IEICE TRANS. INF. & SYST., VOL.E103–D, NO.2 FEBRUARY 2020

LETTER

Constant-Q Deep Coefficients for Playback Attack Detection

Jichen YANG[†*a)], ***Nonmember***, **Longting XU**[††*b)], ***Member***, ***and*** **Bo REN**[†††], ***Nonmember***

SUMMARY Under the framework of traditional power spectrum based feature extraction, in order to extract more discriminative information for playback attack detection, this paper proposes a feature by making use of deep neural network to describe the nonlinear relationship between power spectrum and discriminative information. Namely, constant-Q deep coefficients (CQDC). It relies on constant-Q transform, deep neural network and discrete cosine transform. In which, constant-Q transform is used to convert signal from the time domain into the frequency domain because it is a long-term transform that can provide more frequency detail, deep neural network is used to extract more discriminative information to discriminate playback speech from genuine speech and discrete cosine transform is used to decorrelate among the feature dimensions. ASVspoof 2017 corpus version 2.0 is used to evaluate the performance of CQDC. The experimental results show that CQDC outperforms the existing power spectrum obtained from constant-Q transform based features, and equal error can reduce from 19.18% to 51.56%. In addition, we found that discriminative information of CQDC hides in all frequency bins, which is different from commonly used features.

key words: *playback attack detection, log power spectrum, octave power spectrum, linear power spectrum, constant-Q transform*

1. Introduction

There can be three general types of spoofing attacks when deploying an automatic speaker verification (ASV) system. These are speech synthesis [1], voice conversion [2] and playback attack [3]. In playback attack, a pre-recording of the actual voice of a legitimate client [3] is played back to the ASV system, making it difficult type of attack to detect. As a result, playback attack presents a serious threat to ASV. This motivates the need to develop dedicated spoofing countermeasures. This paper, is mainly focused on playback attack detection.

To date, the most popular feature used in playback attack detection is constant-Q cepstral coefficients (CQCC) [4], such as in [5]. In addition, extended constant-Q cepstral coefficients (eCQCC) [6] was also used.

Manuscript received June 2, 2019.
Manuscript revised September 10, 2019.
Manuscript publicized November 14, 2019.

[†]The author is with Department of Electrical and Computer Engineering, National University of Singapore, Singapore.
[††]The author is with the College of Information Science and Technology, Donghua University, China.
[†††]The author is with Microsoft Search Technology Center Asia, Suzhou, China.
[*]Jichen Yang and Longting Xu have the same contributions for this work and they are the joint first authors.
a) E-mail: NisonYoung@163.com
b) E-mail: xlt@dhu.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2019EDL8115

The method of traditional power spectrum based feature extraction usually applies discrete cosine transform (DCT) on log power spectrum (LPS) and then selects the top ranked coefficients as final feature. However, it's difficult to design rules to extract more discriminative information from LPS for playback attack detection.

Note that deep neural network (DNN) has feature learning ability and can extract the information that handcrafted design can't do [7]. In order to extract more discriminative information from LPS for playback attack detection, DNN is proposed to describe the nonlinear relationship between LPS and discriminative information to discriminate playback speech from genuine speech in this work. In addition, constant-Q transform (CQT) rather than discrete Fourier transform (DFT) is selected to convert speech form the time domain into the frequency domain because CQT is a long-term transform and it can provide more frequency details than DFT.

As discussed above, a new feature is proposed here, which is obtained by combining CQT, DNN and DCT. In which, CQT is used to supply the base to obtain LPS, DNN is used to generate more discriminative information from LPS for playback attack detection and DCT is used to decorrelate among the feature dimensions. We name the feature as constant-Q deep coefficients (CQDC).

The remainder of the paper is organized as follows. Section 2 introduces previous works about power spectrum based feature extraction and Sect. 3 introduces how to extract CQDC. Section 4 gives the experimental results and corresponding analysis, which is based on ASVspoof 2017 version 2.0. Finally, Sect. 5 concludes the paper.

2. Previous Works

In this section, previous works on how to extract features from LPS obtained from CQT will be introduced. To date, there are three methods about how to extract feature from LPS of CQT, they are CQC [6], CQCC [4] and eCQCC [6]. Next, we will introduce them simply.

Figure 1 (a) and Fig. 1 (b) give the diagram of CQC and CQCC extraction, respectively. From Fig. 1 (a), it can be seen that there are four modules in CQC extraction. They are CQT, power spectrum, log and DCT, in which CQT is used to convert signal from the time domain into the frequency domain, followed by the power spectrum at the based of CQT, next LPS can be obtained, finally, DCT is used to decorate among the feature dimensions and CQC

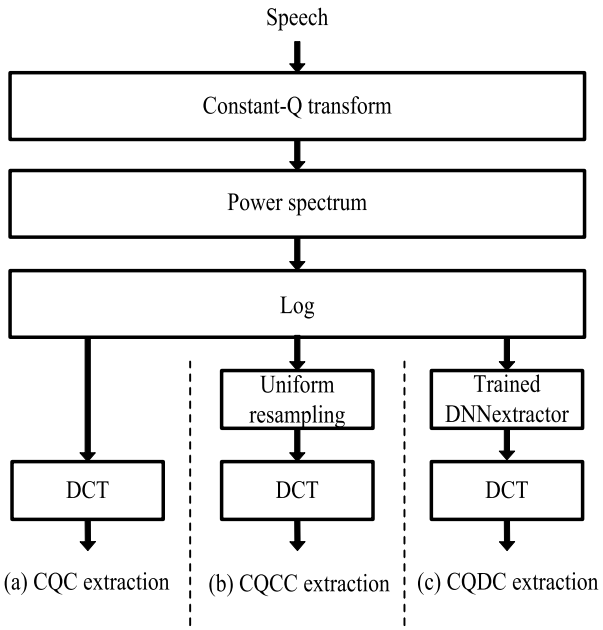

Fig. 1 Schematic diagram of LPS based feature extraction, which includes (a) diagram of CQC extraction, (b) diagram of CQCC extraction and (c) diagram of CQDC extraction.

can be obtained by selecting the first few top ranked coefficients.

Because CQT has geometrically frequency bins, the LPS directly obtained from CQT can be named as octave LPS and CQC can be regarded as feature extraction from octave LPS. Different from CQC, CQCC is extracted from linear LPS, which is shown in Fig. 1 (b). Compared Fig. 1 (a) with Fig. 1 (b), it can be seen that the difference between CQC and CQCC is the module of uniform resampling, which plays the role of converting octave LPS into linear LPS.

As introduced above, we know that CQC is extracted from octave LPS and CQCC is extracted from linear LPS. Octave LPS and linear LPS have different characteristic, for example, octave LPS can reflect some characteristic of human auditory system while linear LPS doesn't have such characteristic. So the information in CQC and CQCC can be complementary with each other. Based on this, eCQCC is proposed by concatenating CQC and CQCC [6].

3. Constant-Q Deep Coefficients Extraction

This section describes how to extract CQDC. Figure 1 (c) is the block diagram of CQDC extraction. From Fig. 1 (c), it can be seen that CQDC extraction contains five modules, namely, CQT, power spectrum, log, trained DNNextractor and DCT. CQT first transforms speech from the time domain into the frequency domain. The power spectrum value is then computed at the power spectrum stage. After octave LPS, trained DNNextractor is then performed to extract discriminative information from octave LPS. Finally, DCT is used to decorrelate among the feature dimensions.

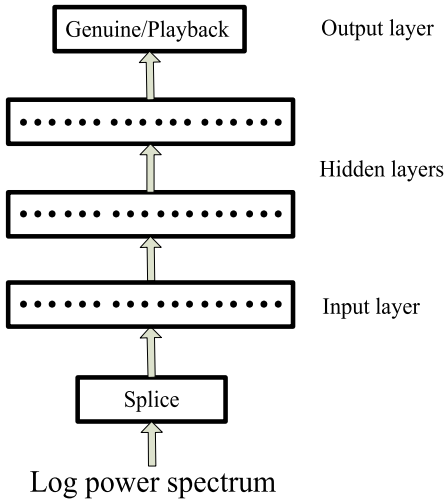

Fig. 2 Schematic diagram of genuine/playback classifier based on deep neural network.

Compared CQDC extraction with CQC and CQCC extraction in Fig. 1, it can be seen that CQC, CQCC and CQDC are all extracted from LPS. In addition, DCT is used in their extraction. However, there are several differences among them:

- CQC and CQDC are extracted from octave LPS, while CQCC is extracted from linear LPS that generated by using the module of uniform resampling.
- Though CQC and CQDC are extracted from octave LPS, CQC is obtained by applying DCT on octave LPS while CQDC is obtained from a trained DNNextractor plus DCT.
- The module of trained DNNextractor is the most difference between CQDC and CQC (CQCC).

Next, how to train DNNextractor will be introduced.

Note that DNN has good feature learning character, which not only eliminates the complex process of hand-crafted feature design but also has the potential of extracting the information that are impossible to extract for hand-crafted design [7]. So DNN can be used here to extract more discriminative information from LPS to discriminate playback speech from genuine speech. That's to say, DNN can play the role of describing the nonlinear relationship between power spectrum and discriminative information. As for CQDC extraction in Fig. 1 (c), the module of DNNextractor plays the role of extracting more discriminative information.

In our work, DNNextractor is obtained from a trained classifier that is based on DNN. In other words, the DNNextractor training consists of two steps: classifier based on DNN training and converting classifier to DNNextractor. The details are as following:

Firstly, a genuine/playback speech classifier based on DNN is trained, which is shown in Fig. 2. From Fig. 2, it can be observed that the classifier consists of input layers, two hidden layers and output layer. In which, the input is

eleven spliced frames centered by current frame, the output is the label of the current frame corresponding to.

Secondly, note that for the DNN, the higher layer, the more semantic. As for the trained DNN, there are more discriminative information in the second hidden layer than the first hidden layer. In other words, the second hidden layer has more semantic function and more discriminative information can be obtained from the layer. So DNNextractor is obtained by removing the output layer of the trained DNN.

4. Experiments and Evaluation

In this section, CQDC is evaluated using ASVspoof 2017 corpus version 2.0 (ASVspoof 2017 V2).

4.1 Database Introduction

ASVspoof 2017 V2 [8] released in 2018, which is constituted by three subsets: training set, development and evaluation set, Table 1 gives some details of ASVspoof 2017 V2.

4.2 Evaluation Rule and Experiment Setup

According to ASVspoof 2017 challenge rule, equal error rate (EER) is used as evaluation metric. In addition, all the parameters in CQT are the same as [4]. The static dimension of CQDC is set to 1024. Since our previous works [6], [9], [12] have shown dynamic features can give better performance and static features will degrade the performance in playback attack detection, so only dynamic features combinations of CQDC are taken into account. Further, D and A stand for delta and acceleration, respectively.

The Computational Network Toolkit (CNTK) [10] is used to train DNNs, which are used as classifier in our experiment. In our experiments, there are two types classifier, one is used to obtain DNNextractor and the other is used for CQDC classifier. For the first type classifier, it has one input layer, two hidden layers and one output layer and the nodes of input layer, 1st hidden layer, 2nd hidden layer and output layers are 9493, 863, 1024 and 2, respectively. After the classifier training is finished, the DNNextractor can be obtained by removing the output layer. For 863-dimension LPS is given to the DNNextractor as the input, 1024-dimension CQDC can be obtained. For the second types of classifier, three four-layer DNNs are trained, which have two hidden layers with 512 nodes at every layer and an output layer with 2 nodes and an input layer constituted by an 11-frame context window of the input feature vector. The feature combinations for CQDC require different input layers. For example, for CQDC-D and CQDC-A, the input layer consisted of 1024×11 nodes inclusive of five frames each on the left and right whereas for CQDC-DA, the input layer was 2048×11.

In our DNNs training, sigmoid network is used for the hidden layers training 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. Totally, 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. A momentum value is set as 0.9.

4.3 Experiment Results and Analysis

Table 2 gives the experimental results on development and evaluation set using dynamic feature combinations of CQDC in terms of EER.

From the performance of CQDC on development set in Table 2, it can be seen that CQDC-DA and CQDC-D can give the first and the second best performance on development set and then CQDC-A performs the worst.

From the performance of CQDC on evaluation set in Table 2, it can be found that CQDC-D and CQDC-DA performs much better than CQDC-A on evaluation set. In addition, CQDC-D gives a little better performance than CQDC-DA on evaluation set. Which is different from that the performance of CQDC-D is worse than CQDC-DA on development set. The reason may be that there are some types playback speech only appear in evaluation set.

4.4 Comparison with Different Dimensions

In traditional features extraction for speech recognition and speaker recognition, 13, 20 and 30 are often selected as static dimension. In our work, CQDC is a high dimension feature, so not only traditional low dimensions such as 13, 20 and 30 but also high dimensions such as 512 and 863 are selected. Table 3 shows the experimental result on evaluation set using CQDC-D under different static dimensions of CQDC in terms of EER.

From Table 3, it can be seen that when static dimension equals 13, 20 and 30, the performance of CQDC-D is very

Table 1 Some details of ASVspoof 2017 V2.

Subset	Num			
	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 2 Experimental results in EER (%) on ASVspoof 2017 V2 development and evaluation set using dynamic features of CQDC.

Feature	Development set			Evaluation set		
	Feature combinations			Feature combinations		
	D	A	DA	D	A	DA
CQDC	15.25	18.50	**14.38**	**9.44**	11.86	9.53

Table 3 Experimental results (EER (%)) on ASVspoof 2017 V2 evaluation set using CQDC-D under different dimensions.

Static dimension	EER	Static dimension	EER
13	20.13	20	19.20
30	19.75	512	20.37
863	21.96	1024	9.44

Table 4 Experimental results (EER (%)) on ASVspoof 2017 V2 evaluation set comparison with some power spectrum based features with different feature combinations.

Power spectrum	Feature	Feature combinations		
		D	A	DA
Octave	CQC	19.49	19.50	18.73
Linear	CQCC	15.01	15.65	15.46
Octave and linear	eCQCC	13.87	16.94	13.38
Octave	CQDC	9.44	11.86	9.53

worse. It means that there is much difference for DCT playing role between CQDC and traditional feature extraction. When static dimension increases to 512 and 863, the performance is still worse. When static dimension equals 1024, in other words, all the frequency bins are used, CQDC-D performs best. It means that there is much difference between discriminative information in CQC (CQCC) and CQDC. There is less discriminative information in the high dimensions of CQC (CQCC), so only a few top ranked coefficients are selected as final features. However, the discriminative information hides in all frequency bins of CQDC, so all the coefficients must be selected as final features.

4.5 Comparison with Power Spectrum Based Features

Table 4 gives the experimental results on evaluation set comparison with some power spectrum based features with different dynamic feature combinations in terms of EER.

From Table 4, several conclusions can be obtained: 1) For feature combinations D, comparing with CQC, CQCC and eCQCC, EER can reduce by 51.56%, 37.11% and 31.94%, respectively. 2) For feature combinations A, comparing with CQC, CQCC and eCQCC, EER can reduce by 19.18%, 24.22% and 29.99%, respectively. 3) For feature combinations DA, comparing with CQC, CQCC and eCQCC, EER can reduce by 49.12%, 38.36% and 28.77%, respectively. 4) In conclusion, the proposed feature based on power spectrum can outperform better than traditional features based on power spectrum, EER can reduce from from 19.18% to 51.56%, which can confirm that our proposed idea is correct.

4.6 Comparison with Some Known Systems

Table 5 compares the performance of our proposed CQDC-D based system with that of existing systems on the evaluation set. In which, CQCCE represents a combination feature by combining CQCC and log energy [8], qDFTspec represents DFT spectrum in q-log domian [11], CMPOC and CQSPIC represent constant-Q magnitude-phase octave coefficients [9] and constant-Q statistics-plus-principal information coefficients [12], respectively. In addition, MFCC represents mel-frequency cepstral coefficients, LFS and MFS represent linear filterbak slope and mel filterbank slope [13], respectively.

From Table 5, it can be seen that CQDC-D based system far outperforms most of existing systems for playback attack detection. However, the performance of our system based on CQDC-D is worse than the systems based on MFCC|LFS|MFS and CQCC|MFCC|LFS|MFS [13]. The reason is that the systems based on MFCC|LFS|MFS and CQCC|MFCC|LFS|MFS are based on decision-level feature switching. In such systems, every feature and its corresponding GMM are selected by a switching method. In other words, such systems are hybrid systems while our system is only a single system.

Table 5 Comparison of CQDC-D based system against some existing systems on the ASVspoof 2017 V2 evaluation set.

Feature	Classifier	EER
CQCCE [8]	GMM	12.24
qDFTspec [11]	GMM	11.19
CMPOC-D [9]	DNN	14.93
eCQCC-DA [6]	DNN	13.38
CQSPIC-DA [12]	DNN	10.45
MFCC|LFS|MFS [13]	GMM	6.23
CQCC|MFCC|LFS|MFS [13]	GMM	6.60
CQDC-D	DNN	9.44

5. Conclusion

In this paper, in order to extract more discriminative information from power spectrum for playback attack detection, CQDC is proposed by means of DNN to describe the nonlinear relationship between power spectrum and discriminative information. Our experimental results show that the proposed CQDC can achieve far better performance on playback attack detection than traditional power spectrum based features. In addition, we found that discriminative information of CQDC hides in all the frequency bins.

Acknowledgments

The work was supported by Shanghai Sailing Program (No.19YF1402000) and the Fundamental Research Funds for the Central Universities (No.2232019D3-52).

References

[1] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statical parametric speech synthesis," IEEE Trans. Audio, Speech, Language Process., vol.21, no.3, pp.587–597, 2013.
[2] X. Tian, S.W. Lee, Z. Wu, E.S. Chng, and H. Li, "An example-based approach to frequency warping for voice conversation," IEEE/ACM Trans. Audio, Speech, Language Process., vol.25, no.10, pp.1863–1876, 2017.
[3] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.A. 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.
[4] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients," Speaker and Language Recognition Workshop (ODYSSEY), pp.283–290, 2016.
[5] M. Kamble and H.A. Patil, "Novel variable length energy separation

algorithm using instantaneous amplitude features for replay detection," 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.646–650, 2018.

[6] J. Yang, R.K. Das, and H. Li, "Extended constant-Q cepstral coefficients for detection of spoofing attacks," Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.1024–1029, 2018.

[7] F. Seide, G. Li, X. Chen, and D. 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.

[8] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K.A. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements," Speaker and Language Recognition Workshop (ODYSSEY), pp.296–303, 2018.

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

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

[11] M.J. Alam, G. Bhattacharya, and P. Kenny, "Boosting the performance of spoofing detection systems of replay attacks using Q-logarithm domain feature normalization," Speaker and Language Recognition Workshop (ODYSSEY), pp.393–398, 2018.

[12] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.651–655, 2018.

[13] S.M. S. and H. Murthy, "Decision-level feature switching as a paradigm for replay attack detection," 19th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.686–690, 2018.