2021 International Conference on
Artificial Intelligence (ICAI)
Islamabad, Pakistan, April 05-07, 2021

Voice Spoofing Countermeasure for Synthetic Speech Detection

Farman Hassan[1]

[1]Department of Software Engineering, University of Engineering and Technology,
Taxila 47050, Pakistan.
farmanhassan555@gmail.com

Ali Javed[2]

[2]Department of Software Engineering, University of Engineering and Technology,
Taxila 47050, Pakistan.
ali.javed@uettaxila.edu.pk

***Abstract*: In the last few years, we have witnessed an exponential growth in voice spoofing attacks. The intruders employ different types of attacks such as speech synthesis where they use the machine generated speech against any target person to fool the automatic speaker verification (ASV) systems for various tasks i.e. home control, bank account access, etc. The availability of modern-day advanced tools has made it convenient to launch such types of voice spoofing attacks. To overcome the challenges associated with bypassing the security of ASV systems using the synthetic speech, we propose an effective synthetic speech detector using a fusion of spectral features. More specifically, we propose a fused feature vector consisting of MFCC, GTCC, Spectral Flux, and Spectral Centroid for audio signal representation. This fused feature set is capable of capturing the traits of speech variation attributes of genuine signal and algorithmic artifacts of synthetic signals. These features are further used to train the bilstm to classify the signal as genuine or spoof. The proposed framework is capable of detecting both the voice conversion and synthetic speech attacks on ASV systems. Performance of our framework is evaluated on ASVspoof 2019 LA dataset. Our experimental results illustrate the effectiveness of the proposed framework for logical access attacks (voice conversion and cloned/synthetic voice) detection.**

***Keywords*: Artificial Intelligence, BiLSTM, Deep Learning, Synthetic speech, Spoofing countermeasure, Voice conversion.**

2021 International Conference on Artificial Intelligence (ICAI) | 978-1-6654-3293-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAI52203.2021.9445238

I. INTRODUCTION

An automatic speaker verification system (ASV) either accepts or rejects the claim of speaker's identity. The authorization is based on the human voice and it is becoming much popular authorization method for the user's identity to authorize access specially these days due to COVID-19. Logical access (LA) attacks such as speech synthesis or speech conversion presents a genuine threat to the ASV systems. By using voice conversion and speech synthesis algorithms, the speech of an original user can be manipulated to breach the security of ASV systems and get an access of some person's bank account or home. Currently, there are only two speech synthesis techniques, statistical parametric speech synthesis [1] and waveform concatenation with unit selection [2]. Statistical speech model is built for predicting speech parameters from user input texts and predicted speech features are employed into a spoof system i.e. vocoder to recreate the acoustic waveforms. This technique is capable of producing highly similar speeches to human beings, but the quality of synthetic speech degrades due to inaccuracy of speech parameters prediction and by the use of vocoder. Speech synthesis uses another technique by concatenating waveform with unit selection i.e. syllables, phones or frames. The idea behind this approach is to select sequence of unit candidates from recorded audio and then append the waveforms of selected units to create a converted or synthetic voice to fool the ASV systems. There exists a strong need to develop a robust synthetic speech detection system that can reliably be used to detect both the voice cloning and conversion attacks.

Existing voice spoofing countermeasures have employed various deep learning models for logical access attacks detections. In [3], deep neural network was used to extract the s-vector features that were then used to train the DNN for voice spoofing detection. Deep-neural-networks have been considered to be highly effective to solve complex artificial intelligence problems including voice-based biometrics applications [4]. Additionally, the computational mechanism of deep learning models make them equally effective for both the back end classifier [5, 6] as well as for feature extraction [7]. The design architecture of the above-mentioned deep feature extractors has been demonstrated to be a determining factor for the development of the voice spoofing countermeasures. In [8], a hybrid light convolutional neural network (LCNN) was employed in addition to the recurrent neural network (RNN) model. The discriminatory attributes of LCNN at the frame-level was combined with the gated recurrent unit (GRU) based RNNs to effectively learn the long-term dependencies in the input audio signal. The resultant design was referred as the Light Convolutional Gated Recurrent Neural Network (LC-GRNN). Traditional log MEL filterbank features were used in [9] with the CNN as back-end classifier to distinguish between the genuine and spoof audio.

Traditional voice spoofing countermeasures [10,11,12,13,14] have been proposed to address logical access attacks using the Gaussian Mixture Model. Hand-crafted features pay attention to feature engineering i.e. Cochlear Filter Cepstral Coefficients Instantaneous Frequency [10], Linear Frequency Cepstral Coefficients (LFCC) [11], and Constant-Q Cepstral Coefficients (CQCC) [12]. CQCC [15] and LFCC [16] features are used with the Gaussian Mixture Model (GMM) classifier to detect the converted and synthetic speeches in the ASVspoof 2019 baseline model. CQCC and LFCC are derived from the magnitude spectrum. These spectral features are reasonable in terms of capturing the traits of the genuine and spoofed audio samples, however, the classification performance of the GMM can be drastically improved by considering the deep learning models. In this paper, we employed a fusion of four spectral features i.e. GTCC, MFCC, Spectral Flux and Spectral Centroid to better select the discriminating attributes of the genuine and spoof samples. Next, we employed the bilstm deep learning model to train these features for improved classification of the genuine and spoofed (synthetic and converted) speeches.

The main contributions of our work are:

- We introduced a spectral features fusion set consisting of GTCC, MFCC, Spectral Flux and Spectral Centroid for input audio presentation.
- The proposed countermeasure successfully detects multiple types of logical access attacks.
- The proposed synthetic spoofing detection system successfully classifies the cloning algorithms used to produce the synthetic and converted speeches from the genuine audio samples.

The remaining paper is organized as follows. Section 2 presents the description of the proposed system. Section 3 provides the details of the dataset and results of the experiments conducted for performance evaluation of the proposed system. Finally, we conclude our work in Section 4.

978-1-6654-3293-1/21/$31.00 ©2021 IEEE

II. PROPOSED SYSTEM

The proposed synthetic spoof detection system takes the audio as input and extract the 30-dimensional integrated spectral features comprising of 14 dimensional MFCCs, 14 dimensional GTCC, 1 dimensional spectral centroid, and 1 dimensional spectral flux. For classification purpose, we train the BiLSTM model using the fused features to classify the genuine and spoof speeches. The process flow of the proposed system is presented in Fig. 1.

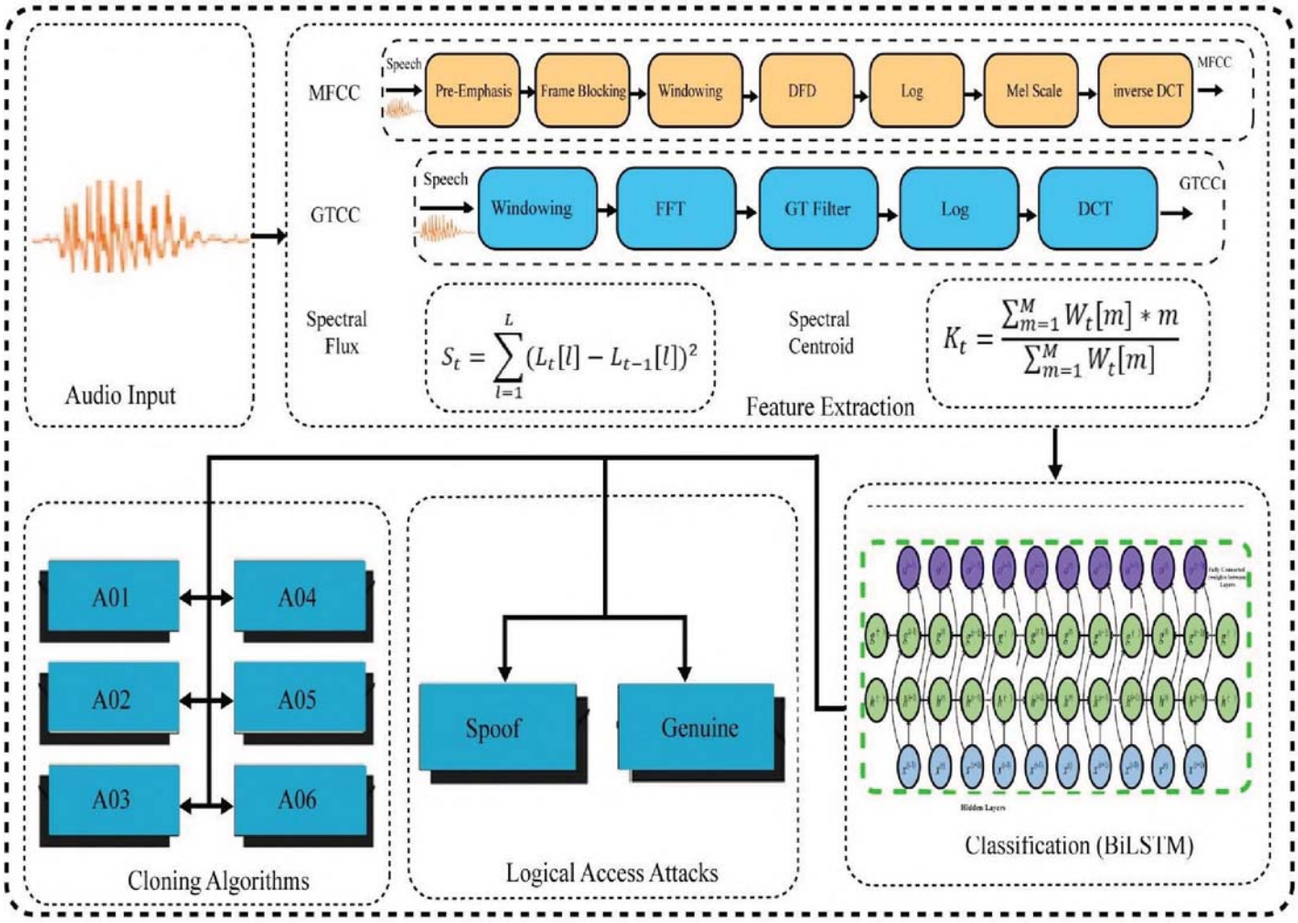

Fig 1: Proposed System.

2.1. Feature Extraction

Effective features extraction is an important step to develop a reliable voice spoofing countermeasure. The extraction of the finest parametric depiction of audio signals is significant to achieve an improved identification performance. For both the LA attacks and cloning algorithms classification, we employed the fused features set comprising of MFCC, GTCC, Spectral Flux, and Spectral Centroid to represent an audio. The details of these features are provided in this section.

2.1.1. MFCC

We extracted the MFCC features by pre-emphasis where speech signal is sent to high-pass filter to balance the spectrum of voice sounds followed by frame blocking in which the speech signal is segmented into frames to examine the speech over a short period of time. In the next step, signals are converted into frames by applying the hamming window to ensure smooth edges, enhanced harmonics, and to decrease the edge effect. Next, we employed the discrete Fourier Transform to convert the speech into magnitude spectrum. The powers of spectrum obtained after computing the Fourier transform is mapped onto mel-scale using the triangular overlapping windows. Next, we computed the log of the powers of spectrum at each mel-frequency. Later, we calculated the Discrete Cosine Transform (DCT) of the list of Mel log powers.

Finally, the amplitude of the resulting spectrum is selected as the MFCCs.

2.1.2. GTCC

We extracted the GTCC features by applying Fast Fourier Transform (FFT) on each speech window followed by employing gammatone filter bank consisting of 48 GT filters to FFT of speech signals and energy of every sub-band is calculated. In next stage, logarithm of each sub-band is calculated followed by employing DCT.

2.1.3. Spectral Flux

Spectral Flux compares the power spectrum of one window against the power spectrum from lagging frame of an audio. We computed the spectral Flux as the square difference between normalized magnitudes and of successive spectral distributions.

$$= \sum_{=1} (\quad [\,] - \quad _{-1}[\,])^2 \qquad (1)$$

where $[\,]$ and $_{-1}[\,]$ are the normalized magnitude of Fourier Transform (FT) at current window *t*, and previous window *t-1*.

2.1.4. Spectral Centroid

Spectral Centroid represents center of gravity of complete power spectrum and the energy distribution across low and high frequency bands. It is computed as below.

$$= \frac{\sum_{=1} \quad [\,] *}{\sum_{=1} \quad [\,]} \qquad (2)$$

where the $[\,]$ is the magnitude of FT at current window *t* and frequency bin *m*.

2.2. Classification

We designed BiLSTM for classification purpose and configured various input parameters for training the network. We used a solver for training, an adam optimizer and set maximum number of epochs to 50, the mini-batch size with 64 observation at every iteration and the gradient threshold value set to 1. The proposed BiLSTM model consists of 100 hidden units, 10 BiLSTM layers, one fully connected layer followed by a SoftMax layer used for classification.

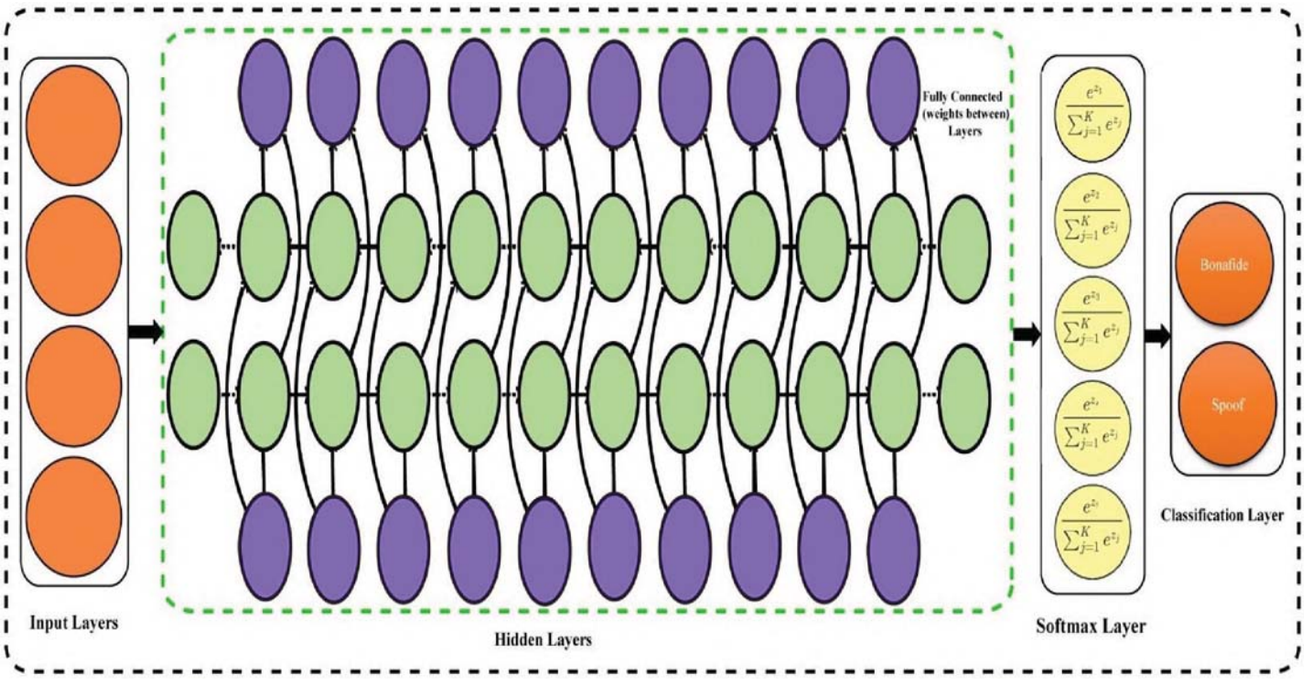

Fig 4: BiLSTM Architecture.

III. EXPERIMENTAL SETUP AND RESULTS

3.1. Dataset

We evaluated the performance of the proposed system on ASVspoof 2019 LA dataset. LA corpus comprises of training, development and evaluation sets. We used the training set for training and evaluation set for testing the model. The details of the genuine and spoof samples used for training and testing are provided in Table 1. Total numbers of 29 TTS/VC attacks have been created for LA database. We used 15,981 audio samples for training and 14,161 samples for testing purposes.

Six cloning algorithms consisting of 2 Text to speech (TTS) and 4 voice conversion (VC) are used to synthesize the genuine samples of ASVspoof 2019 LA dataset. We employed the equal error rate (EER), precision, recall, accuracy, and F1-score to measure the performance of our system.

Table 1. Statistics of ASVspoof 2019 LA dataset.

Logical Access	Training samples	Evaluation samples	#Speaker	
Total samples	15,981	14,161	Male	Female
Bonafide samples	2,580	2,580	8	12
Spoof samples	13,401	11,581	8	12

3.2. Performance Evaluation for Logical Access Attacks Detection

The objective of this experiment is to evaluate the performance of the proposed method on logical access attacks detection. For this purpose, we employed the 30-dimensional spectral fused features set (MFCC + GTCC + Spectral Flux + Spectral Centroid) to train the custom bilstm model for classification of the genuine and spoof (voice cloning and conversion) samples. We used the LA collection of ASVspoof 2019 dataset for measuring our system performance against both the LA attacks. More specifically, we achieved an EER of 3.05%, accuracy of 96.9%, precision of 96.40%, recall of 1, and F1-score of 98.16%. The proposed system outperforms the ASVspoof baseline models by achieving 5.52% and 5.03% lesser EER than CQCC+GMM [15] and LFCC+GMM [16]) respectively. It can be concluded from the results that the proposed spectral features fusion can accurately capture the speaker induced characteristics of the genuine signal and artifacts of the synthetic and converted speech signals.

3.3. Performance Evaluation of Cloning Algorithms Classification

We designed an experiment to investigate the effectiveness of the proposed system in terms of classifying the cloning algorithms used to synthesize the genuine signals. For this experiment, we used 6000 samples of cloning algorithms for model training and 6,000 samples for model testing. We used 1000 samples of each algorithm for training and same for testing. We designed a six class bilstm model for classification of cloning algorithms and results are reported in Table 2. From the Table 2, we can observe that the proposed system performs well on A05 and A06 VC spoofing systems and achieved an EER of 0.1%, whereas, performance of the proposed system is degraded to some extent on A03 TTS spoofing system where we achieved an EER of 2.7%. From the results, we can conclude that the proposed system better detects the voice conversion spoofing over synthetic speech spoofing.

Table 2. Performance evaluation of cloning algorithms classification.

Algorithms	EER%	Accuracy%	Precision%	Recall%	F1 Score%
A01 TTS	0.3%	98.5%	99.1%	100%	99.55%
A02 TTS	2.6%	99.6%	92.76%	100%	96.24%
A03 TTS	2.7%	96.2%	100%	91.9%	95.77%
A04 TTS	0.4%	95.0%	100%	98.8%	99.39%
A05 VC	0.1%	97.6%	99.7%	100%	99.85%
A06 VC	0.1%	95.4%	99.7%	100%	99.85%

3.4. Performance Comparison

We compared the performance of the proposed spoofing countermeasure with other state-of-the-art methods based on EER. For this purpose, we compared the performance of our method against [15,16,17,18] methods, and results are reported in Table. 3. We can observe that our method achieves the lowest EER of 3.05% over comparative models. LFCC-LCNN performs second best and achieves an EER of 5.06%, whereas, the baseline model (CQCC+GMM [15] and LFCC+GMM [16]) achieves the highest EER of 9.57% and 8.09%. These results show that the proposed system outperforms the other state-of-the-art methods and can reliably be used to detect the logical access attacks.

Table 3. Performance comparison with other methods.

System	EER (%) on Eval Set
CQCC + GMM	9.57
LFCC + GMM	8.09
LC-GRNN + SVM	7.12
LC-GRNN + PLDA	6.34
LC-GRNN + LDA	6.28
LFCC-LCNN	5.06
Proposed	**3.05**

IV. CONCLUSION

This paper has presented a spoofing countermeasure to detect both types of LA attacks. We proposed a fused spectral feature set (i.e. GTCC, MFCC, spectral flux, and spectral centroid) and trained the bilstm RNN model for classification of the genuine and spoofed speeches. Additionally, the proposed method is capable of identifying the type of cloning algorithm used to generate the synthesized speech. We measured the performance of our method on a diverse voice spoofing dataset i.e. ASVspoof 2019-LA. The results of our experiments revealed that the proposed system is capable of identifying the discriminative traits of both the genuine and spoofed samples. Our method gives better classification performance than state-of-the-arts methods and shows considerable improvement in terms of EER value which is 6.52% lower than the baseline [15]. In future work, we aim to develop a voice spoofing countermeasure to address both the physical and logical access attacks.

V. REFRENCES

[1] Heiga Zen, Keiichi Tokuda, and Alan W Black. 2009. Statistical parametric speech synthesis. Speech Communication 51, 11 (2009), 1039–1064.

[2] Andrew J Hunt and Alan W Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, Vol. 1. IEEE, 373–376.

[3] Chen, N., et al. *Robust deep feature for spoofing detection—The SJTU system for ASVspoof 2015 challenge*. in *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[4] Yadav, S. and A. Rai. *Learning Discriminative Features for Speaker Identification and Verification*. in *Interspeech*. 2018.

[5] Tian, X., et al. *Spoofing detection from a feature representation perspective*. in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016. IEEE.

[6] Zhang, C., et al. *Joint information from nonlinear and linear features for spoofing detection: An i-vector/DNN based approach*. in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016. IEEE.

[7] Gomez-Alanis, A., et al., *Performance evaluation of front-and back-end techniques for ASV spoofing detection systems based on deep features*. Proc. Iberspeech, 2018: p. 45-49.

[8] Wu, X., et al., *A light cnn for deep face representation with noisy labels*. IEEE Transactions on Information Forensics and Security, 2018. **13**(11): p. 2884-2896.

[9] Gomez-Alanis, A., et al. *A deep identity representation for noise robust spoofing detection*. in *Proc. Interspeech*. 2018.

[10] Tanvina B Patel and Hemant A Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in Sixteenth Annual Conference of the International Speech Communication Association.

[11] Massimiliano Todisco, Hector Delgado, and Nicholas Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in Proc. Odyssey, 2016, vol. 45, pp. 283–290.

[12] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in Sixteenth Annual Conference of the International Speech Communication Association

[13] Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, Daniel Erro, and Tuomo Raitio, "Toward a universal synthetic speech spoofing detection using phase information," IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 810–820, 2015.

[14] Wu, Z., et al., *ASVspoof: the automatic speaker verification spoofing and countermeasures challenge*. IEEE Journal of Selected Topics in Signal Processing, 2017. **11**(4): p. 588-604.

[15] Todisco, M., H. Delgado, and N. Evans, *Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification*. Computer Speech & Language, 2017. **45**: p. 516-535.

[16] Sahidullah, M., T. Kinnunen, and C. Hanilçi, *A comparison of features for synthetic speech detection*. 2015.

[17] Zhang, Y., *ONE-CLASS NEURAL NETWORK FOR ANTI-SPOOFING IN SPEAKER VERIFICATION*.

[18] Gomez-Alanis, A., et al. *A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection*. in *Proc. Interspeech*. 2019.