Expert Systems With Applications 198 (2022) 116770

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Voice spoofing detector: A unified anti-spoofing framework

Ali Javed [a,c], Khalid Mahmood Malik [a,*], Hafiz Malik [b], Aun Irtaza [b,d]

[a] Department of Computer Science and Engineering, Oakland University, Rochester, 48309-4479, MI, USA
[b] Department of Electrical and Computer Engineering, University of Michigan-Dearborn, 4901 Evergreen Road, Dearborn, 48128, MI, USA
[c] Department of Software Engineering, University of Engineering and Technology-Taxila, 47050, Punjab, Pakistan
[d] Department of Computer Science, University of Engineering and Technology-Taxila, 47050, Punjab, Pakistan

ARTICLE INFO

Keywords:
Acoustic ternary co-occurrence patterns
AI for multimedia security
AI for voice-based biometrics in IoT
Anti-spoofing against multiple attack vectors
Deepfakes
Voice spoofing detection

ABSTRACT

Voice controlled systems (VCS) in Internet of Things (IoT), speaker verification systems, voice-based biometrics, and other voice-assistant-enabled systems are vulnerable to different spoofing attacks i.e., replay, cloning, cloned-replay, etc. VCS are not only susceptible to these attacks in a non-network environment, but they are also vulnerable to multi-order spoofing attacks in networked IoT. Additionally, deepfakes with artificially generated audio pose a great threat to the all systems having voice-interfaces. Most of the existing counter-measures against these voice spoofing attacks work for only one specific attack (e.g. voice replay) and fail to generalize this for other classes of spoofing attacks. Additionally, generalization is also crucial for cross-corpora evaluation. Thus, there exists a need to develop a unified voice anti-spoofing framework capable of detecting multiple spoofing attacks. This work presents a unified anti-spoofing framework that uses novel (ATCoP-GTCC) features to combat the variety of voice spoofing attacks. The proposed novel acoustic-ternary co-occurrence patterns (ATCoP) encode the co-occurrence of similar patterns between the center and neighboring samples. Our experiments demonstrate that ATCoP can better capture the microphone induced distortions in replays, unnatural prosody and algorithmic artifacts in cloned samples, and both the distortions and artifacts in cloned-replays including compression on multi-hop attacks in the spoofing samples. The performance of ATCoP could be further enhanced by the Gammatone cepstral coefficients. To evaluate the effectiveness of the proposed anti-spoofing system for multi-order replay and cloned-replay attacks detection, we created a diverse voice spoofing detection corpus (VSDC) containing multi-order replay and cloned-replay audios against the bonafide and cloned audio recordings, respectively. Experimental results obtained on VSDC, ASVspoof 2019, Google's LJ Speech, and YouTube deepfakes datasets illustrate the effectiveness of the proposed system in terms of accurate detection for a variety of voice spoofing attacks.

1. Introduction

Smart Speakers (SS), such as Google Home, Alexa, etc., that manage various Voice Controlled Systems (VCSs) of Internet of Things (IoT) and other voice assistants (e.g. Siri, Cortana, bixbi) are expected to transform our homes, businesses, and vehicles to smart ones due to the advancement of voice recognition system, high accuracy of knowledge-driven question answering engines, and integration of smart speakers with various cyber–physical/intelligent systems. Additionally, automatic speaker verification (ASV) technology has progressed in recent years and its applications are growing in diverse real-world authentication scenarios involving both the logical and physical access (Sahidullah et al., 2019).

In recent years, we have witnessed a tremendous evolution in voice biometrics from a basic security feature to be an enabler for remote communications (Hrabi, 2020). Artificial Intelligence (AI)-enabled secure emerging applications use voice biometrics for access control (e.g. physical facilities), voice controlled systems in IoT at home and office setup (Malik, Malik, & Baumann, 2019), transaction authentication (e.g. toll fraud prevention, bank wire transfers), monitoring (e.g. remote time and attendance logging), information retrieval (e.g. customer information for call centers, forensics (e.g voice sample matching), and so on. Since voice as an authentication mechanism in biometrics security has less potential to spread infections compared to other contemporary authentication methods (e.g. face recognition, finger printing, password entry using keyboard), deployment of ASV and VCS during the COVID-19 pandemic is expected to rise in future generation expert systems. However, VCS and ASV systems pose significant

* Corresponding author.
E-mail addresses: ali.javed@uettaxila.edu.pk (A. Javed), mahmood@oakland.edu (K.M. Malik), hafiz@umich.edu (H. Malik), aun.irtaza@uettaxila.edu.pk (A. Irtaza).

https://doi.org/10.1016/j.eswa.2022.116770
Received 15 August 2021; Received in revised form 4 January 2022; Accepted 25 February 2022
Available online 9 March 2022

security and privacy threats as they may be vulnerable to various voice presentation attacks e.g. replay, cloning, voice conversion, etc. (Malik et al., 2019; Sahidullah et al., 2019). In the near future, these threats are expected to rise due to proliferation of smart speakers and VCS, integration of ASV systems in various online and physical access scenarios, and ease of voice attack generation on them. For example, voice replay attacks can be generated easily because of the access of high-quality recording devices and non-requirement of technical skills (Sahidullah et al., 2019). Likewise, the availability of modern-day tools like Tensorflow or Keras, publicly-available trained models such as WaveNet (Mwiti, 2019), and low-cost computing machines is easing the creation of AI-synthesized speech (a type of deepfake), also known as cloned voice. Voice cloning is becoming a vital component of deepfakes where a source speaker's voice is also cloned besides the video. These deepfakes have immense potential to destroy public trust and empower criminals to exploit business deals or family phone calls. Recently one case has been reported where the robbers used the synthetic voice of a company executive's speech to convince their employees into transferring a massive amount to a confidential account (Harvel, 2019). Therefore, unlike existing approaches like Agarwal et al. (2019) that focus on visual forgeries detection only, audio forgeries should also be detected.

VCSs in IoT are more vulnerable to voice-based spoofing attacks compared to traditional devices with voice interfaces. We have demonstrated that various smart speakers, particularly Amazon smart devices with drop-in feature (Metz, 2019), and VCS are not only vulnerable to replay attacks in non-network environment but are also susceptible to multi-order replay attacks (Malik et al., 2019). An example of a multi-order replay attack is shown in Fig. 1(a) where an intruder uses his phone to play the recorded speech "Alexa, turn off the heat" (first-order replay) on the baby monitor by hacking the wireless LAN using tools such as Aircrack-Ng (2020). Next, this speech is replayed (second-order replay) to the SS of targeted person's home to switch off the heat. Secondly, our analysis shows that VCS in IoT domain are prone to voice cloning attacks, and we emphasize that the speech cloning attacks will be more destructive in IoT environment when intruders will combine their social engineering skills in the process of generating them. Shown in Fig. 1(b) is an example of a voice cloning attack on VCS where a cloned speech is played on VCS through the SS to open the garage door. Thirdly, our experimental analysis confirms that VCS in IoT settings are also prone to a hybrid of cloned and replay attacks—cloned-replay attacks. Shown in Fig. 1(c) is an example of a cloned-replay attack on VCS where a cloned speech is replayed on SS-2 via SS-1 (1st-order cloned-replay attack). Later, this 1st-order cloned-replay is replayed on SS3 via SS-2 to generate the 2nd-order cloned-replay attack that is then used to open the garage door.

Most of the research has focused on developing robust detectors to detect either voice replay or cloned voice attacks on ASV (Nagarsheth, Khoury, Patil, & Garland, 2017; Witkowski, Kacprzak, Zelasko, Kowalczyk, & Galka, 2017). These existing binary-class-based (Bonafide vs Spoofed) detectors are not ready to fully combat the emerging threat of different multiple attacks on ASV systems. For example, results of recent work show that spoofing detectors trained with a certain group of spoofing attacks fail to generalize better for other groups of spoofing attacks (Gonçalves, Violato, Korshunov, Marcel, & Simoes, 2017; Korshunov & Marcel, 2016). In other words, anti-spoofing systems trained with voice cloning based spoofed speech often offer a degraded performance for replay detection (Paul, Sahidullah, & Saha, 2017). Additionally, no effort has been made to address the replay or cloning attacks in multi-hop/multi-vector attack scenarios where multiple smart speakers and microphones are chained/linked together (Fig. 1). Therefore, there exists a strong need to develop a unified anti-spoofing system to reliably detect the replay, cloning and cloned-replay attacks in multi-hop scenario. Unlike traditional binary class detectors, our framework models this task as a multi-class problem because there exists a probability that one SS is robust against replay attacks, receives data from other SS (of different vendor) in a chained scenario that is vulnerable to replay attacks because of a fragile or absent replay detector. Therefore, the received audio will be considered bonafide, and the detector will eventually fail for all the linked devices.

To address this need, we present a unified anti-spoofing framework that can effectively be used to detect multiple categories of voice spoofing attacks (i.e. multi-order replays, multi-order cloned-replays, and cloning) using our novel acoustic ternary co-occurrence patterns (ATCoP) and gammatone cepstral coefficients (GTCC) features. It is important to mention that the human speech contains dynamic attributes due to speaker induced variations, whereas, the synthetic speech contains unusual prosody i.e., absence of natural pauses, lack of unvoiced consonants, unusual pitch, and few mispronunciations, etc. These unnatural prosody in cloned voice and speaker induced variations in bonafide speech demands to develop those features which can analyze these patterns. Thus, we propose time-domain ATCoP features that are capable of analyzing and better capturing those distinctive traits of the bonafide and cloned speech. Further, replay audios include the microphone induced distortions and cloned audios include the artificial 'whine' which can be reliably captured by both the ATCoP and GTCC due to their tolerance against the noise. Thus, we fused the ATCoP with the GTCCs to create a robust feature descriptor for voice anti-spoofing system. The major contributions of our work are:

- We propose a novel acoustic feature descriptor ATCoP to better capture the microphone induced distortions (also known as microphone signature) from the replay samples, dynamic speech variations of bonafide signals and artifacts of cloning algorithms.
- We report that VCS are vulnerable to a hybrid voice spoofing attack i.e., cloned-replay which can be generated by playing the synthetic/cloned audio.
- We present that multi-order replay and cloned-replay attacks are feasible and VCSs are unable to detect them.
- We present the baseline for a unified anti-spoofing framework that is able to detect the multi-order replay-, cloning-, and cloned-replay attacks through our ATCoP-GTCC descriptor.
- Our anti-spoofing method effectively detects the voice spoofing attacks in compressed audio samples along-with the uncompressed audios.
- We have performed rigorous experimentation on four different datasets including the hybrid dataset to signify the effectiveness of our anti-spoofing framework.

2. Related work

VCSs need a unified anti-spoofing framework to counter multiple voice spoofing attacks. The selection of features for audio signal representation is an important step in developing this unified framework. Additionally, none of the existing anti-spoofing methods have considered cloned-replay attacks. This section presents a thorough analysis of existing up-to-date spoofing detection systems.

2.1. Replay spoofing detection techniques

Existing approaches for replay spoofing detection have explored different features using either conventional machine learning classifiers i.e. Gaussian Mixture Model (GMM) or deep learning models like CNN, RNN, etc.

2.1.1. Conventional machine learning (ML) classifiers-based approaches

In Yamagishi et al. (2019), two ASVspoof baseline models based on constant Q-transform cepstral coefficients (CQCC) and linear frequency cepstral coefficients (LFCC) were presented with the GMM classifier for spoofing detection including the replays. In Kumar and Bharathi (2021), a filtering based cepstral coefficients (FBCC) based on the discrete cosine transform of log compressed energy variations of the

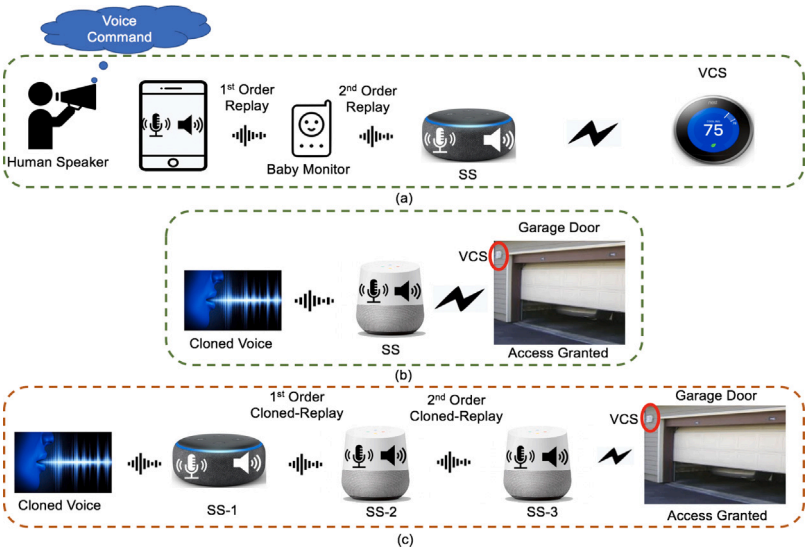

Fig. 1. Examples of audio spoofing attacks.

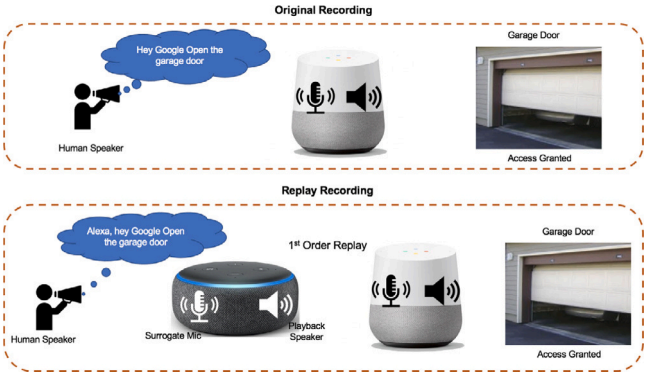

Fig. 2. 1st-order voice spoofing (replay) attack.

audios were employed with the GMM for spoofing detection including the replay attacks.

Few techniques (Nagarsheth et al., 2017; Witkowski et al., 2017) have reported the importance of high-frequency bands analysis to better capture the attributes available in the replay audios. In Nagarsheth et al. (2017), high-frequency cepstral coefficients and CQCC features were employed to generate the embeddings using a deep neural network. Later, these embeddings were used to train the SVM for replay detection. Witkowski et al. (2017) have employed the inverted-MFCC (IMFCC), linear predictive cepstral coefficients (LPCC), LPCCres, CQCC, MFCC, and Cepstrum features to train the GMM for replay detection.

Existing studies (Mishra, Singh, & Pati, 2018; Saranya, Padmanabhan, & Murthy, 2018; Yang & Das, 2019) have also highlighted that reverberation, channel information, recording and playback device characteristics should be investigated for replay spoofing detection. In Saranya et al. (2018), MFCC, CQCC, and Mel-Filterbank-Slope features were employed with GMM to capture the characteristics of channel and reverberation from the audio for replay detection.

2.1.2. Deep learning-based techniques

Deep learning (DL) techniques have also been employed for spoofing detectors apart from the conventional ML based methods. In Cai, Cai, Liu, Li, and Li (2017), the original spectrogram was used instead of CQCC to train a deep residual network for features extraction. This method is taxing due to manual data augmentation and achieves higher equal error rate (EER) due to using only the short time Fourier transform based spectrogram. MFCC and CQCC were employed in Chen, Xie, Zhang, and Xu (2017) with the GMM, DNN and ResNet for replay spoofing detection. Fusion of CQCC-GMM, CQCC-ResNet, and MFCC-ResNet obtained the minimum EER. Fusion of the two deep networks and GMM makes it less practical to deploy on resource constraint VCSs. In Bakar and Hanilçi (2018), long term average spectrum (LTAS) and MFCC features were employed to train the DNN for spoofing detection. Light-weight CNN was employed for audio spoofing detection in Lavrentyeva et al. (2017) and Lavrentyeva et al. (2019). In Monteiro, Alam, and Falk (2020), an end-to-end LCNN ensemble model was proposed based on training a model on the predictions of two separately trained models for replay and cloning attacks respectively. Although this method (Monteiro et al., 2020) outperforms the ASVspoof baseline model (Yamagishi et al., 2019), but with increased features computation cost.

2.2. Voice cloning detection approaches

Existing approaches have employed various magnitude- and phase-oriented features for synthetic/cloned speech detection.

2.2.1. Phase-oriented approaches

In De Leon, Pucher, Yamagishi, Hernaez, and Saratxaga (2012), relative phase shift (RPS) features were extracted from the speech segments of the audio signal and used with the GMM for speech synthesis detection. Similarly, RPS was used with the GMM for synthetic speech detection in Saratxaga, Sanchez, Wu, Hernaez, and Navas (2016). In Janicki (2017), long term prediction residual signals comprised of 23 different parameters were used with the SVM to classify the human and cloned speech. In Wester, Wu, and Yamagishi (2015), MFCC and cosine-normalized phase (cos-phase) features were used with the GMM-Universal background model for voice cloning detection.

2.2.2. Magnitude-oriented approaches

In Patel and Patil (2015), cochlear filter cepstral coefficients (CFCC) and CFCC-instantaneous frequency (CFCCIF) features were used with the GMM for audio spoofing detection. In Wu, Xiao, Chng, and Li (2013), modulation features were used to design a model for synthetic speech detection. For this purpose, MFCC and modified group delay cepstral coefficients (MGDCC) features were extracted from the magnitude and phase spectrums, respectively, and used by the GMM to classify the speech as bonafide or clone. Malik (2019) employed the higher-order spectral analysis (HOSA) features and gaussian and linearity tests to capture the traces of generative models for bonafide and cloned audio detection.

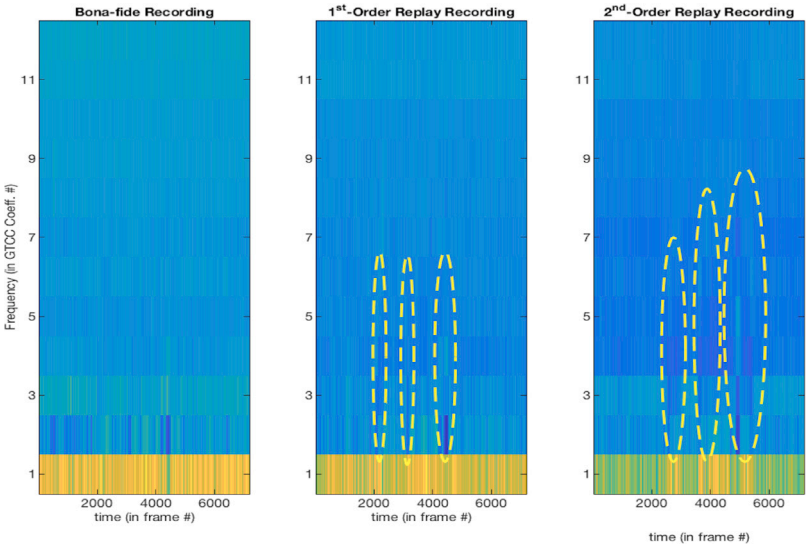

Fig. 3. GTCC features for bona fide, first-, and second-order replay: Twelve coefficients of GTCC features are plotted against the frames of entire audio signal to reveal the details of distortions/artifacts.

3. Analysis of single- and multi-order audio spoofing

Voice spoofing attacks can be employed to exploit both the ASV and VCSs. We categorize these attacks into replays, cloning, and cloned-replay (Fig. 1), and these can be either single- or multi-order.

We model the first-order voice spoofing attack (replay) depicted in Fig. 2 as microphone-speaker-microphone (MSM) processing chain. This is similar to three second-order systems in cascade. Therefore, this MSM chain (demonstrating a 1st-order replay attack) is anticipated to add higher-order non-linearities because of the cascade of the MSM chain. More specifically, this MSM chain introduces beyond 7th-order non-linearity in the replay signal. Higher-order audio replay and cloned-replay attacks are likely to generate stronger higher-order harmonic distortions (HoHDs) in the replay samples (Fig. 1). Conversely, bonafide voice samples lack MSM processing chain and likely to present lesser HoHDs. Therefore, we argue that the HoHDs can be used to discriminate between a bonafide and spoofed voice sample. Spectral features i.e., MFCC, GTCC, etc., or time-domain features i.e., ATCoP can be employed to capture the artifacts of these HoHDs.

Unlike replay and cloned-replay attacks where harmonic distortions exist due to MSM chain, voice cloning attacks are expected to be more linear compared to the bonafide sample. This is because the process of synthetic speech generation is comparatively more linear than the bonafide speech generation process that consists of non-linear sub-processes i.e., respiration, phonation, resonance, and articulation. The bonafide audio recording consists of several components that are input speech signal $s(t)$, environment distortion (reverberant signal $r(t)$ and background noise $\eta(t)$), microphone distortion $\eta_m(t)$, encoding distortion $\eta_e(t)$, and transcoding distortion $\eta_t(t)$. Let $h_m(t)$ be the microphone impulse response and $h_r(t)$ be the room impulse response; we can express the digital audio recording signal as:

$$X(t) = h_r(t) \times h_m(t) \times s(t) + h_m(t) \times \eta(t) + \eta_m(t) + \eta_t(t) \quad (1)$$

Contrarily, cloned voice generation does not include any recording mechanism and thus considered linear compared to the bonafide speech. Additionally, cloned voice will not contain microphone fingerprints like those found in the bonafide audio signal. Therefore, we hypothesize that acoustic and spectral characteristics of cloned signal should be different than the bonafide ones, and ATCoP and GTCC should be able to detect these differences with high accuracy.

In our prior work (Malik, 2012; Malik et al., 2019), we have demonstrated that replay attacks add HoHDs and employed the HOSA to capture these nonlinear distortions. However, HOSA features are less feasible for VCSs because of higher computational cost. Additionally, there exists a need to develop robust audio features which are capable of effectively detecting multiple spoofing attacks. To support our claims and need of robust features, we discuss an example of replay attacks. We created the plots of GTCC features (Fig. 3) for bonafide (left), 1st-order (center), and 2nd-order replay (right) audios to show the effectiveness of our ATCoP-GTCC features to better capture the harmonic distortions. Fig. 3 reveals that replay attacks add harmonic distortions (highlighted ellipses) in the replay samples; and our proposed features can capture these distortions. From Fig. 3, we can also observe that these distortions are more prominent in 2nd-order replay audios as compared to the 1st-order replay audios. This fact endorses our claim that *higher-order audio spoofing attacks are more likely to instigate stronger HoHDs in the audios*.

4. Unified voice spoofing detection framework

This section provides a detailed discussion of the proposed unified anti-spoofing framework. The details of the proposed novel ATCoP-GTCC features are presented in this section. The framework of our system is presented in Fig. 4.

4.1. Features extraction

For accurate spoofing detection, we need to develop robust features that can better extract the unique traits of bonafide and spoofed audios. For this purpose, we introduce a novel hybrid ATCoP-GTCC features to detect various diverse voice spoofing attacks. We provide the details of the proposed features extraction methods below.

4.1.1. Acoustic ternary co-occurrence patterns

The 1-D acoustic patterns i.e. local binary patterns (LBP), local ternary patterns (LTP) (Adnan et al., 2018) have been employed in various audio processing applications including the audio spoofing detection. However, these descriptors have certain limitations such as LBP is sensitive to noise, and possibility of different LBP codes generation for the same class that makes it less effective for bonafide vs spoof classification. On the other hand, LTP employs a fixed threshold-based method that is not much robust over dynamic patterns that exist in the spoofed audios. The limitations of these existing acoustic patterns motivated us to propose a novel feature representation i.e. ATCoP

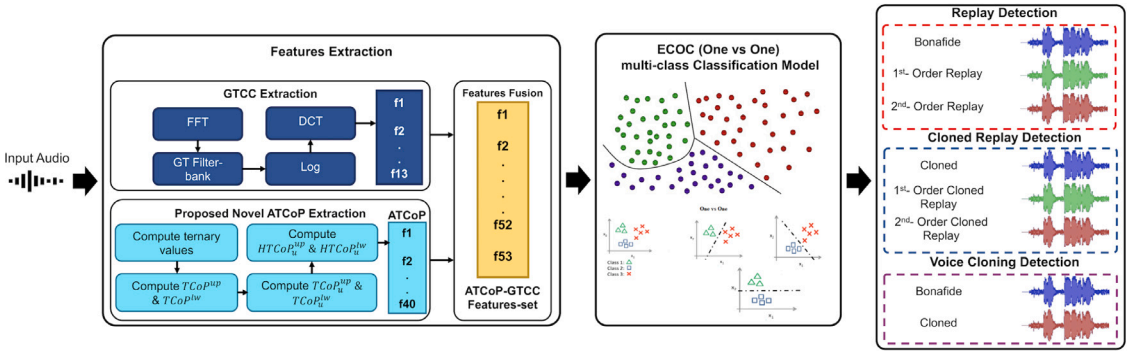

Fig. 4. Architecture of the proposed unified anti-spoofing framework.

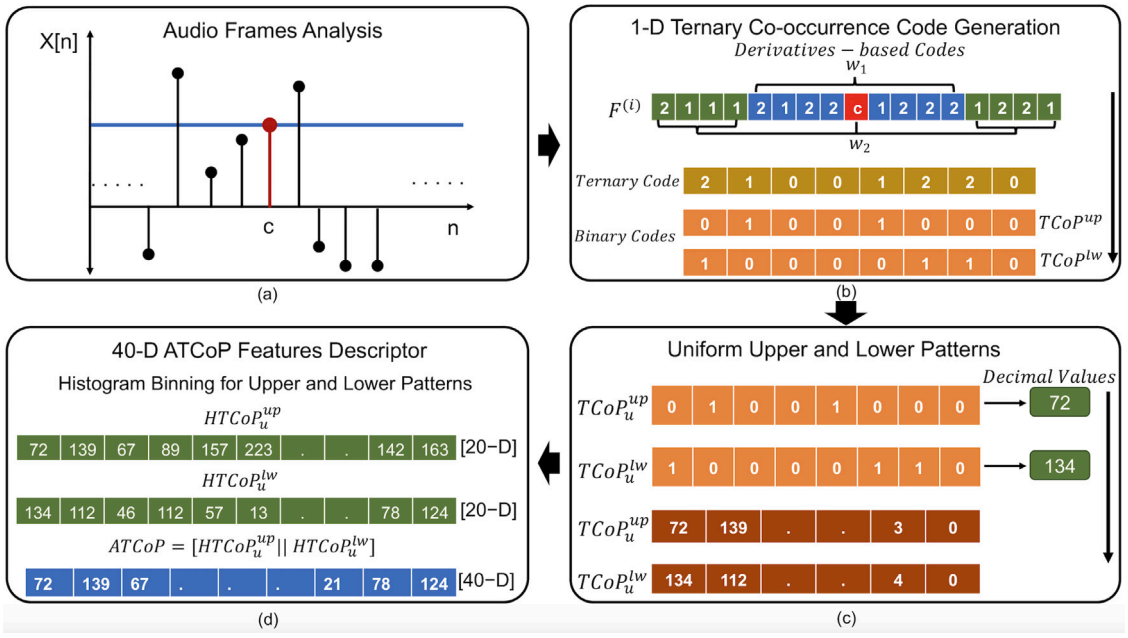

Fig. 5. ATCoP features extraction method.

for 1-D acoustic signals. ATCoP encodes the co-occurrence of similar ternary patterns between the center and neighboring samples without needing any threshold. Additionally, ATCoP provides an effective representation of the audio that can reliably be used to better capture the characteristics of bonafide and spoofed speeches.

Let $X[n]$ be an audio signal with N samples divided into m overlapping frames $F(i)$ having 17 samples in each frame with a step-size of 8, where $i = 1, 2, \ldots, m$. In each frame $F(i)$, c denotes the central sample (Fig. 5(a)). We divide each frame into two windows w_1 and w_2 having adjacent neighbors $z^j_{w_1}$ and far away neighbors $z^j_{w_2}$ as shown in (Fig. 5(b)), where j is the neighboring index w.r.t the sample c. w_1 consists of 4 adjacent neighbors on each side of the central sample c that is highlighted in blue color in $F(i)$ (Fig. 5(b)). Whereas, w_2 consists of remaining 8 samples in $F(i)$ that is highlighted in green color in Fig. 5(b). To compute the ATCoP, we first calculate the 1st-order derivative between the central and each neighboring sample in w_1 and repeat this process for w_2 as shown in Eqs. (2) and (3).

$$D(z^j_{w_1}, c) = z^j_{w_1} - c \tag{2}$$

$$D(z^j_{w_2}, c) = z^j_{w_2} - c \tag{3}$$

where $z^j_{w_1}$ and $z^j_{w_2}$ represent the neighboring samples of w_1 and w_2, respectively. $D(z^j_{w_1}, c)$ and $D(z^j_{w_2}, c)$ represent the first-order derivatives computed between the center and neighboring samples in w_1 and w_2 respectively. Next, we code them according to the sign of first-order derivative as follows:

$$P_l(z^j_{w_1}, c) = \begin{cases} 1, & D(z^j_{w_1}, c) > 0, \\ 2, & D(z^j_{w_1}, c) \leq 0 \end{cases} \tag{4}$$

$$P_l(z^j_{w_2}, c) = \begin{cases} 1, & D(z^j_{w_2}, c) > 0, \\ 2, & D(z^j_{w_2}, c) \leq 0 \end{cases} \tag{5}$$

where $P_l(z^j_{w_1}, c)$ and $P_l(z^j_{w_2}, c)$ represents the assigned codes to the samples of w_1 and w_2, respectively. Next the samples of the corresponding locations in w_1 and w_2 are compared to generate the ternary values as follows:

$$TCoP(z^j, c) = \begin{cases} f(P_1(z^j_{w_1}, c), P_1(z^j_{w_2}, c)), \\ f(P_2(z^j_{w_1}, c), P_2(z^j_{w_2}, c)), \ldots, \\ f(P_8(z^j_{w_1}, c), P_8(z^j_{w_2}, c)) \end{cases} \tag{6}$$

where,

$$TCoP(z^j, c) = \begin{cases} 1, & if\, x = y = 1 \\ 2, & if\, x = y = 2 \\ 0, & if\, x \neq y \end{cases} \tag{7}$$

where $TCoP(z^j$,c) represents the ternary patterns. We further divide the ternary patterns into two binary patterns that are upper patterns

$TCoP_up$ (.) and lower patterns $TCoP_lw$ (.). We retain all values of 1 in $TCoP_up$ (.) and replaced the rest with zeros as follows:

$$TCoP^{up}(z^j, c) = \begin{cases} 1, & if TCoP(z^j, c) = 1 \\ 0, & Otherwise \end{cases} \quad (8)$$

Likewise, we retain all values of 2 in $TCoP_lw$(.) while replacing the rest with zeros as follows:

$$TCoP^{lw}(z^j, c) = \begin{cases} 1, & if TCoP(z^j, c) = 2 \\ 0, & Otherwise \end{cases} \quad (9)$$

Next, we adopt the concept of picking uniform patterns over non-uniform patterns (Ojala, Pietikäinen, & Harwood, 1996). The uniform patterns hold significant attributes of the signal, whereas, non-uniform patterns mostly contain the redundant information. We computed the uniform patterns, $TCoP_u^{up}$(.) and $TCoP_u^{lw}$(.) from the $TCoP^{up}$(.) and $TCoP^{lw}$(.) as depicted in Fig. 5(c), and represented these ternary co-occurrence patterns in decimal form as:

$$TCoP_u^{up}(z^j, c) = \sum_{j=0}^{j=7} TCoP_u^{up}(z^j, c) \times 2^j \quad (10)$$

$$TCoP_u^{lw}(z^j, c) = \sum_{j=0}^{j=7} TCoP_u^{lw}(z^j, c) \times 2^j \quad (11)$$

where the $TCoP_u$ value represents the number of bit-wise transitions (0/1 changes) in the pattern. The co-occurrence patterns with minimal transitions are considered uniform i.e., 11111111 and 00000001 patterns have uniform values of 0 and 1 respectively. After computing the $TCoP_u^{up}$ and $TCoP_u^{lw}$, we calculate the histograms of these uniform patterns. We assign one histogram bin for each uniform pattern and include all non-uniform patterns in a single bin to ensure reducing only the redundant information from the input sample (Fig. 5(d)).

$$HTCoP_u^{up}(TCoP^{up}, n) = \sum_{k=1}^{K} \delta(TCoP_k^{up}, n) \quad (12)$$

$$HTCoP_u^{lw}(TCoP^{lw}, n) = \sum_{k=1}^{K} \delta(TCoP_k^{lw}, n) \quad (13)$$

where n shows the histogram bins corresponding to the uniform ATCoP codes and $\delta(.)$ is the Kronecker delta function. We performed substantial experiments to generate these ATCoP codes by selecting different number of bins for uniform patterns. After detailed experimentation, we observe that the first 20 uniform patterns from each of $TCoP_u^{up}$ and $TCoP_u^{lw}$ were enough to capture the distortions in replay, artifacts in cloning, and dynamic speech variations of bonafide samples. Therefore, we create a 20-D ATCoP code each for $TCoP_u^{up}$ and $TCoP_u^{lw}$ and fused them to generate a 40-D ATCoP features as follows:

$$ATCoP = [HTCoP_u^{up} \bigsqcup HTCoP_u^{lw}] \quad (14)$$

where $\bigsqcup$ represents the concatenation operator for vectors.

Analysis of ATCoP features. Due to the vulnerability of the smart speakers against different voice spoofing attacks, we need a robust anti-spoofing system that should investigate the following facts while designing the features: (i) the microphone introduces a layer of non-linearity because of inter-modulation distortions, which introduce the discernible patterns, (ii) introduction of the higher-order non-linearities in consequent recordings of the given recording make these audios more distinct, (iii) voice cloning methods also add the algorithmic artifacts, and (iv) presence/absence of dynamic speech variations in bonafide/cloned voice. Hence, these facts must be considered while proposing a robust voice anti-spoofing system.

Our ATCoP features are developed to capture the traces of unique attributes of bonafide and spoof audios in time domain. To justify the effectiveness of our ATCoP features for distinct representation of bonafide and various categories of spoof audios, we created the detailed graphs of ATCoP features for the bonafide, cloned, replay, and cloned-replay audios as shown in Fig. 6. For each analysis, we selected the audios of same speaker for both the bonafide and spoof categories for fair comparison. We plotted the ATCoP features of bonafide, 1st-order replay, and second-order replay for VSDC audios in Fig. 6(a). Likewise, features of bonafide and replay for ASV-spoof PA, and bonafide and cloned for ASV-spoof LA corpus are presented in Fig. 6(b) and (c) respectively. Finally, ATCoP features for cloned, 1st-order cloned replay, and second-order cloned replay are shown in Fig. 6(d). By analyzing the peaks of these graphs, we can conclude that our ATCoP features give distinct representation for bonafide and different categories of spoof audios at same feature-points. This analysis demonstrate that ATCoP features can reliably be used to represent the input audios for spoofing detection problem.

4.1.2. Gammatone cepstral coefficients (GTCC)

Spectral features such as GTCC, MFCC, etc., can be employed to capture the non-linearities in frequency scale of the input audio signal. MFCC features have been explored for various audio processing applications due to its effectiveness to capture the significant attributes of the acoustic signal. Recently, GTCC features have also been employed due to their enhanced filter response that better resemble the human auditory system. We employed the GTCC features with our ATCoP features for voice spoofing detection due to two reasons: i) GTCC are more tolerant to noise over MFCC (Cooper, 2013), and (ii) provide marginally better classification performance over MFCC with comparable computational cost. The ability of GT filter to offer more frequency components in low-frequency band and less frequency components in high-frequency band allows us to better capture the non-linearities in the audio signal.

For GTCC extraction, we employed the fast Fourier transform (FFT) on the input audio signal. Next, the gammatone filter bank consisting of different GT filters is applied to the FFT to compute the energy of each sub-band. The discrete cosine transform is applied on the log of each energy band to extract the GTCC features as shown in Fig. 7, where we obtain a 13-D GTCC features vector. It is to be noted that 13 to 20 coefficients are considered enough for optimal audio analysis. Thus, we extracted 13 GTCC coefficients and later fused them with the proposed novel ATCoP features for audio signal representation.

4.2. Classification

To address the multi-class classification problem, we employed the error correcting output codes (ECOC) framework (Escalera, Pujol, & Radeva, 2009) by combining three binary classifiers. ECOC model generates a codeword against each class during encoding and predict the class of given test sample at the decoding phase.

Since we have three classes for replay and cloned-replay detection, we train three binary learners using two classes at a time to obtain a 3-digit codeword for each class. Each bit of the codeword specifies the response of the given binary learner. More precisely, we used three codes -1,0,1 during the encoding to ignore one class and compare the other two in the one vs one approach (Table 1). So, our ternary coding matrix for three classes is shown in Table 1, where the 3-bit error correcting output code word is presented for three-class classification. Each class is assigned a unique 3-bit code-word. One binary classifier is learned for each column during the training. As shown in Table 1, first learner (L-1) is trained to separate class 1 and 2, second learner (L-2) is trained to separate class 1 and 3, whereas, the third learner (L-3) is trained to distinguish class 2 and 3. For each column, 0 is used to ignore the third class while the remaining two classes are used in the classification process. Three binary classifiers are trained in this way.

At decoding, all of these three binary classifiers are evaluated to obtain a 3-bit code. We employed the hamming distance to compute the closest match between this 3-bit code and the assigned code-words of each class. Finally, we select the class of the input audio as the one whose code-word has minimum distance with the 3-bit code of the sample. Our ECOC framework uses three binary SVM learners for spoofing detection.

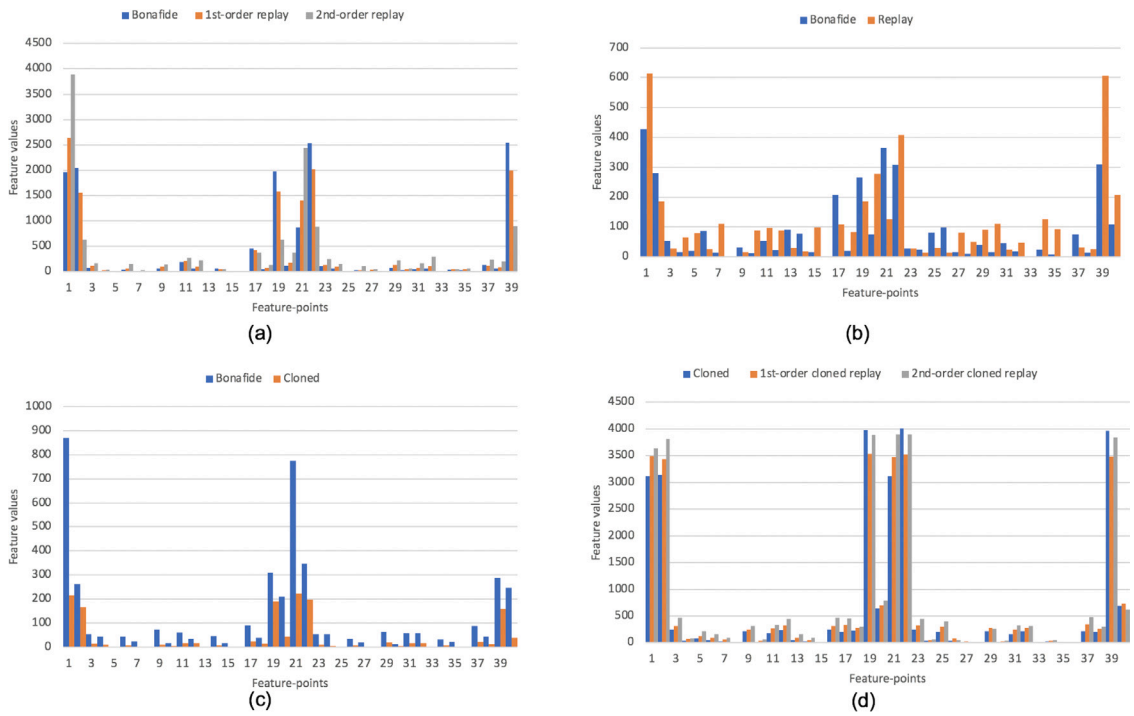

Fig. 6. ATCoP features analysis.

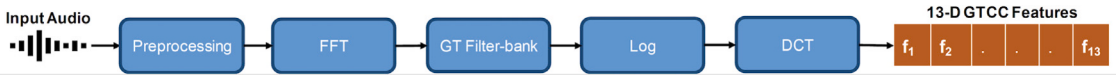

Fig. 7. GTCC features extraction method.

Table 1
Coding matrix design for 3 class classification.

	L-1	L-2	L-3
C-1	1	1	0
C-2	−1	0	1
C-3	0	−1	−1

5. Experimental setup and results

5.1. Dataset

Performance of our system is measured on in-house created VSDC (Baumann et al., 2021), ASVspoof 2019 (Yamagishi et al., 2019), Google's LJ Speech (Ito & Johnson, 2021), and YouTube deepfakes (Agarwal et al., 2019) datasets. VSDC is designed for single- and multi-order voice replay and cloned-replay attacks detection for diverse and challenging scenarios. VSDC comprises both the first- and second-order replay audios against the bonafide ones, unlike ASVspoof (Yamagishi et al., 2019) that contains only the first-order replay samples against the bonafide. Our VSDC (Baumann et al., 2021) is diverse in terms of environment, configurations, speaker genre, recording and playback devices, recording and playback configurations, and number of speakers (Table 2). Each bonafide and replay audio sample in our dataset is 6 s in duration. Since we introduce a new spoofing threat, cloned-replay, that represents the recording of cloned voice sample, we used the ASVspoof synthetic samples to generate the first- and second-order cloned replay samples. Our VSDC is publicly available and more details can be found at (Baumann et al., 2021).

ASVspoof 2019 dataset (Yamagishi et al., 2019) contains the logical access (cloning) and physical access (replay) samples for training, development and evaluation. Training, development, and evaluation sets for replay contain 54 000, 33 534, and 153 522 samples, respectively. Whereas, training, development, and evaluation sets for voice cloning contain 25 380, 24 844 and 71 933 samples, respectively.

LJ Speech is a public domain dataset consisting of 13 100 bonafide audio samples. The duration of voice samples of this dataset varies from 1 to 10 s with a total length of 24 h. Each voice sample is recorded with sampling rate of 22 050 Hz. We employed Google's cloning model (Kang, 2021) to generate 1500 spoofing samples and later used these bonafide and cloned samples for voice cloning detection.

Deepfakes corpus (Agarwal et al., 2019) contains different YouTube videos of various US politicians with average length of 1.5 h. The audio streams of these videos are also forged, and we used them to evaluate our framework.

5.2. Performance evaluation of proposed anti-spoofing system

Performance of the proposed system is measured using the min-tDCF, and EER. We designed separate experiments to detect the replay, cloning, and cloned-replay attacks. For replay experiments, we computed the results on both the ASVspoof 2019 and VSDC. For speech synthesis, we used ASVspoof 2019, LJ Speech, and YouTube deepfakes datasets, whereas, we used the VSDC corpus for cloned-replay detection. The details of datasets division for experimentation are presented in Table 2.

5.2.1. Detection performance of ATCoP, GTCC, and fusion

We conducted an experiment to investigate the performance of ATCoP, GTCC, and their fusion for audio spoofing detection. For this purpose, we employed the ATCoP features with SVM for replay detection (on ASVspoof 2019 and VSDC datasets separately), synthetic speech/cloning detection (on ASVspoof 2019 and LJSpeech datasets separately), and cloned-replay detection on VSDC. The results are

Table 2
Datasets division for experimentation.

Dataset	Training		Testing	
	Division	Number of samples	Division	Number of samples
ASVspoof-PA	Train	54,000	Eval	1,53,522
ASVspoof-LA	Train	25,380	Eval	71,933
VSDC	70%	8397	30%	3603
LJSpeech	70%	18,340	30%	7860
YouTubes deepfakes	70%	63	30%	27

Table 3
Comparative analysis of ATCoP, GTCC, and ATCoP-GTCC.

Spoofing	Dataset	Features	min-tDCF	EER%
Replay	ASVspoof	ATCoP	0.097	2.80
		GTCC	0.211	8.35
		ATCoP-GTCC	0.064	1.00
	VSDC	ATCoP	0.079	2.10
		GTCC	0.21	7.98
		ATCoP-GTCC	0.056	0.90
Synthesis/Cloning	ASVspoof	ATCoP	0.059	0.80
		GTCC	0.132	6.10
		ATCoP-GTCC	0.011	0.10
	LJSpeech	ATCoP	0.007	0.10
		GTCC	0.019	0.28
		ATCoP-GTCC	0.0	0.0
Cloned replay	VSDC	ATCoP	0.05	0.90
		GTCC	0.19	3.99
		ATCoP-GTCC	0.002	0.03

Table 4
Replay detection results.

Dataset	SVM Kernel	min-tDCF	EER%
VSDC	Linear	0.503	25.00
	Quadratic	0.078	2.00
	Cubic	0.0576	0.90
	RBF	0.05	0.75
ASVspoof	Linear	0.068	1.50
	Quadratic	0.064	1.00
	Cubic	0.064	1.00
	RBF	0.068	1.50

provided in Table 3. We repeated this experiment for the evaluation of GTCC and ATCoP-GTCC features. From these results, we found that the ATCoP offers better performance as compared to the GTCC, but ATCoP-GTCC fusion performed the best. Thus, we employed the ATCoP-GTCC features for audio spoofing detection.

5.2.2. Detection performance of the proposed ATCoP-GTCC features for replay attack detection

We used our ATCoP-GTCC features to train the SVM using different kernels on both the VSDC and physical access (PA) collection of ASVspoof datasets and results are presented in Table 4. For VSDC, we obtained the lowest min-tDCF and EER of 0.05 and 0.75% on the radial basis function (RBF) kernel, respectively. For ASVspoof 2019 corpus, we obtained the lowest min-tDCF and EER of 0.064 and 1% on quadratic and cubic kernels. We can observe from Table 4 that the SVM tuned with higher-order polynomial (cubic) and RBF kernels gives superior performance over other kernels on both datasets.

From the results, we observed that SVM tuned on higher-order polynomial kernel better captures the non-linearities exist in multi-order replay audios. As expected, SVM with linear kernel attains the highest min-tDCF. We also found that 3^{rd}-order polynomial (cubic) kernel provides better classification over 2nd-order polynomial (quadratic) for multi-order replay detection. This demonstrates the effectiveness of the cubic kernel in differentiating the non-linearities available in multi-order audio replays. Therefore, we argue that the proposed features with SVM tuned on the cubic kernel effectively detects the non-linearities in the replays.

Table 5
Speech synthesis detection results.

Dataset	SVM Kernel	min-tDCF	EER%
ASVspoof	Linear	0.078	2.00
	Quadratic	0.05	0.75
	Cubic	0.047	0.70
	RBF	0.05	0.75
LJSpeech	Linear	0.0	0.0
	Quadratic	0.0	0.0
	Cubic	0.0	0.0
	RBF	0.0	0.0

Table 6
Cloned replay detection results.

Dataset	SVM Kernel	min-tDCF	EER%
ASVspoof	Linear	0.011	0.17
	Quadratic	0.007	0.10
	Cubic	0.002	0.03
	RBF	0.01	0.15

Table 7
Feature vectors.

Feature-vector	Features
Spectral-MFCC-GTCC 40-D	GTCC [1–13], MFCC [1–13], Spectral (Kurtosis, Skewness, Slope, Centroid, Flatness, Entropy, Decrease, Rolloff point, Flux, Crest, Spread), Energy
ATCoP-Spectral 51-D	ATCoP [40-D], Spectral [11-D]
ATP-MFCC 53-D	ATCoP [40-D], MFCC [13-D]
ATP-GTCC 53-D	ATCoP [40-D], GTCC [13-D]

5.2.3. Detection performance of the proposed ATCoP-GTCC features for speech synthesis (voice cloning) detection

Performance of the proposed anti-spoofing framework is also evaluated on logical access (LA) collection of ASVspoof 2019 and LJ Speech 1.1 datasets for voice cloning detection. For this purpose, we used our features to train the SVM for classification of audio as bonafide or spoof and results obtained on the SVM using different kernels are provided in Table 5.

From the results (Table 5), we can observe that the SVM tuned on higher-order polynomial (cubic and quadratic) and RBF kernels provide remarkable classification performance. It is to be noted that SVM tuned with cubic kernel performs marginally better than quadratic and RBF kernels. More specifically, we obtained min-tDCF and EER of 0.047 and 0.7% on ASVspoof 2019 LA dataset. On the other hand, we obtained the optimal 0.0 min-tDCF and 0% EER for voice cloning detection on LJ Speech dataset. This remarkable performance is attributed to the fact that these cloned samples do not have microphone signatures which our framework successfully detects.

5.2.4. Detection performance of the proposed ATCoP-GTCC features for cloned-replay attack detection

We used the ASVspoof 2019 LA dataset of cloned voices of different speakers to create the first- and second-order cloned replay recordings. We extracted the features from these cloned and cloned-replay samples (1st- and 2nd-order) and train the SVM for classification, and results are reported in Table 6. We conclude from these results that SVM provides remarkable results on all kernels to classify among the cloned, 1st- and 2nd-order cloned audio replays. However, higher-order polynomial kernel (cubic) achieves best results with a small margin. More specifically, we obtained min-tDCF of 0.002 and EER of 0.03% on the cubic kernel of SVM.

Table 8
Comparative analysis of the proposed and other spectral features for replay attacks detection.

Dataset	Features	min-tDCF	EER%
VSDC	MFCC-GTCC-Spectral	0.084	2.33
	ATCoP-Spectral	0.108	4.60
	ATCoP-MFCC	0.065	1.16
	ATCoP-GTCC	0.0576	0.90
ASVspoof	MFCC-GTCC-Spectral	0.137	6.75
	ATCoP-Spectral	0.064	1.00
	ATCoP-MFCC	0.064	1.00
	ATCoP-GTCC	0.064	1.00

Table 9
Comparative analysis of the proposed and other spectral features for speech cloning detection.

Dataset	Features	min-tDCF	EER%
ASVspoof	MFCC-GTCC-Spectral	0.099	3.00
	ATCoP-Spectral	0.053	0.80
	ATCoP-MFCC	0.05	0.75
	ATCoP-GTCC	0.043	0.65
LJSpeech	MFCC-GTCC-Spectral	0.021	0.30
	ATCoP-Spectral	0.0	0.0
	ATCoP-MFCC	0.0	0.0
	ATCoP-GTCC	0.0	0.0

Table 10
Comparative analysis of the proposed and other spectral features for cloned replay detection.

Dataset	Features	min-tDCF	EER%
VSDC	MFCC-GTCC-Spectral	0.028	0.40
	ATCoP-Spectral	0.064	1.00
	ATCoP-MFCC	0.021	0.30
	ATCoP-GTCC	0.002	0.03

5.2.5. Evaluation of proposed ATCoP and spectral features fusions

To justify the robustness of our features for voice spoofing detection, we created different combination of features using our ATCoP and spectral features (Table 7). For classification, we employed the SVM and results are presented in Tables 8 to 10.

The results of replay attack detection are presented in Table 8. From Table 8, we can see that our proposed ATCoP-GTCC features outperform others and attained the min-tDCF and EER of 0.0576 and 0.9% on VSDC, and 0.064 and 1% on the ASVspoof dataset. Whereas, we obtained the highest min-tDCF of 0.084 and 0.137 on MFCC-GTCC-spectral features for VSDC and ASVspoof PA datasets respectively.

Similarly, we provided the results of voice cloning/speech synthesis detection in Table 9. Again, our ATCoP-GTCC features provide better results over other features by attaining min-tDCF of 0.043 and 0.0. Whereas, MFCC-GTCC-spectral achieved the highest min-tDCF of 0.099 and 0.021 for ASVspoof and LJ Speech datasets, respectively. It is important to mention that for LJ Speech dataset, all features fusion containing ATCoP features achieve an optimal min-tDCF of 0.0. However, our ATCoP-GTCC features outperform other features on the ASVspoof LA dataset.

Finally, we evaluated the performance of all feature combinations for cloned-replay detection on VSDC cloned-replay collection, and results are provided in Table 10. Similar to the other experiments, we also achieved the best results for cloned-replay detection using our proposed features. More specifically, we obtained min-tDCF of 0.028 on MFCC-GTCC-spectral, 0.064 on ATCoP-spectral, 0.021 on ATCoP-MFCC, and 0.002 on our ATCoP-GTCC features.

We can observe from the results presented in Tables 8 to 10 that the proposed ATCoP-GTCC features outperform other features-sets. More specifically, we obtained the lowest min-tDCF of 0.0576 for replay attacks detection, optimal 0.0 for voice cloning detection, and 0.002 for cloned-replay detection. These results illustrate the effectiveness of the proposed features for accurate spoofing detection. In short, our novel ATCoP-GTCC features lay the foundation for a unified anti-spoofing framework capable of reliable detection of multiple types of voice spoofing attacks.

5.3. Comparative analysis using different classifiers

We designed an experiment to compare the performance of the proposed features on other machine learning classifiers for replay, cloned-replay, and speech synthesis detection. For this, we used our ATCoP-GTCC features to train the conventional ML and DL classifiers and results are reported in Table 11.

5.3.1. Decision trees classification

For decision trees (Breiman, Friedman, Olshen, & Stone, 2017), we computed the classification results on different depths i.e. fine, medium and coarse, where fine-level has more depth and coarse-level has least depth in tree structure. It is to be noted that decision trees trained at fine-level performed best for all types of spoofing detection. The results on decision trees tuned at fine-level are shown in Table 11.

5.3.2. Naïve Bayes classification

For Naïve Bayes (Fu, Lu, Ting, & Zhang, 2010), we computed the results using the gaussian and kernel distributions. For replay detection, we obtained the min-tDCF of 0.651 and 0.478 with gaussian distribution, and 0.528 and 0.421 with kernel distribution on VSDC and ASVspoof PA datasets, respectively. For voice cloning, we obtained the min-tDCF of 0.139 and 0.0 on gaussian and 0.097 and 0.0 on kernel distribution for ASVspoof LA and LJ Speech datasets. Similarly, for cloned replay detection, we obtained the min-tDCF of 0.106 and 0.067 on gaussian and kernel distributions, respectively. We observed from the experiments that Naïve Bayes using the kernel distribution outperforms the gaussian distribution for voice spoofing detection, however with higher computational cost and memory. The results on Naive Bayes using the kernel distribution are shown in Table 11.

5.3.3. K-nearest neighbor (KNN) classification

For KNN (Zhang, Li, Zong, Zhu, & Wang, 2017) experiments, we tuned three parameters i.e number of neighbors k_n, distance weights d_w, and distance metric d_m. We measured the performance on three different values of k_n (1, 10, 100), three different d_m (Euclidean, cubic, cosine), and two variations of d_w (equal, squared inverse). For all three spoofing categories, we obtained best results on weighted-KNN (k_n=10, d_w = squared inverse, d_m=Euclidean) for all datasets, as shown in Table 11.

5.3.4. Ensemble learning models

We also evaluated our method on different ensemble classifiers (Dietterich, 2000) i.e. bagged trees (Sun, Lang, Fujita, & Li, 2018), boosted trees (Hubáček, Šourek, & Železný, 2019), RUSBoosted trees (Moeyersons, Varon, Testelmans, Buyse, & Van Huffel, 2017), subspace discriminant (Hang, Liu, Song, & Sun, 2015), and subspace KNN (Zhang, Cao, Wang, & Li, 2019). We achieved best results on ensemble bagged trees and worst on RUSboosted trees for all types of spoofing detection. For replay detection, we obtained the min-tDCF of 0.115 and 0.104 on bagged trees, and 0.512 and 0.601 on RUSboosted trees for VSDC and ASVspoof PA datasets, respectively. For cloning detection, we obtained the min-tDCF of 0.068 and 0.0 on bagged trees, and 0.448 and 0.142 on RUSboosted trees for ASVspoof LA and LJ Speech datasets, respectively. For cloned-replay detection, we obtained the min-tDCF of 0.032 and 0.097 on bagged- and RUSboosted-trees. The results on ensemble bagged trees are presented in Table 11.

Table 11
Comparative Analysis of different classifiers with ATCoP-GTCC features.

Dataset	Classifiers	Replay		Cloning		Cloned replay	
		min-tDCF	EER%	min-tDCF	EER%	min-tDCF	EER%
VSDC	Decision trees	0.389	17.83	–	–	–	–
	Naïve Bayes	0.528	27.00	–	–	–	–
	KNN	0.058	0.92	–	–	–	–
	Ensemble Models	0.115	1.83	–	–	–	–
	BiLSTM	0.313	13.10	–	–	–	–
	SVM	0.0576	0.90	–	–	–	–
ASVspoof 2019	Decision trees	0.216	9.00	0.121	5.00	0.048	0.75
	Naïve Bayes	0.421	19.75	0.097	2.80	0.067	1.41
	KNN	0.137	6.75	0.078	2.00	0.032	0.50
	Ensemble Models	0.104	4.50	0.068	1.50	0.032	0.50
	BiLSTM	0.308	12.70	0.081	2.20	0.032	0.50
	SVM	0.064	1.00	0.05	0.75	0.002	0.03
LJSpeech	Decision trees	–	–	0.0	0.0	–	–
	Naïve Bayes	–	–	0.0	0.0	–	–
	KNN	–	–	0.0	0.0	–	–
	Ensemble Models	–	–	0.0	0.0	–	–
	BiLSTM	–	–	0.0	0.0	–	–
	SVM	–	–	0.0	0.0	–	–

Table 12
Performance comparison with existing contemporary anti-spoofing methods.

Spoofing	Dataset	Methods	min-tDCF	EER%
Replay	ASVspoof 2019-PA-Eval	CQCC-GMM baseline (Yamagishi et al., 2019)	0.2454	11.04
		LFCC-GMM baseline (Yamagishi et al., 2019)	0.3017	13.54
		FBCC-GMM (Kumar & Bharathi, 2021)	0.25	10.36
		Stat-SE-Res2Net50 (Li et al., 2021)	0.027	1.00
		LFCC+ProdSpec+MGDCC-CNN (Monteiro et al., 2020)	0.07	2.015
		CQT+LFCC+DCT-LCNN (Lavrentyeva et al., 2019)	0.0122	0.54
		ATCoP+GTCC-SVM (Proposed method)	0.064	1.00
Cloning	ASVspoof 2019-LA-Eval	CQCC-GMM baseline (Yamagishi et al., 2019)	0.236	9.87
		LFCC-GMM baseline (Yamagishi et al., 2019)	0.212	11.96
		FBCC-GMM (Kumar & Bharathi, 2021)	0.155	6.16
		Stat-SE-Res2Net50 (Li et al., 2021)	0.068	2.86
		LFCC+ProdSpec+MGDCC-CNN (Monteiro et al., 2020)	0.198	9.09
		CQT+LFCC+DCT-LCNN (Lavrentyeva et al., 2019)	0.051	1.84
		ATCoP+GTCC-SVM (Proposed method)	0.05	0.75

5.3.5. Deep learning classification

For deep learning, we selected the BiLSTM recurrent deep learning method (Graves & Schmidhuber, 2005). As recurrent DL models are better suited to analyze the sequential and time series data, therefore, we selected the BiLSTM framework among other deep learning models. For experimentation, we tuned the network on 200 hidden units, tanh state activation function, sigmoid gate activation function, maximum epochs of 200, mini-batch size of 64, and 5 hidden layers, as optimal results were obtained on these settings (Table 11).

5.4. Hybrid dataset evaluation

The objective of this experiment is to assess the performance of proposed anti-spoofing framework on more diverse audio samples. Since both the ASVspoof 2019 and VSDC datasets have different characteristics (e.g. sampling rate, speakers, microphone and playback devices, environments, etc.), therefore, we have created a hybrid dataset comprising of bonafide and spoof samples of the ASVspoof and VSDC. Since the VSDC consists of only the bonafide and replay samples, therefore, we have also selected the replay collection (PA) of ASVspoof 2019 dataset. For this experiment, we have taken 8000 audio samples from the training set of ASVspoof PA collection and 8000 bonafide and 1st-order replay audios from the VSDC. Next, we used 70% audios (11,200) for training and rest 30% audios (4800) for testing and achieved the min-tDCF of 0.227 and EER of 9.1%. From the results, we can observe that our anti-spoofing framework achieves better classification performance even on more diverse audio samples.

5.5. Discussion

The proposed ATCoP-GTCC features effectively detect different kinds of voice spoofing attacks that are evaluated on four different datasets. We provided the comparative results obtained on different classifiers for audio spoofing detection in Table 11. From the results, we found that SVM was the best and Naïve Bayes was the worst performer for all types of spoofing. More precisely, SVM achieved the lowest min-tDCF and EER of 0.0 and 0%, 0.0576 and 0.9%, and 0.002 and 0.03% for cloning, replay, and cloned-replay detection, respectively. Thus, we argue that SVM can reliably be used with the proposed features to detect any kind of voice spoofing attack.

We performed one experiment to investigate the effect of compressed replays generation against the bonafide samples. In our dataset, we first used Bluetooth speakers for voice replays generation then we obtained the compressed 1st-order and 2nd-order replay audios. For this, we selected 1269 bonafide, first- and second-order replay audios. We extracted the proposed features for these samples and trained the SVM for replay attacks detection. We obtained lower min-tDCF and EER values on these samples compared to all samples. More specifically, we obtained the lowest min-tDCF and EER of 0.032 and 0.5% for compressed replay samples compared to 0.057 and 0.9% over all samples. We conclude from this experiment that compressed spoofing samples are easier to detect than uncompressed samples due to the fact that the microphone distortions are relatively weak in compressed samples. This also complements our conclusion that cloned replay samples are easier to detect due to weak microphone distortions.

Additionally, it is also important to understand that microphone non-linearities do contribute to microphone induced distortions, but it

is not the only source of distortion. Separation and estimation of each distortion component, e.g., device non-linearity, material or fabrication imperfections, etc., for non-Gaussian inputs (e.g., speech signals) is a challenging task. Intuitively, the replay process amplifies this microphone induced distortions which can be captured via ATCoP-GTCC features. Moreover, human voice holds dynamically induced vocal-tract variations as compared to synthetic speech. For example, natural pauses of the human speech production model are missing from the synthetic speech generated by voice cloning algorithms (Mwiti, 2019). On the other hand, cloned voice sounds similar and contains unusual prosody. The results indicate that our ATCoP-GTCC performs remarkably well for synthetic speech detection which proves that ATCoP-GTCC is able to reliably capture the dynamically variant characteristics of bonafide speech and algorithmic traits of the synthetic speech.

5.6. Comparative analysis with existing methods

To measure the effectiveness of our unified anti-spoofing framework, we compared our system against existing state-of-the-art voice anti-spoofing methods (Kumar & Bharathi, 2021; Lavrentyeva et al., 2019; Li et al., 2021; Monteiro et al., 2020; Yamagishi et al., 2019) on ASVspoof 2019 dataset (Table 12). We employed the ASVspoof 2019-PA-Train/Eval sets for training/testing the proposed and all the comparative methods for replay spoofing and ASVspoof 2019-LA-train/Eval for training/testing the cloning spoofing detection. The proposed method outperforms the contemporary voice cloning detection methods by achieving the lowest min-tDCF of 0.05. Lavrentyeva et al. (2019) was the second best system for cloning detection with min-tDCF of 0.051. For replay detection, Lavrentyeva et al. (2019) was the top performer, whereas, the proposed method along-with Li et al. (2021) was the second best method. Moreover, the LFCC-GMM ASVspoof baseline model (Yamagishi et al., 2019) was the worst performing method for both the replay and cloning detection. It is important to mention that the proposed method performed better over the ASVspoof baseline model by achieving lower EER of 12.54% and 11.21% for replay and cloning detection respectively.

5.7. Performance of proposed features for deepfakes detection

The objective of this evaluation is to quantify the effectiveness of proposed anti-spoofing framework for deepfakes detection. For this experiment, we used deepfakes detection dataset (Agarwal et al., 2019) that comprises of YouTube videos of various US politicians. Agarwal et al. (2019) used this dataset to measure the performance of their visual features oriented deepfakes detection method. We highlighted the fact that we can still develop effective deepfakes detection methods using low-cost audio features. We extracted the audios of the videos from this dataset (Agarwal et al., 2019) comprising of both the bonafide and spoof samples as these videos contain both the visual and audio forgeries. We used our proposed features to train the SVM classifier and obtained min-tDCF of 0.051 and EER of 0.8%. AUC metric was used for performance evaluation in Agarwal et al. (2019). Therefore, we also computed the AUC for this experiment and achieved an AUC of 1 as compared to average AUC of 0.95 achieved by Agarwal et al. (2019). From these results, we argue that our method provides superior classification performance as compared to Agarwal et al. (2019). It is to be noted that our method achieves better performance with low-cost audio features as compared to Agarwal et al. (2019), in which the visual landmark features are employed that are computationally more intensive compared to the proposed ATCoP-GTCC features.

6. Conclusion

This paper has presented a novel unified anti-spoofing framework, that by employing proposed ATCoP-GTCC features, accurately captures the non-linearities introduced in the 1st- and 2nd-order spoofing samples, traces of generative models for both speech synthesis and cloned replay, and dynamic speech variations of bonafide audios. The absence of a multi-order replay spoofing dataset motivated us to develop a diverse voice spoofing detection corpus for multi-order replay and cloned-replay attacks. Additionally, we have presented that hybrid spoofing attacks like cloned-replay can easily be executed in chained scenarios to exploit the VCSs. This research work lays the foundation of addressing multi-order replay, cloning, and cloned-replay voice spoofing attacks, using the unified framework to protect the ASV and VCSs. Experimental results signify the effectiveness of our anti-spoofing framework by achieving optimal results on four datasets having either replay or cloning forgery. This verifies our claim that the proposed features effectively capture the dynamic speech variations and microphone fingerprints of bonafide audio, algorithm artifacts in cloned audio, and non-linear distortions in replay recordings. Additionally, our proposed features also perform remarkably well for deepfakes detection, and this verifies our claim that audio signal analysis is an integral part of deepfakes detection. Based on the fact that the proposed features can effectively capture the traces of manipulated voice attributes i.e., frequencies, etc. in the cloned speech, we argue that our ATCoP-GTCC features can provide superior detection performance even for high-quality synthesized speech samples.

CRediT authorship contribution statement

Ali Javed: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Khalid Mahmood Malik:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Hafiz Malik:** Methodology, Formal analysis, Writing – review & editing, Funding acquisition. **Aun Irtaza:** Methodology, Validation, Formal analysis, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by grant of National Science Foundation of USA via Awards No. (1815724) and (1816019).

References

Adnan, S. M., Irtaza, A., Aziz, S., Ullah, M. O., Javed, A., & Mahmood, M. T. (2018). Fall detection through acoustic local ternary patterns. *Applied Acoustics*, *140*, 296–300.

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting world leaders against deep fakes. In *CVPR workshops, Vol. 1* (pp. 38–45).

Aircrack-Ng (2020). Retrieved from https://www.aircrack-ng.org. Accessed August 4, 2021.

Bakar, B., & Hanilçi, C. (2018). Replay spoofing attack detection using deep neural networks. In *2018 26th signal processing and communications applications conference (SIU)* (pp. 1–4). IEEE.

Baumann, R., Malik, K. M., Javed, A., Ball, A., Kujawa, B., & Malik, H. (2021). Voice spoofing detection corpus for single and multi-order audio replays. *Computer Speech & Language*, *65*, Article 101132.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Cai, W., Cai, D., Liu, W., Li, G., & Li, M. (2017). Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion. In *Interspeech* (pp. 17–21).

Chen, Z., Xie, Z., Zhang, W., & Xu, X. (2017). ResNet and model fusion for automatic spoofing detection. In *Interspeech* (pp. 102–106).

Cooper, D. (2013). *Speech detection using gammatone features and one-class support vector machine* (M.S. thesis), Florida, USA: Dept. Elec. Eng. & Comput. Sc., Univ. Central Florida.

De Leon, P. L., Pucher, M., Yamagishi, J., Hernaez, I., & Saratxaga, I. (2012). Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(8), 2280–2290.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer.

Escalera, S., Pujol, O., & Radeva, P. (2009). Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognition Letters*, *30*(3), 285–297.

Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2010). Learning naive Bayes classifiers for music classification and retrieval. In *2010 20th international conference on pattern recognition* (pp. 4589–4592). IEEE.

Gonçalves, A. R., Violato, R. P., Korshunov, P., Marcel, S., & Simoes, F. O. (2017). On the generalization of fused systems in voice presentation attack detection. In *2017 international conference of the biometrics special interest group (BIOSIG)* (pp. 1–5). IEEE.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5–6), 602–610.

Hang, R., Liu, Q., Song, H., & Sun, Y. (2015). Matrix-based discriminant subspace ensemble for hyperspectral image spatial–spectral feature fusion. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(2), 783–794.

Harvel, D. (2019). An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. Retrieved from https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/. Accessed August 4, 2021.

Hrabi, M. (2020). The future of voice. Retrieved from https://www.biometricupdate.com/202006/the-future-of-voice. Accessed August 8, 2021.

Hubáček, O., Šourek, G., & Železný, F. (2019). Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, *108*(1), 29–47.

[dataset] Ito, K., & Johnson, L. (2021). LJSpeech 1.1 Bonafide samples. Retrieved from https://keithito.com/LJ-Speech-Dataset/. Accessed August 1, 2021.

Janicki, A. (2017). Increasing anti-spoofing protection in speaker verification using linear prediction. *Multimedia Tools and Applications*, *76*(6), 9017–9032.

[dataset] Kang, J. (2021). LJSpeech 1.1 Spoof samples. Retrieved from https://github.com/CorentinJ/Real-Time-Voice-Cloning. Accessed August 1, 2021.

Korshunov, P., & Marcel, S. (2016). Cross-database evaluation of audio-based spoofing detection systems. In *Interspeech* (pp. 1705–1709).

Kumar, S. R., & Bharathi, B. (2021). A novel approach towards generalization of countermeasure for spoofing attack on ASV systems. *Circuits, Systems, and Signal Processing*, *40*(2), 872–889.

Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., & Shchemelinin, V. (2017). Audio replay attack detection with deep learning frameworks. In *Interspeech* (pp. 82–86).

Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., & Kozlov, A. (2019). STC antispoofing systems for the asvspoof2019 challenge. arXiv preprint arXiv:1904.05576.

Li, X., Li, N., Weng, C., Liu, X., Su, D., Yu, D., & Meng, H. (2021). Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6354–6358). IEEE.

Malik, H. (2012). Securing speaker verification system against replay attack. In *Proc. 46th AES conf. audio forensics*.

Malik, H. (2019). Securing voice-driven interfaces against fake (cloned) audio attacks. In *Proc. IEEE conf. multimedia inf. process. Retrieval* (pp. 512–517).

Malik, K. M., Malik, H., & Baumann, R. (2019). Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks. In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 523–528).

Metz, R. (2019). Amazon smart devices. Retrieved from https://edition.cnn.com/2019/09/25/tech/amazon-event/index.html. Accessed August 12, 2021.

Mishra, J., Singh, M., & Pati, D. (2018). Processing linear prediction residual signal to counter replay attacks. In *2018 international conference on signal processing and communications (SPCOM)* (pp. 95–99). IEEE.

Moeyersons, J., Varon, C., Testelmans, D., Buyse, B., & Van Huffel, S. (2017). ECG artefact detection using ensemble decision trees. In *2017 computing in cardiology (CinC)* (pp. 1–4). IEEE.

Monteiro, J., Alam, J., & Falk, T. H. (2020). Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. *Computer Speech & Language*, Article 101096.

Mwiti, D. (2019). A 2019 guide to speech synthesis with deep learning. Retrieved from https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd. Accessed August 10, 2021.

Nagarsheth, P., Khoury, E., Patil, K., & Garland, M. (2017). Replay attack detection using DNN for channel discrimination. In *Interspeech* (pp. 97–101).

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, *29*(1), 51–59.

Patel, T. B., & Patil, H. A. (2015). Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Sixteenth annual conference of the international speech communication association* (pp. 2062–2066).

Paul, D., Sahidullah, M., & Saha, G. (2017). Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2047–2051). IEEE.

Sahidullah, M., Delgado, H., Todisco, M., Kinnunen, T., Evans, N., Yamagishi, J., & Lee, K. A. (2019). Introduction to voice presentation attack detection and recent advances. In *Handbook of biometric anti-spoofing* (pp. 321–361). Cham: Springer.

Saranya, M. S., Padmanabhan, R., & Murthy, H. A. (2018). Replay attack detection in speaker verification using non-voiced segments and decision level feature switching. In *2018 international conference on signal processing and communications (SPCOM)* (pp. 332–336). IEEE.

Saratxaga, I., Sanchez, J., Wu, Z., Hernaez, I., & Navas, E. (2016). Synthetic speech detection using phase information. *Speech Communication*, *81*, 30–41.

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76–91.

Wester, M., Wu, Z., & Yamagishi, J. (2015). Human vs machine spoofing detection on wideband and narrowband data. In *Sixteenth annual conference of the international speech communication association* (pp. 2047–2051).

Witkowski, M., Kacprzak, S., Zelasko, P., Kowalczyk, K., & Galka, J. (2017). Audio replay attack detection using high-frequency features. In *Interspeech* (pp. 27–31).

Wu, Z., Xiao, X., Chng, E. S., & Li, H. (2013). Synthetic speech detection using temporal modulation feature. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7234–7238). IEEE.

Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., & … Nautsch, A. (2019). ASVspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database.

Yang, J., & Das, R. K. (2019). Low frequency frame-wise normalization over constant-Q transform for playback speech detection. *Digital Signal Processing*, *89*, 30–39.

Zhang, Y., Cao, G., Wang, B., & Li, X. (2019). A novel ensemble method for k-nearest neighbor. *Pattern Recognition*, *85*, 13–25.

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1774–1785.