Detection of Replay Attacks using Single Frequency Filtering Cepstral Coefficients

by

Raju Alluri K.N.R.K, Sivanand a, sudarsanareddy.kadiri@research.iiit.ac.in kadiri, Suryakanth V Gangashetty, Anil Kumar Vuppala

in

Interspeech 2017
(*Interspeech 2017*)

Stockholm, Sweden

Report No: IIIT/TR/2017/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
August 2017

Detection of Replay Attacks using Single Frequency Filtering Cepstral Coefficients

K N R K Raju Alluri, Sivanand Achanta, Sudarsana Reddy Kadiri, Suryakanth V Gangashetty, and Anil Kumar Vuppala

Speech Processing Laboratory, KCIS
International Institute of Information Technology, Hyderabad, India
{raju.alluri, sivanand.a, sudarsanareddy.kadiri}@research.iiit.ac.in,
{svg, anil.vuppala}@iiit.ac.in

Abstract

Automatic speaker verification systems are more vulnerable to spoofing attacks. Recently, various countermeasures have been developed for detecting high technology attacks such as speech synthesis and voice conversion. However, there is a wide gap in dealing with replay attacks. In this paper, we propose a new feature for replay attack detection based on single frequency filtering (SFF), which provides high temporal and spectral resolution at each instant. Single frequency filtering cepstral coefficients (SFFCC) with Gaussian mixture model classifier is used for the experimentation on the standard BTAS-2016 corpus. The previously reported best result, which is based on constant Q cepstral coefficients (CQCC) has achieved a half total error rate of 0.67 % on this data-set. Our proposed method outperforms the state of the art (CQCC) with a half total error rate of 0.0002 %.
Index Terms: Replay attack, countermeasures, Gaussian mixture model, single frequency filtering cepstral coefficients.

1. Introduction

Recently developed automatic speaker verification (ASV) systems based on i-vectors [1] and joint factor analysis [2] have achieved a great performance and these systems are less prone to noise and channel effects. These characteristics of ASV systems makes them to ready for mass market adoption in biometric applications like banking transactions, e-commerce, police investigation, and so on [3, 4]. On the other hand, there are concerns about the vulnerability of ASV technology to spoofing attacks [5]. Spoofing attack refers to an attack by a fraudster, who wants to gain the access of authorized user by providing the fake voice samples. In the literature, four types of spoofing attacks were considered [3, 5]. They are impersonation, speech synthesis, voice conversion, and replay.

In impersonation, an attacker will try to mimic the target speaker voice without any help of speech technology. From the studies on impersonation attacks, it was found that there is not much effect on speaker verification performance [3]. In Speech synthesis (SS) and voice conversion (VC) based attacks, the attacker produces the authorized user speech by using different SS and VC techniques. The SS and VC based attacks are known as high technology spoofing attacks. These attacks are extensively studied in ASV spoof 2015 challenge [6] by introducing a standard corpus and protocol to evaluate different countermeasures [7, 8, 9, 10, 11, 12, 13, 14, 15], developed by researchers across the globe on the same platform to provide a generalized solution to spoofing attacks. In the replay attack, the attacker uses a pre-recorded voice collected from genuine target speaker and try to access the speaker verification system. Replay is a simple spoofing attack, which does not require any specific knowledge of speech processing. Availability of low cost and high quality recording devices made replay attack as a significant threat to ASV technology. There are a few studies [16, 17, 18, 19] on replay attack detection in the literature. These studies are based on either comparing the new access voices with previously stored voices or studies based on the channel noise characteristics. These studies are performed on personalized databases, which are publicly not available. In biometrics theory applications and systems (BTAS) 2016, a speaker anti-spoofing challenge [20] was announced to detect replay attacks by introducing a standard BTAS 2016 corpus. Several researchers have addressed their countermeasures [20, 21, 22, 23] for replay attack detection on this corpus. The constant Q cepstral coefficients (CQCC) [24] based anti-spoofing system has achieved good performance for SS [24], VC [24], and replay attacks [23]. In this paper, the focus is on replay attack detection and the proposed countermeasures will be evaluated on BTAS 2016 corpus.

Most of the studies on replay attack use block processing of speech. Hence, they can not capture the instantaneous spectral changes within the analysis block. Also, most of the successful methods for spoof detection either captures the information present in low-frequency or high-frequency regions [13, 20] with the trade off of temporal resolution. Recently proposed countermeasures for ASV spoofing based on CQCC [22, 23, 24] provides a higher frequency resolution at low frequencies and higher temporal resolution at higher frequencies but it lacks in capturing instantaneous spectral changes within analysis block at lower frequencies. In block processing, the characteristics of low SNR instances will get averaged because of high SNR instants present in that block. Our intuition is that a feature representation that can capture instantaneous spectral changes with high spectral and temporal resolution may be a possible solution for replay attack detection. Motivated with this, in this paper we investigated the single frequency filtering (SFF) [25] method of speech analysis, which provides high spectral and temporal resolution. A new feature is proposed by performing cepstral analysis on SFF spectrum to capture instantaneous spectral changes, this feature is named as single frequency filtering cepstral coefficients (SFFCC). In this study, The SFFCC along with Gaussian mixture model is used as a countermeasure for replay attacks.

The remainder of the paper is organized as follows. Section 2 describes the prior works on BTAS 2016 corpus. In Section 3, extraction of SFFCC features is presented. Section 4 describes the experimental setup. In Section 5, results and discussion are presented. Finally, Section 6 provides the conclusions.

2. Prior Works

This section briefly reviews the spoofing countermeasures provided by other researchers on BTAS 2016 corpus. In BTAS 2016 speaker anti-spoofing challenge [20], four teams have submitted their results. The baseline system provided in the challenge uses simple spectrogram based ratios as features and logistic regression as a classifier. The submission by CPqD team used two types of cepstral coefficients as features and deep neural networks (DNNs) as a classifier. CPqD team also used ASVspoof [6] data to enhance the system performance. The submission by SJTUSpeech team used cepstral mean-variance normalization (CMVN) with normalized perceptual linear predictive (PLP) features. This team used two classifiers, a seven layer DNN and four layer bidirectional long short-term (BLSTM). The submission by Idiap team used long-term spectral mean and standard deviation as features. The feature vectors are classified using a linear discriminant analysis (LDA) classifier. The submission by IITKGP-ABSP team is based on score level fusion of mel frequency cepstral coefficients (MFCCs) and IMFCCs using a standard Gaussian mixture model (GMM) classifier. Most of these studies use block processing with different frame sizes (20 - 40 ms) and a common frame shift of 10 ms. Recently in [22, 23], the CQCC features are investigated on BTAS 2016 corpus. From the study [22], it was found that that static CQCC features are performing well on BTAS 2016 corpus, the results are reported in terms of equal error rate (EER). In [23], authors used CQCC static features appended with delta and accelerated coefficients as features and the results are reported in terms of half total error rate (HTER) as per the challenge convention. These two studies [22, 23] used GMM as a classifier. To the best of authors knowledge, the result reported in [23] is the state of the art on BTAS 2016 corpus.

3. Extraction of SFFCC Features

In this study, the features are extracted from recently proposed single frequency filtering (SFF) [25] method of speech analysis. The main objective of SFF is to compute the amplitude envelope of the signal as a function of time. The spectral and temporal resolutions can be adjusted by varying the parameter r, which represents the pole location. Cepstral features are extracted from the SFF envelopes. The block diagram of SFFCC extraction is described in Figure 1. The steps involved in the extraction of SFFCC is as follows [26].

1. Pre-emphasis the input speech signal $s[n]$ to remove any low frequency components introduced during recording.

$$x[n] = s[n] - s[n-1] \tag{1}$$

   where n ranges from 1 to N, N is total number of samples in the signal.

2. The signal $(x[n])$ is multiplied with a complex sinusoidal $e^{j\bar{w_k}n}$, where $\bar{w_k} = \pi - w_k = \pi - \frac{2\pi f_k}{f_s}$, in-order to shift the frequency spectrum $X(w)$ of the signal $x[n]$. The resulting frequency shifted signal is represented by x[k,n], where x[k,n] is

$$x[k,n] = x[n]e^{j\bar{w_k}n} \tag{2}$$

   where k ranges from 0 to K, K is $f_s/2$.

3. The frequency shifted signal x[k,n] is passed through a single-pole filter whose transfer function is *H(z)*, where

$$H(z) = \frac{1}{1 + rz^{-1}} \tag{3}$$

   The root at $z = -r$ in the z-plane is set such that it corresponds to $f_s/2$. For filter to be stable, the location of pole should be near to unit circle. In this study r value is chosen as 0.995.

4. The output $y[k,n]$ of the filter is given by

$$y[k,n] = -ry[k,n-1] + x[k,n] \tag{4}$$

5. The amplitude envelope of the signal $y[k,n]$ is given by

$$v[k,n] = \sqrt{Re(y[k,n])^2 + Im(y[k,n])^2} \tag{5}$$

   where Re, Im represents the real and imaginary parts respectively. The term v[k,n] corresponds to the SFF envelope of the signal at a desired frequency f_k. The magnitude spectrum can be obtained from SFF envelope for each instant of n.

6. Cepstrum c[k,n] is computed from SFF spectrum v[k,n] as follows

$$c[k,n] = IFFT(log(v[k,n])) \tag{6}$$

   From c[k,n], first few cepstral coefficients (p) are considered. In this study, they are named as SFFCC.

4. Experimental Setup

4.1. Database

In this study, BTAS 2016 corpus [20] is considered, which is a subset of AVspoof [1] [27]. This corpus contains three nonoverlapping subsets: train, development, and test. Each subset is further divided into two main parts: (i) genuine data, (ii) different replay attacks. Training and development data contains the similar type of attacks and in the test data, there are two unknown attacks. The number of utterances in each type of attack is given in Table 1. Detailed information of the database can be found in [20].

Table 1: *Description of BTAS 2016 Corpus. RE stands for replay. LP for laptop, HQ means high quality speakers used during replay, PH1 is samsung Galaxy S4 phone, PH2 is iphone 3GS. PH3 is iphone 6S, VC means voice conversion and SS means speech synthesis.*

Data type		# Train	# Dev	# Test
Genuine		4973	4995	5576
RE-LP-LP	R1	700	700	800
RE-LP-HQ-LP	R2	700	700	800
RE-PH1-LP	R3	700	700	800
RE-PH2-LP	R4	700	700	800
SS-LP-LP	R5	490	490	560
SS-LP-HQ-LP	R6	490	490	560
VC-LP-LP	R7	17400	17400	19500
VC-LP-HQ-LP	R8	17400	17400	19500
RE-PH2-PH3	R9	-	-	800
RE-LPPH2-PH3	R10	-	-	800
All attacks		38580	38580	44920

4.2. Parameters used for Feature Extraction

The amplitude envelope of the signal can be computed at any frequency f_k using SFF. In this study, 513 frequencies are considered within the frequency range of 0 to f_s/2 , where fs = 16000 Hz with the spacing of 15.6 Hz. The SFFCC can be obtained at each time instant. In this study, instead of computing at each instant, we computed the SFFCC at selected instants as described below.

[1] https://www.idiap.ch/dataset/avspoof

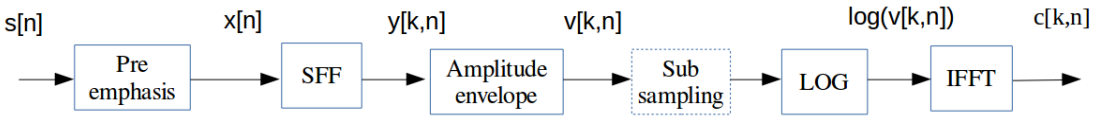

Figure 1: *Block diagram of single frequency filter cepstral coefficients extraction.*

From each 10 ms segment, one instant is considered. This process is named as sub-sampling as shown in Figure 1 (as it is an optional step, dotted lines are used as borders). In this study, the choice of instants for SFFCC extraction are selected based on instantaneous energy. The sum of all the values across frequency in v[k,n] will give instantaneous energy. The instantaneous energy for a speech segment (Figure 2 (a)) is shown in Figure 2 (c). The equation for instantaneous energy E[n] is given below [28],

$$E[n] = \sum_{k=0}^{K} v[k, n]. \tag{7}$$

Three variants of instants were chosen for every 10 ms. They are: (i) At each 10 ms instant, (ii) Lowest energy (low SNR) sample in each 10 ms segment, and (iii) Highest energy (high SNR) sample in each 10 ms segment. The low and high SNR samples are selected by tracing the instantaneous energy at which there is low amplitude and high amplitude, respectively. These three variants of samples are shown in Figure 2 (c) as "|" ,"□", and "∗" respectively. From Figure 2 (b) and 2(c) we can correlate the peaks in E[n] with vertical lines (corresponds to glottal closure instants [26]) in spectrogram which represents high SNR regions. It can also be observed that the selected low SNR instants represent the low SNR regions in the spectrogram. Cepstral coefficients are computed from these three

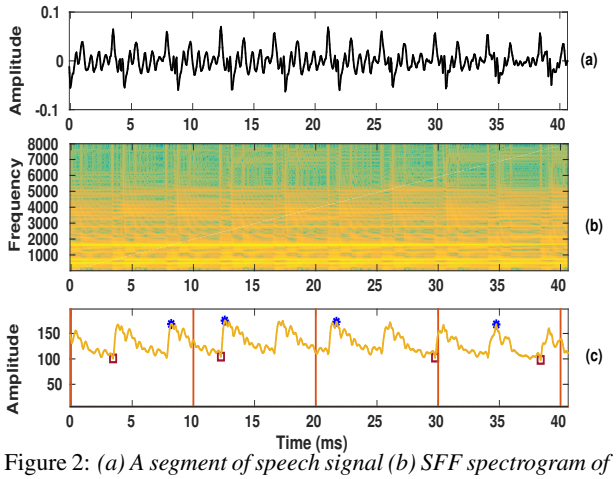

Figure 2: *(a) A segment of speech signal (b) SFF spectrogram of (a). (c) Instantaneous energy of (a) with marked 10 ms instants (|), low (□) and high (∗) SNR samples for each 10 ms segment.*

variants separately. The three variants of SFFCC are explained below by considering the j^{th} speech segment. The instantaneous energy for j^{th} speech segment is represented by $E_j[n]$.

- $SFFCC_{10ms}$: In this, the cepstral coefficients are obtained for the integral multiples of 10 ms instants.
- $SFFCC_{min}$: In this, the cepstral coefficients are computed at the low SNR instants within 10 ms segment. The low SNR instant for each segment is represented by l. where l is

$$l_j = \arg\min_i E_j[i] \tag{8}$$

- $SFFCC_{max}$: In this, the cepstral coefficients are computed at the high SNR instants within 10 ms segment. The high SNR index for each segment is represented by l. where l is

$$l_j = \arg\max_i E_j[i] \tag{9}$$

In this study, different dimensional cepstral coefficients (p = 13, 20 and 30) are considered. From static (S) coefficients, delta (D) and double-delta (A) coefficients are computed. Experiments are performed with different combinations of static and dynamic coefficients.

4.3. Classifier

In this study, GMM is used as a classifier. After the feature extraction, two GMMs are built for each genuine (λ_g) and spoof (λ_s) with 512 mixture components. GMM parameters are estimated with expectation and maximization (EM) algorithm

The log-likelihood score is computed with the following equation,

$$Score(X) = llk(X|\lambda_g) - llk(X|\lambda_s) \tag{10}$$

where X = $\{x_1, x_2, \ldots, x_T\}$ is the feature vector of test utterance, T is the number of sub-sampling instances. where,

$$llk(X|\lambda) = (1/T) \sum_{t=1}^{T} logp(x_t/\lambda) \tag{11}$$

is the average likelihood of X given model λ.

4.4. Evaluation Metrics

The evaluation metrics were considered according to the protocol used in BTAS 2016 speaker anti-spoofing challenge. The results on development data are reported in terms of EER and on the test data in terms of HTER. As the studies in [22] used EER as a metric on both development and test data, EER values are computed for test data. All the EER values were computed with BOSARIS toolkit [29].

5. Results and Discussion

Initial studies are conducted on BTAS test data by using the three variants of SFFCC. The results are reported in Table 2.

Table 2: *Evaluation results (EER in %) for BTAS 2016 test data set using different variations of SFFCC.*

Feature	$SFFCC_{10ms}$	$SFFCC_{min}$	$SFFCC_{max}$
EER	1.02	0.07	1.14

From the results in Table 2, it is evident that $SFFCC_{min}$ is performing better than other two variants. These results suggest us that, $SFFCC_{min}$ which are extracted from low SNR instants is capturing the channel variations effectively than the $SFFCC_{max}$ and $SFFCC_{10ms}$, which are extracted from high SNR instants and for each 10 ms instants respectively. Further experiments are conducted with different feature dimensions of $SFFCC_{min}$, the results are reported in Table 3.

Table 3: *Evaluation results (EER in %) for BTAS 2016 test data set using different dimensions of $SFFCC_{min}$. SDA refers to static appended with dynamic coefficients*

Feature	13-SDA	20-SDA	30-SDA
$SFFCC_{min}$	0.11	0.09	0.07

Table 4: *Individual attack results (in % HTER) of different systems on BTAS test data set.*

	Known attacks									Unknown attacks			All attacks
System	R1	R2	R3	R4	R5	R6	R7	R8	Pool	R9	R10	Pool	Pool
SJTUSpeech [20]	10.34	10.02	1.52	2.05	1.88	1.75	1.73	1.81	2.08	2.84	18.09	10.46	2.20
Idiap [20]	15.83	0.58	0.33	25.18	0.27	0.27	0.33	0.27	1.05	50.08	46.64	48.36	2.04
IITKGP-ABSP [20]	8.58	1.81	0.68	3.59	0.68	0.68	0.74	0.81	0.98	6.49	23.06	14.75	1.26
CQCC-SDA [23]	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	23.92	12.10	0.67
SFFCC-SDA	0.16	**0.04**	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	**0.53**	**0.29**	0.05
SFFCC-DA	14.56	5.46	48.01	51.41	3.20	3.20	3.01	3.01	5.11	48.12	19.36	33.74	5.96
SFFCC-S	**0.01**	0.06	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**	1.19	0.60	**0.0002**

Table 5: *Individual attack results (in % EER) for CQCC and SFFCC features on BTAS test data set.*

	Known attacks									Unknown attacks			All attacks
System	R1	R2	R3	R4	R5	R6	R7	R8	Avg	R9	R10	Avg	Avg
CQCC-S[22]	**0**	0.04	0	0	0	0	0	0	0.005	7.56	**0**	3.78	0.76
SFFCC-SDA	0.17	0.06	0	0	0	0	0	0	0.029	0.04	0.48	0.26	0.07
SFFCC-DA	11.88	5.25	39.10	47.31	1.76	1.28	0.19	0.24	13.37	47.73	19.17	33.45	17.39
SFFCC-S	0.01	**0.02**	0	0	0	0	0	0	**0.004**	**0.01**	0.44	**0.23**	**0.05**

From the results in Table 3, 30-SDA features are performing better than the lower dimensional coefficients (13-SDA, 20-SDA). These results suggest us that, higher order coefficients are useful in spoof detection. In the rest of the paper, experiments will be conducted with 30 dimensional $SFFCC_{min}$. From here onwards the terms $SFFCC_{min}$ and SFFCC are used interchangeably.

The proposed system results are compared with the top five performing systems reported in the literature on BTAS 2016 corpus [20]. The results of SJTUSpeech, Idiap, and IITKGP-ABSP systems are taken from BTAS speaker anti-spoofing challenge [20]. CQCC-SDA system is taken from [23] and CQCC-S from [22]. Our proposed feature with different combinations of static and dynamic coefficients are used in this study. The evaluation results in terms of EER on development data and in terms of HTER on test data are reported in Table 6. From

Table 6: *Evaluation results (error rates in %) of different systems on BTAS 2016 development and test data sets.*

System	Developmet (EER)	Test (HTER)
SJTUSpeech[20]	0.42	2.20
Idiap[20]	0	2.04
IITKGP-ABSP[20]	0	1.26
CQCC-SDA[23]	0	0.67
CQCC-S[22]	0	-
SFFCC-SDA	0	0.05
SFFCC-DA	3.72	5.96
SFFCC-S	0	**0.0002**

the results in Table 6, it can be observed that many of the systems are perfectly tuned to development data (column 2 Development (EER)) by attaining an EER of 0 but the results on test data (column 3 Test (HTER)) are varying across systems. Further analysis is carried out on test data and the individual attack results are reported in Table 4. From the results in Table 4, it can be seen that some of the systems are performing better on specific attacks, for example, Idiap system is performing better for some of the known attacks such as R5, R6, and R8, but it is performing poorly on unknown attacks R9 and R10. CQCC-SDA system is relatively performing better than all the three systems reported in [20]. While dealing with unknown attacks, CQCC-SDA is successful with R9 but it is poorly detecting the R10. The success of CQCC-SDA is because of its unique property of high spectral resolution at lower frequencies and high temporal resolution at higher frequencies. Static SFFCC based system is the best performing system for many of attacks except for R2 and R10. SFFCC-SDA system is detecting R2 and R10 attacks more accurately than the SFFCC-S based system. From the pooled results of known and unknown attacks, it can be seen that for known attacks, SFFCC-S has achieved pooled HTER of 0.01 which is better than the CQCC-SDA based system by a large margin and four times better than the SFFCC-SDA based system. Whereas in the case of unknown attacks SFFCC-SDA system is best performing than the CQCC-SDA based system and SFFCC-S based system.

Similar to the studies in [22], threshold-free EER is computed for test data and the individual attack results are reported in Table 5. From the results in Table 5, the CQCC-S system performed well in many attacks in test data set, for R1 and R10 the CQCC-S system is performing better than the proposed SFFCC based system but in the case of R9 which is an unknown attack, it is not performing well. Whereas proposed system has a low error rate in this case. The overall performance of proposed system is more superior than a CQCC-S system.

From the results reported on SFFCC-DA based system in Table 4 and Table 5, it can be observed that the dynamic coefficients alone are unable to detect the replay attacks efficiently, whereas relatively they are detecting VC and SS based replay attacks (R5 - R8) than the direct replay attacks. The reason for the success of dynamic coefficients in detecting VC and SS based attacks may be because many synthetic speech generating techniques do not use long-term dynamics of speech effectively. Even though static features are able to detect known attacks efficiently, for generalization it is better to use static appended with dynamic coefficients

6. Summary and Conclusions

In this paper, a new feature based on single frequency filtering (SFF) method of speech analysis is proposed for replay attack detection. Cepstral features extracted from low SNR instants of SFF spectrum for each 10 ms segments are more useful than the features extracted from high SNR and 10 ms instants for replay attack detection. Experimental results on BTAS 2016 corpus shows that the proposed system outperforms the state of the art CQCC system with a significant margin. Based on these encouraging results on replay attacks, in future, the proposed features can be explored on other attacks such as VC and SS inorder to provide a generalized solution for spoofing attacks.

7. Acknowledgements

The first author would like to thank the Department of Electronics and Information Technology, Ministry of Communication & IT, Govt of India for granting PhD Fellowship under Visvesvaraya PhD Scheme. The second and third authors would like to thank Tata Consultancy Services (TCS), India for supporting their PhD program.

8. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification,," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification,," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey,," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[4] E. Vale and A. Alcaim, "Adaptive weighting of subband-classifier responses for robust text-independent speaker recognition,," *Electronics Letters*, vol. 44, no. 21, pp. 1280–1282, 2008.

[5] N. Evans, J. Yamagishi, and T. Kinnunen, "Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics,," *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*, 2013.

[6] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,," in *Proc. INTERSPEECH*, 2015, pp. 2037–2041.

[7] X. Xiao, X. Tian, S. Du, H. Xu, E. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge,," in *Proc. INTERSPEECH*, 2015, pp. 2052–2056.

[8] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015,," in *Proc. INTERSPEECH*, 2015, pp. 2072–2076.

[9] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, and V. Shchemelinin, "STC anti-spoofing systems for the ASVspoof 2015 challenge,," in *Proc. ICASSP*, 2016, pp. 5475–5479.

[10] J. Villalba, A. Miguel, A. Ortega, and E. Lleida, "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge,," in *Proc. INTERSPEECH*, 2015, pp. 2067–2071.

[11] A. Janicki, "Spoofing countermeasure based on analysis of linear prediction error,," in *Proc. INTERSPEECH*, 2015, pp. 2077–2081.

[12] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection-the SJTU system for ASVspoof 2015 challenge,," in *Proc. INTERSPEECH*, 2015, pp. 2097–2101.

[13] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: a comparison,," in *Proc. INTERSPEECH*, 2015, pp. 2057–2061.

[14] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech,," in *Proc. INTERSPEECH*, 2015, pp. 2092–2096.

[15] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech,," in *Proc. INTERSPEECH*, 2015, pp. 2062–2066.

[16] W. Shang and M. Stevenson, "Score normalization in playback attack detection,," in *Proc. ICASSP*, 2010, pp. 1678–1681.

[17] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems,," in *Proc. European Workshop on Biometrics and Identity Management*, 2011, pp. 274–285.

[18] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition,," in *Proc. ICMLC*, vol. 4, 2011, pp. 1708–1713.

[19] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification,," in *Proc. APSIPA*, 2014, pp. 1–5.

[20] P. Korshunov, S. Marcel, H. Muckenhirn, A. Gonçalves, A. S. Mello, R. V. Violato, F. Simoes, M. Neto, M. de Assis Angeloni, J. Stuchi *et al.*, "Overview of BTAS 2016 speaker anti-spoofing competition,," in *Proc. BTAS*, 2016, pp. 1–6.

[21] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Proc. INTERSPEECH*, 2016, pp. 1705–1709.

[22] D. Paul, M. Sahidullah and G. Saha, "Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora," in *Proc. ICASSP*, 2017, pp. 2047–2051.

[23] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A Spoofing Countermeasure for Automatic Speaker Verification,," *Computer Speech and Language*, 2017.

[24] ——, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[25] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech,," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[26] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach,," *Speech Communication*, vol. 86, pp. 52–63, 2017.

[27] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing,," in *Proc. BTAS*, 2015, pp. 1–6.

[28] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," *Proc. INTERSPEECH*, pp. 2155–2159, 2016.

[29] N. Brümmer and E. de Villiers, "The BOSARIS Toolkit: Theory, Algorithms and Code for Surviving the New DCF,," *arXiv preprint arXiv:1304.2865*, 2013.