The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.10, pp.513~519

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.10.513

CoNSIST: Consist of New Methodologies on AASIST for Audio Deepfake Detection

Jae Hoon Ha[†] · Joo Won Mun[†] · Sang Yup Lee[††]

ABSTRACT

Advancements in artificial intelligence(AI) have significantly improved deep learning-based audio deepfake technology, which has been exploited for criminal activities. To detect audio deepfake, we propose CoNSIST, an advanced audio deepfake detection model. CoNSIST builds on AASIST, which a graph-based end-to-end model, by integrating three key components: Squeeze and Excitation, Positional Encoding, and Reformulated HS-GAL. These additions aim to enhance feature extraction, eliminate unnecessary operations, and incorporate diverse information. Our experimental results demonstrate that CoNSIST significantly outperforms existing models in detecting audio deepfakes, offering a more robust solution to combat the misuse of this technology.

Keywords : AASIST, ASVspoof, Audio Deepfake, Graph Attention Network

컨시스트: 오디오 딥페이크 탐지를 위한 그래프 어텐션 기반 새로운 모델링 방법론 연구

하재훈[†] · 문주원[†] · 이상엽[††]

요약

인공지능 기술의 발전과 함께 딥러닝 기반의 오디오 딥페이크 기술이 크게 향상되었고, 이를 악용하여 다양한 범죄 활동이 이루어지고 있다. 오디오 딥페이크를 탐지하여 이러한 피해를 예방하기 위해 본 논문은 새로운 컨시스트(CoNSIST) 모델을 제안한다. 이 모델은 그래프 기반의 모델인 AASIST를 기반으로, 세 가지 추가적인 모델링 방법론을 적용하여 오디오 딥페이크 탐지를 한다. 세 가지 추가적인 모델링을 통해 특징 추출을 강화하고, 불필요한 작업을 제거하며, 다양한 정보를 통합하는 것을 목표로 한다. 최종 실험 결과, 컨시스트가 기존 오디오 딥페이크 탐지 모델들보다 더 우수한 성능을 보여 딥페이크의 악용을 방지하기 위해 더 나은 해결책을 제공한다.

키워드 : 어시스트, 에이에스비 스푸프, 오디오 딥페이크, 그래프 어텐션 네트워크

1. Introduction

Recent advances in artificial intelligence (AI) have stimulated the development of audio manipulation techniques that can create realistic and imitated speech[1]. This technology, often referred to as “audio deepfakes,” or “deepvoice,” utilizes machine learning (ML) or deep learning (DL) algorithms to analyze a target speaker's voice and synthesize new audio files that mimic their speech patterns. This technology can be used in beneficial ways such as generating audiobooks and creating assistive tools for individuals with hearing impaired[1]. However, the deepvoice technique also presents significant risks when it is used for criminal purposes. A 2023 prediction made by the American IT research firm Gartner suggests that one-fifth of financial fraud cases could involve this technology in the near future[2].

The increasing prevalence of audio deepfakes provokes the urgency of developing robust detection technologies. To prevent criminal danger using audio deepfake, researchers have actively proposed various detection algo-

※ 본 연구는 2024년도 연세대학교 인공지능 대학원 AI창의자율연구프로그램 지원비를 받아 수행된 연구임.
※ 이 논문은 2024년 ASK 2024의 우수논문으로 "CoNSIST: Consist of New methodologies on AASIST, leveraging Squeeze-and-Excitation, Positional Encoding, and Re-formulated HS-GAL"의 제목으로 발표된 논문을 확장한 것임.

† 준 회 원 : 연세대학교 디지털애널리틱스 융합협동과정 석사과정
†† 정 회 원 : 연세대학교 언론홍보영상학부 부교수

Manuscript Received : July 9, 2024
First Revision : August 29, 2024
Accepted : September 12, 2024

* Corresponding Author : Sang Yup Lee(sangyuplee@yonsei.ac.kr)

rithms using machine learning techniques. However, existing methods often struggle with generalizability, achieving subpar performance when confronted with unseen data[1]. The present study aims to close the gap in the literature by proposing a deep learning model specifically designed to enhance audio deepfake detection accuracy and improve its ability to adapt to unseen data. By proposing new modeling methodologies on existing detection algorithms, we believe that the proposed model mitigate the potential harms associated with audio deepfakes.

One approach to detect audio deepfake using a deep learning algorithm is AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks), which works by analyzing both the frequency and time information in an audio raw-waveform and the relationships between them[1, 3]. AASIST's strong performance in competitions like ASVspoof 2019 and ADD (Audio Deep synthesis Detection) challenge 2023 demonstrates the value of GNNs for audio deepfake detection[1]. Building on AASIST's success, we propose CoNSIST, a new GNN based model that incorporates three additional methodologies, Squeeze-and-Excitation, Posional Encoding, and Re formulated HS-GAL(heterogeneous stacking graph attention layer), to further enhance deepfake detection accuracy.

2. Related Works

2.1 Research on the Detection of Audio Deepfakes

There are two main strategies for detecting audio deepfakes: pipeline and end-to-end algorithms[1]. Pipeline approaches break down the detection process into two steps. First, specific characteristics or features are identified in the audio. Then, these features are used to classify the audio as real or fake. Machine learning algorithms used in pipeline detectors include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), support vector machines (SVM), K-nearest neighbors (KNN), and random forests[4-6]. Currently, the Q-SVM method by Kumar-Singh and Singh is the most successful machine learning model for audio deepfake detection[7]. Pipeline approaches offer flexibility in designing how features are extracted, but they can be slow to handle new types of audio data.

Unlike separate pipeline approaches, end-to-end detection models learn the entire process all at once using a deep neural network[1]. This allows to automatically learn important features from the audio and adapt to unseen data, however, requires a large amount of computing resources. There are three main algorithms of end-to-end models: CNN[8], Transformers[9], and GNN-based[3, 10]. CNN-based models are composed of a single convolutional neural network and multi-layer perceptron layer. While these models have demonstrated strong capabilities in identifying voice conversion (VC) and text-to-speech (TTS) manipulations, they suffer from limited ability to accurately detect other forms of audio deepfakes due to weak generalization performance. ResNet-based models are an improvement over traditional CNN-based models that utilize residual mapping. They have demonstrated a higher level of trainability and have exhibited promising levels of performance. Transformer-based models are composed of convolution layers and transformer's encoder blocks. They capture both local and global features in the audio data, resulting in superior generalization performance.

Among the different model types, GNN-based models stand out for their ability to detect audio deepfakes[11]. GNNs excel at analyzing the complex relationships between data points, which are represented as graphs. Two well-known examples of GNN-based models used for audio deepfake detection are RawGAT-ST[10] and AASIST [3]. RawGAT-ST extracts both sound frequency and timing information from the audio raw-waveform and analyzes the connections between them using the attention mechanism[10]. AASIST, a more recent model based on RawGAT-ST, goes a step further. It uses special graph operations to model the relationships between these features within two separate graphs. This allows AASIST to capture information that indicates the presence of audio deepfakes, resulting in high detection accuracy. For this study, we chose AASIST as the baseline model and built upon it by incorporating three additional components to enhance its performance.

2.2 Architecture of AASIST

CoNSIST, the suggested model in this study, takes AASIST as a basis model and adds three more components to it to enhance deepfaked audio detection performance. We chose AASIST, an end-to-end model, as our baseline because end-to-end models generally exhibit better generalization performance compared to pipeline-based approaches[1]. Among end-to-end models, AASIST has demonstrated exceptional performance in various audio deepfake competitions, performing best in the ASVspoof 2019

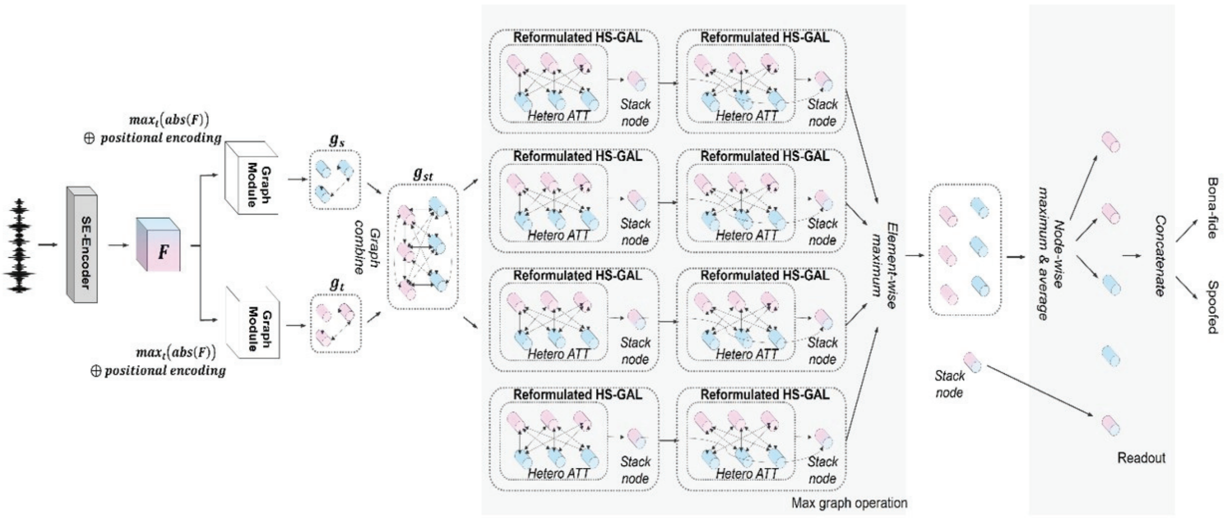

Fig. 1. CoNSIST Architecture

LA dataset and ADD challenge 2023. Therefore, we chose AASIST as our baseline and conducted further research to improve its performance in audio deepfake detection. In this section, we first explain the overall architecture of AASIST and its operation prior to describing the primary elements of CoNSIST.

AASIST uses a RawNet2-based encoder to convert the raw-waveform of an audio into a 3-dimensional feature map. RawNet2 was first proposed by J. Jung, S. Kim, and H. Shim[12] for speaker verification using raw waveforms. H. Tak, J. Jung, and J. Patino[13] and H. Tak, J. Patino, and M. Todisco[11] modified RawNet2 for audio deepfake detection, and AASIST[3] uses an encoder based on this modified RawNet2. RawNet2 consists of a single sinc convolution layer and six residual blocks. The input raw waveform is passed through residual blocks to be converted into a 3D feature map of size (C, S, T): C represents the number of channels, S represents the number of spectral bins, and T represents the number of temporal bins. Finally, the encoder outputs a feature map of size (64, 23, 29).

The 3D feature map is divided into spectral and temporal components and converted into two 2D tensors, each with dimensions (C, S) and (C, T). These tensors are then converted into graph structures using a graph module based on graph attention layers and graph pooling layers[3]. In this process, all nodes in each graph are connected by edges, and the strength of these edges is learned through attention in the graph module. The graph pooling layer then retains important node information and discards unnecessary nodes. Finally, the spectral graph is transformed from (64, 23) to (64, 11) and the temporal graph is transformed from (64, 29) to (64, 20).

After passing through the encoder and graph modules and new spectral and temporal graphs are generated, the nodes of each graph are combined to create a new graph denoted as g_{st}. In this process, all nodes are connected by edges, and this combined graph g_{st} is called a heterogeneous graph[14]. According to the authors, this integration of two different graphs allows for the modeling of two types of graph representations simultaneously. Then, attention between all homogeneous and heterogeneous nodes are performed. During this process, a stack node that accumulates information from both graphs is created. This process of performing three attention operations and creating a stack node in a heterogeneous graph is called HS-GAL (Heterogeneous Stacking-Graph Attention Layer). The HS-GAL layer is passed through twice and the stack node created in the first HS-GAL layer initializes the stack node in the second HS-GAL layer. Additionally, as two HS-GAL layers performed, two nodes from each g_t and g_s are extracted through element-wise and node-wise maximum operations, and one stack node is also extracted, and these are concatenated to perform the final classification task.

3. Architecture of CoNSIST

AASIST[3] 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[15] structure. AASIST uses an HS-GAL structure to enhance the attention between g_s and g_t nodes and two stack nodes. However, our model CoNSIST demonstrated enhanced audio deepfake detection performance than the AASIST. In this section, we will describe the three main modeling methodologies of our CoNSIST.

3.1 SE Encoder

To improve the model’s ability to extract discriminative features from raw audio waveforms, we incorporated Squeeze-and-Excitation[16] from Rawformer[9] into the encoder stage of the AASIST model[3]. The AASIST encoder utilizes a series of residual convolution operations within six residual blocks to transform the raw waveform into a 3D feature map. We strategically inserted the SE block after each residual convolution within these blocks

The SE block operates with three stages: squeeze, excitation, and scale[16]. In the squeeze stage, global average pooling is applied to each channel of the feature map, compressing the feature information into a single scalar value. This value represents the overall importance of channel. The excitation stage then employs a 1x1 convolution layer to generate weights for each channel. These weights essentially capture the relative importance of each channel based on the information obtained in the squeeze stage. Finally, in the scale stage, these learned weights are used to rescale the original feature map element-wisely, emphasizing the channels containing more important features for deepfake detection.

By incorporating the SE block into the baseline encoder, the model gains the capability to dynamically learn the importance of different feature channels. This allows the model to focus on the most discriminative features for audio deepfake classification, potentially leading to improved performance compared to the baseline encoder that solely relies on residual convolutions.

3.2 Positional Encoding

To enhance the model's ability to capture the inherent order within the extracted features, we incorporated positional information during the conversion of the 3D feature map into separate spectral and temporal graphs. By adding a vector containing positional information for each axis (spectral and temporal) of the feature map in an element-wise manner, both graphs are hypothesized to improve the model's generalization performance and broaden its applicability. By explicitly providing the model with knowledge about the order and changes within each graph, we facilitate its ability to effectively learn and utilize the information contained within both spectral and temporal representations.

Positional encoding, leveraging sine and cosine functions as in the Transformer architecture[18], was chosen over positional embedding. This decision aimed to avoid the potential for increased learning complexity and extended training times associated with self-learning positional information vectors through embedding techniques[17].

3.3 Reformulated HS-GAL

The HS-GAL layer within the AASIST model employs two stack nodes and performs three distinct attention operations: self-attention for both the spectral and temporal graphs, and attention between these two graphs[3]. However, we posit that the self-attention operations for the individual spectral and temporal graphs may be redundant. Graph module already performs self-attention to learn the strength of connections between nodes within each graph and the subsequent pooling layer discards information from less important nodes. Consequently, performing self-attention again within the HS-GAL layer might not significantly impact the final classification performance.

To adjust this potential redundancy and enhance model efficiency, we propose a modified HS-GAL layer in our approach. This modified layer removes the self-attention operations for the spectral and temporal graphs, focusing solely on the attention between the two graphs. This reduces the number of unnecessary parameters within the model. Additionally, we increase the number of stack nodes from two to four in the HS-GAL layer. This allows the model to consider a more diverse range of information from the feature graphs, potentially leading to improved performance compared to the baseline AASIST architecture.

4. Experiments and results

4.1 Datasets and Metrics

To ensure a fair comparison between the CoNSIST and the baseline model, we employed the identical LA (Logical Access) dataset utilized in the ASVspoof 2019 challenge. The training set contains 2,580 bonafide and 22,800 spoofed audio samples generated using a combination of

Table 1. 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)

4 TTS and 2 VC techniques. The development set includes 2,548 bonafide and 22,296 spoofed samples. Finally, the evaluation set comprises 7,355 bonafide and 63,882 spoofed samples generated using a wider variety of 7 TTS and 6 VC methods[19].

For evaluation, we adopted the same metrics used in the ASVspoof 2019 challenge: minimum tandem detection cost function (min t-DCF) and equal error rate (EER). A lower value for both metrics indicates a better performance. Min t-DCF prioritizes the accurate classification of spoofed audio[20], while EER assesses the model's ability to achieve a balanced performance in identifying both bonafide and spoofed audio samples.

4.2 Experiments Settings

To comprehensively evaluate the effectiveness of the three proposed methodologies incorporated within CoNSIST (SE block, positional encoding, and Re-formulated HS-GAL), we conducted experiments using all possible combinations of these methodologies. A detailed breakdown of these combinations is provided in Table 5.

To ensure a fair and objective comparison with the baseline, we replicated the experimental conditions employed in the original study. This included utilizing the identical hyperparameters as outlined in Tables 2 and 3. Each of the individual methodologies was evaluated in a single experiment and all other combinations were evaluated three times to calculate both average and best values. Since the performance can differ depending on the random seed, each experiment used a different random seed to calculate min t-DCF and EER[21].

4.3 Results

Under the experimental conditions specified in Table 4 and as the results shown in the Table 6, the baseline AASIST model achieved an average min t-DCF of 0.0393 and an average EER of 1.37. The best min t-DCF and best EER obtained for AASIST were 0.0382 and 1.31. When experimenting with each of the three additional components, the model with positional encoding showed improved performance over the baseline AASIST. When testing models that combined two of the three components, the combinations of SE-encoder with positional encoding and positional encoding with Re HS-GAL both demonstrated better performance than the AASIST.

Our model CoNSIST, which includes all three new components, had about 230k parameters and achieved an average min t-DCF of 0.0288 and an average EER of 0.94, surpassing the performance of the original AASIST. The best min t-DCF and best EER also showed significant improvement at 0.0267 and 0.89.

Even when we compare to the results reported by the authors of AASIST, our model outstands the outcome of audio deepfake detection. As shown in Table 1, A07 to A19 represent various speech synthesis techniques, with the AASIST authors conducting three experiments each to determine their respective average and best values[3]. CoNSIST outperforms AASIST in detecting audio deepfakes across all speech synthesis techniques, except for A09, A13, A14, and A15. As shown in the table 6 and 7, CoNSIST surpasses AASIST in both average and best values for min t-DCF, and average EER values, and outperforms in audio deepfake detection compared to other algorithms.

Table 2. Model Configuration

Model config	
batch size	24
num_epochs	100
number of samp	64,600
first_convolution	128
filters	[70, [1, 32], [32, 32], [32, 64], [64, 64]]
pool_ratios	[0.5, 0.7, 0.5, 0.5]

Table 3. Optimizer Configuration

Optimizer config	
optimizer	Adam
baseline lr	0.0001
minimum lr	0.000005
betas	[0.9, 0.999]
weight_decay	0.0001
scheduler	cosine

Table 4. Experiment Environment

Experiment environment	
OS	CentOS Linux 6 (Core)
CPU	XEON#W-2135
GPU	Galaxy GeForce RTX 3090 SG D6X
RAM	16GB * 8
HDD	SSD 512GB, HDD 8TB

Table 5. Descriptions 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

Table 6. Results of Each Model: 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 7. Results of Different Models

System	Min t-DCF	EER(%)
CoNSIST	0.0267	0.89
AASIST	0.0275	0.83
LCNN-LSTM-sum	0.0524	1.92
Capsule network	0.0538	1.97
GMM	0.0904	3.50
ResNet18-OC	0.0590	2.19
PC-DARTS	0.0914	4.96
MCG-Res2Net50	0.0520	1.78
SENet	0.0368	1.14

5. Conclusion

This study proposes CoNSIST, an enhanced audio deepfake detection model that builds upon the GNN-based AASIST model. CoNSIST integrates three key methodologies to improve upon AASIST: (i) Squeeze-and-Excitation blocks for more efficient feature extraction, (ii) positional encoding to capture the inherent order within the extracted features, and (iii) a reformulated HS-GAL layer that eliminates redundant operations and allows for processing of more diverse information. Experimental results confirm that CoNSIST outperforms AASIST under identical experimental conditions. Additionally, CoNSIST demonstrates greater stability in its performance across various voice synthesis techniques. However, our study has few limitations. First, the model's performance has been evaluated on a limited dataset, which may not fully capture the diversity of real-world audio deepfakes. Second, while CoNSIST has shown improved stability, its generalization on different languages has not been tested. Lastly, hyperparameter settings were not exhaustively explored, potentially leaving room for further optimization. Therefore, we expect further hyperparameter optimization, dataset expansion and augmentation techniques, and exploration of other potential enhancements. These efforts aim to further improve the generalization capabilities of CoNSIST, ultimately leading to even more accurate audio deepfake detection. Further research on hyperparameter tuning, dataset collection and augmentation, and exploration of other potential enhancements have the potential to enhance the generalization performance of CoNSIST, resulting in improved accuracy for audio deepfake detection.

References

[1] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio Deepfake Detection: A Survey." ArXiv (Cornell University), 28 Aug. 2023, https://doi.org/10.48550/arxiv.2308.14970

[2] K. H. Jung and C. H. 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] J. W. Jung 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] A. Hamza et al., "Deepfake Audio Detection via MFCC Features Using Machine Learning." *IEEE Access*, Vol.10, pp.134018-134028, 2022, https://doi.org/10.1109.

[5] M. Lataifeh and A. Elnagar, "Ar-DAD: Arabic Diversified Audio Dataset," *Data in Brief*, Nov. pp.106503, 2020, https://doi.org/10.1016/j.dib.2020.106503.

[6] C. Borrelli, P. Bestagini, F. Antoacci, A. Sarti, and S. Tubaro, "Synthetic Speech Detection through Short-Term and Long-Term Prediction Traces," *EURASIP Journal on Information Security*, Vol. No.1, 6 Apr. 2021, https://doi.org/10.1186/s13635-021- 00116-3.

[7] A. K. Singh and P. Singh, "Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics," *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, Sept. 2021, https://doi.org/10.1109/mipr51284.2021.00076.

[8] A. Chintha et al., "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection," *IEEE Journal of Selected Topics in Signal Processing*, Vol.14, No.5, pp.1024-1037, 2020, https://doi.org/10.1109/jstsp.2020.2999185.

[9] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging Positional-Related Local-Global Dependency for Synthetic Speech Detection," 4 June 2023, https://doi.org/10.1109/icassp49357.2023.10096278.

[10] H. Tak, J. W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "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] H. Tak, J. W. Jung, J. Patino, M. Todisco, and N. Evans, "Graph Attention Networks for Anti-Spoofing." ArXiv (Cornell University), 30 Aug. 2021, https://doi.org/10.21437/interspeech.2021-993.

[12] J. W. Jung, S. B. Kim, H. J. Shim, and J. H. Kim, and H. J. Yu, "Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms." ArXiv (Cornell University), 25 Oct. 2020, https://doi.org/10.21437/interspeech.2020-1011.

[13] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-To-End Anti-Spoofing with RawNet2." HAL (Le Centre Pour La Communication Scientifique Directe), 6 June 2021, https://doi.org/10.1109/icassp39728.2021.9414234.

[14] X. Wang et al., "Heterogeneous Graph Attention Network," The World Wide Web Conference, 13 May 2019, https://doi.org/10.1145/3308558.3313562.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, "Graph Attention Networks," arXiv (Cornell University), Feb. 2018, https://doi.org/10.48550/arXiv.1710.10903.

[16] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 2018, https://doi.org/10.48550/arXiv.1709.01507.

[17] P. Duffer, M. Schmitt, and H. Schütze, "Position Information in Transformers: An Overview," Computational Linguistics, Vol.48, No.3, pp.733-763, 2022, https://doi.org/10.1162/coli_a_00445.

[18] A. Vaswani et al., "Attention is All you Need," arXiv (Cornell University), Vol.30, pp.5998-6008, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[19] X. Wang 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.

[20] T. Kinnunen 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.

[21] X. Wang and J. 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.

Jae Hoon Ha

https://orcid.org/0009-0005-6357-8672

e-mail : gkwogns95@gamil.com

Graduated with a Bachelor's degree in Social Data Analytics from Pennsylvania State University in 2020. Currently pursuing a Master's degree in Digital Analytics at Yonsei University with a focus on natural language processing (NLP), machine learning, deep learning modeling, and data science.

Joo Won Mun

https://orcid.org/0009-0006-7737-9196

e-mail : yustina0514@naver.com

Earned a Bachelor's degree in Information Statistics from Duksung Women's University and is currently pursuing a Master's degree in Digital Analytics at Yonsei University. Research interests include data science, online data analysis, and artificial intelligence (AI).

Sang Yup Lee

https://orcid.org/0000-0002-8869-8428

e-mail : sangyuplee@yonsei.ac.kr

Dr. Lee is an Associate Professor at the Department of Communication, Yonsei University in South Korea. He received his Ph.D and Master's degree from the Media Information Department at Michigan State University. He studied Computer Science for his Bachelor's degree at Yonsei University. His research interests include data science, computational social science, media effects on health, and media psychology.