IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 7, 2021, accepted November 26, 2021, date of publication December 6, 2021, date of current version December 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3133134

Voice Spoofing Countermeasure for Logical Access Attacks Detection

TUBA ARIF[1], ALI JAVED[2], (Member, IEEE), MOHAMMED ALHAMEED[3], (Member, IEEE), FATHE JERIBI[3], AND ALI TAHIR[3]

[1]Department of Software Engineering, University of Engineering and Technology, Taxila, Taxila 47050, Pakistan
[2]Department of Computer Science, University of Engineering and Technology, Taxila, Taxila 47050, Pakistan
[3]College of Computer Science and Information Technology, Jazan University, Jazan 45142, Saudi Arabia

Corresponding author: Ali Javed (ali.javed@uettaxila.edu.pk)

This work was supported by the Grant of Punjab Higher Education Commission (PHEC) of Pakistan under Award PHEC/ARA/PIRCA/20527/21.

ABSTRACT Voice-driven devices (VDDs) like Google Home and Amazon Alexa, which are well-known connected devices in consumer IoT, have applications in various domains i.e., home appliances automation, next-generation vehicles, voice banking, and so on. However, these VDDs that are based on automatic speaker verification systems (ASVs) are vulnerable to voice based logical access (LA) attacks like Text-to-Speech (TTS) synthesis and converted voice signals. Intruders can exploit these attacks to bypass the security of such systems and gain access of victim's bank account or home control. Thus, there exists a need to develop an effective voice spoofing countermeasure that can reliably be used to protect these VDDs against such malicious attacks. This work presents a novel audio features descriptor named as extended local ternary pattern (ELTP) to capture the vocal tract dynamically induced attributes of bonafide speech and algorithmic artifacts in synthetic and converted speeches. We fused our novel ELTP features with the linear frequency cepstral coefficients (LFCC) to further strengthen the capability of our features for capturing the traits of bonafide and spoofed signals. We employ the proposed ELTP-LFCC features to train the deep bidirectional Long Short-Term Memory (DBiLSTM) network for classification of the bonafide and spoof signal (i.e., TTS synthesis, converted speech). Performance of our spoofing countermeasure is measured on the large-scale and diverse ASVspoof 2019 logical access dataset. Experimental results demonstrate that the proposed audio spoofing countermeasure can reliably be used to detect the LA spoofing attacks.

INDEX TERMS Extended local ternary pattern, logical access attacks, text-to-speech synthesis, voice spoofing countermeasure, voice conversion.

I. INTRODUCTION

We have witnessed a tremendous evolution in voice biometrics-based user authentication systems in the last few years. Automatic speaker verification (ASV) systems are commonly embedded in various devices such as mobile phones, smart speakers (Google Home, Amazon Alexa), etc., for user authentication in different application domains i.e. banking, electronic-commerce systems, home automation, apps login [1], etc. For example, Siri in iPhone, Baidu's ASV in lenovo or Google Home receives voice commands from the users and execute different functions such as opening/closing doors, setting reminders, call or text some person, unlock cellphone [2], song play, etc., based on the ASV [3]. In banking sector, we observed many voice-driven based authentications solutions deployed for customers verification like Barclays Wealth and BBVA's bank in Turkey have been using the ASV to verify telephone callers. Whereas, Garanti bank has developed voice-driven interface that allows the users to perform transactions on their app by sending voice commands [4].

The COVID-19 pandemic has resulted in an exponential growth of voice-based authentication systems as lockdown and social distancing measures have restricted the ability to investigate the claimants face-to-face using facial or fingerprint recognition. This pandemic indulges the world to drastically change the verification measures by discouraging human-to-human and human-to-machine interactions

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

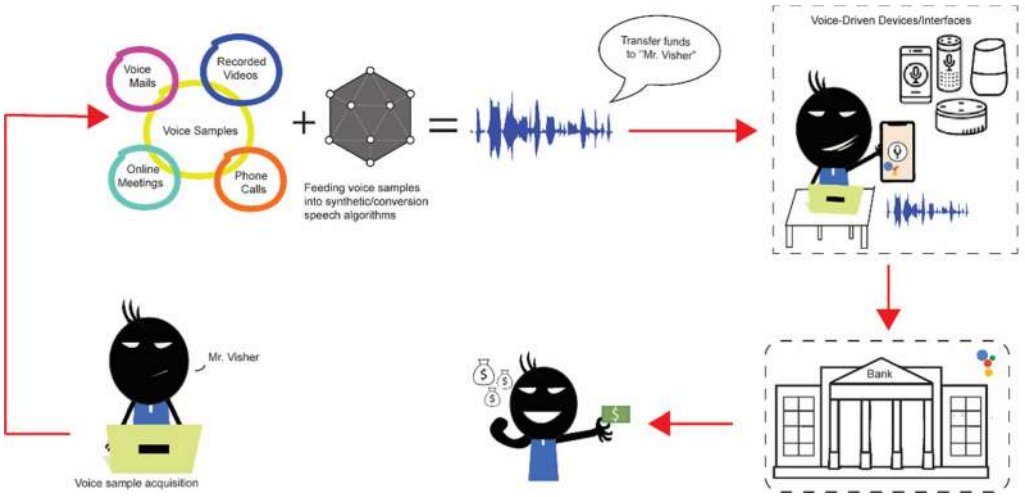

FIGURE 1. Vishing attack scenario in voice banking.

(i.e., fingerprint scanning, password-based verification, etc.). Thus, voice biometrics technology has emerged as a feasible solution among various biometric techniques (i.e., Facial, Iris and, Fingerprint). Moreover, voice biometrics-based authentication systems are considered economical and computationally more efficient over other biometrics systems. Although voice biometrics-based user authentication systems are considered more feasible these days, however, these systems are susceptible to different malicious presentation/spoofing attacks i.e., speech synthesis, voice conversion, replays, etc. These presentation attacks are used to spoof a voice biometric system by a claimant to imitate an authorized person to access the control of someone's home, bank account, device (laptop or mobile), etc. In recent times, three cases were filed in the United States where the imposters used the synthetic voice of CEO's of different organizations to fool their employees and robbed millions of dollars electronically [5]. To address the vulnerabilities of ASV systems, researchers are developing robust voice spoofing countermeasures/detection systems to add a protective layer before ASV systems that can discard the spoof sample before sending those audios to the ASV systems.

Voice spoofing attacks are categorized into logical-access attacks i.e. voice conversion (VC) [6], Text-To-Speech (TTS) synthesis [7] or physical-access attacks i.e. replays [8], impersonation [9]. These spoofing attacks are generated through modifying the bonafide audio signal into a variety of ways. For example, in voice conversion, speech signal spoken by the original speaker is manipulated to sound as if it was spoken by some target speaker while keeping the linguistic information unchanged. Speech synthesis represents the artificial/machine generated voice of the target speaker. As both the voice conversion and TTS synthesis can impersonate a target speaker's voice, thus pose a significant threat to the ASV systems. Additionally, since the converted voice originates from a live person and contains the dynamic variations of human speech as compared to speech synthesis that is void of these variations and contains cloning algorithm artifacts, therefore, we believe that detection of converted speech is more challenging. In replay spoofing, imposter plays a pre-recorded speech in front of the ASV system to get an access on behalf of the bonafide speaker.

With the advent and evolution of generative adversarial networks (GANs) in the last few years, we have witnessed amazing results in synthetic image and audio generation that looks and sounds very realistic. Shown in FIGURE 1 is one practical example of Deepfake voice phishing (Vishing), in which synthetic speech is used to impersonate Google Assistant to initiate the fraudulent transaction Mr. Visher: the intruder, searches the targeted victim and collects his/her voice samples (from online meetings, voice mails, phone calls, etc.). These voice samples are then used to train the voice conversion or speech synthesis algorithms to imitate the victim's voice. Victim's bank account is connected to voice enabled devices like Google Home for Android or iPhone devices. Equipped with the voice synthesis/conversion capability, the intruder can attempt to manipulate the VDDs. It is relatively easy for someone, potentially with malicious intentions, to get access to the VDDs by being in proximity of the victim. In this scenario, we clearly assume that the intruder has permanent or temporary access to VDDs. When attacking VDDs, the intruder can simply send synthetic/converted voice to impersonate himself as a legitimate user and exploit the VDDs into transferring funds to his account by sending the command "Hey Google, I want to transfer funds to Mr. Visher's account". Due to inability to detect the synthetic/converted speech, attacker will be successful in transferring funds into his account. This scenario shows that current VDDs are unable to differentiate between the bonafide and spoofed voice samples reliably. This demands to develop a reliable spoofing detection system for VDDs that can provide a

protective spoofing countermeasure layer in front of the ASV systems.

In recent years, various research efforts have been made to detect the spoofing (synthetic/converted) attacks [10] in conventional ASV systems to authenticate the legitimate user in financial sector [11]. The idea of voice texture is a relatively new concept of voice characterization as spectral analysis reveals that the texture of cloned voice signals varies as compared to the bonafide ones.

The concept of texture is well explored in image processing domain. Texture descriptors such as local binary patterns (LBP) and local ternary patterns (LTP) are found to be effective for texture-based classification of images. Later, this texture descriptor was introduced to develop an acoustic LBP-based voice spoofing countermeasure. LBP has two main limitations, i) noise sensitive, and ii) possibility of different LBP patterns assignment into the same class, which reduces its discriminating property. We proposed the acoustic-Local Ternary Patterns (LTP) [12] to overcome these limitations. However, acoustic-LTP features are vulnerable for certain scenarios that must be addressed. The potential limitations of this fixed threshold-based approach of our prior acoustic-LTP method are: (a) non-robust over dynamic pattern detection—spectral analysis of the synthetic voice reveals that the signal has dynamic repetition pattern that can be effectively captured using a dynamic threshold approach. However, the acoustic-LTP uses a static threshold for computing the LTP codes, therefore, there exists a need to improve the existing acoustic-LTP features for ASV applications. (b) brute-force optimization—as in acoustic-LTP we need a brute-force approach for threshold optimization, which makes it difficult to achieve better accuracy in real-time applications under diverse conditions. (c) intolerance over non-uniform noise— acoustic-LTP is robust against the consistent uniform noise that is available in the indoor audios experienced in fall detection applications, whereas we experience the non-uniform noise in the outdoor environments for applications like voice spoofing detection. Therefore, static threshold-based acoustic-LTP features are not robust under non-uniform noise and hence, not reliable for voice spoofing detection in diverse environments. The motivation behind the proposed work is to develop an effective features representation scheme that is robust to above-mentioned limitations and can reliably detect the logical-access (LA) attacks in diverse scenarios. To address these issues, we develop a novel audio features descriptor named extended local ternary pattern (ELTP) where we propose an automated threshold computation approach based on calculating the standard deviation locally for each audio frame. Our ELTP features analyze the patterns of the audios in time domain by using the dynamically computed automatic threshold approach capable of capturing the algorithmic artifacts in the synthetic speech signals and vocal tract induced variations in the genuine signals. Moreover, we exploited the ability of linear frequency cepstral coefficients (LFCC) features to effectively extract the significant information from the low- and high-frequency bands of the audios. Thus, we integrated the frequency-domain LFCC with our novel time-domain ELTP features to better enhance the features representation in terms of capturing the vocal tract induced variations of bonafide voice and algorithmic artifacts of synthesized speech. The proposed ELTP-LFCC features are later used to train a BiLSTM model to reliably detect the LA attacks. The main contributions of our research work are:

1. We propose a novel extended local ternary pattern feature descriptor to effectively capture the traits of speaker induced variations in bonafide audio and algorithmic artifacts in converted and synthetic audio.

2. Our novel ELTP features are robust to non-uniform noise and dynamic patterns detection that makes them to perform well for voice spoofing detection in diverse indoor and outdoor environmental conditions.

3. We integrated our ELTP features with the LFCC to develop a more effective descriptor that further strengthens the performance of our spoofing countermeasure.

4. Rigorous experimentation was performed to illustrate the significance of the proposed countermeasure for detection of LA based voice spoofing attacks.

The rest of the paper is organized as follows. Section II investigates the existing state-of-the-art voice spoofing countermeasures. Section III explains the proposed voice spoofing detection framework. Section IV comprises the details of dataset and experiments conducted to measure the performance of our countermeasure. Lastly, Section V presents the conclusion.

II. RELATED WORK

This section presents a critical investigation of current state-of-the-art voice spoofing countermeasures for logical-access attacks detection. Existing spoofing countermeasures have used various conventional machine learning-based [13]–[17] or deep learning-based spoofing detection systems [18]–[23] for LA attacks. The researchers in ASV and spoofing detection community have focused more on developing robust acoustic features to detect the voice spoofing attacks [13], [15], [16]. Existing works have explored a variety of audio features based on phase spectrum, magnitude spectrum, pitch, group delay, etc., to distinguish between the spoofed and human speech. Furthermore, Gaussian mixture model (GMM), its variants and support vector machine (SVM) classifiers [15]–[17], [24], [25] have been explored in various studies for voice spoofing detection.

A. SHALLOW MACHINE LEARNING-BASED APPROACHES

Existing methods have heavily explored the GMM along with different variants to develop various algorithms for synthetic speech and converted voice detection. In [13], Constant Q-transform cepstral coefficients (CQCC) were used to classify the speech samples as synthetic or bonafide. Few works have highlighted the significance of modified group delay function (MGDF) in synthetic/converted speech signals. In [14], MGDF-based and relative phase shift features were

employed for synthetic speech detection. Similarly, in [16], a features-set comprised of mel-frequency cepstral coefficients (MFCC)-cosine-normalized phase-based cepstral coefficients (CNPCC), and linear prediction residual cepstral coefficients (LPRCC) along with some existing features i.e. Modified group delay cepstral coefficients (MGDCC) and CNPCC was used to train a bi-class GMM to distinguish spoofed (synthetic/converted) speech from the bonafide. In [15], mean pitch stability (MPS), mean pitch stability range (MPSR), and jitter were computed by analyzing the pitch pattern to distinguish between the genuine and synthetic speech. The integration of multiple features makes these solutions [15], [16] computationally complex for real time applications. Few works [17], [26] have used LBP, MGDF, and CNPF to detect the LA attacks. Since the LBP is sensitive to noise that results in generation of similar patterns for both classes, thus, makes them less effective to better differentiate between the bonafide and spoof samples. Similarly, [27] highlighted the significance of relative phase information derived from Fourier spectrum and fusion of relative phase information with existing phase-based features for voice spoofing (synthetic/converted) detection. In [28], authors used the fusion of long-term modulation and short-term spectral features to discriminate between the bonafide and synthetic speech. This method uses filter-bank energies to reduce dimensionality that might result in the loss of some detailed information in modulation features. In [29], a combination of cochlear filter cepstral coefficients (CFCC) and change in instantaneous frequency (IF) was used to capture the traits of natural and spoofed (synthetic/converted) speech. The classification performance of CFCCIF features was increased when used in combination with the MFCC. An anti-spoofing system based on calculating linear predictive coding (LPC) pair-wise distances between genuine and converted speech was proposed in [30]. This countermeasure takes advantage of the prior knowledge of the attack. A spoofing countermeasure based on high-order spectral analysis specifically quadrature phase coupling (QPC), Gaussianity and linearity test statistics was used in [7] for cloned audio detection. In [31], authors investigated an utterance level feature termed as longer contexts or high level feature (HLF) and voice assessment tool (p563) which calculates Mean Opinion Score to detect artificial signals. The latter approach was unable to discriminate between the genuine and artificially produced signals effectively.

B. DEEP LEARNING-BASED APPROACHES

In recent years, ASV research community have widely explored the deep learning-based methods for logical access attacks detection. In [18], MFCC, CQCC, and STFT were employed to train the ResNet model for audio spoofing attacks detection. It was demonstrated that the fusion of three variants of residual convolutional neural networks: MFCC-ResNet, CQCC-ResNet and Spec-ResNet achieve better classification performance than the ASVspoof baseline spoofing detection methods (LFCC-GMM, CQCC-GMM). In [20], spoofing-discriminant network was employed to obtain the spoofing vector (s-vector) for each utterance. Later, mahalanobis distance with normalization was applied to s-vectors for spoofing (synthetic/converted) detection. Fusion of two magnitude-based features was used with the multilayer perceptron classifier in [19] to detect the LA attacks. This method attained improved classification performance but at higher features computation cost. In [21], a deep dense convolutional network with 135 layers was used to detect the converted voice spoofing. Similarly in [23], two low-level acoustic features i.e. log power magnitude spectra (logspec) and CQCC were employed to train the deep neural network (DNN) models based on several variants of Squeeze-Excitation network and residual networks to classify between the spoofed and bonafide speech. This method [23] achieves better classification results than other contemporary methods however, fusion of several DNN models significantly increase the training time of these methods.

Besides extracting the spectral features like MGDF, MFCC and others, which are then fed to different machine learning or deep learning model for classification, few works [22], [32], [33] have also employed machine learned features. In [32], a DNN was used to generate a bottleneck feature and frame level posteriors to discriminate between the bonafide and spoofed (synthetic/converted) samples. In this method, GMM classifier was trained using both the extracted and machine learned features. In [22], authors used the fusion of Light convolutional neural network (LCNN) and a deep feature extractor termed as Gated Recurrent neural network (GRNN). Extracted deep features were then used to train three different classifiers i.e., linear discriminant analysis (LDA), and its probabilistic version (PDLA), and SVM for voice spoofing detection. Similarly in [33], DNN-based frame-level features and RNN-based sequence-level features were extracted to train different classifiers i.e. LDA, gaussian density function (GDF), and SVM for LA attacks detection. More specifically, for DNN, authors employed three model structures that are stacked autoencoder, spoofing discriminant deep neural network (DNN), and multi-task joint-learned DNN. Whereas, in RNN-based system, LSTM-RNN and bidirectional LSTM-RNN were implemented. Autoencoder compresses the information that can result in loss of relevant content. These methods achieve better classification performance however, with increased features computation cost. TABLE 1 presents the details of existing spoofing countermeasures for LA attacks detection.

III. PROPOSED FRAMEWORK

A detailed description of our voice spoofing countermeasure is presented in this section. We proposed a novel audio feature descriptor ELTP to represent the input audio signals. The details of ELTP feature descriptor are also provided in this section. We fused our ELTP features with the LFCC for audio signal representation. We designed a bidirectional LSTM (DBiLSTM) recurrent neural network and train it

TABLE 1. Literature review summary.

Countermeasures	Method		Datasets
	Feature	Classification	
[13]	CQCC	GMM	ASvspoof2015
[14]	MGD and RPS	GMM	ASVspoof2015 and Blizzard challenge
[15]	MPS, MPSR and jitter	GMM	2008 and 2011 Blizzard Challenge, NIST2002 corpus, Switchboard corpus, Resource Management Corpus, Wall Street Journal corpus.
[16]	MFCC-CNPCC, LPRCC, MGDCC and CNPCC	GMM	ASVspoof2015
[17]	LBP, MGDF, and CNPF	SVM	ASVspoof2015
[24]	LBP	Five different ASV systems 1. GMM-UBM 2. GMM Supervector Linear Kernel 3. GMM Supervector Linear Kernel-Nuisance attribute projection 4. Factor Analysis 5. GMM Supervector Linear Kernel - Factor Analysis	NIST05, NIST06
[27]	RP, MFCC, MGD,	GMM	ASVspoof2015
[28]	Phase modulation, magnitude modulation, MFCC, MGD	GMM	Wall Street Journal corpora (WSJ0+WSJ1)
[29]	CFCC, CFCCIF, MFCC	GMM	ASVspoof2015
[30]	LPC	Calculated pair-wise distances of bonafide and converted voice samples.	male subset of the NIST'05 dataset and NIST'06 dataset.
[31]	HLF and voice assessment tool (p563)	ASV systems using combinations of different parameterizations and classifiers: 1. GMM_m33 2. FA_m33 3. GMM_m50 4. FA_m50 5. SVM_m33	NIST
[7]	QPC, Gaussianity and linearity test statistics		Baidu Silicon Valley AI Lab cloned audio dataset
[18]	MFCC, CQCC, and STFT	Residual convolutional network	ASVspoof2019
[19]	LMS, RLMS, GD, MGD, IF, BPD, and PSP	Multilayer perceptron	ASVspoof 2015
[20]	Deep neural network as feature extractor	Mahalanobis distance	ASVspoof 2015
[21]	Deep features are automatically extracted by the 135-layer DenseNet		Timit, NIST, UME
[23]	logspec and CQCC	variants of squeeze-excitation and residual networks	ASVspoof2019
[22]	LC-GRNN (feature extractor)	LDA, PDLA, and SVM	ASVspoof2019
[32]	DNN bottleneck feature and frame level posteriors	GMM	ASVspoof 2015
[33]	DNN-based frame-level features and RNN-based sequence-level features	LDA, GDF, and SVM	ASVspoof 2015

by using our ELTP-LFCC features for classification of the bonafide and synthetic/converted signals. The architecture of the proposed framework is presented in FIGURE 2.

A. FEATURE EXTRACTION

For accurate detection of logical access attacks, we need to develop a robust audio feature descriptor that can effectively capture the algorithmic artifacts in synthesized signals and dynamic speech attributes of human speaker in the bonafide speech. Moreover, audio features must also be robust over non-uniform noise which is quite apparent in the outdoor environments where voice samples can be recorded and later used for voice spoofing detection. To better address these concerns, we proposed a novel audio features representation method ELTP that is robust to non-uniform noise and dynamic patterns detection, and capable of capturing the dynamic varying attributes in genuine speech and algorithmic artifacts in the cloned voice. Moreover, we integrated the LFCC features with our ELTP features to further enhance the performance of spoofing detection.

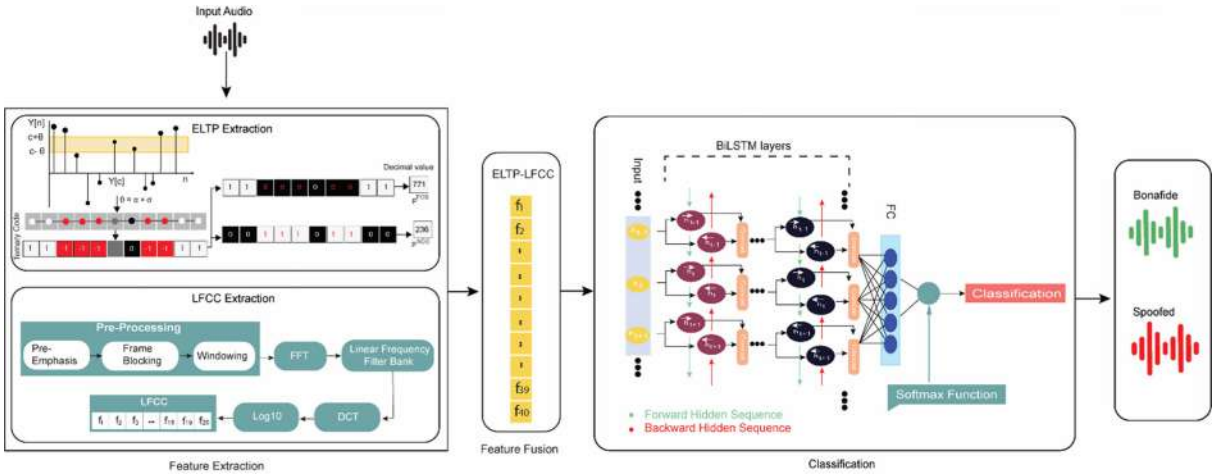

FIGURE 2. Architecture of proposed framework.

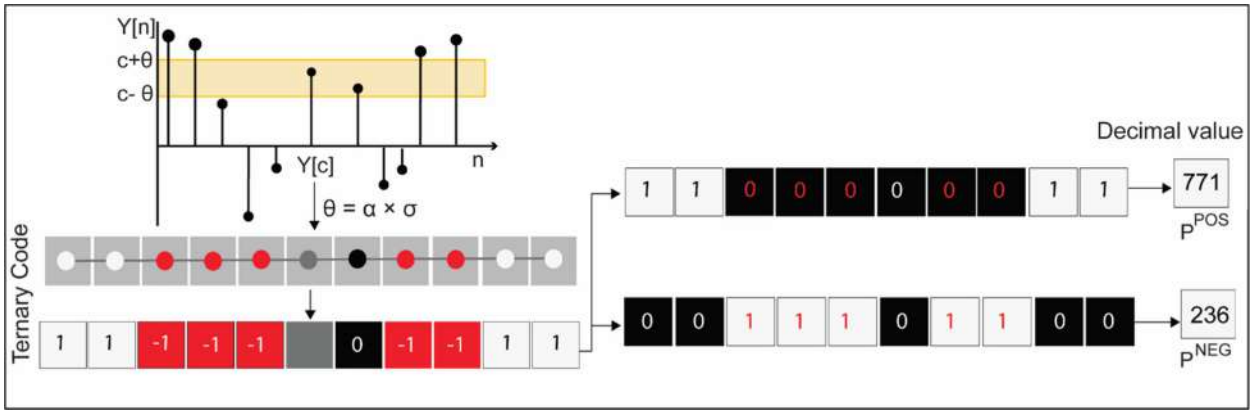

FIGURE 3. Illustration of ELTP descriptor.

1) EXTEND LOCAL TERNARY PATTERNS (ELTP)

We partition the input audio signal $Y[n]$ having N samples into non-overlapping frames of length l. The concept to generate the ELTP features are taken from the image processing research that consider the closest neighborhood of a pixel comprising of the 8 surrounding pixels in a 3×3 window for 2D LTP features [34]. However, for 1D audio signals, we employed different number of neighbors and found better features representation with 10 neighbors. Thus, we selected 10 neighbors around a central sample c to create each frame of length 11 (FIGURE 3) in the input audio. LTP extends LBP to 3 valued codes, which quantize the width $\pm\theta$ around c, ones above and below this are quantized to 1 and -1 respectively. The process for converting a region into its ELTP representation is as follows:

$$P\left(s^i, \mathrm{c}, \theta\right) = \begin{cases} 1, & s^i \geq c + \theta \\ 0, & |(s^i - \theta)| < \theta \\ -1, & s^i \leq (\mathrm{c} - \theta) \end{cases} \tag{1}$$

where $P(s^i, c, \theta)$ represents the acoustic signal, c is the central sample of the frame F with s^i neighbors where i represents neighbor index and θ represents the threshold. To compute the ELTP, we compute the magnitude difference between central sample c and the 10 surrounding audio samples s^i by applying θ around the c. In our prior work of 1D LTP features [12], we used the fixed threshold that is not much robust to noise. To overcome this limitation, we develop an automatic scheme to calculate the threshold dynamically using an auto-adaptive method instead of using a fixed threshold θ. We computed this auto-adapted threshold as follows:

$$\theta = \alpha \times \sigma (0 < \alpha \leq 1) \tag{2}$$

where σ is the standard deviation computed for each frame of the audio, and α is a scaling factor. We employed a linear searching mechanism to optimize the value of the scaling factor α by finding the convergence point between 0 and 1. We found this optimized value of α =0.6 as we achieved the best results on this value. Thus, we used α =0.6 for computing the threshold θ.

Next, we split each ternary pattern of ELTP into its positive (P^{POS}) and negative (P^{NEG}) halves. All values quantized to +1 or -1 are maintained in P^{POS} and P^{NEG} respectively. And replacing all other values with zeros using the (3) and (4):

$$P^{POS}\left(s^i, c, \theta\right) = \begin{cases} 1, & if\ P\left(s^i, c, \theta\right) = +1 \\ 0, & Otherwise \end{cases} \tag{3}$$

$$P^{NEG}\left(s^i, c, \theta\right) = \begin{cases} 1, & if\ P\left(s^i, c, \theta\right) = -1 \\ 0, & Otherwise \end{cases} \tag{4}$$

Inspired from the concept of uniform patterns in image processing research [35], we used this idea for voice signals since they provide valuable information about the signal. In contrast to non-uniform patterns, which provide less significant signal information, uniform patterns include substantial signal information. It is also worth noting that uniform patterns are more prevalent than non-uniform patterns. We computed positive uniform $ELTP_u^{POS}$ and negative uniform $ELTP_u^{NEG}$ patterns from the earlier mentioned P^{POS} and P^{NEG} and represented these patterns in decimal forms using the (5) and (6) as follows:

$$ELTP_u^{POS}(s^i, c, \theta) = \sum_{i=0}^{9} 2^i \times P_u^{POS}\left(s^i, c, \theta\right) \tag{5}$$

$$ELTP_u^{NEG}(s^i, c, \theta) = \sum_{i=0}^{9} 2^i \times P_u^{NEG}\left(s^i, c, \theta\right) \tag{6}$$

Next, we compute the histogram of $ELTP_u^{POS}$ and $ELTP_u^{NEG}$ separately to obtain the details of both the patterns. The number of bins is significantly reduced by assigning all non-uniform patterns to one bin, without losing too much data. Histograms are calculated as follows:

$$h^{POS}(ELTP^{POS}, n) = \sum_{k=1}^{k} \left(ELTP_k^{POS}, n\right) \tag{7}$$

$$h^{NEG}(ELTP^{NEG}, n) = \sum_{k=1}^{k} \left(ELTP_k^{NEG}, n\right) \tag{8}$$

Here n is the histogram bins. Through extensive experimentation, we concluded that first 10 uniform patterns from both categories were sufficient to capture the distinctive traits in bonafide and spoof samples. Therefore, we used the 10-dimensional ELTP code each for positive and negative uniform patterns. Finally, (7) and (8) are concatenated to create a 20-dimensional ELTP features as follows:

$$ELTP = [h^{POS} + h^{NEG}] \tag{9}$$

2) LINEAR FREQUENCY CEPSTRAL COEFFICIENTS (LFCC)

Recently, we have seen many methods that employed different spectral features alone or in combination for voice spoofing detection. Spectral features such as MFCC, GTCC, CQCC, etc., have been employed to develop features representation schemes for anti-spoofing methods [36], [37]. ASVspoof community have provided two baseline models, one using the CQCC and other the LFCC for physical- and logical-access attacks detection. MFCC features were proposed based on the resemblance to human auditory system. LFCC is identical to MFCC in terms of feature extraction computation but with the difference of linear filter bank. Furthermore, according to speech production theories, some characteristics of speaker associated with the anatomy of vocal tract are greatly reflected in high frequency areas of the speech [38]. This argues the use of linear scale frequency for speaker identification and spoofing detection. Additionally, a comparative analysis in [39] performed for synthetic speech detection demonstrates the effectiveness of LFCC in terms of capturing the distinctive traits available in the high-frequency bands over other cepstral coefficients. This fact motivated us to integrate the LFCC with our novel ELTP features to better capture the vocal tract induced variations of bonafide voice and algorithmic artifacts of synthesized speech. For this work, we extracted the 20-dimensional LFCC features with software implementation in MATLAB provided by the ASVspoof 2019 challenge [10] and fused them with our ELTP features for acoustic signal representation. The process of LFCC extraction that returns 20 dimensional LFCC coefficients is represented in FIGURE 4. Pre-processing block includes the tasks of framing and windowing. We obtained the spectrum of each audio frame using the Fast Fourier Transform. A set of linear filters is applied to Fast fourier transform of the audio signals and gain (g_k) is calculated. Next, log of each $g_k$is calculated and discrete cosine transform is applied to obtain the LFCC features. LFCC features are calculated as follows:

$$\sum_{k=1}^{K} log(g_k) cos(\frac{(2k-1)i\pi}{2K}), \quad 1 \le i \le I \tag{10}$$

where, K and I represent the number of filters and number of LFCC respectively. We integrated this 20-dimensional LFCC features with our 20-dimensional ELTP features to create the final 40-dimensional ELTP-LFCC feature vector.

B. CLASSIFICATION

Since audio is a time series signal and BiLSTM is suitable for extracting the time series data. Therefore, we employed the BiLSTM model in this work for classification of the bonafide and spoofed samples. We used our ELTP-LFCC features-set to train the BiLSTM classifier for logical-access attacks detection. For a given input sequence $x = [x_1, x_2, \ldots, x_T]$, RNN computes the hidden vector $h = [h_1, h_2, \ldots, h_T]$ and the output vector $y = [y_1, y_2, \ldots, y_T]$ by iterating the (11) and (12) from $t = 1$ to T:

$$h_t = H\left(W_{xh}x_t + W_{hh}h_{t-1} + B_h\right) \tag{11}$$

$$y_t = W_{hy}h_t + b_y \tag{12}$$

where W represents the weight matrices (e.g., W_{xh} is the input-hidden weight matrix), B is the bias vectors (e.g., B_h is hidden bias vector) and H is the hidden function. For the

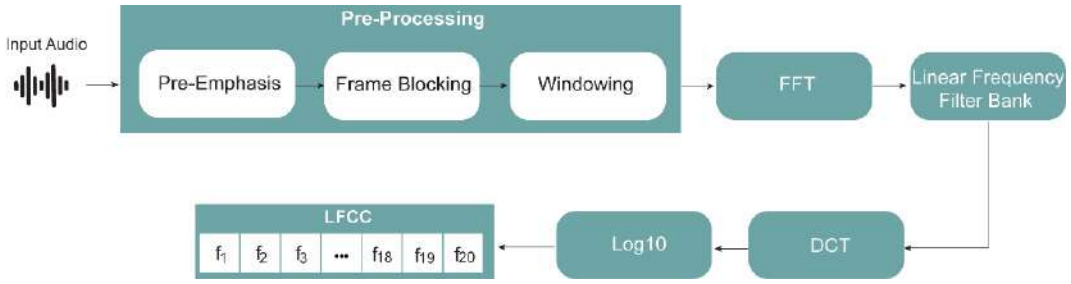

FIGURE 4. Illustration of LFCC descriptor.

LSTM network, we computed the hidden function at time t using (13) to (17) as follows:

$$f_t = \sigma_g \left(W_{xf} \times x_t + W_{hf} \times h_{t-1} + W_{cf} \times c_{t-1} + B_f\right) \quad (13)$$

$$i_t = \sigma_g \left(W_{xi} \times x_t + W_{hi} \times h_{t-1} + W_{ci} \times c_{t-1} + B_i\right) \quad (14)$$

$$o_t = \sigma_g \left(W_{xo} \times x_t + W_{ho} \times h_{t-1} + W_{co} \times c_t + B_o\right) \quad (15)$$

$$c_t = f_t c_{t-1} + i_t tanh \left(W_{xc} \times x_t + W_{hc} \times h_{t-1} + B_c\right) \quad (16)$$

$$h_t = o_t tanh(c_t) \quad (17)$$

where σ_g is the hard-sigmoid function, f, i, o, c and h are forget gate, input gate, output gate, cell memory, and hidden vector respectively.

LSTM is also commonly employed for classification of time series data, however, LSTM network is limited as it uses the previous context only. Bidirectional RNN (BRNN) [40] successfully overcomes this issue by accessing the data in both directions. Thus, we employed the BiLSTM network in the proposed method. As illustrated in FIGURE 2, a forward hidden sequence $\vec{h}$, backward hidden sequence $\overleftarrow{h}$ and output sequence is computed by iterating the forward layer from $t = 1$ to T, backward layer from $t = T$ to 1. The output layer is updated by concatenating the outputs of forward and backward hidden sequences as follows:

$$y_t = W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + B_y \quad (18)$$

Our model used 10 bidirectional LSTM layers, each with 64 hidden units. Extracted ELTP-LFCC features are fed to the first BiLSTM layer. The outputs of one BiLSTM layer are concatenated and passed to the next BiLSTM layer. Feature vector from the 10^{th} BiLSTM layer is passed into a fully connected (FC) layer. The output of FC layer is propagated to a softmax layer and finally to a classification layer that assigns each input to one of the mutually exclusive classes as shown in FIGURE 2. We used Adam optimizer [41] to tune our network with initial learning rate set to 0.001 and squared gradient decay factor set to 0.999. We tuned various parameters during the network training. Specifically, we tunned state and gate activation functions, mini-batch size, maximum epochs and number of hidden units. We performed experiments by setting number of hidden units equal to 64,100 and 150 and found best results with 64 hidden units. For network training, mini-batch size was tuned at values of 128, 64 and 30 and received best results on 30 mini-batch size. Maximum number of epochs was set to different values and finally selected as 100 epochs as optimal results were achieved on this setting. We also tuned the system on *tanh* and *soft sign* for state activation function where *tanh* outperforms the *soft sign* in almost all experiments, as *tanh* delivers better training performance for multilayer neural networks [42]. Similarly, we tuned the system on sigmoid and hard-sigmoid for gate activation function and found best results on the hard-sigmoid.

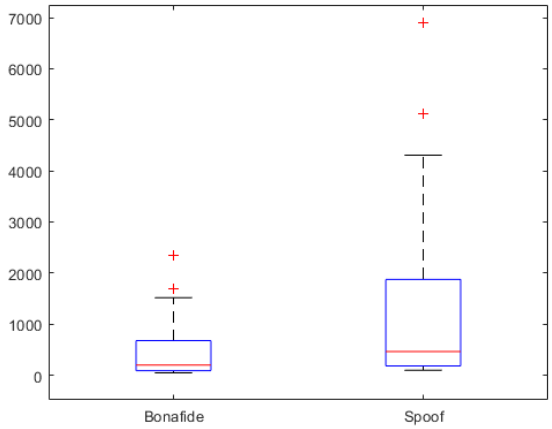

FIGURE 5. Distributional variance of bonafide and spoof feature values.

C. ADDRESSING THE LIMITATIONS OF ACOUSTICS-LBP AND ACOUSTICS-LTP APPROACHES

As we discussed in the Introduction section, existing approaches like acoustics-LBP are sensitive to noise and hard-coded threshold-based acoustics-LTP features are non-robust over dynamic pattern detection that makes it difficult to achieve better accuracy in real-time applications under diverse conditions. As the proposed ELTP features analyze the patterns of audios in time domain, therefore, reliably captures the algorithmic artifacts of synthetic samples and dynamic vocal tract traits of the genuine audios for effective classification of the genuine/bonafide and cloned samples. To demonstrate the effectiveness of our ELTP features for distinctive representation of bonafide and synthetic/cloned sample, we generated the box plots of ELTP for the bonafide and synthetic samples of the same speaker as shown in FIGURE 5. From the FIGURE 5, we can see that the

TABLE 2. Statistics of training, development, and evaluation subsets of ASVSPOOF 2019 LA dataset.

Subset	Audio Samples		Human Speakers		Spoofing Algorithms			Sampling Rate	Spoofing Systems
	Logical Access		Total: 107		VC	TTS	VC and TTS		
	Bonafide	Spoof	Male	Female					
Training	2,580	22,800	8	12	2	4	×	16 kHz	Known: 6 Unknown: 11
Development	2,548	22,296	8	12					
Evaluation	7,355	63,882	30	37	5	11	3		

spoof sample has a larger distributional variance over the bonafide sample of the same speaker. Moreover, most of the feature values of spoof samples are high as compared to the bonafide samples. These facts signify the effectiveness of our ELTP features for more distinctive representation of the bonafide and spoof samples. Moreover, our ELTP features also address the limitation (non-robustness against noise) of acoustics-LBP approach. We can prove from FIGURE 3 that our proposed ELTP features are robust against the noise. As the noise can enhance or reduce the value of central sample within a frame resulting in generation of wrong code, however, we can see from the audio frame shown in FIGURE 3 that the value of c now remains within the $c+\theta$ and c-θ range, thus, achieves more robustness against the noise.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the details of experiments conducted to measure the performance of our technique. We also provided a discussion on the results of these experiments. Moreover, details of the dataset used for performance evaluation is also presented. The evaluation plan of ASVspoof 2019 dataset considers tandem detection cost function (t-DCF) and equal error rate (EER) as primary and secondary evaluation metrics, respectively. Thus, we also used the t-DCF and EER to measure the performance of the proposed countermeasure. For experimentation, we used the training subset of ASVspoof 2019 LA dataset for training and evaluation subset for testing.

A. DATASET

Performance of our proposed countermeasure is investigated on the logical access subset of ASVspoof 2019 dataset. It comprises training, development, and evaluation subsets. Each subset contains bonafide and spoofed samples where spoofed samples are generated from genuine speech samples using several spoofing algorithms (A01–A19) [43]. Genuine speech samples are collected from 107 speakers. The training subset contains 25,380 samples, the development subset contains 24,986 samples, and the evaluation (eval) subset contains 71,933 audio samples. The statistics of ASVspoof 2019 LA dataset in terms of number of spoofed and bonafide samples in each subset, number of male and female speakers, spoofing algorithms, and sampling rate are listed in TABLE 2. The duration of each utterance is in the range of one to two seconds and all audio files in these three subsets are stored in flac format. The details can be found at [43].

TABLE 3. Detection performance on synthetic speech, voice conversion, and combined LA-EVAL subset.

Spoofing Category	t-DCF	EER (%)
Voice conversion	0.39	33.28
Synthetic speech	0.00002	0.002
Overall LA-eval subset	0.008	0.74

B. PERFORMANCE EVALUATION OF ELTP AND LFCC FEATURES

We performed an experiment to investigate the performance of our proposed ELTP features, LFCC features and ELTP-LFCC features fusion for LA spoofing detection. For this, we employed the proposed ELTP features, LFCC features and ELTP-LFCC features separately to train the DBiLSTM model for LA attacks detection. We achieved an EER and t-DCF of 2.45% and 0.067, 19.85% and 0.409, and 0.74% and 0.008 on ELTP, LFCC, and ELTP-LFCC features respectively. These results show that our proposed ELTP features achieved remarkable performance for LA spoofing detection. LFCC performed the worst, whereas, ELTP-LFCC features fusion performed the best by achieving the lowest t-DCF of 0.008 for LA evaluation corpus. Thus, we used the ELTP-LFCC features to train the DBiLSTM model for LA attacks detection.

C. PERFORMANCE EVALUATION OF PROPOSED COUNTERMEASURE

We designed an experiment to measure the performance of our countermeasure for voice conversion, TTS synthesis, and overall LA spoofing detection. For this, we employed the proposed ELTP-LFCC features set to train the DBiLSTM model for voice conversion, TTS synthesis, and overall LA spoofing detection separately. The results of this experiment are presented in TABLE 3. We achieved an EER and t-DCF of 33.28% and 0.39, 0.002% and 0.00002, and 0.74% and 0.008 for voice conversion, synthetic speech, and overall LA evaluation subset respectively. From the results presented in TABLE 3, we can see that the proposed system attains better performance on synthetic speech over the converted voice samples. This might be due to the reason that input resource in speech synthesis is text in digitized form which is converted into speech. Whereas VC systems (A05, A06, A17, A18, A19) use human speech as source and preserve the prosodic qualities of the speaker which might be missing in the synthetic speech. Since A04 and A16 are unit-selection

TABLE 4. Performance comparison of proposed feature-set on different classifiers.

Spoofing Category	Classifier	t-DCF	EER (%)
Voice Conversion	SVM (RBF)	0.39	33.30
	LSTM	0.374	31.97
	DBiLSTM	0.39	33.28
Synthetic Speech	SVM (RBF)	0.21	17.30
	LSTM	0.264	22.59
	DBiLSTM	0.00002	0.002
Overall LA-eval subset	SVM (RBF)	0.12	10.59
	LSTM	0.119	10.22
	DBiLSTM	0.008	0.74

based TTS systems, they may preserve the acoustic features of bonafide audios. However, these systems might fail if the required segment (phrase/word) is missing from the database. Such strong dependence on the bonafide signal makes this approach less effective. Overall, our spoofing detection framework achieves remarkable performance for LA attacks detection that illustrates the effectiveness of our method for reliable LA spoofing detection.

D. PERFORMANCE COMPARISON ON DIFFERENT CLASSIFIERS

We performed an experiment to investigate the classification performance of the proposed ELTP-LFCC features against different classifiers. For this purpose, we used our ELTP-LFCC features to train the conventional machine learning and deep learning classifiers. More specifically, we employed our features to train the SVM, LSTM and DBiLSTM separately and reported the results in the TABLE 4. Again, we used the training set of ASVspoof LA dataset for training and evaluation subset for testing. We tuned these classifiers on different settings and selected the parameters where we obtained the best results.

For the synthetic speech detection, the DBiLSTM classifier achieved the best results, whereas, SVM obtained the highest EER and t-DCF. For voice conversion, all the classifiers achieved almost similar performance where the LSTM performed marginally better than the rest. For the overall LA-eval collection, the DBiLSTM achieved the best results with a significant margin as compared to the LSTM and SVM classifiers. From the results of this experiment, we can conclude that the proposed ELTP-LFCC features offer the best performance with the DBiLSTM model. Thus, we used the DBiLSTM model for classification of the bonafide and spoof audios.

E. PERFORMANCE COMPARISON AGAINST EXISTING LA SPOOFING DETECTION METHODS

To measure the robustness of the proposed method for LA spoofing detection, we performed a comparative analysis of our method against existing LA spoofing detection methods including the baseline methods provided by ASVspoof challenge. For this purpose, we compared our method with these existing methods [18], [22], [37] on the ASVspoof 2019 LA dataset for LA spoofing detection. The results of the proposed and comparative methods in terms of t-DCF and EER are provided in TABLE 5.

TABLE 5. Comparative analysis of proposed and existing LA detection methods.

Systems		EER (%)	t-DCF
[43]	CQCC+GMM (ASVspoof baseline)	11.04	0.2454
	LFCC+GMM (ASVspoof baseline)	13.54	0.3017
[18]	MFCC+ResNet	9.33	0.2042
	Spec+ResNet	9.68	0.2741
	CQCC+ResNet	7.69	0.2166
	Fusion	6.02	0.1569
[22]	LC-GRNN + SVM	7.12	0.1873
	LC-GRNN + PLDA	6.34	0.1552
	LC-GRNN + LDA	6.28	0.1523
[37]	Contrastive-1 (CQCC, ICQCC, eCQCC, IFCC, CQ-EST, CQ-OST, CQSPIC, eCQ-OST, and CMC + DNN / GMM)	4.13	0.1246
	Primary (CQCC, ICQCC, eCQCC, IFCC, CQ-EST, CQ-OST, CQSPIC, and CMC +DNN)	8.82	0.2417
	Single (eCQCC + DNN)	11.08	0.2852
Proposed Method	ELTP-LFCC+DBiLSTM	0.74	0.008

TABLE 6. Performance comparison of baseline and proposed features on DBILSTM.

Spoofing Category	Baseline				Proposed	
	CQCC		LFCC		ELTP-LFCC	
	EER (%)	t-DCF	EER (%)	t-DCF	EER (%)	t-DCF
Voice Conversion	30.61	0.339	46.5	0.714	33.28	0.39
Synthetic Speech	12.78	0.2850	0.04	0.0002	0.002	0.00002
Overall LA-eval subset	7.54	0.208	19.85	0.409	0.74	0.008

From the classification results, we can see that the proposed countermeasure outperforms the existing methods including the ASVspoof baseline methods for LA attacks detection. Thus, we argue that our method can effectively be used to detect the LA voice spoofing attacks.

F. PERFORMANCE COMPARISON OF PROPOSED ELTP-LFCC AND BASELINE FEATURES FOR LA SPOOFING DETECTION

Since we proposed a novel features descriptor for voice spoofing detection, therefore, features wise comparison against the existing baseline features (CQCC and LFCC) on the same classifier is important to evaluate the significance of our ELTP-LFCC features set. For this, we compared the performance of our features against the ASVspoof baseline features CQCC and LFCC on DBiLSTM classifier and results

are shown in TABLE 6. From the results, we can observe that the proposed ELTP-LFCC features provide better detection performance over the CQCC and LFCC alone when trained with the DBiLSTM on the ASVspoof LA dataset as a whole. More specifically, we achieved lesser t-DCF of 0.2 and EER of 6.8% as compared to CQCC, and 0.401 and 19.11% as compared to LFCC for LA attack's detection. However, CQCC performs better on the converted voice samples. These results demonstrate the effectiveness of the proposed ELTP-LFCC features for LA attacks detection.

V. CONCLUSION

This paper has presented an effective voice spoofing countermeasure using the novel ELTP-LFCC features and Deep Bidirectional LSTM to combat the TTS synthesis and converted voice samples of logical-access attacks. We presented a novel audio features descriptor ELTP and fused it with LFCC to better capture the characteristics of the vocal tract speech dynamics of bonafide voice and cloning algorithm artifacts. Performance evaluation on the diverse ASVspoof 2019-LA dataset demonstrates the significance of our system for reliable detection of logical access spoofing attacks. Performance comparison against the baseline and existing contemporary methods shows that our spoofing countermeasure provides better detection performance over the existing voice spoofing countermeasures. The fact that the ASVspoof evaluation set contains the unknown bonafide and spoof samples and voice samples of unseen human speakers indicates that our system can provide better performance on cross-dataset scenario. Experimental analysis showed encouraging results on TTS synthesis attacks however, we found that converted voice samples are more difficult to detect due to the fact that voice conversion algorithms take voice samples as input over the TTS which takes the digitized text as input. This makes the voice conversion algorithms to better preserve the prosodic qualities of the speaker in synthesized samples which might be missing in the synthetic speech generated using the TTS algorithms. In the future, we plan to improve the performance of our countermeasure against the voice conversion attacks.

REFERENCES

[1] *Voiceprint: The New WeChat Password*. Accessed: Apr. 4, 2015. [Online]. Available: https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/

[2] S. Millward. *Open Sesame: Baidu Helps Lenovo Use Voice Recognition to Unlock Android Phones*. Accessed: Mar. 12, 2012. [Online]. Available: https://www.techinasia.com/baidu-lenovo-voice-recognition-android-unlock

[3] A. Vigderman. *What is Home Automation and How Does it Work*. Accessed: Apr. 16, 2017. [Online]. Available: https://www.security.org/home-automation/

[4] L. Fernández. *EFMA Recognizes Garanti Bank's Mobile Voice Assistant*. Accessed: Mar. 3, 2017. [Online]. Available: https://www.bbva.com/en/efma-recognizes-garanti-banks-mobile-voice-assistant/

[5] D. Harwell. *An Artificial-Intelligence First: Voice-Mimicking Software Reportedly Used in a Major Theft*. Accessed: Jan. 5, 2019. [Online]. Available: https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/

[6] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," 2018, *arXiv:1804.04262*.

[7] H. Malik, "Securing voice-driven interfaces against fake (Cloned) audio attacks," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 512–517.

[8] T. Gunendradasan, S. Irtza, E. Ambikairajah, and J. Epps, "Transmission line cochlear model based AM-FM features for replay attack detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6136–6140.

[9] R. G. Hautamáki, T. Kinnunen, V. Hautamáki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: On vulnerability of speaker verification systems against voice mimicry," in *Proc. Interspeech*, 2013, pp. 930–934.

[10] *Automatic Speaker Verification, Spoofing and Countermeasures Challenge*. Accessed: Sep. 7, 2021. [Online]. Available: https://www.asvspoof.org/index2019.html

[11] J. Kollewe. *HSBC Rolls Out Voice and Touch ID Security for Bank Customers*. Accessed: Mar. 20, 2021. [Online]. Available: https://www.theguardian.com/business/2016/Feb./19/hsbc-rolls-out-voice-touch-id-security-bank-customers

[12] S. M. Adnan, A. Irtaza, S. Aziz, M. O. Ullah, A. Javed, and M. T. Mahmood, "Fall detection through acoustic local ternary patterns," *Appl. Acoust.*, vol. 140, pp. 296–300, Nov. 2018.

[13] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Odyssey*, 2016, pp. 283–290.

[14] I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas, "Synthetic speech detection using phase information," *Speech Commun.*, vol. 81, pp. 30–41, Jul. 2016.

[15] P. L. D. Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. Interspeech*, Sep. 2012, pp. 370–373.

[16] M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in *Proc. Interspeech*, Sep. 2015, pp. 1–5.

[17] Y. Liu, Y. Tian, L. He, J. Liu, and M. T. Johnson, "Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2015, pp. 1–5.

[18] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," 2019, *arXiv:1907.00501*.

[19] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. Interspeech*, Sep. 2015, pp. 2052–2056.

[20] N. Chen, Y. Qian, H. Dinkel, B. Chen, and K. Yu, "Robust deep feature for spoofing detection—The SJTU system for ASVspoof 2015 challenge," in *Proc. Interspeech*, Sep. 2015, pp. 1–5.

[21] Y. Wang and Z. Su, "Detection of voice transformation spoofing based on dense convolutional network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2587–2591.

[22] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection," *Proc. Interspeech*, 2019, pp. 1068–1072.

[23] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: Anti-spoofing with squeeze-excitation and residual networks," 2019, *arXiv:1904.01120*.

[24] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Lyon, France, 2013, p. 5.

[25] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2012, pp. 1–6.

[26] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Proc. Interspeech*, Lyon, France, 2013, pp. 1–5.

[27] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1–5.

[28] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7234–7238.

[29] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1–5.
[30] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3068–3072.
[31] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1–5.
[32] M. J. Alam, P. Kenny, V. Gupta, and T. Stafylakis, "Spoofing detection on the ASVspoof2015 challenge corpus employing deep neural networks," in *Proc. Odyssey*, 2016, pp. 270–276.
[33] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Commun.*, vol. 85, pp. 43–52, Dec. 2016.
[34] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.
[35] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.
[36] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A light-weight replay detection framework for voice controlled IoT devices," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 982–996, Aug. 2020.
[37] R. K. Das, J. Yang, and H. Li, "Long range acoustic features for spoofed speech detection," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1058–1062.
[38] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Jan. 2011, pp. 559–564.
[39] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 2087–2091.
[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[41] D. P. Kingma and J. Ba, "Adam: A methodfor stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[42] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018, *arXiv:1811.03378*.
[43] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. Lee, V. Vestman, and A. Nautsch, "ASVspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database," Centre Speech Technol. Res., Univ. Edinburgh, Tech. Rep., 2019, doi: 10.7488/ds/2555.

TUBA ARIF received the bachelor's degree in software engineering from UET Taxila, Pakistan, in 2018, where she is currently pursuing the M.S. degree with the Department of Software Engineering. Her research interests include deep learning, machine learning, and audio forensics.

ALI JAVED (Member, IEEE) received the B.Sc. degree (Hons.) (third position) in software engineering and the M.S. and Ph.D. degrees in computer engineering from UET Taxila, Pakistan in 2007, 2010, and 2016, respectively.

He is currently serving as an Associate Professor for the Computer Science Department, UET Taxila. Previously, he served as an Assistant Professor for the Software Engineering Department, UET Taxila. He has also served as the HOD for the Software Engineering Department, UET Taxila, in 2014. He served as a Postdoctoral Scholar for the SMILES Laboratory, Oakland University, USA, in 2019, and a Visiting Ph.D. Scholar for the ISSF Laboratory, University of Michigan, USA, in 2015. His research interests include digital image processing, computer vision, multimedia forensics, video content analysis, machine learning, and multimedia signal processing.

Dr. Javed has been a member of Pakistan Engineering Council, since 2007. He was a recipient of various research grants from HEC Pakistan, National ICT Research and Development Fund, NESCOM, and UET Taxila. He received the Chancellor's Gold Medal for his M.S. degree in computer engineering. He got selected as the Ambassador of the Asian Council of Science Editors from Pakistan, in 2016.

MOHAMMED ALHAMEED (Member, IEEE) received the M.Sc. degree in computer science from the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, in 2013, and the Ph.D. degree in computer science from the Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, in December 2018. He is currently an Assistant Professor at the College of Computer Science and Information Technology, Jazan University, Saudi Arabia. Moreover, his research interests include computer engineering and networks, intelligent systems, vehicular *ad hoc* networks, and beaconing protocol design. He is a member of ACM.

FATHE JERIBI received the B.S. degree in information systems from Jazan University, Jazan, Saudi Arabia, in 2010, the M.S. degree in computer science and information technology from Sacred Heart University, CT, USA, in 2014, and the Ph.D. degree in information technology from Towson University, MD, USA, in 2018. Moreover, he has worked as an Adjunct Faculty with Towson University for two years. He is an Assistant Professor at the College of Computer Science and Information Technology, Jazan University. His research interests include computer networks, SDN, software engineering, machine learning, wireless *ad hoc* networks, and distributed computing.

ALI TAHIR received the B.S. degree in computer engineering and the M.S. degree in telecom engineering from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2006 and 2010, respectively, and the Ph.D. degree in computer science from COMSATS University Islamabad (CUI), Wah Campus, Pakistan, in 2021. He served at Nokia Siemens Network for two years as a BSS Engineer. He also served at SCB and AHQ, Islamabad, as a Network Engineer. Currently, he is working as a Senior Lecturer at the College of Computer Science and Information Technology, Jazan University, Jazan, Saudi Arabia. His research interests include wireless networks, distributed systems, software defined networking, network security, machine learning, and software engineering.