AEXANet: An End-to-End Deep Learning based Voice Anti-spoofing System

Zarish Anwar [1[0000-0001-5623-3610]], Ali Javed [2*[0000-0002-1290-1477]] and Khalid Mahmood Malik [3[0000-0002-7927-3436]]

[1,2] Software Engineering Department, University of Engineering and Technology, Taxila, Punjab, Pakistan
zarish.anwar@students.uettaxila.edu.pk
ali.javed@uettaxila.edu.pk
[3]Dept. of Computer Science & Engineering, Oakland University, Rochester, MI, USA
mahmood@oakland.edu

Abstract. Exponential growth in the use of smart speakers (SS) for the automation of homes, offices, and vehicles has brought a revolution of convenience to our lives. However, these SSs are susceptible to a variety of spoofing attacks, known/seen and unknown/unseen, created using cutting-edge AI generative algorithms. The realistic nature of these powerful attacks is capable of deceiving the automatic speaker verification (ASV) engines of these SSs, resulting in a huge potential for fraud using these devices. This vulnerability highlights the need for the development of effective countermeasures capable of the reliable detection of known and unknown spoofing attacks. This paper presents a novel end-to-end deep learning model, AEXANet, to effectively detect multiple types of physical- and logical-access attacks, both known and unknown. The proposed countermeasure has the ability to learn low-level cues by analyzing raw audio, utilizes a dense convolutional network for the propagation of diversified raw waveform features, and strengthens feature propagation. This system employs a maximum feature map activation function, which improves the performance against unseen spoofing attacks while making the model more efficient, enabling the model to be used for real-time applications. An extensive evaluation of our model was performed on the ASVspoof 2019 PA and LA datasets, along with TTS and VC samples, separately containing both seen and unseen attacks. Moreover, cross corpora evaluation using the ASVspoof 2019 and ASVspoof 2015 datasets was also performed. Experimental results show the reliability of our method for voice spoofing detection.

Keywords: ASVspoof2019, logical access, physical access, spoofing countermeasure, text-to-speech synthesis, voice conversion.

1 Introduction

Smart speakers (SS) equipped with intelligent voice assistants, like Google Home, Apple's Siri, and Amazon's Alexa are being used nowadays to "smarten" our offices, homes, and automobiles through the use of speaker verification systems. These sys-

tems have become an integral part of cyber intelligent systems through the inclusion of cutting-edge and highly precise knowledge engines for speaker verification. Automatic Speaker Verification (ASV) systems are routinely used to accept or reject the speaker's claimed identification. Although the quality of modern-day ASV systems has increased significantly, they continue to be susceptible to audio spoofing attacks due to the extremely sophisticated nature of synthetic speech generative algorithms. The ASVspoof community has categorized voice spoofing attacks into two main types: physical access (PA) and logical access (LA) [1]. The PA scenario involves speech samples that are captured in a physical reverberant space. Attacks use replayed samples in a simulated setup, captured surreptitiously by recording bonafide speech, and then replayed to the microphones of ASV systems. On the other hand, spoofing attacks are directly injected into the ASV system in the LA scenario. These attacks are generated either by text-to-speech synthesis (TTS) or voice conversion (VC) algorithms. These attacks generate samples that are perceptually identical to the real voice of a verified subject. TTS synthesis spoofing uses the text command, whereas VC uses the audio samples as input and feeds to the generative algorithms for the creation of LA attacks. High-quality digital recording and playback devices, minimal effort for replay creation, and advanced AI generative algorithms have encouraged attackers to produce and use these spoofing attacks for scamming. Thus, anti-spoofing systems capable of the reliable detection of both LA and PA attacks are strongly needed. These countermeasures have many applications across multiple voice biometric [2,3] domains.

Existing voice spoofing detection approaches use evolving reliable acoustic spectral characteristics with either the traditional machine learning (ML) or the deep learning (DL) algorithms. Current acoustic features are based on factors like pitch pattern, phase spectrum, group delay, and spectral magnitude. The Gaussian mixture model (GMM) and its variations, along with the support vector machine (SVM) classifiers [4] [5], are being heavily leveraged for audio spoofing detection. Techniques that use pitch patterns for detection are mean pitch stability range (MPSR) and mean pitch stability (MPS) [6]. Time-domain acoustic features such as local binary patterns (LBP) [7] and our previously proposed acoustic ternary patterns (ATP) [8] have also been used to develop voice spoofing countermeasures. However, LBP features are more sensitive to noise, and ATP features, by employing a fixed threshold value, are not robust to dynamic pattern detection, and thus, they are unable to achieve better performance in a real-time scenario. Another work [9] highlights the importance of phase information for voice spoofing detection, which can be derived using the Fourier spectrum and its fusion with other existing phase-based features. Existing ML-based approaches have potential limitations concerning time complexity, inaccurate data interpretation, and high error-susceptibility. All must be taken into consideration when creating a successful spoofing detection application.

The ASVspoof research community also employed DL models as front-end feature extractors e.g., Gated Recurrent neural network (GRNN) is the fusion of Light convolutional neural network and gated recurrent RNNs (GRNN) [10]. A GRNN extracted the deep features and used them to train classifiers like SVM, linear discriminant analysis (LDA), etc., for voice spoofing detection. Another technique involved using

backend classifiers with various acoustical features, e.g., Mel frequency cepstral coefficients (MFCC), constant Q cepstral coefficients (CQCC), short-time Fourier transform (STFT), and their fusion, for audio spoofing detection. To overcome the issues of our ATP features, we proposed extended-local ternary patterns (ELTP) in our prior work [11] and used them with a bidirectional LSTM for LA attack detection. Although we successfully addressed the limitations of our ATP-based method, this system was unable to achieve good results on VC samples. These deep learning variants showed significant performance improvements over many baseline spoofing detection approaches developed by the ASVspoof community, i.e., CQCC-GMM, and LFCC-GMM [5].

More recently, we have seen a trend in the development and usage of frontend feature extractors with either traditional classifiers, like GMM, or DL classifiers like an LSTM/BiLSTM. A recent study has shown that frontend features-based spoofing detectors often fail to generalize better to unseen attacks [12]. Another observation has shown that despite the performance of unforeseen attacks is not comparable to, or even superior to that of known attacks, there is significant variation in performance for known attacks of diverse nature [13]. To address these challenges, the ASV community has worked on developing effective end-to-end DL detectors for various classes of spoofing attacks, including unseen attacks. The use of these newly optimized end-to-end representations has shown an assembly of multiple frequency responses, extracted at every kernel of convolutional layers [14], instead of using general fixed-bandwidth decomposition methods such as Fourier-based analysis [15].

In view of these challenges, we present an end-to-end DL-based anti-spoofing model. In addition, we address the limitations of our previous countermeasure [11] by enhancing the detection performance on VC attacks. This work presents a reliable countermeasure for the effective identification of seen and unseen spoofing attacks. The significant contributions of our work are in the following areas:

- We propose an effective end-2-end DL-based voice anti-spoofing system, AEXANet, which is capable of capturing a low-level representation of raw waveforms to effectively detect multiple types of seen and unseen spoofing attacks.
- Our proposed model employs a dense convolutional network, which gives a compelling advantage for propagating diversified raw waveform features with minimal complexity and strengthens feature propagation.
- Our proposed system employs a maximum feature map (MFM) activation function which makes it computationally more efficient.
- Rigorous experiments were performed on the PA and LA collections including the cross-corpora evaluation to show the strength of our method for voice spoofing detection.

2 PROPOSED COUNTERMEASURE

This section shows the details of our Audio Examiner RawNet model (AEXANet) for voice spoofing detection (Fig. 1). ASVspoof baseline solutions are unable to achieve better detection on most forms of spoofing attacks and are less robust to dynamic

voice spoofing patterns [16]. Our model extracts dense features from the raw waveforms in the higher layers of the dense convolutional network, which gives an advantage in improved detection performance against diverse spoofing attacks. We introduce an MFM activation function in our architecture, which demonstrates robustness against unseen spoofing attacks and helps to learn audio cues easily.

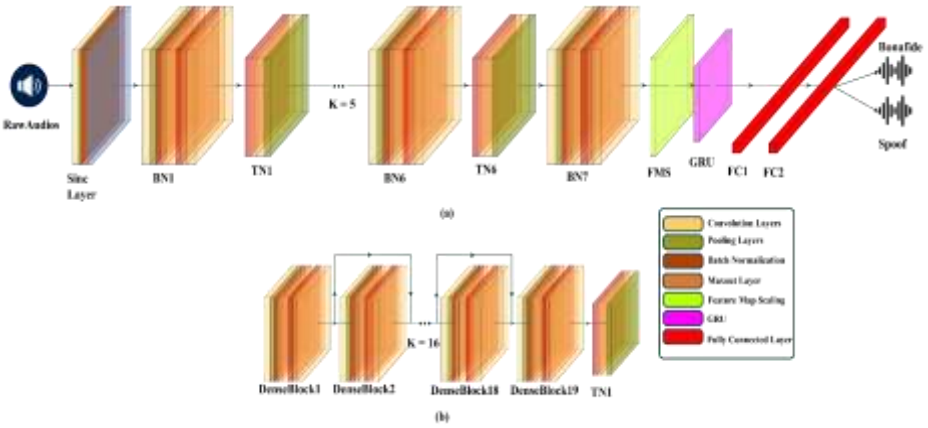

Fig. 1. Architecture diagram of proposed AEXANet model.

* Fig 1(a). shows the bottleneck layer (BN) and transition block (TN), and K = 5 indicates five more bottlenecks and transition layers in the architecture, (b) shows the working of densenet blocks in every bottleneck layer; K = 16 shows 16 additional densely connected blocks (not shown), for a total of 19 in the network.

2.1 Details of the AEXANet Architecture

AEXANet is a neural speaker embedding extractor that takes raw waveforms as input and generates speaker embeddings specifically designed for speaker verification. We employed our deep neural network (DNN) to directly derived the speaker embeddings from raw waveforms with hidden layers to yield more discriminative information. The first layer that makes use of the raw audio is SincNet, which makes use of a parametrized sinc function to implement the band-pass filters. Network usage is encouraged because the first convolutional layer provides informative filters within higher and lower cut-off frequencies. The SincNet architecture was created for speaker and speech recognition tasks, and therefore we consider it to be suitable for audio examination, especially for the artifacts spoofed by TTS and VC attacks. A DenseNet connectivity pattern was introduced in [17] to enhance the optimal information flow within higher layers. The network enhances computational efficiency by utilizing the same feature map size to generate direct connections from any layer to all of the preceding layers. The l^{th} layer receives the feature-maps of all preceding layers $x_0, \ldots, x_{l-1}$ as input:

$$x_l = H_l\left([\, x_0, x_1, \ldots, x_{l-1}]\right) \quad (1)$$

The concatenation of the feature maps produced in layers $0, \ldots, l-1$ is denoted by $[\, x_0,\ x_1, \ldots, x_{l-1}\,]$. $H_l\,(.)$ represents the concatenation of several inputs into a single tensor. $H_L(.)$ is a composite function created using three consecutive operations: batch normalization (BN), max feature map [5], and convolution (Conv) layer. For consistent and improved performance, the DenseNet architecture uses computationally efficient transition blocks after every dense block. This reduces feature redundancy and optimizes parameter learning. The transition layers of our model use a BN layer, maxout layer, and 3 × 3 convolutional layers, followed by 3 × 3 average pooling layers. The DNN ensures maximum information flow between the deep layers, allowing them to rely more on high-level characteristics rather than low-level features.

Further improvement to the network can be achieved by progressively raising the growth rate. The increasing growth rate (IGR) technique places more parameters in the deeper layers of the model. This may decrease the parameter efficiency and increases the computational efficiency substantially in some cases. To make our model computationally more efficient, we employ BN and MFM [18] as an activation function to further down-sample the network layers. MFM activation is introduced in the proposed model to provide a more effective replacement for non-linear activation functions such as ReLU. Existing threshold-based non-linear functions can cause information loss and may not generalize well to unknown data distributions, particularly in the first few convolutional layers. To solve this problem, MFM uses an elementwise max operation instead of ReLU's non-linearity to build an economical connection between feature maps without any dependency on the threshold or parameters. This makes our end-2-end spoofing detector model generalize well even for different data distributions. Moreover, MFM initializes layers with the same dimension and selects the most important values among items in the layers. This improves the efficiency of our model and makes it lighter. The MFM applied feature map is created elementwise by applying the $max\ (a_1, a_2)$ as:

$$y_{i,j}^{k} = max(x_{i,j}^{k} + x_{i,j}^{k+n}) \quad (2)$$

For an input convolution layer $x^n \in R^{H \times W}$, W and H represent the height and width of input tensor where $n = \{1, \ldots, 2N\}$. The channel of the input convolution layer is $2N, 1 \le k \le N,\ 1 \le i \le H,\ 1 \le j \le W$. The output $y_{i,j}^{k}$ via the MFM function is $R^{H \times W \times N}$. The outputs of the denseNet are fed independently to the FMS layers, where the most informative filters are extracted using a hybrid additive and multiplicative feature scaling technique [19]. A gated recurrent unit (GRU) architecture with 1024 hidden nodes aggregates the frame-level features into a speech-level representation that precedes the fully connected layers, giving the final timestep. Finally, a softmax function is used to predict the output class, i.e., bonafide or spoof. The complete AEXANet architecture is presented in Table 1.

3 PERFORMANCE EVALUATION

The performance of our system is evaluated using the ASVspoof 2019 LA and PA datasets. We used the following parameters to train our AEXANet model: an ADAM optimizer, 100 epochs, a learning rate of 0.0001, with a mini-batch size of 8 for training and 16 for testing. The minimum tandem detection cost function (min t-DCF) and equal error rate (EER) were used as the primary and secondary metrics, per the evaluation plan of the ASVspoof 2019 dataset [20]. Thus, we also used min t-DCF and EER for performance evaluation. Moreover, we used the ASVspoof 2015 dataset for cross-corpus evaluation.

Table 1. DETAILS OF AEXANET ARCHITECTURE

Layers	Input≈64000 samples	Output shape
Fixed Sinc filters	Conv(1024,1,20) MaxPooling(3) BN & LeakyReLU	(21192,20)
BottleNeck block	Conv(7,1,20) MaxPooling(3) BN & Maxout Conv(1,1,20) BN & Maxout Conv(7,1,20) *6	(5296,20)
Transition block	BN & Maxout Conv(3,1,20) AvgPooling(3) *6	
BottleNeck block	Conv(7,1,20) MaxPooling(3) BN & Maxout Conv(1,1,20) BN & Maxout Conv(7,1,20)	(165,20)
	FMS	
GRU	GRU (1024)	(1024)
FC	1024	(1024)
Output	1024	2

*For convolutional layers, numbers inside parentheses refer to kernel size, stride size, and the number of filters. Each bottleneck is densely connected with 20 dense blocks following a transition block. AEXANet comprises seven bottleneck blocks and six transitions. For the gated recurrent unit (GRU) and fully connected layers, numbers inside the parentheses indicate the output for the voice conversion dataset.

3.1 Dataset

ASVspoof 2019 is a large and diverse audio spoofing dataset comprising two main collections, LA and PA. Each of these two collections is further broken into three independent partitions i.e., training, development, and evaluation. The bonafide and spoof samples are generated using 17 diverse TTS and VC systems for the LA subset. The evaluation set contains 2 known and 11 unknown spoofing attacks. The training and development sets contain only known spoofing attacks [21]. The PA subset is developed in a reverberant acoustic environment and simulates [11] offering replay

samples to the microphone of an ASV system. There are 27 distinct acoustic and 9 different replay configurations in the training and development sets. The evaluation set is generated in the same way but uses varying acoustics and playback configurations. Table 2 shows the statistics of the ASVspoof 2019 dataset.

Table 2. STATISTICS FOR ASVSPOOF 2019 LA AND PA DATASET

Subsets	**Logical Access**		**Physical Access**	
	Bonafide	Spoofed	Bonafide	Spoofed
	#Utterances	*#Utterances*	*#Utterances*	*#Utterances*
Training	2,580	22,800	5,400	48,600
Development	2,548	22,296	5,400	24,300
Evaluation	7,355	63,882	18,090	199,367

3.2 Performance Evaluation of Proposed Countermeasure

Effectiveness of the proposed countermeasure is evaluated on the VC, TTS, LA, and PA datasets. The objective of these experiments is to measure the countermeasure's ability to learn and detect spoofed speech. Experiments are conducted separately using the AEXANet model for VC, TTS, and complete LA and PA sets, and the results are shown in Table 3. Our experiments attained an EER and min t-DCF of 12.10% and 0.40 for VC, 0.60% and 0.08 on TTS, 4.93% and 0.17 on the LA evaluation subset, and 5.29% and 0.2 for the PA evaluation subset, separately. From Table 3, we can argue that our AEXANet countermeasure is very effective in terms of classifying the TTS LA spoofing. The reason better performance is achieved on TTS is the synthetic nature of the text data in a digitalized and completely synthetic structure. TTS models use RNNs for waveform generation and have sequential data patterns with rich information which makes successful prediction more likely. Thus, our model has less accuracy against voice conversion than other sets. NN-based VC systems use VAE-based, GMM-UBM, and i-vector PLDA with MFCC acoustic models. These models not only preserve the prosodic characteristics of the speaker but also create realistic spoofed pitch. This in turn makes it difficult for countermeasures to detect channel variability and increases the impostor acceptance rate in ASV systems [22]. Based on our prior work [11], we have attained improved results by reducing the EER to 21.18% for VC spoofing.

For evaluation of the LA set, our proposed model exhibited a substantial improvement over other systems with an EER of 4.9%. Even the best performing system suggests that reliable performance against the LA scenario depends upon the fusion of complementary sub-systems with an ensemble of classifiers because of the diversity of attacks (TTS, VC, and hybrid).

For the PA scenario, we also achieved good results despite the fact that the PA-Eval set used for testing contains samples of unseen speakers, recording, and playback devices. Moreover, the replay spoofing attacks in the PA dataset are generated according to different replay configurations (acoustic environment and replay devices) rather than with different spoofing algorithms. These results, considering this degree of diversity and the challenging conditions, show the robustness of the proposed

method for replay spoofing detection against both seen and unseen attacks. Table 3 shows the results of the proposed countermeasure in terms of min t-DCF and EER for VC, TTS, LA, and PA evaluation subsets.

Table 3. PERFORMANCE EVALUATION OF PROPOSED COUNTERMEASURE

Spoofing Category	min t-DCF	EER (%)
Text-to-Speech (TTS)	0.08	0.61
Voice Conversion (VC)	0.40	12.10
Overall LA eval dataset	0.17	4.93
Overall PA eval dataset	0.2	5.29

3.3 Performance comparison on different activation functions

In the proposed AEXANet architecture, we employed the MFM activation function because MFM achieves better computational efficiency while maintaining good accuracy. MFM activation is based on the Max-Out activation function. Our DNN with MFM is capable of choosing reliable features that not only capture the distinctive traits of the signal but also make the model computationally more efficient. To better investigate the effectiveness of MFM activation, we designed an experiment to test our AEXANet model with different activation functions, i.e., ReLU, leakyReLU, SiLU, and MFM separately, and the results are shown in Table 4. Our experiments attained an EER and min t-DCF of 4.93% and 0.17, 6.75% and 0.21, and 7.24% and 0.21, on MFM, LeakyRelu, SiLU, and Relu, respectively. All were evaluated using the ASVspoof 2019 LA-Eval dataset. In Table 4, it should be noted that the AEXANet model achieved the highest results with MFM activation and the second-best performance with leakyReLU. Moreover, AEXANet with ReLU attained the worst results. Dense connections make the network more compact by substantially reducing the parameters, and MFM further suppresses those compact representations and performs max-out feature filter selection to separate the noisy and informative low-activation neurons in each layer. These results justify the selection of MFM activation in our model.

Table 4. EVALUATION ON VARYING ACTIVATION FUNCTIONS

Activation Function	min t-DCF	EER (%)
MFM	0.17848	4.93319
LeakyRelu	0.21226	6.75754
SiLU	0.21299	7.24647
Relu	0.3199	9.39521

3.4 Performance comparison against existing systems

To check the effectiveness of our system against the contemporary methods, we compared our countermeasure with these spoofing detection systems [1], [12], [10], [23], [24], [25], [26], [27], [28], [29], [30], and [31], including the ASVspoof baseline

methods. Table 5 highlights the results of the proposed and comparative methods on the LA and PA sets of ASVspoof 2019.

Table 5. PERFORMANCE COMPARISON AGAINST EXISTING SYSTEMS ON THE ASVSPOOF 2019 LA AND PA DATASET

Methods	Logical Access (LA)		Physical Access (PA)	
	min t-DCF	**EER (%)**	**min t-DCF**	**EER (%)**
BaselineRawNet2	0.415	8.95	0.9999	46.03
CQCC:B01 [1]	0.237	9.57	0.2454	11.04
LFCC:GMM B02 [1]	0.212	8.09	0.3017	13.54
Chadha et al. [28]	-	9.055	-	9.951
Zeinali et al. [12]	-	8.01	-	-
Yang et al. [27]	-	-	0.2081	11.44
Lavrentyeva et al. [24]	0.1827	7.86	-	-
Das et al. [26]	0.184	7.70	-	-
Chettri et al. [25]	0.179	7.66	0.1492	6.11
Patil et al. [29]	0.1718	6.87	0.2499	11.44
Lai et al. [23]	-	6.70	-	-
Gomez-Alanis et al. [10]	-	6.28	-	-
Borzi et al. [30]	-	5.00	-	-
Gao et al. [31]	-	4.03	-	-
Proposed AEXANet	**0.17**	**4.93**	**0.2061**	**5.29**

The outcome of this experiment shows that our system outperforms many current methods including the ASVspoof baseline methods on both the LA and PA subsets. All of the compared methods also used the ASVspoof 2019 dataset and follow similar experimentation protocols to the proposed method. A few systems contributed to the ASVspoof 2019 LA and PA challenge and outline their best outcomes using the model ensembles, classifiers, and data augmentation, however, our use of a DNN with feature learning produces improved results. Systems that produce an EER below 4% on the LA evaluation set or an EER below 5% on the PA evaluation set are rare due to the increasing challenge in the detection of unforeseen spoofing attacks. The results for Baseline RawNet2 were not available for the ASVspoof 2019 dataset, therefore, we also computed the results of the baseline RawNet2 model on the ASVspoof 2019 LA and PA datasets and included those results to better conclude this comparative analysis. All of the baseline models, i.e., RawNet2, CQCC-GMM, and LFCC-GMM, attained higher min t-DCF and EER values than the other compared methods. It should be noted that "VGG + SincNet," adopted in [12] showed poorer performance due to a mismatch of attacks between the training and evaluation sets. The system based on SVM and DBN classifiers from [25] showed a substantial increase in performance due to its focus on phase and wavelet features. As a result of feature engineering and highly optimized DNN models, [10], [23], and [26] provided a promising performance on the LA and PA sets. Still, for LA and PA attacks, our system outperformed the comparative methods for audio spoofing detection.

3.5 Cross-Dataset Testing

To examine the generalization power of our AEXANet model, we conducted a two-stage cross-dataset experiment on the ASVspoof 2015 and ASVspoof 2019 datasets. For this experiment, we used only the LA collection of the ASVspoof 2019 dataset, as ASVspoof 2015 only incorporates LA spoofing attacks. In the first stage, we trained our model using the ASVspoof 2015 training and development subsets and evaluated it on the ASVspoof 2019 LA-Eval set. In the second stage, we trained our model using the ASVspoof 2019 LA training and development collections and evaluated it on the ASVspoof 2015 LA-Eval set. The results are shown in Table 6. Despite the fact that both training and evaluation sets contain different speakers, synthetic audio generation algorithms, environments, and microphones, we discovered that our countermeasure achieved excellent results on the ASVspoof 2019 LA and PA sets. However, the results in Table 6 show that our proposed countermeasure is unable to generalize effectively when entirely different databases are used. More specifically, across the ASVSpoof 2019 dataset, we can see that ASVSpoof 2015 generalizes poorly to unseen conditions, with EER and min t-DCF reaching 37% and 0.91, respectively. In comparison to the first stage experiment, ASVSpoof 2019 performed well, with EER and min t-DCF of 19% and 0.58, respectively. It should be noted that the ASVspoof 2015 LA training set uses obsolete speech synthesis algorithms that generalize poorly, resulting in softmax probability distributions that are indistinguishable for real and false speech samples. This eventually leads to poor performance on the ASVSpoof 2019 LA-Eval set. The average EER for the experiment with ASVSpoof 2019 as a trained set is much lower due to the fact that the trained set covers the advanced and diverse nature of the attacks, giving a better generalization capability that provides good results on the ASVspoof 2015 LA-Eval set. Cross-dataset evaluation demonstrates that when there is a considerable domain mismatch between the training and testing sets, the current countermeasures fail to a significant extent. Therefore, aside from having a large variation in the training dataset, domain adaptation techniques may be useful in real-time applications where attacks are unseen.

Table 6. CROSS DATASET TESTING BETWEEN ASVSPOOF2015 AND ASVSPOOF2019

Datasets		**min t-DCF**	**EER (%)**
Train Test	ASVspoof2015 ASVspoof2019	0.91	37.60
Train Test	ASVspoof2019 ASVspoof2015	0.58	19.71

4 Conclusion

In this study, we have presented a reliable voice anti-spoofing system for ASV systems. We have shown that AEXANet, as a single end-2-end DNN, yielded the best results for spoofed speech detection. The proposed method makes use of low-level acoustic features to capture important attributes in the bonafide and spoofed audio samples. The proposed method is fused with a dense connectivity pattern to provide

better parameter efficiency and strengthen the propagation of audio features. We have also introduced MFM activation in our model, which not only selects the reliable features but also makes the model computationally more efficient. Our method has shown great potential by giving the min t-DCF and EER of 0.17 and 4.93% on LA, and 0.20 and 5.29% on the PA dataset, respectively. The proposed algorithm demonstrated significant performance against the TTS and VC spoofing attacks with the min t-DCF and EER of 0.08 and 0.6% on TTS and 0.4 and 12% on VC, respectively, compared to the baseline and other contemporary methods. Our direction for future work is to improve the generalization power of the AEXANet model for cross-corpora evaluation.

Acknowledgment

This work was supported by the Punjab Higher Education Commission of Pakistan via Award No. PHEC/ARA/PIRCA/20527/21, National Science Foundation of USA via Award No. 1815724, and Michigan Translational Research and Commercialization (MTRAC) Advanced Computing Technologies (ACT) Grant Case number 292883.

References

1. Todisco, M., Wang, X., Vestman V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T. and Lee, K.A.: ASVspoof 2019: Future horizons in spoofed and fake audio detection. In: International Speech Communication Association, pp. 1904.05441 (2019)
2. Meng, Y., Wang, Z., Zhang, W., Wu, P., Zhu, H., Liang, X., and Liu, Y.: Wivo: Enhancing the security of voice control system via wireless signal in iot environment. In: Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 81-90 (2018)
3. Wang, Q., Lin, X., Zhou, M., Chen, Y., Wang, C., Li, Q., and Luo, X.: Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications, pp. 2062-2070 (2019)
4. Muckenhirn, H., Korshunov, P., Magimai-Doss, M., Marcel, S.: Long-term spectral statistics for voice presentation attack detection. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25(11), pp. 2098-2111 (2017)
5. Alegre, F., Amehraye, A., and Evans, N.: Spoofing countermeasures to protect automatic speaker verification from voice conversion. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3068-3072 (2013)
6. Alam, M. J., Kenny, P., Bhattacharya, G., and Stafylakis, T.: Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015. In: Sixteenth annual conference of the international speech communication association (2015)
7. F., Alegre, Vipperla, R., Amehraye, A., and Evans, N.: A new speaker verification spoofing countermeasure based on local binary patterns. In: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon: France (2013), p. 5p (2013)

8. A., Javed, Malik, K. M., Irtaza, A., and Malik, H.: Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. In: Applied Acoustics, vol. 183, p. 108283 (2021)
9. Wang, L., Yoshida, Y., Kawakami, Y., and Nakagawa, S.: Relative phase information for detecting human speech and spoofed speech. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
10. A., Gomez-Alanis, Peinado, A. M., Gonzalez, J. A., and Gomez, A. M.: A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In: Proc. Interspeech, vol. 2019, pp. 1068-1072 (2019)
11. Arif, T., Javed, A., Alhameed, M., Jeribi, F., and Tahir, A.: Voice Spoofing Countermeasure for Logical Access Attacks Detection. In: IEEE Access, vol. 9, pp. 162857-162868 (2021)
12. Zeinali, H., Stafylakis, T., Athanasopoulou, G., Rohdin, J., Gkinis, I., Burget, L., and Černocký, J.: Detecting spoofing attacks using vgg and odenet: but-omilia submission to asvspoof 2019 challenge. In: Proc. INTERSPEECH (2019)
13. Nautsch, A., Wang, X., Evans, N., Kinnunen, T.H., Vestman, V., Todisco, M., Delgado, H., Sahidullah, M., Yamagishi, J., and Lee, K.A.: ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. In: IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3 (2), pp. 252-265 (2021)
14. Jung, J.W., Heo, H.S., Yang, I.H., Shim, H.J. and Yu, H. J.: Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification. In: Proc. of Interspeech, vol. 8(12), pp. 23-24 (2018)
15. Muckenhirn, H., Doss, M. M., and Marcell, S.: Towards directly modeling raw speech signal for speaker verification using CNNs. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4884-4888 (2018)
16. Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., and Larcher, A.: End-to-end anti-spoofing with RawNet2. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6369-6373 (2021)
17. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708 (2017)
18. Cai, M., Shi, Y., and Liu, J.: Deep maxout neural networks for speech recognition. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 291-296 (2017)
19. Jung, J.W., Kim, S.b., Shim, H.J., Kim, J.H., and Yu, H.J.: Improved rawnet with filter-wise rescaling for text-independent speaker verification using raw waveforms. In: Proc. Interspeech, pp. 1496-1500 (2020)
20. Kinnunen, T., Lee, K.A., Delgado, H., Evans, N., Todisco, M., Sahidullah, M., Yamagishi, J., and Reynolds, D.A.: t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. In: Proc. The Speaker and Language Recognition Workshop, pp. 312-319 (2018)
21. Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., Kinnunen, T., Lee, K.A., Vestman, V., and Nautsch, A.: Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database. In: University of Edinburgh. The Centre for Speech Technology Research (2019)
22. Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K.A., and Juvela, L.: ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. In: Computer Speech & Language (CSL), vol. 64, p. 101114 (2020)

23. Lai, C.I., Chen, N., Villalba, J., and Dehak, N. J.: ASSERT: Anti-spoofing with squeeze-excitation and residual networks. In: Proc. Interspeech, pp. 1013-1017 (2019)
24. Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., and Kozlov, A.: STC antispoofing systems for the ASVspoof2019 challenge. In: Proc. Interspeech, pp. 1033-1037 (2019)
25. Chettri, B., Stoller, D., Morfi, V., Ramírez, M.A., Benetos, E. and Sturm, B.L.: Ensemble models for spoofing detection in automatic speaker verification. In: Proc. Interspeech, pp. 1018-1022 (2019)
26. Das, R. K., Yang, J., and Li, H.: Long range acoustic and deep features perspective on ASVspoof 2019. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1018-1025 (2019)
27. Yang, J., Xu, L., Ren, B., Ji, Y.: Discriminative features based on modified log magnitude spectrum for playback speech detection. In: EURASIP Journal on Audio, Speech, and Music Processing vol. 2020, no. 1, pp. 1-14 (2020)
28. A., Chadha, Abdullah, A., Angeline, L. A.: A Comparative Performance of Optimizers and Tuning of Neural Networks for Spoof Detection Framework. In: International Journal of Advanced Computer Science and Applications, vol. 13, no. 4 (2022)
29. Patil, T., Acharya, R., Patil, H. A., Guido, R. C.: Improving the potential of Enhanced Teager Energy Cepstral Coefficients (ETECC) for replay attack detection. In: Computer Speech & Language vol. 72, p. 101281 (2022)
30. Borzì, S., Giudice, O., Stanco, F., and Allegra, D.: Is Synthetic Voice Detection Research Going Into the Right Direction?. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 71-80 (2022)
31. Gao, Y., Vuong, T., Elyasi, M., Bharaj, G., and Singh, R.: Generalized spoofing detection inspired from audio generation artifacts. In: Proc. Interspeech (2021)