A Survey on Replay Attack Detection for Automatic Speaker Verification (ASV) System

Hemant A. Patil and Madhu R. Kamble
Speech Research Lab, Dhirubhai Ambani Institute of Information
and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India
E-mail: {hemant_patil and madhu_kamble }@daiict.ac.in

***Abstract*—In this paper, we present a brief survey of various approaches used to detect replay attack for Automatic Speaker Verification (ASV) system. The replay spoofing attack is the most challenging task to detect as only few seconds of audio samples are required to replay genuine speaker's voice. Due to large availability and the widespread usage of the mobile/smart gadgets, recording devices, it is easy and simple to record and replay the genuine speaker's voice. The challenging task, in replay spoof attack is to detect the acoustical characteristics of the speech signal between the natural and replayed version. The speech signal recorded with the playback device contains the convolutional and additive distortions from the intermediate device. Background noise and channel degradations seriously constrain the performance of the system. The goal of this paper is to provide an overview of the replay attack focusing on 2^{nd} ASVspoof 2017 challenge which is an emerging research problem in the field of anti-spoofing. This paper presents critical analysis of state-of-the-art techniques, various countermeasures, databases, and also aims to present current limitations along with road map ahead, i.e., future research directions in this technological challenging problem.**

I. Introduction

The biometric security system has the pattern of using our own biometric identification and discards the use of cards, passwords as well as the interrogation Q&A that comes when customers forget passwords [1], [2]. Researchers from many fields have applied the recent techniques in each area of biometrics to improve the performance of biometric systems [3]. These technologies have made the use of biometrics in various diverse areas, such as forensics, border and access control, surveillance, e-commerce, etc. [4]. Attacks to such systems are carried out in realistic scenarios for various applications. The research found that 90 % of people are eager to use voice biometric solutions in place of traditional methods of authentication [1]. Among various vulnerabilities, extensive efforts were focused on direct or spoofing attacks. The spoofing gets the advantage of the biometric identification when the data is available at public level and hence, it is one of the most drawbacks of biometrics as well, i.e., "*biometric traits are not secrets*" [4]. Such publicly availability of biometrics is one of the reasons that spoofing has attracted huge interest [5].

The problem nowadays is not the question of whether or not the biometrics can be manipulated, rather the question is to what extent the systems are robust to such kind of attacks and if they are not then what are the possible countermeasures to detect them. Among current concerns of a threat to the systems one of the vulnerabilities is *spoofing* and it is defined as, the speaker masquerading as others in order to gain the access of protected data [6], [7]. The study of spoofing or anti-spoofing have shown interest in recent years, however, the problem is much far away from being solved in real practical applications. Thus, the study of spoofing requires greater attention to be solved.

The goal of the Automatic Speaker Verification (ASV) system is to determine or verify the identity of an individual speaker's voice. The spoofing attacks in ASV or biometrics field in general are considered as part of *presentation attacks* as per International Organization for Standardization (ISO) and International Electro-technical Commission (IEC) [8]. The various spoofing attacks in ASV system are Speech Synthesis (SS) [9], Voice Conversion (VC) [10]–[12], replay [13]–[15], impersonation [16], [17] and twins [18]. Among all the spoofing attacks, replay, SS, and VC present the great risk to the ASV system [19]. The replay Spoof Speech Detection (SSD) task might be the easiest, and common technique to spoof the system that presents a great risk for both text-dependent and text-independent ASV systems [19], [20]. Recently, the 2^{nd} ASVspoof 2017 Challenge [21] was organized as a special session in INTERSPEECH 2017 with a follow up of last two special sessions on spoofing and countermeasures for ASV held during INTERSPEECH 2013 [7] and INTERSPEECH 2015 [22]. The first ASVspoof 2015 Challenge was built upon the text-independent system and concentrated on SS and VC spoofing attacks. The 2^{nd} ASVspoof 2017 Challenge was focused only on replay attack with text-dependent system [23]. This paper is mainly focused on the studies reported on ASVspoof 2017 Challenge database, its problem, various countermeasures proposed, limitations of the database etc.

II. Replay Spoofing Attack

The task of replay spoof detection is to identify whether a given speech signal is recorded from a genuine speaker or an intermediate (recording + playback (spoofed speaker)) device. The genuine speech signal $s[n]$ can be modeled as a convolution of glottal airflow, $p[n]$ and vocal tract impulse response $h[n]$ [24].

$$s[n] = p[n] * h[n]. \quad (1)$$

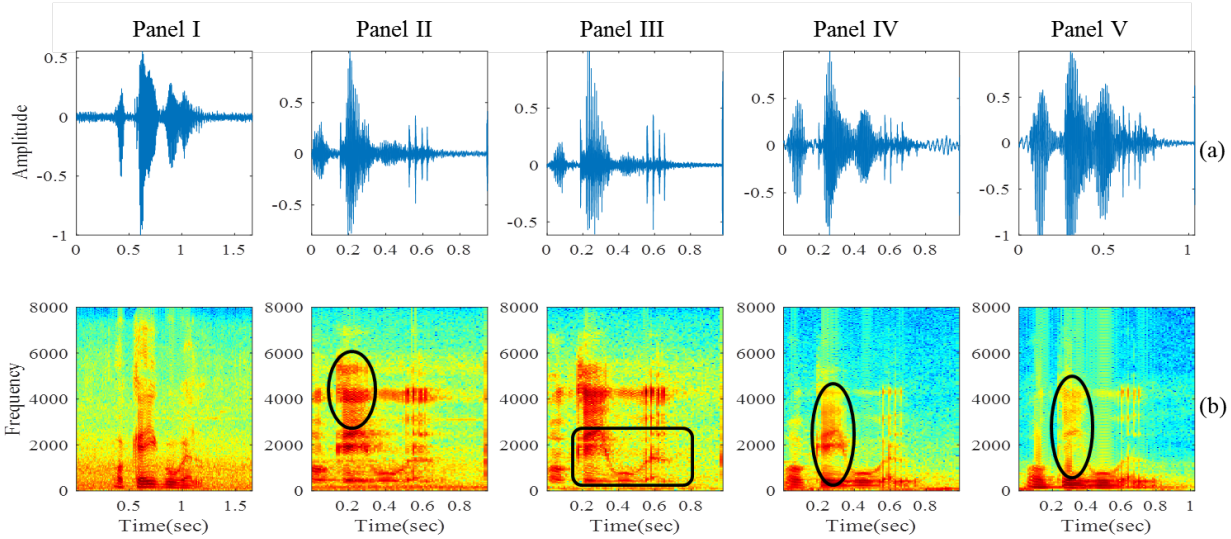

Fig. 1. Spectrographic analysis of natural (Panel I) and replayed speech recorded in various environment conditions, such as Panel II (Balcony), Panel III (Bedroom), Panel IV (Canteen) and Panel V (Office) of utterance, *'Actions speak louder than words'*. Fig. 1(a) time-domain speech signal of natural and replayed speech and Fig. 1(b) corresponding spectral density

Whereas, the replay speech signal $r[n]$ can be modeled as the convolution of the genuine speech signal $s[n]$ and the impulse response of the intermediate devices $\eta[n]$ (playback and recording device) along with propagating environment and is given by:

$$r[n] = s[n] * \eta[n]. \tag{2}$$

where the $\eta[n]$ is the extra convolved components which is combination of impulse responses of recording device $h_{mic}[n]$, recording environment $a[n]$, playback device (multimedia speaker) $h_{spk}[n]$, and playback environment $b[n]$.

$$\eta[n] = h_{mic}[n] * a[n] * h_{spk}[n] * b[n]. \tag{3}$$

The challenging task, in replay spoof attack, is to detect the acoustic characteristics as there is a *imperceptible* difference of the speech signal between the natural and replayed speech. The speech signal recorded with the playback device contains the convolutional and additive distortions from the intermediate device. The most crucial part in the detection of replay attack is the process of feature extraction. To obtain the discriminatory information between natural and replayed speech signal, the focus while extracting the features should represent the spectral characteristics of the intermediate device. The Eq. (3) presents the convolution term that transforms to the additive relation in the cepstral domain and is given by [25]:

$$\mathbf{r} = \mathbf{s} + \mathbf{h}, \tag{4}$$

where **r**, **s** and **h** represents the cepstral vectors of replay, natural speech signal and the impulse response of device, respectively. The features obtained from the vector **h** can be used by subtracting the natural speech signal from the replayed speech signal.

The Fig. 1 shows the spectrographic analysis of natural speech with four different conditions of environment, such as balcony, bedroom, canteen, and office of ASVspoof 2017 Challenge database [21]. The Panel I of Fig. 1 is of the natural speech signal with the corresponding spectrogram of the original speech signal for the utterance, *"Actions speak louder than words"*. It can be observed from the Fig. 1 that there is a difference in terms of temporal-domain as well as in spectral-domain for all Panel I (Natural), Panel II (Balcony), Panel III (Bedroom), Panel IV (Canteen) and Panel V (Office). The spectral pattern varies with different acoustical conditions. The spectral energy density obtained from the speech signal recorded in balcony has the energy largely focused in the higher frequency regions. However, may be because of the acoustic conditions and the intermediate device, the spectral resolution is blurred, and the harmonic pattern is not well preserved for the replayed signal compared to the natural speech signal. The blurring in spectral resolution is may be due to the widening of -3 dB bandwidth of formant peaks of replayed speech as observed in [25].

III. Replay Databases

This section gives an overview of the current existing publicly available databases and the results for the anti-spoofing evaluation purpose. Currently, there are two standard databases that are focused on the replay spoof attacks, namely, AVspoof [6] and ASVspoof 2017 Challenge database [21], [26]. The details of each database are given below:

A. AVspoof Database

The AVspoof database is the first standard database that introduces replay spoofing attacks along with SS and VC spoofing attacks. This database was used in the BTAS 2016

Challenge and the details of the database are given in [6], [27]. The statistics of the database are given in Table I. This database reports a comprehensive variety of presentation attacks. The 'unknown' attacks were introduced in the test set to make the competition more challenging. The organizers of the challenge provided a baseline system which is based on the open source Bob toolbox. The baseline system consist of simple spectrogram-based ratios as features and logistic regression as a classifier.

TABLE I
STATISTICS OF AVSPOOF DATABASE [6]

Subset	# Speakers	# Utterances		
		Genuine	PA Attacks	LA Attacks
Training	10(M)-4(F)	4973	38580	17890
Development	10(M)-4(F)	4995	38580	17890
Evaluation	11(M)-5(F)	5576	43320	20060

PA: physical access, LA: logical access, M: male speaker, F: female speaker

Few of the countermeasures proposed or approached for the AVspoof database are reported in Table II.

TABLE II
COMPARISON OF RESULTS ON EVAL DATASET OF AVSPOOF DATABASE

Feature Set	LA	PA
SCFC [28]	0.00	5.34
RFCC [28]	0.04	3.27
LFCC [28]	0.00	4.73
MFCC [28]	0.00	5.43
IMFCC [28]	0.00	4.09
SSFC [28]	0.70	4.70
SCMC [28]	0.01	3.95
CQCC [28]	0.66	3.8
μ [29]	0.51	0.04
σ [29]	2.03	4.65
μ, σ [29]	0.18	0.04

B. ASVspoof 2017 Challenge Database

This database is mainly based on the RedDots corpus, and its replayed version, which is basically text-dependent database [26], [30]. The spoofed data was recorded through a variety of different environments in the H2020-funded OCTAVE project. The RedDots corpus was replayed through different replay configurations consisting of varied devices, recording devices, and the loudspeakers [21]. However, the organizers released a modified ASVspoof 2017 Version 2.0 database. The organizers corrected data anomalies (something that deviates from what is expected) detected in the post evaluation. These were patched in a second version of the database released in [31]. The difference in the data of original and modified database with the speech signal from the evaluation set (E_1010850) is shown in Fig. 2. We can clearly see the difference of two speech signal that are differed. Along with the corrected data more detailed description of recording, playback devices and acoustic environments are also corrected and reported in version 2.0 database [31].

The number of speakers and the number of trials in each subset of ASVspoof 2017 Version 2.0 database are summarized in Table III. This database is much smaller compared to

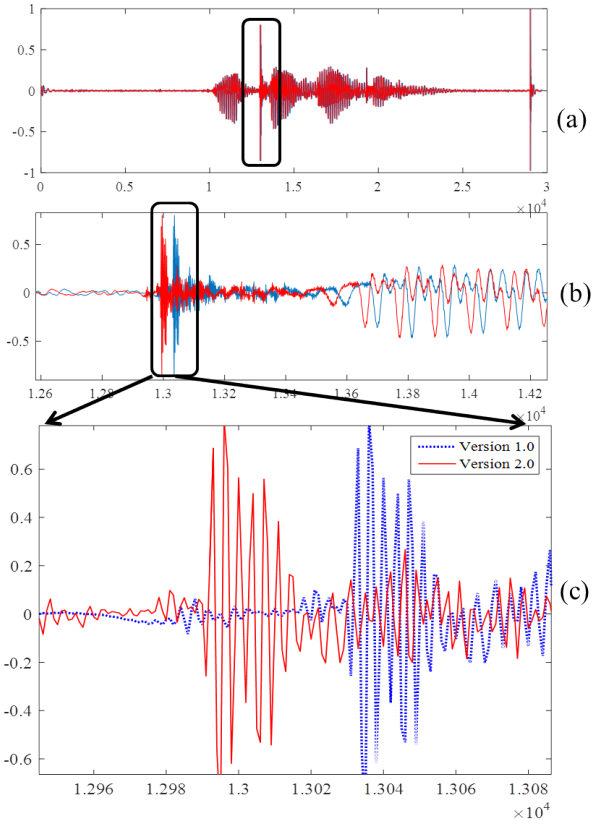

Fig. 2. Speech signal from evaluation set (E_1010850) of ASVspoof 2017 challenge database version 1.0 (blue signal) and version 2.0 (ref signal).

TABLE III
STATISTICS OF ASVSPOOF 2017 CHALLENGE VERSION 2.0 DATABASE [21], [31]

Subset	# Speakers	# Utterances	
		Genuine	Spoofed
Training	10	1507	1507
Development	8	760	950
Evaluation	24	1298	12008

the ASVspoof 2015 Challenge database [22] and thus, makes easier to handle the database and requires less time complexity for performing the experiments and to investigate the problem. This database contains only male speakers with 10 speakers in training set having 1507 utterances of natural and spoofed speeches. The development set, consists of 760 and 950 for natural and spoofed trials, while evaluation set consists of 1298 and 12008 for natural and spoofed trials.

C. Performance Evaluation Metrics

1) **Equal Error Rate (EER):** The performance of a biometric system is generally calculated by the Equal Error Rate (EER). It corresponds to a threshold at which the False Acceptance Rate (FAR) is *equal* to the False Rejection Rate (FRR). The FAR and FRR of a verification system define different operating points on

the Detection-Error Trade-off (DET) curve [32]. In both speaker verification and spoofing detection, two types of errors exist, namely, FAR, i.e., impostor or spoofed speech is accepted as natural or human and FRR, i.e., natural or human speech is accepted as impostor or spoofed speech. There is a trade-off between FAR and FRR and both error rates are found to be equal to a threshold value. The operating point of the error at which both FAR and FRR are equal is known as the Equal Error Rate (EER). In DET curve, the lower the FAR means higher security against the spoof speech and lower the EER means higher convenience to the ASV system performance.

2) **Tandem-Detection Cost Function (t-DCF):** The studies have reported the performance for the ASV system when it is used with the countermeasures. The countermeasure used should not affect the performance of the ASV system. However, the features used should have lower FRR when they are used with joint ASV system to provide user convenience by lesser rejections of genuine trials. Hence, with the progress made in the research of spoofing detection, a evaluation metrics must evolve to reflect the whole system performance. The study reported in [33] proposed a tandem detection cost function (t-DCF) metric it is an elegant solution to the assessment of combined spoofing countermeasures and ASV system.

IV. Replay Attack Detection Approaches

The replay spoofing attack came up in a very big way during the 2nd ASVspoof 2017 Challenge organized by INTERSPEECH 2017 [21]. The challenge brought many of the researchers to a common platform to discuss their individual approaches to solve the replay attack problem. Several approaches were proposed in the challenge, however, the replay attack was found to be the challenging task to achieve a unique countermeasure. To the best of the authors' knowledge, yet the replay problem is far away to obtain a better countermeasures. Given the imperceptible nature of the replay speech, various approaches are currently being explored. We discuss two aspects in the field of replay research: (i) design of feature extractor for replay speech signal detection and (ii) pattern classifiers. It also includes machine learning (in particular, deep learning) algorithms for capturing and modeling the feature patterns.

A. Baseline System

A baseline system was provided to the participants by the ASVspoof 2017 Challenge organizers. The system uses Constant Q Cepstral Coefficients (CQCC) features and Gaussian Mixture Model (GMM) as a classifier [21], [34]. The CQCC feature sets are extracted with F_{max}=FNY_Q, where FNY_Q is the Nyquist frequency of 8 kHz. Features extracted with *30*-DCT static coefficients (with log-energy), resulting in total *90*-D feature vector. The baseline system gave an EER on development and evaluation set of 10.35 % and 28.48 %, respectively. However, the baseline system do not perform better for replay detection task and hence, cannot be used as a good countermeasures. Furthermore, the enhanced baseline system on ASVspoof 2017 version 2.0 database was reported in the form of log-energy coefficients and cepstral mean and variance normalization (CMVN) in addition to an alternative i-vector backend [31]. When compared to the previous baseline score of 28.48 % the best enhanced result was 12.2 % and was found to have relative improvement of approximately 50 %.

B. Features and Classifiers for Replay SSD

Several acoustic feature sets were approached to detect the replay spoof speech during the challenge [35]–[39]. The researchers found that the higher frequency regions are more important than the lower frequency regions [38], [40]. It is due to the fact that the spectral energy density in the higher frequency regions have high energy and have more *blunt/smearing* of the harmonics for replayed speech compared to the natural spectral energy density as shown in Figure 1. The Instantaneous Frequency (IF)-based features along with the significance of Temporal Fine Structure (TFS) in the form of IF were also explored in [37], [41]. The use of CMVN technique was found to be effective to reduce the EER further, as the additional channel effects accumulated due to playback and recording device of the speech signal in different acoustical environments are attenuated that helps to detect the replay attack [35], [38], [42]. However, it was observed that the features which performed best on the ASVspoof 2015 Challenge database (that were focused on only SS and VC) such as, CQCC, LFCC, CFCC-IF, etc. did not give better results on ASVspoof 2017 Challenge database. The summary of various feature sets, classifiers and feature dimension on development and evaluation set are reported in Table IV. It can be observed from Table IV that most of the participants have explored their front-end features with GMM as a classifier. However, very few of the systems used complex classifiers, such as Deep Neural Network (DNN), Resnet Neural Network (ResNet), Convolutional neural network (CNN), Recurrent Neural Network (RNN), etc, to obtain a lower EER.

V. Limitations and Technological Challenges

In this Section, we summarize the current findings in this problem, and also discuss some of their limitations along with possible future research directions.

1) **Why Replay Detection is Challenging ?**
As discussed in Section 2. Eq. (3) indicates that replayed speech is expressed via a convolutional model. In order to detect replay speech, we need to capture the impulse response of the intermediate device, acoustic environment, etc. Thus, w.r.t Eq. (3) this is a *blind* deconvolution problem and hence, getting exact deconvolution of $h[n]$ from $r[n]$ is still a challenge in signal processing .

2) **Robustness in ASV implies Vulnerability for Replay Spoof Speech Detection**

TABLE IV
COMPARISON OF RESULTS ON DEV AND EVAL DATASET OF ASVSPOOF 2017 CHALLENGE DATABASE

Feature Set	Classifier	Dev	Eval
CQCC (BL) [21]	GMM	10.35	28.48
ESA-IFCC [43]	GMM	04.12	12.79
VESA-IACC [44]	GMM	6.12	11.94
LFCC [35]	GMM	10.31	16.54
SCFC [35]	GMM	24.51	24.83
SSFC [35]	GMM	12.81	22.38
IMFCC [35]	GMM	03.85	30.91
SCMC [35]	GMM	09.32	11.49
RFCC [35]	GMM	06.91	11.90
MFCC [35]	GMM	07.76	27.12
CQCC [36]	DNN	05.18	19.41
CQCC [36]	ResNet	05.05	18.79
MFCC [36]	ResNet	10.95	16.26
CQCC (6-8 kHz) [40]	GMM	05.13	17.31
Norm. CQCC [38]	GMM	13.70	28.50
HFCC [38]	GMM	05.9	23.90
DA-CQCC [45]	GMM	07.01	19.18
DA-CQCC [45]	ResNet	06.32	23.14
SFCC-D [39]	GMM	02.35	20.20
SFCC-D [39]	BLSTM	03.66	22.40
VESA-IFCC [37]	GMM	4.61	14.06
CFCC-IF [37]	GMM	06.80	34.49
FFT features [42]	LCNN	04.53	07.37
CQT features [42]	LCNN	04.80	16.54
LPCC [42]	SVM *i-vector*	09.80	12.54
SCC [46]	GMM	3.16	19.79
ConvRBM-CC [47]	GMM	0.82	8.89
LFMGDCC [48]	GMM	20.70	20.84
EMDCC [49]	GMM	28.48	28.06
LFRCC [50]	GMM	8.38	22.28
GD Spectrum [51]	Attention ResNet	**0.0**	**0.0**

In practice, we would like ASV system to be robust against variations, such as microphone and transmission channel, intersession, acoustic noise, speaker aging, etc. This robustness makes ASV system to be vulnerable to various spoofing attacks (especially replay) as it tries to nullify these effects and makes replayed speech more closer to the natural speech. Thus, this robustness in ASV system makes replay detection all the more difficult i.e., technologically challenging which is why newer approaches are required to alleviate this difficulty.

3) **Exploiting of Specific Frequency Region: Why ?**
In practical scenarios, a replay will be done by and enlarge in air medium that contains the air particles having mass and springiness and thus, a slug of air will be responsive to a particular frequency band which will emphasize onto the spectrum of the replayed speech. The investigation for replay detection has shown the significance of selecting particular frequency regions [38], [40].

4) **Lack of Exploiting Excitation Source Information**
Less amount of work is done in using excitation source assuming that the Glottal Closure Instants (GCI) are having sharp impulse-like nature for voiced speech. The spectrum of the glottal source (Glottal Flow Waveform (GFW)) for voiced speech is expected to have *harmonic* structure in the frequency-domain. In particular, the influence of formant structures is suppressed (if not removed completely) due to nonlinear source filter interaction. Thus, any deviation from degradation in the harmonic structure could capture the signature of replayed speech than its natural counterpart. To best of authors knowledge, there is no any work reported in analyzing this particular aspects. We believe several source information such as Linear Prediction (LP) residual, Teager Energy Operator (TEO) profile and its Variable length version (VTEO) profile, etc, could be explored.

5) **Exploring Phase-based Features**
It is important to note that phase-based features (either time-domain analytic or frequency-domain) could capture different kind of information in spoofed speech depending upon type of spoof. For example, in unit selection-based synthesis (USS), when the speech sound units are picked up by optimizing target cost, since these units are recorded in different sessions and hence, in synthesized voice it will have *linear phase* informations [52]. On the other hand, in replayed speech, the impulse response of acoustic environment (say room), gets convolved with the natural speech. Impulse response of an acoustic system (in this case room) is infinite duration, i.e., IIR in nature (due to infinite transmissions and reflections). Thus, the nonlinear phase in frequency-domain of this acoustic system is added to the phase of natural speech. Thus, there is need for more deeper investigation in the phase-based research for replay SSD.

6) **Which Classifier: Conventional or Neural Network ?**
Conventional classifiers (GMM, GMM-UBM, SVM) do not have feature abstraction capability that is found in deeper models (DNN, CNN, RNN, LSTM). However, there has been not much study of classifier-level fusion. To explore the possible complementary information that is captured by a specific mathematical structure of a particular classifier. For example, GMM classifier captures only first and second-order statistics of feature vectors. On the other hand, neural network-based classifier captures non-linear aspects of features. Thus, this demands a more deeper investigations on the suitability of a particular classifier for this problem.

7) **Joint Protocol for Spoof Detection and ASV System**
The integrated system accepts a claimed identity only if it is been accepted by the ASV system and is classified as human speech by the countermeasure. A good countermeasure reduces the FARs by rejecting a spoofed speech generated by the machines, i.e., non-human speech and hence, spoofing attacks are reduced significantly and thus, indicates the need of effective countermeasure for detecting spoofing attacks. A joint protocol approach having a countermeasure and speaker verification system retains low computational complexity.

8) **Comparison of Human *vs.* Machine-learning**
To compare the countermeasure results obtained from automatic machine-based techniques there is need to confirm the results by verifying the speech samples by human listeners as done for the earlier ASVspoof 2015 Challenge database (SS and VC spoofs) reported in [53].

9) **Exploring Frequency Scales**
In ASVspoof 2017 Challenge campaign not much attention was given on investigating effectiveness of linear frequency scale as opposed to state-of-the-art Mel frequency scale. As the replayed speech signal gets affected in medium and high frequency regions it is important to have good resolution in those frequency regions. However, this is not at all possible with the traditional Mel frequency scale. Thus, replay spoof detection problem demand an investigation of appropriate frequency scale.

10) **Exploiting Knowledge of Speech Production and Random Process**
When a person repeats an utterance each repetition is unique (due to naturalness) and is referred to as sample function of speech production mechanism that is modeled as a random process. GMM requires only first two statistical moments (mean and standard deviation) that captures *almost* all the characteristics of the features. The models obtained from the training set of natural speech signals have the distribution that captures great variability in various sample functions of natural speech mechanism. The replayed speech signals are lesser in duration compared to corresponding natural speech and do not have exactly similar variability in its generation rather it replicates the sample function of natural speech.

In summary, although significant progress is made for detecting spoofed speech in a clean environment, the problem is yet to be solved especially for signal degradation and channel mismatch conditions. This study presented an brief technological challenges associated with design of replay detector.

ACKNOWLEDGMENTS

The authors would like to thank the organizers of the special session on replay attack APSIPA ASC 2018. In addition, they also thank University Grants Commission (UGC) for providing Rajiv Gandhi National Fellowship (RGNF) and authorities of DA-IICT Gandhinagar for their kind support to carry out this research work. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

REFERENCES

[1] B. Beranek, "Voice biometrics: success stories, success factors and what's next," *Biometric Technology Today*, vol. 2013, no. 7, pp. 9–11, 2013.

[2] N. Memon, "How biometric authentication poses new challenges to our security and privacy [in the spotlight]," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 196–194, 2017.

[3] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: A tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.

[4] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[5] S. Bruce, "Inside risks: the uses and abuses of biometrics," *Communications of the ACM*, vol. 42, no. 8, 1999.

[6] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Virginia, USA, 2015, pp. 1–6.

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

[8] J. Koppell, "International organization for standardization," *Handb Transnatl Gov Inst Innov*, vol. 41, p. 289, 2011.

[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[10] Y. Stylianou, "Voice transformation: A survey," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, Taipei, Taiwan, China, 2009, pp. 3585–3588.

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

[12] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Voice anti-spoofing," *Handbook of biometric antispoofing, S. Marcel, SZ Li, and M. Nixon, Eds. Springer*, 2014.

[13] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH, Lyon, France*, 2013, pp. 940–944.

[14] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *IEEE Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA)*, Chiang Mai, Thailand, 2014, pp. 1–5.

[15] 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, 2016.

[16] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, 2004, pp. 145–148.

[17] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *INTERSPEECH*, Lyon, France, 2013, pp. 930–934.

[18] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[19] 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.

[20] Villalba, Jesús and Lleida, Eduardo, "Detecting replay attacks from far-field recordings on speaker verification systems," in *European Workshop on Biometrics and Identity Management*. Roskilde, Denmark: Springer, 2011, pp. 274–285.

[21] 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 *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1–6.

[22] Z. Wu, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge,," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2037–2041.

[23] T. Kinnunen, N. Evans, J. Yamagishi, K. A. Lee, M. Sahidullah, M. Todisco, and H. Delgado, "ASVspoof 2017: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," 2017. [Online]. Available: Last Access 11-Jan-2018, http://www.asvspoof.org/

[24] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice. 1^{st} Edition*. Pearson Education India, 2006.

[25] B. S. M. Rafi, K. S. R. Murty, and S. Nayak, "A new approach for robust replay spoof detection in asv systems," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Montreal, Canada: IEEE, 2017, pp. 51–55.

[26] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, 2017, pp. 5395–5399.

[27] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. Mello, R. Violato, F. Simões, M. Uliani Neto, M. de Assis Angeloni, J. A. Stuchi *et al.*, "Overview of BTAS 2016 speaker anti-spoofing competition," Idiap, Tech. Rep., 2016.

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

[29] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Presentation attack detection using long-term spectral statistics for trustworthy speaker verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016, pp. 1–6.

[30] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The RedDots data collection for speaker recognition." in *INTERSPEECH*, Dresden, Germany, 2015, pp. 2996–3000.

[31] 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*, Les Sables d'Olonne, France, 2018, pp. 296–303.

[32] A. Martin *et al.*, "The DET curve in assessment of decision task performance," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 1895–1898.

[33] 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 *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, pp. 312–319.

[34] 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.

[35] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 7–11.

[36] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 102–106.

[37] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel variable length Teager energy separation based instantaneous frequency features for replay detection," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 12–16.

[38] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 97–101.

[39] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 107–111.

[40] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Gałka, "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 27–31.

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

[42] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 82–86.

[43] M. R. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay spoof speech detection," *to appear in INTERSPEECH*, Hyderabad, India, 2018.

[44] M. R. Kamble and H. A. Patil, "Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection," *to appear in INTERSPEECH*, Hyderabad, India, 2018.

[45] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 17–21.

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

[47] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," *To appear in INTERSPEECH*, Hyderabad, India, 2018.

[48] K. Srinivas and H. A. Patil, "Relative phase shift features for replay spoof detection system," *to appear in SLTU*, pp. 1–5, 2018.

[49] P. A. Tapkir and H. A. Patil, "Novel empirical mode decomposition cepstral features for replay spoof detection," *to appear in INTERSPEECH*, Hyderabad, India, September 2-6, 2018.

[50] H. Tak and H. A. Patil, "Novel linear frequency residual cepstral features for replay attack detection," *to appear in INTERSPEECH*, Hyderabad, India, September 2-6, 2018.

[51] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," *to appear in INTERSPEECH*, Hyderabad, India, 2018.

[52] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 232–239, 2001.

[53] Z. Wu *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.