CoNSIST : Consist of New methodologies on AASIST, leveraging Squeeze-and-Excitation, Positional Encoding, and Re-formulated HS-GAL

Jae-Hoon Ha[1], Joo-Won Mun[1], Sang-Yup Lee[2]

[1]Dept. of Digital Analytics, Yonsei University
[2]Dept. of Communication, Yonsei University

Jae Hoon Ha and Joo Won Mun contributed equally to this work

Abstract

With the recent advancements in artificial intelligence (AI), the performance of deep learning-based audio deepfake technology has significantly improved. This technology has been exploited for criminal activities, leading to various cases of victimization. To prevent such illicit outcomes, this paper proposes a deep learning-based audio deepfake detection model. In this study, we propose CoNSIST, an improved audio deepfake detection model, which incorporates three additional components into the graph-based end-to-end model AASIST: (i) Squeeze and Excitation, (ii) Positional Encoding, and (iii) Reformulated HS-GAL, This incorporation is expected to enable more effective feature extraction, elimination of unnecessary operations, and consideration of more diverse information, thereby improving the performance of the original AASIST. The results of multiple experiments indicate that CoNSIST has enhanced the performance of audio deepfake detection compared to existing models.

1. Introduction

Audio deepfake technology utilizes artificial intelligence (AI) and machine learning (ML) methods to generate or synthesize human-like audio clip that imitate the voice of a particular person[1]. With the recent advancements in AI, deep learning-based audio deepfake technology has seen significant improvements in its performance, contributing to the betterment of people's lives through various applications such as audiobooks and assistive tools for the hearing impaired[1]. However, audio deepfake technology is also being exploited for criminal purposes, leading to various cases of victimization. In 2023, the American IT research firm Gartner predicted that 20% of financial fraud crimes would involve the use of such deepfake technology[2].

It is becoming clear that research on audio deepfake detection technologies is necessary to prevent such illegal results. Recently, various detection algorithms based on machine learning and deep learning have been actively proposed. However, the existing research results have a common limitation that the generalization performance of proposed methos for the unseen data is low[1]. In this regard, this study proposes a new model that improves audio deepfake detection performance and enhances generalization performance using deep learning technology, thereby contributing to the prevention of damage caused by audio deepfakes.

Deep learning-based audio deepfake detection techniques can be broadly classified into three categories: CNN-based, Transformer-based, and GNN-based[1]. Among these, GNN-based AASIST[3] is an end-to-end algorithm that effectively detects audio deepfakes by extracting spectral and temporal features from the raw waveform of the audio and performing graph attention operations between the two types[1]. AASIST has demonstrated good performance in the ASVspoof 2019[1] audio deepfake detection competition and the ADD challenge 2023[1], proving the effectiveness of graph attention-based algorithms in audio deepfake detection. Given the effectiveness of AASIST, we propose a new model named CoNSIST. This model enhances the audio deepfake detection performance by incorporating three methodologies into the AASIST, which serves as the baseline model in this study.

2. Related Works

There are two main types of algorithms used for audio deepfake detection: pipeline detectors and end-to-end algorithms[1]. Pipeline-based detection models divide the data processing process into multiple stages to perform feature extraction and then classification. Pipeline models typically use machine learning-based algorithms such as Linear and Quadratic Discriminant[4], Linear SVM[5], KNN[5], and Random Forest[6] for classification. Currently, the Quadratic Support Vector Machine(Q-SVM) proposed by Kumar-Singh and Singh[7] is the best performing machine learning model. Pipeline-based approaches have the

advantage of allowing flexible design of the audio data feature extraction process. However, they can be time-consuming to apply to new unseen datasets.

End-to-end based detection models learn the entire process of data processing, feature extraction, and classification in a single integrated deep neural network(DNN). This allows the model to automatically learn the features of the audio data and quickly adapt to unseen datasets, however, require significant computational resources. End-to-end models can be broadly classified into these three categories: CNN[8], Transformer[9], and GNN-based[3][10]. Among these, GNN-based models have shown promising performance in audio deepfake detection due to their capability to model complex relationships between graphs[11]. GNN-based algorithms include RawGAT-ST[10], which extracts spectral and temporal features from the raw waveform and applies graph attention, and AASIST[3], which was proposed by Jung et al. in 2022 as an extension of RawGAT-ST. AASIST models the relationship between spectral and temporal features through graph operations (HS-GAL layer) between two heterogeneous graphs, effectively capturing information about the presence of deepfakes and achieving good performance. In this study, we used AASIST as the baseline model and added three components to further improve its performance.

3. Model Architecture

AASIST extracts a 3D feature map from the raw waveform of the audio using an encoder block consisting of six residual blocks, and effectively utilizes the characteristics of both spectral and temporal graphs through a GAN[12] structure. In particular, the performance of the model is improved by using an HS-GAL structure to enhance the attention between g_s and g_t nodes and two stack nodes.

3.1 SE Encoder

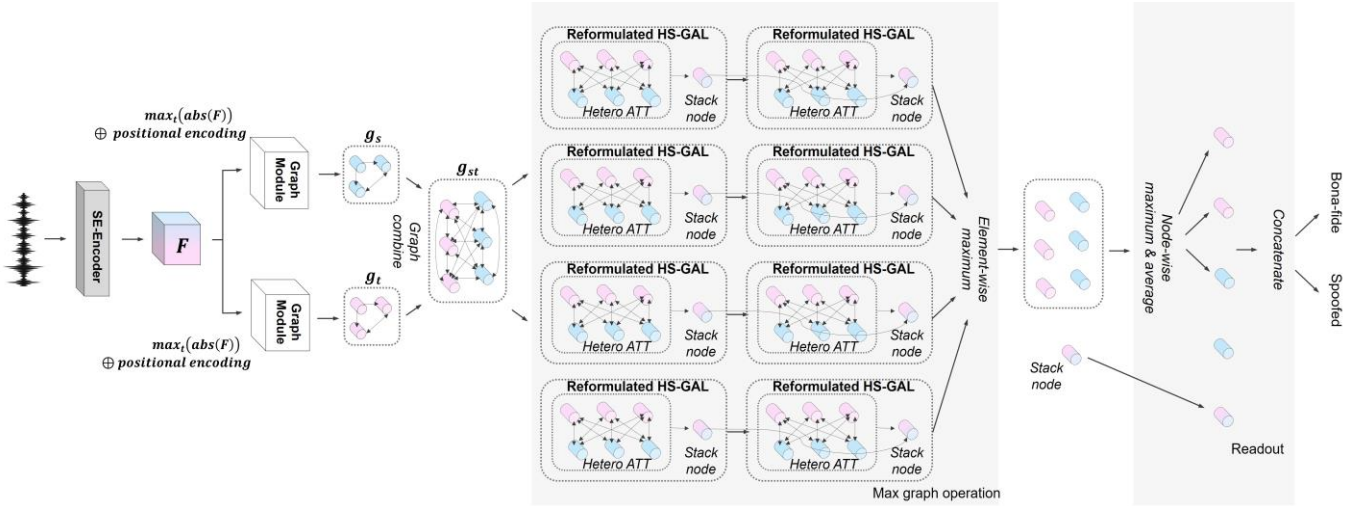
First, a Squeeze-and-Excitation(SE) block from Rawformer[9] is added to the encoder block to extract new feature maps. In the encoder stage that transforms the raw waveform into a 3D feature map, the squeeze-and-excitation operation[13], which is mainly used in CNN architectures, is added after each residual convolution operation of the six residual blocks of AASIST. The SE encoder is divided into three main stages. First, in the squeeze stage, the features of each channel are compressed into a single number through global average pooling. In the excitation stage, 1x1 convolution is used to learn weights for each channel to represent the importance of each channel. Finally, in the scale stage, the weights for each channel are used to readjust the existing feature map. By additionally using the SE operation on the encoder consisting of six residual blocks of AASIST, weights are applied to each element of the feature map according to the importance of each channel, and the important features are learned and given a greater weight. By adding the SE method, the model can learn on its own which features to focus on more, and thus improve the performance of the model compared to the encoder of the original AASIST, which is only composed of residual blocks.

3.2 Positional Encoding

Positional information was element-wisely added during the process of extracting and converting the 3D feature map into a graph. Before graph module, a vector containing positional information for each spectral and temporal axis of the feature map was added element-wisely. By adding positional information on both graphs, the model's generalization performance and range of applicability were improved, and the model was set to better recognize and learn the order and changes in each graph by making it easier for the model to recognize the information of both graphs[14]. The positional encoding method was chosen instead of the positional embedding method, which requires self-learning of vectors containing positional information, in anticipation that the learning process would become more complex, and the learning time would increase. The formula for each added positional information vector is the same as the formula for positional encoding in Transformer[15], using sin and cos functions.

3.3 Reformulated HS-GAL

AASIST's HS-GAL uses two stack nodes to stack two layers and performs three attention operations: self-attention for spectral and temporal graphs, and attention between spectral and temporal graphs. However, since the graph module already performs self-attention to learn the connection strength between nodes, and the pooling layer discards unimportant node information, we hypothesized that performing self-attention redundantly in HS-GAL would not have a significant impact on the final classification. Therefore, in this study, we removed the self-attention operation for spectral and temporal graphs in the HS-GAL stage and only used attention between the two graphs to reduce the number of unnecessary parameters. We also increased the number of stack nodes from two to four to allow the model to consider more diverse information, thereby improving the performance of the model.



(Picture 1) CoNSIST Architecture.

4. Experiments and results

4.1 Datasets and metrics

This study employed the LA (Logical Access) dataset obtained from the ASVspoof 2019 challenge. AASIST also utilized the LA dataset. By utilizing the identical dataset, our aim is to objectively compare the outcomes of the proposed model with those of AASIST. The training set contains 2,580 bonafide and 22,800 spoofed audio data generated using 4 TTS(text to speech) and 2 VC(voice conversion) speech synthesis techniques. The development set contains 2,548 bonafide and 22,296 spoofed audio data. The evaluation set contains 7,355 bonafide and 63,882 spoofed audio data generated using 7 TTS and 6 VC[16].

To evaluate the performance of the model, the minimum tandem detection cost function(min t-DCF) and equal error rate(EER) were used, which are the same evaluation metrics as the ASVspoof 2019 challenge. A lower value of both evaluation metrics indicates higher performance of the model. Min t-DCF focuses on the classification of spoofed audio data[17], while EER evaluates the balance of performance between real and spoofed audio data.

4.2 Experiments settings

Experiments were conducted using all possible combinations of the three methodologies proposed in CoNSIST as shown in Table 1. To guarantee a fair comparison with AASIST, all experiments were carried out using same hyperparameters and conditions as those utilized by AASIST. Each single application model (SE, POS, RE HS-GAL) was experimented with once, and all other combinations were experimented with three times. The average and best values of min-tDCF and EER were measured for comparison. Since the performance of the model can differ depending on the random seed, each experiment used a different random seed to calculate min-tDCF and EER[18].

<Table 1> Description of each model

	Model explanation
Only SE	AASIST + Squeeze and Excitation
Only Pos	AASIST + Positional Encoding
Only Re HS-GAL	AASIST + Reformulated HS-GAL
Con_v1	AASIST + SE + Pos
Con_v2	AASIST + Pos + Re HS-GAL
Con_v3	AASIST + SE + Re HS-GAL
CoNSIST*(ours)*	AASIST + SE + Pos + Re HS-GAL

4.3 Results

<Table 2> Results of experiments on each architecture: avg(best)

System	**Min t-DCF**	**EER(%)**
AASIST*(baseline)*	0.0393 (0.0382)	1.37 (1.31)
Only SE	0.0447	1.36
Only POS	0.0345	1.22
Only Re HS-GAL	0.0429	1.34
Con_v1	0.0373 (0.0344)	1.2646 (1.18)
Con_v2	0.0339 (0.0317)	1.19 (1.00)
Con_v3	0.0489 (0.0422)	1.45 (1.32)
CoNSIST*(ours)*	0.0288 (0.0267)	0.94 (0.89)

Table 2 shows that results of the experiments. Based on min t-DCF, all models except SE, Re HS-GAL single model, and Con_v3 outperformed AASIST. CoNSIST had the lowest min t-DCF with an average of 0.0288 and the best of 0.0267, indicating that when all three methodologies are applied, it improves AASIST's performance the most. In terms of EER, all models except Con_v3 surpassed AASIST's performance. CoNSIST also performed better than AASIST based on EER, indicating that in the same environment, all the three methodologies of CoNSIST improved AASIST's performance.

<Table 3> Comparison of CoNSIST and AASIST

system	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	Min t-DCF	EER(%)
AASIST	0.80	0.44	**0.00**	1.06	0.31	0.91	**0.1**	**0.14**	**0.65**	0.72	1.52	3.40	0.62	0.0374(0.0275)	1.13(**0.83**)
CoNSIST	**0.66**	**0.27**	0.01	**0.98**	**0.23**	**0.71**	0.14	0.20	0.71	**0.46**	**1.48**	**2.28**	**0.48**	**0.0288(0.0267)**	**0.97**(0.89)

Even when compared to the final results of AASIST claimed by authors, CoNSIST demonstrates an improvement in performance by employing the three methodologies. A07 to A19 represent different speech synthesis techniques, and the authors of AASIST conducted three experiments each to measure their respective average and best values[3]. CoNSIST outperforms AASIST in detecting deepfake voices for all speech synthesis techniques except A09, A13, A14 and A15. Based on min t-DCF, CoNSIST surpasses AASIST in both average and best values, while in terms of EER, there is a significant difference in the average values.

5. Conclusion

In this study, we proposed CoNSIST, an improved audio deepfake detection model based on the GNN-based end-to-end model AASIST. CoNSIST incorporates three model improvement methodologies into AASIST: (i) Squeeze and Excitation, (ii) Positional Encoding, and (iii) Reformulated HS-GAL. These methodologies enable more effective feature extraction, elimination of unnecessary operations, and consideration of more diverse information, thereby improving the performance of the original AASIST. The results of the experiments demonstrate that CoNSIST outperforms AASIST under the same experimental conditions. CoNSIST also exhibits more stable performance across different voice synthesis systems. We expect that further research on hyperparameter tuning, dataset collection and augmentation, and other aspects can further improve the generalization performance of CoNSIST, leading to enhanced accuracy in audio deepfake detection.

References

[1] Yi, Jiangyan, et al. “Audio Deepfake Detection: A Survey.” *ArXiv (Cornell University)*, 28 Aug. 2023,

[2] Jung, Kyunghoon, and Changhyun Kim. “Beware of Voice Cloning: Deep Voice Crime Steals 400 Billion Won.” *Moneytoday*, 11 Feb. 2023, news.mt.co.kr/mtview.php?no=2023020913433930492.

[3] Jung, Jee-weon, et al. “AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks.” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23 May 2022, https://doi.org/10.1109/icassp43922.2022.9747766.

[4] Hamza, Ameer, et al. “Deepfake Audio Detection via MFCC Features Using Machine Learning.” *IEEE Access*, vol. 10, 2022, pp. 134018–134028, https://doi.org/10.1109.

[5] Lataifeh, Mohammed, and Ashraf Elnagar. “Ar-DAD: Arabic Diversified Audio Dataset.” *Data in Brief*, Nov. 2020, p. 106503, https://doi.org/10.1016/j.dib.2020.106503.

[6] Borrelli, Clara, et al. “Synthetic Speech Detection through Short-Term and Long-Term Prediction Traces.” *EURASIP Journal on Information Security*, vol. 2021, no. 1, 6 Apr. 2021, https://doi.org/10.1186/s13635-021-00116-3.

[7] Arun Kumar Singh, and Priyanka Singh. “Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics.” *ArXiv (Cornell University)*, 3 Sept. 2020. Accessed 12 Apr. 2024.

[8] Chintha, Akash, et al. “Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection.” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, Aug. 2020, pp. 1024–1037, https://doi.org/10.1109/jstsp.2020.2999185.

[9] Liu, Xiaohui, et al. Leveraging Positional-Related Local-Global Dependency for Synthetic Speech Detection. 4 June 2023, https://doi.org/10.1109/icassp49357.2023.10096278.

[10] Tak, Hemlata, et al. “End-To-End Spectro-Temporal Graph Attention Networks for Speaker Verification Anti-Spoofing and Speech Deepfake Detection.” ArXiv (Cornell University), 1 Jan. 2021, https://doi.org/10.48550/arxiv.2107.12710.

[11] Tak, Hemlata, et al. “Graph Attention Networks for Anti-Spoofing.” ArXiv (Cornell University), 30 Aug. 2021, https://doi.org/10.21437/interspeech.2021-993. Accessed 3 Apr. 2024.

[12] Veličković, Petar, et al. “Graph Attention Networks.” arXiv (Cornell University), Feb. 2018, doi:10.17863/cam.48429.

[13] Hu, Jie, et al. “Squeeze-and-Excitation Networks.” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, doi:10.1109/cvpr.2018.00745.

[14] Dufter, Philipp, et al. “Position Information in Transformers: An Overview.” *Computational Linguistics*, vol. 48, no. 3, 2022, pp. 733–763, https://doi.org/10.1162/coli_a_00445. Accessed 4 Dec. 2022.

[15] Vaswani, Ashish, et al. “Attention is All you Need.” arXiv (Cornell University), vol. 30, June 2017, pp. 5998–6008, arxiv.org/pdf/1706.03762v5.

[16] Wang, Xin, et al. “ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech.” ArXiv (Cornell University), 4 Nov. 2019, https://doi.org/10.48550/arxiv.1911.01601.

[17] Kinnunen, Tomi, et al. “T-DCF: A Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification.” *Odyssey 2018 the Speaker and Language Recognition Workshop*, 26 June 2018, www.isca-speech.org/archive/Odyssey_2018/pdfs/68.pdf, https://doi.org/10.21437/odyssey.2018-44.

[18] Wang, Xin, and Junichi Yamagishi. “A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection.” *ArXiv (Cornell University)*, 30 Aug. 2021, https://doi.org/10.21437/interspeech.2021-702.