Instantaneous Phase and Excitation Source Features for Detection of Replay Attacks

Rohan Kumar Das and Haizhou Li
Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
E-mail: {rohankd, haizhou.li}@nus.edu.sg

***Abstract*—In present era, the spoof detection has become an integral part of biometric systems and speaker verification is no exception to it. The replay attacks are the most common, where the attacker plays the recorded speech of a user to validate a false identity claim. Currently, constant-Q cepstral coefficient (CQCC) feature based system represents the standalone benchmark for spoof detection. However, we hypothesize that the phase and excitation source information of speech may carry additional artifacts that are useful for identifying the replay attacks. In this regard, instantaneous frequency cosine coefficients and two source features namely, discrete cosine transform of integrated linear prediction residual and residual mel frequency cepstral coefficients are explored. The studies are conducted on ASVspoof 2017 Version 2.0 database designed for the replay attacks. The results reveal that the phase and source features although perform poorer than CQCC, their fusion helps to achieve an improved performance. This indicates the complementary nature of information carried by the stated features is useful for detecting replay attacks. Further, an analysis on the behavior of each of these features under different replay configurations is also presented to highlight their effect in different scenarios.**

I. Introduction

In the current decade, the area of speaker verification (SV) has passed through a breakthrough in research showing potential for deployable systems [1]–[4]. Particularly, the text-dependent SV has gained attention for practical application oriented systems due to lesser time complexity and high performance associated with it. However, spoofing attacks have become a concern when such systems are used for having access to an intended service [5]. It has been found that the SV systems are greatly affected by spoofing as they are quite vulnerable to such attacks [6]. The spoofing attacks are of several kinds, broadly they are categorized as impersonation, voice conversion, test-to-speech synthesis (TTS) and replay attacks [7]. The voice conversion and TTS based spoof attacks depend on technology and require technical insight of the attacker. On the other hand, impersonation and replay attacks can be considered as common attacks for spoofing.

As SV is increasingly adopted as a biometric means, the security of SV systems against spoofing attack becomes one of the focal points of the research community. In this regard, special challenges has been organized with standard datasets to spearhead this field. This all started with a special session on spoofing and countermeasure in Interspeech 2013 [8], [9]. It is followed by the first edition of spoof challenge called as ASVspoof 2015, which focuses on synthetic speech based attacks for SV systems [10]. The synthetic speech in this database comprises of speech generated using voice conversion and TTS synthesized system. Subsequently, an anti-spoofing competition, BTAS 2016 was organized that includes both replay and synthetic speech attacks. The overview of this competition can be found in [11]. Then another challenge ASVspoof 2017 explicitly based on the replay attacks is organized to study their behavior [12]. There are few anomalies found on the examples of this ASVspoof 2017 database on replay attacks in terms of presence of beep sounds and broken files, which are omitted in its second version release ASVspoof 2017 Version 2.0 database [13].

The performance on ASVspoof 2015 database on synthetic speech has reached a benchmark after a lot of explorations by different groups. A novel feature constant-Q cepstral coefficient (CQCC) is introduced in [14], [15] for detection of spoofed speech, which proved to dominate all other features in this field. This feature provides a very high performance for all the 10 conditions of ASVspoof 2015 database. However, the replay attack based speech on ASVspoof 2017 database is found to be more challenging as can be seen from the submissions to the ASVspoof 2017 challenge [12]. Further, although the CQCC feature performs better compared to other features for replay attacks, its performance is not comparable to that obtained in case of synthetic speech detection. Thus, replay speech detection needs much more in depth investigation in order to understand the discriminating behavior of the replay signal from a genuine speech [13].

In this work, we study the detection of replay speech and investigate few features that can capture potential information for identifying replay attacks. The replay speech possesses the information of playback device, recorder and the environment in which the attack is performed. Thus, the features which can have definite information of these aspects can be useful for identifying such attacks. This motivates us to explore features that capture additional/complementary information from that carried by the conventional features for spoof detection in order have supplementary useful artifacts.

In speech signal processing, most of the features are derived from the magnitude of the signal. However, phase of a signal also possesses definite characteristic information which can be useful for many applications. The phase information have been found to be useful for detection of synthetic speech as explored in [16]. There are different attempts made to capture the phase information of a signal in the literature. One of them is the instantaneous frequency cosine coefficient (IFCC) features that are derived from analytic phase of speech [17]. The IFCC features are computed over long range information of the speech signal and hence carry long duration instantaneous variations in speech. In this regard, they are expected to capture significant information to detect spoof speech that can be useful like the conventional CQCC features that too depends on the long range characteristic information of the speech signal.

The excitation source features are very sensitive to the variation in the channel/session that affects SV performance as reported in [18]. Thus, in case of spoof detection this characteristics of source features can be useful to recognize replay attacks, the playback device and recorder plays a crucial role apart from the background environment. Further, the work of [19] shows that the evidence obtained from the throat microphone is very useful in identifying the replay attacks. The throat microphone generally captures the glottal source information and hence projects their scope for use in replay speech detection. This motivated us to explore two source features namely, discrete cosine transform of linear prediction residual (DCTILPR) and residual mel frequency cepstral coefficients (RMFCC) for the detection of replay attacks. Further, an analysis on the behavior of each of investigated features under different replay configurations is also presented to show their effect for different scenarios. The contribution of this work lies in projecting the scope of instantaneous phase and excitation source features for replay speech detection along with their ability in different replay configurations to handle the replay attacks.

The remaining part of the paper is organized as: Section II mentions regarding the works related to replay attacks and the current benchmark for handling such attacks. Section III investigates the instantaneous phase and excitation source features for the detection of replay based spoof attacks. In Section IV, the experimental studies related to the explored features are reported. An analysis on the behavior of the considered features based on different replay configurations are also discussed in this section. Finally, Section V provides the conclusion to this work highlighting the future directions associated with it.

II. Related Work and Current Benchmark

The replay attacks have been studied by different research groups across the globe over the time. The authors of [20], [21] have studied replay attacks made with far-field recordings. They considered noise and reverberation information for preventing such attacks. Similarly, channel noise has been used a characteristics for discriminating replay attacks as mentioned in [22]. A spectral bitmap based approach is adopted by the authors of [23] to detect replay speech in a text-dependent SV framework. This work is extended to text-independent SV in terms of creation of average spectral bitmap models for identifying replay based spoof attacks [24]. In [25], a playback detection algorithm is proposed based on spectral features and score normalization that is useful to capture the acoustic behavior of the environment.

The ASVspoof 2017 challenge on replay attacks became successful as a large number of submissions are made by various groups [12]. The authors of [26] presents an analysis on the use of different features for replay attack detection on the first version of the ASVspoof 2017 database. They have considered linear frequency cepstral coefficients, inverted mel frequency cepstral coefficients (IMFCC), rectangular filter cepstral coefficients, linear prediction cepstral Coefficients, subband spectral flux coefficients, subband spectral centroid frequency coefficients and subband spectral centroid magnitude coefficients apart from conventional CQCC and mel frequency cepstral coefficient (MFCC) features. These features are found have additional information that helps in improving the performance. Similarly, in [27] the authors proposed novel features based on epoch strength and peak to side lobe ratio that is useful when fused with existing features for replay attacks. A novel variable length Teager energy separation based instantaneous frequency feature is proposed for detection of replay attacks in [28]. The authors of [29] reported regarding the importance of high frequency based features for detecting the replay attacks. Again, the hierarchical scattering decomposition coefficients and IMFCC features are found to be useful for handling the replay attacks [30]. All these works project the scope of alternative features having additional information in case of replay based spoof detection.

The ASVspoof database released as a part of ASVspoof 2017 challenge had few broken files. Further, the beep sound occurring at the beginning of replay speech makes it as a characteristics of such spoofed speech and helps in easier detection. The creators of the database have worked upon these issues to remove all those shortcomings and a new database is released as the ASVspoof 2017 Version 2.0 database [13]. In their recent work [13], the authors have presented two systems based on Gaussian mixture model (GMM) and i-vector modeling [31], [32]. Further, the authors have introduced log-energy features along with CQCC features and performed cepstral mean and variance normalization (CMVN) [33] that is found to improve the performance for replay speech detection. It is also observed from their work that the GMM based system performs better than the i-vector based system. Therefore, the baseline system in this current work considers CQCC features and log-energy coefficient, which are normalized in the cepstral domain with a GMM based classifier.

III. Instantaneous Phase and Excitation Source Features for Replay Attacks

In this section, we discuss the instantaneous phase and excitation source features. A basic description of each of the considered feature along with their cues to detect replay attacks is presented.

A. Instantaneous Phase Features

As discussed in the introduction, there have been many attempts to capture the phase information. One of them is the IFCC features that are derived from analytic phase of speech signal [17]. The Fourier transform properties are used to extract instantaneous frequency (IF) to avoid the problem of phase warping. To compute IF, the narrow band components of the speech signal are taken in the following way,

$$\theta'[n] = \frac{2\pi}{N} Re\left\{ \frac{F_d^{-1} kZ[k]}{F_d^{-1} Z[k]} \right\} \tag{1}$$

where F_d^{-1} represents inverse discrete Fourier transform (IDFT), N is the length of the narrowband signal and $Z[k]$ is the DFT of the analytic signal $z[n]$, obtained from the narrowband component of speech signal as mentioned in [34].

The DCT is then used on top of the IF for extracting the IFCC features. These features are found to carry complementary information than that in conventional MFCC features and are useful in fusion for SV studies [17]. Recently, the IFCC features are also proven to be significant for language recognition [35]. The IFCC features discussed here are computed over long range information of the speech signal. The short term processing is only applied at the end to obtain frame based features. Thus, they are expected to have definite characteristic information that are useful for detection of spoofed speech. Figure 1 shows a pair of genuine speech and its replayed version with their corresponding pyknogram representations generated using instantaneous frequency. A pyknogram represents the scatter plot of all the IF from all the bands of speech signal. It is observed that the phase information in genuine and replay speech has different characteristics. Therefore, the instantaneous phase information can be used for discriminating the genuine and replay speech.

B. Excitation Source Features

From the source/filter modeling of speech signal, the excitation source and the vocal tract have complementary information which can be utilized together for many applications. The conventional features like MFCC generally capture the vocal tract information in terms of shape/size, change in shape, rate of change in shape, etc. Thus, the source features that represent the excitation source characteristics may be useful for capturing the additional information. Further, it has been found that the source features posses different attributes of excitation source that on combination is useful for recognizing speakers [36], [37]. We give a brief description of the two source features used in this work in the following subsections.

1) Discrete Cosine Transform of Integrated LP Residual: The feature DCTILPR is derived from integrated linear prediction residual (ILPR) that closely resembles the glottal signal [38]. The speech regions of a given utterance are first identified to extract the epochs for those locations. Then a voiced/unvoiced classification is applied to consider the epochs from the voiced regions as glottal closure instant (GCI)s to perform pitch synchronous DCT in the interval between one GCI to the successive one. The first 24 coefficients are considered to obtain the DCTILPR features [18], [39].

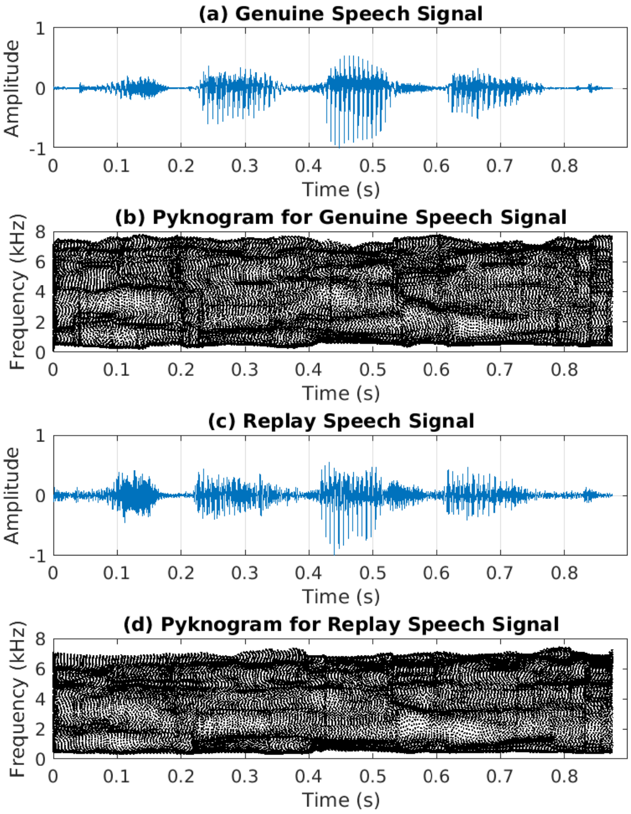

Fig. 1. (a)-(b) Genuine speech signal and corresponding pyknogram (c)-(d) replay speech signal and corresponding pyknogram.

2) Residual Mel Frequency Cepstral Coefficients: Cepstral analysis over linear prediction (LP) residual is attractive due to its simplicity [40]. The approach involving cepstral analysis can be improved by use of spectral subband energies. Since the spectrum of LP residual is flat in nature, if the spectral energies are accumulated over the subbands the benefit of using them as features can be achieved. The source feature RMFCC is extracted from log-magnitude spectrum of LP residual by performing short term processing. A non-uniform mel filterbank is used, through which the LP residual spectrum is passed and then inverse discrete time Fourier transform (IDFT) of the resultant signal is taken to obtain RMFCC features [41].

The LP residual of speech contains the excitation information and it has a noise like structure. In case of replay speech, it is expected that due to involvement of a playback and a recorder, the the LP residual signal of the replay speech becomes more distorted which will lead to discriminate it from that of the genuine speech. Similarly, the GCI locations are expected to detect spuriously for replay speech that will affect the ILPR signal which can also be used as a discriminating information from the genuine speech. Figure 2 shows the voiced portion of a speech signal and its replayed version along with their corresponding LP residual and ILPR signals. It can be

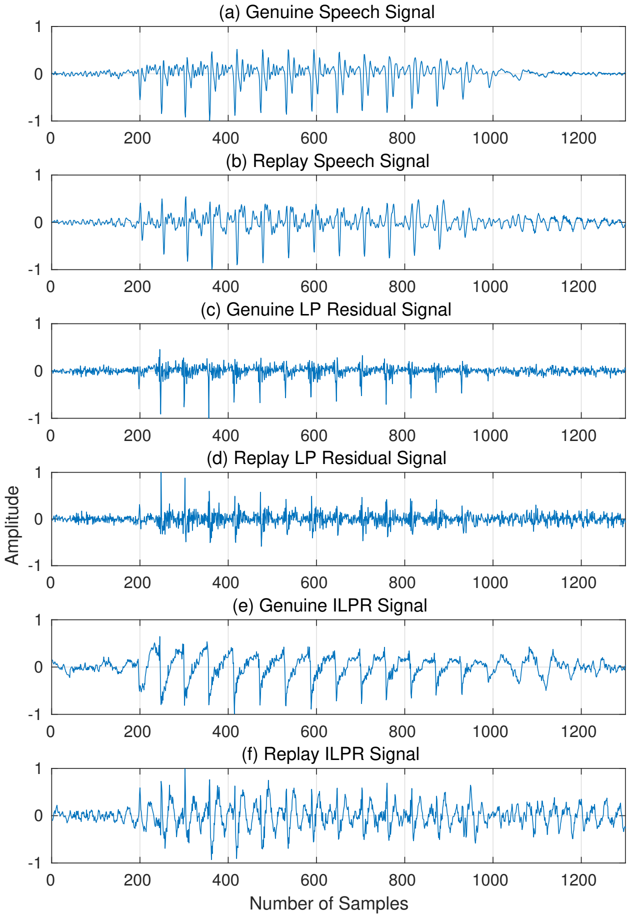

Fig. 2. Figure showing differences in genuine and its corresponding replay signal (a)-(b) speech signal, (c)-(d) LP residual signal (e)-(f) ILPR signal.

observed that there is little discriminative trace between the waveforms of genuine and replay speech. However, the LP residual of replay speech has more noisy structure. Further, the ILPR of replay speech is quite distinguishing from that of the genuine speech. This depicts that the excitation source information can be helpful for the detection of replay based spoof attacks.

IV. Experimental Studies

This section presents the details of the experimental studies related to different features considered in this work for detecting the replay attacks. The analysis on the fusion of different features is also included in this section.

A. Database

The ASVspoof 2017 Version 2.0 database is a modified version of the original ASVspoof 2017 database which is created using the RedDots corpus designed for text-dependent SV studies [13], [42]. The speech examples are stored with the sampling rate of 16 kHz with resolution of 16 bit per sample. There are three subsets, namely, train, development and evaluation set in the current version of the database. The train set contains labeled genuine and replay examples for training the models of genuine and replay based spoof speech. In development set, the trials are constituted with both genuine and replay speech which are to be evaluated against the trained model. Similarly, the evaluation set also contains a set of trials, but these trials are more challenging due to the configurations used for replay attacks. The train and development set contains only 3 and 10 replay configurations, respectively. On the contrary, there are 57 different replay configurations in the evaluation set. Table I summarizes the details of ASVspoof 2017 Version 2.0 database.

TABLE I
ASVspoof 2017 Version 2.0 database details.

Database Subset	**# Speakers**	**# Replay Configurations**	**# Utterances**	
			Genuine	**Spoofed**
Train	10	3	1,507	1,507
Development	8	10	760	950
Evaluation	24	57	1,298	12,008

B. Experimental Setup

The authors of [13] have put forward a benchmark system using CQCC features for the detection of replay speech on ASVspoof 2017 Version 2.0 database. 19-dimensional CQCC features and log-energy coefficient (E_l) along with their delta (Δ) and delta-delta ($\Delta\Delta$) derivatives are considered. Further, CMVN is performed on top the features as it found to be useful as mentioned in [13]. GMM based framework is chosen for building the models for genuine and replay based spoofed speech as it yields improved results than that of i-vectors [13]. Two 512 component based GMMs are trained for genuine and replay speech. For the studies under development set, the utterances of train set are used to build these models. On the other hand, during the evaluation set based studies, the utterances of train as well as development set are used together to train the respective models of genuine and replay speech. Given a test speech its log-likelihood ratio (LLR) is computed considering the genuine and spoof speech models. Equal error rate (EER) is used as a metric for reporting the results as per the protocol of the database [13].

While investigating the other features considered in this work, the respective features are extracted and then 512 component based GMM models are build similarly as done for the case of CQCC features. The remaining framework for evaluating the system performance remains the same for each feature based studies. Bosaris toolkit is used for score fusion based studies performed in this work [43]. It is to be noted that the parameters for fusion are learned on the development set and then applied on the evaluation set.

The MFCC features are the most common features used in the domain of speech processing. With the success of MFCC features in a wide range of speech related applications, it has been investigated as a common reference feature in

TABLE II
DIFFERENT FEATURES AND THEIR SPECIFICATIONS USED.

Feature	Configuration	CMVN
MFCC	(E_l+19-static) + 20-Δ + 20-$\Delta\Delta$	Yes
Long Range Features		
CQCC	(E_l+19-static) + 20-Δ + 20-$\Delta\Delta$	Yes
IFCC	20-static + 20-Δ + 20-$\Delta\Delta$	Yes
Source Features		
DCTILPR	24-static	No
RMFCC	13-static + 13-Δ + 13-$\Delta\Delta$	Yes

TABLE III
PERFORMANCE ANALYSIS IN TERMS OF EER (%) FOR DIFFERENT FEATURES ON ASVSPOOF 2017 VERSION 2.0 DATABASE.

Feature	Development	Evaluation
MFCC	18.04	20.78
Long Range Features		
CQCC	8.93	12.20
IFCC	16.20	15.90
Source Features		
DCTILPR	22.69	14.03
RMFCC	23.58	20.49

TABLE IV
PERFORMANCE ANALYSIS IN TERMS OF EER (%) FOR DIFFERENT FEATURE COMBINATIONS ON ASVSPOOF 2017 VERSION 2.0 DATABASE.

Combination	Development	Evaluation
Fusion with CQCC		
CQCC+MFCC	7.99	11.06
CQCC+IFCC	8.70	11.33
CQCC+DCTILPR	7.21	9.36
CQCC+RMFCC	7.27	10.45
Long Range Features vs. Source Features		
CQCC+IFCC	8.70	11.33
DCTILPR+RMFCC	19.63	13.75
All Features Fusion	**6.05**	**9.01**

this work for the detection of replay speech. The CQCC and IFCC features are grouped together to observe their effects as they contain the long range information of speech signal. On the other hand, the excitation source features DCTILPR and RMFCC are put under one category as they represent source information. In Table II, we summarize the specifications of different features.

C. Results and Discussions

This subsection reports the results associated with different studies and the related observations.

1) Performance of Individual Features: Table III reports the results for different features considered in this work. The MFCC feature is used as a common reference feature, the long range information features CQCC and IFCC are grouped together. Similarly, the performance of the source features are put together. It can be observed that CQCC dominates to provide the best results as an individual feature for replay attack detection. Further, the long range information based features work better compared to the source features as well as conventional MFCC features.

2) Performance under Feature Combinations: The studies in [44] mentioned that the score fusion of multiple systems is very useful for detection of spoof attacks. Hence, we have performed score fusion on the scores obtained from different features investigated in this work. Table IV shows the results under fusion of different feature combinations. As CQCC provides the individual feature based benchmark result, its combination with the phase and source features is investigated. Table IV shows the results of fusion of different features with CQCC. It is observed that the complementary information of CQCC and IFCC is comparatively less as both of them are based on long range information of the speech signal. On the other hand, the fusion of CQCC with the source features as well as conventional MFCC feature is found to be more effective due to different nature of characteristics carried by them. Additionally, we make a comparison of performance for long range features against the source features. The long range features clearly show their effectiveness for replay speech detection over the source features. Finally, all the features are fused to show the combined performance. The combined fusion outperforms any other combinations showing the usefulness of both instantaneous phase and excitation source features for the detection of replay attacks.

3) Effect of Environment and Devices: The quality of the replay speech greatly depends on the playback and recording devices as well as the recording environment as mentioned in [13]. The ASVspoof 2017 Version 2.0 database is collected in 26 different recording environments with 26 playback and 25 recording devices. The authors of [13] have classified the recording environment, playback and recording devices in three different categories namely, low, medium and high based on their effect on replay speech. In case of recording environment, the low noise condition has a higher threat ('high' category) than with relatively higher noise ('medium' category) and very high noise ('low' category) condition based environments. Similarly, for playback and recording devices, the high quality devices produce a high quality signal which makes them a bigger threat ('high' category) than the comparatively lower quality ('medium' category) and very low quality devices ('low' category).

Table V shows the performance of different features with respect to the different categories of environment, playback and recording devices. The results depict that the recording environment plays the most crucial role in case of every feature and their combined fusion as the replay detection performance severely degrades in a clean recording environment ('high' category) of the replay speech. Additionally, the trend of the features are not the same between the development and evaluation set as the conditions of evaluation set are quite different. However, the inferences stated from this analysis are not much precise as the quality of replay speech depends on environment, playback and recording devices collectively at the same time. Therefore, a combination of environment, playback and recording device, which is referred to as a replay configuration should be analyzed in detail. In other words, it is also necessary to get the information regarding the category of playback and recording devices while evaluating the performance under different categories of environment.

TABLE V
PERFORMANCE ANALYSIS IN TERMS OF EER (%) FOR DIFFERENT FEATURES ON ENVIRONMENT (E), PLAYBACK (P) AND RECORDING (R) DEVICE OVER ASVSPOOF 2017 VERSION 2.0 DATABASE.

Feature	Configuration	Development Set			Evaluation Set		
		Low	Medium	High	Low	Medium	High
MFCC	E	15.95	19.20	-	7.60	17.16	43.30
	P	0.25	2.78	22.29	2.02	12.01	30.89
	R	11.30	31.95	19.03	19.88	15.53	22.66
CQCC	E	8.22	9.25	-	9.49	11.66	18.07
	P	6.04	6.95	10.15	10.29	11.77	14.51
	R	6.47	14.16	9.51	11.54	8.11	13.99
IFCC	E	15.24	16.84	-	11.31	15.26	23.92
	P	8.73	15.57	18.14	14.73	13.85	17.76
	R	11.33	23.73	18.07	16.11	14.33	16.60
DCTILPR	E	21.16	23.46	-	11.96	13.28	24.81
	P	11.34	6.84	26.29	5.33	4.11	19.28
	R	14.76	35.93	26.02	9.16	15.78	18.19
RMFCC	E	26.11	21.17	-	9.64	17.46	38.12
	P	18.07	5.11	26.98	4.22	10.72	29.44
	R	21.63	44.06	19.87	19.60	14.97	22.52
Fusion	E	4.96	6.64	-	5.80	7.88	16.62
	P	0.76	1.10	7.38	5.44	7.52	11.84
	R	2.07	13.76	6.24	8.14	5.39	10.78

TABLE VI
PERFORMANCE ANALYSIS IN TERMS OF EER (%) FOR DIFFERENT FEATURES ON DIFFERENT REPLAY CONFIGURATIONS BASED ON LOW, MEDIUM (MED) AND HIGH (HIG) CATEGORY OF E-P-R ON ASVSPOOF 2017 VERSION 2.0 DATABASE.

Configuration E-P-R	MFCC	CQCC	IFCC	DCTILPR	RMFCC	Fusion
Development Set						
low-low-low	0.26	6.91	10.52	11.81	19.40	0.86
low-hig-low	16.84	6.40	13.03	14.08	32.06	2.60
low-hig-hig	20.14	9.70	18.61	28.98	25.40	7.38
med-low-low	0.23	3.96	6.23	9.95	16.13	0.32
med-med-hig	2.78	6.95	15.57	6.84	5.11	1.10
med-hig-low	15.42	6.64	14.43	17.89	13.00	3.94
med-hig-med	31.95	14.16	23.73	35.93	44.06	13.76
med-hig-hig	22.62	10.72	17.97	27.76	18.91	6.81
Evaluation Set						
low-low-low	0.39	8.51	10.41	6.36	6.35	3.49
low-low-hig	5.63	13.82	17.59	13.44	7.11	8.51
low-hig-low	13.60	6.87	9.39	9.22	16.66	5.31
low-hig-hig	13.67	6.00	5.37	19.35	15.23	4.31
med-low-low	1.23	10.56	16.52	4.40	4.09	5.54
med-low-med	2.23	8.47	11.79	0.14	1.68	4.09
med-low-hig	2.33	11.20	15.61	5.18	4.21	5.61
med-med-low	6.20	8.03	9.90	2.94	8.18	5.10
med-med-hig	16.99	15.27	17.34	4.86	13.30	9.50
med-hig-low	24.46	17.24	17.68	12.20	27.78	12.08
med-hig-med	31.78	6.28	19.55	31.28	31.03	6.85
hig-hig-low	42.84	13.94	20.78	10.19	33.73	11.41
hig-hig-hig	43.66	26.43	29.44	49.40	46.41	26.09

4) Effect of Different Recording Configurations: The variation in replay configuration in the train and development set of the database is very little compared to that of the evaluation set as can be observed from Table I. If we categorize the replay configurations based on the three categories (low, medium and high) of environment, playback and recording devices we come across 8 broad replay configurations for development set and 13 for evaluation set.

Table VI shows the performance of different features under different replay configurations on ASVspoof 2017 Version 2.0 database. It shows that the long range features CQCC and IFCC are more stable across different conditions and perform better individually compared to the source features as well as MFCC in most of the cases. The MFCC features perform well when the replay condition is less challenging by yielding EER as low as 0.39% on the evaluation set when environment, playback and recording devices all belong to 'low' category. However, under challenging conditions its performance is very poor. Again, the source features DCTILPR and RMFCC perform better on the evaluation set than in development set. This may be due to the involvement of different environment, playback and recording devices in the evaluation set. The models are trained using both train and development set for the studies related to evaluation set. Thus, there is a lot of mismatch in the replay configurations of the evaluation set from that used in for training models for genuine and spoof speech. This in turn helps more for the source features in evaluation set based studies. Further, the results for fusion of the features show that the fusion helps in obtaining an improved performance due to alternate artifacts obtained from each feature. However, the improvement is very little compared to the performance with CQCC feature under the condition when environment, playback and recording devices all belong to 'high' category. This thereby signifies the challenges associated in this field and the focus of future research along such challenging conditions of replay attacks.

V. Conclusions

The replay attack is the most common way to have a spoofing attack by replaying the user's voice to have an unauthorized access to a service. The CQCC features are the most reliable features to identify the spoof attacks as mentioned in the literature. In this work, apart from this CQCC features, instantaneous phase and excitation source features are investigated due to their potential for having some additional artifacts to detect replay speech. The IFCC feature is used as a representation of instantaneous frequency that is computed over long range information of the signal. Similarly, the excitation source features DCTILPR and RMFCC are considered for the detection of replay attacks. The studies with theses features as well as conventional CQCC features are conducted on standard ASVspoof 2017 Version 2.0 database. It has been found that the phase and source features posses definite characteristic information that can be utilized for replay speech detection. Additionally, their fusion all together improves the baseline performance based on CQCC features by a large margin depicting their importance. An analysis on the behavior of these features for different replay configurations is also presented in this work to illustrate their role to specific scenario. The future work will focus on exploring the scope of these phase and source features for spoof detection based on synthetic speech.

Acknowledgment

This research is supported by Programmatic Grant No. A1687b0033 from the Singapore government's Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

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] S. Dey, S. Barman, R. K. Bhukya, R. K. Das, Haris B C, S. R. M. Prasanna, and R. Sinha, "Speech biometric based attendance system," in *National Conference on Communications (NCC) 2014, IIT Kanpur*, 2014.

[4] Rohan Kumar Das, S. Jelil, and S. R. Mahadeva Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, Sep 2017.

[5] 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.

[6] 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.

[7] 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.

[8] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 925–929.

[9] 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.

[10] 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, Dresden, Germany*, 2015.

[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] 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.

[13] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in *ODYSSEY 2018, The Speaker and Language Recognition Workshop, June 26-29, 2018, Les Sables d'Olonne, France*, Les Sables d'Olonne, France, 2018.

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

[15] 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.

[16] 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, phase-Aware Signal Processing in Speech Communication.

[17] 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, phase-Aware Signal Processing in Speech Communication.

[18] Rohan Kumar Das, Abhiram B., S. R. M. Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *Interspeech 2014, Singapore*, 2014, pp. 1836–1840.

[19] M. Sahidullah, D. A. L. Thomsen, R. G. Hautamaki, T. Kinnunen, Z.-H. Tan, R. Parts, M. Pitkanen, M. Sahidullah, D. A. L. Thomsen, R. Gonzalez Hautamaki, T. Kinnunen, Z.-H. Tan, R. Parts, and M. Pitkanen, "Robust voice liveness detection and speaker verification using throat microphones," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 1, pp. 44–56, Jan. 2018.

[20] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *International Carnahan Conference on Security Technology (ICCST) 2011*, pp. 1–8.

[21] ——, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*. Springer, 2011, pp. 274–285.

[22] Z. F. Wang, G. Wei, and Q. H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *International Conference on Machine Learning and Cybernetics 2011*, vol. 4, July 2011, pp. 1708–1713.

[23] Z. Wu, S. Gao, E. S. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. Asia-*

Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2014.

[24] A. Paul, R. K. Das, R. Sinha, and S. R. M. Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *International Conference on Signal Processing and Communications (SPCOM) 2016*, June 2016, pp. 1–5.

[25] J. Gaka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143 – 153, 2015.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] K. Sriskandaraja, G. Suthokumar, V. Sethu, and E. Ambikairajah, "Investigating the use of scattering coefficients for replay attack detection," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 2017*, Dec 2017, pp. 1195–1198.

[31] D. A. Reynolds and R. 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, Jan 1995.

[32] N. Dehak, P. 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, May 2011.

[33] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[34] S. L. Marple, "Computing the discrete-time 'analytic' signal via fft," in *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers (Cat. No.97CB36136)*, vol. 2, Nov 1997, pp. 1322–1325 vol.2.

[35] K. Vijayan, H. Li, H. Sun, and K. A. Lee, "On the importance of analytic phase of speech signals in spoken language recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, Alberta, Canada, April*, 2018, pp. 5194–5198.

[36] Rohan Kumar Das, Debadatta Pati, and S. R. M. Prasanna, "Different aspects of source information for limited data speaker verification," in *National Conference on Communications (NCC) 2015, IIT Bombay*, 2015.

[37] Rohan Kumar Das and S. R. Mahadeva Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016.

[38] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, issue 12, pp. 2471 – 2480, 2013.

[39] A. G. Ramakrishnan, B. Abhiram, and S. R. M. Prasanna, "Voice source characterization using pitch synchronous discrete cosine transform for speaker identification," *JASA Express Letters*, vol. 137, pp. EL469–EL475, 2015.

[40] P. Thvenaz and H. Hgli, "Usefulness of the lpc-residue in text-independent speaker verification," *Speech Communication*, vol. 17, no. 12, pp. 145 – 157, 1995.

[41] D. Pati and S. R. M. Prasanna, "Speaker information from subband energies of linear prediction residual," in *National Conference on Communications (NCC), 2010*, Jan 2010, pp. 1–4.

[42] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. A. L. Thomsen, A. K. Sarkar, Z. Tan, H. Delgado, M. Todisco, N. W. D. Evans, V. Hautamäki, and K. Lee, "Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 539–5399.

[43] The BOSARIS toolkit, (accessed on 10th Dec. 2013). [Online]. Available: www.sites.google.com/ site/bosaristoolkit/

[44] P. Korshunov and S. Marcel, "Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 695–705, June 2017.