Extended Constant-Q Cepstral Coefficients for Detection of Spoofing Attacks

Jichen Yang, Rohan Kumar Das and Haizhou Li
National University of Singapore, Singapore
E-mail: nisonyoung@gmail.com, {rohankd, haizhou.li}@nus.edu.sg

***Abstract*—The constant-Q cepstral coefficients (CQCC) feature is one of the most effective feature in the field of spoof speech detection. The extraction of this feature involves constant-Q transform that captures long range information from the signal. It is followed by uniform resampling of the octave power spectrum to have linear power spectrum from which the CQCC features are obtained. However, we hypothesize that the information obtained from octave power spectrum is complementary with that captured by the linear spectrum. 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 spoof attacks. The studies for spoof detection are conducted on both synthetic voice and replay based spoofing attacks using ASVspoof 2015 and ASVspoof 2017 Version 2.0 database, respectively. The studies confirm that the proposed eCQCC feature consistently outperforms the baseline CQCC feature in all tasks.**

I. 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 [1], [2], voice conversion [3], [4], replay [5]–[8] and impersonation [9], [10]. In order to make ASV systems practically viable, there is a need to detect such attacks. To effectively detect spoofing voice, it is very important to seek the features that can discriminate natural and spoofed speech [3]. For synthetic speech detection, the goal is to seek the artifacts between natural and spoofed speech, which is generated by text-to-speech (TTS) system or voice converted speech [1]–[3]. While for replay speech detection, the goal is to seek the device and environment information between genuine and playback speech, which gets added to the genuine speech in the process of playback speech generation because of environment effect and the usage of playback and recording devices [11]–[13].

Many countermeasures have been proposed for the detection of spoofing attacks. These are either based on front-end feature or back-end classifier. Further, feature plays the role of extracting effective representation and classifier plays the role of binary classifier in the task of spoof speech detection. The work [14] 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 features level countermeasures. According to the features used, countermeasures may be mainly classified into two categories: one is based on power spectrum [14]–[16] and another is based on phase spectrum [17], [18]. In spoofing attack detections, the performance of phase spectrum based feature is worse than traditional power spectrum based features (for example, mel-frequency cepstral coefficients (MFCC) and constant-Q cepstral coefficients (CQCC)). Therefore, the phase spectrum based features are often combined with power spectrum based features for enhanced performance [17], [18].

There have been studies of different power spectrum based features for spoofing attack detections, for example, MFCC, inverted mel-frequency cepstral coefficients, mel-warped overlapped block transformation [15], speech-signal frequency cepstral coefficients [15] and CQCC [14], etc. Among them, MFCC and CQCC are the most widely used features in spoofing attack detections. In addition, CQCC have shown effectiveness for spoofing attack detections [14], [19]–[21]. The reason may be that it can seek some artifacts in synthetic speech detection and also capture some devices and environment information in replay speech detection.

Traditional features are mostly extracted based on linear power spectrum. The CQCC as studied in [14], [16], 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 the features. We note that the octave power spectrum doesn't offer the same level of detail as linear power spectrum, but octave power spectrum and linear power spectrum may offer complementary information. Further, octave power spectrum can reflect some characteristics of human auditory system, for instance, higher frequency resolution at low frequency and higher temporal time resolution at high frequency, unlike the linear spectrum. We hypothesize that if information from octave spectrum can be collectively used with that obtained with uniform sampling,

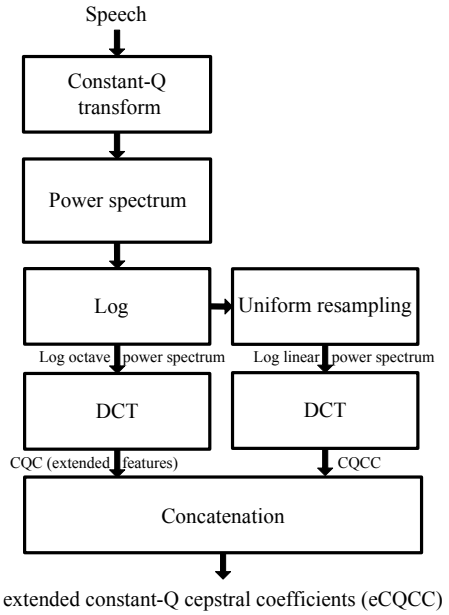

Fig. 1. Schematic diagram for extraction of extended constant-Q cepstral coefficients (eCQCC).

it can contribute towards better discrimination of natural and spoofed speech.

With the stated motivation in order to have a better discriminative characteristics for detection of spoof attacks, the information from linear power spectrum and octave power spectrum are used together. In other words, we propose a feature, 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), which is the main contribution of this work. A deep neural network (DNN) based classifier is used in the back-end as it not only has a classifier function but also has a feature learning ability [22]. The studies are performed for both synthetic and replay based spoofing attacks using ASVspoof 2015 and ASVspoof 2017 V2.0 database, respectively.

The remainder of the paper is organized as follows. Section II introduces the feature eCQCC in detail. Section III and IV mention the experimental results and their analysis are reported based on synthetic and replay speech detection, respectively. Finally, Section IV concludes the paper.

II. Introducing Extended constant-Q cepstral coefficients

In this section, we introduce the extraction process of eCQCC features in detail. Fig. 1 shows the block diagram for the extracting eCQCC features. It can be observed that there are six modules involved in the process, which consists of CQT, power spectrum, log, uniform resampling, DCT and feature concatenation.

The module of CQT is used to transform speech from the time domain into the frequency domain. Then power spectrum is used to calculate octave power spectrum value on the basis of CQT. The module of Log is used to obtain logarithm octave power spectrum, followed by uniform resampling to convert logarithm octave power spectrum into logarithm linear power spectrum. Finally, DCT is used to de-correlate 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. Next, we discuss the extraction process in detail.

A. Constant-Q transform

CQT is proposed in [23] and [24]. 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 $\mathbf{x}(n)$, its CQT $\mathbf{Y}(k,n)$ is defined as:

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

where $k = 1, 2, ..., K$ is the frequency bin index, N_k are the variable window lengths, $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$. The basic functions of $a_k(n)$ are complex-valued time-frequency atoms and are defined by

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

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

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

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.

B. Uniform resampling

Uniform resampling is used to convert logarithm octave power spectrum into logarithm linear power spectrum, its more details can be found in [14]. For $\mathbf{Y}(k,n)$, its logarithm octave power spectrum is $\log|\mathbf{Y}(k,n)|^2$, in which $\log(\cdot)$ represents logarithm operation. In addition, we consider that logarithm linear power spectrum of $\log|\mathbf{Y}(k,n)|^2$ is $\log|\mathbf{Y}(l,n)|^2$.

C. 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 $\mathbf{Y}(k,n)$ as an example. After DCT is employed on $\log|\mathbf{Y}(k,n)|^2$ and $\log|\mathbf{Y}(l,n)|^2$, we obtain the coefficients as

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

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

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

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

where $C_O(0)$ and $C_O(z)$ represent 0^{th} and z^{th} order coefficients obtained from octave spectrum; $C_L(0)$ and $C_L(z)$ represent 0^{th} and z^{th} order coefficients obtained for linear spectrum, respectively; 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 $\log|\mathbf{Y}(k,n)|^2$ and $\log|\mathbf{Y}(l,n)|^2$, respectively. In addition, l represents linear frequency bin number, $l = 1, 2, ..., N_l$.

D. Concatenation

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

$$\mathbf{eCQCC}_x = \begin{bmatrix} C_O(0) & C_O(z) & C_L(0) & C_L(z) \end{bmatrix} \tag{10}$$

where z ranges from 1 to Z-1.

III. Studies on synthetic speech detection

In this section, the studies related to synthetic speech detection using eCQCC features are reported on ASVspoof 2015 database. We describe the experimental setup and report the results next.

TABLE I
ASVspoof 2015 database specifications

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

TABLE II
Results (AEER(%)) on ASVspoof 2015 development set using eCQCC features under different configurations.

SDN	FC	S1	S2	S3	S4	S5	AEER
26	D	0.0	0.0	0.0	0.0	0.010	0.002
	A	0.0	0.0	0.0	0.0	0.0	0.0
	DA	0.0	0.0	0.0	0.0	0.028	0.006
40	D	0.0	0.0	0.0	0.0	0.0	0.0
	A	0.0	0.0	0.0	0.0	0.0	0.0
	DA	0.0	0.0	0.0	0.0	0.0	0.0
60	D	0.0	0.0	0.0	0.0	0.0	0.0
	A	0.0	0.0	0.0	0.0	0.0	0.0
	DA	0.0	0.0	0.0	0.0	0.0	0.0

A. Database description

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 conversation and speech synthesis algorithms. There are 10 spoofing-attack algorithms (referred as S1 to S10) to generate the spoofed utterances, their more details can be found in [1], [2]. 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 I summarizes the composition of the database.

B. Evaluation protocol

According to the ASVspoof 2015 challenge protocol, there are 3,750 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.

C. Experiment setup

In CQT, all parameters are set according to [14], 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, high number, for example, 30, can be used to investigate whether higher order coefficients contain additional useful information [14]. 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

TABLE III
EXPERIMENT RESULTS (AEER(%)) ON ASVSPOOF 2015 EVALUATION SET USING ECQCC-A UNDER DIFFERENT SDNS.

SDN	Known attack					Unknown attack					AEER
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
26	0.0	0.007	0.0	0.0	0.005	0.005	0.0	0.004	0.0	0.30	0.035
40	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.037	0.0	0.791	0.083
60	0.0	0.006	0.0	0.0	0.005	0.005	0.005	0.083	0.004	0.756	0.087

dimensions from both the power spectra that results in 26, 40 and 60 as the feature SDN in case of eCQCC.

As found in [14], static features may be counter effective in spoofing detection. Here we only use delta (D) and acceleration (A) in eCQCC. We are interested in the performance of eCQCC-A, eCQCC-D and eCQCC-DA as different feature combinations. The computational network Toolkit (CNTK) is used to train DNN, which is used as classifier in our experiment. In addition, during the DNN training process, stochastic gradient descent (SGD) is used. 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.

D. ASVspoof 2015 development set: Results and analysis

Table II shows the experimental results on the development set of ASVspoof 2015 database using different feature combinations (FC) of eCQCC features under different SDNs. We have several observations from Table II: (1) When SDN equals 26, eCQCC-A performs much better than eCQCC-D and eCQCC-DA according to AEER. (2) When SDN equals 40 or 60, the performance of eCQCC-A, eCQCC-D, eCQCC-DA is the same, which suggests that eCQCC captures the artifacts well in ASVspoof 2015 development set. (3) Finally, eCQCC-A consistently outperforms others, suggesting that it is a more reliable representation. Therefore, we have decided to use eCQCC-A as the feature for run-time testing on ASVspoof 2015 evaluation set.

E. ASVspoof 2015 evaluation set: Results and analysis

In this subsection, eCQCC-A is used as a feature to evaluate eCQCC performance on synthetic speech detection under ASVspoof 2015 evaluation set. Table III shows the experiments under different SDNs using eCQCC-A. It can be seen that 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. In addition, we can observe that the higher order of 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.

F. Features based on different power spectra: A comparison

Now we compare the group of features based on power spectrum for synthetic speech detection. In Fig. 1, let us consider the DCT coefficients obtained using only the octave power spectrum and we refer as constant-Q coefficients (CQC). Additionally, when we do not consider the modules related to the octave power spectrum in Fig. 1, the CQCC feature is obtained using only the linear power spectrum. We remind here that as the mentioned earlier, eCQCC feature is obtained using both linear power spectrum and octave poser spectrum.

TABLE IV
COMPARISON OF ECQCC-A WITH CQC-A AND CQCC-A ON ASVSPOOF 2015 EVALUATION SET IN TERMS OF AEER(%).

Feature	Power spectrum	AEER
CQC-A	Octave	0.52
CQCC-A	Linear	0.11
eCQCC-A	Octave and linear	0.04

Table IV provides a comparison among CQC-A, CQCC-A and eCQCC-A, in which the SDN of CQC, CQCC and eCQCC are considered as 13, 13 and 26, respectively. In addition, CQC-A and CQCC-A have their own DNN classifiers for ASVspoof 2015 evaluation set. Their training methods are the same as eCQCC DNN classifiers for ASVspoof 2015 evaluation set.

From Table IV, it can be seen that 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. Additionally, with respect to CQCC-A the AEER of eCQCC-A reduces by 64%, this too proves the additional information carried by both the power spectra. In conclusion, the linear power spectrum has complementary information from octave power spectrum for synthetic speech detection. Thus, when the information obtained from both of them are combined, it results into an improvement that confirms our hypothesis.

IV. STUDIES ON REPLAY SPEECH DETECTION

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

A. Database description

The ASVspoof 2017 corpus is collected using 26 playback devices and 25 recording devices in 26 different environments [12], [13]. It was originally released for the ASVspoof 2017 challenge [11]. However, the organizers found some zero-value samples and silence in ASVspoof 2017 corpus

TABLE V
ASVSPOOF 2017 V2 DATABASE SPECIFICATIONS.

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

TABLE VI
EXPERIMENTAL RESULTS (EER(%)) ON ASVSPOOF 2017 V2 DEVELOPMENT SET USING DIFFERENT FEATURE COMBINATIONS OF ECQCC DYNAMIC FEATURES UNDER DIFFERENT SDN, RESPECTIVELY.

SDN	Feature combinations		
	D	A	DA
26	18.53	14.91	16.35
40	17.23	13.90	17.59
60	17.26	13.43	13.97

that can affect the result of playback detection. In 2018, the organizers updated ASVspoof 2017 by removing those zero-value samples and silence, and named the corrected version as ASVspoof 2017 V2 [13]. This database is constituted by three subsets: training, development and evaluation set. Table V summarizes the composition of the ASVspoof 2017 V2 database.

B. Evaluation protocol

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 metrics.

C. Experiment setup

The experimental setup for replay attack based studies follows the same that is considered for synthetic speech detection.

D. ASVspoof 2017 V2 development set: Results and analysis

Table VI shows the experimental results on ASVspoof 2017 development set using different feature combinations of eCQCC dynamic features. The table infers to the following: (1) For all SDN setups, eCQCC-A always gives the best performance followed by eCQCC-DA. (2) When SDN equals 60, the EER of eCQCC-A reaches minimum. In conclusion, when SDN equals 60, eCQCC-A and eCQCC-DA can be used as features to evaluate ASVspoof 2017 V2 evaluation set.

E. ASVspoof 2017 V2 evaluation set: Results and analysis

Fig. 2 demonstrates the experimental results on ASVspoof 2017 V2 evaluation set using eCQCC-A and eCQCC-DA when SDN equals 60. It can be seen 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 replay and playback devices along with the environments used are very different for evaluation set than that used in development set.

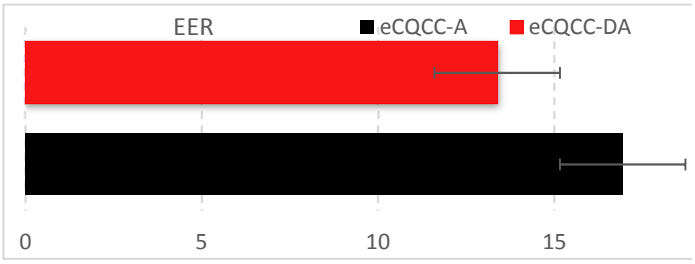

Fig. 2. Experimental results (EER(%)) on ASVspoof 2017 V2 evaluation set using eCQCC-A and eCQCC-DA.

TABLE VII
COMPARISON OF ECQQCC-DA WITH CQC-DA AND CQCC-DA ON ASVSPOOF 2017 V2 EVALUATION SET IN TERMS OF EER(%).

Feature	Power spectrum	EER
CQC-DA	Octave	18.73
CQCC-DA	Linear	15.46
eCQCC-DA	Octave and linear	13.38

F. Features based on different power spectra: A comparison

In this subsection, the performance of CQC-DA, CQCC-DA and eCQCC-DA is compared to observe if it follows similar to that obtained for synthetic speech detection. Table VII shows the comparison of eCQCC-DA with CQC-DA and CQCC-DA ASVspoof 2017 V2 evaluation set, in which the SDN of CQC, CQCC and eCQCC are 30, 30 and 60, respectively. Further, CQC-DA and CQCC-DA based systems have their individual DNN classifiers for ASVspoof 2017 V2 evaluation set. Their training methods are the same as that of eCQCC based DNN classifiers for ASVspoof 2017 V2 evaluation set.

From Table VII, it can be seen that eCQCC-DA 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. Additionally, 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 spoof detection similar to the case of synthetic speech detection. Finally, the studies under both the databases for synthetic and replay attacks confirms the importance of having a feature representation in terms of eCQCC feature for improved detection of spoofing attacks.

V. CONCLUSIONS

This work proposes a new feature referred to as eCQCC, which is obtained by using both linear power spectrum and octave power spectrum. The conventional CQCC features extracted only using linear power spectrum are found to dominate in the field of spoof detection. The proposed eCQCC feature is hypothesized to carry additional information due to use of coefficients extracted using octave power spectrum along with that obtained from linear power spectrum. The

studies are conducted on both synthetic and replay attack based databases ASVspoof 2015 and ASVspoof 2017 V2, respectively. The experiments depict that the eCQCC feature is able to have better discriminative ability for detection of spoofing attacks than the original CQCC features. This shows the complementary and useful information carried by the octave power spectrum. The future work will focus on combining the information from octave and linear power spectrum in a more effective manner to have the maximum benefit out of it.

ACKNOWLEDGMENT

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).

REFERENCES

[1] Zhizheng Wu, Phillip L. De Leon, Cenk Demiroglu, Ali Khodabakhsh, Simon King, Zhen-Hua Ling, Daisuke Saito, Bryan Stewart, Tomoki Toda, Mirjam Wester and Junichi Yamagishi, "Anti-spoofing for text-independent speaker verification: an initial database,comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol.20, no.8, pp.768–783, 2016.

[2] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Md Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco and Héctor Delado, "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing,* vol.11, no.4, pp.588–604, 2017.

[3] Junichi Yamagishi, Tomi Kinnunen, Nicolas Evans and Phillip L. De Leon, "Introduction to the issues on spoofing and countermeasures for automatic speaker verification," *IEEE Journal of Selected Topics in Signal Processing,* vol.11, no.4, pp.585–587, 2017.

[4] Xiaohai Tian, Siu Wa Lee, Zhizheng Wu, Eng Siong Chng and Haizhou Li, "An example-based approach to frequency warping for voice conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol.25, no.10, pp.1863–1876, 2017.

[5] Anupama Paul, Rohan Kumar Das, Rohit Sinha and S. R. Mahadeva Prasanna, "Countermeasure to handle replay attacks in pratical speaker verification systems," *International Conference on Signal processing and Communications (SPCOM) 2016,* pp.1–5, 2016.

[6] Sarfaraz Jelil, Rohan Kumar Das, S. R. Mahadeva Prasanna, and Rohit Sinha, Spoof detection using source, instantaneous frequency and cepstral features, *Annual Conference of the International Speech Communication Association (INTERSPEECH),* pp. 22–26, 2017.

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

[8] Rohan Kuman Das and Haizhou Li, "Instantaneous Phase and Excitation Source Features for Detection of Replay Attacks," *Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference (ASC) 2018,* Honolulu, Hawaii, USA, November 2018.

[9] Rosa Gonzalez Hautamaki, Tomi Kinnunen, Ville Hautamaki, Timo Leino, Anne-maria Laukanen, "I-vector meet imitators: On vulnerability of speaker verification systems against voice mimicry," *Annual Conference of the International Speech Communication Association (INTERSPEECH),* pp.930–934, 2013.

[10] Rosa Gonzalez Hautamaki, Tomi Kinnunen, Ville Hautamaki, Timo Leino and Anne-maria Laukanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication,* vol.72, pp.13–31, 2015.

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

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

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

[14] Massimiliano Todisco, Héctor Delgado and Nicholas Evans, "A new feature for automatic speaker verification anti-spoofing:constant Q cepstral coefficients," *In speaker and language recognition workshop (ODYESSEY),* pp.283–290, 2016.

[15] Dipjyoti Paul, Monisankha Pal and Goutam Saha, "Spectral features for synthetic speech detection," *IEEE Journal of Selected Topics in Signal Processing,* vol.11, no.4, pp.605–617, 2017.

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

[17] Zhizheng Wu, Eng Siong Chng and Haizhou Li, "Detecting conveerted speech and natural speech for anti-spoofing attack in speaker recognition," *Annual Conference of the International Speech Communication Association (INTERSPEECH),* pp.1700–1703, 2012.

[18] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng and Haizhou Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," *Annual Conference of the International Speech Communication Association (INTERSPEECH),* pp.2052-2056, 2015.

[19] Zhuxin Chen, Zhifeng Xie, Weibin Zhang and Xiangmin Xu, "Resnet and model fusion for automatic spoofing detection," *Annual Conference of the International Speech Communication Association (INTERSPEECH),* pp.102–106, 2017.

[20] Zhe Ji, Zhi-Yi Li, Peng Li, Maobo An, Shengxiang Gao, Dan Wu and Faru Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in asvspoof2017," *Annual Conference of the International Speech Communication Association (INTERSPEECH),* pp.87–91, 2017.

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

[22] Frank Seide, Gang Li, Xie Chen and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* pp.24–29, 2011.

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

[24] Judith C. Brown, "An efficient algorithm for the calculation of a constant Q spectral transform," *Journal of Acoustical Society of America,* vol.92, no.5, pp.2698–2701, 1992.