known systems on ASVspoof 2017 V2 evaluation set.

Feature	Classifier	EER
DSia [25]	GMM	6.40
FFTL [23]	RNN	7.37
MFCC [41]	GMM	16.26
CQCC [41]	ResNet	18.79
(CQCC-E)-SDA [18]	GMM	12.24
SCMC [52]	GMM	11.49
CMPOC-D [37]	DNN	14.93
CMPOC-DA [37]	DNN	14.99
CQSPIC-DA [39]	DNN	11.09
CQNSC-DA [29]	DNN	10.63
CQCC-DA	DNN	15.46
eCQCC-DA	DNN	13.38
(eCQCC-STSSI)-DA	DNN	10.07

Table 15
The performance of CQCC in combination with some commonly used features on synthetic speech detection on ASVspoof 2015 evaluation set and playback speech detection on ASVspoof 2017 V2 evaluation set in terms of AEER (%) and EER (%), respectively. Note that the coefficient configurations of the features for ASVspoof 2015 and ASVspoof 2017 V2 are A and DA, respectively.

Features and Combinations	ASVspoof	
	2015	2017 V2
CQCC	0.112	15.46
MFCC	2.228	24.40
IMFCC	2.204	18.53
MOBT	1.395	24.42
IMOBT	3.083	18.57
Dfea	0.307	9.95
CQCC + MFCC + IMFCC + MOBT + IMOBT	0.252	14.75
CQCC + MFCC + IMFCC + MOBT + IMOBT + Dfea	0.288	11.78
eCQCC	0.035	13.38
eCQCC-STSSI	0.032	10.07

according to their performance on ASVspoof 2017 V2 evaluation set. In addition, we observe that eCQCC-STSSI performs better than CQNSC.

5. Discussion

The proposed eCQCC-STSSI is a combined feature that possesses information from the octave and linear power spectrum as well as short-term spectrum statistics information. These variations information are complementary to one another due to their way of extraction by processing the power spectrum derived using CQT. Literature shows that complementary information can be useful to improve performance of various tasks [34,53] and our proposed eCQCC-STSSI is no exception to it. The different attributes discussed when combined yields an improved results as observed from the studies in previous section.

We are now interested to study whether the proposed combined feature is effective enough compared to various combinations of existing features. In this regard, we study some of the well known features and their combined performance with CQCC. The features considered for this study are MFCC, IMFCC, MBOT, IMBOT and deep feature (Dfea). It is to be noted that the Dfea is extracted using a deep feature extractor that considers log power magnitude based on CQT as the input. We have implemented and evaluated them on ASVspoof 2015 and ASVspoof 2017 V2 corpora for synthetic and playback speech detection, respectively. Table 15 shows the performance comparison of these features and their combinations with CQCC. We find that our proposed eCQCC-STSSI outperforms the combinations of various existing features altogether with CQCC. This may be due to the more complementary information present in the octave power spectrum based coefficients and short-term spectrum statistics compared to the other features. Further, we note that when the other common features are combined to CQCC, the performance improves for the playback speech detection by some margin, unlike the case with synthetic speech detection. The complementary information from octave, linear power spectrum and short-term spectrum statistics in our proposed feature eCQCC-STSSI helps to achieve the best performance in case of both synthetic and playback speech detection.

6. Conclusions

This work focuses on proposal of novel features for improving performance of anti-spoofing systems. The long-term CQT based CQCC features are one of the most strong features for spoofing attack detection. However, they do not use the information from octave power spectrum, instead resamples it to linear power spectrum to obtain the feature. In this regard, we proposed a novel feature referred to as eCQCC by combination of the coefficients obtained from octave and linear power spectrum that offer complementary attributes. Further, we consider the STSSI from the octave magnitude spectrum and use along with eCQCC feature to derive a hybrid feature referred to as eCQCC-STSSI. The studies for synthetic and playback speech detection conducted on ASVspoof 2015 and ASVspoof 2017 V2 databases show the importance of the proposed features for detection of spoofing attacks. Both eCQCC and eCQCC-STSSI outperform the baseline with CQCC feature as well as many known systems for spoof detection. We note that the addition of STSSI helps to enhance the performance of eCQCC feature that reflects in terms of the best performance achieved in terms of eCQCC-STSSI feature compared to existing features as well as their combinations.

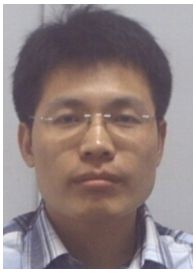
Acknowledgements

This research is supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain) and in part by the National Nature Science Foundation of China Grant U180120050, Grant 61702192, and Grant U1636218.

References

[1] De Leon Phillip L, Pucher Michael, Yamagishi Junichi, Hernaez Inma, Saratxaga Ibon. Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans Audio, Speech, Language Process 2012;20(8):2280–90.
[2] Yamagishi Junichi, Nose Takashi, Zen Heiga, Ling Zhen-Hua, Toda Tomoki, Tokuda Keiichi, et al. Robust speaker-adaptative HMM-based text-to-speech synthesis. IEEE Trans Audio, Speech, Language Process 2009;17(6):1208–30.
[3] Zen Heiga, Gales Mark JF. Yoshihiko Nankaku, and Keiichi Tokuda, Product of experts for statistical parametric speech synthesis. IEEE Trans Audio, Speech, Language Process 2012;20(3):794–805.
[4] Shannon Matt, Zen Heiga, Byrne William. Autoregressive models for statical parametric speech synthesis. IEEE Trans Audio, Speech, Language Process 2013;21(3):587–97.
[5] Erro Daniel, Moreno Asuncion, Bonafonte Antonio. Voice conversation based on weighted frequency warping. IEEE Trans Audio, Speech, Language Process 2010;18(5):922–31.
[6] Erro Daniel, Navas Eva, Hernaez Inma. Parametric voice conversation based on bilinear frequency warping plus amplitude scaling. IEEE Trans Audio, Speech, Language Process 2013;21(3):556–66.
[7] Tian Xiaohai, Lee Siuwa, Zhizheng Wu, Chng Eng Siong, Li Haizhou. An example-based approach to frequency warping for voice conversation. , IEEE/ACM Trans Audio, Speech, Language Process 2017;25(10):1863–75.
[8] Shang Wei, Stevenson Maryhelen. A preliminary study of factors affecting the performance of a playback attack detector. Proceddings of Canadian Conference on Electrical and Computer Engineering (CCECE), Niagara Falls, ON, Canada. p. 459–64.
[9] Paul Anupama, Das Rohan Kumar, Sinha Rohit, Prasanna SRM. Countermeasure to handle replay attacks in practical speaker verification systems. In: International Conference on Signal Processing and Communications. p. 12–5.
[10] Kinnunen Tomi, Sahidullah Md, Falcone Mauro, Costantini Luca, Hautamaki Rosa Gonzalez, Thomsen Dennis, et al. RedDots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), NEW ORLEANS, USA. p. 5395–9.
[11] Tomi Kinnunen Md, Sahidullah Héctor Delgado, Todisco Massimiliano, Evans Nicholas, Yamagishi Junichi, Lee Kong Aik. The ASVspoof challenge: assessing the limits of replay spoofing attack detection. 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden. p. 5395–9.
[12] Jelil Sarfaraz, Das Rohan Kumar, Prasanna SRM. Rohit Sinha, Spoof detection using source, instantaneous frequecny and cepstral features. In: 18th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 22–6.
[13] Liu Yitong, Das Rohan Kumar, Li Haizhou. Multi-band spectral entropy information for detection of replay attacks. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China.
[14] Hautamäki Rosa González, Kinnunen Tomi. Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen, I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In: 14th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 930–4.

[15] Hautamäki Rosa González, Kinnunen Tomi, Hautamäki Ville, Laukkanen Anne Marial. Automatic versus human speaker verification:the case of voice mimicrys. Speech Commun. 2015;72:13–31.
[16] Yamagishi Junichi, Kinnunen Tomi, Evans Nicolas, De Leon Phillip L. Introduction to the issues on spoofing and countermeasures for automatic speaker verification. IEEE J Selected Topics Signal Process 2017;11:585–7.
[17] Zhizheng Wu, Yamagishi Junichi, Tomi Kinnunen Md, Sahidullah Aleksandr Sizov, Evans Nicholas, Todisco Massimiliano, Delado Hector. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. IEEE J Selected Topics Signal Process 2017;11:588–604.
[18] Delgado Héctor, Massimilliano Todisco Md, Sahidullah Nicholas Evans, Kinnunen Tomi, Lee KongAik, Yamagishi Junichi. ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements. Speaker and Language Recognition Workshop(ODYSSEY). p. 296–303.
[19] Todisco Massimiliano, Delgado Héctor, Evans Nicholas. A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. Speaker and language recognition workshop (ODYSSEY), Bilbao, Spain. p. 283–90.
[20] Todisco Massimiliano, Delgado Héctor, Evans Nicholas. Constant Q cepstral coefficients: a spoofing countermeasure for automatic Speaker verification. Computer Speech Language 2017;45:516–35.
[21] Qian Yanmin, Chen Nanxin, Kai Yu. Deep feature for automatic spoofing detection. Speech Commun. 2016;85:43–52.
[22] Zhang Chunlei, Chengzhu Yu, Hansen John HL. An investigation of deep learning frameworks for speaker verification anti-spoofing. IEEE J Selected Topics Signal Process 2017;11:684–94.
[23] Lavrentyeva Galina, Novoselov Sergey, Malykh Egor, Kozlov Alexander, Kudasher Oleg, Shchemelinin Vadim. Audio replay attack detection with deep learning framework. In: 18th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 82–6.
[24] Tom Francis, Jain Mohit, Dey Prasenjit. End-to-end audio replay attack detection using deep convolutional networks with attention. In: 19th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 681–5.
[25] Sriskandaraja Kaavya, Sethu Vidhyasaharan, Ambikairajah Eliathamby. Deep siamese architecture based replay detection for secure voice biometric. In: 19th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 671–5.
[26] Paul Dipjyoti, Pal Monisankha, Saha Goutam. Spectral features for synthetic speech detection. IEEE J Selected Topics Signal Process 2017;11:605–17.
[27] Jichen Yang, Rohan Kumar Das, and Haizhou Li, Significance of subband features for synthetic speech decetion, IEEE Transactions on Information Forensics and Security (Accepted with Minor Revision). .
[28] Yang Jichen, He Qianhua, Yongjian Hu, Pan Weiqiang. CBC-based synthetic speech detection. Int J Digital Crime Forensics 2019;11(2):63–74.
[29] Yang Jichen, Das Rohan Kumar. Low frequency frame-wise normalization over constant-Q transform for playback speech detection. Digital Signal Process 2019;80:30–9.
[30] Yang Jichen, Liu Leian, He Qianhua. Discriminative feature based on FWMW for playback speech detection. Electron. Lett. 2019;55(15):861–4.
[31] Das Rohan Kumar, Yang Jichen, Li Haizhou. Long range acoustic features for spoofed speech detection. In: 20th Annual Conference of the International Speech Communication Association (INTERSPEECH).
[32] Das Rohan Kumar, Yang Jichen, Li Haizhou. Long range acoustic and deep features perspective on ASVspoof 2019. Automatic Speech Recognition Understanding Workshop (ASRU). Submitted for publication.
[33] Sahidullah Md, Kinnunen Tomi, Hanilci Cemal. A comparison features for synthetic speech detection. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, Germany. p. 2087–91.
[34] Xiao Xiong, Tian Xiaohai, Steven Du, Haihua Xu, Chng Eng Siong, Li Haizhou. Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof challenge. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, Germany. p. 2052–6.
[35] Das Rohan Kumar, Li Haizhou. Instantaneous phase and excitation source features for detection of replay attacks. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, Hawaii. p. 1030–7.
[36] Srinivas Kantheti, Das Rohan Kumar, Patil Hemant A. Combining phase-based features for replay spoof detection system. International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, Taiwan. p. 151–5.
[37] Yang Jichen, Liu Leian. Playback speech detection based on magnitude-phase spectrum. Electron. Lett. 2018;54(14):901–3.
[38] Tanvina Patel B, Patil Hemant A. Cochlear filter and instantaneous frequency based features for spoofed speech detection. IEEE J Selected Topics Signal Process 2017;11:618–31.
[39] Yang Jichen, You Changhuai, He Qianhua. Feature with complementarity of statistics and principal information for spoofing detection. In: 19th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 651–5.
[40] Wang Xianliang, Xiao Yanhong, Zhu Xuan. Feature selection based on CQCCs for automatic speaker verification spoofing. In: 18th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 32–6.
[41] Chen Zhuxin, Xie Zhifeng, Zhang Weibin, Xu Xiangmin. ResNet and model fusion for automatic spoofing detection. In: 18th Annual Conference of the International Speech Communication Association(INTERSPEECH). p. 102–6.
[42] Ji Zhe, Li Zhi-Yi, Li Peng, An Maobo, Gao Shengxiang, Dan Wu, Zhao Faru. Ensemble learning for countermeasure of audio repaly spoofing attack in ASVspoof2017. In: 18th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 87–91.
[43] Kamble Madhu, Tak Hemlata, Patil Hemant A. Effectiveness of speech demodulation-based features for replay detection. In: 19th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 641–5.
[44] Sailor Hardik B, Kamble MadhuR, Patil Hemant A. Auditory filterbank learning for temporal modulation features in replay spoof speech detection. In: 19th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 666–70.
[45] Yang Jichen, Das Rohan Kumar, Li Haizhou. Extended constant-Q cepstral coefficients for detection of spoofing attacks. In: Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). p. 1024–9.
[46] Seide Frank, Li Gang, Chen Xie, Dong Yu. Feature engineering in context-dependent deep neural networks for conversational speech transcription. Workshop on Automatic Speech Recognition and Understanding (ASRU), Big Island, Hawaii, USA. p. 24–9.
[47] Youngberg James, Boll S. Constant-Q signal analysis and synthesis. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Tulsa, Oklahoma, USA. p. 375–8.
[48] Judith Brown C. An efficient algorithm for the calculation of a constant Q spectral transform. J Acoust Soc Am 1992;92(5):2698–701.
[49] Zhizheng Wu, De Leon Phillip L, Demiroglu Cenk, Khodabakhsh Ali, King Simon, Ling Zhen-Hua, et al. Anti-spoofing for text-independent speaker verification: an initial database, comparison of countermeasures, and human performance. IEEE/ACM Trans Audio, Speech, Language Process 2016;20(8):768–83.
[50] Seide Frank, Amit Agarwal CNTK. Microsoft's open-source deep learning toolkit, in. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2135–2135.
[51] Pal Monisakha, Paul Dipjyoti, Saha Goutam. Synthetic speech detection using fundamental frequency variation and spectral features. Computer, Speech Language 2018;48:31–50.
[52] Font Roberto, Espin Juan M, Cano Maria José. Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 challenge. In: 18th Annual Conference of the International Speech Communication Association (INTERSPEECH). p. 7–11.
[53] Das Rohan Kumar, Mahadeva Prasanna SR. Exploring different attributes of source information for speaker verification with limited test data. J Acoust Soc Am 2016;140(1):184–90.



Jichen Yang received Ph.D degree in Communication and Information System from South China University of Technology (SCUT), Guangzhou, China in 2010. He was a Postdoctoral Research Fellow from October 2011 to March 2016 in SCUT. Since April 2016, he has been a Postdoctoral Researcher Fellow initially at the Department of Human Language Technology, Institute for Infocomm Research (I^2R), A*STAR, Singapore and then in the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests mainly include anti-spoofing and forensics, speaker recognition, speaker diarization, emotion recognition and deep learning.

Rohan Kuman Das received Ph.D degree from Indian Institute of Technology (IIT) Guwahati in the year 2017 and B. Tech degree in Electronics and Communication Engineering from North-Eastern Hill University (NEHU), Shillong, India in the year 2010. His Ph.D. work focused on speaker verification using short utterances from the perspective of practical application oriented systems. Prior to his research in the field of speech processing, he has been a Project Scientist at Assam Science Technology and Environment Council (ASTEC) from 2010 to 2011. After completing doctoral studies he worked as a Data Scientist in a multinational company called Kovid Research Labs (now acquired by Kaliber Labs) and involved in speech analytics based application services in 2017. Currently, he is a Postdoctoral Research Fellow at National University of Singapore and continuing post-doctoral research work. He has published over 50 research papers in peer reviewed journals and conferences. His research interests are speech signal processing, speaker verification, anti-spoofing, machine learning and pattern recognition.