Interspeech 2018
2-6 September 2018, Hyderabad

Feature with Complementarity of Statistics and Principal Information for Spoofing Detection

Jichen Yang[1], *Changhuai You*[2], *Qianhua He*[1*]

[1]School of Electronic and Information Engineering, South China University of Technology, China
[2]Institute for Infocomm Research, A*STAR, Singapore

eenisonyoung@scut.edu.cn, echyou@i2r.a-star.edu.sg, eeqhhe@scut.edu.cn

Abstract

Constant-Q transform (CQT) has demonstrated its effectiveness in anti-spoofing feature analysis for automatic speaker verification. This paper introduces a statistics-plus-principal information feature where a short-term spectral statistics information (STSSI), octave-band principal information (OPI) and full-band principal information (FPI) are proposed on the basis of CQT. Firstly, in contrast to conventional utterance-level long-term statistic information, STSSI reveals the spectral statistics at frame-level, moreover it provides a feasibility condition for model training while only small training database is available. Secondly, OPI emphasizes the principal information for octave-bands, STSSI and OPI creates a strong complementarity to enhance the anti-spoofing feature. Thirdly, FPI is also of complementary effect with OPI. With the statistical property over CQT spectral domain and the principal information through discrete cosine transform (DCT), the proposed statistics-plus-principal feature shows reasonable advantage of the complementary trait for spoofing detection. In this paper, we setup deep neural network (DNN) classifiers for evaluation of the features. Experiments show the effectiveness of the proposed feature as compared to many conventional features on ASVspoof 2017 and ASVspoof 2015 corpus.

Index Terms: constant-Q transform, anti-spoofing countermeasure, automatic speaker verification

1. Introduction

Conventional speaker verification system becomes frail or incompetent while facing attack from spoofed speech. There are three main challenging attacks from different sources, synthetic speech [1, 2, 3], voice converted speech [4, 5, 6], and playback speech [7, 8, 9].

Countermeasure of spoofing attacks has been studied presently, focusing on feature and classifier respectively. The features used for anti-spoofing detection can be generalized into three categories: Long-term spectral statistics based feature [10], phase spectrum based feature [11, 12] and power spectrum based feature. In [13], two types of long-term spectral statistics, i.e. first and second order statistics over the entire utterance in each of DFT frequency bin, are concatenated to form a single vector representation of an utterance. Typical phase spectrum based features are the cosine normalized phased feature (CNPF), group delay (GD)[14], instantaneous frequency (IF), and instantaneous frequency cosine coefficients. There are many variants of the power spectrum based feature such as the scattering cepstral coefficients (SCC) [15], speech-signal frequency cepstral coefficients (SSFCC) [3], and constant-Q cepstral coefficients (CQCC) [16, 17]. CQCC is the most widely used feature; it was firstly applied in synthetic and voice converted speech detection [18], then used in playback speech detection [19, 20, 21]. CQCC adopts a constant-Q transform (CQT) for the spectral analysis. The CQT employs geometrically spaced frequency bins. In contrast to the Fourier transform which imposes regular spaced frequency bins and hence leads to variable Q factor, the CQT ensures a constant Q factor across the entire spectrum. This trait allows the CQT to provide higher spectral resolution at lower frequencies while providing a higher temporal resolution at higher frequencies, as a result the distribution of the CQT time-frequency resolution is consistent with human hearing characteristics. Founded upon the basis of CQT, the CQCC has been reported to achieve effective performance for speech synthesis and voice conversion spoofing detection [18].

In this paper, we aim to study complementarity of sub-features that are used to form concatenated features through constant-Q transformation. Different from the conventional CQCC feature, each sub-feature is of complementary information to one another. The first sub-feature is STSSI that is considered to carry the statistic information at frame level, in which the first- and second-order statistics over different CQT-spectral bins are obtained. The second sub-feature is OPI, which is to provide the octave principle information, where octave segmentation and discrete cosine transform (DCT) are applied. And the third sub-feature is from FPI, it formulates the full-band principle information from the CQT spectrum. Finally, the three sub-features are combined to generate its delta and acceleration coefficients as a feature for spoofing detection. We refer to the proposed feature as constant-Q statistics-plus-principal information coefficient (CQSPIC). In this paper, we adopt deep neural network (DNN) as the means for the feature evaluation.

The remainder of the paper is organized as follows. The CQT is briefly introduced in Section 2. In Section 3, we describe in detail the proposed CQSPIC feature. Section 4 gives the experimental results and corresponding analysis, which is based on ASVspoof 2017 corpus and ASVspoof 2015 corpus. Finally, Section 5 concludes the paper.

2. Constant-Q Transform

CQT is related to the discrete Fourier transform (DFT) [22]. Different from DFT, the ratio of center frequency to bandwidth, Q, is constant, which makes CQT spectrum have a higher frequency resolution in low frequency and higher temporal resolution in higher frequency.

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

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

10.21437/Interspeech.2018-1693

where $k = 1, 2, ..., K$ denotes the frequency bin, l is the time frame index and M the frame shift size so that $n = lM$, a_k^* is the complex conjugate of a_k, and $\lfloor \cdot \rfloor$ rounds a value to the nearest integer towards negative infinity. The basic function a_k is complex-valued time-frequency atom

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

where f_k is the centre frequency of the kth bin, f_s is the sampling frequency, and $\nu(t)$ is a window function (e.g. Hanning window). $\phi_{\mathbf{k}}$ is the phase offset. C is a scaling factor given below

$$C = \sum_{m=-\lfloor \frac{N_k}{2} \rfloor}^{\lfloor \frac{N_k}{2} \rfloor} \nu(\frac{m + \frac{N_k}{2}}{N_k}) \quad (3)$$

Since a bin spacing is desired to be of equal temperament, the center frequency f_k is set by

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

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

Recently, CQCC was reported to be sensitive to the general form of spoofing attack so it becomes an effective spoofing countermeasure [18].

3. Proposed Constant-Q Statistics-plus-Principal Information Coefficient (CQSPIC)

In this paper, we aim to seek an effective feature with different complementary characteristics for spoofing detection on the basis of the advantages of CQT. Consequently, we propose a constant-Q statistics-plus-principal information coefficient (CQSPIC) that includes three characteristics: STSSI, OPI and FPI.

3.1. Short-term Statistics Information

In spoofing detection, we face a situation where there is insufficient prior knowledge about the characteristics to distinguish a spoofed speech from genuine speech. It is known that the two kinds of speech signals have two different statistical characteristics.

In [23], long-term spectral statistics (LTSS) is reported to be effective for spoofing detection in speaker verification system. It is believed that the mean and variance of the spectral amplitude distributed over either a long-term period of certain spectrum or a range of frequencies at a time frame can provide good traits to distinguish the two different kinds of speech signals. However, LTSS is not suitable for small training database due to insufficient feature data generated. In this paper, we propose a short term statistics at frame level for the purpose of solving the small training data issue and build complementary characteristics on the basis of CQT.

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

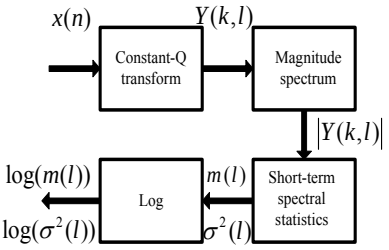

Figure 1: *Block diagram of short-term statistics extraction.*

is used to obtain mean and variance in log-scale. Fig. 1 shows block diagram of short-term information statistics extraction. To estimate STSSI cross frequency bins at frame-level, one is to estimate the statistics over full frequency-band, the other is to compute the statistics over each individual subband such as the octave-band. To generalize the statistics formula, we give the subband statistics as follows. Supposing $|Y(k,l)|$ is a frame magnitude spectrum of $Y(k,l)$ The mean of the CQT spectral amplitude over subband, $\mathbf{m}_s$, is defined by

$$\mathbf{m}_s(l) = \frac{1}{K_s - K_{s-1}} \sum_{k=K_{s-1}+1}^{K_s} |Y(k,l)|, \quad s = 1, ..., S \quad (5)$$

And the variance of the CQT spectrum amplitude over subband is defined by

$$\sigma_s^2(l) = \frac{1}{K_s - K_{s-1}} \sum_{k=K_{s-1}+1}^{K_s} (|Y(k,l)| - \mathbf{m}_s(l))^2 \quad (6)$$

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

Experiments on ASVspoof 2017 database show the octave-band statistics is not competent with full-band statistics for spoofing detection. It may be because that there are insufficient frequency bins to approximate the statistics in an octave-band. Subsequently, we only focus on reporting the performance with full-band statistics.

3.2. Octave-band Principal Information

The term 'octave' is derived from the western musical scale and is therefore common in audio electronics [24, 25]. The Law of Octaves states that we can use an octave of a frequency to the same effect as the frequency itself. An octave is the doubling or halving of a certain frequency. The speech frequency range can be separated into unequal segments called octaves. A band is defined to be an octave in width when the upper band frequency is twice the lower band frequency.

On the other hand, in contrast to DFT where frequency region of each frequency bin is equal, the frequency region of different frequency bin in CQT is different. The centre frequency bin of CQT complies with a nonlinear distribution with (4), we have

$$f_{nB+k} = 2^{\frac{nB+k-1}{B}} f_1 = 2^n f_k = 2f_{(n-1)B+k} \quad n = 1, ..., N \quad (7)$$

where N denotes the number of octave-bands. So we have $K = N * B$. From (7) we can see that $f_{B+1} = 2f_1$, $f_{2B+1} = 2f_{B+1}$, ... , $f_{NB+1} = 2f_{(N-1)B+1}$. Therefore, B frequency

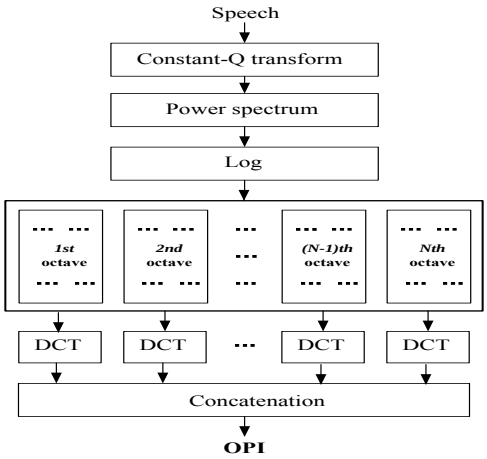

Figure 2: *Procedure of the OPI extraction.*

bins (i.e. $f_1, f_2, ..., f_B$) between f_1 and f_{B+1} form the first octave band; B frequency bins between f_{B+1} and f_{2B+1} (i.e. $f_{B+1}, f_{B+2}, ..., f_{2B}$) form the second octave band; ...; and B frequency bins (i.e. $f_{(N-1)B+1}, f_{(N-1)B+2}, ..., f_{NB}$) between $f_{(N-1)B+1}$ and f_{NB+1} form the N-th octave band. As a result, there are B frequency bins in each of octave-band with CQT. The higher an octave-band is, the larger frequency region the corresponding frequency bin occupies.

In this paper, we propose an octave principal information (OPI) on the basis of CQT. In OPI, octave segmentation is applied, and it is followed by a DCT to generate principal information. In particular, OPI includes five modules: CQT, power spectrum, octave segmentation, log and DCT. The p-th principal coefficients of the n-th octave-band is given using discrete cosine transform as follows:

$$X_{np}(l) = \sum_{k=(n-1)B+1}^{nB} \log\big(|(Y(k,l)|^2\big) cos\big[\frac{\pi}{B}(k+\frac{1}{2})p\big] \quad p = 1, 2, ..., P \tag{8}$$

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

3.3. Full-band Principal Information

In this paper, we propose a full-band principal information (FPI) as complementary characteristics of the OPI. Different from the CQCC with linearized log power spectrum resampling, the FPI directly applies DCT on logarithm power spectrum in CQT domain. For the FPI feature extraction, there are four modules including CQT, power spectrum, logarithm and DCT. In the FPI, the r-th principal coefficients are given via DCT as follows:

$$Z_r(l) = \sum_{k=1}^{K} \log\big(|(Y(k,l)|^2\big) cos\big[\frac{\pi}{K}(k+\frac{1}{2})r\big] \tag{9}$$

$r = 1, 2, ..., R$ where R is the number of principal coefficients. Fig. 3 shows the block diagrams of the FPI procedure.

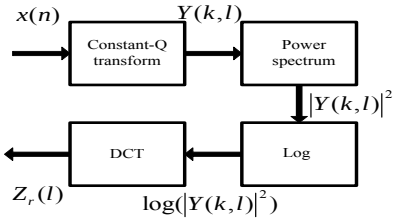

Figure 3: *Block diagram of FPI extraction.*

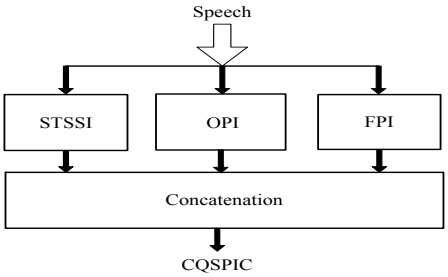

Figure 4: *Block diagram of the extraction of the proposed constant-Q statistics and principal information coefficient.*

3.4. Combination, Delta and Acceleration

The proposed CQSPIC is formed by combining the three subfeatures: STSSI, OPI and FPI. OPI and FPI are complementary because they represent octave spectral information and full-band spectral information respectively. STSSI represents statistics, it is of complementarity with both OPI and FPI.

The STSSI (either mean or variance), OPI and FPI are concatenated, delta and double-delta of the concatenated feature are applied to produce the final CQSPIC feature. Fig.4 illustrates the block diagram of CQSPIC feature extraction.

In playback speech detection, our experiment shows that the STSSI mean from STSSI has discriminative property rather than variance. In synthetic or voice converted speech detection, the STSSI variance can capture the dynamics between natural and synthetic speech. Therefore, we select the STSSI mean, OPI and FPI to form the CQSPIC feature for playback spoofing detection, while we select the STSSI variance, OPI and FPI to form the CQSPIC feature for synthetic or voice conversion speech detection.

4. Performance Evaluations

In this paper, the anti-spoofing performance of the proposed CQSPIC feature is evaluated in terms of equal error rate (EER) and average EER (AEER) on two automatic speaker verification (ASVspoof) databases: ASVspoof 2015 [1] and ASVspoof 2017 [26, 27]. In CQT computation, all configuration parameters are set to be the same as those in [18]. For OPI, we set $P = 12, N = 9$, as a result, there are 108 dimensions of static OPI. For FPI, R is set to 12, it means the FPI has 12 dimensions of its principal vector. In the feature evaluation, we trained DNN models with stochastic gradient descent (SGD) as spoofing detection platform using computational network toolkit (CNTK) [28]. In particular, different DNN models are trained corresponding to different features such as MFCC, CQCC, proposed OPI and final proposed CQSPIC. Here, the static dimension of CQCC and MFCC are 12 and 13 respectively. In this evaluation, the input layer of the DNN is the feature coefficients of eleven spliced frames centred by the current

Table 1: *The experiment results for ASVspoof 2015 evaluation set using CQSPIC-D, CQSPIC-DA and CQSPIC-A.*

Feature	Known attack					Unknown attack					AEER
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	
CQSPIC-D	0	0.004	0	0	0.024	0.018	0.004	0.009	0	0.860	0.092
CQSPIC-DA	0	0	0	0	0.009	0.006	0	0.008	0	0.820	0.084
CQSPIC-A	0	0	0	0	0.004	0	0	0.008	0	0.368	0.038

frame. The feature coefficients of each frame can be the static feature coefficient, or its delta, or its double delta (i.e. acceleration), or their combining feature. In our experiment, it is observed that delta or double-delta or their concatenated features without static coefficients may give better performance than those with static coefficients in spoofing detection; and similar phenomenon is also reported in [29] and [18]. During evaluation, we use D and A to represent delta and acceleration respectively.

4.1. ASVspoof 2015 Evaluation

The ASVspoof 2015 database only contains speech synthesis and voice conversion attacks produced through logical access. Only five types of attacks are in the training set marked as S1, S2, ..., S5, while ten types are in the evaluation set marked as S1, S2, ..., S10. It creates known and unknown attacks for evaluation. For evaluation on ASVspoof 2015, we use 16,375 training utterances to train the deep neural network (DNN) model, which has four hidden layers with 512 nodes per layer and one output layer with 2 nodes indicating genuine and spoofed speech.

For speech synthesis and voice conversion, the variance component of STSSI is found to give good performance and therefore used to form the proposed CQSPIC. In other words, the CQSPIC for ASVspoof 2015 platform comes from the combination of OPI, FPI and variance of STSSI, i.e. OPI+FPI+STSSIv. Table 1 shows the experiments result (EER) on ASVspoof 2015 evaluation set using CQSPIC-D, CQSPIC-A and CQSPIC-DA. It can be seen that CQSPIC-A performs the best with AEER of 0.038%. In the next experiments for ASVspoof 2015, we will use acceleration (i.e. 'A') as the final features. Table 2 shows the comparison between different features for ASVspoof 2015 under the same DNN structure.

Table 2: *Performance comparison with different features on ASVspoof 2015 in terms of AEER(%).*

Feature	AEER	Feature	AEER
FPI	0.392	MFCC	2.602
OPI+FPI	0.042	CQCC	0.184
OPI+FPI+STSSIm	0.046	OPI	0.134
OPI+FPI+STSSImv	0.045	OPI+CQCC	0.066
OPI+FPI+STSSIv	0.038	OPI+CQCC+STSSIv	0.062

4.2. ASVspoof 2017 Evaluation

Different from ASVspoof 2015 which focuses merely on speech synthesis and voice conversion, ASVspoof 2017 is designed to detect playback attack. In ASVspoof 2017 evaluation, 4,726 utterances in both training and development sets are used to train the model which is used for evaluation set. A series of four-layer DNN including two hidden layers of 512 nodes each layer are trained, while the input and output layers are the same as the DNN models in the ASVspoof 2015 evaluation.

It is observed that the mean from STSSI is more helpful than variance for the playback situation. The CQSPIC for ASVspoof 2017 evaluation is from the combination of OPI, FPI, and STSSI's mean, i.e. OPI+FPI+STSSIm. We investigate the performance of delta (D), accelration (A) and their concatenated (DA), and Fig.5 shows the experimental results. We can see that the CQSPIC-DA is the best in terms of EER. In the next experiments for ASVspoof 2017, we will use DA as the final features. Table 3 shows the comparison between different features for ASVspoof 2017 under the same DNN structure.

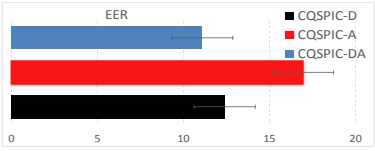

Figure 5: *Experimental result (EER(%)) comparison among CQSPIC-D, CQSPIC-A and CQSPIC-DA on ASVspoof 2017 evaluation set.*

Table 3: *Performance comparison with different features on ASVspoof 2017 in terms of EER(%).*

Feature	EER	Feature	EER
FPI	24.67	MFCC	18.36
OPI+FPI	13.81	CQCC	15.05
OPI+FPI+STSSImv	11.19	OPI	14.08
OPI+FPI+STSSIv	11.66	OPI+CQCC	13.77
OPI+FPI+STSSIm	11.09	OPI+CQCC+STSSIm	11.40

From the above experimental results, we can see that the proposed CQSPIC (i.e. OPI+FPI+STSSIv for ASVspoof 2015 and OPI+FPI+STSSIm for ASVspoof 2017) greatly outperforms the conventional CQCC and MFCC.

5. Conclusion

On the basis of the advantage of CQT, we proposed a useful feature, CQSPIC, by extracting information from octave-band, full-band and short-term statistics for spoofing detection in speaker verification system. The complementarity of the sub-features have been investigated for the different types of spoofing attacks: synthetic speech, voice converted speech, and playback speech. Compared to conventional MFCC and CQCC features, CQSPIC brings more channel information in playback speech detection and more artifacts in synthetic (voice converted) speech detection. The experimental results show that the CQSPIC outperforms CQCC and MFCC. And the complementarity of FPI to OPI+STSSI is better than that of CQCC. The combination of OPI, FPI and STSSI is reasonable and useful for spoofing detection.

6. Acknowledgment

This work is partly supported by National Nature Science Foundation of China (61571192), Natural Science Foundation of Guangdong Province (2015A030313600), Science and Technology Planning Projects of Guangdong Province (2017B010110009), and China Scholarship Council (CSC). In addition, Qianhua He is the corresponding author of the paper.

7. 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] Junichi Yamagishi, Kinnunen Tomi, Nicholas Evans, Phillip De Leon, and Isabel Trancoso, "Introduction to the issues on spoofing and countermeasures for automatic speaker verification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 588–604, 2017.

[3] Dipjyoti Paul, Monisankha Pal, and Goutam Saha, "Spectral features for synthetic speech detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 605–617, 2017.

[4] Zhizheng Wu, Junichi Yamagishi, Kinnunen Tomi, Md Sahidullah, Aleksandr Sizov, Nicholas Evans, Massimiliano Todisco, and Hector Delado, "Asvspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 588–604, 2017.

[5] Xiaohai Tian, Lee Siu Wa, Zhizheng Wu, Eng Siong Chng, and Haizhou Li, "An examplar-based approach to frequency warping for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1875, 2017.

[6] Chunlei Zhang, Shivesh Ranjan, Mahesh Kumar Nandwana, Qian Zhang, Abhinav Misra, Gang Liu, Finnian Kelly, and John H. L. Hansen, "Joint information from nonlinar and linear features for spoofing detection: an i-vector based approach," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5035–5038, 2016.

[7] Wei Shang and Maryhelen Stevenson, "A preliminary study of factors affecting the performance of a playback attack detector," in *Proceddings of Canadian Conference onElectrical and Computer Engineering(CCECE),*, 2008, pp. 459–464.

[8] Zhifeng Wang, Qianhua He, Xueyuan Zhang, Haiyu Luo, and Zhuosheng Su, "Playback attack detection based on channel pattern noise," in *Journal of South China University of Technology (Natural Science Edition)*, 2011, pp. 1708–1713.

[9] Parav Nagarshenth, Elie Khoury, kailash Patil, and Matt Garland, "Replay attack detection using dnn for channel discrimination," *18th Annual Conference of the International Speech Communication Association(INTERSPEECH)*, pp. 97–101, 2017.

[10] Zhifeng Wang, Gang Wei, and Qianhua He, "Channel pattern noise based on playback attack detection algorithm for speaker recognition," in *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics*, 2011, vol. 39, pp. 5–12.

[11] Sarfaraz Jelil, Rohan Kumar Das, S. R. M. Prasanna, and Rohit Sinha, "Spoof detection using source, instantaneous frequecny and cepstral features," *18th Annual Conference of the International Speech Communication Association(INTERSPEECH)*, pp. 22–26, 2017.

[12] Cochleara B. patel and Hemant A. Patil, "Cochlear filter and instantaneous frequency based features for spoofed speech detecton," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 618–631, 2017.

[13] Hannah Muckenhirn, Pavel Korshunov, Mathew Magimai-Doss, and Sebastein Marcel, "Long-term spectral statistics for voice presentation attack detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2098–2111, 2017.

[14] Xiong Xiao, Xiaohai Tian, S. 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)*, 2015.

[15] Kaavya Sriskandaraja, Vidhyassharan Sethu, Eliathamby Ambikairajah, and Haizhou Li, "Front-end for anti-spoofing countermeasures in speaker verification: scattering spectral decomposition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 632–643, 2017.

[16] Massimiliano Todisco, Hector Delgado, and Nicholas Evans, "Constant q cepstral coefficients: a spoofing countermeasure for automatic speaker verification," *Computer, speech and language*, pp. 759–762, 2017.

[17] Zhuxin Chen, Zhifeng Xie, Weibin Zhang, and Xiangmei Xu, "Resnet and model fusion for automatic spoofing detection," *18th Annual Conference of the International Speech Communication Association(INTERSPEECH)*, pp. 102–106, 2017.

[18] Massimiliano Todisco, Hector Delgado, and Nicholas Evans, "A new feature for automatic speaker verification anti-spoofing:constant q cepstral coefficients," in *the speaker and language recognition workshop(ODYSSEY)*, 2016.

[19] Xianliang Wang, Yanhong Xiao, and Xuan Zhu, "Feature selection based on cqccs for automatic speaker verification spoofing," *18th Annual Conference of the International Speech Communication Association(INTERSPEECH)*, pp. 32–36, 2017.

[20] Marcin Withowski, Stanislaw Kacprasko, Piotr Zelasko, Konrad Kowlczyk, and Jakub Galka, "Audio replay attack detection using high-frequency features," *18th Annual Conference of the International Speech Communication Association(INTERSPEECH)*, pp. 27–31, 2017.

[21] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudasher, and Vadim Shchemelinin, "Audio replay attack detection with deep learning framework," *18th Annual Conference of the International Speech Communication Association(INTERSPEECH)*, pp. 82–86, 2017.

[22] Judith C. Brown, "An efficient algorithm for the calculation of a constant q spectral transform," *Journal of Acoustical Society of America*, vol. 92, 1992.

[23] Hannah Muckenhirn, Pavel Korshunov, Mathew Magimai-Doss, and Sebastein Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," *In proceeding of International Conferences of Biometrics Special Interest Group*, pp. 1–6, 2016.

[24] Leon Crickmore, "New light on the babylonian tonal system," *Proceedings of the International Conference of Near Eastern Archaeomusicology (ICONEA 2008)*, pp. 11–12, 2008.

[25] L. Demany and F. Armand, "Cntk:microsoft's open-source deep learning toolkit," *Journal of Acoustical Society of America*, vol. 76, pp. 57–66, 1984.

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

[27] Tomi Kinnunen, and Hector Delgado Md Sahidullah, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee, "The asvspoof 2017 challenge: assessing the limits of replay spoofing attack detection," in *Annual Conference of the International Speech Communication Association(INTERSPEECH)*, 2017.

[28] Frank Seide and Amit 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.

[29] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilci, "A comparison features for synthetic speech detection," *Annual Conference of the International Speech Communication Association(INTERSPEECH)*, 2015.