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

IIIT-H Spoofing Countermeasures for Automatic Speaker Verification Spoofing and Countermeasures Challenge 2019

K N R K Raju Alluri, Anil Kumar Vuppala

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

Abstract

The ASVspoof 2019 challenge focuses on countermeasures for all major spoofing attacks, namely speech synthesis (SS), voice conversion (VC), and replay spoofing attacks. This paper describes the IIIT-H spoofing countermeasures developed for ASVspoof 2019 challenge. In this study, three instantaneous cepstral features namely, single frequency cepstral coefficients, zero time windowing cepstral coefficients, and instantaneous frequency cepstral coefficients are used as front-end features. A Gaussian mixture model is used as back-end classifier. The experimental results on ASVspoof 2019 dataset reveal that the proposed instantaneous features are efficient in detecting VC and SS based attacks. In detecting replay attacks, proposed features are comparable with baseline systems. Further analysis is carried out using metadata to assess the impact of proposed countermeasures on different synthetic speech generating algorithm/replay configurations.

Index Terms: Spoofing, countermeasures, voice conversion, speech synthesis, replay attack, Gaussian mixture model, single frequency filtering, zero time windowing, instantaneous frequency.

1. Introduction

State-of-the-art automatic speaker verification (ASV) systems are developed to distinguish between genuine speakers and zero-effort impostors, by compensating session, channel and noise variations. However, the performance of the ASV system is vulnerable to spoofing attacks [1]. Spoofing refers to an attempt by a fraudster to get access of authorized user by providing the fake voice samples to the ASV system. In the literature, four types of spoofing attacks were considered [1, 2]. They are impersonation, speech synthesis (SS), voice conversion (VC), and replay. Majority of the spoofing countermeasures on non-standard datasets are presented in [1].

The ASVspoof 2019 challenge is in continuation to the previous special sessions conducted in Interspeech 2013 [2], 2015 [3] and 2017 [4]. The first special session [2] aims to highlight the need for collaboration among researchers to develop standard data sets and common evaluation protocols to move spoofing research forward. In the second special session [3], organizers came up with a standard text independent data set and a common protocol to deal with high technology VC and SS attacks. Several researchers have developed countermeasures for these attacks; the results are summarised in [3, 5]. In 2017 special session, the organisers focused on replay attacks and corresponding countermeasures developed are summarised in [4].

The ASVspoof 2019 challenge focuses on countermeasures for all major spoofing attacks, namely SS, VC and replay spoofing attacks. The spoofed data is collected from the recently developed state-of-the-art voice conversion and speech synthesis systems for VC and SS based attacks, and in more controlled setup for replay attacks. These advancements made spoofed speech more natural and it is more difficult to distinguish from genuine speech to that of previous challenges [3, 4]. The goal of ASVspoof 2019 challenge is to differentiate between genuine and spoofed speech. Based on the generation of spoofed speech, ASVspoof 2019 challenge has two subtasks i.e, logical access (LA) and physical access (PA). In LA sub-task, spoofed speech refers to the speech generated from SS and VC methods whereas in PA sub-task it is replayed speech. For the successful development of countermeasures, it is essential to select a feature which captures the artifacts related to the spoofing attack followed by modeling these feature representation with pattern recognition techniques[6]. Based on the principles mentioned above, several researchers have developed countermeasures for spoof detection [6, 7, 8].

Successful countermeasures developed for SS and VC attacks are based on the following theories: (i) natural phase information is missing in the synthesized speech [9], (ii) the long term dynamics of speech are not properly modelled by the majority of the synthetic speech generating techniques [6], and (iii) the cues related to synthetic speech are spread across all frequency regions [7, 10]. Successful countermeasures developed against replay attacks are based on the acoustic characterization of channel characteristics using high resolution features [6, 11, 12, 13].

By considering the aforementioned fundamental ideas involved in the implementation of successful countermeasures for all the major spoof attacks (VC, SS and replay), in this study we investigated a set of instantaneous cepstral features. They are (a) single frequency cepstral coefficients (SFFCC)[11, 12], (b) zero time windowing cepstral coefficients (ZTWCC)[13], and (c) instantaneous frequency cepstral coefficients (IFCC)[14, 15] for both LA and PA tasks. The main motivation to use these features as spoofing countermeasures are as follows,

- All the three features are computed from the instantaneous spectrum.
- All the features avoid conventional frame-based analysis while computing spectrum. They use short term analysis on the instantaneous spectrum to get frame-based features.
- Both SFFCC and IFCC are computed over long-range information of speech.
- Both IFCC and ZTWCC features carry phase information.
- SFFCC has high spectral resolution with a moderate temporal resolution, and ZTWCC have high temporal resolution with moderate spectral resolution.

http://dx.doi.org/10.21437/Interspeech.2019-1623

The primary objective of this paper is to find the robust counter measures for spoofing attacks. In this regard, instantaneous cepstral features are investigated for both LA and PA tasks using GMM modelling/classifiers. In this work firstly, the effectiveness of proposed features for individual tasks is studied. Later, the effectiveness of these features for combined (GMM models are trained on merged training set) LA and PA task investigated, and finally, the usefulness of these features for cross-task analysis (models are trained on one dataset and tested on other datasets) is explored.

The organization of this paper is as follows: Section 2 describes the proposed approach which along with the motivation behind the selection of front-end features and back-end classifiers used. The experimental setup is presented in Section 3. Results are discussed in Section 4. Finally, Section 5 gives our conclusions.

2. Proposed approach

In this section, the basic components of the proposed system, i.e., different front-end features and classifiers are presented.

2.1. Front-end features

In this work, three instantaneous cepstral features are studied for the task of spoof detection. The details of each feature are described as follows:

2.1.1. Single Frequency Filter Cepstral Coefficients (SFFCC):
Single frequency filtering analysis of speech produces output at any desired frequency at each instant of time [16]. The spectro-temporal resolution of spectrum can be adjusted with the parameter r, which represents the pole location of a single pole filter in z-plane [17]. Recently, cepstral coefficients computed from low SNR instants of SFF spectrum are explored for replay attack detection [11, 12]. In this study, we modified the SFFCC features proposed in [11] and named it as modified SFFCC (MSFFCC). As described in our previous work [12], MSFFCC will also contain three steps in its computation. They are: SFF spectrum estimation, sub-sampling, and cepstrum computation, respectively. The SFF spectrum estimation and cepstrum computation steps are same in both the methods, but the sub-sampling step is modified in MSFFCC.

- Sub-sampling in MSFFCC computation:
  In the proposed MSFFCC, the sub-sampling step is done by averaging the spectra in each 10 ms segment. The averaged spectrum of j^{th} segment is given by

$$l_j = \sum_{n=1}^{N} v_l[k, n] \tag{1}$$

  where $v_l[k, n]$ represents the j^{th} segment instantaneous spectra. The resultant signal after sub-sampling is v[k,l], for l < n.

2.1.2. Zero Time Windowing Cepstral Coefficients (ZTWCC)
Zero time windowing (ZTW) analysis of speech is primarily proposed to extract instantaneous vocal-tract features within 5 ms of speech [18]. ZTW analysis is explored for different applications [19, 20]. Recently, cepstral coefficients extracted from low SNR instants of ZTW spectrum for each 10 ms are explored for replay spoofing countermeasures [13]. In this study, modified ZTWCC (MZTWCC) are explored for developing countermeasures for both LA and PA tasks. There are three steps involved in the extraction of ZTWCCs which are similar to the steps involved in SFFCC extraction in [12]. The instantaneous spectrum computation and cepstral coefficients extraction steps are same in ZTWCC and MZTWCC, and the only difference is the sub-sampling step. The sub-sampling step in the previous section is adapted to get MZTWCC, i.e, instead of selecting low SNR instant in each 10 ms speech segment, we have taken the average of spectrum in each 10 ms segment as described in equation 1.

2.1.3. Instantaneous Frequency Cepstral Coefficients (IFCC)
Instantaneous frequency cepstral coefficients are primarily explored to study the importance of analytic phase for automatic speaker recognition problem [14]. Later IFCCs are explored for language identification [15] and replay attacks detection [21, 22]. As phase information is useful in detecting spoofing attacks [23, 24], in this study IFCCs are considered for both the tasks. The IFCC extraction in this study is similar to that of IFCC extraction in [14, 21, 22].

2.2. Back-end classifiers

2.2.1. Gaussian Mixture Model (GMM):
Gaussian mixture model (GMM) [25] is a generative model, which can effectively capture the low dimensional feature distribution for classification problems.

In this study, for the given training data, two separate GMMs are built for genuine ($\lambda_{genuine}$) and spoof (λ_{spoof}) data. Expectation maximization (EM) algorithm [26] is used to estimate the model parameters for each class individually by using maximum likelihood (ML) criteria. Once the models are build, the scoring for the test utterance (X) is computed as follows,

$$Score(X) = llk(X|\lambda_{genuine}) - llk(X|\lambda_{spoof}) \tag{2}$$

where X = $\{x_1, x_2, ..., x_T\}$ is the feature vector of test utterance, T represents number of frames. Here $llk(X|\lambda)$ represents the average likelihood of X given model λ.

$$llk(X|\lambda) = (1/T) \sum_{t=1}^{T} log(p(x_t/\lambda)) \tag{3}$$

3. Experimental setup

3.1. Database and protocol

In ASVspoof 2019 challenge there are two separate tasks named LA and PA. For each task, there is a separate database, which includes three mutually exclusive subsets named train, development and evaluation/test data. Further information regarding the number of utterances in each subset and spoofing configurations are given in [27]. First time in this challenge the affect of spoofing countermeasures on ASV performance is studied with the recently proposed metric named tandem detection cost function (t-DCF) [28]. Results are presented in terms of equal error rate (EER) and min-tDCF [28] according to challenge protocol.

3.2. Parameters used for feature extraction

The parameters used for computing MSFFCCs are pole location of single pole filter (r), frequencies at which amplitude envelopes have to compute and dimensionality of the feature vector. In this study, the r value is considered as 0.995. The amplitude envelopes are calculated at every 15.6 Hz frequencies within the range of 0 to Nyquist frequency ($f_s/2$) which results in 513 envelopes. Thirty dimensional static features appended with dynamic coefficients are considered for experimentation.

The parameters used for computing MZTWCCs are the length of the heavily decaying window and dimensionality of the feature vector. In this study, 5 ms decaying window is considered and 30-dimensional static features appended with dynamic coefficients are considered for experimentation. These parameters are same to that of our work presented in [13].

The parameters used to compute IFCCs are number of channels for filter bank analysis and dimensionality of features. In this study, 40-channel filter bank is considered and 20-dimensional static features appended with dynamic coefficients are considered for experimentation.

For convenience, MSFFCC, MZTWCCs are denoted as SFFCC and ZTWCCs in the rest of the paper.

3.3. Classifiers

In this study 512 mixtures are considered for building the GMMs with 10 iterations of EM algorithm for both genuine and spoof classes. Score level fusion of different sub-systems are performed using BOSARIS toolkit [29].

4. Results and discussion

Experiments performed on ASVspoof 2019 dataset are grouped into three categories. They are (i) experiments on individual tasks, (ii) experiments on combined task, (iii) experiments on cross-task data. The detailed analysis of these tasks is presented in the following sub-sections.

4.1. Result analysis on individual tasks

In this category, experiments are conducted for each task (LA and PA) separately by considering proposed instantaneous cepstral features along with baseline features. For all the experiments in this section, GMMs are built using training data of particular task, and the results are reported on the development and evaluation data set of the corresponding task.

Table 1: *Results for different systems on LA task.*

		Development		Evaluation	
Description	Feature	t-DCF	EER	t-DCF	EER
Baseline	CQCC [30]	0.0123	0.43	0.2366	9.57
Baseline	LFCC [30]	0.0663	2.71	0.2116	8.09
Single	ZTWCC	0.0005	0.04	0.1414	6.13
Contrasive-1	IFCC	0.0002	0.01	0.2886	10.21
Contrasive-2	SFFCC	0.0034	0.12	0.1295	5.22
Primary	ZTWCC+CQCC	0	0	0.1239	4.92

Experiments conducted on LA task using five features and the results submitted to the challenge are reported in Table 1. From the results in Table 1, it is evident that the proposed features can detect the SS and VC based attacks better than baseline features. In particular on development set, IFCCs and ZTWCCs are performing better than the rest of the features. These results are in line with our hypothesis that the phase based features are useful in detecting attacks based on VC and SS. Whereas for evaluation data IFCCs are performing poorly compared to SFFCCs and ZTWCCs. This result indicates that high resolution features are more generalisable to the new attacks than the IFCCs. Further analysis is performed on these features using metadata provided by the organizers, and the results on development data are reported in the first half of Table 2.

In Table 2, SS-1, SS-2, SS-4, US-1, VC-1, and VC-4 are the different types of algorithms used to generate spoofed speech. From the results in the first half of Table 2, it can be observed that SS-1, SS-2 and SS-4 type spoofed speech is easily distinguished from genuine speech using all the features. Whereas the proposed features get benefit from the remaining attacks to perform well for the overall task. From the results one more interesting point can be observed that these features exhibit complementary nature, for example, CQCCs are unable to detect VC-1 efficiently, but it can identify US-1 and VC-4 effectively whereas ZTWCC can detect VC-1 effectively than CQCC while underperforming for other two types. This motivated us to do score level fusion across features, and the results are reported in the second half of Table 2. All the ten combinations of fusion results are presented in Table 2. From the results, it is evident that each combination get benefit with fusion and the fused system result is better than their counterparts. The score level fusion of CQCC and ZTWCC based sub-systems correctly detected the spoofing types, and the same is submitted as our primary system for LA task.

Table 2: *Individual attack results (in % EER) of different systems on LA task development data set.*

Replay Configuration	SS-1	SS-2	SS-4	US-1	VC-1	VC-4	Pooled
CQCC	0	0	0.08	0	0.94	0.03	0.43
LFCC	0.03	0	0	4.90	0.16	5.27	2.71
SFFCC	0	0	0	0.16	0.03	0.08	0.12
ZTWCC	0	0	0	0.11	0	0.03	0.04
IFCC	0	0	0	0	0.03	0.08	0.01
CQCC+LFCC	0	0	0	0	0.11	0.02	0.05
CQCC+SFFCC	0	0	0	0	0.04	0	0.01
CQCC+ZTWCC	0	0	0	0	0	0	0
CQCC+IFCC	0	0	0	0	0.02	0.05	0.01
LFCC+SFFCC	0	0	0	0.27	0	0.15	0.15
LFCC+ZTWCC	0	0	0	0.40	0	0.17	0.12
LFCC+IFCC	0	0	0	0	0.02	0.08	0.02
SFFCC+ZTWCC	0	0	0	0.08	0	0.02	0.03
SFFCC+IFCC	0	0	0	0	0.02	0.05	0.01
ZTWCC+IFCC	0	0	0	0	0	0.05	0.01

Table 3: *Results for different systems on PA task.*

		Development		Evaluation	
Description	Feature	t-DCF	EER	t-DCF	EER
Baseline	CQCC [30]	0.1953	9.87	0.2454	11.04
Baseline	LFCC [30]	0.2555	11.96	0.3017	13.54
primary/Single	ZTWCC	0.2169	10.11	0.2810	12.20
Contrasive-1	IFCC	0.2926	13.45	0.3573	15.59
Contrasive-1	SFFCC	0.2359	11.09	0.3232	13.97

Experiments conducted on PA task using the above mentioned five features and the results submitted to the organisers are reported in Table 3. From the results in Table 3, it can be observed that the proposed features SFFCC and ZTWCC are able to detect the replay attacks similar to that of baseline features, whereas IFCC is under-performing to that of baseline features on both development and evaluation data sets. Here in PA task, we consider ZTWCC as our primary submission. In order to investigate the reasons for the performance degradation, further analysis is performed using the metadata provided by the organisers and the results are reported in Table 4.

Here the duple (D_a, Q) is constructed based on the attacker to talker distance and the replay device quality. Each entity has three (A, B, C) categories and the detailed description is given in metadata. From A to C, the severity decreases, i.e., AA represents it is challenging to differentiate genuine and replay speech. Similarly, CC represents it is very easy to distinguish genuine and replay speech. Results for all the 9 combinations of the replay configurations are presented in Table 4. From the results in Table 4, it can be observed that in each subsystem, it is challenging to detect AA and easy to recognize CC. These results are in line with the information provided in the metadata. From the results of the first three rows in Table 4, it can be observed that the error rates are very high compared to the remaining six replay configurations for both data sets. From these results, it can be observed that, the replay device quality should be weighted more than the attacker to talker distance in developing the spoofing countermeasures. So, further analysis is performed using the recording device quality, and the results on development data set are reported in Table 5.

Table 4: *Individual configuration type results (in % EER) of different systems on PA task.*

Replay configuration	CQCC		LFCC		SFFCC		ZTWCC		IFCC	
	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
AA	25.62	25.28	31.22	32.48	28.24	30.62	27.72	30.02	30.54	33.11
BA	18.72	21.87	21.06	24.59	19.14	23.75	16.59	20.74	24.07	29.79
CA	17.99	21.10	17.84	21.63	18.19	22.16	14.11	17.78	20.70	25.72
AB	2.26	6.16	2.83	4.40	2.60	9.61	2. 52	6.22	2.10	5.45
BB	1.46	5.26	2.29	4.29	1.87	7.49	1.35	4.34	2.07	5.30
CB	1.74	4.70	1.92	3.92	2.01	6.87	1.30	3.59	1.92	5.10
AC	0.92	2.13	1.57	3.95	1.12	4.13	1.04	4.01	1.28	3.31
BC	0.97	1.61	1.24	3.20	1.09	2.97	0.78	2.62	1.66	3.32
CC	0.61	1.79	1.23	3.06	0.73	3.04	0.46	2.37	1.61	3.24
Pooled	9.87	11.04	11.96	13.54	11.09	13.97	10.11	12.20	13.45	15.59

Table 5: *Results (in % EER) in terms of replay device quality for different systems on PA task development data set.*

Replay Configuration	Low	High	Perfect	Pooled
CQCC	0.63	1.39	21.53	9.87
LFCC	1.08	2.01	22.51	11.96
SFFCC	1.17	2.60	23.39	11.09
ZTWCC	0.81	1.94	20.55	10.11
IFCC	1.42	2.00	25.94	13.45
CQCC+LFCC	0.43	1.05	20.26	9.12
CQCC+SFFCC	0.58	1.40	20.96	9.50
CQCC+ZTWCC	0.48	1.16	20.32	9.04
CQCC+IFCC	0.41	0.89	20.17	9.77
LFCC+SFFCC	0.90	2.04	22.00	10.68
LFCC+ZTWCC	0.77	1.76	20.44	9.93
LFCC+IFCC	0.94	1.59	23.06	11.88
SFFCC+ZTWCC	0.95	2.13	21.56	10.55
SFFCC+IFCC	0.84	1.5	22.19	11.24
ZTWCC+IFCC	0.79	1.51	22.03	11.23

Based on the second entity in the duple (D_a, Q) three conditions are considered for experimentation. They are low (C), high (B) and perfect (A). Similar to the LA task, here also score level fusion is performed across subsystems, and the results are reported in Table 5. For each feature from low to perfect error rates are increasing. A similar observation is there for score-level fusion of different subsystems. Here we made an interesting observation is that, the score level fusion of sub-systems get benefit in case of low and high-quality recording devices, whereas in case of perfect, the error rates are still high. This limitation in the case of perfect recording device quality is affecting the overall system performance. From the Results in Table 5, it can be observed that, high-resolution features (CQCC, SFFCC, and ZTWCC) are useful in detecting low and high quality recording device based attacks.

4.2. Results on combined task

The primary motivation behind this work is to study the effect of merged data on individual tasks. Here the combined task means the models are built by using the merged training data of both the tasks and the results are reported on individual development data sets. Experimental results are reported in Table 6.

Table 6: *Results (in % EER) of different systems trained on merged data set.*

Trained on Combined data	LA		PA	
	EER	t-DCF	EER	t-DCF
CQCC	0.51	0.015	9.57	0.183
LFCC	3.73	0.107	9.98	0.213
SFFCC	1.10	0.032	11.51	0.247
ZTWCC	0.71	0.024	10.48	0.229
IFCC	0.27	0.008	13.91	0.283

From the results in Table 6, in case of LA task, the models developed on merged data resulted in high error rates to that of individual counterparts. For CQCC the increased in error rates are relatively low to that of other features. Whereas in case of PA task the error rates are almost equal to that of individual counterparts. LFCC get benefit with merged data in case of PA, and the error rate decreased from 11.96 % to 9.98 %. From these results, it is clear that the development of individual countermeasures for LA and PA tasks are beneficial than that of combined countermeasures i.e., we can not combine all the spoofing attacks into one group for successful spoofing countermeasures.

4.3. Results on the cross evaluation of tasks

The primary motivation to conduct these experiments is to investigate the common cues for both the tasks. In this study, models built with one task are tested with other task development data and vice versa. Results for a cross-task are presented in Table 7.

Table 7: *Results (in % EER) of different systems for cross task on development data*

Feature type	Train: LA Test: PA		Train: PA Test: LA	
	EER	t-DCF	EER	t-DCF
CQCC	40.0	0.819	25.74	0.614
LFCC	16.98	0.380	35.83	0.752
SFFCC	19.63	0.430	27.68	0.70
ZTWCC	20.24	0.476	31.47	0.72
IFCC	32.46	0.770	9.61	0.2816

It can be observed from Table 7 that, the error rates for cross task are much higher compared to that of the individual results. LFCC, SFFCC, and ZTWCCs are performing better than CQCC and IFCC in case of models trained on LA and tested on PA task. Whereas in other case IFCC and CQCC are performing better than other features. These results suggest that the developed countermeasures are not completely generalizable, and the best results can be obtained when a matched task and attacks are used in both training and testing.

5. Summary and conclusion

This study presents IIIT-H submission for ASVspoof 2019 challenge. In this study, three instantaneous cepstral features named SFFCC, ZTWCC, and IFFCC are studied for the LA and PA tasks. The proposed features are able to detect SS and VC based attacks better than the baseline features, whereas in case of PA task, these are comparable with the baseline features. Experiments conducted on merged data resulted in higher error rates in case of LA task and similar results in case of PA task, when compared with individual task results. Experiments performed on cross-task suggest that the cues captured by proposed features are task specific. Further analysis needs to be done on evaluation data to develop generalizable spoofing countermeasures.

6. Acknowledgements

The first author would like to thank the Department of Electronics and Information Technology, Ministry of Communication & IT, Govt of India for granting Ph.D. Fellowship under Visvesvaraya Ph.D. Scheme.

7. References

[1] 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.

[2] 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.

[3] 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.

[4] 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, pp. 2–6.

[5] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilci, 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*, 2017.

[6] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

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

[8] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. A. L. Thomsen, A. K. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco *et al.*, "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *Proc. ICASSP*, 2017, pp. 5395–5399.

[9] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. INTERSPEECH*, 2012, pp. 1700–1703.

[10] 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.

[11] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "Detection of replay attacks using single frequency filtering cepstral coefficients," in *Proc. INTERSPEECH*, 2017, pp. 2596–2600.

[12] ——, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *Proc. INTERSPEECH*, 2017, pp. 107–111.

[13] K. R. Alluri and A. K. Vuppala, "Replay spoofing countermeasures using high spectro-temporal resolution features," *International Journal of Speech Technology*, pp. 1–11, 2019.

[14] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[15] K. Vijayan, H. Li, H. Sun, and K. A. Lee, "On the importance of analytic phase of speech signals in spoken language recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5194–5198.

[16] 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.

[17] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," 2018, pp. 441–445.

[18] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[19] R. Prasad and B. Yegnanarayana, "Acoustic segmentation of speech using zero time liftering (ZTL)," in *Proc. INTERSPEECH*, 2013, pp. 2292–2296.

[20] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ZTWCCs)," in *Proc. INTERSPEECH*, 2018, pp. 232–236.

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

[22] R. K. Das and H. Li, "Instantaneous phase and excitation source features for detection of replay attacks," in *Proc. APSIPA ASC*. IEEE, 2018, pp. 1030–1037.

[23] I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas, "Synthetic speech detection using phase information," *Speech Communication*, vol. 81, pp. 30–41, 2016.

[24] 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.

[25] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[26] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.

[27] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan."

[28] 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," *arXiv preprint arXiv:1804.09618*, 2018.

[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.

[30] T. Massimiliano, Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, Todisco, and H. Delgado, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," *submitted to INTERSPEECH 2019*.