Computer Speech & Language 65 (2021) 101140

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Detection of replay spoof speech using teager energy feature cues

Madhu R. Kamble*, Hemant A. Patil

Speech Research Lab Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

ARTICLE INFO

Article History:
Received 31 October 2019
Revised 18 July 2020
Accepted 11 August 2020
Available online 14 August 2020

Keywords:
Automatic speaker verification
Spoof
Replay
Reverberation
TEO profile

ABSTRACT

The vulnerability of Automatic Speaker Verification (ASV) systems to spoofing or presentation attacks is still an open security issue. In this context, replay spoofing attacks pose a great threat to an ASV system since they can be easily performed (using a playback device, and without needing any technical skill). In this paper, we analyze replay speech signals in terms of reverberation that may occur during recording of the speech signal. Such reverberation introduces delay and changes in amplitude, producing close copies of speech signals, which significantly influences the replay components. To that effect, we propose to exploit the capabilities of the Teager Energy Operator (TEO) to compute a running estimate of subband energies for replay *vs.* genuine signals. We have used a linearly-spaced Gabor filterbank to obtain a narrowband filtered signal. The TEO has the ability to track the instantaneous changes of a signal. Experiments are performed on the ASVspoof 2017 Challenge version 2.0 database using a Gaussian Mixture Model (GMM) as pattern classifier. Furthermore, we compared our results with state-of-the-art feature sets, namely, Constant Q Cepstral Coefficients (CQCC), Linear Frequency Cepstral Coefficients (LFCC), Mel Frequency Cepstral Coefficients (MFCC), and used their score-level fusion with the proposed feature sets, i.e., Teager Energy Cepstral Coefficients (TECC), in order to obtain possible *complementary* information that further reduces the Equal Error Rate (EER). Relatively low EERs are obtained with score-level fusion of CQCC, MFCC, LFCC, and TECC feature sets, resulted in 6.68% and 10.45% on development and evaluation sets, respectively. Moreover, for the evaluation dataset, we also studied the performance of the TECC feature set on different Replay Configurations (RC), namely, for acoustic environments, playback, and recording devices. For all the levels of threat conditions (i.e., low, medium, and high-level) to an ASV system, the proposed feature set performed better compared to existing state-of-the-art feature sets. In addition to the ASVspoof 2017 Challenge database, we also performed experiments on other spoofing databases, namely, the ASVspoof 2015 Challenge, BTAS 2016, and ASVspoof 2019 Challenge databases. For all the spoofing databases used in this study, the proposed TECC feature set perform significantly better than the other feature sets.

*Corresponding author at Ph.D. Research Scholar at DA-IICT, Gandhinagar, Gujarat, India
E-mail addresses: madhu_kamble@daiict.ac.in, mk310191@gmail.com (M.R. Kamble).

https://doi.org/10.1016/j.csl.2020.101140
0885-2308/

1. Introduction

An Automatic Speaker Verification (ASV) system (also known as voice biometrics) is a speaker authentication system to verify a claimed speaker's identity with the help of machines (Evans et al., 2013). The biometric traits that have been successfully used in practical applications include voice (Jain et al., 2006), face (Jain and Li, 2011), fingerprint (Maltoni et al., 2009), iris (Daugman, 2003), palmprint (Connie et al., 2005), and hand geometry (Sanchez-Reillo et al., 2000; Jain et al., 2016). However, ASV systems are exposed to the possibility of being attacked by spoofed speech signals, such as replay (Alegre et al., 2014), impersonation (Lau et al., 2004), speech synthesis (SS) (Zen et al., 2009), voice conversion (VC) (Stylianou, 2009), and speech from twins (Rosenberg, 1976). Spoofing attacks in biometrics are also known as *presentation attacks* as per the International Organization for Standardization (ISO), and the International Electro-technical Commission (IEC) (Koppell, 2011). In practice, we would like an ASV system to be robust against variations such as microphone and transmission channel, intersession effects, acoustic noise, and speaker aging. Due to recent technological developments, it is possible to generate spoof data that resemble very closely their natural counterparts, such as replay spoofing.

Among all the spoofing attacks, replay attacks are a major threat to any ASV system since they can be easily mounted (using playback of recorded voice, without requiring specialized technical skills) (Alegre et al., 2014). To promote the research in development of countermeasures for the replay spoof speech detection (SSD) task, first of its kind the ASVspoof 2017 Challenge was organized as part of a special session during INTERSPEECH 2017 (Kinnunen et al., 2017a). The goal of this challenge was to develop replay SSD using the acoustical characteristics of the genuine *vs.* replay speech. The replayed speech may contain unknown background noises, reverberation, and channel noise.

Various attempts have been made in the ASVspoof 2017 Challenge to detect replay attacks. In particular, various acoustical features, such as spectral, time-domain, excitation source, and neural network-based features have been investigated for the replay SSD task. Such features include Subband Spectral Centroid Magnitude Coefficients (SCMC) (Font et al., 2017), Single Frequency Filter (SFF) (Alluri et al., 2017), Instantaneous Frequency (IF)-based features (Patil et al., 2017; Jelil et al., 2017), Inverse-Mel Cepstral Coefficients (IMFCC) (Font et al., 2017), Rectangular Filter Cepstral Coefficients (RFCC) (Font et al., 2017), and Scattering Coefficients (Sriskandaraja et al., 2017). The importance of signal mass is studied to give a more precise estimate of a signal's energy using an Enhanced Teager Energy Operator (ETEO) (Acharya et al., 2019). Speech demodulation features are also studied for various spoofing databases in Kamble et al. (2019). Some approaches used deep learning techniques (for both feature extraction and classification) along with the feature normalization method (Lavrentyeva et al., 2017; Chen et al., 2017).

A special session focusing on Replay Attack Anti-Spoofing Measures for ASV Systems was held during APSIPA ASC 2018 ASV. In this special session, various countermeasures focusing on the replay database were investigated. One of the approaches reported in Yang et al. (2018) used a combined feature referred to as extended CQCC (eCQCC), which is extracted using an octave power spectrum along with linear spectrum, and it is found to capture complementary information beyond the CQCC feature set. In Das and Li (2018), phase and excitation source-based information was used to capture additional artifacts that are useful for identifying replay attacks. In particular, they used instantaneous frequency cosine coefficients and two source features, namely, Discrete Cosine Transform of the integrated linear prediction residual and residual Mel frequency cepstral coefficients. Another approach reported in Suthokumar et al. (2018) investigates the potential advantages of sharing speaker data between the ASV system and the replay detection system. They found benefits of using the claimed speaker's model in place of the common genuine speaker's model. Furthermore, the importance of using power function-based features, namely, Power Normalized Cepstral Coefficients (PNCC) and Q-Log Normalized Cepstral Coefficients (QLNCC), was studied in Kim and Stern (2016), Pardede (2015) and Tapkir et al. (2018). The PNCC and QLNCC feature sets are noise robust and they are able to capture speaker-specific information in noisy environments. Concluding the overall special session, a survey paper focusing on replay attack, in particular, the 2nd ASVspoof 2017 Challenge, was studied in Patil and Kamble (Hawaii, USA, 2018). This paper presents critical analysis of state-of-the-art techniques, various countermeasures, and databases, and also aims to present current limitations along with the road map ahead, i.e., future research directions in this technological challenging problem.

Recent research focuses on improving the novel features or improving the back-end modeling for the SSD task. The features computed have to deal with speaker variability, phonetic variability, channel effects, and acoustical environment (Suthokumar et al., 2018). Replay speech is mainly distorted because of background noise (such as air-conditioners, computers, or any other source in room A or B); echo (signal from room A returning to this room through speaker-microphone coupling in room B); and reverberation (acoustical properties of rooms A or B) (de Lima et al., 2008). In recent years, there has been significant progress on reverberant speech processing in the field of speech recognition and audio processing. Research in these fields focuses particularly on combining ideas from room acoustics, optimal filtering, machine learning, speech modeling, enhancement, and recognition (Yoshioka et al., 2012). Several dereverberation techniques were developed, such as blind deconvolution and nonnegative matrix factorization, for audio processing research. This study provides a deeper analysis of replay speech signals from the view-point of reverberation. Replay speech gets distorted by both interfering sounds and reverberation caused because of the recording of a target speaker's voice from a distance. While natural/genuine speech may also has reverberation (especially if it not recorded in an infinite space or rooms with signal absorbing surfaces). However, replay speech is reverberated *twice* as it is recorded twice and thus, the relative degree of reverberation in genuine *vs.* replay speech is different and thus this can serve as a discriminatory feature to aid in the SSD task.

This paper is an extension of Kamble and Patil (2019) in terms of depth of literature search, quantitative analysis of proposed idea, additional experiments, and results on additional datasets. In particular, this paper presents the below-mentioned novelty:

- In Kamble and Patil (2019), a qualitative analysis of reverberation effect was presented whereas in this paper, we present a detailed literature search and quantitative mathematical analysis to support our hypothesis. In particular, this paper provides analysis in terms of the modeling of replay signals and associated reverberation mechanisms. To that effect, we investigate how the reverberation affects the genuine signal by producing close copies of the genuine signal, which are apparently related to the transmissions and reflections of the recording acoustic environment.
- In this work, we analyzed genuine *vs.* replay (reverberated) speech signals in the time-domain by exploiting the mathematical structure of the TEO in order to capture characteristics of reverberation. In particular, extra pulses are observed when the replay speech signal is recorded in a close room, such as bedroom and office, whereas they are not observed when the signal is recorded in an open acoustic environment. This analysis of TEO profiles around Glottal Closure Instant locations shows the effect of reverberation that depends on the closed and open acoustic environment conditions. The idea behind using the TEO is the nonlinear modeling of speech production. The TEO computes the true total energy source of a resonance signal and preserves both the amplitude and frequency information (Maragos et al., 1991). The supplementary information improves the time and frequency resolution, and in addition, the TEO also has noise suppression capability (due to it's mathematical structure) that further helps to detect a replay signal from its natural counterpart (Maragos et al., 1992; Jabloun and Cetin, 1999; Sailor and Patil, 2017). Given this, we exploit capability of TEO to capture nonlinear aspects of natural speech production (e.g., properties of airflow pattern, violation of acoustic impedance from linear acoustics, 180 degree phase shift in airflow at different locations in vocal tract, modulations in acoustical energy, etc.), and also to capture characteristics of reverberation via impulse-like TEO energy profile.
- Additional Results on ASVSpoof 2017 version 2 Dataset: The experiments are performed for effect of number of subband filters and their bandwidths (which were kept fixed in Kamble and Patil (2019)) for proposed TECC feature extraction. The frequency resolution of the linearly-spaced Gabor filterbank is explicitly related to the number of subband filters in a filterbank. By increasing the number of subband filters, the frequency resolution is improved and thus, it is found to capture more detailed spectral characteristics. In addition, the relevant theory of Heisenberg's uncertainty principle is presented in the context of choice of *optimal* Gabor filterbank for TECC. Furthermore, results are compared with and without applying filterbank; and for effect of shape of subband filters, in particular, linearly-spaced Gabor filterbank *vs.* linear-spaced Mel filterbank. The analysis of spectral energy densities of natural and corresponding replay signal is also presented. The score-level fusion with two feature sets at a time (as in Kamble and Patil, 2019) is extended to four feature sets, i.e., CQCC, LFCC, MFCC, and TECC. For evaluation set, the performance of TECC feature set is compared with LFCC, and MFCC feature sets than with CQCC alone (as in Kamble and Patil, 2019).
- Experiments with Other Datasets: To investigate the performance of proposed TECC feature set, experiments are performed on additional datasets, namely, ASVspoof 2015, BTAS 2016 ASVspoof 2019, and recently released ASVSpoof 2019 real PA database. In addition, the comparison of real replay (from ASVspoof 2017 version 2.0) *vs.* synthetic replay (ASVspoof 2019 PA task) in terms of TEO profiles is also analyzed. Furthermore, we also observe the distribution patterns for log-likelihood ratio (LLR) scores for the ASVspoof 2017, and ASVspoof 2019 challenge databases in order to justify suitability of proposed feature set for replay SSD problem.

In addition, reverberation introduces delay in genuine speech components corresponding to different reflections that further depends on the environmental conditions. The reverberation transforms a mono-component signal into a multi-component one, where they are spectrally very close and hence, we cannot separate the natural components from the replay components (Arroabarren et al., 2006). We attempt to avoid treating replay detection as a "machine learning black-box" rather to first understand the properties of reverberated speech, and exploit that for the replay SSD task. Furthermore, in this work, we analyzed genuine *vs.* replay (reverberated) speech signals in the time-domain. We further analyze the individual replay configuration in terms of EER with the proposed Teager Energy Cepstral Coefficients (TECC) feature set and compare the results with state-of-the-art feature sets, such as Constant-Q Cepstral Coefficients (CQCC), Mel Frequency Cepstral Coefficients (MFCC), and Linear Frequency Cepstral Coefficients (LFCC).

2. Effect of reverberation on the replay speech signal

2.1. Basics of the replay speech signal

The task of replay spoof detection is to identify whether a given speech sample is genuine speech or whether it is a recorded version of the genuine speech through intermediate (recording + playback) devices. An *intermediate device* in this context means the device used during the recording and playing it back in order to obtain the replayed speech signal. In particular, during recording, different kinds of microphone, speaker, and tape recorder are used. The scenario of the replay spoof speech detection (SSD) system is shown in Fig. 1. In particular, Fig. 1 shows the process of generation of the replay speech signal in two different acoustic environments, i.e., during recording and playback different kinds of microphone, speaker, tape recorder are used.

The genuine speech signal, $s[n]$, can be modeled as a convolution of glottal airflow, $p[n]$, with the impulse response of the vocal tract system, $h[n]$ (Quatieri, 2006), i.e.,

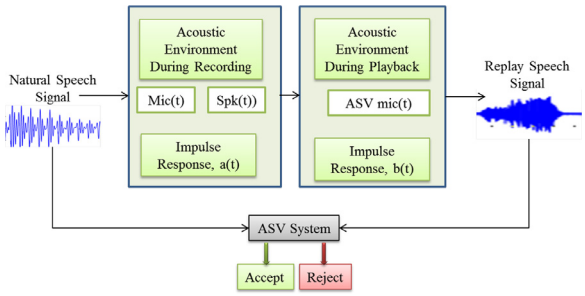

Fig. 1. Illustration of replay spoof attack scenario at ASV system.

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

It should be noted that as the convolution operation requires the assumption of a linear time-invariant (LTI) system, Eq. (1) and subsequent analysis in this paper is valid either for a segment (10-30 ms) of speech signal $s[n]$ or the impulse response of vocal tract system is fixed for a speech frame, i.e., $h[n]$ is dependent on the index of a speech frame.

On the other hand, the replay speech signal, $r_e[n]$, can be modeled as the convolution of the natural speech signal, $s[n]$, and the impulse response of intermediate devices, $h_1[n]$, (playback and recording device) along with the propagating acoustic environment and it is given by:

$$r_e[n] = s[n] * h_1[n], \tag{2}$$

where $h_1[n]$ has extra convolved components due to replay; in particular, it is the combination of the impulse responses of the recording device, $h_{mic}[n]$, recording environment, $a[n]$, playback device (speaker), $h_{spk}[n]$, and playback environment, $b[n]$, i.e., $h_1[n]$ *lumps* together all three impulse responses. In particular,

$$h_1[n] = h_{mic}[n] * a[n] * h_{spk}[n] * b[n]. \tag{3}$$

If there is a presence of extra additive noise, $\eta[n]$, then Eq. (2) becomes, however, here $\eta[n]$ is assumed to vanish ($\eta[n] = 0$).

$$r_e[n] = s[n] * h_1[n] + \eta[n]. \tag{4}$$

The speech signal recorded with the playback device contains convolutional and additive distortions from the intermediate devices. The most crucial part in the detection of replay attack occurs during the process of feature extraction. To obtain the discriminatory information for genuine and replay speech signals, the focus should be on the representation of the spectral characteristics obtained from the intermediate devices. Eq. (2) represents the convolution term that transforms to an additive relation when converted to the real cepstral domain (by ignoring phase information), and it is given by Rafi et al. (2017):

$$\mathbf{r_e} = \mathbf{s} + \mathbf{h_1}, \tag{5}$$

where $\mathbf{r_e}$, $\mathbf{s}$, and $\mathbf{h_1}$ represent the cepstral vectors of the replay, the genuine speech signal, and the impulse response of intermediate devices, respectively. Eq. (5) indicates that genuine and replayed speech are related by an additive relation due to the LTI assumption used in this work. The features extracted from the replay signal are also affected by the recording process.

The acoustical behavior of the speech signal recorded in different environments have differences in the speech signal. The speech signal when recorded in noisy environment will have distortion in the signal. However, its effect on the acoustical characterization of replay has yet to be analyzed. Replay speech is affected by the reverberation, which is included during recording of the speech signal and hence, the basics of reverberation effect are explained in the next Section 2.2.

2.2. Basics of reverberation

The replay speech signal is a re-recording of a target speaker's voice captured unknowingly from a distance with the help of a recording device. The recording can be done at different places, such as bedroom, balcony, canteen, office, or home. When the recording is done within a closed room, reverberation may be introduced severely by the replay mechanism. Reverberation is the phenomenon of sound that is a result of multiple reflections from surfaces, such as furniture, people, or air media in a closed surface (Rev). Sound energy builds up with each reflection and decays gradually as they are absorbed by the surfaces of objects in the enclosed space. The reflections can be 1st order (with only one deviation) or 2nd order (with two deviations) from walls, surfaces, and direct paths without any deviations.

The impulse response of a transmission is known to carry information of the acoustic environment, however, under an assumption of a Linear Time-Invariant (LTI) system (Houtgast and Steeneken, 1985; Blesser and Salter, 2009; Kuttruff, 2016). Conventionally, a replay signal (with reverberation), $s_{rev}[n]$, is modeled as a convolution of the natural speech signal, $s[n]$, with the impulse response of the acoustic environment, $r[n]$, (Kuc, 1988; Kinoshita et al., 2016), i.e.,

$$s_{rev}[n] = s[n] * r[n]. \tag{6}$$

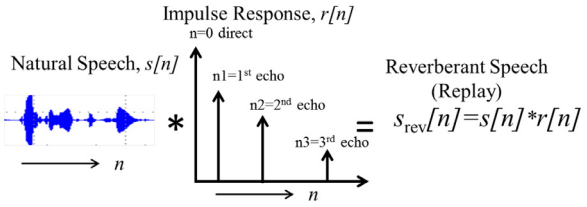

Fig. 2. Convolution of genuine speech with an impulse response (i.e., sequence of impulses at different echo locations) to obtain the reverberant replay speech signal.

If additive noise $\eta[n]$ is present, then Eq. (6) becomes Eq. (7), however, here $\eta[n]$ is assumed to vanish ($\eta[n]=0$).

$$s_{rev}[n]=s[n]*r[n]+\eta[n]. \tag{7}$$

The natural speech is repeated, time-shifted, and scaled for every non-zero point in the impulse response and the resulting signals are summed as shown via a schematic representation in Fig. 2.

The non-stationary mono-component signal can be mathematically expressed as (Boashash, 1991; Cohen, 1995):

$$s[n]=a[n]\ \cos\ \phi[n], \tag{8}$$

where $a[n]$ is the slowly-varying instantaneous amplitude and $\phi[n]$ is the instantaneous phase (Cohen, 1995). The non-stationary multi-component signal can be defined as the superposition of `M' numbers of mono-component signals given as:

$$s_{multicompnent}[n]=\sum_{i=1}^{M} a_i[n]\ \cos\ \phi_i[n]. \tag{9}$$

As discussed earlier, reverberation includes delay and changes in amplitude, forming close copies of a genuine signal that correspond to different reflections, modeling the reverberation and understanding how the parameters related to the model affect both physical and perceptual properties of reverberation (Wen, 2009; Traer and McDermott, 2016). Different types of reverberation models are a time-frequency room model, novel signal-based measurement, reverberation decay tail measure, and colouration measure (Wen, 2009). From a signal processing viewpoint and under the assumption of a fixed acoustic environment, reverberation can be modelled as a linear time-invariant (LTI) system with room impulse response (RIR), $h[n]$, with the input signal, $s[n]$, to give the output signal $s_{rev}[n]$. The reverberation process can then be written as the convolution between the input and the RIR:

$$s_{rev}[n]=\sum_{i} k_i s[n-n_i], \tag{10}$$

where $s[n]$ and $s_{rev}[n]$ are the genuine and reverberated signals, respectively, and k_i and n_i are the change in amplitude and delay of each sample, respectively, for i number of reflections that occurred in the closed room. When we compare Eq. (9) and Eq. (10), we can say that reverberation changes a mono-component signal into multi-component signals. The duplicates are spectrally very close to each other (Arroabarren et al., 2006).

Reverberation introduces delay and attenuation to produce close copies of the genuine signal corresponding to the different reflections of the acoustical signal in the environment (Arroabarren et al., 2006). It can be observed from Fig. 3 that the replay speech samples are shifted from the genuine components, and the amplitude also varies compared to the genuine signal.

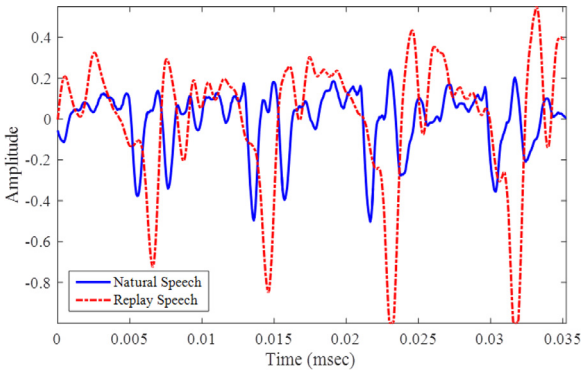

Fig. 3. Segment of speech signal showing the effect of reverberation for a replay signal (dotted line) in terms of delay in each speech sample, and changes in amplitude compared to the genuine speech segment (solid line).

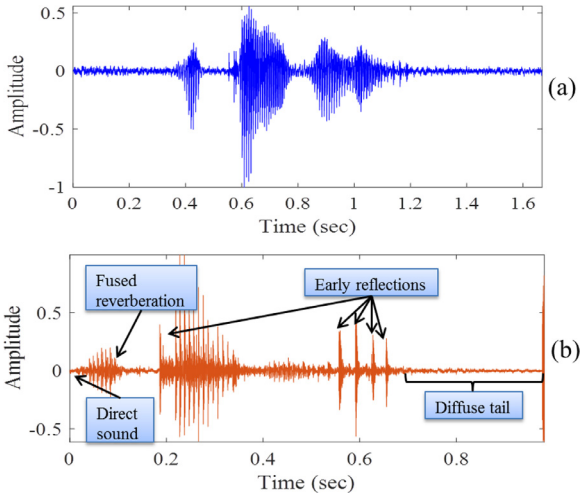

Fig. 4. Time-domain speech signal for speech that is (a) genuine and (b) reverberated (replay). After (Kamble and Patil, 2019; Houtgast and Steeneken, 1985).

Discrete early reflections (1st or 2nd order reflections) are typically involved in the early regions of an impulse response. The discrete early reflections can be simulated by means of a tapped delay line, which allows replicating some versions of the input signal, each delayed by a different amount (Traer and McDermott, 2016). The time-domain speech signals are shown for both genuine (Fig. 4(a)) and reverberated speech signals in (Fig. 4(b)). The reflections further become densely packed in the time-domain, composing a diffuse tail (as seen in Fig. 4(b)) (Traer and McDermott, 2016). The time of the peak indicates 'how long is the delay that a reflected signal will arrive at the recording device?', and the amplitude of the peak shows the amplitude of the reflected signal (Traer and McDermott, 2016). The first peak of the reverberated signal corresponds to the signal that arrives directly from the source of the recording, which arrives with the shortest possible delay. The other subsequent peaks arrive because of reflections, each related to its particular path that comes in its way. Eventually, the reflections become sufficiently dense that they indeed overlap in time. Because energy is absorbed by environmental surfaces with each reflection (as well as by air), longer paths produce lower amplitudes, and the overlapping echoes produce a "tail" in the impulse response that decays with time (Traer and McDermott, 2016).

If a room does not have any signal absorbing surfaces, such as wall, ceiling, and floor, the signal bounces back from surfaces and takes very long (theoretically, infinite) time for the signal to end. In such a room, the listener or the recording device will hear/record both the direct signal as well as the repeated reflected signal waves. If these reverberations are excessive, the sound may run together with a loss of articulation, and it may become *muddy* and also *garbled* (Rev). Larger rooms have few reflections, resulting in a slow decay of reverberated signals, and the decay rates are also affected by material, such as carpet, curtains, sofa-sets. Reverberation is also found to distort the structure of source signals in the spectral energy density as shown in Fig. 5 via spectrograms (Patil and Kamble, Hawaii, USA, 2018; Traer and McDermott, 2016; Kamble et al., 2018; Kamble and Patil, 2018). The time-domain speech signals for genuine (Panel I) and replay speech (Panel II) are shown in Fig. 5(a) corresponding to their spectral energies in Fig. 5(b). The highlighted regions in spectral energy densities show the distortions that are included because of reverberation. Distortion is due to *decay* in spectrum and hence, a kind of energy loss (Tak and Patil, 2018).

3. Analysis of reverberation using the TEO

An algorithm derived by Teager uses a nonlinear energy tracking operator (Teager, 1980; Teager and Teager, 1990; Kaiser, 1990). For a mono-component discrete-time signal, $x[n]$, the TEO, $(\Psi_d\{\cdot\})$, is defined as (Kaiser, 1990):

$$E_n = \Psi_d\{x[n]\} = x^2[n] - x[n-1]x[n+1], \tag{11}$$

where E_n gives the running estimate of the signal's energy that is under consideration. The TEO cannot be applied directly on a speech signal as it is the summation of multi-component signals. Hence, the speech signal is bandpass filtered to obtain '*N*' number of narrowband filtered signals, and then the TEO is applied on the *i*th narrowband filtered signals, i.e., $\Psi_d\{x_i[n]\}$.

Further simulation is performed to observe the effect of reverberation on the Teager energy profiles of synthetic speech (i.e., simulated genuine), and corresponding replay signals in Fig. 6. The train of impulses (Fig. 6(a)) is convolved with a damped sinusoid signal (Figure 6(b)), producing a convolved signal (Fig. 6(c)). Now, assume the convolved signal in Fig. 6(c)) is a simulated genuine speech signal. Now, to obtain a reverberated signal, we convolved the simulated genuine speech signal (Fig. 6(e)) with a train of impulses representing echo (Fig. 6(d)), and obtained a synthetic reverberated signal having close copies of the original genuine signal (Fig. 6(f)).

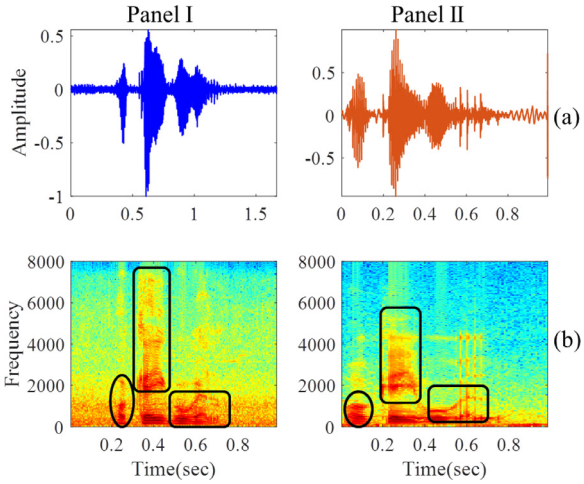

Fig. 5. (a) Time-domain speech signal, and (b) corresponding spectral energy densities via spectrograms of genuine (Panel I) and replay (Panel II) speech signals, highlighted regions oval, and boxes show the distorted spectral regions.

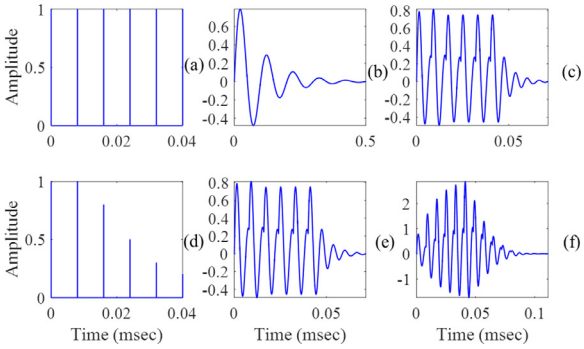

Fig. 6. (a-d) Train of impulses and echoes to model reverberation; (b) damped sinusoid signal; (c-e) convolved signal obtained from (a) and (b); (f) convolved signal obtained from (d) and (e).

The impulse response for the reverberated signal in Fig. 6 had an echo kept with a particular time interval, i.e., an impulse arrives after every 8 ms. However, in a real case scenario, it may not be the case, i.e., the echo impulses may arrive with a small interval gap or it may arrive with a large delay as well depending upon aspects of the acoustic environment.

We observed that the Teager energy traces of replay speech signal segment recorded for different environments, such as (Panel I) balcony, (Panel II) bedroom, (Panel III) canteen, and (Panel IV) office are as shown in Fig. 7(b). The extra pulses are observed when the replay speech signal is recorded in a closed room, such as a bedroom and office as shown in Fig. 7 (Panel II and Panel IV (b)). On the other hand, extra impulse-like energy traces are not observed for replay speech recorded on a balcony and canteen environments Fig. 7 (Panel I and Panel III (b)).

The TEO profiles show high energy pulses around the Glottal Closure Instant, because of impulse-like excitation to the vocal tract system, and this sudden glottal closure of the vocal folds produces high energy and thus, the TEO produces high energy around these regions (Patil and Parhi, 2010). Along with high Teager energy pulses, bumps are observed around the energy pulses, indicating significant contribution of nonlinear effects during the speech production process (Patil and Parhi, 2010).

In this Section, we studied modulations of energy estimated via a TEO profile to emphasize the impulse that arrives because of echo/reverberation. Furthermore, we observed that for different environments, the Teager energy traces obtained are different. In particular, for a closed room (such as bedroom, office) extra energy traces are observed because of echo impulse. Hence, these observations motivated us to extract features that are based on the energy traces, and thus, the proposed Teager Energy Cepstral Coefficients (TECC), which is discussed in the next Section.

3.1. Teager energy features

The functional block diagram of the proposed TECC feature extraction is shown in Fig. 8. Here, the input speech signal is passed through a pre-emphasis filter having a system function, $H(z) = 1 - az^{-1}$, with a typical value of $a = 0.97$ (Deng and O'Shaughnessy, 2003), to emphasize high frequency regions (Witkowski et al., 2017).

This pre-emphasized speech signal is then passed through a Gabor filterbank to obtain narrowband filtered signals.

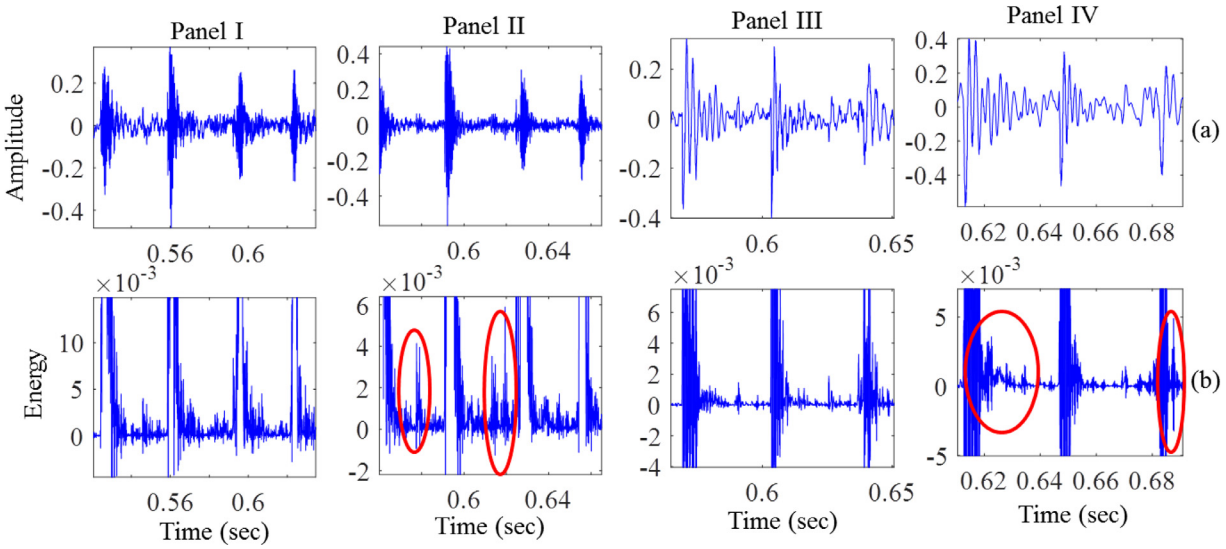

Fig. 7. (a) Time-domain speech segment of replay signal recorded in balcony (Panel I), bedroom (Panel II), canteen (Panel III), and office (Panel IV) along with their corresponding Teager energy profiles (b). Highlighted ovals show the extra impulse-like Teager energy traces observed replay speech recorded in closed room. The center frequency of a filter used in the Gabor filterbank is around 1st formant = 500 Hz to extract the Teager energy profiles.

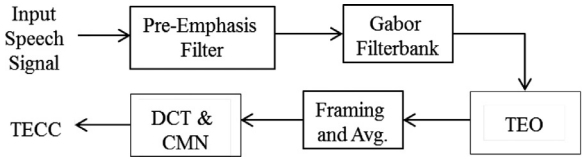

Fig. 8. Functional block diagram of the proposed TECC feature set. After Kamble and Patil (2019).

The Gabor filter is compact, smooth, and also has *optimal* joint time-frequency resolution, and thus, distortions and noise present in distinct locations, time or frequency, do not significantly interfere with the filter responses (Mallat, 1999). The optimal criteria here is to be able to achieve the minimum time-bandwidth product that is dictated by Heisenberg's uncertainty principle in a signal processing framework (Mallat, 1999); in particular, the following statement. The temporal variance σ_t^2 and the frequency variance σ_ω^2 of a signal $f(t) \in L^2(R)$ (Hilbert space of finite energy signals) satisfy

$$\sigma_t^2.\sigma_\omega^2 \geq 1/4. \tag{12}$$

This inequality becomes equality if and only if $f(t)$ is Gaussian, where $\sigma_t^2.\sigma_\omega^2$ is called the time-bandwidth product (which is also the area of the Heisenberg box). Studies in Kamble et al. (2018) and Kamble and Patil (2017) found that the linearly-spaced center frequencies have good resolution in both the lower and higher frequency regions that make it more reliable to compute the spectral information. Hence, the narrowband filtered signals are obtained at center frequency, which are linearly-spaced between f_{min}=10 Hz, and f_{max}=8000 Hz. The impulse response, $h(t)$, of each Gabor filter is given by (Maragos et al., 1991) :

$$h(t) = exp(-b^2t^2)cos(\omega_c t), \tag{13}$$

where ω_c is the center frequency (in Hz) of the filter chosen as per the frequency scales of Equivalent Rectangular Bandwidth (ERB), Mel, and linear. The parameter b controls the bandwidth of a subband filter. The Gabor filterbank has linear phase response characteristics and hence, it maintain the same pattern (shape) of the filtered speech signal (within the passband of each filter) with a delay in time that is equal to the group delay function (in seconds) of the filter (Klapper and Harris, 1959).

The center frequencies for ERB and Mel scale have a number of cut-off frequencies in the lower frequency regions. Motivated by the studies of auditory perception in humans, the center frequencies of the ERB and Mel scales have narrow and wider bandwidth in the lower and higher frequency regions, respectively (Glasberg and Moore, 1990; Stevens et al., 1937). In the case of a linear scale, all the sub-band filters have almost *equal* bandwidth and hence, have good resolution in the lower and higher frequency regions, which makes it more reliable to estimate the spectral information.

Moreover, the original concept of the TEO was developed on a mono-component signal and thus, application of the TEO demands at least the signal under consideration to be bandpass filtered through a narrowband filter. Hence, we have used a linear-spaced filterbank rather than its triangular counterpart, which is known to have wider frequency intervals (due to the Mel warping), at higher frequency regions. The filtered subband signals obtained from the linearly-spaced Gabor filterbank are applied to the TEO block and compute the instantaneous energy of each subband filtered speech signal. Furthermore, these

Table 1
Statistics of ASVspoof 2017 Version 2.0 challenge database. After (Delgado et al., 2018).

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

Teager energy profiles are passed through frame-blocking, and averaged with a short window of 20 ms and with a window shift of 10 ms followed by a logarithm operation to compress the data. To obtain a low-dimensional representation that has compact energy, a Discrete Cosine Transform is applied along with Cepstral Mean Normalization (CMN) (also known as Cepstral Mean Subtraction (CMS)) to reduce the channel mismatch/distortion conditions (Molau et al., 2003). Finally, the retained few Discrete Cosine Transform coefficients, i.e., Teager Energy Cepstral Coefficients (TECC) are appended along with their Δ and $\Delta\Delta$ features to obtain a higher-dimensional feature vector.

4. Experimental setup

4.1. Database

We perform experiments on the ASVspoof 2017 Challenge version 2.0 database, which is mainly based on the RedDots corpus, and its replay speech version (Kinnunen et al., 2017a; Delgado et al., 2018; Lee et al., 2015; Kinnunen et al., 2017b). The version 2.0 database presents in-depth analysis of the replay speech detection performance along with a detailed description of playback and recording devices. All utterances in the database have a sampling rate of 16 kHz, and resolution of 16-bits per sample. The statistics of the database are summarized in Table 1 (Delgado et al., 2018).

4.2. Feature parameterization

The parameters selected for different feature sets depends upon the earlier studies as given below:

Details of CQCC (Delgado et al., 2018; Todisco et al., 2016; 2017): To extract CQCC features, a constant Q transform (CQT) is used that employs variable time-frequency resolution. Furthermore, CQCC features are extracted with $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 8 kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \approx 15Hz$. The number of bins per octave B is set to 96. Features are extracted with 30 DCT static coefficients (with log-energy), resulting in a total 90-D feature vector.

Details of MFCC (Patil et al., 2017; Kamble and Patil, 2019; Kamble et al., 2018; Davis and Mermelstein, 1980): To extract MFCC features, speech signal is passed through the frame-blocking with 20 ms window with 10 ms frame shift. We used 40 triangular Mel subband filters. Features extracted with 13-dimensional static features plus 13-delta and 13-double-delta features to get 39-dimensional feature vector. Similarly, LFCC features are extracted along with difference in frequency scale (i.e., linear scale is used) and the features extracted with 40-dimensional static features plus 40-delta and 40-double-delta features to get 120-dimensional feature vector (Kamble and Patil, 2019; Sahidullah et al., 2015).

From our earlier studies, we found that the TECC feature set performed better with the 40 subband filtered signals with a 120-D feature vector (Kamble and Patil, 2019). If we reduce the number of subband filters or feature dimension, the performance of the SSD system degrades and hence, we choose the parameters as mentioned in the Table 2.

The Gaussian Mixture Model (GMM) classifier is used for modeling the genuine and replay classes with 512 Gaussian components in the GMM. The final scores are represented in terms of Log-Likelihood Ratio (LLR). The decision of the test speech being genuine or replay is based on the scores of LLR, i.e.,

$$LLR = \log \frac{P(X|H_0)}{P(X|H_1)}, \tag{14}$$

where $P(X|H_0)$ and $P(X|H_1)$ are the likelihood scores of genuine and replay trials, respectively. To explore the possible complementary information present in various feature sets, we computed score-level fusion of two feature sets, given by Murty and Yegnanarayana (2005):

Table 2
Details of features extraction parameters.

Parameters	CQCC	LFCC	MFCC	TECC
Freq. Scale	-	Linear	Mel	Linear
Subband Filters	-	40	40	40
Fmin (in Hz)	15	0	0	10
Fmax (in Hz)	8000	8000	8000	8000
Dimension	90	120	39	120

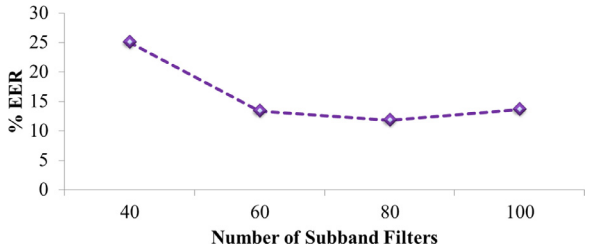

Fig. 9. Results in % EER on development set with varying the number of subband filtered signals in a filterbank.

$$LLK_{fused\_feature} = \alpha LLK_{feature1} + (1-\alpha) LLK_{feature2}, \tag{15}$$

where $LLK_{feature1}$, and $LLK_{feature2}$ are the log-likelihood scores of feature1 and feature2, respectively. The fusion parameter α lies between 0 and 1.

4.3. Performance measures

The ASV system uses a standard evaluation metric, i.e., Equal Error Rate (EER), which is indicated on the Detection Error Trade-off (DET) curve (Martin et al., 1997). The DET curve is used to study the performance of the SSD system. When the operating point in the DET curve of the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) or miss probability is *equal*, then it is referred to as EER.

5. Experimental results

5.1. Results on ASVspoof 2017 V2.0 Database

5.1.1. Results on the development set

This section presents the experiments performed on the development set in order to optimize the parameters on the evaluation set, such as the approximate number of subband filtered signals, and the choice of bandwidth of subband filters in the Gabor filterbank.

Effect of the Number of Subband Filters:

The human auditory system carries several thousands of filters, which results in a dense filterbank in the frequency-domain (Dimitrios et al., 2005; Vijayan et al., 2014; 2016). To compute the Teager energy features accurately, we increased the number of subband filters in the Gabor filterbank. Results with the increase in the number of subband filtered signals are shown in Fig. 9. It can be observed that with 40 subband filters in the filterbank, we obtain a very high EER of 25.07% on the development set. As we increase the number of subband filters in the filterbank, the EER goes on decreasing about 50% from the EER obtained using 40 subband filters. The low EER of 11.82% was obtained with 80 subband filter signals. However, when we further increase the subband filtered signals to 100, the EER increases. The filters overlap with each other and hence, discriminative information is lost that results in a degrading of the SSD performance.

Effect of Bandwidth in Subband Filters:

The formant transitions, in particular for the higher formants, are important when it comes to speaker-related information (namely, a speaker identification or verification task). The higher formants, or the energy present in higher frequency, indeed help to distinguish the replay speech signal from its natural counterpart. The higher spectral energy information depends on the process of how it is extracted, in particular, the frequency scale used in the filterbank, the bandwidth of a subband filter, and the shape of subband filters. The choice of the bandwidth in a subband filter should not be too narrow, nor should it be wider. If the bandwidth of the subband filter is too small then the filtered signal may not capture the formant transition well, whereas if the bandwidth is too large, the features extracted are inaccurate (Vijayan et al., 2016). Hence, after a certain bandwidth, further widening of the bandwidth results in poor frequency resolution and hence, it degrades the performance (Kamble et al., 2018; Vijayan et al., 2014; 2016). Using 80 subband filters as it gave lower EER (discussed in Section 5.1.1), we performed further experiments by varying the bandwidth of a filter from 50 Hz to 400 Hz. The results obtained by varying the bandwidth are shown in Fig. 10. Using 100 Hz bandwidth, we obtained lower EER of 10.80% on the development set compared to other choices of bandwidths.

In addition, we also performed an experiment without integrating the Gabor filterbank, to investigate the importance of the filterbank for the proposed approach. Table 3 shows the results (with and without the filterbank), on the development and evaluation sets for the TECC feature sets. It can be observed that the EER obtained for the TECC feature set gave high EER for both development and evaluation sets. Hence, for the proposed feature sets, the bandpass filtering is a must to estimate the corresponding narrowband component signals. This is primarily due to the fact that the TEO works on mono-component signals (as the TEO

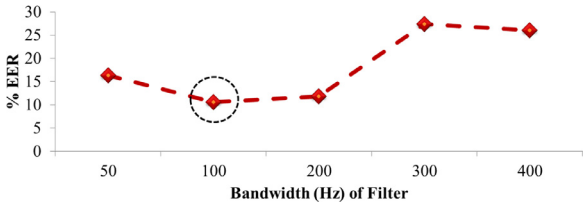

Fig. 10. Results with varying the bandwidth on the development set.

represents a running estimate of signal's energy, which models simple harmonic motion (Kaiser, 1990)), and the fact that speech is inherently a multi-component AM-FM signal and thus, requires bandpass filtering to separate various components.

Furthermore, we performed the experiments by replacing the linearly-spaced gabor filterbank with the linearly-spaced Mel filterbank during the TECC feature extraction process. The EER obtained are shown in Table 4. It can be observed that though the frequency spacing is kept linear for both Gabor and Mel filterbank the shape of the filter do matter during feature extraction process. The EER obtained from Mel filterbank do not perform better compared to the Gabor filterbank.

5.1.2. Results on the evaluation set

Based on the experiment performed on the development set, parameters were optimized based on the development set, and later carried forward to the evaluation set. In particular, 80 subband filters using 100 Hz bandwidth in the filterbank was used to compute the TECC feature set. In addition, we analyzed the effect of EER depending on replay configurations, in particular: change in acoustic environment, playback and recording devices on the evaluation set.

5.1.3. Results using score-level fusion

Table 5 shows the results in EER on the development and evaluation datasets. We compared the TECC feature set with other existing feature sets, namely, CQCC, MFCC, and LFCC. On the ASVspoof 2017 version 2.0 database, the post-evaluation baseline was modified from the earlier baseline in the form of having the log-energy coefficients, and the Cepstral Mean Variance Normalization (CMVN) method; the enhanced baseline results are reported in Table 5. While the post-evaluation CQCC baseline system

Table 3
Results in terms of EER (%) on ASVspoof 2017 database obtained with and without applying filterbank.

Feature Set	Without Filterbank		With Filterbank	
	Dev	Eval	Dev	Eval
TECC	40.94	42.33	**10.80**	**11.41**

Table 4
Results (in % EER) of TECC feature sets using linearly-spaced Gabor filterbank *vs.* linear-spaced Mel filterbank .

Filterbank Parameters	Dev	Eval
Linearly-Spaced Mel Filterbank	24.80	29.49
Linearly-Spaced Gabor Filterbank	**10.80**	**11.41**

Table 5
The final results (in % EER) on the development and evaluation sets.

Feature Set	Dev	Eval
CQCC (Post Eval)	9.06	13.74
CQCC (Our baseline1)	12.81	19.04
MFCC	24.19	26.90
LFCC (Our baseline2)	16.76	13.90
TECC	**10.80**	**11.41**
TECC+CQCC	8.90	11.77
TECC+MFCC	13.13	13.64
TECC+LFCC	8.10	10.49
CQCC+LFCC+MFCC	7.37	12.06
CQCC+LFCC+MFCC+TECC	**6.68**	**10.45**

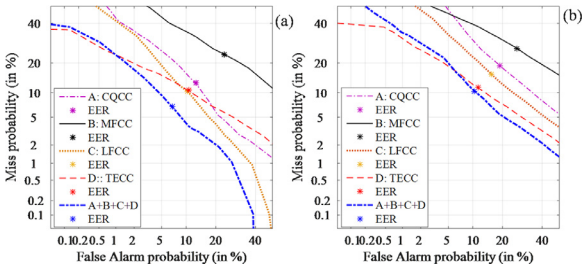

Fig. 11. DET curves for (a) development, and (b) evaluation set of ASVspoof 2017 V2.0 database.

Table 6
Comparison of feature sets on the evaluation set in % EER on different replay configurations (RC).

Feature Set	Acoustic Environment	Playback Device	Recording Device
CQCC	17.85	16.43	18.06
MFCC	26.34	24.15	24.49
LFCC	15.18	14.48	14.85
TECC	**11.41**	**10.42**	**11.20**

includes log-energy coefficients (Delgado et al., 2018), our baseline does not use it. In addition, we also considered LFCC as our second baseline, since the TECC feature set is extracted with the linear frequency scale. The MFCC feature set was also compared as it is one of the state-of-the-art feature sets used in the speech literature. From Table 5, it can be observed that the relatively low EER obtained with the TECC feature set resulted in 10.80% and 11.41% on the development and evaluation set, respectively.

Furthermore, in order to increase the performance of the replay SSD task, we further performed score-level fusion as per Eq. (15) to obtain possible complementary information. The low EER obtained is with score-level fusion of the TECC and LFCC feature sets that resulted in 8.10% and 10.49% EER at fusion weight of $\alpha = 0.7$ on the development and evaluation sets, respectively. In addition, we also fused the scores of all the feature sets used and observed the importance of the TECC feature set. It can be observed from Table 5, with the score-level fusion of the TECC along with CQCC, MFCC, and LFCC, that the performance of replay detection is better compared to the other fusion of the feature set, indicating that the proposed feature set captures *complementary* information better than other feature sets alone or their fusion. The low EERs obtained were 6.68% and 10.45% on the development and evaluation sets, respectively.

The performance evaluation is also shown by the DET curves for the CQCC, MFCC, LFCC, and TECC feature sets along with their best score-level fusion results in Fig. 11. It is observed that the miss probability of CQCC, MFCC, and LFCC is very high for the given FAR, which is not good for an ASV system. There is significant decrease in miss probability fusing the TECC feature set on the development set as shown in Fig. 11 (left side), which is further reduced when the scores are fused with the LFCC feature set. We observe a similar pattern for the development set; the evaluation set is also shown in Fig. 11 (right side). However, the TECC feature set and it's score-level fusion with LFCC has relatively lower FAR compared to the other feature sets.

5.1.4. *Effect of replay configurations (RC)*

The updated ASVspoof 2017 Challenge version 2.0 database provides the detailed description of replay configuration, in particular, acoustic environment, playback, and recording devices (Delgado et al., 2018). There are in total 61 distinct different replay configurations. The replay utterances encompass those of a playback and recording device along with an acoustic environment through which sound propagates (Delgado et al., 2018). On the evaluation set, the EERs with all the feature sets for different replay configurations are shown in Table 6. The overall performance of different replay configurations has low EER using the TECC feature set. Hence, the TECC feature set is able to detect different replay configurations better compared to the other feature sets. Furthermore, we analyzed the individual replay configurations discussed in the next sub-section.

The acoustic environment listed in Delgado et al. (2018) is the actual space in which the original speech data were re-recorded. The ASVspoof 2017 Challenge version 2.0 database has in total 26 different environments, denoted E01-E26. Different

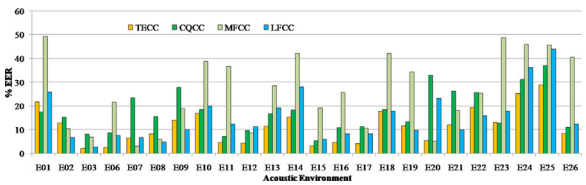

Fig. 12. Individual % EER for different environmental conditions with the CQCC, MFCC, LFCC, and TECC feature sets. After (Delgado et al., 2018).

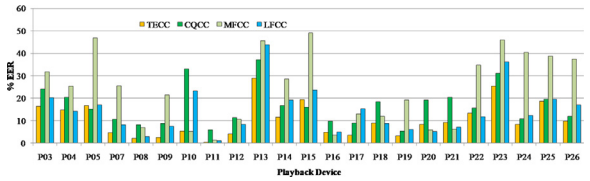

Fig. 13. Individual % EER for different playback devices with the CQCC, MFCC, LFCC, and proposed TECC feature sets. After (Delgado et al., 2018).

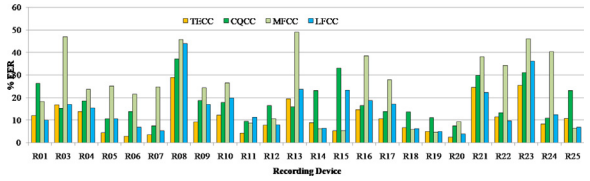

Fig. 14. Individual % EER for different recording devices with the CQCC, MFCC, LFCC, and TECC feature sets. After (Delgado et al., 2018).

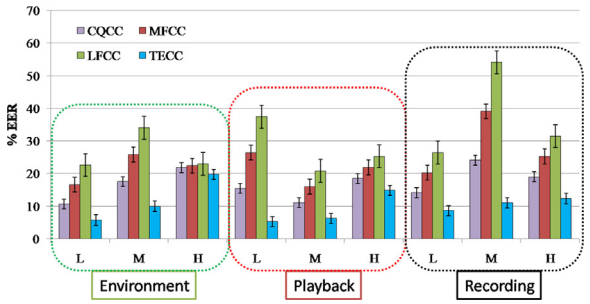

Fig. 15. Different levels of threats, namely, low (L), medium (M), and high (H), for the CQCC, LFCC, MFCC, and TECC feature sets on all replay configurations.

environments have the variations included with the levels of additive ambient, convolution, and reverberation noise. The level of noise in the environment is assumed to be *inversely* proportional to the threat that they pose to the ASV system.

Fig. 12 shows the individual EER for various environmental conditions on the evaluation dataset. We can observe that, using the MFCC and CQCC feature sets, the EER for most of the environments is relatively higher compared to the LFCC and TECC feature sets. However, the TECC feature set shows a lower EER for different environments.

Similar to different acoustical environments, there are 26 different playback devices denoted by P01-P26 (Delgado et al., 2018). The EERs for all the different playback devices using all the feature sets are shown in Fig. 13. Similar to the acoustical environments, the TECC feature set gave relatively lower EER for different playback devices compared to the other feature sets.

There are 25 different recording devices used during collection of replay speech denoted by R01-R25 (Delgado et al., 2018). Fig. 14 shows the EER for different recording devices with all the feature sets on the evaluation dataset. Similar to acoustical environments and playback devices, the pattern of lower EER is observed with the TECC feature set for different recording devices compared to the other feature sets.

The acoustic environment, playback devices, and recording devices are classified into three different levels of threat, namely, low, medium, and high. Fig. 15 shows the EER for different levels of threat using the CQCC, LFCC, MFCC, and TECC feature sets for all different replay configurations. The high-level threats are difficult to detect because professional audio equipment, such as active studio monitors and studio headphones, were used to produce replay samples. In addition, samples collected from studio quality condenser microphones or hand-held recorders are assumed to be of higher quality and hence, give higher EER for high-level threats. As the level of threat continues to increase, the EER also increases. The TECC feature set has lower EER for all levels of threats compared to the other systems.

Finally, we compared the TECC feature set with other systems that were proposed for the replay SSD task on the ASVspoof 2017 Challenge version 2.0 database. Very few studies have been reported on the modified database as listed in Table 7. However, it should be noted that strict comparison of various SSD systems reported in Table 7 from the SSD literature is not possible, primarily due to differences in implementations, data partitions, feature dimensions, ways of computing the EERs, etc.

5.2. Results on the ASVspoof 2015 challenge

The performance of the TECC feature set is also evaluated on other spoofing databases, such as the ASVspoof 2015 Challenge, BTAS 2016, and ASVspoof 2019 Challenge. The ASVspoof 2015 Challenge database was created for the ASV spoofing and countermeasure challenge, and it is comprised of genuine and spoof speech data, in particular, synthetic speech and voice conversion

Table 7
Comparison of results (in % EER) on ASVspoof 2017 version 2.0 Challenge database.

Feature Set	Classifier	Dev	Eval
CQCC (Delgado et al., 2018) (BL)	GMM	9.06	13.74
LFCC	GMM	10.58	16.62
MFCC	GMM	24.19	26.90
PNCC (Tapkir et al., 2018)	GMM	20.78	23.74
QLNCC (Tapkir et al., 2018)	GMM	21.81	24.67
CILPR (Jelil et al., 2018)	GMM	19.68	20.66
PSRMS (Jelil et al., 2018)	GMM	33.38	28.16
eCQCC-DA (Yang et al., 2018)	DNN	13.97	13.38
CQCC (Das and Li, 2018)	GMM	8.93	12.20
IFCC (Das and Li, 2018)	GMM	16.20	15.90
DCTILPR (Das and Li, 2018)	GMM	22.69	14.03
RMFCC (Das and Li, 2018)	GMM	23.58	20.49
Proposed:TECC	GMM	10.80	**11.41**

Table 8
Comparison of results (in % EER) on the ASVspoof 2015 Challenge database. Feature parameters extracted are the same as mentioned in Table 2. After (Kamble et al., 2020).

Feature Set	Development	Evaluation
MFCC (Kamble and Patil, 2017)	6.14	9.15
TECC	0.38	5.95
CQCC (Todisco et al., 2016)	**0.0381**	**0.255**

Table 9
Results in (% EER) for the BTAS 2016 Database. Feature parameters extracted are the same as mentioned in Table 2. After (Kamble et al., 2020).

Subset	Baseline	MFCC	CQCC	TECC
Dev	5.91	3.66	3.05	2.25
Eval	-	7.59	18.86	4.51
Fusion with TECC				
Dev	-	**2.20**	2.25	-
Eval	-	**4.43**	4.51	-

(Wu et al., 2015a). Brief details of the database are given in Wu et al. (2015a,b). The results obtained on the development and evaluation sets of the ASV spoof 2015 Challenge database are reported in Table 8. It can be observed that, on the development set, the TECC feature set results in 0.38% EER (Kamble et al., 2020). However, the best performing feature set, i.e., CQCC, gave a lower EER of 0.038%.

5.3. Results on BTAS 2016

The detailed statistics of the database are given in Korshunov et al. (2016). The organizers of the BTAS 2016 Challenge provided a baseline system that uses the simple spectrogram-based ratio as feature, and logistic regression as classifier. The results on the development and evaluation sets are shown in Table 9. We compared our results with the baseline system, MFCC, and CQCC feature sets. It can be observed that the TECC feature set results in lower EERs of 2.25% and 4.51% on the development and evaluation sets, respectively, compared to the baseline system, MFCC, and CQCC feature sets (Kamble et al., 2020).

We further used score-level fusion of MFCC and CQCC with the TECC feature set to obtain possible complementary information, and reduce the % EER further (as shown in Table 9). The score-level fusion reduced the EER to 2.20% with the MFCC and TECC feature sets (with the fusion factor, α= 0.8 in Eq. 15), and with the CQCC feature set, it reduced to 2.25% (with fusion factor α= 0.9 in Eq. 15). On the other hand, on the evaluation set, the score-level fusion reduced only with the fusion of the MFCC and TECC results in 4.43% EER (with fusion factor α= 0.9 in Eq. 15) whereas with the CQCC feature set, the EER did not reduce.

5.4. Results on the ASVspoof 2019 challenge

The organizers of the ASVspoof 2019 Challenge provided a baseline system for both Logical and Physical Access (LA and PA) tasks (Todisco et al., 2019). We observed that the spectral energy density obtained from the Teager energy-based approach has

Table 10
Comparison of the TECC feature set with other systems for the LA task of the ASVspoof 2019 Challenge database. Feature parameters extracted are the same as mentioned in Table 2.

	Dev		Eval	
Feature Sets	EER	t-DCF	EER	t-DCF
CQCC	0.43	0.0123	9.57	0.2366
LFCC	2.71	0.0663	8.09	0.211
TECC	**0**	**0**	**7.51**	**0.1940**

Table 11
Comparison of the TECC feature set with other systems for the PA task of the ASVspoof 2019 Challenge database. Feature parameters extracted are the same as mentioned in Table 2.

	Dev		Eval	
Feature Sets	EER	t-DCF	EER	t-DCF
CQCC	**9.87**	**0.1953**	11.04	0.2456
LFCC	11.96	0.2554	13.54	0.3017
TECC	24.7	0.62441	43.62	0.8085

high energy across entire frequency regions (because of the linearly-spaced Gabor filterbank) as compared to the spectral energy density obtained from the traditional spectrogram, and MGD spectrum. The baseline system utilizes two feature sets, namely, CQCC and LFCC with 512 mixtures for modeling genuine and corresponding spoof models in GMM. The ASVspoof 2019 Challenge uses the minimum Tandem-Detection Cost Function (t-DCF) as the evaluation metric along with EER (Todisco et al., 2019). Due to the computational load for the available hardware, fewer Gaussians mixtures were used for the TECC feature extraction (i.e., 256 Gaussian mixtures for LA and only 64 Gaussians mixtures for the PA task). From Table 10, it can be observed that the TECC feature set performed better than the baseline systems. Furthermore, it should be that direct performance comparison of TECC for SSD on ASVSpoof 2017 database is not possible primarily due to use of number of 512 Gaussian in GMM (as opposed to just 64 Gaussian on PA task of ASVspoof 2019). The results for the PA task are reported in Table 11. The training set for the PA task contains twice the number of training files that were present in the LA set, which in turn increases the computational load on the hardware and hence, we reduced the number of Gaussian mixture further. From Table 10, it can be observed that the TECC feature set did not perform well for the PA task, though it performed relatively best on the LA task. This is primarily due to the fact that TECC is able to capture differences in real replay signal than the synthetic (simulated) replay signal w.r.t. corresponding bonafide or natural signal. In particular, TEO being known to capture characteristics of natural speech production, brings out more distinct TEO profile for real replay signal. In this context, we observed the differences between the natural, actual replay, and simulated replay speech signal along with its corresponding TEO profiles in Fig. 16.

The comparison is made for the speech signals of ASVspoof 2017 (Panel I and Panel II) and ASVspoof 2019 (Panel III and Panel IV) challenge databases. It can be observed from the TEO profiles that the replay signal from ASVspoof 2017 database preserves the similar pattern of bumps as of natural speech signal. On the other hand, for ASVspoof 2019 replay signal it can be observed that the TEO profiles are distorted because of the simulated noise added in the natural signal, and hence makes SSD system to degrade the performance. This analysis shows that the simulated replay signal do not contain the nonlinearities of speech production as it is preserved for the natural or actual replay signal. Furthermore, we also observe the score distribution pattern for the ASVspoof 2017 and ASVspoof 2019 challenge database as shown in Fig. 17. For the simulated replay, TEO profile gives noise-like signal which we believe distorts the log-likelihood scores of genuine and impostor trials.

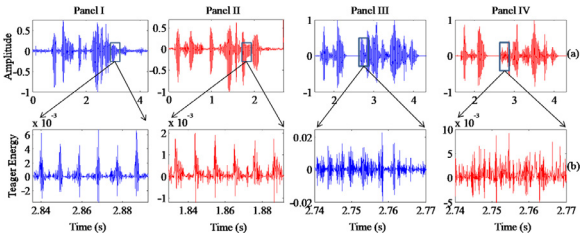

Fig. 16. (a) Time domain speech signals Pane I and Panel II: natural and actual replay signal form ASVspoof 2017 database. Panel III and Panel IV: natural and simulated replay signal form ASVspoof 2019 database. (b) shows the corresponding Teager energy profiles.

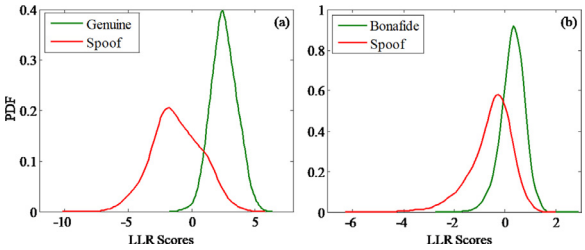

Fig. 17. LLR scores distribution for TECC feature set on (a) ASVspoof 2017, and (b) ASVspoof 2019 corpus.

Table 12
Results in % EER on ASVspoof 2017, ASVspoof 2019, and the Real PA challenge database. Feature parameters extracted are the same as mentioned in Table 2.

Feature	ASVspoof 2017		ASVspoof 2019		Real PA
Sets	Dev	Eval	Dev	Eval	Eval
CQCC	12.81	19.04	9.7	11.04	15.71
TECC	10.80	11.41	24.7	43.62	39.16

In addition, a comparison of the ASVspoof 2017, ASVspoof 2019, and real PA of the ASVspoof 2019 Challenge databases are shown in Table 12. On the ASVspoof 2017 Challenge database, EERs of 10.80% and 11.41% are obtained with the TECC feature set on the development and evaluation sets, respectively. The similar set of features did not perform well on the controlled acoustic environment, i.e., for the ASVspoof 2019 Challenge database, it results in 24.7% and 43.62% EER on the development and evaluation sets, respectively. The absolute difference on the development set for both replay databases is approximately 15%, which is a huge difference for the SSD task. The performance of the SSD system degrades in the case of the ASVspoof 2019 Challenge database, as this database is simulated and has controlled acoustical conditions. This proves that the same feature set does not work for different acoustical conditions, and hence, there is a need for more generalized features for the SSD task. Furthermore, when the experiments were performed on the real PA database of ASVspoof 2019, the EER reduced from 43.62% to 39.16%.

6. Summary and conclusions

This study analyzed the effect of reverberation using the TEO for the replay SSD task. We first studied the basic mechanism of the replay signal, the factors involved during the replay process, the mathematical model of reverberation, and relevant discussion of the TEO. For the replay speech signal, the concept of reverberation was also analyzed and we observed the delay and change in amplitude components for the replay speech signal. The delay and change of amplitude in the replay speech signal occur because of the reverberation. The reverberated signal is also affected by material present in the recording environment, the shape and size of the room, and the sound absorbing property of the material kept in the room.

In order to emphasize high-frequency regions, features were extracted using a pre-emphasis filter. The frequency resolution of the linearly-spaced Gabor filterbank is explicitly related to the number of subband filters in filterbank. By increasing the number of subband filters, the frequency resolution is improved, and thus it captures more detailed spectral characteristics. Furthermore, we used the energy of the subband filtered speech signal estimated via the TEO to calculate the EER of individual replay configurations. The TECC feature set gave lower EER for all replay configurations and also for different levels of threats for the ASV system.

We further compared the performance of the TECC feature set with state-of-the-art feature sets, namely, CQCC, LFCC, and MFCC. The experiments demonstrated on the ASVspoof 2017 Challenge version 2.0 database with the TECC feature set gave better results than the other systems. The TECC feature set when fused at the score-level fusion with the CQCC, MFCC, and LFCC feature set improved the system performance compared to that for individual feature sets. Furthermore, on the evaluation set, we investigated the replay configurations, namely, acoustic environment, playback and recording devices with the proposed feature set. We observe that the TECC feature set gave lower EER for all the different conditions of threat compared to the other system. For the high-level threat and high quality devices used during playback and recording, the EERs are quite high. This needs further investigation to detect the high level replay configuration threat.

In addition, we also performed the experiments on the other state-of-the-art spoofing databases to analyze the performance of the proposed TECC feature set, namely, ASVspoof 2015 Challenge, BTAS 2016, and ASVspoof 2019 Challenge database. As the proposed TECC feature set captures the characteristics of natural speech production, the performance varies depending on the type of database used to evaluate the replay SSD system performance. In particular, the performance for machine-generated speech signals (such as SS and VC) gave better results when compared to the recorded (replay) speech signals. The TECC feature set gave better performance on the VC and SS spoof signals as observed for the ASVspoof 2015 and ASVspoof 2019 LA task. Also, for the real replay spoof signals (of ASVspoof 2017 version 2 database), TECC performed better compared to the other feature sets. However, it did not perform well for the synthetic generated replay signals (i.e., ASVspoof 2019 PA task) as TEO has the

capability to capture the nonlinear aspects of natural speech production (Maragos et al., 1991; 1992; Quatieri, 2006; Teager and Teager, 1990). For synthetic replay speech signals, these nonlinear aspects of speech production is not present and hence, TECC fails to perform for these kind of spoof signals.

In the future, we plan to perform a detailed representation and investigation of reverberation effects on the replay signals that are collected using high quality recording and playback devices. The amount of reverberation might be even more in some of the bonafide far-field samples compared to near-field high quality replay speech and thus directly relying on the amount of reverberation to test the replay spoof may be a little risky. In this context, a more detailed study and generalized countermeasure is required to overcome the replay detection task. Furthermore, we plan to investigate the effect of reverberation on AM-FM components and its relation to the proposed TEO framework for the different acoustic environments and intermediate device conditions in the replay speech. Since this study is solely based on considering reverberation as a key factor to distinguish natural *vs.* replay spoof, our method is not likely to produce good results for replay recorded in outdoor environments. In addition, our approach ignores the fact bonafide utterances might also contains reverberant noises, such as in smart speakers.

Declaration of Competing Interest

There are no interests to declare.

Acknowledgments

The authors would like to thank the organizers of the special issue on Advances in Automatic Speaker Verification Anti-spoofing. In addition, they also thank authorities of DA-IICT Gandhinagar for their kind support to carry out this research work and University Grants Commission (UGC) for providing Rajiv Gandhi National Fellowship (RGNF). We would like to sincerely thank Prof. (Dr.) Douglas O'Shaughnessy (PhD, MIT, USA) (IEEE Fellow, ISCA Fellow, ASA Fellow) for his valuable suggestions to improve the presentation and language-related corrections.

References

Acharya, R., Patil, H.A., Harsh, K., 2019. Novel enhanced Teager energy based cepstral coefficients for replay spoof detection. IEEE ASRU. Sentosa, Singapore, pp. 342–349.
Alegre, F., Janicki, A., Evans, N., 2014. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. IEEE International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany, pp. 1–6.
Alluri, K.R., et al., 2017. SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017. INTERSPEECH. Stockholm, Sweden, pp. 107–111.
Arroabarren, I., Rodet, X., Carlosena, A., 2006. On the measurement of the instantaneous frequency and amplitude of partials in vocal vibrato. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1413–1421.
Blesser, B., Salter, L.-R., 2009. Spaces Speak, are you Listening?: Experiencing Aural architecture. MIT Press.
Boashash, B., 1991. Time-Frequency Signal Analysis, second ed. Prentice Hall.
Chen, Z., Xie, Z., Zhang, W., Xu, X., 2017. ResNet and model fusion for automatic spoofing detection. INTERSPEECH 2017. Stockholm, Sweden, pp. 102–106.
Cohen, L., 1995. Time-Frequency Analysis, first ed. 778. Prentice Hall PTR Englewood Cliffs, NJ.
Connie, T., Teoh, A., Goh, M., Ngo, D., 2005. Palmhashing: a novel approach for cancelable biometrics. Inf. Process. Lett. 93 (1), 1–5.
Das, R.K., Li, H., 2018. Instantaneous phase and excitation source features for detection of replay attacks. V Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Hawaii, USA, pp. 1030–1037.
Daugman, J., 2003. The importance of being random: statistical principles of iris recognition. Pattern Recognit. 36 (2), 279–291.
Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Proc. 28 (4), 357–366.
Delgado, H., Todisco, M., Sahidullah, M., Evans, N., Kinnunen, T., Lee, K.A., Yamagishi, J., 2018. ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements. Odyssey The Speaker and Language Recognition Workshop. Les Sables d'Olonne, France, pp. 296–303.
Deng, L., O'Shaughnessy, D., 2003. Speech Proc. – A Dynamic and Optimization-Oriented Approach, first ed. Marcel Dekker Inc..
Dimitrios, D., Petros, M., Alexandros, P., 2005. Auditory Teager energy cepstrum coefficients for robust speech recognition. INTERSPEECH. Lisboa, Portugal, pp. 3013–3016.
Evans, N.W., Kinnunen, T., Yamagishi, J., 2013. Spoofing and countermeasures for automatic speaker verification. INTERSPEECH. Lyon, France, pp. 925–929.
Font, R., Espín, J.M., Cano, M.J., 2017. Experimental analysis of features for replay attack detection results on the ASVspoof 2017 Challenge. INTERSPEECH. Stockholm, Sweden, pp. 7–11.
Glasberg, B.R., Moore, B.C., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47 (1), 103–138.
Houtgast, T., Steeneken, H.J., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. 77 (3), 1069–1077.
Jabloun, F., Cetin, A.E., 1999. The Teager energy based feature parameters for robust speech recognition in car noise. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 1, pp. 273–276.Phoenix, AZ, USA
Jain, A.K., Li, S.Z., 2011. Handbook of Face Recognition. Springer.
Jain, A.K., Nandakumar, K., Ross, A., 2016. 50 years of biometric research: accomplishments, challenges, and opportunities. Pattern Recognit. Lett. 79, 80–105.
Jain, A.K., Ross, A., Pankanti, S., 2006. Biometrics: A tool for information security. IEEE Trans. Inf. ForensicsSecur. 1 (2), 125–143.
Jelil, S., Das, R.K., Prasanna, S.M., Sinha, R., 2017. Spoof detection using source, instantaneous frequency and cepstral features. INTERSPEECH. Stockholm, Sweden, pp. 22–26.
Jelil, S., Kalita, S., Prasanna, S.M., Sinha, R., 2018. Exploration of compressed ILPR features for replay attack detection. INTERSPEECH. Hyderabad, India, pp. 631–635 .
Kaiser, J.F., 1990. On a simple algorithm to calculate the energy of a signal. IEEE ICASSP. Albuquerque, New Mexico, USA, pp. 381–384.
Kamble, M., Patil, H., 2018. Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection. INTERSPEECH. Hyderabad, India, pp. 646–650.
Kamble, M., Tak, H., Patil, H., 2018. Effectiveness of speech demodulation-based features for replay detection. INTERSPEECH. Hyderabad, India, pp. 641–645.
Kamble, M.R., Patil, H.A., 2017. Novel energy separation based instantaneous frequency features for spoof speech detection. IEEE EUSIPCO. Kos Island, Greece, pp. 106–110.

Kamble, M.R., Patil, H.A., 2019. Analysis of reverberation via Teager energy features for replay spoof speech detection. IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP). Brighton, UK, pp. 2607–2611.
Kamble, M.R., Patil, H.A., 2017. Effectiveness of Mel Scale-Based ESA-IFCC Features for Classification of Natural vs. Spoofed Speech, Kolkata, India, December 17-20. Springer, pp. 308–316.
Kamble, M.R., Pulikonda, A.K.S., Maddala, V.S.K., Patil, H.A., 2020. Analysis of Teager energy profiles for spoof speech detection. Speaker Odyssey. Tokyo, Japan, pp. 304–311.
Kamble, M.R., Sai, P.A.K., Krishna, M.V.S., Patil, A.T., Acharya, R., Patil, H.A., 2019. Speech demodulation-based techniques for replay and presentation attack detection. IEEE Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC). Lanzhou, China, pp. 1545–1550.
Kim, C., Stern, R.M., 2016. Power-normalized cepstral coefficients PNCC for robust speech recognition. IEEE/ACM IEEE Trans. Audio Speech Lang. Process. 24 (7), 1315–1329.
Kinnunen, T., et al., 2017. The ASVspoof 2017 Challenge: assessing the limits of replay spoofing attack detection. INTERSPEECH. Stockholm, Sweden, pp. 1–6.
Kinnunen, T., et al., 2017. Reddots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research. IEEE ICASSP. New Orleans, Louisiana, USA, pp. 5395–5399.
Kinoshita, K., et al., 2016. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP J. Adv. Signal Process. 1, 1–19.
Klapper, J., Harris, C., 1959. On the response and approximation of Gaussian filters. IEEE IRE Trans. Audio 7 (3), 80–87.
Koppell, J., 2011. International organization for standardization. Handbook of Transnational Governance: Institutions and Innovations, 41, p. 289.
Korshunov, P., Marcel, S., Muckenhirn, H., Gonçalves, A., Mello, A.S., Violato, R.V., Simões, F.O., Neto, M.U., de Assis Angeloni, M., Stuchi, J.A., et al., 2016. Overview of BTAS 2016 speaker anti-spoofing competition. IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). Niagara Falls, New York, USA, pp. 1–6.
Kuc, R., 1988. Introduction to Digital Signal Processing, first ed. McGraw-Hill, Inc..
Kuttruff, H., 2016. Room Acoustics, first ed. CRC Press.
Lau, Y.W., Wagner, M., Tran, D., 2004. Vulnerability of speaker verification to voice mimicking. IEEE International Symposium on Intelligent Multimedia, Video and Speech Proc.. Hong Kong, pp. 145–148.
Lavrentyeva, G., et al., 2017. Audio replay attack detection with deep learning frameworks. INTERSPEECH. Stockholm, Sweden, pp. 82–86.
Lee, K.A., et al., 2015. The RedDots data collection for speaker recognition. INTERSPEECH. Dresden, Germany, pp. 2996–3000.
de Lima, A.A., Freeland, F.P., Esquef, P.A., Biscainho, L.W., Bispo, B.C., de Jesus, R.A., Netto, S.L., Schafer, R.W., Said, A., Lee, B., et al., 2008. Reverberation assessment in audioband speech signals for telepresence systems. SIGMAP. Porto, Portugal, pp. 257–262.
Mallat, S., 1999. A Wavelet Tour of Signal Proc., third ed. Academic press.
Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S., 2009. Handbook of Fingerprint Recognition. Springer Science & Business Media.
Maragos, P., Quatieri, T.F., Kaiser, J.F., 1992. On separating amplitude from frequency modulations using energy operators. IEEE ICASSP, 2, pp. 1–4.San Francisco, California, USA
Maragos, P., et al., 1991. Speech nonlinearities, modulations, and energy operators. IEEE ICASSP. Toronto, Canada, pp. 421–424.
Martin, A., et al., 1997. The DET curve in assessment of decision task performance. EUROSPEECH. Rhodes, Greece, pp. 1895–1898.
Molau, S., Hilger, F., Ney, H., 2003. Feature space normalization in adverse acoustic conditions. IEEE ICASSP. Hong Kong, China.I–656–659–I
Murty, K.S.R., Yegnanarayana, B., 2005. Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process. Lett. 13 (1), 52–55.
Pardede, H.F., 2015. On noise robust feature for speech recognition based on power function family. IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). Bali, Indonesia, pp. 386–390.
Patil, H.A., Kamble, M.R., Hawaii, USA, 2018. A survey on replay attack detection for automatic speaker verification (ASV) system. IEEE Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC), pp. 1047–1053.
Patil, H.A., Kamble, M.R., et al., 2017. Novel variable length Teager energy separation based instantaneous frequency features for replay detection. INTERSPEECH. Stockholm, Sweden, pp. 12–16.
Patil, H.A., Parhi, K.K., 2010. Development of TEO phase for speaker recognition. IEEE International Conference on Signal Processing and Communications (SPCOM), pp. 1–5.
Quatieri, T.F., 2006. Discrete-Time Speech Signal Proc.: Principles and Practice, first ed. Pearson Education India.
Rafi, B.S.M., Murty, K.S.R., Nayak, S., 2017. A new approach for robust replay spoof detection in ASV systems. IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, Montreal, Canada, pp. 51–55.
Replay attack anti-spoofing measures for ASV systems, https://apsipa2018.org/Papers/PublicSessionIndex3.asp?Sessionid=1050. Last Accessed: 2018-11-27.
Reverberation, https://byjus.com/physics/reverberation/. {Last Accessed: 2018-10-24}.
Rosenberg, A.E., 1976. Automatic speaker verification: a review. Proc. IEEE 64 (4), 475–487.
Sahidullah, M., Kinnunen, T., Hanilçi, C., 2015. A comparison of features for synthetic speech detection. INTERSPEECH. Dresden, Germany, pp. 2087–2091.
Sailor, H.B., Patil, H.A., 2017. Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition. J. Acoust. Soc. Am. 141 (6), EL500–EL506.
Sanchez-Reillo, R., Sanchez-Avila, C., Gonzalez-Marcos, A., 2000. Biometric identification through hand geometry measurements. IEEE Trans. Pattern Anal. Mach. Intell. (10), 1168–1171.
Sriskandaraja, K., Suthokumar, G., Sethu, V., Ambikairajah, E., 2017. Investigating the use of scattering coefficients for replay attack detection. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, Kuala Lampur, Malaysia, pp. 1195–1198.
Stevens, S.S., Volkmann, J., Newman, E.B., 1937. A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. 8 (3), 185–190.
Stylianou, Y., 2009. Voice transformation: A survey. IEEE International Conference on Acoustics, Speech and Signal Proc., (ICASSP). Taipei, Taiwan, China, pp. 3585–3588.
Suthokumar, G., Sriskandaraja, K., Sethu, V., Wijenayake, C., Ambikairajah, E., Li, H., 2018. Use of claimed speaker models for replay detection. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Hawaii, USA, pp. 1038–1046.
Tak, H., Patil, H., 2018. Novel linear frequency residual cepstral features for replay attack detection. INTERSPEECH. Hyderabad, India, pp. 726–730.
Tapkir, P.A., Kamble, M.R., Patil, H.A., 2018. Replay spoof detection using power function based features. IEEE Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA-ASC) Hawaii, USA, pp. 1019–1023.
Teager, H., 1980. Some observations on oral airflow during phonation. IEEE Trans. Acoust. Speech Signal Proc. 28 (5), 599–601.
Teager, H., Teager, S., 1990. Evidence for nonlinear sound production mechanisms in the vocal tract. Speech Production and Speech Modelling. Springer, pp. 241–261.
Todisco, M., Delgado, H., Evans, N., 2016. A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. Speaker Odyssey Workshop, 25, pp. 249–252.Bilbao, Spain
Todisco, M., Delgado, H., Evans, N., 2017. Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification. Computer Speech & Lang. 45, 516–535.
Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T.H., Lee, K.A., 2019. ASVspoof 2019: future horizons in spoofed and fake audio detection. INTERSPEECH. Graz, Austria, pp. 1008–1012.
Traer, J., McDermott, J.H., 2016. Statistics of natural reverberation enable perceptual separation of sound and space. Proc. Natl. Acad. Sci. 113 (48), E7856–E7865.
Vijayan, K., Kumar, V., Murty, K.S.R., 2014. Feature extraction from analytic phase of speech signals for speaker verification. INTERSPEECH, pp. 1658–1662.
Vijayan, K., Reddy, P.R., Murty, K.S.R., 2016. Significance of analytic phase of speech signals in speaker verification. Speech Comm. 81, 54–71.
Wen, J., 2009. Reverberation: Models, Estimation and Application. Imperial College London.

Witkowski, M., Kacprzak, S., Zelasko, P., Kowalczyk, K., Gałka, J., 2017. Audio replay attack detection using high-frequency features. INTERSPEECH. Stockholm, Sweden, pp. 27–31.
Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2015. Spoofing and countermeasures for speaker verification: a survey. Speech Comm. 66, 130–153.
Wu, Z., et al., 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. INTERSPEECH. Dresden, Germany, pp. 2037–2041.
Yang, J., Das, R.K., Li, H., 2018. Extended constant-Q cepstral coefficients for detection of spoofing attacks. IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Hawaii, USA, pp. 1024–1029.
Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., Kellermann, W., 2012. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. 29 (6), 114–126.
Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Comm. 51 (11), 1039–1064.