International Journal of Digital Crime and Forensics
Volume 11 • Issue 2 • April-June 2019

CBC-Based Synthetic Speech Detection

Jichen Yang, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Qianhua He, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Yongjian Hu, School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

Weiqiang Pan, Information and Network Engineering and Research Centre, South China University of Technology, Guangzhou, China

ABSTRACT

In previous studies of synthetic speech detection (SSD), the most widely used features are based on a linear power spectrum. Different from conventional methods, this article proposes a new feature extraction method for SSD from octave power spectrum which is obtained from constant-Q transform (CQT). By combining CQT, block transform (BT) and discrete cosine transform (DCT), a new feature is obtained, namely, constant-Q block coefficients (CBC). In which, CQT is used to transform speech from the time domain into the frequency domain, BT is used to segment octave power spectrum into many blocks and DCT is used to extract principal information of every block. The experimental results on ASVspoof 2015 corpus shows that CBC is superior to other front-ends features that have been benchmarked on ASVspoof 2015 evaluation set in terms of equal error rate (EER).

KEYWORDS

Block Transform, CBC, Constant-Q Transform, DCT, Synthetic Speech Detection

INTRODUCTION

Automatic speaker verification (ASV) is the task to accept or reject an identity claim based on a person's speech sample (Kinnunen & Li, 2008), which has received wide spread attention over the recent 30 years. Most ASV systems assume natural human speech as input. However, ASV systems are often attacked by synthetic speech (Wu, et al, 2016), which is usually obtained by speech synthesis (SS) and voice conversation (VC) (Wu & Li, 2014). In order to protect ASV systems safe, it is necessary to detect synthetic speech from input speech. In addition, in the field of criminal investigators for forensics, SSD is helpful.

Generally speaking, there are two types of countermeasures for SSD: front-end feature and back-end model.

In terms of feature, features based on power spectrum, combining magnitude with phase and so on. The most widely used features based on power spectrum in SSD are mel-frequency cepstral coefficients (MFCC) (Sahidullah, Kinnunen & Hanilci, 2015) and constant-Q cepstral coefficients (CQCC) (Todisco, Delgado & Evans, 2016). In 2017, Paul et al. proposed several types of transformation for SSD in (Paul, Pal & Saha, 2017), they are speech-signal frequency cepstral coefficients (SFCC), mel-warped overlapped block transformation (MOBT), speech-signal-based overlapped block transformation (SOBT), inverted speech-signal frequency cepstral coefficients (ISFCC), inverted mel-warped overlapped block transformation (IMOBT). In addition, inverted mel frequency cepstral

DOI: 10.4018/IJDCF.2019040105

coefficients (IMFCC) (Chakroborty, Roy & Saha, 2007) is also used in (Sahidullah, Kinnunen & Hanilci, 2015). However, those features are all based on linear power spectrum that every frequency bin has the same frequency region.

Phase features were often combined with magnitude features in SSD because the performance of phase features is usually worse than commonly used features based on power spectrum. For example, In 2015, Xiao et al. used logarithm magnitude spectrum (LMS) + residual logarithm magnitude spectrum (RLMS) + group delay (GD) + modified group delay (MGD) + instantaneous frequency (IF) + baseband phase difference (BPD) + pitch synchronous phase (PSP) in (Xiao, Tian, Du, et al, 2015), Novoselov et al. used modified group delay cepstral coefficients (MGDCC) + MFCC + Mel-frequency principal coefficients (MFPC) in (Novoselov, Kozlov, et al, 2016).

In addition, there are some other features used in SSD. For example, Zhang et al. employed Teager energy operator critical band autocorrelation envelope plus perceptual minimum variance distortionless response (TCAEP) and spectrogram in SSD (Zhang, Ranjan, Nandwana, et al, 2016, Zhang,Yu, & Hansen, 2017). Sriskandaraja et al. proposed scattering cepstal coefficients (SCC) (Sriskandaraja, Sethu, Ambikairajah & Li, 2017) in SSD, respectively. Patel and Patil proposed to use fundamental frequency, strength of excitation and cochlear filter cepatral coefficients and instantaneous frequency (CFCC-IF) (Patel & Patil, 2015, Patel & Patil, 2016) in SSD. In (Sahidullah, Kinnunen & Hanilci, 2015), a series of features were compared in SSD by Md Sahidullah et al. They are rectangular filter cepstral coefficients (RFCC) (Hasen, Sadjadi, Liu, Shokouhi, Boril, & Hansen, 2013), linear frequency cepstral coefficients (LFCC) (Alegre, Amehraye, & Evans, 2013), linear prediction cepstral coefficients (LPCC) (Furui, 1981), perception linear prediction cepstral coefficients (PLPCC) (Hermansky, 1990), subband spectral fux coefficients (SSFC) (Scheirer & Slaney, 1997), spectral centroid magnitude coefficients (SCMC) (Kua, Thiruvaran, Nosratighods, Ambikairajah, Epps, 2010), subband centroid frequency coefficients (SCFC) (Kua, Thiruvaran, Nosratighods, Ambikairajah, Epps, 2010).

Gaussian mixture model (GMM), support virtual machine (SVM) and deep learning based are often used as classifier in SSD. In the previous study, GMM is the most widely used model in SSD. For example, CQCC is followed by GMM in (Todisco, Delgado & Evans, 2016), in (Sriskandaraja, Sethu, Ambikairajah & Li, 2017), the authors also used GMM to model SCC. In addition, (Hasen, Sadjadi, Liu, Shokouhi, Boril, & Hansen, 2013, Todisco, Delgado & Evans, 2016, Patel & Patil, 2016) all used GMM as classifier. Some classifiers based on deep learning were employed in SSD, for example, multilayer perceptron neural network (MLPNN) was used in (Xiao, Tian, Du, et al, 2015). In addition, convolutional neural network (CNN) plus recurrent neural network (RNN) were used in (Zhang, Yu, & Hansen, 2017), in which CNN was used to learn feature and RNN was used as classifier.

Since the work in (Todisco, Delgado & Evans, 2016) has shown that more gains can be obtained from feature rather than model. So we only focus feature level in this work.

As mentioned above, there is no report about how to extract discriminative information from octave power spectrum in previous SSD study, in which full frequency-band is segmented into several octaves and different frequency bins have different frequency region in the same or different octaves. We want to explore a new feature from octave power spectrum in this study. In addition, in order to capture more discriminative information from octave power spectrum, block transform (BT) is firstly used to segment logarithm octave power spectrum (LOPS) into many blocks and discrete cosine transform (DCT) is applied to extract principal information of every block, which is main contribution of the work.

In addition, because constant-Q transform (CQT) (Youngberg & Boll, 1978) has high frequency resolution and it can capture more detail of frequency, so CQT is combined with BT and DCT to extract a discriminative feature for SSD. We name the new feature as constant-Q block coefficients (CBC).

Because DNN not only has classifier function but also has feature learning ability (Seide, Li, et al, 2011), in order to make CBC detect synthetic speech better, DNN is utilized as classifier in this work.

The remainder of the paper is organized as follows. CBC extraction is introduced in Section 2. Section 3 gives the experimental results and analysis of ASVspoof 2015 using CBC. Section 4 gives the conclusion.

CBC EXTRACTION

In this section, how to extract CBC is introduced. An illustrative block diagram of the proposed feature is shown in Figure 1. From Figure 1, we can see that there are five modules in CBC extraction: CQT, power spectrum, Log, BT and DCT. In which, the module of CQT is used to transform speech from the time domain into the frequency domain, power spectrum is used to obtained octave power spectrum, log is used to obtain LOPS, BT is used to segment LOPS into many blocks and DCT is used to extract principal information of every block. A detailed description of each module is as following.

Constant-Q Transform

CQT was proposed in (Youngberg & Boll, 1978) and (Brown,1991). In which, Q is defined as the ratio of center frequency to bandwidth, which is as following:

$$\mathrm{Q} = \frac{f_k}{\delta f} \tag{1}$$

where f_k is center frequency, k represents $k\text{-}th$ frequency bin and δf is the bandwidth. f_k is defined as:

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

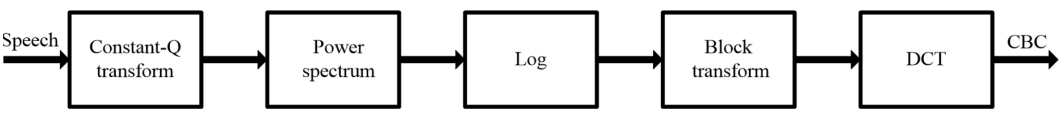
where f_1 is the center frequency of the lowest-frequency bin, B is the number of bins of per octave. Then δf can be obtained:

$$\delta f = f_{k+1} - f_k = f_1 2^{\frac{k}{B}} - f_1 2^{\frac{k-1}{B}} = f_1 2^{\frac{k-1}{B}} \left(2^{\frac{1}{B}} - 1 \right) \tag{3}$$

For a discrete time domain signal $x(n)$, its CQT $Y(k,n)$ can be calculated as:

$$Y(k,n) = \sum_{\mathrm{j}=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j - n - N_k / 2) \tag{4}$$

Figure 1. Schematic diagram of CBC extraction



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

$$a_k(n)=\frac{1}{\mathrm{C}}\omega(\frac{n}{N_k})\exp\left[i(2\pi n\frac{f_k}{f_s}+\phi_k)\right] \tag{5}$$

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

$$\mathrm{C}=\sum_{n_1=-\lfloor N_k/2\rfloor}^{\lfloor N_k/2\rfloor}\omega\left(\frac{n_1+N_k/2}{N_k}\right) \tag{6}$$

From Formula (2) and (3), we can see that the more k, the more f_k and δf.

Power Spectrum and Log

The module of power spectrum is used to obtained octave power spectrum. Now we introduce how to calculate octave power spectrum:

$$M_Y=\{ms_1,ms_2,...,ms_{K-1},ms_K\} \tag{7}$$

where M_Y represents for magnitude spectrum of $Y(k,n)$, ms_k represents for $k\text{-}th$ frequency bin and k is from 1 to K:

$$O_Y=\{os_1,os_2,...,os_{K-1},os_K\} \tag{8}$$

where O_Y represents for octave power spectrum of $Y(k,n)$, os_k represents for $k\text{-}th$ frequency bin and k is from 1 to K. In addition:

$$os_k=ms_k\ \ k=1,2,...,K \tag{9}$$

The module of Log is used to obtain octave power spectrum in log-scale, which is as following:

$$LO_Y=\left\{\log(os_1),\log(os_2),...,\log(os_{K-1}),\log(os_K)\right\} \tag{10}$$

where LO_Y represents for O_Y in log-scale and $\log(\cdot)$ is used to calculate value in log-scale.

Block Transform and Discrete Cosine Transform

In order to extract more discriminative information from LOPS, BT is used here to segment LOPS, then LOPS is segmented into many blocks by BT. In order to make blocks continuous, there is overlapping frequency bins between neighboring blocks.

After LOPS is segmented into blocks, DCT is applied on every block to extract principal information of every block, then the former R coefficients obtained from DCT are selected as the feature for every block, finally, the R coefficients obtained from every DCT is concatenated to form final feature. Figure 2 shows the detail of back-step feature extraction by using BT for LOPS, DCT for blocks and concatenation.

From Figure 2, we can see that LOPS is segmented into many blocks using BT, which is as following:

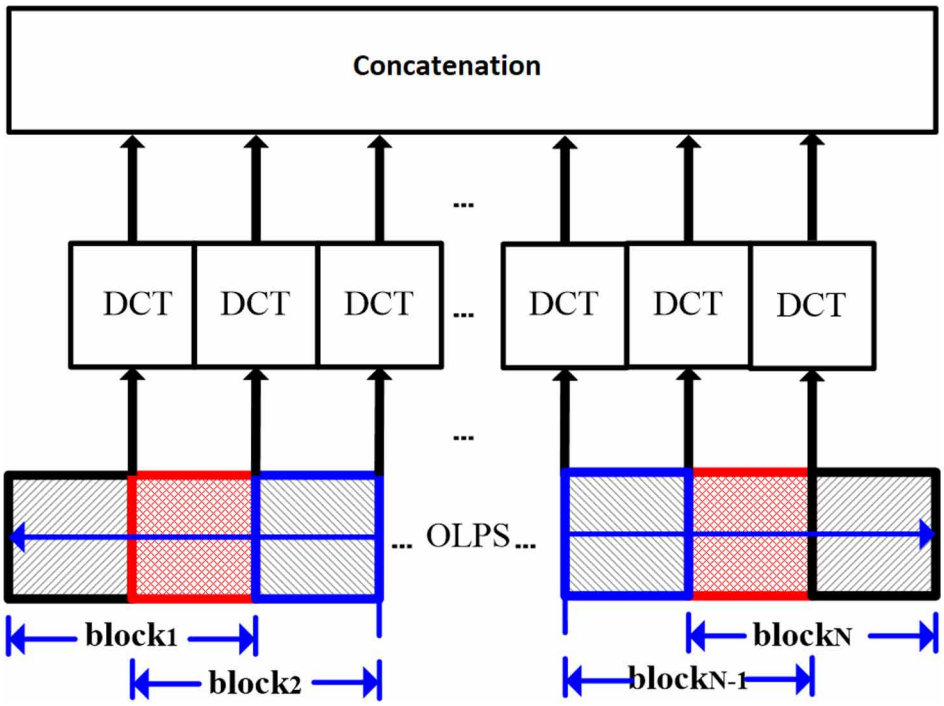
$$LOPS=\left[Block_1, Block_2, ..., Block_{N-1}, Block_N\right] \tag{11}$$

In addition, there is overlapping frequency bins (red part in Figure 2) between neighboring blocks. DCT is supplied on every block, final feature can be obtained by concatenating the former R coefficients obtained from every DCT.

Supposing the dimension number of every block is M , CBC can be calculated as following:

$$CBC(r)=\left[C_1(0), C_1(r), C_2(0), C_2(r), ..., C_{N-1}(0), C_{N-1}(r), C_N(0), C_N(r)\right] \tag{12}$$

Figure 2. BT for LOPS, DCT for every block and concatenation



where r represents $r\text{-}th$ DCT coefficients, and r is from 1 to R-1, R represents principal coefficients of DCT and:

$$
\begin{aligned}
C_1(0) &= \frac{1}{M}\sum_{m=0}^{M-1} Block_1(m) \\
C_1(r) &= \sqrt{\frac{2}{M}}\sum_{m=0}^{M-1} Block_1\left(m\right)\cos\left(\frac{(2m+1)r\pi}{2M}\right) \\
C_2(0) &= \frac{1}{\sqrt{M}}\sum_{m=0}^{M-1} Block_2(m) \\
C_2(r) &= \sqrt{\frac{2}{M}}\sum_{m=0}^{M-1} Block_2\left(m\right)\cos\left(\frac{(2m+1)r\pi}{2M}\right) \\
C_{N-1}(0) &= \frac{1}{\sqrt{M}}\sum_{m=0}^{M-1} Block_{N-1}(m) \\
C_{N-1}(r) &= \sqrt{\frac{2}{M}}\sum_{m=0}^{M-1} Block_{N-1}\left(m\right)\cos\left(\frac{(2m+1)r\pi}{2M}\right) \\
C_N(0) &= \frac{1}{\sqrt{M}}\sum_{m=0}^{M-1} Block_N(m) \\
C_N(r) &= \sqrt{\frac{2}{M}}\sum_{m=0}^{M-1} Block_N\left(m\right)\cos\left(\frac{(2m+1)r\pi}{2M}\right)
\end{aligned}
\tag{13}
$$

In addition, in order to study the role of the module of BT in feature extraction, we can suppose the module of BT is removed from Figure 1 and the obtained feature can named as constant-Q coefficients (CQC).

EXPERIMENTS AND ANALSIS

Experimental Data

ASVspoof 2015 corpus is constituted by three subsets: training data, development data and evaluation data, every part consists of human and spoofed speech that spoofed speech is generated from original genuine speech with different VC and SS algorithms. Table 1 gives some detail of every subset. In addition, all the data in ASVspoof 2015 is sampled at 16KHz and saved as mono channel wav formats. Totally, there are total ten spoofing-attack types speech (name as S1 to S10) in ASVspoof 2015 (Wu, et al., 2016). In addition, all the three subsets contain spoofing type S1-S5, which are denoted as known attack while S6 -S10 only appear in the evaluation subset and denoted as unknown attack.

Experiment

In CQT, there are several important parameters, which will affect the final performance. They are B which is number of bins in an octave, octave number, sampling period, gamma, respectively. In our experiments, the same as (Todisco, Delgado & Evans, 2016), B is set as 96, octave number is set as 9, sampling period is set as 16 and gamma is set as 3.3026.

After DCT on every block, the former 12-dimension coefficients are used as feature and the other dimension coefficients are discarded, in other words, R is set as 12 in the experiments. According to ASVspoof 2015 challenge rule, 16,375 utterances in training set is used to train model, which can be used to evaluate the performance of the proposed feature on ASVspoof 2015 development and evaluation set.

Table 1. The detail of ASVspoof 2015 corpus

Subset	Number			
	Male	Female	Genuine	Spoofed
Training	10	15	3,750	12,625
Development	15	20	3,497	48,875
Evaluation	20	26	9,404	184,000

A type of classifier with six-layer DNN is trained, which has four hidden layers with 512 nodes at every layer and output layer with 2 nodes.

In BT, overlapping length of block is set as a half of block length. In addition, equal error rate (EER) and average equal error rate (AEER) are used as evaluation metrics.

Since the work in (Todisco, Delgado & Evans, 2016) has shown static features degrade the performance, CBC-DA is used as feature to evaluate CBC performance on ASVspoof 2015 corpus. In which D and A stands for delta and acceleration, respectively.

Experimental Results on ASVspoof 2015 Development Set Using CBC and CQC

Table 2 gives the experimental results on ASVspoof 2015 development set using CBC-DA and CQC-DA under different block length and CQC.

From Table 2, several conclusions can be obtained:

1. For CBC-DA, when block length is 144, its performance is very bad, the reason may be that there is only a little discriminative information is extracted;
2. For CBC-DA, when block length is less than 144, its performance is satisfied, however, when block length is 84, its performance will degrade;
3. Compared with the performance of CQC-DA and CBC-DA on ASVspoof 2015 development set, it can be shown that CBC-DA performs much better than CQC-DA at most situations except when block length is 144, which means that more discriminative information is obtained by using BT for CBC-DA extraction.

Experiment Result on ASVspoof 2015 Evaluation Set Using CBC-DA and CQC-DA

In ASVspoof 2015 evaluation set, there are five known attacks (S1-S5) and five unknown attacks (S6-S10), AEER of the ten attacks can be used as evaluation metric. Figure 3 gives the relationship between block length and average equal error rate (AEER) on ASVspoof 2015 evaluation set using CBC-DA.

From Figure 3, it can be seen that: AEER increases continuously when block length decreased from 132 to 84, especially when block length decreases from 96 to 84, the trend is very vast, AEER reach its minimal when block length equals 132, which means that discriminative information obtained is less when block length declines from 132 to 84.

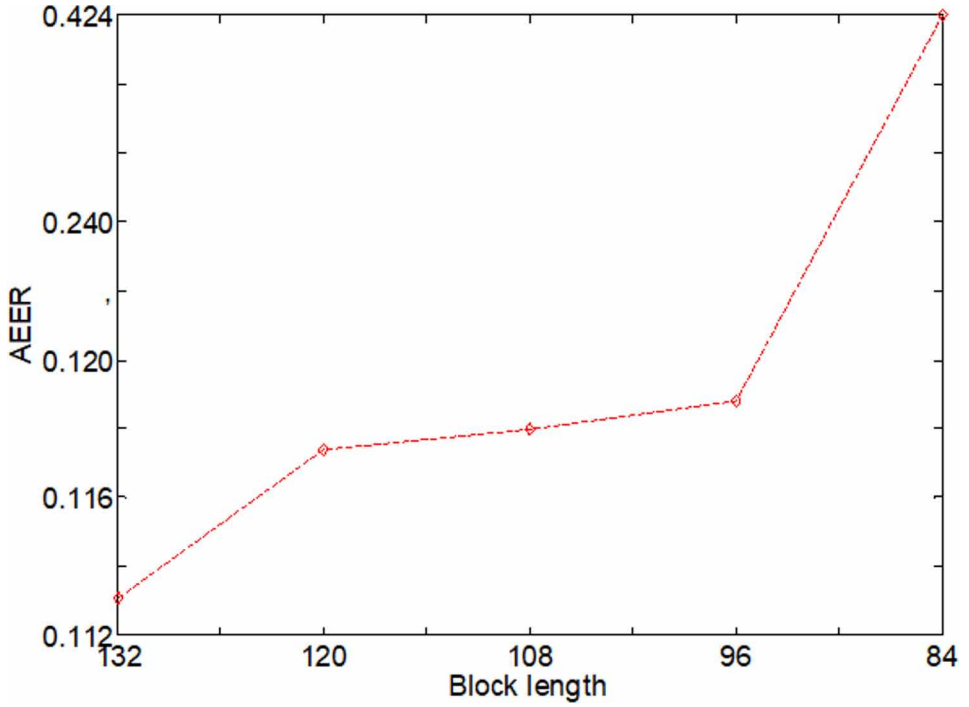
Figure 4 shows AEER comparison between CBC-DA and CQC-DA on ASVspoof 2015 evaluation set.

From Figure 4, it can be seen that the performance of CBC-DA is much better than CQC-DA on ASVspoof 2015 evaluation set, the reason may be that there is more discriminative information extracted by using BT in CBC extraction. It confirms that our idea is correct.

Table 2. Experimental result (EER(%)) on ASVspoof 2015 development set using CBC-DA under different block length and CQC-DA

Feature	Block Length	Overlapping Length	EER					
			S1	S2	S3	S4	S5	Ave.
CQC-DA			0.0090	0.0197	0	0.0148	0.2898	0.0667
CBC-DA	144	72	13.5548	29.5388	2.1397	1.9013	37.6657	16.9813
	132	66	0	0.0148	0	0	0.0474	0.0125
	120	60	0	0.0085	0	0	0.0539	0.0125
	108	54	0	0	0	0	0.0217	0.0043
	96	48	0	0	0	0	0.0236	0.0047
	84	42	0	0.0147	0	0	0.1580	0.0345

Figure 3. The relationship between AEER and block length on ASVspoof 2015 evaluation set using CBC-DA



Comparison With Some Other Systems

Table 3 gives the comparison with some other known systems on ASVspoof 2015 evaluation set in terms of EER.

From Table 3, two conclusions can be obtained: (1) In terms of average EER, for known spoofing type, IFCC-IF performs the worst, then MFCC and SSFC; for unknown spoofing type, magnitude + phase features give the worst performance, then SOBT, MOBT and ISOBT. (2) The system based on CBC-DA, outperformed all previously reported systems not only on known attack but also on unknown attack type. The reason may be that CQT can supply more frequency detail, BT can supply the base to obtain more discriminative information and DCT is used to extract information to form final feature.

Figure 4. AEER (%) comparison between CBC-DA and CQC-DA on ASVspoof2015 evaluation set

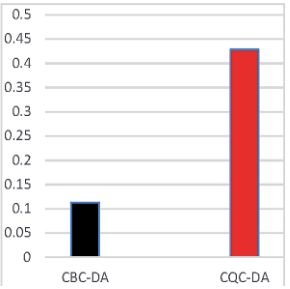


Table 3. Comparison with some other systems on ASVspoof 2015 evaluation set in terms of EER (%)

System	Known	Unknown	Ave.
CFCC-IF (Patel & Patil, 2015)	0.41	2.01	1.21
MGDCC+MFC+MFPC (Novoselov, Kozlov, et al, 2016)	0.00	3.92	1.97
Magnitude + phase features (Xiao, Tian, Du, et al, 2015)	0.00	5.22	2.61
RFCC (Sahidullah, Kinnunen & Hanilci, 2015)	0.12	1.92	1.02
LFCC (Sahidullah, Kinnunen & Hanilci, 2015)	0.11	1.67	0.89
MFCC (Sahidullah, Kinnunen & Hanilci, 2015)	0.39	3.84	2.12
IMFCC (Sahidullah, Kinnunen & Hanilci, 2015)	0.15	1.86	1.01
LPCC (Sahidullah, Kinnunen & Hanilci, 2015)	0.11	2.31	1.21
SSFC (Sahidullah, Kinnunen & Hanilci, 2015)	0.30	1.96	1.13
SCFC (Sahidullah, Kinnunen & Hanilci, 2015)	0.07	8.84	4.46
SCMC (Sahidullah, Kinnunen & Hanilci, 2015)	0.17	1.71	0.94
APGDF (Sahidullah, Kinnunen & Hanilci, 2015)	0.16	2.34	1.25
TCAEP (Zhang,Yu, & Hansen, 2017).	0.27	2.66	1.47
SFCC (Paul, Pal & Saha, 2017)	0.15	1.83	1.05
MOBT (Paul, Pal & Saha, 2017)	0.08	3.72	1.85
SOBT (Paul, Pal & Saha, 2017)	0.28	4.48	2.49
ISFCC (Paul, Pal & Saha, 2017)	0.05	1.63	0.86
IMOBT (Paul, Pal & Saha, 2017)	0.01	2.87	1.46
CQCC (Todisco, Delgado & Evans, 2016)	0.05	0.46	0.26
SCC (Sriskandaraja, Sethu, Ambikairajah & Li, 2017)	0.02	0.33	0.18
CBC-DA	0.01	0.22	0.11

CONCLUSION

In pursuit of capturing discriminative information from octave power spectrum and seeking an effective feature to detect synthetic speech, CBC, a new feature is proposed in this paper, which is based on CQT, BT and DCT. In which CQT is used to generate octave power spectrum, BT is used to supply the base to obtain more discriminative information and DCT is used to extract information to form final feature. Then the proposed feature is evaluated with ASVspoof 2015 corpus.

Experimental results showed that the proposed approach can achieve encouraging result. EER can reach 0.11%, which indicates that CBC-DA performs much better than some commonly used features and outperforms all previously reported systems not only on known attack but also on unknown attack type on ASVspoof 2015 evaluation set in terms of EER.

Though the result is encouraged, there is still room to improve. For example, 1) CBC is obtained from octave power spectrum obtained from CQT, there are too more frequency bins to describe low frequency bins and too fewer frequency bins to describe high frequency bins in octave power spectrum. If we can seek a method to transform octave power spectrum into linear power spectrum, there is gain that can be obtained because there are enough frequency bins to describe high frequency information. 2) After DCT on every block, only the former 12-dimension coefficients are selected as feature in our experiments, if much or less dimensions coefficients are selected as final features, how will affect the final performance. These can be our study directions in the near future.

ACKNOWLEDGMENT

This work was supported National Natural Science Foundation of China (NSFC) (61571192), Natural Science Foundation of Guangdong Province (2015A030313600) and the Science and Technology Planning Project of Guangdong Province (2017B010110009).

REFERENCES

Alegre, F., Amehraye, A., & Evans, N. (2013). A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *IEEE sixth International Conference on Biometrics: Theory, Applications and System (BATS)* (pp. 1–8). doi:10.1109/BTAS.2013.6712706

Wu, Z., De Leon, P. L., Demiroglu, C., Khodabakhsh, A., King, S., Ling, Z. H., ... & Yamagishi, J. (2016). 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*, *20*(8), 768–783.

Brown, J. (1991). Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, *89*(1), 425–434. doi:10.1121/1.400476

Chakroborty, S., Roy, A., & Saha, G. (2007). Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. *International Journal of Signal Processing*, *4*(2), 114–122.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(2), 254–272. doi:10.1109/TASSP.1981.1163530

Hasen, T., Sadjadi, S., Liu, G., Shokouhi, N., Boril, H., & Hansen, J. (2013). CRSS systems for 2012 NIST speaker recognition evaluations. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 6783–6787). doi:10.1109/ICASSP.2013.6638975

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. The Journal of the Acoustical Society of America, 87(4), 1738-1752. PMID:2341679

YoungbergJ. & Boll S. (1978). Constant-q signal analysis and synthesis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 375-378).

Kinnunen, T., & Li, H. (2008). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12–40. doi:10.1016/j.specom.2009.08.009

Kua, J., Thiruvaran, T., Nosratighods, M. E., Ambikairajah, E., & Epps, J. (2010). Investigation of spectral centroid magnitude and frequency for speaker recognition. In The speaker and language recognition workshop (ODYSSEY).

Novoselov, S., & Kozlov, A. et al.. (2016). STC anti-spoofing systems for the ASVspoof 2015 challenge. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 5475–5479). doi:10.1109/ICASSP.2016.7472724

Patel, T., & Patil, H. (2015). Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 2062-2066).

Patel, T., & Patil, H. (2015). Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Patel, T., & Patil, H. (2016). Effective of fundament frequency (f0) and strength of excition for spoofed speech detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (pp.5105–5109).

Paul, D., Pal, M., & Saha, G. (2017). Spectral features for synthetic speech detection. *IEEE Journal of Selected Topics in Signal Processing*, *11*(4), 605–617. doi:10.1109/JSTSP.2017.2684705

Sahidullah, M., Kinnunen, T., & Hanilci, C. (2015). A comparison features for synthetic speech detection. In *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 2087-2091).

Scheirer, E., & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 1331–1334). doi:10.1109/ICASSP.1997.596192

Seide, F., Li, G., Chen, X., & Yu, D. (2011, December). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (pp. 24-29). IEEE. doi:10.1109/ASRU.2011.6163899

Sriskandaraja, K., Sethu, V., Ambikairajah, E., & Li, H. (2017). Front-end for anti-spoofing countermeasures in speaker verification: Scattering spectral decomposition. *IEEE Journal of Selected Topics in Signal Processing*, *11*, 632–643.

Todisco, M., Delgado, H., & Evans, N. (2016). A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. In *The speaker and language recognition workshop*. Bilbao, Spain: ODYSSEY.

Wu Z. & Li H. (2014). Voice conversation versus speaker verification: an overview. *APSIPA Transactions on Signal and Information Processing*, 1-16.

Xiao, X., Tian, X., & Du, S. et al.. (2015). Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. In *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 2052-2056).

Zhang, C., Ranjan, S., & Nandwana, M. et al.. (2016). Joint information from nonlinear and linear features for spoofing detection: an i-vector based approach. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (pp. 5035-5038). doi:10.1109/ICASSP.2016.7472636

Zhang, C., Yu, C., & Hansen, J. (2017). An investigation of deep learning frameworks for speaker verification anti-spoofing. *IEEE Journal of Selected Topics in Signal Processing*, *11*(4), 684–694. doi:10.1109/JSTSP.2016.2647199

Jichen Yang is a post-doc of SCUT.

Qianhua He received the B. S. Degree in physics from Hunan Normal University in 1987, the M. S. Degree in medical instrument engineering from Xi'an Jiaotong University in 1990, and the Ph. D degree in communication engineering from South China University of Technology in 1993. Since 1993, he works with South China University of Technology. From 1994 to 1996, he worked with City University of Hong Kong. From 2007.11 to 2008.10, he worked with University of Washington in Seattle as a visiting scholar. His research interests include speech processing, digital audio forensic, speech coding, multimedia retrieval and audio event analysis. Email: eeqhhe@scut.edu.cn.

Yongjian Hu is a Prof. of SCUT.