INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Long Range Acoustic Features for Spoofed Speech Detection

Rohan Kumar Das, Jichen Yang, Haizhou Li

Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
{rohankd, eleyji, haizhou.li}@nus.edu.sg

Abstract

Speaker verification systems in practice are vulnerable to spoofing attacks. The high quality recording and playback devices make replay attack a real threat to speaker verification. Additionally, the furtherance in voice conversion and speech synthesis has produced perceptually natural sounding speech. The ASVspoof 2019 challenge is organized to study the robustness of countermeasures against such attacks, which cover two common modes of attacks, logical and physical access. The former deals with synthetic attacks arising from voice conversion and text-to-speech techniques, whereas the latter deals with replay attacks. In this work, we explore several novel countermeasures based on long range acoustic features that are found to be effective for spoofing attack detection. The long range features capture different aspects of long range information as they are computed from subbands and octave power spectrum in contrast to the conventional way from linear power spectrum. These novel features are combined with the other known features for improved detection of spoofing attacks. We obtain a tandem detection cost function of 0.1264 and 0.1381 (equal error rate 4.13% and 5.95%) for logical and physical access on the best combined system submitted to the challenge.

Index Terms: anti-spoofing, logical access, physical access, ASVspoof 2019 challenge

1. Introduction

The progress in automatic speaker verification (ASV) research has led to many real-world deployments [1–3]. Security of ASV system has become an important topic, therefore research on anti-spoofing countermeasures has attracted much attention recently [4, 5]. Broadly these attacks are grouped under four categories, which are impersonation, voice conversion (VC), text-to-speech (TTS) and replay attacks [6]. Among these categories, impersonation is a behavioral attack and less vulnerable compared to the other kinds of attacks.

The inception of ASVspoof[1] series of challenges took place to spearhead the research on anti-spoofing countermeasures [7, 8]. The previous edition of challenges focused on synthetic and replay attacks in the year 2015 and 2017, respectively [9, 10]. Another challenge, BTAS 2016 was organized considering both synthetic and replay attacks [11]. However, the recent advancements in VC and TTS have produced perceptually very natural sounding speech, which prompts us to look into their threat for spoofing [12, 13]. Further, consideration of uncontrolled replay configurations in ASVspoof 2017 challenge made the results difficult to analyze [14, 15]. To gain better insights into the research problem, the current ASVspoof 2019 challenge focuses on the synthetic speech generated from recent state-of-the-art VC and TTS systems along with replay attacks simulated by various replay devices under controlled acoustic conditions.

ASVspoof 2019 is different from the previous editions. Firstly, it runs two tracks, namely, logical access and physical access [14]. The former deals with synthetic attacks, whereas the latter deals with replay attacks. Secondly, the challenge introduces a new performance metric that measures the integrated performance of spoofing detection and ASV [14,16]. Two baseline systems are released by the organizers as a part of the challenge. The first one is developed using popular constant-Q cepstral coefficient (CQCC) with Gaussian mixture model (GMM) as back-end [17,18]. Another system based on linear frequency cepstral coefficient (LFCC) with GMM is reported as a contrast system for the challenge [19].

As synthetic speech and replay are considered separately in the ASVspoof evaluations, there have been studies on anti-spoofing countermeasures for them individually. In practice, we are not informed of which type of attacks is coming our way. It would be interesting to study acoustic features that are applicable for both type of attacks. The CQCC feature derived from long-term constant-Q transform (CQT) remains as a strong baseline for both type of attacks [18, 20]. The importance of long range features is also supported by effectiveness of cochlear filter cepstral coefficients and instantaneous frequency feature that produced the best performance in the first edition of anti-spoofing challenge on synthetic speech attacks [21]. Features with complementary information and their fusion have shown promising results in previous explorations on synthetic attacks [22, 23]. Similarly, several works on ASVspoof 2017 challenge have proved the importance of alternative features having different discriminating characteristics can be useful for detection of replay attacks [24–28].

In this work, we would like to investigate several novel countermeasures that are applicable for both tracks, logical and physical access, in ASVspoof 2019 challenge. We focus on long range acoustic features extracted from CQT domain as such information has been found to be useful for capturing the artifacts for spoof detection [17, 18]. Features derived from subband information and octave power spectrum of CQT are two another directions that we investigated. We consider a total of 10 different front-ends for both tracks of ASVspoof 2019 challenge. Further, each front-end is evaluated with GMM and deep neural network (DNN) based two different back-ends to have 20 subsystems for each track. Finally, we combine different aspects of long range features and other well-known features to have an effective detection of spoofing attacks.

The rest of the paper is organized as follows. Section 2 mentions about the different long range acoustic features explored for spoof detection. In Section 3, we present a summary of system description. Section 4 details the experiments and Section 5 reports the results with discussion. Finally, Section 6 provides the conclusion to the work.

[1] http://www.asvspoof.org/

http://dx.doi.org/10.21437/Interspeech.2019-1887

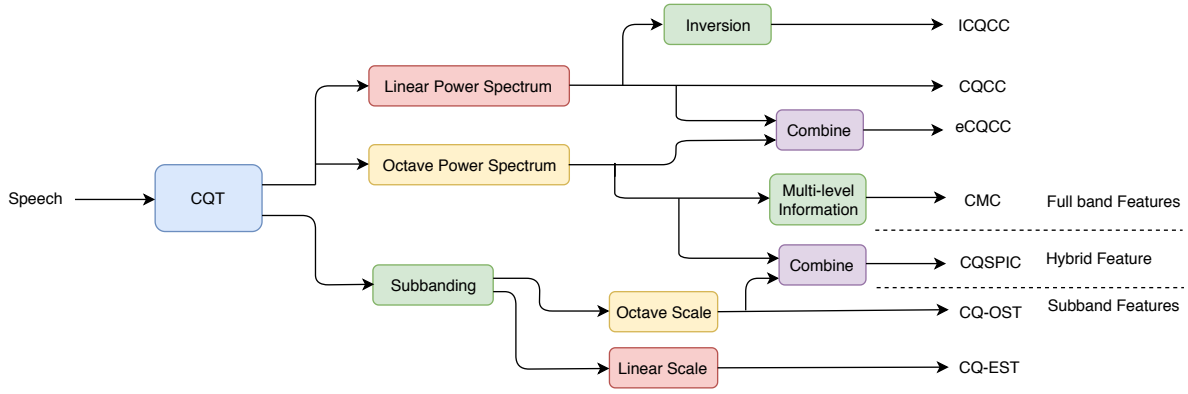

Figure 1: *Block diagram showing CQT based different long range acoustic features.*

2. Long Range Acoustic Features

The conventional acoustic features are derived using short-term processing of speech signal. A window of few tens of milliseconds is considered to derive such features regardless of the length of the signal. However, this may not be the best way to detect spoofed speech as the temporal signal variations over a long range can be more reliable to discriminate from natural speech. The perceptually motivated CQT is a long-term window transform to capture signal characteristics. Unlike the fixed time-frequency resolution of Fourier transform methods, it can give a higher frequency resolution for lower frequencies along with a higher temporal resolution for higher frequencies [17, 18]. Further, in case of CQT, the centre frequencies of each filter and the octaves are geometrically distributed. We refer the artifacts from signal captured in such a manner as long range information due to the consideration of better time-frequency resolution compared to short-term processing.

In this work, we make use of long-term CQT to explore several novel countermeasures for identifying spoofing attacks. Figure 1 illustrates the relationship among various long range acoustic features and their different perspectives. We group them based on their nature to capture frequency content from the spectrum. The three categories are full band, subband and hybrid features. Next, we discuss the details for each of them.

2.1. CQCC

The CQCC feature is undoubtedly one of the promising features for detection of spoofing attacks [17, 18]. It has been found that consideration of energy coefficient with CQCC can improve the performance [20]. In this regard, we consider the magnitude and phase of energy (MPE) computed from constant-Q domain and then append as additional feature dimensions [29].

2.2. CQ-EST

We have recently proposed a novel subband feature, namely, constant-Q equal subband transform (CQ-EST) [30] that is extracted considering subbands over linear scale. We compute MPE and append as additional feature dimensions.

2.3. CQ-OST

Another subband feature, constant-Q octave subband transform (CQ-OST) extracted from subbands over octave scale is explored [30]. Further, variance as a short-term spectral statistics information (STSSI) is also considered along with MPE and then appended as additional feature dimensions [31].

2.4. eCQCC

This feature refers to extended CQCC (eCQCC) that is derived by combining the coefficients from octave power spectrum with that of conventional CQCC feature obtained from linear power spectrum [32]. Additionally, we use MPE and then append to the feature dimensions.

2.5. CQSPIC

The constant-Q statistics-plus-principal information coefficient (CQSPIC) is obtained using extended coefficients of eCQCC features (coefficients generated from octave power spectrum) and SSTI with CQ-OST feature [31]. The MPE is also considered for appending to the feature dimensions.

2.6. ICQCC

Another long range feature, inverted constant-Q cepstral coefficient (ICQCC) is considered [33]. It is extracted from inverted linear power spectrum from CQT domain. We note that STSSI and MPE are used and appended to the feature dimensions.

2.7. CMC

The constant-Q multi-level coefficient (CMC) features that capture the multiple levels of information from octave power spectrum in CQT domain are found to be effective as investigated in [29]. This is another long range information based front-end.

3. System Description

In our studies, apart from the long range acoustic features based on CQT, we considered three another front-ends as contrast systems. Further, two different modeling techniques are used as the back-ends in combination with all the front-ends as summarized in Table 1 and 2. Next, we brief regarding the contrast front-ends and the two back-ends.

3.1. Contrast Front-ends

Just like LFCC serves as a contrast system to CQCC system for the challenge [19], we consider instantaneous frequency cosine coefficient (IFCC) based phase features as a contrast front-end due to their success in previous studies [25, 34]. Further, widely used mel frequency cepstral coefficient (MFCC) features are also used in our studies [35].

3.2. Back-ends

The GMM based back-end is referenced as a part of the given baseline system for ASVspoof 2019 challenge. Two GMMs are trained for bonafide and spoofed speech examples of the train subset for both tracks. Given a test speech, its likelihoods with respect to bonafide and spoofed models are computed to obtain the log-likelihood ratio score. With the advancements of deep learning techniques, they have been successfully applied to various domains. In this regard, we use DNN based system as another back-end for the studies. The architecture basically follows the one mentioned in [32]. The dimension of input features decides the number of input nodes to the DNN, whereas there are two nodes in the output layer as it is a binary task.

4. Experiments

This section describes the details of the ASVspoof 2019 database and the experimental setup.

4.1. Database

The ASVspoof 2019 database contains both synthetic and replay speech attacks that are categorized as logical access and physical access attacks, respectively. There are three subsets under these two tracks, namely, train, development and evaluation set. The genuine speech data for both tracks are collected from a population of 107 speakers consisting of 46 male and 61 female speakers from VCTK[2] database. Two recent VC methods and four TTS algorithms are used to create the spoofed speech for train and development subset for logical access. On the other hand, the evaluation set examples are created with diverse unseen spoofing algorithms.

The train and development subset of physical access attacks are created with simulated room acoustics involving 3 room sizes, 3 levels of reverberation and 3 speaker-to-ASV microphone distances. Further, there are total 9 recording configurations generated with 3 attacker-to-speaker recording distances and 3 different quality loudspeakers. The evaluation data contains unseen replay configurations to observe the effect and to find generalizable countermeasures. Recently proposed ASV-centric metric tandem detection cost function (t-DCF) is used as the primary metric and equal error rate (EER) as a secondary measure to report the results on the corpus [14, 16]. The ASV scores for both the tracks are provided by the challenge organizers to obtain the t-DCF metric [14].

4.2. Experimental Setup

The 10 different front-ends used in this work are investigated for logical and physical access tracks. We used the delta and delta-delta coefficients along with the static coefficients for logical attack scenario in case of all the 10 front-ends. On the other hand, we considered only the static coefficients of CQ-EST, CQ-OST, eCQCC, eCQ-OST and CMC for physical access track.

The features on the train set are used to build GMM and DNN systems. We do not perform any kind of normalization on the features in case of GMM systems. Two 512 component GMMs are trained for all the front-ends followed by log-likelihood ratio computation under both logical and physical access. On the other hand, we used a window of 11-frame features as the input for DNN training. The number of hidden layers and hidden layer nodes are obtained empirically on the development set for every front-end considered. The hidden layers are trained with sigmoid network and cross-entropy with softmax is used as training criterion. It is to be noted that mean and variance normalization is performed on the input features to the DNN. We used the Microsoft Cognitive Toolkit for DNN training [36].

The scores generated from each system are combined with the ASV system scores given from the organizers of the challenge under each track to obtain the ASV-centric primary metric t-DCF. The weights for fusion of different systems are learned on development set. We have used Bosaris[3] toolkit for fusion and calibration of multiple systems [37].

[2] http://dx.doi.org/10.7488/ds/1994

[3] https://sites.google.com/site/bosaristoolkit/

Table 1: *Performance for logical access attacks on the development set of ASVspoof 2019 corpus.*

Front-ends	Back-ends			
	GMM		DNN	
	EER (%)	t-DCF	EER (%)	t-DCF
CQCC	0.431	0.012	**0.000**	**0.000**
CQ-EST	4.983	0.167	0.040	0.013
CQ-OST	3.845	0.129	0.042	0.001
eCQCC	0.941	0.025	**0.000**	**0.000**
CQSPIC	4.428	0.143	0.039	0.013
ICQCC	0.393	0.010	**0.000**	**0.000**
CMC	9.965	0.224	0.071	0.001
LFCC	2.827	0.077	1.569	0.046
IFCC	0.042	0.001	0.004	0.001
MFCC	7.062	0.169	12.087	0.221

5. Results and Discussion

The long range features discussed in Section 2 are developed as synthetic attack countermeasures in our previous works. In this work, we also study them for replay attacks to investigate the scope of generalized countermeasures for practical scenario. We have evaluated all the 10 front-ends using two different back-ends, totaling 20 systems each for logical and physical access scenario. Next, we discuss the results of the studies.

5.1. Logical Access

5.1.1. Primary System

Table 1 shows the performance for different front-ends with GMM and DNN modeling for logical access attacks. The studies on the development set showed that the three systems CQCC, eCQCC and ICQCC with DNN classifier perform accurately. Further, the front-ends IFCC, CQ-EST, CQ-OST, CQSPIC and CMC with DNN classifier perform very well. Therefore, we consider the fusion of these 8 front-ends (except LFCC and MFCC) with DNN classifier as our primary system.

5.1.2. Single System

We then investigate on the distributions of the bonafide and spoof scores of the three best systems on the development set. The separability for eCQCC feature is found to the best. Therefore, eCQCC feature with DNN classifier is submitted as our single system.

5.1.3. Contrastive-1 System

This is the fusion of 9 front-ends (except MFCC feature) with 2 different back-ends, totaling 18 subsystems. We do not consider MFCC features as the performance is much lower than all other systems on the development set.

Table 2: *Performance for physical access attacks on the development set of ASVspoof 2019 corpus.*

Front-ends	Back-ends			
	GMM		DNN	
	EER (%)	t-DCF	EER (%)	t-DCF
CQCC	10.037	0.192	9.850	0.182
CQ-EST	7.575	0.167	**1.891**	**0.043**
CQ-OST	7.148	0.137	2.187	0.048
eCQCC	8.518	0.169	4.094	0.084
CQSPIC	7.409	0.145	2.292	0.059
ICQCC	9.553	0.193	9.998	0.181
CMC	9.444	0.188	2.825	0.069
LFCC	11.797	0.253	13.263	0.256
IFCC	13.963	0.299	12.074	0.251
MFCC	11.834	0.276	11.588	0.241

Table 3: *Results for logical access on ASVspoof 2019 corpus.*

Submission	Development Set		Evaluation Set	
	EER (%)	t-DCF	EER (%)	t-DCF
Contrastive-1	**0.00**	**0.0000**	**4.13**	**0.1264**
Primary	**0.00**	**0.0000**	8.82	0.2417
Single	**0.00**	**0.0000**	11.08	0.2853
Given Challenge Baseline				
CQCC	0.43	0.0123	9.57	0.2366
LFCC	2.71	0.0663	8.09	0.2116

Table 4: *Results for physical access on ASVspoof 2019 corpus.*

Submission	Development Set		Evaluation Set	
	EER (%)	t-DCF	EER (%)	t-DCF
Contrastive-1	1.44	0.030	6.56	0.1590
Primary	**1.26**	**0.027**	**5.95**	**0.1381**
Single	1.89	0.043	8.23	0.2147
Given Challenge Baseline				
CQCC	9.87	0.1953	11.04	0.2454
LFCC	11.96	0.2554	13.54	0.3017

5.2. Physical Access

5.2.1. Primary System

Table 2 shows the results on the development set for different front-ends with the two back-ends for physical access attacks. We have fused all the 10 front-ends with 2 different back-ends, totaling 20 subsystems for primary submission to make it effective on the evaluation set.

5.2.2. Single System

The CQ-EST subband feature with DNN classifier performed the best as a single system on development set. Hence, we consider this as our single system under physical access scenario.

5.2.3. Contrastive-1 System

We consider all the 10 front-ends with DNN classifier for fusion and submitted as the contrastive system.

5.3. Observations on ASVspoof 2019 Submission

Table 3 and 4 show the results of the submitted systems with the given baseline results on the development and evaluation sets of ASVspoof 2019 database for logical access and physical access attacks, respectively. The results on the development set show that our systems work without error for logical access attacks.

Table 5: *Comparison in terms of average EER (AEER) for different long range features on ASVspoof 2015 evaluation set.*

Feature	Classifier	AEER (%)
CQCC [17]	GMM	0.255
CQCC [32]	DNN	0.110
eCQCC [32]	DNN	0.035
CQ-EST [30]	DNN	0.056
CQ-OST [30]	DNN	0.107
CQSPIC [31]	DNN	0.038
ICQCC [33]	DNN	0.099
CMC [29]	DNN	0.026

However, the evaluation set results depict that the best result is obtained with the contrastive-1 system, which is a fusion of 18 subsystems. This shows the nature of mismatch between the train and evaluation set on ASVspoof 2019 corpus. Similarly, for physical access attacks, the primary system obtained by fusion of 20 subsystems shows the best result.

At this stage, we are unable to make a detailed analysis of individual systems on the evaluation set as the metadata and keys for the challenge has not been released. However, the studies show that different aspects of long range features considered in this work helped us to achieve improved results than the single system submitted to the challenge. Further, our combined systems outperform the CQCC and LFCC baseline results given from the organizers by a large margin. We note that the our countermeasures behave in a similar manner to both logical and physical access attacks as their combined performances are close. This depicts that our countermeasures are useful for practical applications, where the attacks are unknown.

5.4. Evaluation on ASVspoof 2015 Corpus

As the metadata and keys for evaluation set of ASVspoof 2019 challenge have not been made available, we therefore present the performance comparison of individual features on previous anti-spoofing corpus. Table 5 summarizes the results for each CQT based long range features on ASVspoof 2015 database that deals with synthetic attacks. It can be observed that our long range features in general outperforms conventional CQCC features that validates the idea of using long range acoustic features in spoofing attack detection.

6. Conclusions

This work focuses on investigating few novel countermeasures on recent ASVspoof 2019 database for spoofing attack detection. The corpus contains both synthetic and replay attacks that are generated with latest available techniques. Our explorations on the countermeasures are based on long range information useful for capturing the artifacts of spoofed speech. The different aspects of long range acoustic features are combined along with few contrast systems that helps to have strong countermeasure generic to both synthetic and replay attacks. We intend to make a detailed analysis after release of the metadata and keys of ASVspoof 2019 challenge.

7. Acknowledgements

This research work is supported by Programmatic Grant No. A1687b0033 from the Singapore Government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain). All the authors are corresponding authors of this paper.

8. References

[1] K.-A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," in *SLTC Newsletter*, February 2013.

[2] K.-A. Lee, A. Larcher, H. Thai, B. Ma, and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *Interspeech*, 2011, pp. 3317–3318.

[3] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Development of multilevel speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, Sep 2017.

[4] Z. Wu and H. Li, "On the study of replay and voice conversion attacks to text-dependent speaker verification," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5311–5327, May 2016.

[5] P. L. D. Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[6] 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.

[7] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Interspeech 2013*, 2013, pp. 925–929.

[8] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanili, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, June 2017.

[9] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc i, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech 2015*, pp. 2037–2041.

[10] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech 2017*, 2017, pp. 2–6.

[11] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonalves, A. G. S. Mello, R. P. V. Violato, F. O. Simoes, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, H. Dinkel, N. Chen, Y. Qian, D. Paul, G. Saha, and M. Sahidullah, "Overview of BTAS 2016 speaker anti-spoofing competition," in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS) 2016*, Sept 2016, pp. 1–6.

[12] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey 2018*, 2018, pp. 195–202.

[13] T. Kinnunen, J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, and Z. Ling, "A spoofing benchmark for the 2018 voice conversion challenge: Leveraging from spoofing countermeasures for speech artifact assessment," in *Proc. Odyssey 2018*, 2018, pp. 187–194.

[14] "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2019.

[15] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech 2019*, 2019.

[16] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Odyssey 2018*, 2018, pp. 312–319.

[17] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Odyssey 2016*, 2016, pp. 283–290.

[18] ——, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[19] M. Sahidullah, T. Kinnunen, and C. Hanili, "A comparison of features for synthetic speech detection," in *Proc. Interspeech 2015*, 2015, pp. 2087–2091.

[20] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: metadata analysis and baseline enhancements," in *Proc. Odyssey 2018*, 2018, pp. 296–303.

[21] 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 *Interspeech 2015*, Dresden, Germany, 2015, pp. 2062–2066.

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

[23] M. Pal, D. Paul, and G. Saha, "Synthetic speech detection using fundamental frequency variation and spectral features," *Computer, Speech & Language*, vol. 48, pp. 31–50, 2018.

[24] R. Font, J. M. Espn, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge," in *Proc. Interspeech 2017*, 2017, pp. 7–11.

[25] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequency and cepstral features," in *Proc. Interspeech 2017*, 2017, pp. 22–26.

[26] H. A. Patil, M. R. Kamble, T. B. Patel, and M. H. Soni, "Novel variable length teager energy separation based instantaneous frequency features for replay detection," in *Proc. Interspeech 2017*, 2017, pp. 12–16.

[27] M. Witkowski, S. Kacprzak, P. elasko, K. Kowalczyk, and J. Gaka, "Audio replay attack detection using high-frequency features," in *Proc. Interspeech 2017*, 2017, pp. 27–31.

[28] J. Yang and R. K. Das, "Low frequency frame-wise normalization over constant-Q transform for playback speech detection," *Digital Signal Processing*, vol. 89, pp. 30 – 39, June 2019.

[29] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spectra information for spoofing attack detection," *IEEE/ACM Transactions on Audio. Speech and Language Processing (Under Revision)*, 2019.

[30] J. Yang, R. K. Das, and H. Li, "Subband features for synthetic speech decetion," *IEEE Transactions on Information Forensics and Security (Submitted)*, 2019.

[31] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," in *Proc. Interspeech 2018*, 2018, pp. 651–655.

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

[33] J. Yang and R. K. Das, "Long-term high frequency features for synthetic speech detection," *Digital Signal Processing, Elsevier (Submitted)*, 2018.

[34] R. K. Das and H. Li, "Instantaneous phase and excitation source features for detection of replay attacks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, Hawaii, 2018, pp. 1030–1037.

[35] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

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

[37] N. Brümmer and E. de Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," *CoRR*, vol. abs/1304.2865, 2013.