Journal of King Saud University – Computer and Information Sciences 34 (2022) 7300–7312

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

A robust voice spoofing detection system using novel CLS-LBP features and LSTM

Hussain Dawood [a,*], Sajid Saleem [a], Farman Hassan [b], Ali Javed [b]

[a] Department of Computer and Network Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 23890, Saudi Arabia
[b] Department of Software Engineering, University of Engineering and Technology Taxila, 47050 Punjab, Pakistan

ARTICLE INFO

Article history:
Received 22 October 2021
Revised 23 February 2022
Accepted 26 February 2022
Available online 22 March 2022

Keywords:
ASV system
Center-Lop-Sided Local Binary Patterns
LSTM
Logical access attacks
Physical access attacks
Voice spoofing detection

ABSTRACT

Automatic Speaker Verification (ASV) systems are vulnerable to a variety of voice spoofing attacks, e.g., replays, speech synthesis, etc. The imposters/fraudsters often use different voice spoofing attacks to fool the ASV systems to achieve certain objectives, i.e., bypass the security of someone's home or stealing money from a bank account, etc. To counter such fraudulent activities on the ASV systems, we propose a robust voice spoofing detection system capable of effectively detecting multiple types of spoofing attacks. For this purpose, we propose a novel feature descriptor *Center Lop-Sided Local Binary Patterns* (CLS-LBP) for audio representation. CLS-LBP effectively analyzes the audios bidirectionally to better capture the artifacts of synthetic speech, microphone distortions of replay, and dynamic speech attributes of the bonafide signal. The proposed CLS-LBP features are used to train the long short-term memory (LSTM) network for detection of both the physical- (replay) and logical-access attacks (speech synthesis, voice conversion). We employed the LSTM due to its effectiveness to better process and learn the internal representation of sequential data. More specifically, we obtained an equal error rate (EER) value of 0.06% on logical-acess (LA) while 0.58% on physical-access (PA) attacks. Additionally, the proposed system is also capable of detecting the unseen voice spoofing attacks and also robust enough to classify among the cloning algorithms used to synthesize the speech. Performance evaluation on the ASVspoof 2019 corpus signify the effectiveness of the proposed system in terms of detecting the physical- and logical-access attacks over existing state-of-the-art voice spoofing detection systems.

1. Introduction

Automatic speaker verification (ASV) systems are commonly used these days in a variety of devices, e.g., cellphones, intelligent speakers (Amazon Alexa, Google Home), etc., to authenticate the identity of any person for various application domains, i.e., banking, call centers, forensic laboratories, e-commerce systems, etc. For instance, in an iPhone, a Siri or a Google Home gets voice-based commands from its users to perform several actions, e.g., scheduling reminders, searching on internet, call or text someone, unlock cellphone, weather check, etc., (Delfino, 2021). The ASV based authentication is gaining importance due to the recent COVID-19 situation where other biometrics verification like fingerprint scanning and password-based verification systems are discouraged due to health concerns. Hence, voice biometrics technology is becoming popular for user authentication. Apart from the benefits of ASV for user authentication, ASV systems are vulnerable to numerous voice spoofing attacks such as replays, speech synthesis, voice conversion (VC), etc., that can be employed to achieve certain tasks such as controlling the home or bank account of someone, etc. In recent times, we have witnessed some cases where intruders have employed different voice spoofing attacks to spoof the ASV systems for potential frauds. Recently in the US, a complaint was reported where robbers employed an artificial speech of the CEO of a company to deceive their employees to transfer funds into a secret account (Harwell, 2021). In order to deal with the potential limitations of ASV systems, the research community is focusing to develop robust voice anti-spoofing systems to provide a protective layer for ASV systems against different voice spoofing attacks.

* Corresponding author.
E-mail address: hdaoud@uj.edu.sa (H. Dawood).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2022.02.024
1319-1578/

Spoofed voice samples can be produced by changing the original audio signal using recording, manipulating, or mimicking. Existing voice spoofing attacks can be classified into physical-access attacks, i.e., replay (Alegre et al., 2014; Rosenberg, 1976), or logical-access attacks, i.e., speech synthesis (Yamagishi et al., 2009; Lindberg and Blomberg, 1999), VC (Evans et al., 2009; Zen et al., 2009). In voice conversion, the voice spoken by the registered speaker is synthetically generated to very similar sound of the already enrolled speaker. Speech synthesis represents the machine generated voice of the target speaker. In replay spoofing attack, impersonator records the voice of enrolled speaker and play to the ASV system for granting an access on behalf of a registered speaker. We have mentioned two scenarios in Fig. 1, where intruders can employ the replay and speech synthesis/cloning attacks to exploit the vulnerability of ASV against spoofing attacks and gain access of someone's home or organization. Shown in Fig. 1(a) is the scenario of a replay attack, where devices such as air conditioner (AC) in a home are remotely accessible via a mobile application. These home devices are controllable through the smart speakers, e.g., Google Home, etc., where we can send various commands to the smart speaker via mobile application. Consider a case where an intruder uses some covert device to record the voice command of the genuine speaker and later playback the replay audio in front of the Google Home device to control the AC system of the home. Next, we present a scenario of a voice cloning attack as shown in Fig. 1(b), where the staff of a healthcare company uses a clinical app in the clinic which is remotely accessible via a mobile application. Clinical app is controllable through the smart speakers, e.g., Sonos One, Apple HomePod, Amazon Echo & Alexa, etc., where clinical staff can send various commands to the smart speaker using a mobile application. The staff uses the clinical app to enter large amount of data verbally and remotely to minimize the mistakes and omissions faced in manual data entry. Consider a spoofing scenario where an intruder artificially generates the synthesized voice sample against a bonafide speaker from text or voice samples using the sophisticated cloning algorithms. Later, the intruder plays the synthesized voice in front of Sonos One to get access to the clinical app.

Existing voice spoofing countermeasures have been proposed to address the physical-access (PA) and logical-access (LA) attacks. In Witkowski et al. (2017), inverted mel-frequency cepstral coefficients (IMFCC), linear prediction cepstral coefficients (LPCC), LPCCres features were employed to analyze the high-frequency bands for audio representation. These three spectral features were fed to the Gaussian mixture model (GMM) for classification of bonafide and replay samples. Yang et al. (2018) explored the extended Constant-Q cepstral coefficients (eCQCC) extracted from the constant-Q transform and fixed re-sampling of octave power spectrum to obtain the linear power spectrum. The coefficients of both octave and linear spectrum were concatenated to obtain the eCQCC features. Next, these features were employed with a deep neural network (DNN) for classifying the bonafide and spoof

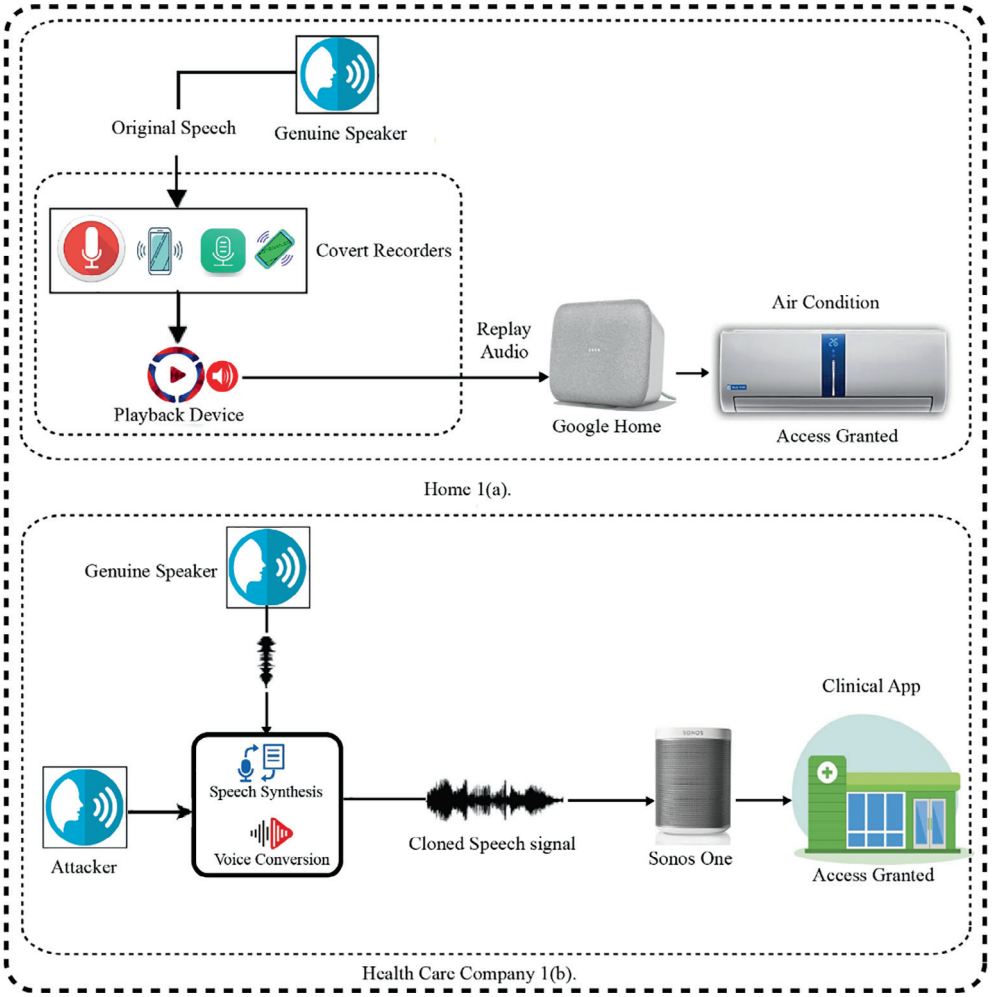

Fig. 1. Example of voice replay and cloning attacks.

samples. Malik (2019) employed the higher order spectral analysis (HOSA) features to capture the discriminative attributes of bonafide and cloned voice samples. Linearity statistical tests and Gaussianity were performed on the HOSA features to differentiate between the bonafide and cloned samples. Chettri and Sturm (2018) used a hybrid feature vector comprising of IMFCC, Mel-frequency cepstral coefficients (MFCC), Linear filter cepstral coefficients (LFCC), rectangular filter cepstral coefficients (RFCC), spectral centroid magnitude coefficients (SCMC), and CQCC features to classify the bonafide and spoof audios by employing the GMM. This model delivers improved recognition performance on development set as compared to the evaluation set of PA collection of ASVspoof 2019 dataset. Moreover, this method (Chettri and Sturm, 2018) is computationally complex due to increased features computation cost. Recently, the researchers worked to detect the voice replay (Kamble and Patil, 2020; Malik et al., 2020; Lin et al., 2020; Kamble and Patil, 2021; Phapatanaburi et al., 2020) and speech synthesis attacks (Elsaeidy et al., 2020; Gritsenko et al., 2020; Krishna et al., 2020; Helali et al., 2020; Bird et al., 2020; Raju et al., 2020). However, existing voice anti-spoofing methods have certain limitations, i.e., speech samples lack intentional speaker-oriented modifications, high features computation cost, single spoofing type detectors, etc. In practice, the type of attacks on ASV systems are mostly unknown. Unfortunately, generalized countermeasures that can cope with unknown voice spoofing attacks have not been thoroughly explored yet. Therefore, spoofing countermeasures need to be trained in most generalized way to effectively capture distinct nature of characteristics for numerous PA and LA attacks. Still, there is a need to develop a robust spoofing countermeasure that can accurately detect a variety of PA and LA voice spoofing attacks.

In this paper, we introduce an effective voice spoofing detection system to detect both the PA and LA attacks. For this, we propose a novel features representation scheme Center Lop-Sided Local Binary Patterns (CLS-LBP) to better capture the characteristics of genuine and spoofed audio samples. Later, we used our CLS-LBP features to train the long-short term memory (LSTM) network for classification. The significance of LSTM network for better analysis of time-based series data has encouraged us to use it for classification purpose. Moreover, the proposed system is able to accurately classify the cloning algorithms used to synthesize the bonafide samples. We evaluated the performance of the proposed technique on both PA (bonafide and replay samples) and LA sets (bonafide, speech synthesis, and voice conversion) of ASVspoof 2019 corpus. The major contributions of our paper are as follows:

- We propose novel acoustic CLS-LBP features to reliably capture the traits of bonafide as well as spoofed samples by extracting the information bi-directionally from the audio signal.
- We propose a robust voice spoofing detection system that can dependably be employed to detect both the physical- and logical-access attacks.
- Our spoofing detection system is also capable of detecting the unseen synthetic speech attacks.
- Our system has the capability of detecting the type of algorithm used to synthesize the bonafide audio.
- We provide rigorous experimentation on ASVspoof 2019 corpus to evaluate the significance of our spoofing detector over existing techniques.

The remaining paper is structured as follows. Section 2 provides a detailed analysis on the existing voice spoofing detection systems. Section 3 presents the details of our method. Section 4 has the details of experiments and discussion while the Section 5 presents the conclusion.

2. Related work

The related work section presents a significant analysis and discussion of the existing methods for voice spoofing countermeasures. Existing spoofing countermeasures have employed either the spectral or deep features for audio signal representation. Moreover, current methods have used either the traditional machine learning or deep learning classifiers-based approaches. We have discussed all these variants in this section.

2.1. Spectral features based methods

The ASV research community has proposed various voice spoofing countermeasures (Kamble and Patil, 2020; Malik et al., 2020; Lin et al., 2020; Kamble and Patil, 2021; Phapatanaburi et al., 2020; Banaras et al., 2021) to address the PA attacks. Kamble and Patil (2020) employed the variable length teager energy cepstral coefficients with the GMM to detect the voice replay attacks. Malik et al. (2020) introduced the acoustic ternary pattern-gammatone cepstral coefficients (ATP-GTCC) features for replay spoofing detection in voice controlled IoT devices. Error correcting output codes model was employed for training the multi-class support vector machine (SVM) classifier on the ATP-GTCC features. Lin et al. (2020) employed the Teager energy operator (TEO) to determine running approximation of sub-band energies and utilized these features for training the GMM to classify bonafide and replay signals. Kamble and Patil (2021) modified the conventional CQCCs using linear prediction residual (LPR) signal instead of raw speech signal. Linear prediction residual constant-Q cepstral coefficients (LPR-CQCC) features were employed in combination of CQCCs features for training the GMM to distinguish between the authentic and fake audio. In (Alluri and Vuppala, 2019), three features such as single frequency cepstral coefficients (SFCC), zero-time windowing cepstral coefficients, and instantaneous frequency cepstral coefficients are employed to detect the synthetic speech attacks. GMM was employed as a back-end classifier to classify the authentic and spoof audio. In Alluri et al. (2017), single-frequency filtering that provides spectral and high temporal resolution at each instant was employed to detect the replay attacks. SFCCs were fed into GMM to classify the authentic and spoof audio.

Existing countermeasures (Elsaeidy et al., 2020; Gritsenko et al., 2020; Krishna et al., 2020; Helali et al., 2020; Bird et al., 2020; Raju et al., 2020; Hassan and Javed, 2021; Qadir et al., 2022) have also explored various spectral features for LA attacks detection. Gritsenko et al. (2020) explored the energy difference between diffusions of cloned and bonafide speech signal. Linguistic and pitch features were fed to a deep neural network (DNN) for classification. Krishna et al. (2020) explored the electroencephalography (EEG) for speech synthesis and used the recurrent neural network (RNN) regression model for classification. Helali et al. (2020) fused the perceptual wavelet packet (PWP) and MFCCs to train the SVM for classification of bonafide and synthetic speech. In De Leon et al. (2012), modified group delay (MGD), relative phase shift (RPS) and MFCC features were employed for training the GMM to detect the synthetic speech.

Existing spoofing detection methods have also been proposed to address both the PA and LA attacks using either the spectral or deep features. Long-range acoustic features derived from long term constant-Q transform (CQT) were used in Das et al. (2019) for PA and LA attacks detection. Spectral features, i.e., MFCC, LFCC, CQCC, instantaneous Frequency cosine coefficient, and eCQCC were used in Das et al. (2019) to train the GMM and DNN classifiers for detection of the PA and LA attacks. Fusion of these features show better performance with DNN classifier. However, performance of the

fusion of MFCC and LFCC degrades on the development set of ASVspoof 2019 LA corpus. Tak et al. (2004) used the CQCC and LFCC to train the GMM for classification of spoof and bonafide samples. It was concluded in Tak et al. (2004) that the performance on linearly scaled CQCC and LFCC was worst for A17 spoofing algorithm of LA set. Das et al. (2020) explored CQCC, eCQCC, and constant-Q statistic-plus-principal information coefficients features to train the DNN for detection of both the PA and LA attacks. This method provides better performance on development set over the evaluation set of the ASVspoof 2019 corpus.

2.2. Analysis of deep learning based methods

The significance of deep learning has also been utilized by the ASV research community and proposed various deep learning-based voice spoofing countermeasures to deal with both the PA as well as LA attacks.

Existing countermeasures (Zhai and Vamvoudakis, 2020; Singh and Pati, 2020; Huang and Pun, 2020; Adiban et al., 2019; von Platen et al., 2002; Gong et al., 2020; Aravind et al., 2008; Wang et al., 2019; Zhang et al., 2020; Saranya and Murthy, 2018; Suthokumar et al., 2018; Chettri et al., 2018; Białobrzeski et al., 2019) have explored various deep learning methods to detect the replay attacks. Tak et al. (Wang et al., 2019) employed the linear frequency residual cepstral coefficients (LFRCC) with the CNN for voice replay detection. LFRCC provides better detection performance on the development set of PA, however, unable to perform well on the evaluation set. Zhang et al. (2020) introduced channel consistency DenseNeXt by integrating the ResNeXt and DenseNet for voice replay attacks. MFCC, LFCC, CQCC features were employed for training the DNN to classify the spoof and authentic audio. Saranya and Murthy (2018) introduced mel filterbank slope (MFS) and linear filterbank slope (LFS) features with the GMM to detect the replay attacks. MFS captures low frequency while the LFS captures high frequency information which corresponds to low- and high-quality recording devices, respectively. In Suthokumar et al. (2018), short-term spectral features and long-term spectral average features were extracted from the modulation spectrum to analyze the static and dynamic characteristics of the signal. Long-term spectral average captures the static characteristics of modulation spectrum of the speech signal. GMM was used to classify the replay and bonafide signals. Chettri et al. (2018) employed the instantaneous frequency cosine coefficients, discrete cosine transforms, and residual mel frequency cepstral coefficients to train the convolutional neural network for classification of bonafide and replay signal. Białobrzeski et al. (2019) explored the Bayesian neural network (BNN) and light convolutional neural network (LCNN) to detect the replay attacks. The performance of BNN was better on small-scale dataset, however, unable to generalize well on a large-scale dataset like ASVspoof 2019.

Research community has worked on various voice spoofing detectors (Janyoi and Seresangtakul, 2020; Michelsanti et al., 2004; Valle et al., 2005; Koriyama and Saruwatari, 2020; Zhou et al., 2020) for LA attacks. Janyoi and Seresangtakul (2020) presented a fundamental frequency (F_0) model based on RNN and combined their linguistic features to represent supra-segmental characteristics of F_0 contour. Valle et al. (2005) presented a generative network Flowtron for synthetic speech detection. This model learns the inverse mapping of data that can be changed to control different aspects of speech synthesis (i.e., tone, speech-rate, accent, pitch, etc). Koriyama and Saruwatari (2020) introduced a deep guassian process (DGP) model for audio sequence modeling. DGP comprises of many layers known as Bayesian kernel regression. Bayesian models can be trained with consideration of model complexity. Simple recurrent unit was used to classify the bonafide and spoof samples.

The research community has also worked on numerous deep learning techniques (Malik, 2019; Gomez-Alanis et al., 2019; Lavrentyeva et al., 1904; Zeinali et al., 2019) to address both the PA and LA attacks. In Malik (2019), LCNN based system was used to detect both the PA and LA attacks. The potential benefit of LCNN architecture is use of Max-Feature-Map-Activation function that was used to reduce the computational cost of deep learning model. Alanis et al. (2019) employed the light convolutional gated recurrent neural network (LC-GRNN) for deep features extraction that were then employed to train the SVM, linear discriminant analysis, and probabilistic Linear discriminative analysis for detection of both the PA and LA attacks. Lavrentyeva et al. (1904) explored the efficiency of using simple energy based speech activity detector and LFCC features to train thGomez-Alanis ee LCNN for classification. Zeinali et al. (2019) employed the mel-filter bank, MFCC, Constant Q-transform, CQCC and power spectrogram with visual geometry group for detection of genuine and replay/cloned voices.

3. Proposed methodology

This section provides a discussion on the proposed voice spoofing countermeasure. The details of our novel feature extraction scheme, i.e., CLS-LBP is also discussed in detail. We have designed an LSTM network consisting of 10 LSTM layers, which is trained using the proposed CLS-LBP features to categorize the bonafide and spoof audio. The architecture of our method is shown in Fig. 2.

3.1. Motivation of proposed feature

To develop a robust method that can accurately detect a variety of voice spoofing attacks such as PA (replays) and LA (voice conversion, TTS synthesis) attacks, we need a feature descriptor that can capture the dynamic properties of bonafide speaker vocals, generative algorithmic traits, and microphone fingerprints. For this purpose, we propose a novel CLS-LBP feature descriptor that analyzes the local variations of time-domain audio signals in both the forward and backward directions. By analyzing only, the 8 neighboring samples of the central sample, our CLS-LBP features can effectively extract even the minute details of vocal dynamic traits of bonafide speech, microphone fingerprints, and generative algorithm artifacts. Thus, makes it a reliable method for voice spoofing detection.

3.2. Feature extraction

For a robust voice anti-spoofing system, we need to develop an effective feature descriptor capable of capturing the distortions of replay signals, cloning algorithm artifacts in synthesized/cloned signals and dynamic attributes of human speaker vocal tract in genuine audio. To accomplish this objective, we propose a novel CLS-LBP descriptor for audio representation. CLS-LBP features extract the distinctive information bidirectionally from the audios to better capture the distortions of replays, artifacts of synthetic speech and dynamic speech variations of the bonafide signals. Fig. 3 depicts the framework of the proposed CLS-LBP algorithm.

An input audio signal $Y[n]$ with N samples is partitioned into $i = \{1,2,\ldots,k\}$ non-overlapping windows $W^{(i)}$ with length $l = 9$. In each window $W^{(i)}$, p represents the central sample in a frame and have four right neighbors $q^{(i)}_{right}$ and four left neighbors $q^{(i)}_{left}$, where i represents the index of neighboring samples. CLS-LBP features are computed by encoding each window $W^{(i)}$ of an audio signal $Y[n]$.

To compute the CLS-LBP pattern, we compare the right and left neighboring samples with the central sample p and set it to 1 or 0 depending on the values of left neighbor $q^{(i)}_{left}$ and right neighbor

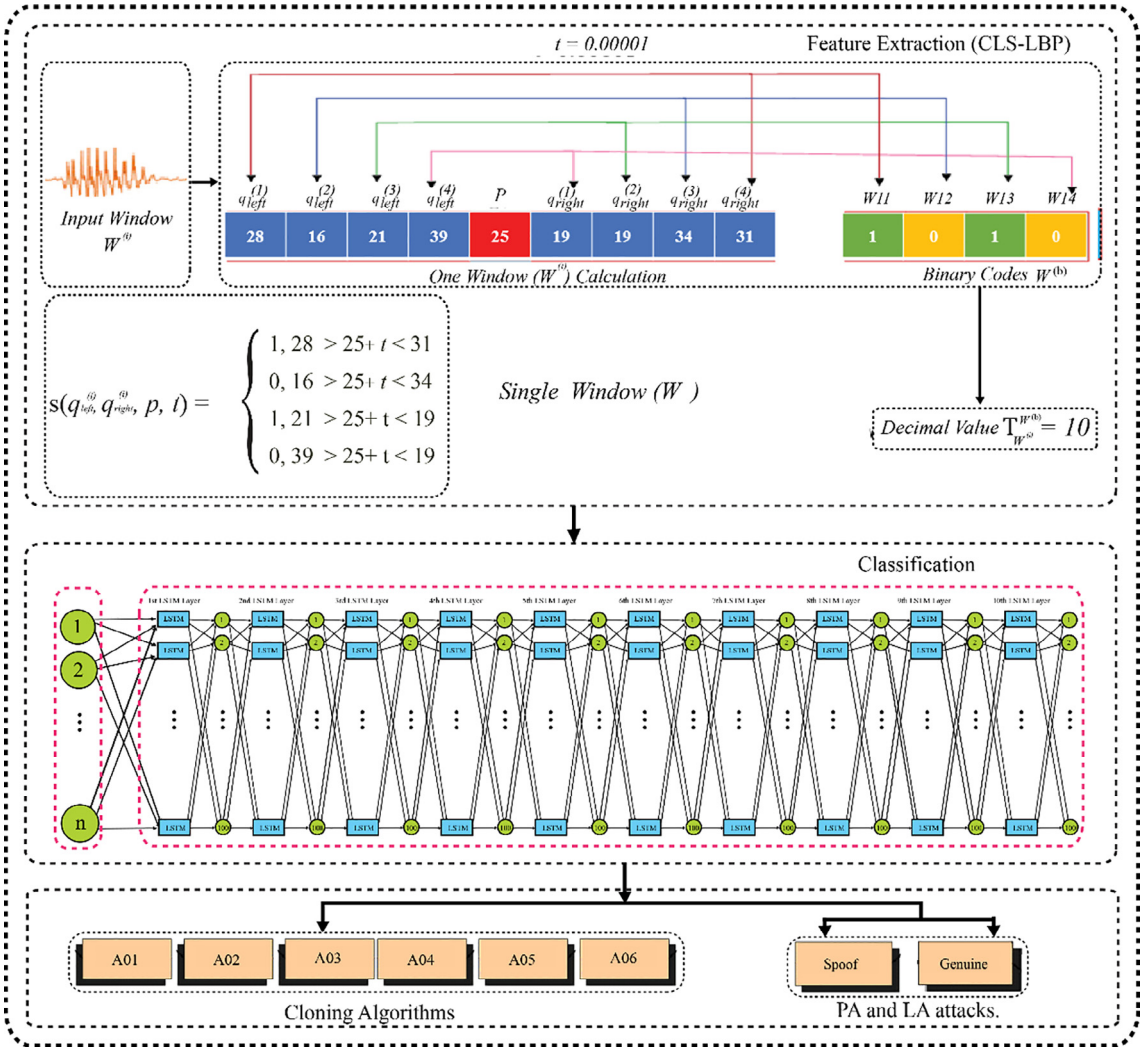

Fig. 2. Proposed Voice Spoofing Detection System.

$q^{(i)}_{right}$ around p using the threshold value t. We used a linear searching approach to determine the value of t. For this purpose, we initialize the value of t to zero and optimize it to find the convergence point between 0 and 1. More precisely, we obtained the value of t = 0.00001 as optimal results were obtained on this threshold value. If the magnitude of both the $q^{(i)}_{left}$ and $q^{(i)}_{right}$ is greater than $(p + t)$ or the magnitude of both $q^{(i)}_{left}$ and $q^{(i)}_{right}$ is smaller than $(p + t)$, then value of the sample is set to 1. Next, if the magnitude of $q^{(i)}_{left}$ is greater than $(p + t)$ and the magnitude of $q^{(i)}_{right}$ is smaller than $(p + t)$, then we set the value of sample to 0. Similarly, if the magnitude of $q^{(i)}_{left}$ is smaller than $(p + t)$ and the magnitude of $q^{(i)}_{right}$ is greater than $(p + t)$, then we also set it to 0. By following this process, we generate a binary code of four bits against each window of the audio signal.

Fig. 3(a) illustrates the computation of CLS-LBP features for one window $W^{(i)}$of the audio signal. We compute the CLS-LBP codes against each window $W^{(i)}$ in four steps. In the first step, we compare the magnitude of $q^{(1)}_{left}$ and $q^{(4)}_{right}$ with $(p + t)$. As the magnitude of $q^{(1)}_{left}$ and $q^{(4)}_{right}$ is greater than $(p + t)$, so we assign the binary code of 1, shown as W_{11} in Fig. 3. In the second step, we compare the magnitude of $q^{(2)}_{left}$ and $q^{(3)}_{right}$ with $(p + t)$, where we can observe that the magnitude of $q^{(2)}_{left}$ is smaller and magnitude of $q^{(3)}_{right}$ is greater than $(p + t)$, therefore, we assign the code of 0 to W_{12} (Fig. 3a). In the third step, we compare the magnitude of $q^{(3)}_{left}$ and $q^{(2)}_{right}$ against $(p + t)$, and as the magnitude of both neighbors is smaller than $(p + t)$, so we assign the code of 1 to W_{13} (Fig. 3a). Finally, in the last step, we compare $q^{(4)}_{left}$ and $q^{(1)}_{right}$ with $(p + t)$. As the magnitude of $q^{(4)}_{left}$ is greater and the magnitude of $q^{(1)}_{right}$ is smaller than $(p + t)$, so we assign the code of 0 to W_{14} (Fig. 3a). This process is repeated for all windows of the audio to compute the CLS-LBP features for the entire acoustic signal. The two-valued function of proposed CLS-LBP algorithm is computed as follows:

$$S\left(q^{(i)}_{left}, q^{(i)}_{right}, p, t\right) = \left\{ \begin{array}{c} 1, if q^{(i)}_{left} > p + t and q^{(i)}_{right} > p + t \\ or \\ q^{(i)}_{left} < p + t and q^{(i)}_{right} < p + t \\ 0, if q^{(i)}_{left} < p + t and q^{(i)}_{right} > p + t \\ or \\ q^{(i)}_{left} > p + t and q^{(i)}_{right} < p + t \end{array} \right\}, \quad (1)$$

where $S\ (q^{(i)}_{left}, q^{(i)}_{right}, p, t)$ represents the acoustic signal using two-valued center lop-sided local binary pattern. Next, we compute and encode the patterns in their decimal values as

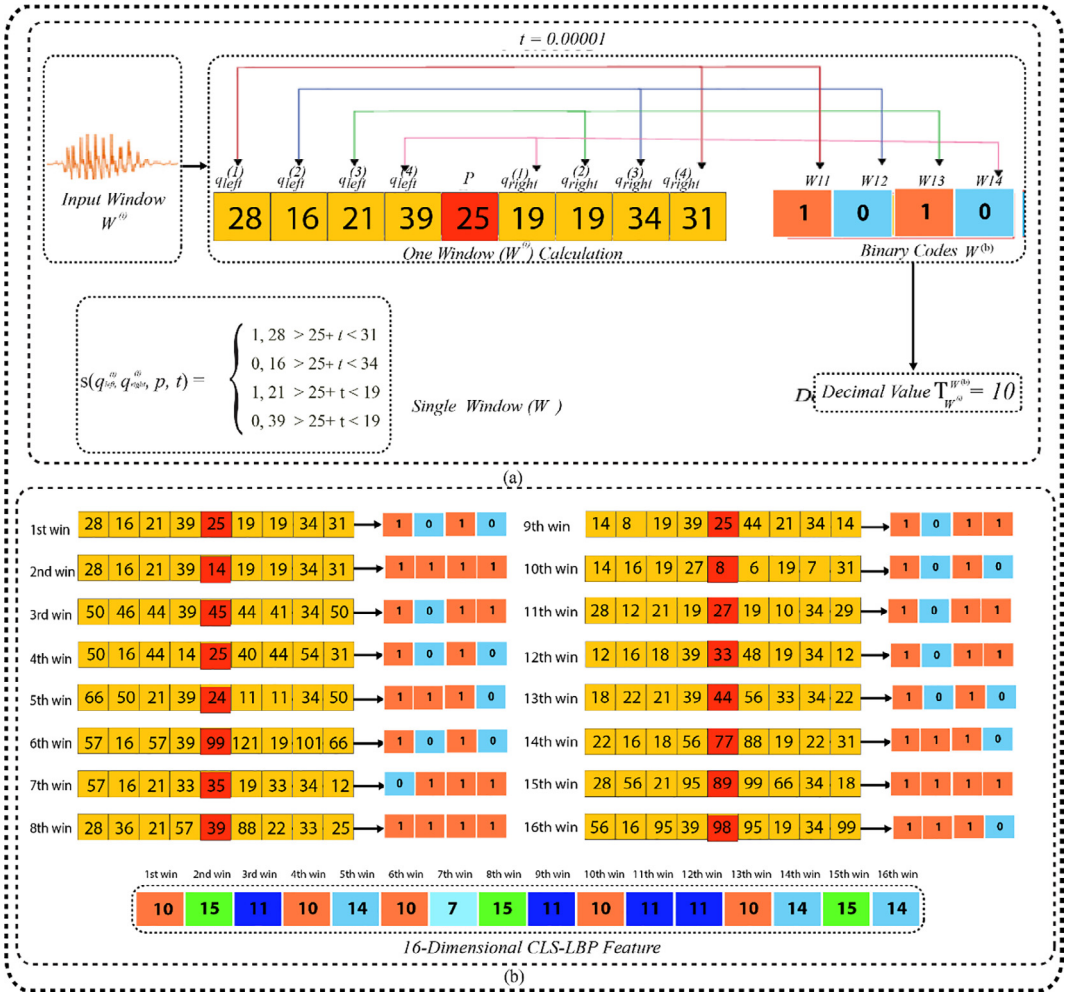

Fig. 3. Acoustic Center Lop-sided Local Binary Pattern (CLS-LBP) Computation.

$$T^r = \sum_{i=0}^{3} S(qi, p, t) \times 2^i, \tag{2}$$

where the T^r represents the uniform CLS-LBP codes in decimal form. Finally, we compute the histogram of T^r and assign one histogram bin for each uniform pattern and include all non-uniform patterns into single bin as the uniform patterns contains maximum information of the signal (Malik et al., 2020). We compute the histogram as follows:

$$h(x) = \sum_{m=1}^{M} \delta(T_m^r, x), \tag{3}$$

where x denotes the histogram bins corresponding to uniform acoustic CLS-LBP codes and $\delta(.)$ is the Kronecker delta function. We performed different experiments to determine that the first 16 uniform patterns are enough to capture maximum characteristics of the bonafide and spoof signals. Therefore, we selected the histogram of these uniform patterns to create a 16-dimension CLS-LBP features descriptor as illustrated in Fig. 3(b).

3.3. Classification

3.3.1. Long Short-Term memory networks (LSTM)

The RNN has achieved remarkable performance in sequential modeling tasks. To process the random sequences of different inputs, RNN uses internal memory which allows the information to remember as the information is stored in all memory cells. Moreover, LSTM can memorize the information for a long period and is designed to prevent the long-term dependencies among elements within the input sequence. Thus, it is better able to analyze the information about the input sequences. We used the LSTM in the proposed work for the classification task. The architecture is composed of a memory part of the LSTM unit (cell) and three different regulators or gates, i.e., input gate, output gate and a forget gate.

Fig. 4 demonstrates the stream of the information at time step s involving the gates to update, forget, and output the cell and hidden states. The learning weights of the layer are the input weights

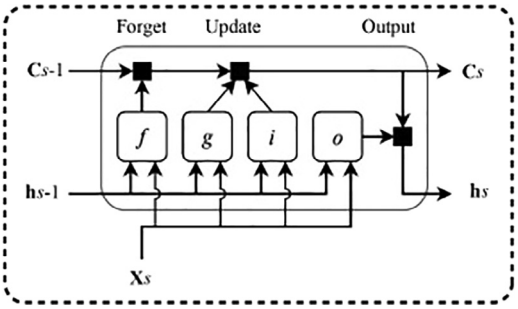

Fig. 4. LSTM cell (Hochreiter and Schmidhuber, 1997).

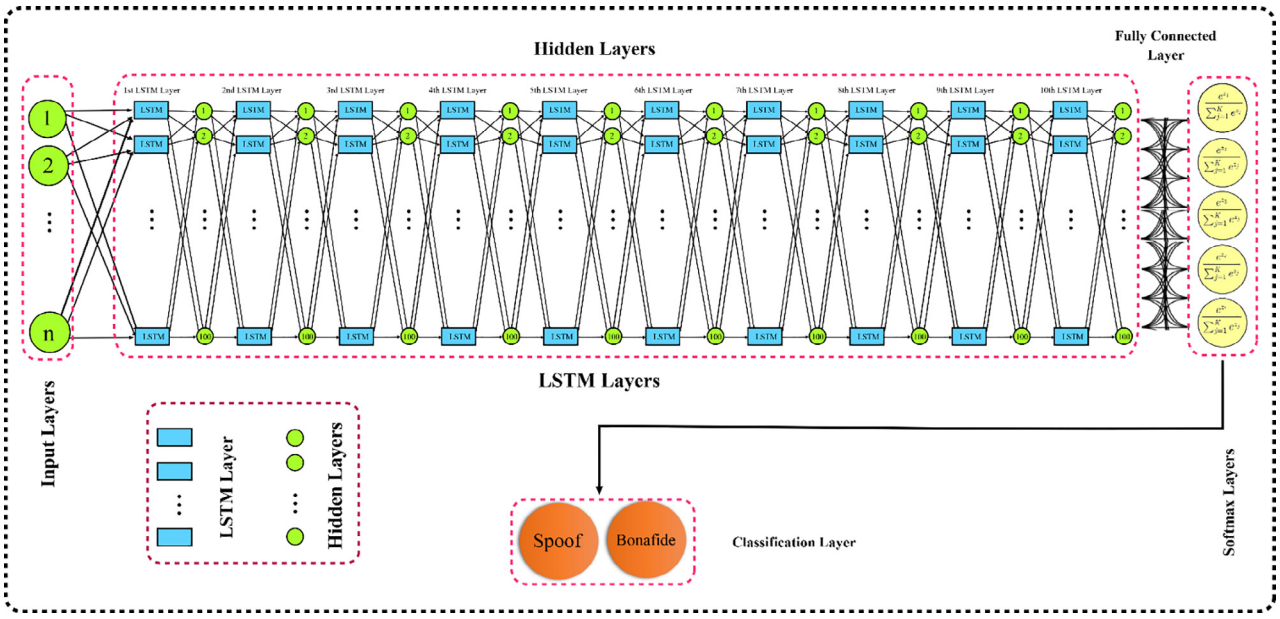

Fig. 5. LSTM Architecture.

Z, the recurrent weights *R*, and the bias *b*. Following three matrices *Z*, *R*, and *b* are concatenation of *Z*, *R*, and *b* of every module. *Z*, *R*, and *b* are concatenated as (Hochreiter and Schmidhuber, 1997):

$$Z = \begin{bmatrix} Zi \\ Zf \\ Zg \\ Zo \end{bmatrix}, R = \begin{bmatrix} Ri \\ Rf \\ Rg \\ Ro \end{bmatrix}, b = \begin{bmatrix} bi \\ bf \\ bg \\ bo \end{bmatrix}, \quad (4)$$

where *i*, *f*, *g*, and *o* represent an input gate, forget gate, cell candidate, and the output gate, respectively. Cell in the state at some time step *s* is given by $\mathbf{c}_s = f_s \odot c_s - 1 + i_s \odot g_s$, where $\odot$ represents the element-wise multiplication of vectors (Hadamard product). The hidden state at some time step *s* is given by $\mathbf{h}_s = o_s \odot \sigma_c(\mathbf{c}_s)$, where σ_c represents the state activation function.

The time-series data has more complexity of sequence dependence among the input variables as compared to the regression modeling. Since audio is a time-series signal, therefore, there exists a need to effectively analyze the sequence of unique patterns in the audio. LSTM is appropriate to analyze the time-series data due to the ability to learn both the short- and long-term sequential dependencies of the audio signal. LSTM employ various gates to boost the capability to capture the nonlinear relationships and feedback connections to better analyze the sequential patterns of input data. The recurrent state in LSTM helps to learn order dependencies of time-series audio signal by retaining the previous information in the signal. Consequently, in this work, we used the LSTM network for the classification purpose. We used our proposed CLS-LBP features-set to train the LSTM network for LA, PA, known and unseen attacks detection. In the proposed system, we employed the LSTM network comprising of 10 LSTM layers with 100 hidden units in each layer, fully connected layer followed by a SoftMax layer and a classification layer. We have added a SoftMax layer at the end of the network for the classification of bonafide and spoof audio. SoftMax is an activation function that is used to normalize the outputs by converting the weighted sum values into probabilities (machinelearningmastery.com, 2020). The softMax function yields the actual probability scores for both the classes, i.e., spoof and bonafide in our work. If the probability score of the bonafide class is greater than the spoof one, the SoftMax layer predicts the sample as bonafide, whereas, predicts the sample as spoof in case probability score of the spoof class is greater than the bonafide class. Moreover, we set the maximum epochs to 20, mini batch size to 64 at each iteration, gradient threshold to 1, and trained the model using an Adam optimizer as we obtained best results on these hyperparameters settings. The details of our LSTM architecture are provided in Fig. 5.

4. Experimental setup and results

This section presents the details of the experiments performed for evaluating the performance of the proposed voice spoofing detection system. We evaluated the performance of proposed system on the PA and LA sets of ASVspoof 2019 dataset using equal error rate (EER), min-tDCF, accuracy, precision, and recall.

We have used the MATLAB 2019 for implementation purposes. Moreover, we used a computing machine with these specifications: Core i5 7th generation, 12 GB RAM.

4.1. Dataset

ASVspoof challenge was started with the development of ASVspoof 2015 corpus (Wu et al., 2015) which was developed to evaluate the speech synthesis/cloning detection systems. Two years later, ASVspoof 2017 corpus (Kinnunen et al., 2017) was released to evaluate the replay detection systems. In 2019, ASVspoof challenge came up with the development of a large and diverse public dataset ASVspoof 2019 (Wang et al., 2019) to counter both the physical and logical access attacks. ASVspoof 2019 corpus (Wang et al., 2019) consists of two sets, i.e., LA and PA. The LA set of ASVspoof 2019 contains voice conversion and speech synthesis samples along-with the bonafide audio samples, whereas, the PA set of ASVspoof 2019 corpus (Wang et al., 2019) includes the bonafide and replay samples. Moreover, both the PA and LA sets are further partitioned into three subsets, i.e., training set, development (dev) set, and evaluation (eval) set.

LA dataset consists of spoofed and bonafide speech data created by 17 distinct TTS and VC systems. The data, which is utilized to train the VC and TTS systems is derived from the voice cloning toolkit VCTK database (Veaux et al., 2020). Six spoofing systems are labelled as well-known attacks and remaining 11 as anonymous

Table 1
Details of Cloning Algorithms used for ASVspoof 2019 LA Dataset.

Logical access	Training samples	Development samples	Spoof System	Algorithm Type	Input	Input processor	Conversion	Speaker Represent	Outputs	Waveform Generator
Total samples	22,800	22,296	–	–	–	–	–	–	–	–
A01	3,800	3,716	TTS	Neural waveform model	Text	NLP	AR RNN*	VAE*	MCC, F0	WaveNet*
A02	3,800	3,716	TTS	Vocoder	Text	NLP	AR RNN*	VAE*	MCC, F0, BAP	WORLD
A03	3,800	3,716	TTS	Vocoder	Text	NLP	FF*	One hot embed	MCC, F0, BAP	WORLD
A04	3,800	3,716	TTS	Waveform concatenation	Text	NLP	CART	–	MCC, F0	Waveformconcat.
A05	3,800	3,716	VC	Vocoder	Speech (human)	WORLD	VAE*	One hot embed	MCC, F0, BAP	WORLD
A06	3,800	3,716	VC	spectral filtering	Speech (human)	LPCC/MFCC	GMM-UBM	–	LPC	Spectral filtering + OLA

Table 2
Statistics of ASVspoof 2019 PA Dataset.

PA samples	Total Samples	Environment Definition	Labels			Attack Definition	Labels			Replay Device Quality	OB	minF	LNRL
			a	b	c		A	B	C				
Training	54,000	–				–				–			
Dev	29,700	S: Room size (m^2)	2–5	5–10	10–20	Da: Attacker-to-talker	10–50	50–100	>100	Perfect	inf	0	inf
		R: T60ms	50–200	200–600	600–1000			distance (cm)		High	>10	<600	>100
Eval	1,34,730	Ds: Talker-to	10–50	50–100	100–150	Q: Replay device	perfect	high	low quality	Low	<10	>600	<100
			ASV distance (cm)										

Table 3
Statistics of ASVspoof 2019 LA Dataset.

LA samples	–	Spoof Systems	Input and Input Processor	Waveform Generator
Training	25,380	–	–	–
Dev	24,844	A01, A02, A03, A04, A05, A06, A07, A08, A09, A10, A11, A12, A13, A14, A15, A16, A17, A18, A19	Text, Speech (human), Speech (TTS), NLP, WORLD, LPCC/MFCC, CNN + bi + RNN*, ASR*	WaveNet*, WORLD, Waveform concat, Spectral filtering
Eval	71,237			+OLA, Vocaine, STRAIGHT.

attacks. The training and dev set comprises of well-known attacks, whereas the eval set consists of 2 known and 11 anonymous attacks. LA set includes 4 TTS and 2 VC systems. VC systems use neural network based and spectral filtering-based approaches (Matrouf et al., 2006), whereas, the TTS systems use either waveform concatenation or neural network-based speech through traditional source-filter vocoder (Morise et al., 2016) or WaveNet based vocoder (Oord et al., 2016). The 11 anonymous spoofing systems containing 6 TTS, 3 hybrid VC-TTS and 2 VC systems employed numerous waveform generation techniques that are conventional vocoding, GriffinLim (Griffin and Lim, 1984), generative adversarial networks (Tanaka et al., 2018), neural waveform models (Oord et al., 2016; Wang et al., 2019), waveform concatenation, waveform filtering (Kobayashi et al., 2014), spectral filtering, and their combination.

ASVspoof 2017 (Kinnunen et al., 2017) dataset consists of real replay recordings, whereas ASVspoof 2019 (Wang et al., 2019) consists of simulated (Janicki et al., 2016; Campbell et al., 2005; Novak et al., 2015) replay recording in reverberant acoustic environment in order to enhance ASV reliability in reverberant conditions (Ko et al., 2017; Roomsimove, 2020). Training and dev sets are generated according to 27 acoustic and 9 replay configurations. The room sizes are classified into three intervals, i.e., small rooms, medium rooms, and large rooms. There are 3 groups of talker to ASV distance (Ds), i.e., short distance, medium distance, and large distance. Each physical space exhibit reverberation variability among spaces, i.e., ceiling, floor wall, and position in the room. Level of reverberation is mentioned in terms of T60 reverberation time with three different categories, i.e., short, medium, and high. Recordings are made in three different zones (A, B, C), each represents different distance (Da) from the talker. The recordings which are captured in Zone A is believed to be of better quality compared to those in zones B and C. The eval set is also created in the same way as dev and training datasets. The statistics of cloning algorithms are provided in Table 1 and the statistics of the PA and LA sets of ASVspoof 2019 corpus (Wang et al., 2019) are provided in Table 2 and Table 3, respectively.

4.1.1. Experimental protocols

In this section, we present details of the experimental protocol used during the experiments. For evaluation on the PA dataset, we used the training set having 54,000 samples (5400 bonafide and 48,600 spoof) to train the model. Whereas we tested our model on both the dev and eval sets. The dev set consists of 29,700 samples (5,400 bonafide and 24,300 spoof) and eval set consists of 1,34,730 samples (18,090 bonafide and 116,640 spoof). For evaluation on the LA dataset, we used the training set comprising of 25,380 samples (2,580 bonafide and 22,800 spoof) to train the model. We tested our model on both the dev and eval set. The eval set of LA dataset contains 71,237 samples (63,882 bonafide and 7,355 spoof), whereas, the dev set contains 24,844 samples (2548 bonafide and 22,296 spoof).

For evaluation of the cloning algorithms classification, we used spoofed samples of the entire training set (22800) to train our model and spoofed samples of entire dev set (22296) for testing the model.

4.2. Results and discussion

4.2.1. Performance evaluation on physical access attacks

The major goal of this experiment is to check the performance of our spoofing detector for PA attacks detection. For this purpose, we represented the audio samples of PA set using the proposed CLS-LBP features to train the LSTM model for classification of bonafide and replay samples. We obtained an EER of 0.58% and 2.91%, and min-tDCF of 0.016 and 0.072 on the eval and dev sets, respectively as shown in Table 4. From the results, we can examine that our spoofing detector achieved remarkable performance specifically on the eval set. Our spoofing detection system achieved better classification performance over the CQCC-GMM baseline model (Todisco et al., 2017). Particularly, we obtained an EER of 10.46% smaller than the EER obtained on eval set using the CQCC-GMM baseline model. These experimental findings signify that our spoofing detector is better to detect the physical access attacks on a diverse and large-scale ASVspoof 2019 corpus (Wang et al., 2019). We can conclude from this experiment that our CLS-LBP features are capable of effectively capturing the microphone distortions and fingerprint information available in the replay samples.

4.2.2. Performance evaluation on logical access attacks

The major goal of this experiment is to evaluate the performance of our spoofing detector on LA attacks. For experimentation on LA dataset, we utilized the proposed CLS-LBP features and train the LSTM to categorize bonafide and spoof samples (i.e., speech synthesis, voice conversion, etc.). The results obtained on ASVspoof 2019 LA dataset is provided in Table 5. We achieved an EER and min-tDCF of 0.06% and 0.0017 on eval set, and 0.35% and 0.0079 on dev set, as shown in Table 5. From the results, we can perceive that the proposed spoofing detector also achieved remarkable detection performance specifically on the eval set. Our system achieved better classification performance over the CQCC-GMM baseline model (Todisco et al., 2017). More specifically, we achieved an EER of 9.51% smaller than the EER obtained on the CQCC-GMM baseline model. The experimental findings signify

Table 4
Results on ASVspoof 2019 PA Dataset.

Corpus	EER%	min-tDCF	Accuracy%	Precision%	Recall%
Eval set	0.58	0.0160	99.42	99.97	99.33
Dev set	2.91	0.0720	96.18	99.89	95.49

Table 5
Results on ASVspoof 2019 LA Dataset.

Corpus	EER %	min-tDCF	Accuracy %	Precision %	Recall %
Eval set	0.06	0.0017	99.81	99.97	99.95
Dev set	0.35	0.0079	99.65	99.75	99.85

Table 6
Results on Cloning Algorithms of ASVspoof 2019.

Algo	Accuracy%	Precision%	Recall%
A01	98.5	98.54	98.57
A02	99.6	99.64	91.25
A03	96.2	92.97	99.43
A04	95.0	94.96	100
A05	97.6	96.79	98.38
A06	95.4	95.35	98.30

the significance of our spoofing detector to better detect the LA attacks on a diverse and large-scale ASVspoof 2019 corpus. We can conclude from this experiment that our CLS-LBP features effectively capture the dynamic speech variations of the bonafide samples along-with the artifacts available in the synthesized samples.

4.2.3. Performance evaluation of voice cloning algorithms detection

The objective of this experiment is to determine the algorithm type used to synthesize the bonafide samples of ASVspoof 2019 LA corpus. The LA set contains both the synthesized and voice conversion samples. Six different cloning algorithms are used for speech synthesis in the ASVspoof 2019 LA corpus (i.e., A01 TTS neural waveform model, A02 TTS vocoder, A03 TTS vocoder, A04 TTS waveform concatenation, A05 VC vocoder and A06 VC spectral filtering). For this experiment, we used the training set of LA collection (22,800 samples) to train our model and used the dev set of LA collection (22,296 samples) for model testing. We obtained an EER of 0.7% for A01, 2.26% for A02, 2.01% for A03, 1.32% for A04, 1.21% for A05 and 1.62% for A06 algorithm. The detailed results in terms of EER, min-tDCF, accuracy, precision, and recall are shown in Table 6.

From Table 6, we can observe that the proposed system performed best on A02 algorithm and achieved the lowest accuracy among all on A04 and A06 algorithms. As A02 is vocoder that uses WORLD waveform generator for speech synthesis, therefore, we can conclude from these results that our spoofing detector better captures those cloning artifacts introduced by WORLD waveform generator. On the other hand, A04 is waveform concatenation model that uses Waveformconcat waveform generator for speech synthesis while A06 is spectral filtering model that uses Spectral filtering + OLA waveform generator for speech synthesis. So, these results show that our system is slightly less effective to capture the cloning artifacts of Waveformconcat and Spectral filtering + OLA waveform generators over other synthetic models. Overall, we obtained excellent results for classification of the cloning algorithms. Thus, we conclude from this experiment that the cloning algorithms add their algo-specific artifacts in the synthesized audio

Table 7
Results on synthetic speech and voice conversion.

Spoofing category	EER %	min-tDCF	Accuracy %	Precision %	Recall %
TTS	0.64	0.0166	99.90	99.88	99.33
VC	20.31	0.4137	79.70	98.76	77.17
Overall LA	0.06	0.0017	99.81	99.97	99.95

Table 8
Results on unseen LA attacks.

Attacks	EER %	min-tDCF	Accuracy%	Precision%	Recall%
A07	0.37	0.0086	99.6	99.38	99.67
A08	0.38	0.0089	99.4	99.36	99.67
A09	9.47	0.2409	90.5	76.65	99.57
A10	0.37	0.0086	99.6	99.38	99.67
A11	0.37	0.0086	99.6	99.38	99.67
A12	0.38	0.0089	99.6	99.36	99.67
A13	0.37	0.0086	99.6	99.38	99.67
A14	0.39	0.0091	99.6	99.34	99.67
A15	0.43	0.0098	99.6	99.24	99.67
A16	0.41	0.0094	99.6	99.28	99.67
A17	39.31	0.4706	60.7	21.57	86.88
A18	40.18	0.4808	59.8	99.80	0.00
A19	40.18	0.4808	59.8	99.80	0.00

samples that can be captured well using our robust CLS-LBP features. This capability of not only detecting the spoofing type but also the cloning algorithms used to generate the spoofed audio makes our spoofing detection system more effective and useful for audio forensics applications.

4.2.4. Performance evaluation of synthetic speech and voice conversion

The main aim of this experiment is to investigate the performance of our spoofing detector on TTS and VC. For this purpose, we employed 16-dim CLS-LBP features to train the model to categorize the bonafide and spoof samples of TTS and VC separately. There are four TTS spoofing systems, i.e., A01, A02, A03, and A04 and two VC spoofing systems, i.e., A05 and A06 that are used to produce the spoof samples of training set of the LA dataset. For the evaluation set of LA dataset, there are 13 spoofing systems that consists of 7 TTS, i.e., A07, A08, A09, A10, A11, A12, A16, 3 VC spoofing systems i.e., A17, A18, A19 and 3 VC-TTS spoofing systems, i,e, A13, A14, and A15 that are used to create the spoof samples. We conducted a multi-stage experiment to separately evaluate the effectiveness of our method for TTS and VC spoofing detection.

In the first stage of this experiment, we used the bonafide and spoof samples (TTS) of training set of LA dataset to train the model and used the bonafide and spoof samples of evaluation set of LA dataset (TTS) to evaluate the model. The results are reported in Table 7. We achieved an EER of 0.64% and min-tDCF of 0.0166. In the second stage of this experiment, we used the bonafide and spoof samples (VC) of the training set of LA dataset to train the model and used the bonafide and spoof samples (VC) of eval set of LA dataset to evaluate the model. We achieved an EER of 20.31% and min-tDCF of 0.4137. The detailed results of VC spoofing detection are given in Table 7. It can be observed from these results that the proposed system performs better on the TTS spoofing detection as compared to the VC detection. The proposed system better captures the artifacts generated by neural waveform, griffin lim, and Vocoder TTS. We believe that this might be due to the reason that VC spoofing systems use the original voices as a source preserving the periodic characteristics of the speaker, which is not available in the TTS samples. Overall, the proposed system performs well on LA dataset and achieved an EER of 0.06% that shows the effectiveness of our system.

4.2.5. Performance evaluation of unseen attacks detection

This experiment is designed to evaluate the performance of the proposed system on unseen LA attacks, i.e., A07, A08, A09, A10, A11, A12, A13, A14, A15, A16, A17, A18, and A19. These spoofing systems are used to synthesize the samples of evaluation set of LA dataset and there are 63,895 samples of unseen attacks in the

Table 9
Performance Comparison against existing Voice Spoofing Detection Systems.

System	*LA Eval Set*		*PA Eval Set*	
	EER (%)	min-tDCF	EER (%)	min-tDCF
Baseline: CQCC + GMM (Todisco et al., 2017)	9.57	0.2366	11.04	0.2454
Baseline: LFCC + GMM (Todisco et al., 2017)	8.09	0.2116	13.54	0.3017
CQT + SE-Res2Net50 + CE (Li, et al., 2010)	2.502	0.0743	0.459	0.0116
Spec + LCGRNN + GKDE-Softmax(Gomez-Alanis et al., 2020)	3.77	0.0842	1.06	0.0222
Spec + LCGRNN + GKDE-Triplet (Gomez-Alanis et al., 2020)	3.30	0.0776	0.92	0.0198
sm-ALTP-Asymmetric Bagging (Aljasem et al., 2021)	5.22	0.132	1.1	0.0335
Ours: CLS-LBP + LSTM	**0.06**	**0.0017**	**0.58**	**0.0160**

Table 10
Results of Existing features on LSTM.

Features	*LA Eval Set*		*PA Eval Set*	
	EER (%)	min-tDCF	EER (%)	min-tDCF
CQT	15.92	0.3590	7.28	0.1620
CQT-ICQT	49.37	0.5823	6.57	0.1510
ICQT	23.28	0.1178	1.56	0.0397
LFCC	76.99	0.7423	49.86	0.5829
LFCC + CQT	17.44	0.3927	39.84	0.3829
CQCC	1.18	0.0520	11.5	0.2457
Ours: CLS-LBP	**0.06**	**0.0017**	**0.58**	**0.0160**

evaluation set. We used the bonafide and spoof samples of LA training set for training the model and used the bonafide of evaluation set and spoof samples of specific unknown attacks for the evaluation purpose. The results are given in Table 8. These results show that our method performed well on A07, A10, A11, and A13 and achieved an EER of 0.37%, min-tDCF of 0.0086, precision of 99.38%, recall of 99.67%, and an accuracy of 99.6%. Moreover, our method obtained the worst results on A17, A18 and A19 by achieving an EER of 39.31% for A17, 40.18% for A18 and A19, and min-tDCF of 0.4706 for A17, 0.4808 for A18 and A19. These results illustrate that TTS-base synthetic speech is easier to detect than VC-based synthetic speech. Comparison of the waveform generation methods proves that synthetic speech generated by waveform filtering based techniques (A17, 18, and A19 attacks used VC waveform filtering, VC Vocoder and spectral filtering) are the most difficult to detect than the other types of attacks. Although, our method struggled to better capture the artifacts produced by VC attacks, however, performed well on the overall LA attacks.

4.2.6. Robustness of the proposed system

To demonstrate the robustness of the proposed system, we tested our system on the diverse and largescale ASVspoof 2019 dataset. It is important to mention that this corpus contains the voice samples of 87 unseen speakers used for evaluation purposes as compared to the voice samples of 20 speakers, which were used for training. Similarly, spoofed samples used for the training purposes were cloned using only 6 algorithms, whereas the spoofed samples used for evaluation purposes were generated through 19 voice cloning algorithms including the 13 new cloning algorithms. More specifically, we have performed experiments to detect the unseen voice spoofing attacks that are generated by using powerful spoofing algorithms such as A07, A08, A09, A10, A11, A12, A13, A14, A15, A16, A17, A18, and A19. The experimental results presented in Table 8 show the robustness of our method for unseen attacks detection. It is also important to mention that the ASVspoof training and evaluation sets contain the speech samples of different speakers, different algorithms for logical access attacks (voice conversions and text-to-speech synthesis), different microphones, background environments, etc., for physical access attacks. Thus, the evaluation sets of both the PA and LA collections contain much more diversity and challenging conditions as compared to the training sets. The experimental results in Tables 4, 5, and 6 indicate the robustness of our method for unseen speakers, microphones, background environments, and cloning algorithms. By testing our method on the evaluation set that contains the unseen speakers, unknown attacks, different microphones and background environments and still getting such excellent results indicate the robustness of the proposed method for reliable voice spoofing detection. Moreover, our method is robust to variations in spoofing attacks and able to reliably detect all types of physical and logical access attacks, i.e., text-to-speech, voice conversion, and voice replay attacks.

4.2.7. Performance comparison with existing methods

This experiment is performed to compare our spoofing detector against the existing techniques for voice spoofing detection. To justify the effectiveness of the proposed CLS-LBP features for better detection of the distortions in replay samples, artifacts in the cloning algorithms, and vocal tracts based dynamic speech characteristics of the bonafide samples, we performed a comparative analysis of our method with the models listed in Table 8. The results in terms of an EER and min-tCDF of the proposed and existing methods on the ASVspoof 2019 PA and LA corpus are provided in Table 9. On the PA-Eval set, (Li et al., 2010) performs best and achieved 0.459% EER and min-tDCF of 0.0116, whereas, our method performs second-best and achieved 0.58% EER and min-tDCF of 0.0160, whereas, the CQCC-GMM baseline performs the worst by achieving an.

EER of 11.04% and min-tDCF of 0.2454. On the LA-Eval set, our method achieves the best results (0.06% EER and min-tDCF of 0.0017) and CQCC-GMM baseline is the worst (9.57% EER and min-tDCF of 0.2366). From this comparative analysis, we can conclude that the proposed system outperforms the existing contemporary voice spoofing detectors and able to reliably detect a variety of voice spoofing attacks along-with the cloning algorithms used to synthesize the samples of LA dataset.

4.2.8. Performance comparison with existing features

The objective of this experiment is to evaluate the performance of our CLS-LBP features over the existing baseline features on the same classifier. For this purpose, we have compared our CLS-LBP features with the CQCC, LFCC, CQT, ICQT, CQT + ICQT, LFCC + CQT features using the same LSTM classifier. Results on both PA and LA dataset are given in Table 10. We achieved the best results on LA dataset for our features CLS-LBP (min-tDCF of 0.0017 and an EER and of 0.06%) and worst for LFCC (min-tDCF of 0.7423 and an EET of 76.99%). On the other hand, for the PA dataset our CLS-LBP features performed best (min-tDCF of 0.016 and an EER of 0.58) and LFCC was the worst (min-tDCF of 0.5829 and an EER of 49.86) when used with the LSTM classifier.

5. Conclusion

This paper has presented a robust voice spoofing detection system using the novel CLS-LBP features and LSTM to counter various LA and PA spoofing attacks. We proposed a novel features representation scheme CLS-LBP to effectively capture the attributes of bonafide speech dynamics, cloning algorithm artifacts, and microphone distortions of the replay signals. Experimental results on a large-scale and diverse ASVspoof 2019 corpus illustrate that the proposed system can reliably be used to detect various types of voice spoofing attacks. More specifically, our method achieved an EER of 0.06% and 0.58% on LA and PA attacks, respectively. Additionally, our system also detects the cloning algorithms used to generate the synthetic voices. Our comparative analysis reveals that our voice spoofing detection system provides better detection performance over state-of-the-art voice spoofing detectors. It is important to mention that the evaluation set of ASVspoof corpus includes the data of unseen speakers. We obtained remarkable results on the evaluation set of ASVspoof 2019 corpus that signify the effectiveness of our method for cross dataset evaluation. Our future work aims to investigate the performance of our method on cross dataset scenario using two different voice spoofing datasets entirely.

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.

Acknowledgement

The work was funded by the University of Jeddah, Saudi Arabia under Grant No (UJ-21-DR-28). The authors, therefore, acknowledge with thanks the university's technical and financial support.

References

Devon Delfino, Google smart lock Retrieved June 09, 2021, from: https://get.google.com/smartlock/.

Drew Harwell, An-artificial intelligence first: Voice-mimicking software reportedly used in a major theft. Retrieved June 19, 2021, from: https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking- software- reportedly-used-major-theft.

Alegre, F., Janicki, A., Evans, N. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In: 2014 International Conference of the Biometrics Special Interest Group (BIOSIG). 2014. IEEE.

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

Yamagishi, J. et al., 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio Speech Language Process 17 (1), 66–83.

Lindberg, J., Blomberg, M., 1999. Vulnerability in speaker verification-a study of technical impostor techniques. Sixth European Conference on Speech Communication and Technology.

Evans, N. et al., 2009. Anti-spoofing: voice conversion. Encycl Biometr, 1–10.

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

Witkowski, M. et al., 2017. Audio replay attack detection using high-frequency features. Interspeech.

Yang, J., Das, R.K., Li, H., 2018. Extended constant-Q cepstral coefficients for detection of spoofing attacks. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE.

Malik, H., 2019. Securing voice-driven interfaces against fake (Cloned) audio attacks. 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE.

Chettri, B., Sturm, B.L., 2018. A deeper look at Gaussian mixture model based anti-spoofing systems. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

Kamble, M.R., Patil, H.A., 2020. Novel variable length teager energy profiles for replay spoof detection. Energy 32, 33.

Malik, K.M. et al., 2020. A light-weight replay detection framework for voice controlled iot devices. IEEE J. Selected Topics Signal Process. 14 (5), 982–996.

Lin, L. et al., 2020. A robust method for speech replay attack detection. KSII Trans. Internet Inf Syst. 14 (1).

Kamble, M.R., H.A. Patil, *Detection of replay spoof speech using teager energy feature cues.* Computer Speech Language. 65: p. 101140, 2021.

Phapatanaburi, K. et al., 2020. Linear prediction residual-based constant-Q cepstral coefficients for replay- attack detection. 2020 8th International Electrical Engineering Congress (iEECON). IEEE.

Elsaeidy, A.A. et al., 2020. Replay attack detection in smart cities using deep learning. IEEE Access 8, 137825–137837.

Gritsenko, A.A., et al., *A spectral energy distance for parallel speech synthesis.* arXiv preprint arXiv:2008.01160, 2020.

Krishna, G. et al., 2020. Speech synthesis using eeg. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

Helali, W., Hajaiej, Z., Cherif, A., 2020. Real time speech recognition based on PWP thresholding and MFCC using SVM. Eng. Technol. Appl Sci. Res. 10 (5), 6204–6208.

Bird, J.J. et al., 2020. Overcoming data scarcity in speaker identification: dataset augmentation with synthetic MFCCs via character-level RNN. 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). IEEE.

Raju, K.P., Krishna, A.S., Murali, M., *Automatic speech recognition system using MFCC-based LPC approach with back propagated artificial neural networks.*

De Leon, P.L. et al., 2012. Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans. Audio Speech Language Process. 20 (8), 2280–2290.

Das, R.K., Yang, J., Li, H., 2019. Long range acoustic features for spoofed speech detection. Interspeech.

Das, R.K., Yang, J., Li, H., 2019. Long range acoustic and deep features perspective on ASVspoof 2019. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE.

Tak, H., et al., *An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification.* arXiv preprint arXiv:2004.06422, 2020.

Das, R.K., Yang, J., Li, H., 2020. Assessing the scope of generalized countermeasures for anti-spoofing. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

Zhai, L., Vamvoudakis, K.G., 2020. A data-based private learning framework for enhanced security against replay attacks in cyber-physical systems. Int. J. Robust Nonlinear Control.

Singh, M., Pati, D., 2020. Replay attack detection using excitation source and system features. In: Advances in Ubiquitous Computing. Elsevier, pp. 17–44.

Huang, L., Pun, C.-M., 2020. Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM network. IEEE/ACM Trans. Audio Speech Language Process. 28, 1813–1825.

Adiban, M., Sameti, H., Shehnepoor, S., 2019. Replay spoofing countermeasure using autoencoder and siamese networks on ASVspoof 2019 challenge. In: Computer Speech & Language, p. 101105.

von Platen, P., Tao, F., Tur, G., *Multi-Task Siamese Neural Network for Improving Replay Attack Detection.* arXiv preprint arXiv:2002.07629, 2020.

Gong, Y., Yang, J., Poellabauer, C., 2020. Detecting replay attacks using multi-channel audio: a neural network-based method. IEEE Signal Process. Lett.

Aravind, P., Nechiyil, U., Paramparambath, N., *Audio Spoofing Verification using Deep Convolutional Neural Networks by Transfer Learning.* arXiv preprint arXiv:2008.03464, 2020.

Wang, Q. et al., 2019. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE.

Zhang, C. et al., 2020. Improving replay detection system with channel consistency DenseNeXt for the ASVspoof 2019 challenge. In: Proc. Interspeech 2020, pp. 4596–4600.

Saranya, M., Murthy, H.A., 2018. Decision-level feature switching as a paradigm for replay attack detection. Interspeech.
Suthokumar, G., et al. *Modulation dynamic features for the detection of replay attacks.* in *Interspeech*- 2018.
Chettri, B., et al., *A study on convolutional neural network based end-to-end replay anti-spoofing.* arXiv preprint arXiv:1805.09164, 2018.
Białobrzeski, R. et al., 2019. Robust bayesian and light neural networks for voice spoofing detection. Proc. Interspeech 2019, 1028–1032.
Janyoi, P., Seresangtakul, P., 2020. Tonal contour generation for isarn speech synthesis using deep learning and sampling-based F0 representation. Appl. Sci. 10 (18), 6381.
Michelsanti, D., et al., *Vocoder-Based Speech Synthesis from Silent Videos.* arXiv preprint arXiv:2004.02541, 2020.
Valle, R., et al., *Flowtron: An Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis.* arXiv preprint arXiv:2005.05957, 2020.
Koriyama, T., Saruwatari, H., 2020. Utterance-level sequential modeling for deep gaussian process based speech synthesis using simple recurrent unit. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.
Zhou, X., Ling, Z.-H., Dai, L.-R., 2020. Learning and modeling unit embeddings using deep neural networks for unit-selection-based mandarin speech synthesis. ACM Trans. Asian Low-Resour. Language Inf. Process. (TALLIP) 19 (3), 1–14.
Lavrentyeva, G., et al., *Stc antispoofing systems for the asvspoof2019 challenge.* arXiv preprint arXiv:1904.05576, 2019.
Zeinali, H., et al., *Detecting spoofing attacks using vgg and sincnet: but-omilia submission to asvspoof 2019 challenge.* arXiv preprint arXiv:1907.12908, 2019.
Wu, Z. et al., 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures– challenge. Sixteenth Annual Conference of the International Speech Communication Association.
Kinnunen, T., et al., *The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.* 2017.
Retrieved October 20, 2020, from: Wang, X., et al., *ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech.* Computer Speech & Language, 2020: p. 101114.
Retrieved October 21, 2020, from VCKT database Veaux, C., J. Yamagishi, and K. MacDonald, *Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.* 2016.
Matrouf, D., Bonastre, J.-F., Fredouille, C., 2006. Effect of speech transformation on impostor acceptance. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. IEEE.
Morise, M., Yokomori, F., Ozawa, K., 2016. WORLD: a vocoder-based high-quality speech synthesis system- for real-time applications. IEICE Trans. Inf. Syst. 99 (7), 1877–1884.
Oord, A.v.d., et al., *Wavenet: A generative model for raw audio.* arXiv preprint arXiv:1609.03499, 2016.
Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoustics Speech Signal Process. 32 (2), 236–243.
Tanaka, K. et al., 2018. Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial-networks. 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE.
Wang, X., Takaki, S., Yamagishi, J., 2019. Neural source-filter-based waveform model for statistical parametric speech synthesis. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.
Kobayashi, K. et al., 2014. Statistical singing voice conversion with direct waveform modification based on the spectrum differential. Fifteenth Annual Conference of the International Speech Communication Association.
Janicki, A., Alegre, F., Evans, N., 2016. An assessment of automatic speaker verification vulnerabilities to replay-spoofing attacks. Security Commun. Networks 9 (15), 3030–3044.
Campbell, D., Palomaki, K., Brown, G., 2005. A matlab simulation of" shoebox" room acoustics for use in research and teaching. Comput. Inf. Syst. 9 (3), 48.
Novak, A., Lotton, P., Simon, L., 2015. Synchronized swept-sine: theory, application, and implementation. J. Audio Eng. Soc. 63 (10), 786–798.
Ko, T. et al., 2017. A study on data augmentation of reverberant speech for robust speech recognition. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.
Rectreived: November, 18, 2020. From the web page: http://homepages.loria.fr/evincent/software/Roomsimove 1.4.zip.
Todisco, M., Delgado, H., Evans, N., 2017. Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification. Computer Speech Language 45, 516–535.
Gomez-Alanis, A., Peinado, A.M., Gonzalez, J.A., Gomez, A.M., 2019. A Light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. Proc. Interspeech 2019, 1068–1072.
Li, X., et al., Replay and Synthetic Speech Detection with Res2net Architecture. arXiv preprint arXiv:2010.15006, 2020.
Gomez-Alanis, A., Gonzalez-Lopez, J.A., Peinado, A.M., 2020. A kernel density estimation-based loss function and its application to ASV-spoofing detection. IEEE Access 8, 108530–108543.
Aljasem, M. et al., 2021. Secure Automatic Speaker Verification (SASV) System through sm-ALTP features and asymmetric bagging. IEEE Transactions on Information Forensics and Security.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput 9 (8), 1735–1780.
Alluri, K.R., Vuppala, A.K., 2019. IIIT-H spoofing countermeasures for automatic speaker verification spoofing and countermeasures challenge 2019. Interspeech.
Alluri, K.R. et al., 2017. Detection of replay attacks using single frequency filtering cepstral coefficients. Interspeech.
Available online on 1/11/2022, accessed online at: https://machinelearningmastery.com/softmax-activation-function-with-python/.
Hassan, F., Javed, A., 2021. Voice spoofing countermeasure for synthetic speech detection. In: 2021 International Conference on Artificial Intelligence (ICAI). IEEE, pp. 209–212.
Qadir, G., Zareen, S., Hassan, F., Rahman, A.U., 2022. Voice spoofing countermeasure based on spectral features to detect synthetic attacks through LSTM. Int. J. Innovat. Sci. Technol. 3, 153–165.
Banaras, Y., Javed, A., Hassan, F., 2021. Automatic speaker verification and replay attack detection system using novel glottal flow cepstrum coefficients. In: 2021 International Conference on Frontiers of Information Technology (FIT), pp. 149–153. https://doi.org/10.1109/FIT53504.2021.00036.