*SIP* (2020), vol. 9, e2, page 1 of 18 © The Authors, 2020.

doi:10.1017/ATSIP.2019.21

OVERVIEW PAPER

# Advances in anti-spoofing: from the perspective of ASVspoof challenges

MADHU R. KAMBLE,[1] HARDIK B. SAILOR,[2] HEMANT A. PATIL[1] AND HAIZHOU LI[3]

*In recent years, automatic speaker verification (ASV) is used extensively for voice biometrics. This leads to an increased interest to secure these voice biometric systems for real-world applications. The ASV systems are vulnerable to various kinds of spoofing attacks, namely, synthetic speech (SS), voice conversion (VC), replay, twins, and impersonation. This paper provides the literature review of ASV spoof detection, novel acoustic feature representations, deep learning, end-to-end systems, etc. Furthermore, the paper also summaries previous studies of spoofing attacks with emphasis on SS, VC, and replay along with recent efforts to develop countermeasures for spoof speech detection (SSD) task. The limitations and challenges of SSD task are also presented. While several countermeasures were reported in the literature, they are mostly validated on a particular database, furthermore, their performance is far from perfect. The security of voice biometrics systems against spoofing attacks remains a challenging topic. This paper is based on a tutorial presented at APSIPA Annual Summit and Conference 2017 to serve as a quick start for those interested in the topic.*

**Keywords:** Automatic speaker verification (ASV), Spoofing attacks, Database, Performance evaluation metric, Countermeasures

Received 19 February 2019; Revised 9 October 2019

## I. INTRODUCTION

A biometric system aims to verify the identity of an individual from their behavioral and/or biological characteristics [1,2]. The body traits that can be used for biometric recognition are classified into anatomical and behavioral characteristics [3]. Anatomical traits include face [4], fingerprint [5], iris [6], palmprint [7], hand geometry [8], and ear shape [9]; while gait [10], signature [11], and keystroke [12] dynamics are some of the behavioral characteristics [13]. Voice biometrics can be considered either as an anatomical or as a behavioral characteristics [3]. Robustness and security are two important factors as far as system deployment is concerned.

Speaker recognition usually refers to both speaker identification and speaker verification. A speaker identification system identifies who the speaker is, while an automatic speaker verification (ASV) system decides if an identity claim is true or false. The former is a multi-class classification problem, while the latter is a hypothesis test. A general ASV system is robust to zero-effort impostors, they are vulnerable to more sophisticated attacks. Such vulnerability represents one of the security concerns of ASV systems.

Spoofing involves an adversary (attacker) who *masquerades* as the target speaker to gain the access to a system [14–16]. The spoofing attacks against an ASV system or biometric system in general are considered as a part of *presentation attacks* as per International Organization for Standardization (ISO) and International Electro-technical Commission (IEC) [17]. As biometric information of a person can be easily obtained, spoofing attacks are inevitable [18]. Such spoofing attacks can happen to various biometric traits, such as fingerprints, iris, face, and voice patterns. Figure 1 shows some examples how the original biometric patterns can be spoofed with different techniques. In this paper, we are focusing only on the voice-based spoofing and anti-spoofing techniques for ASV system.

The spoofed speech samples can be obtained through speech synthesis, voice conversion, or replay of recorded speech. Depending upon how the spoof samples are presented to ASV system the attacks are broadly classified into two categories, namely, direct attacks and indirect attacks. In direct attacks (also called as Physical Access (PA) attacks), the samples are applied as input to the ASV system through the sensor, i.e. at the microphone and transmission-level. In indirect attacks (also called as Logical Access (LA) attacks), the samples are involved by passing the sensor, i.e. ASV system software process, access during feature extraction, interfering with the models, and at the decision or score computation as shown in Fig. 2 [21].

[1] Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India
[2] University of Sheffield, UK
[3] National University of Singapore (NUS), Singapore

**Corresponding author:**
Madhu R. Kamble
Email: madhu_kamble@daiict.ac.in, mk310191@gmail.com

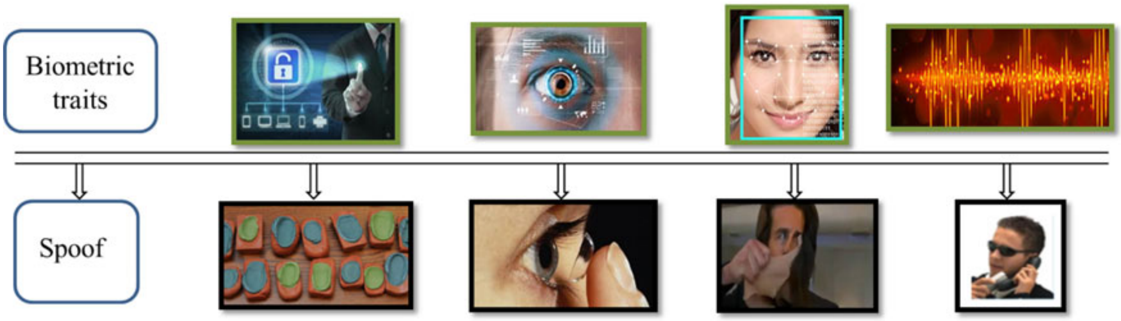

**Fig. 1.** Biometric identification along with spoofing techniques for fingerprint, iris, face, and voice. (Images are adapted from [19].)

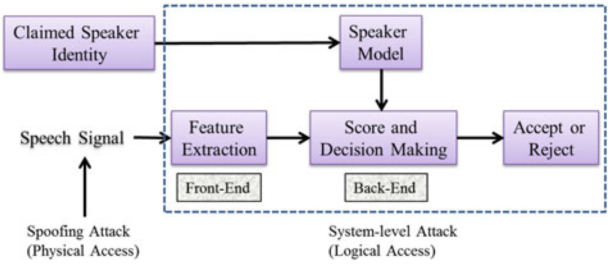

**Fig. 2.** Brief illustration of an Automatic Speaker Verification (ASV) system. After [20].

To objectively report the research progress, there is a need to provide a common dataset along with performance metric to evaluate the spoofing countermeasures. This was also discussed in the special session on spoofing and countermeasures for ASV held during INTERSPEECH 2013 [15]. This special session motivated the researchers to organize the first ASVspoof 2015 Challenge held in INTERSPEECH 2015 [22]. The database released in this challenge contains two types of spoofing attacks, namely, synthetic speech (SS) and voice conversion (VC). As a follow-up, the second and third challenges were organized during INTERSPEECH 2017 and INTERSPEECH 2019, respectively [24,23]. The historical developments and key milestones of the ASVspoof initiative are illustrated in Fig. 3.

We have seen a surge of research papers on spoofing detection in scientific conferences, such as APSIPA Annual Summit and Conference [24], ICASSP, INTERSPEECH, and special issues in scientific journals, such as IEEE Transactions of Information Forensics special issues on Biometrics Spoofing and Countermeasures [25], IEEE Signal Processing Magazine special issue on Biometric Security and Privacy Protection [26], IEEE Journal on Selected Topics in Signal Processing special issue on Spoofing and Countermeasures for Automatic Speaker Verification [27], Special issue on Speaker and language characterization and recognition: voice modeling, conversion, synthesis, and ethical aspects [28], and Special issue on Advances in Automatic Speaker Verification Anti-spoofing [29]. This article provides an overview of the recent advances, and discusses the challenges.

A general discussion on biometrics and spoofing attacks was presented in [2]. The first survey paper on the ASVspoof challenge [20] discusses the past work and identifies priority research directions for the future [30] and presents the details of the dataset, protocols, and metrics of the ASVspoof 2015 challenge. It also provides a detailed analysis of the participating systems in the challenge.

A recent survey paper [31] compares different countermeasures for both SS and replay detection. In particular, it discusses various classical representation learning approaches for SS and replay detection. This overview is an extension to [31] that provides both historical and technological perspectives about the recent progress.

The organization of rest of the paper is as follows: The discussion of various spoofing attacks is presented in Section II. The discussion of different spoofing challenges and performance evaluation metrics are discussed in Section III. Furthermore, in Section IV and Section V, we discussed different countermeasures approaches for synthetic and replay spoof speech detection (SSD) task. In this section, we represented the countermeasures in both classical and representation learning approaches for

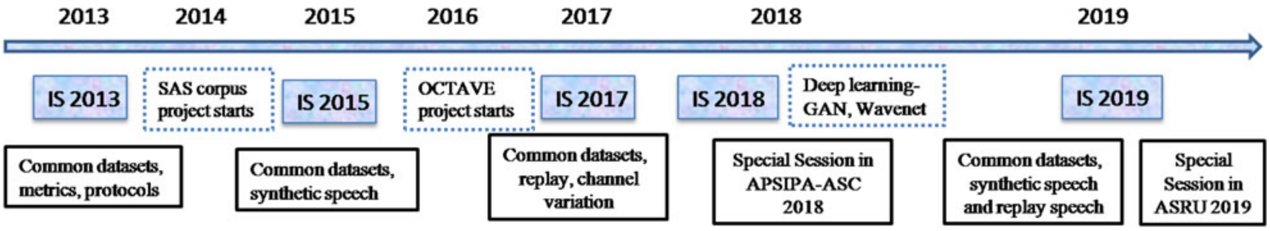
2013
2014
2015
2016
2017
2018
2019
IS 2013
SAS corpus project starts
IS 2015
OCTAVE project starts
IS 2017
IS 2018
Deep learning-GAN, Wavenet
IS 2019
Common datasets, metrics, protocols
Common datasets, synthetic speech
Common datasets, replay, channel variation
Special Session in APSIPA-ASC 2018
Common datasets, synthetic speech and replay speech
Special Session in ASRU 2019

**Fig. 3.** The selected chronological progress in ASVspoof for voice biometrics. In INTERSPEECH 2013, a special session was organized and Spoofing and Anti-Spoofing (SAS) corpus of speech synthesis and voice conversion spoofing data was created. The first ASVspoof challenge was held in INTERSPEECH 2015. In 2016, the OCTAVE project started which focused on only replay spoofing data resulting in the second edition of ASVspoof challenge in INTERSPEECH 2017. The follow-up third ASVspoof 2019 challenge was on physical and LA attacks going to be held during INTERSPEECH 2019 [23]. IS indicates INTERSPEECH.

https://doi.org/10.1017/ATSIP.2019.21

spoofing detection. The Section VI describes the limitation, technological challenges along with future research directions in spoofing research, and finally, we summarize the paper in Section VII.

## II. ASV SYSTEM: SPOOFING ATTACKS

In the literature, the spoofing attacks are broadly classified into four types, namely, impersonation, SS, VC, and replay. Few of the spoofing algorithms used for spoofing attacks are shown in Fig. 4. The detailed description of each spoofing attack is discussed next.

### A) Impersonation

Impersonation is defined as the process of producing the similar voice pattern and speech behavior of the target speaker's voice [33–35]. This can be done either by professional mimics/impersonator (by utilizing behavioral characteristics) or by twins (by utilizing physiological characteristics) [36]. The impersonators do not require any technical background or machines to imitate the target speaker. The study in [37] found that if the impostor is aware of the claimed speaker's voice and also carries similar voice pattern could crack the biometric system. For better imitation, the professional imitator tries to mimic the prosodic features of a target speaker [38]. Professional voice imitator intend to mimic the claimed speaker's prosody, accent, pronunciation, lexicon, and other high-level speaker traits. Such imitation may mislead human perception, however, it is less effective in attacking speaker verification systems because most speaker verification systems are based on spectral features to make decisions. Just like twins attacks, in impersonation attacks, the system is presented with natural human speech. A system to detect unnatural speech does not help. As it takes special training to impersonate someone's voice, impersonation attack is not considered as a common threat to speaker verification systems.

In speaker recognition, we aim to extract the unique speaker features from speech data. However, the speaker features become less unique between the twins [39]. Generally, spectrographic analysis is used to identify the speaker's voice. In the case of identical twins, the same technique fails to perform [40]. The study reported in [41] states that the pattern of speech signals, pitch ($F_o$) contours, formant contours, and spectrograms for identical twin speakers are very similar, if not identical. Due to lack of uniqueness, the FAR increases for identical twins verification. Recently, the Voice ID service was launched by HSBC's phone banking business [42,43]. It failed to recognize true speaker [44]. Similar twins fraud was studied in other biometrics literature as well [39]. The identical twins do have a similar spectrographic pattern, however, the speaker verification technology has seen a significant reduction in fraud, and has proven to be more secure than PINS, passwords, and memorable phrases. In twins attacks, the system is presented with natural human speech, a SS detection mechanism will not enhance the security of the system. To distinguish between the twins, further study on discriminative speaker features is required or more study in this direction is required as observed four decades earlier in [36].

### B) Synthetic speech

SS is also known as Text-To-Speech (TTS), which takes text as input and generate speech as output. It emulates a human vocal production system and represents a genuine threat. SS is now able to generate high-quality voice due to recent advances in unit selection [45], statistical parametric [46], hybrid [47], and DNN-based TTS methods. Recently, deep learning-based techniques, such as Generative Adversarial Network (GAN) [48], Tacotron [49], Wavenet [50], etc., are able to produce very natural sounding speech both in timbre

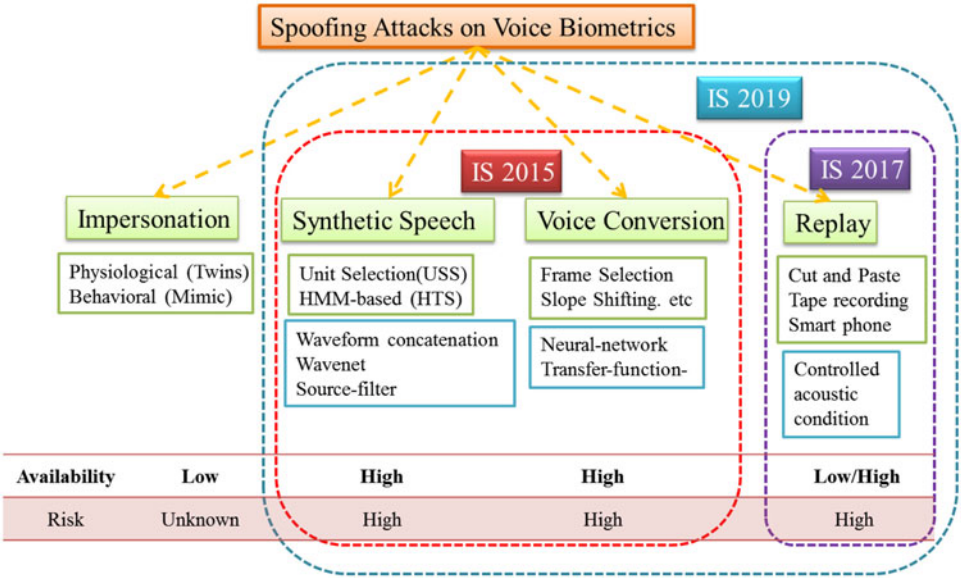

**Fig. 4.** Different spoofing attacks on voice biometrics along with their availability and risk factor. IS: INTERSPEECH. Adapted from [32].

https://doi.org/10.1017/ATSIP.2019.21

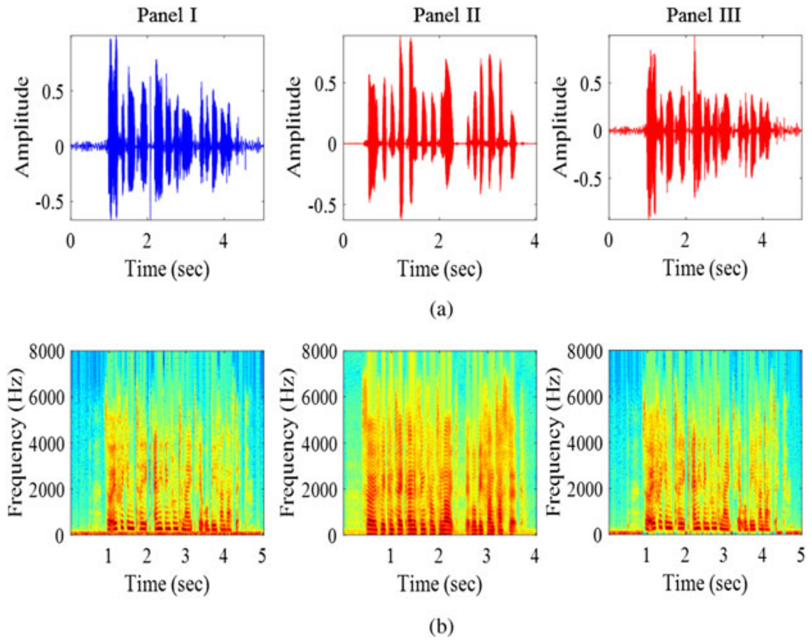

**Fig. 5.** Spectral energy densities of natural (Panel I), synthetic speech (Panel II), and voice converted speech (Panel III). (a) Time-domain speech signal, and (b) corresponding spectral energy density.

and prosody. SS uses properties of a claimed speaker's voice characteristics and spectral cues of the natural speech. The spectral energy density of natural (Panel I) and synthetic (Panel II) speech signal are shown in Fig. 5. It is clearly observed that the distributions of spectral energies are very different between the natural speech and SS. The research on SS detection has been focused on how to detect the artifacts that exist in the SS samples. More technical description of algorithms are reported in [51,52].

## C) Voice conversion

VC is the process of converting the source speaker's voice to a sound similar to the target speaker's voice [34,53,54]. VC deals with the information that relates to the segmental and suprasegmental features and keep the language content similar [55]. Earlier studies include statistical techniques, such as Gaussian Mixture Model (GMM) [56], Hidden Markov Model (HMM) [57], unit selection [58], principal component analysis (PCA) [59], and Non-negative matrix factorization (NMF) [60] for VC task. Recently, DNN [61], Wavenet [50], and GAN [48] represent a technology leap.

Studies also reported in the area of signal processing techniques, such as vector quantization [62] and frequency warping [63]. The research on VC detection has also been focused on how to detect the artifacts arising from the VC process. One example of the converted speech is illustrated in Panel III of Fig. 5. More technical description of converted voices are reported in [51,55].

## D) Replay

One of the most accessible spoofing is replay attack. The attacker replays a pre-recorded voice from the target speaker to the system to gain access [64–66]. Such attack is meaningful only for text-dependent speaker verification systems. With high-quality record-replay audio device, the replayed speech is highly similar to the original speech, spectral content will change slightly due to device impulse response. Hence, replay is a serious adversary to text-dependent speaker verification system.

The genuine speech signal $s[n]$ can be modeled as a convolution of glottal airflow, $p[n]$ and vocal tract impulse response, $h[n]$ [67], i.e.

$$s[n] = p[n] * h[n]. \tag{1}$$

On the other hand, the replay speech signal, $r[n]$ can be modeled as the convolution of the genuine speech signal $s[n]$, and the impulse response, $\eta[n]$ of the intermediate devices (playback and recording device) along with propagating environment and additive noise, $N[n]$ of [68] and is given by:

$$r[n] = s[n] * \eta[n] + N[n], \tag{2}$$

where the $\eta[n]$ is the extra convolved components which is a combination of impulse responses of recording device $h_{mic}[n]$, recording environment $a[n]$, playback device (multimedia speaker) $h_{spk}[n]$, and playback environment $b[n]$ [68].

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

In replayed speech detection, we hope todetect the presence of the channel and noise distortion to the original speech signal [69]. The speech signal recorded with the playback device contains the convolutional and additive distortions from the intermediate device and background [68]. An important part in the detection of replay attack is the process of feature representation. To obtain the discriminative information between natural and replayed speech

https://doi.org/10.1017/ATSIP.2019.21

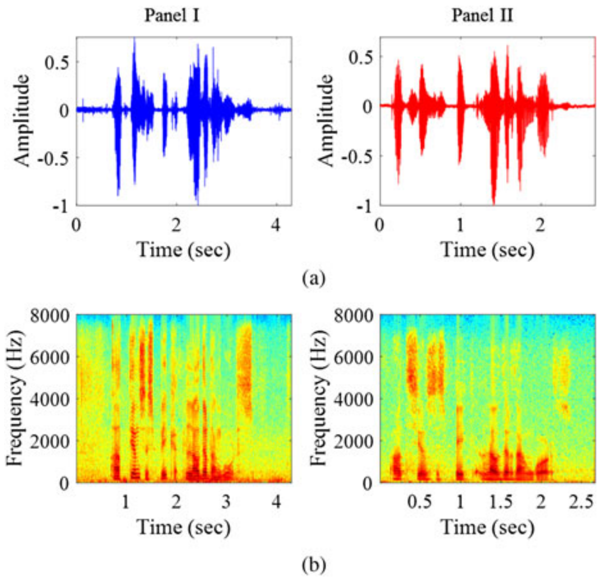

**Fig. 6.** Spectral energy densities of natural (Panel I) and replay speech (Panel II). (a) Time-domain speech signal, and (b) spectral energy density.

signal, one should focus on the spectral characteristics that represent the information of the intermediate devices [70]. Figure 6 shows the spectrographic analysis of natural speech and replay speech signal taken from the ASVspoof 2017 Challenge database [24]. The Panel I of Fig. 6 shows the natural speech signal with the corresponding spectrogram of the natural speech signal for the utterance, "*Actions speak louder than words*", and similarly Panel II is for the replayed speech signal. It can be observed from Fig. 6 that there is a difference in temporal as well as in spectral representation between Panel I (natural) and Panel II (replay) speech signal due to the channel and noise distortion as shown in equation (2).

## III. DATABASES AND PERFORMANCE EVALUATION METRICS

The early studies of spoofing attacks used different speech and speaker recognition databases, such as YOHO, NIST, and WSJ. The databases used for anti-spoofing studies are reported in Table 1. Since 2015, the research community has released multiple evaluation databases, that include SAS, ASVspoof 2015, ASVspoof 2017, ASVspoof 2019 challenge, AVspoof, RedDots Replayed databases. The AVspoof database introduces replay spoofing attacks along with SS and VC spoofing attacks. It was designed to simulate the attacks via LA and PA. This database was used in the BTAS 2016 Challenge [14,76]. RedDots [77] database is developed originally for text-dependent ASV research that was re-developed from replay attacks. This database is derived from the original RedDots database under various recording and playback conditions. However, standard impersonation database is not yet available publicly, the study reported in [79] used the YOHO database that was designed for ASV system. In this paper, we focus on the description of ASVspoof challenge datasets. Next, we will discuss about challenge databases in details.

**Table 1.** Various corpus on spoofing attacks to ASV system.

| Spoofing attacks | Corpus used |
|---|---|
| Impersonation [34] | YOHO |
| Voice Mimicry [35] | NIST |
| SS [71] | WSJ |
| SS [52] | WSJ |
| VC [53] | NIST SRE 2006 |
| VC [54] | NIST SRE 2006 |
| VC [72] | NIST SRE 2006 |
| VC, SS and Artificial Spoof [73] | NIST SRE 2006 |
| SS and VC [51] | SAS |
| Replay [74] | RSR2015 |
| VC and Replay [75] | RSR2015 |
| SS and VC [20] | ASV Spoof 2015 |
| SS, VC and Replay [76] | AV Spoof |
| Replay [77] | RedDots |
| Replay [24] | ASVspoof 2017 |
| SS, VC and Replay [78] | ASVspoof 2019 |

**Table 2.** A summary of ASVspoof 2015 Challenge database [22].

| | # Speakers | | # Utterances | |
|---|---|---|---|---|
| Subset | Male | Female | Genuine | Spoof |
| Training | 10 | 15 | 3750 | 12625 |
| Development | 15 | 20 | 3497 | 49875 |
| Evaluation | 20 | 26 | 9404 | 193404 |

## A) ASVspoof 2015 challenge

The ASVspoof 2015 Challenge database was the first major release for spoofing and countermeasures research [20]. The database consists of natural and spoofed speech, which is generated via speech synthesis and VC, for LA attacks. There are no remarkable channel or background noise effects. The database is divided into three subsets, namely, training, development, and evaluation. The evaluation subset consists of *known* and *unknown* attacks. They include the same five algorithms used to generate the development dataset and hence, called as *known (S1-S5)* attacks. In addition, other spoofing algorithms are included in *unknown (S6-S10)*, attacks which were used directly in the test data. The number of speakers in the database is reported in Table 2. The detailed description of the database can be found in [22,51,20]

## B) AVspoof database

AVspoof database introduces replay spoofing attacks along with synthetic speech and VC spoofing attacks. It was designed to simulate the attacks via LA and PA. This database was used in the BTAS 2016 Challenge [14,76]. The statistics of the database are summarized in Table 3. This database reports a comprehensive variety of presentation attacks including attacks when a genuine data is played back to an ASV system using laptop speakers, high-quality

https://doi.org/10.1017/ATSIP.2019.21

**Table 3.** A summary of AVspoof Database [14].

| | # Utterances | | |
|---|---|---|---|
| Subset | Genuine | PA attacks | LA attacks |
| Training | 4973 | 38580 | 17890 |
| Development | 4995 | 38580 | 17890 |
| Evaluation | 5576 | 43320 | 20060 |

PA, Physical Access; LA, Logical Access

**Table 4.** A summary of ASVspoof 2017 Challenge version 2.0 [24,81].

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

**Table 5.** The summary of ASVspoof 2019 Challenge database [82].

| | # Speakers | | # Utterances | | | |
|---|---|---|---|---|---|---|
| | | | Logical access | | Physical access | |
| Subset | Male | Female | Natural | Spoof | Natural | Spoof |
| Training | 8 | 12 | 2580 | 22800 | 5400 | 48600 |
| Development | 8 | 12 | 2548 | 22296 | 5400 | 24,300 |
| Evaluation | – | – | 71747 | | 137457 | |

**Table 6.** Data volume of the ReMASC corpus (*indicates incomplete data due to recording device crashes).

| Environment | Subjects | Genuine | Replayed |
|---|---|---|---|
| Outdoor | 12 | 960 | 6900 |
| Indoor 1 | 23 | 2760* | 23104 |
| Indoor 2 | 10 | 1600 | 7824 |
| Vehicle | 10 | 3920 | 7644 |
| Total | 55 | 9240 | 45472 |

speakers, and two mobile phones. SS attacks, such as speech synthesis and VC replayed with laptop speakers, are also included [76]. The “unknown” attacks were introduced in the test set to make the competition more challenging [76]. The organizers of the challenge provided a baseline system which is based on the open source Bob toolbox [76]. The baseline system consists of simple spectrogram-based ratio as features and logistic regression as a pattern classifier [76].

## C) ASVspoof 2017 challenge

The ASVspoof 2017 Challenge database was built on the RedDots corpus [80], and its replayed speech [77], which is therefore a replay database, and the speech is text-dependent. The number of speakers in training, development, and evaluation subset with corresponding number of genuine and spoofed utterances are summarized in Table 4. The detailed description of the database can be found in [24,81].

There were some anomalies in the original ASVspoof 2017 database. The problem was fixed in the ASVspoof 2017 Version 2.0 release [81]. Along with the corrected data, more detailed description of recording and playback devices as well as acoustic environments was also reported.

## D) ASVspoof 2019 challenge

The ASVspoof 2019 challenge is an extension of the previously held two challenges which focuses on countermeasures for all the three major attack types, namely, SS, VC, and replay. In this particular challenge database, there are two sub-challenges, namely, LA and PA. The statistics of the database are summarized in Table 5 [82]. The training dataset includes genuine and spoofed speech from 20 speakers (eight male and 12 female). The spoof speech signals are generated using one of the two VC and four speech synthesis algorithms. The data conditions for earlier ASVspoof 2017 challenge were created in an uncontrolled setup, and hence, this condition made the results challenging to analyze the signal due to varying additive and convolutive noise. This uncontrolled condition was taken care in the present challenge by creating a simulated and controlled acoustic environment conditions. Unlike previous challenge editions, ASVspoof 2019 adopts a recently-proposed Tandem Detection Cost Function (t-DCF) as the primary performance metric along with % EER [83].

## E) ReMASC

The ReMASC (Realistic Replay Attack Microphone Array Speech Corpus) is the first publicly available database that is designed specifically for the protection of voice-controlled systems (VCSs) against various replay attacks in various conditions and environments [84]. The ASVspoof 2019 challenge consists of simulated data for clear theoretical analysis of audio spoofing attacks in physical environments, however, it brings a simulation-to-reality gap. Recent increase for the VCSs depends on voice input as the primary user–machine interaction modality such as, intelligent personal assistants (e.g. Amazon Echo, Samsung Bixby, and Google Home) allow users to control their smart home appliances and complete many other tasks with ease. The VCSs also began to be used in vehicles to allow drivers to control their cars’ navigation systems and other vehicle services. The number of speakers and the environment conditions are summarized in Table 6.

## F) Performance evaluation metrics

For effective comparison between algorithms, we need both the standard databases and common evaluation metrics. Given a test speech sample, four possible decisions can be made in SSD, which are summarized in Table 7, where a False Acceptance Rate (FAR) and a False Rejection Rate (FRR) represent two types of classification errors. FAR and FRR are also called as false alarm and miss detection, respectively [85]. A system’s performance can also be described by an Equal Error Rate (EER) at which the FAR (false alarm) and FRR (miss detect) equals [85,15].

https://doi.org/10.1017/ATSIP.2019.21

**Table 7.** Decision of four possible outcomes in the ASV system [20].

| | Decision | |
|---|---|---|
| | Acceptance | Rejection |
| Genuine | Correct acceptance | False rejection/ (miss detection) |
| Impostor | False acceptance/ (false alarm) | Correct rejection |

For a particular ASV system, the detection scores are computed with a false alarm and miss rate, denoted respectively, as $P_{\mathrm{fa}}(\theta)$ and $P_{\mathrm{miss}}(\theta)$ at decision threshold $\theta$, and are given as follows:

$$P_{\mathrm{fa}}(\theta) = \frac{\#\{\text{spoof trials with score} > \theta\}}{\#\{\text{total spoof trials}\}}, \tag{4}$$

$$P_{\mathrm{miss}}(\theta) = \frac{\#\{\text{genuine trials with score} \leq \theta\}}{\#\{\text{total genuine trials}\}}, \tag{5}$$

where $P_{\mathrm{fa}}(\theta)$ and $P_{\mathrm{miss}}(\theta)$ are monotonically decreasing and increasing functions of $\theta$. The EER corresponds to the threshold $\theta_{EER}$ at which the two detection error rates coincide, i.e.

$$P_{\mathrm{fa}}(\theta_{EER}) = P_{\mathrm{miss}}(\theta_{EER}). \tag{6}$$

Impostor trials corresponding to a score higher than the threshold will be misclassified as genuine trials, whereas genuine trials with a score lower than the threshold will be misclassified as impostor trials. Since the two errors are inversely related, it is often desirable to illustrate the performance as a function of the threshold $\theta$. One such measure is the Half Total Error Rate (HTER) [76]:

$$HTER(\theta) = \frac{P_{\mathrm{fa}}(\theta) + P_{\mathrm{miss}}(\theta)}{2}. \tag{7}$$

Performance can also be illustrated graphically with Detection-Error Trade-off (DET) curve [86]. The DET curve illustrates the behavior of a system for different decision threshold, $\theta$, that also allows us to observe a trade-off between the FAR and the FRR.

The Detection Cost Function (DCF) is defined in terms of the cost of miss and false alarms along with the prior probability for the target speaker hypothesis. The standard DCF is designed for the assessment of a single ASV system whereas t-DCF metric combines the assessment of ASV system and the spoofing countermeasures [83]. One of the initial attempts in this direction was reported for the spoof detection task for professional mimics [87]. The t-DCF metric was used in ASVspoof 2019 evaluation [83].

## IV. COUNTERMEASURES FOR SYNTHETIC SPOOFING ATTACKS

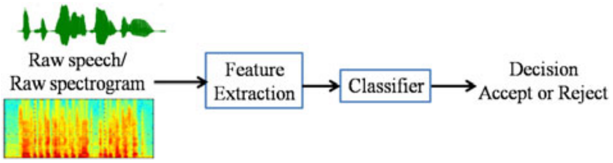

**Fig. 7.** Spoofing detection framework.

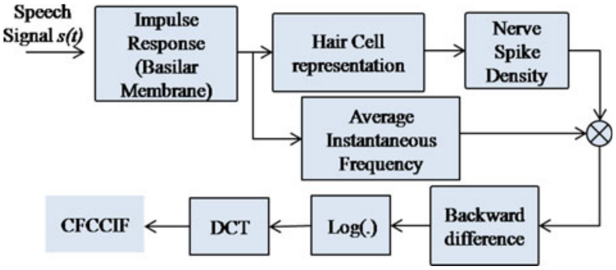

**Fig. 8.** Block diagram of the CFCCIF feature extraction process. After [93].

We now give an overview of system construction for anti-spoofing against SS that includes synthesized and converted voices. The ASVspoof 2015 Challenge provided a common platform to study the effectiveness of countermeasures. Similar to other pattern classification system, a traditional spoof detection system consists of two parts, namely, feature extraction and pattern classifier as shown in Fig. 7. We will discuss the traditional approach and the end-to-end approach in more detail in this section.

### 1) TRADITIONAL APPROACHES

There have been several early studies on finding features that reflect the artifacts in the SS. For example, one study considers that the pitch ($F_0$) pattern of SS is more rigid than that of natural speech in [88], temporal structure of SS is different from that of natural speech [89], and SS contains phase distortions [90]. However, these features were observed on *ad hoc* databases, and moreover, they were not evaluated using a common performance evaluation metric. Hence, there was a need to develop a shared task for SS detection, that motivated the ASVspoof 2015 Challenge [20,22].

In ASVspoof 2015 Challenge, it was observed that the efforts on better features were more effective than the complex classifiers [91]. Furthermore, long-term features are more effective than short-term features that are derived from short-term windows. The Constant-Q Cepstral Coefficients (CQCC) [92], and Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) [93] offer state-of-the-art performance on ASVspoof 2015 database. The CFCCIF feature extraction, which is described by Speech Research Lab DA-IICT in Fig. 8, represents the best performance in ASVspoof 2015.

The CQCC features are extracted with the constant-Q transform (CQT), a perceptually-inspired alternative to Fourier-based approaches for time-frequency analysis. The CQCC features found to be generalized across three different databases (i.e. ASVspoof 2015 Challenge, AVspoof, and RedDots replayed database) and it delivered the state-of-the-art performance in each case [94].

https://doi.org/10.1017/ATSIP.2019.21

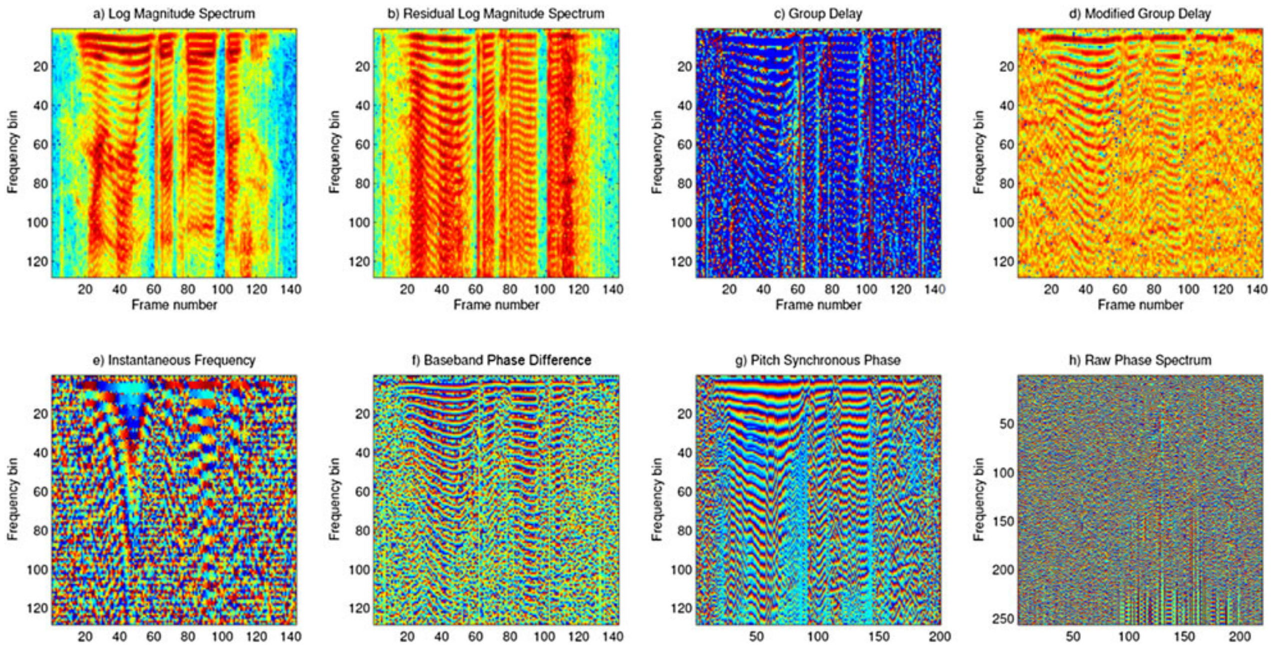

**Fig. 9.** Demonstration of eight different types of features is shown for a natural utterance D15_1000931 from the development set of ASVspoof 2015 challenge dataset. For each feature type, only the low half of the FFT frequency bins are shown. Adapted from [95].

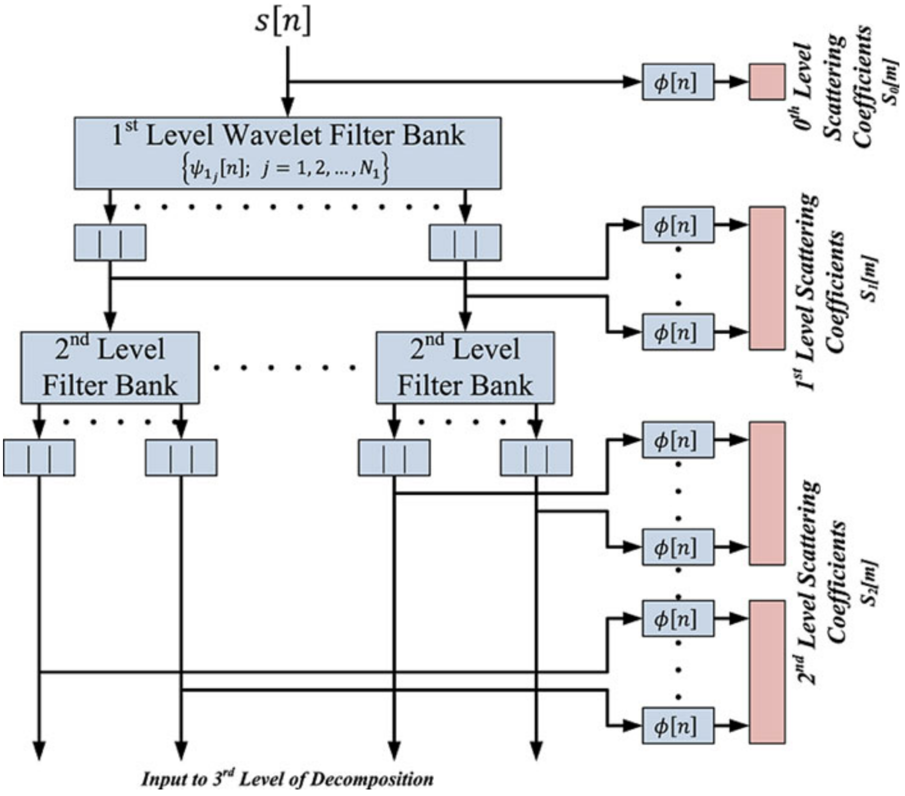
$s[n]$
1st Level Wavelet Filter Bank
$\{\psi_{1_j}[n];\ j = 1, 2, \ldots, N_1\}$
$| \; |$
2nd Level Filter Bank
$\phi[n]$
0th Level Scattering Coefficients $S_0[m]$
1st Level Scattering Coefficients $S_1[m]$
2nd Level Scattering Coefficients $S_2[m]$
Input to 3rd Level of Decomposition

**Fig. 10.** Block diagram of two-level scattering decomposition. Adapted from [98].

Other effective features include high-dimensional magnitude-based features, and phase-based features as reported in a comparative study [95]. The magnitude-based features include Log Magnitude Spectrum, and Residual Log Magnitude Spectrum; the phase-based features include Group Delay Function, Modified Group Delay Function, Baseband Phase Difference, Pitch Synchronous Phase, Instantaneous Frequency Derivative in Fig. 9.

The features extracted using subband processing were also explored, such as Linear Frequency Cepstral Coefficients (LFCC) [96], Energy Separation Algorithm-Instantaneous Frequency Cepstral Coefficients (ESA-IFCC)

[91], and Constant-Q Statistics-plus-Principal Information Coefficient (CQSPIC) [97]. The basic motivation behind subband processing is that artifacts of SS manifest differently in different subbands. Temporal features, such as instantaneous frequency and envelop, are sensitive to those artifacts. Another technique for subband processing is to perform a two-level scattering decomposition through a wavelet filterbank to derive a scalogram as shown in Fig. 10 [98].

### 2) Representation learning approaches

The representation learning approaches work either in the form of feature learning or as a pattern classifier. With feature learning, it was observed that the use of DNN for representation learning followed by GMM or SVM classifier was more successful than using DNN as a classifier. The hidden layer representation obtained from DNN was used as features (called as spoofing vectors or *s*-vectors), and Mahalanobis distance for classification [99]. The CNN and RNN classifiers were explored along with three features, namely, Teager Energy Operator (TEO) Critical Band Autocorrelation Envelope (TEO-CB-Auto-Env), Perceptual Minimum Variance Distortionless Response (PMVDR), and raw spectrograms [100].

In [101], feature learning is followed by LDA and GMM classifiers. The frame-level and sequence-level features were extracted using DNN and RNN, respectively, resulted in 0% EER for all the attack types from S1 to S9, and 1.1% EER on all the averaged conditions [101]. Bottleneck features extracted from the DNN hidden layers were also used with GMM classifier in [102]. In [103], the Convolutional Restricted Boltzmann Machine (ConvRBM) is used for auditory filterbank learning that performed better than traditionally handcrafted filterbanks. The study [103] shows that ConvRBM learns better low-frequency subband filters on ASVspoof 2015 dataset than on TIMIT. Supervised auditory filterbank learning using DNN was also studied in [104]. The first- and second-order Long-Term Spectral Statistics (LTSS) were used for synthetic SSD task along with various classifiers with DNN outperforming others. [105].

Recently, end-to-end DNN approaches have emerged for various speech and audio processing applications [106], [107]. The goal of the end-to-end DNN is to learn acoustic representation from the raw speech and audio signals as well as perform classification task in a DNN network [108], [109]. For synthetic SSD task, Convolutional Neural Network (CNN) was used for feature learning from raw speech signals and binary classification task [110]. Along with CNN layers, Long-Short Term Memory (LSTM) layers were used in an architecture called Convolutional LSTM DNN (CLDNN) trained directly on raw speech signals [111,112].While CLDNN achieves 0% EER, it has not worked well for S10 set. The end-to-end DNN approach represents a new direction of anti-spoofing study.

In ASVspoof 2015 Challenge, the systems in [95,113] use DNN as the classifiers. In [114], a DNN classifier with novel human log-likelihoods (HLL) scoring method was proposed that performed significantly better and achieved an average EER of all the attack types to 0.04%. It was shown in [114] that HLL scoring method is more suitable for the SSD task than the classical LLR scoring method, especially when the spoofed speech is very similar to the human speech [114]. The output softmax layer consists of neurons representing spoofing and human (natural) speech labels. According to the literature, the system performances on ASVspoof 2015 Challenge database are summarized in Table 8, with the system of CQCC feature set, and DNN-HLL classifier representing the best performance.

**Table 8.** Comparison of results (in % EER) on ASVspoof 2015 Challenge Database.

| Feature Set | Classifier | Dev | Eval |
|---|---|---|---|
| CQCC [94] | GMM | 0.00 | 0.26 |
| CFCCIF [93] | GMM | 2.29 | 1.21 |
| LFCC [96] | GMM | 0.66 | 0.89 |
| RFCC [96] | GMM | 075 | 1.02 |
| MFCC [96] | GMM | 1.09 | 3.0 |
| SCFC [96] | GMM | 0.25 | 4.45 |
| SCMC [96] | GMM | 0.95 | 0.94 |
| LPCC [96] | GMM | 0.68 | 1.21 |
| IMFCC [96] | GMM | 0.48 | 1.00 |
| RPS [96] | GMM | 0.37 | 5.30 |
| SCC [98] | GMM | – | 0.18 |
| DMCC-BNF [102] | GMM | – | 2.15 |
| ESA-IFCC [91] | GMM | 1.89 | 6.79 |
| ConvRBM-CC [103] | GMM | 2.53 | 4.47 |
| DNN-IGFCC [104] | GMM | 0.12 | 0.56 |
| LF RPS [113] | SVM | 1.34 | 6.11 |
| DNN, RNN features [101] | LDA, GMM | – | 1.1 |
| LF Spectrum [113] | DNN | 0.03 | 4.38 |
| TEO [100] | DNN | 2.31 | – |
| PMVDR [100] | DNN | 1.44 | – |
| LTSS [105] | DNN | – | 0.25 |
| E2E CNN [110] | DNN | – | 2.89 |
| LTSS and E2E CNN [110] | DNN |  | 0.157 |
| E2E CLDNN [112] | DNN | – | 4.56 |
| CQCC [114] | DNN−HLL | – | **0.04** |
| Spectrogram [100] | CNN | 0.36 | 3.07 |
| Spectrogram [100] | RNN | 1.04 | 2.46 |
| Spectrogram [100] | CNN+RNN | 0.42 | 1.86 |

## V. COUNTERMEASURES FOR REPLAY SPOOFING ATTACKS

We now give an overview of system construction for anti-spoofing against replay attacks.

### 1) Acoustic features

Theuse of high-fidelity recording devices represents a serious threat, therefore countermeasures were proposed to guard against such attack. The spectral peak mapping method was proposed as a countermeasure to detect the replay attack on a remote telephone interaction [115]. Replay attacks with far-field recordings were addressed in [66].

The ASVspoof 2017 Challenge paid a special attention to replay speech detection. The baseline system with CQCC features and GMM classifier were provided by the organizers as it performs well in the earlier evaluation [24].

https://doi.org/10.1017/ATSIP.2019.21

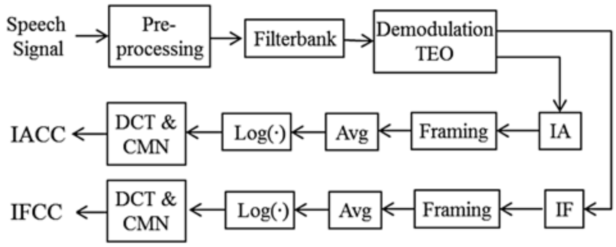

**Fig. 11.** Schematic block diagram of short-time instantaneous amplitude and frequency modulation (AM-FM) feature set. After [122].

The acoustic features, such as Rectangular Filter Cepstral Coefficients (RFCC), Subband Spectral Centroid Magnitude Coefficients (SCMC), Subband Spectral Centroid Frequency Coefficients (SCFC), Subband Spectral Flux Coefficients (SSFC) studied. It is found that the SCMC followed by feature normalization method outperforms other acoustic features [116]. With the analysis on Inverse Mel Frequency Cepstral Coefficients (IMFCC), Linear Prediction Cepstral Coefficients (LPCC), and LP residual features, it is found that high-frequency regions have more discriminative information than the other frequency regions [117]. The effect of mean and variance normalization of CQCC feature set with Support Vector Machines (SVM) classifier was studied in [81,118]. One of the approaches used Single Frequency Filtering (SFF) and found the importance of high-resolution temporal features [119].

The short-time AM-FM features set obtained using Energy Separation Algorithm (ESA) were studied in [120,121] as shown in Fig. 11. The features were also developed with subband filter analysis using CFCCIF [120], IFCC [123], Empirical Mode Decomposition Cepstral Coefficients (EMDCC) [124], transmission line cochlear model [125], auditory inspired spatial differentiation filterbank [126], and ESA-IF-based feature estimation using Cochlear filter in [127]. Excitation source-based features were studied in [128], wavelet-based features in [129], and phase-based features in [130]. The concept of feature switching at the decision-level, along with information from the non-voiced segments were studied in [131].

The study in [132] shows that some phonemes carry more replay artifacts than others, therefore consequently judicious use of phoneme-specific models can improve replay detection. The analysis of full-frequency bands with F-ratio and multi-channel feature extraction using attention-based adaptive-filters (AAF) is studied in [133]. The analysis of replay speech signal using reverberation concept and Teager energy profile is studied in [134].

### 2) Representation of learning approaches

The three key observations from ASVspoof 2017 Challenge are the use of spectral information in the higher frequency regions, feature normalization, and representation learning approach. It was shown that many representation learning-based approaches did well in the ASVspoof 2017 Challenge.

First, we describe the representation learning approaches used in ASVspoof 2017 Challenge. End-to-end replay spoofing detection was proposed using deep residual network (ResNet) and raw spectrograms of speech signals [135]. It was also shown that data augmentation in DNN significantly improves the performance [135]. In one of the approaches, DNN was trained to discriminate between the various channel conditions available in the ASVSpoof 2017 Challenge database, namely, recording, playback, and session conditions [136]. In [136], the DNN features were learned from CQCC and HFCC features followed by an SVM classifier. The model fusion strategies using ResNet, GMM, and DNN were also explored and found to perform better compared to individual systems [137]. In particular, the ASVspoof 2017 Challenge winner system used CNN and RNN for representation learning from STFT spectrograms followed by a GMM classifier [138].

The use of ConvRBM to learn auditory filterbank followed by the AM-FM demodulation using ESA for the replay SSD task was studied in [139]. The ConvRBM learns subband filters that represent high-frequency information in a much better way when used with pre-emphasized speech signals as shown in Fig. 12. Combining representation learning and signal processing techniques there is significant improvement of 0.82 and 8.89% EER on the development and evaluation set. A novel algorithm called NeuroEvolution of Augmenting Topologies (NEAT) was used in an end-to-end anti-spoofing network [140]. The NEAT framework also introduces a new fitness function for DNN that results in better generalization than the baseline system and improves the relative performance by 22% on the ASVspoof 2017 database [140].

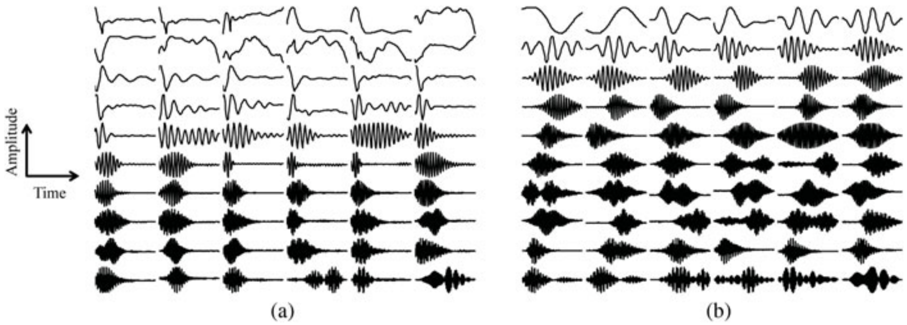

**Fig. 12.** The ConvRBM subband filters in temporal-domain (a) without, and (b) with pre-emphasis, respectively. After [139].

**Table 9.** Comparison of results (in % EER) on ASVspoof 2017 Challenge Database.

| Feature Set | Classifier | Dev | Eval |
|---|---|---|---|
| CQCC (BL) [24] | GMM | 10.35 | 28.48 |
| ESA-IFCC [122] | GMM | 4.12 | 12.79 |
| VESA-IACC [121] | GMM | 6.12 | 11.94 |
| AWFCC [143] | GMM | 6.37 | 11.72 |
| LFCC [116] | GMM | 10.31 | 16.54 |
| SCFC [116] | GMM | 24.51 | 24.83 |
| SSFC [116] | GMM | 12.81 | 22.38 |
| SCMC [116] | GMM | 9.32 | 11.49 |
| RFCC [116] | GMM | 6.91 | 11.90 |
| VESA-IFCC [120] | GMM | 4.61 | 14.06 |
| CFCCIF [120] | GMM | 6.80 | 34.49 |
| CQCC (6–8 kHz) [117] | GMM | 5.13 | 17.31 |
| HFCC [136] | GMM | 5.9 | 23.90 |
| SFCC [119] | GMM | 2.35 | 20.20 |
| SCC [129] | GMM | 3.16 | 19.79 |
| qDFTspec [144] | GMM | – | 11.43 |
| qPspec [144] | GMM | – | 11.85 |
| EOC$_m$ [140] | – | – | 18.2 |
| ConvRBM-CC [139] | GMM | 0.82 | 8.89 |
| LFMGDCC [130] | GMM | 20.70 | 20.84 |
| EMDCC [124] | GMM | 28.48 | 28.06 |
| LFRCC [128] | GMM | 8.38 | 22.28 |
| DLFS [131] | GMM | 6.68 | 19.16 |
| PNCC [145] | GMM | 20.78 | 23.74 |
| PSRMS [146] | GMM | 33.38 | 28.16 |
| CF [126] | GMM | – | 10.84 |
| CM [126] | GMM | – | 10.93 |
| TLC_AM [125] | GMM | – | 8.68 |
| TLC_FM [125] | GMM | – | 11.30 |
| TECC [134] | GMM | 9.55 | 11.73 |
| PPWS [132] | GMM | – | 10.70 |
| PPRFWS_LR [132] | GMM | – | 9.28 |
| ESA-IFCC [147] | GMM+CNN | 1.90 | 10.42 |
| LPCC [138] | SVM *i-vector* | 9.80 | 12.54 |
| CQCC [137] | DNN | 5.18 | 19.41 |
| CQCC [137] | ResNet | 5.05 | 18.79 |
| MFCC [137] | ResNet | 10.95 | 16.26 |
| GD Spectrum [141] | ResNet | **0.0** | **0.0** |
| CQCC [135] | ResNet | 6.32 | 23.14 |
| AF-DRN [142] | ResNet | 6.55 | 8.99 |
| SFCC [119] | BLSTM | 3.66 | 22.40 |
| FFT features [138] | LCNN | 4.53 | 7.37 |
| CQT features [138] | LCNN | 4.80 | 16.54 |
| AFCC [133] | – | 4.01 | 27.80 |
| ARP [133] | – | 9.11 | 12.65 |

BL, baseline

A novel visual attention mechanism is employed in deep ResNet architecture using the group delay features (GD spectrum) that resulted in EER of 0% on both the development and evaluation sets, respectively [141]. In [142] attention-based filtering is used that enhances the feature representation in both time and frequency-domains and used ResNet-based classifier. Class activation maps (CAM) using global average pooling (GAP) utilizes the implicit attention mechanism present in CNN. Hence, representation learning approaches are very promising directions for the replay SSD compared to the synthetic SSD task. According to the literature, the system performances on ASVspoof 2017 Challenge dataset are summarized in Table 9.

## VI. LIMITATIONS AND TECHNOLOGICAL CHALLENGES

In this section, we hope to discuss some topics that are worthy of further inquiry and possible future direction.

(i) **Logical and Physical Access**: The PA is the actual spoofing where the speech is played back through a microphone into the ASV system. However, the ASVspoof database gave special attention to LA attacks. For such attacks, it is assumed that the spoofed samples are directly injected into the system through a software-based process [21]. Hence, PA attacks are more realistic than the logical access attacks, where the attacker plays back a recorded utterance to the system. This utterance can be either obtained from the real speaker or can be forged using VC or synthetic speech (SS) algorithms. This motivates the study on PA attacks and evaluation database development.

(ii) **Diversity of Spoofing Attacks**: The ASVspoof 2015 Challenge database consists of only VC and synthetic speech spoofing algorithms. This database consists of variation of seven VC spoofing techniques and only three synthetic speech techniques. It is noted that ASVspoof 2017 Challenge database focuses only on replay spoof. While ASVspoof 2015 and 2017 database includes SS, VC, and replay spoofing techniques, the spoofing voice database is not developed using the latest neural voice generation techniques.

(iii) **Performance of Joint Protocol with ASV Systems**: The current studies of countermeasures and ASV systems are carried out separately. What user would like to have is a secure and accurate ASV system. However, a more robust ASV system to noise and channel variations may become less secure against spoofing attacks. As there is no guarantee of having a better performing countermeasure that provides lower EER and also reliable for the ASV system performance. Hence, with the progress made in the research of spoofing detection, evaluation metrics must evolve to reflect the joint protocol system performance.

(iv) **Liveness Detection**: The use of high-quality recording loudspeaker or playback device to record/playback the speech signal, in this process the quality of signal captured becomes indistinguishable from live human voice. This high-quality device makes the speech signal impossible to detect that depends on the acoustic cues. This gives rise to investigate further on the liveness detection of human voice.

(v) **Signal Degradation Conditions**: Current publicly available spoofing databases are developed in clean conditions. However, the recent replay database was recorded under various acoustic environmental conditions. For ASVspoof 2015 Challenge database, the noisy database was developed by adding various noises at different Signal-to-Noise Ratio (SNR) levels. Further investigations are required as to how

https://doi.org/10.1017/ATSIP.2019.21

the diversity of different noise types affects the SSD performance. In addition, the study is required to observe the effect on SSD when the additive noise is added manually, and when the noise is added naturally via the acoustic environment. For example, a study was done for a replay database under different background, microphone, etc. [81]. Hence, the countermeasures must be developed that it should be robust to signal degradation conditions as well.

(vi) **Robustness in ASV implies Vulnerability**: In practice, we would like ASV system to be robust against variations, such as microphone and transmission channel, intersession, acoustic noise, speaker aging, etc. A robust ASV system may become vulnerable to various spoofing attacks as it tries to nullify these effects and normalize the spoofing speech toward the natural speech. Thus, robustness and anti-spoofing security should be addressed separately. It is worth to study how features, classifiers, and systems are designed to be both robust and secure.

(vii) **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. Thus, any deviation from the degradation in the harmonic structure could capture the signature of spoof speech [128]. To the best of authors' knowledge, there is no study reported in analyzing this particular aspect. We believe several source information, such as Linear Prediction (LP) residual, Teager Energy Operator (TEO) profile and its Variable length (VTEO) profile, etc., could be explored in the framework of recent study reported in [128].

(viii) **Exploring Phase-based Features**: It is important to note that phase-based features (either time-domain analytic or frequency-domain) could capture a different kind of information in spoofed speech depending upon the type of spoof. For example, in USS system, when the speech sound units are picked up by optimizing the target cost, in the synthesized voice, it will have *linear phase* mismatches (since these units are recorded in different sessions) [148]. On the other hand, for replay speech, the impulse response of the acoustic environment (say room) gets convolved with the natural speech. The impulse response of an acoustic system (in this case room) is infinite in duration, i.e. Infinite Impulse Response (IIR) in nature (due to infinite transmissions and reflections). Thus, the non-linear phase in frequency-domain of this acoustic system is added to the phase of natural speech. In addition, corresponding effects of this non-linear phase could be observed in temporal-domain, such as non-integer delay in frequency components. There have been many studies in phase features in SS detection. Phase study remains a research topic that is worth more investigations.

(ix) **Comparison of Human versus Machine-learning**: It is of great interest to know whether human perception is important in identifying spoofing, and hence, humans can achieve better performance than automatic approaches in detecting spoofing attacks. There was a benchmark study comparing automatic systems against human performance on a speaker verification and SS spoofing detection tasks (SS and voice conversion spoofs) [149]. It was observed that human listeners detect spoofing less well than most of the automatic approaches except USS speech [149]. In a similar study, it was found that both the humans and machines have difficulties in spoofing detection when narrowband speech signals were used (8 kHz sampling frequency) [150]. Hence, for telephone line speech signals, it is more challenging to do SSD due to the lower available bandwidth up to 4 kHz. It may be of great interest to study human performance for replay SSD task.

(x) **Robustness to High-Quality Speech Synthesizers**: Recently, many representation learning-based high-quality speech synthesis techniques were proposed that achieved significantly better naturalness. The Wavenet [50], GAN [48], and other end-to-end speech synthesis architectures [151] produce high-quality synthesized speech. It is also shown that low-quality publicly available database can be used to produce high-quality spoof data using GAN-based speech enhancement [152]. Such high-quality SS and VC techniques may further increase the difficulties in synthetic SSD. This technique could be used to generate spoof speech database in the next edition of ASVspoof challenge [23].

## VII. SUMMARY AND CONCLUSIONS

This article provided an overview of the SSD task. We reviewed different countermeasure approaches for synthetic and replay detection, in particular, classical and representation learning approaches. The study also reported various technological challenges involved during spoofing detection and also discussed various spoofing databases with their limitations. The article also discusses the recent advances in the spoofing area for ASV task.

A significant amount of research has been carried out to assess the vulnerability of ASV systems to different spoofing attacks. It is especially challenging to recreate real attacking conditions during the development of various spoofing database. Under particular controlled conditions, different spoofing attacks are developed, as they are unfeasible to collect a database with all different possibilities that are available in the market. The performance metric is usually distributed into train, development, and test set, where these

https://doi.org/10.1017/ATSIP.2019.21

individual sets have almost similar examples of spoofs in all the sets. However, real-world scenario for ASV represents an open set evaluation without having any constraints on the spoofs used to attack given ASV system.

In the current spoofing context described in this article and lessons learned in more than 10 years of spoofing research, there are still few open questions that need to be answered. They are: What are the future challenges that arise further in voice biometric spoofing? What are the issues which are yet to be looked into and need to be explored further? Where do we go from here?

Presently, one of the most urgent needs is to define a clear methodology to assess the spoofing attacks. This is not a straightforward issue, as many new variables are involved during the development of spoofing algorithms. Another observation is that there does not exist superior anti-spoofing technique that performs uniformly along all the spoofs. Approaching only with one countermeasure will depend on the nature of the attack scenario and data acquisition conditions. Hence, there should be another complementary countermeasure followed by fusion approaches to develop high-performance countermeasure over different spoofing data. In addition, practical considerations should not be left out. As technology progresses, new techniques continue to emerge in the form of hardware devices and signal processing methods. Hence, it is important to keep a track of such technological progress, since this advancement could be the key to develop a novel and efficient countermeasure.

Finally, though a significant amount of work is now being reported in the field of spoofing detection, different methodologies and attacks have also evolved that became more and more sophisticated. As a consequence, yet there are many big challenges that are to be faced to protect against spoofing attacks, hopefully, that will be lead in the upcoming years with a new generation of more secure voice biometric systems.

## ACKNOWLEDGMENTS

Authors would like to thank the authorities of DA-IICT Gandhinagar, India and NUS Singapore for their support to carry out this research work. First author would like to thank the University Grants Commission (UGC), New Delhi, India for providing Rajiv Gandhi National Fellowship (RGNF) for her doctoral studies.

## REFERENCES

1 Jain A.K.; Nandakumar K.; Ross A.: 50 years of biometric research: accomplishments, challenges, and opportunities. *Pattern. Recognit. Lett.*, **79** (2016), 80–105.

2 Hadid A.; Evans N.; Marcel S.; Fierrez J.: Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. *IEEE Signal. Process. Mag.*, **32** (5) (2015), 20–30.

3 Jain A.K.; Ross A.; Pankanti S.: Biometrics: a tool for information security. *IEEE Trans. Inf. Foren. Secur.*, **1** (2) (2006), 125–143.

4 Jain A.K.; Li S.Z.: Handbook of Face Recognition, Springer, 2011.

5 Maltoni D.; Maio D.; Jain A.K.; Prabhakar S.: Handbook of Fingerprint Recognition, Springer Science & Business Media, 2009.

6 Daugman J.: The importance of being random: statistical principles of iris recognition. *Pattern. Recognit.*, **36** (2) (2003), 279–291.

7 Connie T.; Teoh A.; Goh M.; Ngo D.: Palmhashing: a novel approach for cancelable biometrics. *Inf. Process. Lett.*, **93** (1) (2005), 1–5.

8 Sanchez-Reillo R.; Sanchez-Avila C.; Gonzalez-Marcos A.: Biometric identification through hand geometry measurements. *IEEE Trans. Pattern Anal. Mach. Intell.*, **10** (2000), 1168–1171.

9 Yan P.; Bowyer K.W.: Biometric recognition using 3D ear shape. *IEEE Trans. Pattern Anal. Mach. Intell.*, **29** (8) (2007), 1297–1308.

10 Yazdanpanah A.P.; Faez K.; Amirfattahi R.: Multimodal biometric system using face, ear and gait biometrics, in *IEEE Int. Conf. on Information Sciences Signal Processing and their Applications (ISSPA)*, Kuala Lumpur, Malaysia, 2010, 251–254.

11 Nalwa V.S.: Automatic on-line signature verification. *Proc. IEEE*, **85** (2) (1997), 215–239.

12 Monrose F.; Rubin A.: Authentication via keystroke dynamics, in *ACM Conf. on Computer and Communications Security*, Zurich, Switzerland, 1997, 48–56.

13 Jain A.K.; Nandakumar K.; Nagar A.: Biometric template security. *EURASIP J. Adv. Signal. Process.*, (2008), 113.

14 Ergünay S.K.; Khoury E.; Lazaridis A.; Marcel S.: On the vulnerability of speaker verification to realistic voice spoofing, in *IEEE Int. Conf. on Biometrics Theory, Applications and Systems (BTAS)*, Virginia, USA, 2015, 1–6.

15 Evans N.W.; Kinnunen T.; Yamagishi J.: Spoofing and countermeasures for automatic speaker verification, in *INTERSPEECH*, Lyon, France, 2013, 925–929.

16 Evans N.: Handbook of Biometric Anti-spoofing: Presentation Attack Detection, Springer, 2019.

17 Koppell J.: International organization for standardization. *Handb. Transnatl. Gov. Inst. Innov.*, **41** (2011), 289.

18 Galbally J.; Marcel S.; Fierrez J.: Biometric antispoofing methods: a survey in face recognition. *IEEE Access*, **2** (2014), 1530–1552.

19 Biometric and spoofing identification, https://www.google.com/search?q=SPOOFING+BIOMETRIC+IDENTIFICATION&source=lnms&tbm=isch&sa, last accessed: 2018-10-15.

20 Wu Z.; Evans N.; Kinnunen T.; Yamagishi J.; Alegre F.; Li H.: Spoofing and countermeasures for speaker verification: a survey. *Speech Commun.*, **66** (2015), 130–153.

21 Muckenhirn H.; Magimai-Doss M.; Marcel S.: Presentation attack detection using long-term spectral statistics for trustworthy speaker verification, in *IEEE Int. Confe. of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2016, 1–6.

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

23 Kinnunen T.; Evans N.; Yamagishi J.; Lee K.A.; Sahidullah M.; Todisco M.; Delgado H.: ASVspoof 2019: automatic speaker verification spoofing and countermeasures challenge, Last accessed = 15 Oct 2018. [Online]. Available: http://www.asvspoof.org/.

24 Kinnunen T.; Sahidullah M.; Delgado H.; Todisco M.; Evans N.; Yamagishi J.; Lee K.A.: The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection, in *INTERSPEECH*, Stockholm, Sweden, 2017, 1–6.

25 Evans N.; Li S.Z.; Marcel S.; Ross A.: Guest editorial: special issue on biometric spoofing and countermeasures. *IEEE Trans. Inf. Foren. Secur.*, **10** (4) (2015), 699–702.

26 Evans N.; Marcel S.; Ross A.; Teoh A.B.J.: Biometrics security and privacy protection [from the guest editors]. *IEEE Signal Process. Mag.*, **32** (5) (2015), 17–18.

27 JSTSP Special Issue on Spoofing and Countermeasures for Automatic Speaker Verification, https://signalprocessingsociety.org/blog/jstsp-special-issue-spoofing-and-countermeasures, last accessed: 2019-07-27.

28 Special Issue on Speaker and Language Characterization and Recognition: Voice modeling, Conversion, Synthesis and Ethical Aspects, https://www.journals.elsevier.com/computer-speech-and-language/call-for-papers/special-issue-on-speaker-and-language-characterization, last accessed: 2019-07-27.

29 Special Issue on Advances in Automatic Speaker Verification Anti-spoofing, https://www.journals.elsevier.com/computer-speech-and-//language/call-for-papers/advances-in-automatic-speaker, last accessed: 2019-07-27.

30 Wu Z.; Yamagishi J.; Kinnunen T.; Hanilçi C.; Sahidullah M.; Sizov A.; Evans N.; Todisco M.: ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE J. Sel. Top. Signal. Process.*, **11** (4) (2017), 588–604.

31 Patil H.A.; Kamble M.R.: A survey on replay attack detection for automatic speaker verification (ASV) system, *to appear in Asia-Pacific Signal and Information Processing Association, Annual Summit and Conf. (APSIPA-ASC)*, Hawaii, USA, 2018

32 Li H.; Patil H.A.; Kamble M.R.: Tutorial on spoofing attack of speaker recognition, in *Asia-Pacific Signal and Information Processing Association, Annual Summit and Conf. (APSIPA-ASC)*, Kuala Lumpur, Malaysia, 2017.

33 Markham D.: Phonetic imitation, accent, and the learner, Linguistics and Phonetics, vol. **33**, 1997.

34 Lau Y.W.; Wagner M.; Tran D.: Vulnerability of speaker verification to voice mimicking, in *IEEE Int. Symp. on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, 2004, 145–148.

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

36 Rosenberg A.E.: Automatic speaker verification: a review. *Proc. IEEE*, **64** (4) (1976), 475–487.

37 Lau Y.W.; Tran D.; Wagner M.: Testing voice mimicry with the YOHO speaker verification corpus, in *Int. Conf. on Knowledge-Based and Intelligent Information and Engineering Systems, Springer*, Melbourne, VIC, Australia, 2005, 15–21.

38 Farrús M.; Wagner M.; Anguita J.; Hernando J.: How vulnerable are prosodic features to professional imitators?, in *Odyssey The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, 2008, 1–4.

39 Jain A.K.; Prabhakar S.; Pankanti S.: On the similarity of identical twin fingerprints. *Pattern Recognit.*, **35** (11) (2002), 2653–2663.

40 Kersta L.; Colangelo J.: Spectrographic speech patterns of identical twins. *J. Acoust. Soc. Am.*, **47** (1A) (1970), 58–59.

41 Patil H.A.; Parhi K.K.: Variable length Teager energy based mel cepstral features for identification of twins, in *Int. Conf. on Pattern Recognition and Machine Intelligence, Springer*, Berlin, Heidelberg, Germany, 2009, 525–530.

42 HSBC reports high trust levels in biometric tech as twins spoof its voice ID system, *Biometric Technology Today*, vol. 2017, no. 6, p. 12, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969476517301194.

43 BBC fools HSBC voice recognition security system, https://www.bbc.com/news/technology-39965545, last accessed: 2018-10-15.

44 Twins fool HSBC voice biometrics - BBC, https://www.finextra.com/newsarticle/30594/twins-fool-hsbc-voice-biometrics--bbc, last accessed: 2018-10-15.

45 Hunt A.J.; Black A.W.: Unit selection in a concatenative speech synthesis system using a large speech database, in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, Georgia, USA, 1996, 373–376.

46 Zen H.; Tokuda K.; Black A.W.: Statistical parametric speech synthesis. *Speech Commun.*, **51** (11) (2009), 1039–1064.

47 Qian Y.; Soong F.K.; Yan Z.-J.: A unified trajectory tiling approach to high quality speech rendering. *IEEE Trans. Audio. Speech Lang. Process.*, **21** (2) (2013), 280–290.

48 Saito Y.; Takamichi S.; Saruwatari H.: Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **26** (1) (2018), 84–96.

49 Wang Y. *et al.*: Tacotron: towards end-to-end speech synthesis, *arXiv preprint arXiv:1703.10135*, last accessed: 2018-10-17, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135.

50 van den Oord A. *et al.*: Wavenet: a generative model for raw audio, in *ISCA Speech Synthesis Workshop (SSW), Sunnyvale, California, USA*, 2016, 1–15.

51 Wu Z. *et al.*: SAS: a speaker verification spoofing database containing diverse attacks, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, 2015, 4440–4444.

52 De Leon P.L.; Pucher M.; Yamagishi J.; Hernaez I.; Saratxaga I.: Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Trans. Audio Speech Lang. Process.*, **20** (8) (2012), 2280–2290.

53 Bonastre J.F.; Matrouf D.; Fredouille C.: Artificial impostor voice transformation effects on false acceptance rates, in *INTERSPEECH*, Antwerp, Belgium, 2007, 2053–2056.

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

55 Wu Z.; Li H.: Voice conversion versus speaker verification: An overview. *APSIPA Trans. Signal Inf. Process.*, **3** (2014).

56 Stylianou Y.; Cappé O.; Moulines E.: Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process.*, **6** (2) (1998), 131–142.

57 Kim E.-K.; Lee S.; Oh Y.-H.: Hidden Markov model based voice conversion using dynamic characteristics of speaker, in *European Conf. on Speech Communication and Technology*, Rhodes, Greece, 1997, 1–4.

58 Sundermann D.; Hoge H.; Bonafonte A.; Ney H.; Black A.; Narayanan S.: Text-independent voice conversion based on unit selection, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, 2006, I–81–I–84.

59 Wilde M.M.; Martinez A.B.: Probabilistic principal component analysis applied to voice conversion, in *IEEE Asilomar Conf. on Signals, Systems and Computers*, vol. **2**, Pacific Grove, California, 2004, 2255–2259.

60 Zhang S.; Huang D.; Xie L.; Chng E.S.; Li H.; Dong M.: Non-negative matrix factorization using stable alternating direction

method of multipliers for source separation, in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA)*, 2015, 222–228.

61 Desai S.; Raghavendra E.V.; Yegnanarayana B.; Black A.W.; Prahallad K.: Voice conversion using artificial neural networks, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, 3893–3896.

62 Abe M.; Nakamura S.; Shikano K.; Kuwabara H.: Voice conversion through vector quantization. *J. Acoust. Soc. Jpn.*, **11** (2) (1990), 71–76.

63 Erro D.; Moreno A.: Weighted frequency warping for voice conversion, in *INTERSPEECH*, Antwerp, Belgium, 2007, 1965–1968.

64 Lindberg J.; Blomberg M.: Vulnerability in speaker verification-a study of technical impostor techniques, in *EUROSPEECH*, vol. *99*, Budapest, Hungary, 1999, 1211–1214.

65 Villalba J.; Lleida E.: Speaker verification performance degradation against spoofing and tampering attacks, in *FALA Workshop*, Vigo, Spain, 2010, 131–134.

66 Villalba J.; Lleida E.: Detecting replay attacks from far-field recordings on speaker verification systems, in *European Workshop on Biometrics and Identity Management*, Roskilde, Denmark, 2011, 274–285.

67 Quatieri T.F.: Discrete-Time Speech Signal Processing: Principles and Practice, 1st ed., *Pearson Education India*, 2006.

68 Alegre F.; Janicki A.; Evans N.: Re-assessing the threat of replay spoofing attacks against automatic speaker verification, in *IEEE Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2014, 1–6.

69 Janicki A.; Alegre F.; Evans N.: An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks. *Secur Commun Netw*, **9** (15) (2016), 3030–3044.

70 Rafi B.S.M.; Murty K.S.R.; Nayak S.: A new approach for robust replay spoof detection in ASV systems, in *IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, 2017, 51–55.

71 De Leon P.L.; Pucher M.; Yamagishi J.: Evaluation of the vulnerability of speaker verification to synthetic speech, in *Odyssey The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

72 Wu Z.; Kinnunen T.; Chng E.S.; Li H.; Ambikairajah E.: A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case, in *IEEE Asia-Pacific Signal & Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, Hollywood, California, 2012, 1–5.

73 Alegre F.; Amehraye A.; Evans N.: A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns, in *IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, Washington DC, USA, 2013, 1–8.

74 Wu Z.; Gao S.; Cling E.S.; Li H.: 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 Conf. (APSIPA)*, Chiang Mai, Thailand, 2014, 1–5.

75 Larcher A.; Lee K.A.; Ma B.; Li H.: Text-dependent speaker verification: classifiers, databases and RSR2015. *Speech Commun.*, **60** (2014), 56–77.

76 Korshunov P. *et al.*: Overview of BTAS 2016 speaker anti-spoofing competition, Idiap, Tech. Rep., 2016.

77 Kinnunen T. *et al.*: Reddots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, Louisiana, USA, 2017, 5395–5399.

78 Todisco M. *et al.*: Asvspoof 2019: future horizons in spoofed and fake audio detection, *arXiv preprint arXiv:1904.05441*, 2019.

79 Campbell J.P.: Speaker recognition: a tutorial. *Proc. IEEE*, **85** (9) (1997), 1437–1462.

80 Lee K.A. *et al.*: The RedDots data collection for speaker recognition, in *INTERSPEECH*, Dresden, Germany, 2015, 2996–3000.

81 Delgado H. *et al.*: ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements, in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, 296–303.

82 ASVspoof 2019: automatic speaker verification spoofing and countermeasures challenge evaluation plan, 2019.

83 Kinnunen T. *et al.*: 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, 312–319.

84 Gong Y.; Yang J.; Huber J.; MacKnight M.; Poellabauer C.: ReMASC: realistic replay attack corpus for voice controlled systems, *arXiv preprint arXiv:1904.03365*, 2019.

85 Bimbot F. *et al.*: A tutorial on text-independent speaker verification. *EURASIP J. Adv. Signal Process.*, **2004** (4) (2004), 101962.

86 Martin A. *et al.*: The DET curve in assessment of decision task performance, in *European Conf. on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, 1997, 1895–1898.

87 Patil H.A.; Dutta P.; Basu T.: Effectiveness of LP based features for identification of professional mimics in Indian languages, in *Int. Workshop on Multimodal User Authentication, MMUA06, Toulouse, France*, 2006, 11–18.

88 Leon P.L.D.; Stewart B.; Yamagishi J.: Synthetic speech discrimination using pitch pattern statistics derived from image analysis, in *INTERSPEECH*, Portland, Oregon, 2012.

89 Wu Z.; Xiao X.; Chng E.S.; Li H.: Synthetic speech detection using temporal modulation feature, in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, 7234–7238.

90 De Leon P.L.; Hernaez I.; Saratxaga I.; Pucher M.; Yamagishi J.: Detection of synthetic speech for the problem of imposture, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, 4844–4847.

91 Kamble M.R.; Patil H.A.: Novel energy separation based instantaneous frequency features for spoof speech detection, in *IEEE European Signal Processing Conf. (EUSIPCO)*, Kos Island, Greece, 2017, 106–110.

92 Todisco M.; Delgado H.; Evans N.: A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients, in *Speaker Odyssey Workshop*, Bilbao, Spain, vol. **25**, 2016, 249–252.

93 Patel T.B.; Patil H.A.: Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural *vs.* spoofed speech, in *INTERSPEECH*, Dresden, Germany, 2015, 2062–2066.

94 Todisco M.; Delgado H.; Evans N.: Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification. *Comput. Speech Lang.*, **45** (2017), 516–535.

95 Xiao X.; Tian X.; Du S.; Xu H.; Chng E.S.; Li H.: Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge, in *INTERSPEECH*, Dresden, Germany, 2015, 2052–2056.

96 Sahidullah M.; Kinnunen T.; Hanilçi C.: A comparison of features for synthetic speech detection, in *INTERSPEECH*, Dresden, Germany, 2015, 2087–2091.

97 Yang J.; You C.; He Q.: Feature with complementarity of statistics and principal information for spoofing detection, in *INTERSPEECH*, Hyderabad, India, 2018, 651–655.

98 Sriskandaraja K.; Sethu V.; Ambikairajah E.; Li H.: Front-end for antispoofing countermeasures in speaker verification: scattering spectral decomposition. *IEEE J. Sel. Top. Signal Process.*, **11** (4) (2017), 632–643.

99 Chen N.; Qian Y.; Dinkel H.; Chen B.; Yu K.: Robust deep feature for spoofing detection – the SJTU system for ASVspoof 2015 challenge, in *INTERSPEECH*, Dresden, Germany, 2015, 2052–2056.

100 Zhang C.; Yu C.; Hansen J.H.: An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE J. Sel. Top. Signal Process.*, **11** (4) (2017), 684–694.

101 Qian Y.; Chen N.; Yu K.: Deep features for automatic spoofing detection. *Speech Commun. Elsevier*, **85** (2016), 43–52.

102 Alam M.J.; Kenny P.; Gupta V.; Stafylakis T.: Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks, in *Odyssey 2016*, Bilbao, Spain, 2016, 270–276.

103 Sailor H.B.; Kamble M.R.; Patil H.A.: Unsupervised representation learning using convolutional restricted Boltzmann machine for spoof speech detection, in *INTERSPEECH*, Stockholm, Sweden, 2017, 2601–2605.

104 Yu H.; Tan Z.-H.; Zhang Y.; Ma Z.; Guo J.: DNN filter bank cepstral coefficients for spoofing detection. *IEEE Access*, **5** (2017), 4779–4787.

105 Muckenhirn H.; Korshunov P.; Magimai-Doss M.; Marcel S.: Long-term spectral statistics for voice presentation attack detection. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **25** (11) (2017), 2098–2111.

106 Zhang Z.; Cummins N.; Schuller B.: Advanced data exploitation in speech analysis: an overview. *IEEE Signal Process. Mag.*, **34** (4) (2017), 107–129.

107 Heittola T.; çakır E.; Virtanen T.: The machine learning approach for analysis of sound scenes and events, in *Computational Analysis of Sound Scenes and Events*, *Springer*, 2018, 13–40.

108 Tokozume Y.; Harada T.: Learning environmental sounds with end-to-end convolutional neural network, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, 2721–2725.

109 Chiu C.-C. *et al.*: State-of-the-art speech recognition with sequence-to-sequence models, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, 1–5.

110 Muckenhirn H.; Magimai-Doss M.; Marcel S.: End-to-end convolutional neural network-based voice presentation attack detection, in *IEEE Int. Joint Conf. on Biometrics (IJCB)*, Denver, Colorado, USA, 2017, 335–341.

111 Dinkel H.; Chen N.; Qian Y.; Yu K.: End-to-end spoofing detection with raw waveform CLDNNS, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, 4860–4864.

112 Dinkel H.; Qian Y.; Yu K.: Investigating raw wave deep neural networks for end-to-end speaker spoofing detection. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **26** (11) (2018), 1–13.

113 Villalba J.; Miguel A.; Ortega A.; Lleida E.: Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge, in *INTERSPEECH*, Dresden, Germany, 2015, 2067–2071.

114 Yu H.; Tan Z.-H.; Ma Z.; Martin R.; Guo J.: Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features. *IEEE Trans. Neural Netw. Learn. Syst.*, **29** (10) (2017), 1–12.

115 Shang W.; Stevenson M.: Score normalization in playback attack detection, in *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, Adam's Mark Hotel Dallas, TX, USA, 2010, 1678–1681.

116 Font R.; Espín J.M.; Cano M.J.: Experimental analysis of features for replay attack detection results on the ASVspoof 2017 challenge, in *INTERSPEECH*, Stockholm, Sweden, 2017, 7–11.

117 Witkowski M.; Kacprzak S.; Zelasko P.; Kowalczyk K.; Gałka J.: Audio replay attack detection using high-frequency features, in *INTERSPEECH*, Stockholm, Sweden, 2017, 27–31.

118 Wang X.; Xiao Y.; Zhu X.: Feature selection based on CQCCs for automatic speaker verification spoofing, in *INTERSPEECH*, Stockholm, Sweden, 2017, 32–36.

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

120 Patil H.A.; Kamble M.R.; Patel T.B.; Soni M.: Novel variable length Teager energy separation based instantaneous frequency features for replay detection, in *INTERSPEECH*, Stockholm, Sweden, 2017, 12–16.

121 Kamble M.R.; Patil H.A.: Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection, in *INTERSPEECH*, Hyderabad, India, 2018, 646–650.

122 Kamble M.R.; Tak H.; Patil H.A.: Effectiveness of speech demodulation-based features for replay spoof speech detection, *INTERSPEECH*, Hyderabad, India, 2018, 641–645.

123 Jelil S.; Das R.K.; Prasanna S.M.; Sinha R.: Spoof detection using source, instantaneous frequency and cepstral features, in *INTERSPEECH*, Stockholm, Sweden, 2017, 22–26.

124 Tapkir P.A.; Patil H.A.: Novel empirical mode decomposition cepstral features for replay spoof detection, in *INTERSPEECH*, Hyderabad, India, September 2–6, 2018, 721–725.

125 Gunendradasan T.; Irtza S.; Ambikairajah E.; Epps J.: Transmission line cochlear model based am-fm features for replay attack detection, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, 6136–6140.

126 Wickramasinghe B.; Ambikairajah E.; Epps J.; Sethu V.; Li H.: Auditory inspired spatial differentiation for replay spoofing attack detection, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, 6011–6015.

127 Patil A.T.; Rajul A.; Sai P.A.K.; Patil H.A.: Energy sepration-based instantaneous frequency estimation for cochlear cepstral feature for replay spoof detection, in *INTERSPEECH*, Graz, Austria, 2019, 1–5, accepted.

128 Tak H.; Patil H.A.: Novel linear frequency residual cepstral features for replay attack detection, in *INTERSPEECH*, Hyderabad, India, September 2–6, 2018, 726–730.

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

130 Srinivas K.; Patil H.A.: Relative phase shift features for replay spoof detection system, in *Spoken Language Technologies for Under-resourced languages (SLTU)*, Gurugram, India, 2018, 1–5.

131 Saranya M.S.; Padmanabhan R.; Murthy H.A.: Replay attack detection in speaker verification using non-voiced segments and decision level feature switching, in *IEEE Int. Conf. on Signal Processing and Communications (SPCOM)*, Indian Institute of Science (IISc), Bangalore, 2018, 1–5.

132 Suthokumar G.; Sriskandaraja K.; Sethu V.; Wijenayake C.; Ambikairajah E.: Phoneme specific modelling and scoring techniques for anti spoofing system, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, 6106–6110.

133 Liu M.; Wang L.; Dang J.; Nakagawa S.; Guan H.; Li X.: Replay attack detection using magnitude and phase information with attention-based adaptive filters, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, 6201–6205.

134 Kamble M.R.; Patil H.A.: Analysis of reverberation via teager energy features for replay spoof speech detection, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, 2607–2611.

135 Cai W.; Cai D.; Liu W.; Li G.; Li M.: Countermeasures for automatic speaker verification replay spoofing attack: on data augmentation, feature representation, classification and fusion, in *INTERSPEECH*, Stockholm, Sweden, 2017, 17–21.

136 Nagarsheth P.; Khoury E.; Patil K.; Garland M.: Replay attack detection using DNN for channel discrimination, in *INTERSPEECH*, Stockholm, Sweden, 2017, 97–101.

137 Chen Z.; Xie Z.; Zhang W.; Xu X.: ResNet and model fusion for automatic spoofing detection, in *INTERSPEECH*, Stockholm, Sweden, 2017, 102–106.

138 Lavrentyeva G.; Novoselov S.; Malykh E.; Kozlov A.; Kudashev O.; Shchemelinin V.: Audio replay attack detection with deep learning frameworks, in *INTERSPEECH*, Stockholm, Sweden, 2017, 82–86.

139 Sailor H.B.; Kamble M.R.; Patil H.A.: Auditory filterbank learning for temporal modulation features in replay spoof speech detection, *INTERSPEECH*, Hyderabad, India, 2018, 666–670.

140 Valenti G.; Delgado H.; Todisco M.; Evans N.; Pilati L.: An end-to-end spoofing countermeasure for automatic speaker verification using evolving recurrent neural networks, in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, 288–295.

141 Tom F.; Jain M.; Dey P.: End-to-end audio replay attack detection using deep convolutional networks with attention, in *INTERSPEECH*, Hyderabad, India, 2018, 681–685.

142 Lai C.-I.; Abad A.; Richmond K.; Yamagishi J.; Dehak N.; King S.: Attentive filtering networks for audio replay attack detection, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, 6316–6320.

143 Kamble M.R.; Patil H.A.: Novel amplitude weighted frequency modulation features for replay spoof detection, in *Int. Symp. on Chinese Spoken Language Processing (ISCSLP)*, Taipei, Taiwan, 2018, to appear.

144 Alam M.J.; Bhattacharya G.; Kenny P.: Boosting the performance of spoofing detection systems on replay attacks using q-logarithm domain feature normalization, in *Odyssey The Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, 393–398.

145 Tapkir P.A.; Kamble M.R.; Patil H.A.: Replay spoof detection using power function based features, in *Asia-Pacific Signal and Information Processing Association, Annual Summit and Conf. (APSIPA-ASC)*, Hawaii, USA, 2018, to appear.

146 Jelil S.; Kalita S.; Prasanna S.M.; Sinha R.: Exploration of compressed ILPR features for replay attack detection, in *INTERSPEECH*, Hyderabad, India, 2018, 631–635.

147 Kamble M.R.; Tak H.; Maddala S.K.; Patil H.A.: Novel demodulation-based features using classifier-level fusion of GMM and CNN for replay detection, in *Int. Symp. on Chinese Spoken Language Processing (ISCSLP)*, Taipei, Taiwan, 2018, to appear.

148 Stylianou Y.: Removing linear phase mismatches in concatenative speech synthesis. *IEEE Trans. Speech Audio Process.*, **9** (3) (2001), 232–239.

149 Wu Z. *et al.*: Anti-spoofing for text-independent speaker verification: an initial database, comparison of countermeasures, and human performance. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **24** (4) (2016), 768–783.

150 Wester M.; Wu Z.; Yamagishi J.: Human *vs.* machine spoofing detection on wideband and narrowband data, in *INTERSPEECH 2015*, Dresden, Germany, September 2015, 2047–2051.

151 Wang X.; Lorenzo-Trueba J.; Takaki S.; Juvela L.; Yamagishi J.: A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis, in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, 1–15.

152 Lorenzo-Trueba J.; Fang F.; Wang X.; Echizen I.; Yamagishi J.; Kinnunen T.: Can we steal your vocal identity from the internet? Initial investigation of cloning obama's voice using gan, wavenet and low-quality found data, in *Odyssey*, Les Sables d'Olonne, France, 2018, 240–247.

**Madhu R. Kamble** is a Ph.D. student at DA-IICT, Gandhinagar. She did her M. Tech. degree from Cummins College of Engineering, Pune, Maharashtra, India in 2015 in Signal Processing specialization and the B.Tech degree from P.V.P.I.T, Budhgaon, Sangli, Maharashtra in 2012. She has been awarded with Rajiv Gandhi National Fellowship (RGNF) for her doctoral research studies. Her research interest is in voice biometrics, in particular, analysis of spoofing attacks and development of countermeasures. Recently, she offered a tutorial jointly with Prof. Patil on the same topic in IEEE-WIE Conference, at AISSM's Pune in Dec 2016. She was the co-instructor for a tutorial in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Kuala Lumpur, Malaysia, 2017. She is a research intern at Samsung Research Institute, Bangalore (SRI-B), India during May-Nov 2019. She is a student member of ISCA, student member of IEEE, IEEE Signal Processing Society, and APSIPA. She is a reviewer for Computer, Speech and Language and Nerocomputing Journal, Elsevier. She received ISCA and IEEE SPS student travel grant to present her papers during INTERSPEECH, 2017 and ICASSP 2019.

**Hardik B. Sailor** is a Post Doctoral researcher in the University of Sheffield, UK. He completed his Ph.D. degree in 2018 at DA-IICT, Gandhinagar, India. He received the B.E. degree from Government Engg. College (GEC), Surat in 2010. In 2013, he received the M.Tech. degree from DA-IICT, Gandhinagar. He was also a project staff member of MeitY, Govt. of India sponsored consortium project, "Automatic Speech Recognition for Agricultural Commodities Phase-II", (April 2016–March 2018). At DA-IICT, he was a project staff member of MeitY, Govt. of India sponsored project on, "Development of Text-to-Speech (TTS) Synthesis Systems for Indian languages Phase-II", from May 2012 to March 2016. His research area includes representation learning, auditory processing, Automatic Speech Recognition (ASR), and sound classification. His main research is focused on developing representation learning to model the auditory processing. He has published 25 research papers in top conferences and peer-reviewed journals. He is a student member of IEEE, IEEE Signal Processing Society, and International Speech Communication Association (ISCA). He is a reviewer for IEEE/ACM Transactions in Audio, Speech, and Language Processing, IEEE Signal Processing Letters, and Applied Acoustics, Elsevier. Recently, he received ISCA student travel grant 650 euros to present his three co-authored papers during INTERSPEECH 2018.

**Hemant A. Patil** received the B.E. degree from the North Maharashtra University, Jalgaon, India, in 1999, the M.E. degree from Swami Ramanand Teerth Marathwada University, Nanded, India, in 2000, and the Ph.D. degree from the Indian Institute of Technology, Kharagpur, India, in 2006. He is currently a Professor at Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India. He has coedited a book with Dr. A. Neustein (Editor-in-Chief, IJST, Springer) on Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism (New York, NY, USA: Springer). He served as PI/Co-PI for three MeitY and two DST sponsored projects. Prof. Patil was a chair of satellite workshop committee during INTERSPEECH, 2018, Hyderabad, India. He is selected as APSIPA Distinguished Lecturer (DL) for 2018–2019 and delivered 21 APSIPA DLs in three countries, namely, India, Canada, and China. Recently, he is elected as ISCA DL for 2020–2021. He is an affiliate member of the IEEE SLTC and a member of the IEEE Signal Processing Society, the IEEE Circuits and Systems Society (Awards), and the International Speech Communication Association (ISCA).

**Haizhou Li** (M'91-SM'01-F'14) received the B.Sc., M.Sc., and Ph.D degrees in electrical and electronic engineering from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is currently a Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore. He is also a Conjoint Professor at the University of New South Wales, Kensington, NSW, Australia. Prior to joining NUS, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor with CRIN in France (1994–1995), Research Manager with the Apple-ISS Research Centre (1996–1998), Research Director with Lernout & Hauspie Asia Pacific (1999–2001), Vice President with InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology with the Institute for Infocomm Research, Singapore (2003–2016). His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. He is currently the Editor-in-Chief for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2015–2018), a Member of the Editorial Board of Computer Speech and Language (2012–2018). He was an elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012 and INTERSPEECH 2014. He was the recipient of the National Infocomm Award 2002 and the Presidents Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation.