Spoofing Detection Employing Infinite Impulse Response - Constant Q Transform-based Feature Representations

Jahangir Alam
Computer Research Institute of Montreal
Montreal, Quebec, Canada
jahangir.alam@crim.ca

Patrick Kenny
Computer Research Institute of Montreal
Montreal, Quebec, Canada
Patrick.kenny@crim.ca

***Abstract*—Speaker recognition researchers acknowledge that systems which aim to verify speakers automatically based on their pronunciation of an utterance are vulnerable to spoofing attacks using voice conversion and speech synthesis technologies. The first automatic speaker verification spoofing and countermeasures challenge (ASVspoof2015) was designed to stimulate interest in this problem among the speaker recognition communities. In the course of the challenge and subsequently, it became clear that the most effective countermeasures against spoofing attacks are low-level acoustic features (typically extracted at 10 ms intervals) designed to detect artifacts in synthetic or voice converted speech. In this work, we demonstrate the effectiveness of the infinite impulse response - constant Q transform (IIR-CQT) spectrum-based cepstral coefficients (ICQC) as anti-spoofing front-end. The IIR-CQT spectrum is estimated by filtering the multi-resolution fast Fourier transform with an infinite impulse response filter. These features can be used on their own with a standard Gaussian mixture model backend to detect spoofing attacks or they can be used in tandem with bottleneck features which are extracted from a bottleneck layer in a deep neural network designed to discriminate between synthetic and natural speech. We show that the ICQC features are capable of producing very low equal error rates on the individual spoofing attacks in the ASVspoof2015 data set (0.02% on the known attacks, 0.23% on the unknown attacks, and 0.13% on average). Moreover, with a single decision threshold (common to all of the attacks), the ICQC front end yielded an equal error rate of 0.20%.**

Keywords—spoofing detection; ASVspoof2015; GMM; bottleneck features; ICQC;

I. INTRODUCTION

Spoofing refers to a situation in which a person or computer program successfully impersonates a legitimate user of an authentication system. Impersonation, replay, speech synthesis and voice conversion are some examples of spoofing attacks. Impersonation or human mimicking requires a mimic to imitate a target speaker's voice and it does not pose a genuine threat to automatic speaker verification system [1-5]. Replay attacks consist of playing back the pre-recorded voice of a target speaker to spoof the system. Liveness detection and detection of channel differences are found to effective in this situation [1, 5].

Given the availability of open-source toolkits online, spoofing attacks based on speech synthesis and voice conversion techniques are potentially more serious [1-4]. This problem has attracted the interest of both speech synthesis and speaker recognition researchers. Various Countermeasures have been developed and investigated since the susceptibility of voice biometrics to spoofing attacks has been recognized by the research community. In most systems, prior knowledge about the specific spoofing type plays a vital role for spoofing detection [2, 6]. The automatic speaker verification spoofing and countermeasures challenge ASVspoof2015 [4] provided a common framework for the evaluation of spoofing countermeasures in the presence of known and unknown spoofing attacks. These spoofing attacks were generated using different voice conversion and speech synthesis techniques. During and after the ASVspoof2015 challenge, many countermeasures based on spectral amplitude, phase [5-7, 10-19], and combined amplitude-phase [8-9], have been used for spoofing detection. Some recent studies using the ASVspoof2015 corpus include constant Q cepstral coefficients [7], pitch contour and strength of excitation [19] for spoofing detection, and analyses of robustness of spoofing detection systems in the presence of additive and convolutive noise [15, 16].

In this work, our main goal is to demonstrate the effectiveness of a new feature representation derived from the infinite impulse response - constant Q transform by recursively filtering the multi-resolution fast Fourier transform of the signal [20, 21]. We refer to these features by the acronym ICQC for Infinite impulse response Constant Q transform Cepstrum. The constant Q transform (CQT) [28], widely used in music signal processing, is the direct evaluation of the discrete Fourier transform (DFT) in a way which keeps the "quality factor" Q constant by varying the channel bandwidth proportionally to its center frequency. Hence the CQT provides a finer frequency resolution for low frequencies whereas temporal resolution improves with increasing frequency.

ISBN 978-0-9928626-7-1 © EURASIP 2017

This feature makes the CQT well suited for audio signals as it better reflects the resolution in the human auditory system than the uniform frequency resolution provided by the fast Fourier transform (FFT) used in STFT (short-time Fourier transform) analysis.

Based on a direct evaluation of CQT and then converting geometric space to linear space, the constant Q cepstral coefficients (CQCC) feature was introduced as a spoofing countermeasure in [7]. In order to avoid the computational expense of directly evaluating the CQT, we use an efficient approximation introduced in [21] to compute the ICQC features that we propose in this paper for anti-spoofing. ICQC features have already been applied in [23] for speech recognition on the 4-th CHiME speech separation and recognition challenge tasks [24]. **I**IR-**C**QT spectrum-based **M**el frequency **C**epstral coefficients (ICMC) have also been used in [25] for a joint utterance and text-dependent speaker verification task.

II. CONSTANT Q TRANSFORM-BASED FEATURES

A. The Contant Q transform (CQT)

The CQT transforms a time-domain signal $x(n)$ into the time-frequency domain so that the center frequencies of the frequency bins are geometrically spaced and the quality factor Q remains constant. Mathematically, the *k*-th spectral component of the CQT is expressed as:

$$X^{cq}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x(n) w(n,k) e^{-j\frac{2\pi Q n}{N_k}}, \qquad (1)$$

where *n* and *k* are time and frequency domain indices, $w_k(n)$ is an analysis window of length N_k and the Q-factor Q, which depends on number of bins per octave *b*, is given by

$$Q = \left(2^{\frac{1}{b}} - 1\right)^{-1}. \qquad (2)$$

The Q-factor is defined as the ratio of center frequency to the bandwidth of each window. Because of this constant Q-factor the CQT provides better frequency resolution for low frequencies and the temporal resolution is better for high frequencies. This feature makes the CQT well suited for audio signals as it better reflects the resolution in the human auditory system than the uniform frequency resolution provided by the fast Fourier transform used in the short-time Fourier transform analysis [21]. The CQT is widely used in music signal processing as the center frequencies of analysis are aligned with the equal tempered scale and it enables music signals to be analyzed with a frequency resolution high enough to separate different notes within an octave [32].

B. Contant Q Cepstral coefficients (CQCC)

In [7], the CQCC feature was proposed for the spoofing detection task and a significant reduction in EER (equal error rates) was demonstrated. The best performance reported in [7] was accomplished when only 20-dimensional acceleration coefficients (denoted as CQCC-A in [7]) were used as countermeasure. Fig. 1 presents a schematic diagram showing the various steps to extract CQCC features as described in [7]. After estimating CQT spectra, logarithmic compression is applied. A spline interpolation is applied to the estimated spectra to convert the geometric frequency scale to a linear scale [7]. Finally, CQCC features are obtained by applying the discrete cosine transform. Similar to [7], the number of bins per octave was set to b = 96 so that the corresponding quality factor is $Q = 1/(2.\hat{}(1/b) - 1) = 138$.

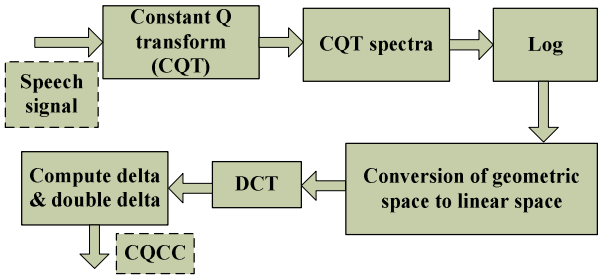

Fig. 1. Extraction of constant Q cepstral coefficients (CQCC) features as proposed in [7]. 40-dimensional delta + double delta and 20-dimensional double delta (i.e., acceleration) coefficients are used as countermeasures and denoted here as the CQCC and CQCC-A [7], respectively.

III. IIR-CQT-BASED CEPSTRAL COEFFICIENTS

In this section, we propose to compute a spoofing countermeasure which is based on an infinite impulse response - constant Q transform (IIR-CQT) spectrum. The IIR-CQT spectrum is estimated by recursive filtering of the multi-resolution fast Fourier transform of signal [20, 21]. We denote this features as the **I**IR-**CQ**T spectrum-based **C**epstrum (ICQC) features.

Direct evaluation of CQT, i.e., eqn. (1), is very time consuming. Since the frequency bins in CQT spectra are geometrically spaced it is very difficult work with than the time-frequency representations obtained by using the short-time Fourier transform. It was shown in [22] that by taking advantage of the fast Fourier transform an approximation of CQT can be computed efficiently. The IIR-CQT, which is used here compute ICQC features, shows a good compromise between the flexibility of efficient CQT [22] and the low computational cost of multi-resolution fast Fourier transform [20].

A. Computation of ICQC

In order to compute ICQC features we first estimate the IIR-CQT spectra in the following steps:

1) Design an infinite impulse response (IIR) filterbank that has constant Q behavior. The location of the poles of the IIR filterbank vary for each frequency bin along the real axis to obtain different time window widths resulting in a multi-resolution behavior of the transform. The window width is wider for lower frequency and narrower for higher frequency.

2) After the computation of poles, a simple and effective design of the linear time varying (LTV) IIR filterbank consists in choosing for each frequency bin the corresponding pole of the IIR filterbank, that is pole varies with frequency

ISBN 978-0-9928626-7-1 © EURASIP 2017

$p[n] = p(k)$ [21]. Therefore, the recursive equation of the filter to approximate a LTV IIR filterbank can be expressed as [21]:

$$Y(k) = X(k) + X(k+1) + p(k)Y(k-1), \qquad (3)$$

where $p(k)$ is the pole corresponding to the k-th frequency bin, $X(k)$ is the DFT spectrum of the speech signal.

The filter is applied in the forward direction followed by reverse filtering to obtain the IIR-CQT spectrum $Y(k)$. After estimating the IIR-CQT spectrum the next step is to compute ICQC features. These are obtained by applying the DCT to the estimated spectrum followed by logarithmic compression. Alternatively, We can compute ICQC features from the IIR-CQT log spectrum by applying PCA instead of DCT as shown in fig. 2. In this work, we carried out experiments with both types of ICQC features and they are denoted here as ICQC (DCT) and ICQC (PCA).

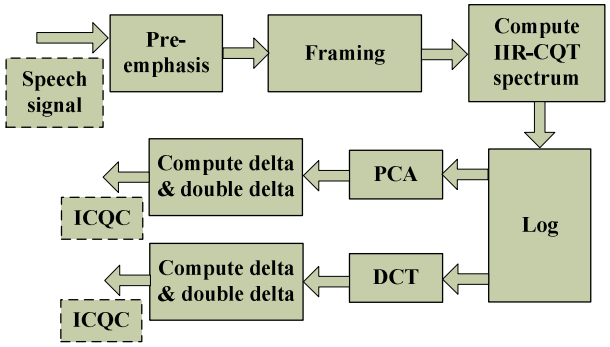

Fig. 2. Extraction of ICQC spoofing countermeasure features from the IIR-CQT spectrum [23] when the discrete cosine transform (DCT) and principal component analysis (PCA) are used for decorrelation. In this work we empirically choose quality factor Q = 13. 40-dimensional delta + double delta and 30-dimensional double delta (i.e., acceleration) coefficients are used as countermeasures.

In fig. 3 we present a comparison of spectrograms estimated using the DFT and IIR-CQT for a spoof signal (D1_1003515.wav) randomly selected from the ASVspoof2015 corpus. It is observed from fig. 3 that in the low frequency band, where there is a higher density of components, the IIR-CQT provides a better discrimination. This is because the time windows of IIR-CQT are flatter than the typically used windows such as Hamming window. The IIR-CQT also captures the artifacts present in the nonspeech regions (shown by the rectangles in fig. 3 (c)).

In order to report a real time (RT) factor for ICQC and CQCC (both implemented in MATLAB) we conducted experiments on an Intel(R) Xeon(R) CPU E5-2630 0 @ 2.30GHz with a total memory of 126GB. For this experiment, we selected randomly 5000 recordings (total duration = 18677.55 sec) from the ASVspoof2015 corpus, extracted features and execution times are recorded 10 times for each countermeasure. The execution time for the extraction of CQCC [7] features in a single thread is 1.24 times faster than real time using a memory of 3.99GB. On the contrary execution time for the extraction of IQCC features in a single thread is 20.08 times faster than the real time using a memory of 2.16GB. The real-time factor was obtained by dividing the total processed segments duration with the average execution time.

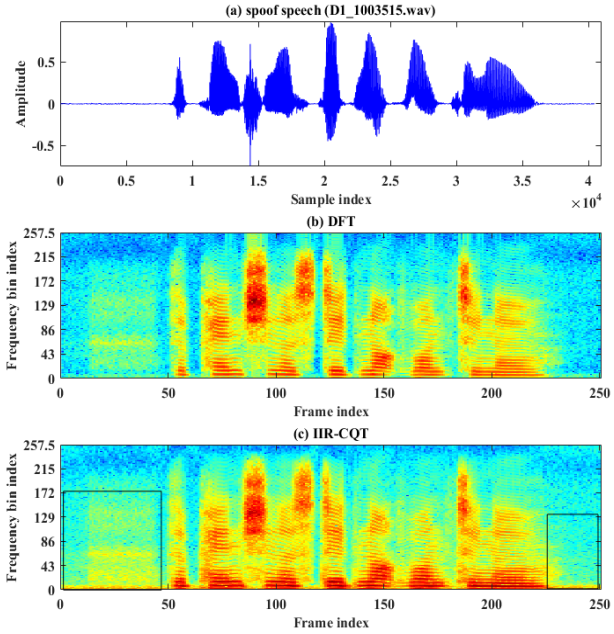

Fig. 3. Comparison of speech spectrograms of (a) spoof signal (D1_1003515.wav) obtained using the (b) DFT and (c) IIR-CQT algorithms. The recording D1_1003515.wav is taken from the ASVspoof2015 corpus.

IV. Extraction of Bottleneck and Tandem Features

In order to extract high level features for the spoofing detection task, we trained a DNN (designed to discriminate between human and spoofed speech signals) on the spoofing challenge training data by using low level features in the input layer of the DNN [8]. The DNN has 5 hidden layers and the 5-th layer is the bottleneck layer. Each hidden layer has 1000 neurons and uses sigmoid activations with the exception of 5-th layer which is linear and has 64 nodes. The output layer is a softmax of dimension 2 i.e., one output for human speech signals and one for spoof signals. After extracting bottleneck features from all the ASVspoof 2015 corpus using the trained DNN, the tandem features [26] are computed by concatenating the ICQC features and bottleneck features and reducing the feature dimension by principal component analysis.

V. Performance Evaluation

In order evaluate the performance of ICQC we carried out spoofing detection experiments using a GMM backend on the ASVspoof2015 corpus [4] and results are reported on the evaluation set. The ASVspoof2015 corpus comprised of ten spoofing attacks denoted by S1, S2, ... S10. S1-S5 attacks are referred as known and S6-S10 are called unknown attacks. For more detail about the corpus and type of spoofing attacks please see [4]. The equal error rate (EER) is used as a metric for performance assessment of the spoofing countermeasures. Spoofing detection scores were evaluated against each spoofing attack as well as using a single decision threshold (common to all spoofing attacks) and results were reported for

ISBN 978-0-9928626-7-1 © EURASIP 2017

known (average of S1-S5), ***unknown*** (average of S6-S10), ***average*** (average of S1-S10) and ***all*** conditions. Besides CQCC features we also considered following features for comparison purpose - Mel frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), linear predictive cepstral coefficients (LPCC), modified group delay [27]-based cepstral coefficients (MGDC), spectral flux cepstral coefficients [10] and product spectral cepstral coefficients (PSCC) [8-9]. All features are of 40-dimensional (delta + double delta coefficients) except for CQCC-A [7] and ICQC (PCA)-A it is 20 and 30, respectively.

It is observed by comparing the performance of LFCC and MFCC of table 1 that integration of an Mel-filterbank is harmful for anti-spoofing. It seems that an auditory filterbank integration smooths out (due to summing operation over frequency bins) some spoofing artifacts. These findings motivated us not to use VAD or an auditory filterbank in any of our spoofing countermeasures considered here (i.e., table 2). Comparing the performances of LFCC (without VAD) versus LFCC (without VAD, 8 kHz) and MFCC (without VAD) versus MFCC (without VAD, 8 kHz) we can conclude that there are significant artifacts in the high frequency (frequency > 4 kHz) regions of the spoof signal and they carry more discriminative information for spoofing detection. This is because most of the spoofing algorithms try to model low frequency content (up to 4 kHz) of speech signal precisely. This agrees with the experience of other participants in the ASVspoof2015 challenge [4, 8-9, 14]. Therefore, we did not perform reduce the sampling rate of ASVspoof2015 corpus for any of our systems reported in table 2.

TABLE 1
Comparison of spoofing detection performance (in terms of EER) of the MFCC and LFCC features on ASVspoof2015 evaluation set with/without VAD (voice activity detection) to remove non-speech frames and without or with downsampling of speech signals. The LFCC (without VAD, 8 kHz) and MFCC (without VAD, 8 kHz) represent systems with downsampling of signals to 8 kHz.

	EER (%)			
	Known	***Unknown***	***Average***	***all***
LFCC (without VAD)	**0.265**	**1.10**	**0.684**	**1.02**
LFCC (with VAD)	0.266	5.52	2.89	5.21
MFCC (without VAD)	1.208	1.905	1.557	2.339
MFCC (with VAD)	1.197	5.688	3.443	6.211
LFCC (without VAD, 8 kHz)	1.612	5.531	3.57	4.869
MFCC (without VAD, 8 kHz)	3.631	5.512	4.575	6.595

In table 2 we report EERs achieved by all the spoofing countermeasures considered in this work for the *known*, *unknown*, *average* and *all* evaluation conditions. The proposed ICQC (PCA) features outperformed all other systems on the *unknown*, *average* and *all* evaluation conditions. On the *known* condition, the Bottleneck and tandem systems demonstrated the best performance with an EER of 0.0%. With ICQC (PCA) features, we achieved EERs of 0.04%, 0.286%, 0.16%, and 0.227% on the *known*, *unknown*, *average* and *all* evaluation conditions, respectively. Compared to the constant Q transform-based features (CQCC-A), the proposed ICQC (PCA) countermeasure provided relative improvements of 54% and 56% on the *unknown* (Table 2, 3rd column) and *average* conditions (Table 2, 4-th column), respectively. Since in [7] the performance of CQCC-A was not reported on the *all* evaluation condition we were not able to compare it in this work. With the ICQC (PCA)-A features we were able to get EERs of 0.02%, 0.23%, 0.125%, and 0.195% on the *known*, *unknown*, *average* and *all* evaluation conditions, respectively. As mentioned in section III (A) that in the low frequency band the proposed ICQC provides a better discrimination and helps to capture the artifacts present in the nonspeech regions of spoof speech signals.

By comparing the performance of ICQC (PCA) and ICQC-A from table 2 it is apparent that the using only double delta coefficients as a countermeasure instead of delta + double delta coefficients helps to reduce the EER further. This agrees well with the finding of [7]. The relative improvements achieved by the ICQC (PCA)-A over ICQC (PCA) on the *known*, *unknown*, *average* and a*ll* evaluation conditions are 50.0%, 19.5%, 24.2% and 14.1%, respectively. For spoofing detection task the dynamic coefficients (e.g., delta, acceleration or delta + acceleration) as countermeasure outperformed the static and combination of static + dynamic coefficients [7-8, 10]. The is because spoofing techniques focus on modeling smooth version (both temporal and spectral) of the natural speech. Smooth temporal structure means temporal dynamic is less and spectral details are missing in smooth spectral structure [14].

TABLE 2
Spoofing detection performance (in terms of EER) of the proposed ICQC features and comparison with other spoofing countermeasures on the ASVspoof2015 evaluation set. Results are reported on the *known* (average of S1-S5), *unknown* (average of S6-S10), *average* (average of S1-S10), and *all* condition. The lowest EERs are highlighted in bold face. The EERs of **CQCC [7]** and **CQCC-A [7]** are taken from [7].

	EER (%)			
	Known	***Unknown***	***Average***	***all***
LFCC	0.265	1.10	0.684	1.02
LPCC	0.262	1.10	0.684	0.987
MGDC	0.292	1.63	0.964	1.47
SFCC	0.362	1.18	0.773	0.96
PSCC	0.266	1.56	0.915	1.41
ICQC (PCA)	0.04	**0.286**	**0.165**	**0.227**
ICQC (PCA)-A	0.02	**0.23**	**0.125**	**0.195**
ICQC (DCT)	0.05	0.78	0.416	0.64
Bottleneck	**0.0**	0.865	0.432	0.58
Tandem	**0.0**	0.60	0.30	0.46
CQCC	0.15	0.79	0.472	0.80
CQCC [7]	0.0334	0.92	0.4768	
CQCC-A [7]	0.067	0.653	0.36	

Earlier work [1-3] on spoofing detection showed that phase related features perform better than amplitude-based features. This is because natural phase information is almost entirely lost in spoofed speech realized using voice conversion and speech synthesis approaches [1-3]. It is observed from the

ISBN 978-0-9928626-7-1 © EURASIP 2017

results of tables 2 that amplitude-based features (e.g., LFCC, CQCC, and IQCC) can provide better or at least comparable results to that of the phase related features. This is due to the presence of spoofing artifacts in the amplitude spectra of voice converted or synthesized spoof signals.

VI. CONCLUSION

In this work, we introduced a new anti-spoofing front end, namely infinite impulse response - constant Q-transform spectrum (IIR-CQT)-based cepstral coefficients (ICQC) extracted by filtering the fast Fourier transform of a speech signal with an infinite impulse response filter. Based on the discrete cosine transform and principal component analysis decorrelation techniques two variants of ICQC features were proposed. Since the proposed ICQC feature is based on IIR-CQT spectra, in addition to its simplicity it is also computationally efficient. We carried out standalone spoofing detection experiments using a GMM backend on the ASVspoof2015 challenge corpus. The proposed ICQC features (when principal component analysis is used for decorrelation instead of discrete cosine transform) demonstrated the best performance on all spoofing attacks (S1-S10) in both evaluation cases. Acceleration coefficients as anti-spoofing feature outperformed the delta + acceleration coefficients. The bottleneck and tandem features were very successful in detecting vocoded spoofing attacks (S1-S9) and provided an EER of almost zero on the average. Integration of auditory filterbank and removing non-speech frames were found to be detrimental. Features corresponding to frequency content greater than 4 kHz were found helpful for the spoofing detection. This is because most of the speech synthesis and voice conversion approaches focus on modeling low frequency content (up to 4 kHz) of speech precisely.

REFERENCES

[1] Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Zhizheng Wu, Federico Alegre and Phillip De Leon, "Speaker recognition anti-spoofing," in the *Handbook of Biometric Anti-spoofing*, Springer, S. Marcel, S. Li and M. Nixon, Eds., 2014.

[2] T. Kinnunen, Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4401-4404, 2012.

[3] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in Proc. IEEE Speaker and Language Recognition Workshop (Odyssey), pp. 151-158, 2010.

[4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, "ASVspoof 2015: the First ASV Spoofing and Countermeasures Challenge," in proc. of INTERSPEECH, 2015. http://www.spoofingchallenge.org/is2015_asvspoof.pdf

[5] Nanxin Chen, Yanmin Qian, Heinrich Dinkel, Bo Chen, Kai Yu, "Robust Deep Feature for Spoofing Detection - The SJTU System for ASVspoof 2015 Challenge", in proc. of Interspeech, 2015.

[6] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP) , 2013.

[7] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, Bilbao, Spain, 2016.

[8] Md. Jahangir Alam, Patrick Kenny, Vishwa Gupta and Themos Stafylakis, "Spoofing Detection on the ASVSpoof2015 Challenge Corpus Employing Deep Neural Networks," Proc. Odyssey Speaker and Language Recognition Workshop, Bilbao, Spain, June 2016.

[9] Md. Jahangir Alam, Patrick Kenny, Gautam Bhattacharya and Themos Stafylakis, "Development of CRIM System for the Automatic Speaker Verification Spoofing and Countermeasures Challenge 2015," Proc. Interspeech, Dresden, Germany, Sept. 2015.

[10] M. Sahidullah, T. Kinnunen, C. Hanilçi, "A Comparison of Features for Synthetic Speech Detection," Proc. of Interspeech, 2015.

[11] Tanvina B. Patel, Hemant A. Patil, "Combining Evidences from Mel Cepstral, Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs. Spoofed Speech", Interspeech 2015.

[12] S.K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel. On the vulnerability of speaker verification to realistic voice spoofing. In Proc. Int. Conf. On Biometrics: Theory, Applications and Systems (BTAS), 2015.

[13] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, Haizhou Li, "Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU Approach for ASVspoof 2015 Challenge", Interspeech 2015.

[14] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, Haizhou Li, "Spoofing detection from a feature representation perspective", in proc. of ICASSP, 2016.

[15] C. Hanilçi, T. Kinnunen, M. Sahidullah, A. Sizov, "Spoofing Detection Goes Noisy: An Analysis of Synthetic Speech Detection in the Presence of Additive Noise", Speech Communication, vol. 85, pp. 83-97, December 2016.

[16] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, Haizhou Li, "An investigation of spoofing speech detection under additive noise and reverberant conditions", in proc. of INTERSPEECH, 2016.

[17] Pavel Korshunov and Sébastien Marcel, "Cross-database evaluation of audio-based spoofing detection systems," proc. of Interspeech, San Francisco, USA, 2016.

[18] Dipjyoti Paul, Monisankha Pal, Goutam Saha, "Novel speech features for improved detection of spoofing attacks," https://arxiv.org/pdf/1603.04264.pdf, 2016.

[19] T. Patel and H. Patil, "Effectiveness of fundamental frequency (F0) and strength of excitation (SOE) for spoofed speech detection," in Proc. ICASSP, 2016.

[20] K. Dressler, "Sinusoidal Extraction Using and Efficient Implementation of a Multi-Resolution FFT," in Proc. of the DAFx, Montreal, Canada, 2006.

[21] P. Cancela, M. Rocamora, E. Lopez, "An efficient multi-resolution spectral transform for music analysis," in proc. of the ISMIR, 2009.

[22] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," JASA, vol. 92, no. 5, pp. 2698–2701, 1992.

[23] Md. Jahangir Alam, Vishwa Gupta, and Patrick Kenny, "CRIM's Speech Recognition System for the 4th CHIME Challenge," Proc. of 4th CHIME Challenge, pp. 63-67, San Francisco, CA, September 2016. http://spandh.dcs.shef.ac.uk/chime_workshop/chime2016proceedings.pdf

[24] The 4th CHiME speech separation and recognition challenge, 2016.

[25] H. Delgado, M. Todisco, M. Sahidullah, A. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan; "Further optimizations of constant Q cepstral processing for integrated utterance verification and text-dependent speaker verification"; in proc. of IEEE workshop on Spoken Language Technology (to appear), San Diego, CA, December 2016.

[26] Md. Jahangir Alam, Patrick Kenny, and Vishwa Gupta, "Tandem features for text-dependent speaker verification on the RedDots corpus," *Proc. Interspeech,* San Francisco, August 2016.

[27] H. Murthy and V. Gadde. The modified group delay function and its application to phoneme recognition. In Proc. of ICASSP, vol. 1, p. 68–71, 2003.

[28] J. Brown, "Calculation of a constant Q spectral transform," The Journal of the Acoustical Society of America, vol. 89,no. 1, pp. 425–434, 1991.

ISBN 978-0-9928626-7-1 © EURASIP 2017