FuseAST: Fusing Noise and Spectrograms for Audio Deepfake Detection

Min Park[1], **Gordon Lim**[1], **Qirui Jin**[1], **William Zheng**[1], **Bill Shao**[1]

[1]University of Michigan
{phyunmin, gbtc, qiruijin, willizhe, szx}@umich.edu

Abstract

In recent years, Artificial Intelligence (AI) technologies have shown great promise in various domains, such as video recommendation, bioinformatics, and speech recognition. Our research focuses on audio speech verification (ASV) systems, which are widely used in critical security systems and real-world applications. However, with the increasing pervasiveness and capability of AI systems, the security and integrity of many ASVs have come into question. In particular, AI-enabled attacks via text-to-speech synthesis and voice conversion can generate deepfake audio to mimic individuals' voices, posing significant threats through spoofing attacks. To address these vulnerabilities, we present FUSEAST, based on state-of-the-art Audio Spectrogram Transformers (AST) for distinguishing real audio from deepfake audio. FUSEAST introduces two different feature engineering strategies: (1) fusing Log-Mel spectrograms with delta and delta-delta features in three channels, representing the first and second temporal derivatives, interpreting the temporal dynamics of the audio, and (2) fusing features of high and low-frequency regions in the spectrogram with weight masking, providing different perspectives for analysis.

Our evaluation showed that the integration of the two feature engineering strategies resulted in a significant 6% improvement in performance accuracy over the Vallina AST model as evaluated on the ASVspoof challenges. In addition, FUSEAST significantly outperforms the baseline LFCC-GMM model on the Equal Error Rate (EER) metric. FUSEAST underscores the effectiveness of innovative feature engineering in strengthening the security of ASV systems against AI-driven spoofing attacks.

Introduction

As audio deepfakes become increasingly realistic, it has become increasingly difficult to distinguish them by ear alone. In this context, deepfake technology has been exploited for malicious purposes, as seen in a 2019 case where fraudsters used AI-based software to mimic a CEO's voice and successfully stole over $243,000 via a phone call (TrendMicro 2019). To address this growing concern, deep learning methods have emerged as a promising solution for audio deepfake detection. Two primary approaches have been explored for processing audio data with neural networks: (1) converting audio into spectrograms (visual representations of sound) and applying traditional computer vision techniques (Firc, Malinka, and Hanáček 2024), and (2) extracting features directly from raw audio (Zhang, Wen, and Hu 2024). The latter approach has shown great promise in recent studies (Zhang, Wen, and Hu 2024).

Specifically, we experiment with the state-of-the-art Audio Spectrogram Transformer (AST) (Gong, Chung, and Glass 2021) of its effectiveness in detecting audio deepfake. AST is a Vision Transformer (ViT)-based model (Dosovitskiy et al. 2021) that represents the state-of-the-art (SOTA) in audio classification tasks, offering a novel approach in comparison to existing methodologies like LightCNN (Wu et al. 2018), ResNet (He et al. 2015), and LSTM (Hochreiter and Schmidhuber 1997) models. It has demonstrated strong performance in multi-class audio classification tasks and achieves state-of-the-art results on AudioSet, 95.6% accuracy on RSC-50, and 98.1% on Speech Commands V2. Each audio clip is converted into an audio spectrogram. Then, the AST implementation concatenations three audio spectrograms into a three-channel image and use this image as an input of the model. To classify audio spectrograms with variable sizes, (Gong, Chung, and Glass 2021) uses linear interpolation to map the oversized audio spectrogram to the original positional embedding of ViT.

Since AST has demonstrated strong performance in multi-class audio classification tasks, we propose adapting it for audio deepfake detection. Our proposed approach, FUSEAST introduces two innovations:

- To capture a broader range of audio features and improve detection performance, we propose integrating a multi-spectrogram fusion technique with AST. In particular, we leverage log-mel, delta, and delta-delta spectrograms as separate channels, each spectrogram has a different focus. Although variants of this method have been explored with other architectures and applications (Zheng et al. 2017; Liu and Fang 2023), to the best of our knowledge, this is the first work to utilize multi-spectrogram fusion with AST for the task of audio deepfake detection.
- During our naive evaluation, we found out that although deepfake tools generate convincing voices, they often fail to imitate the naturalness of background noise, which corresponds to the low and high frequency regions in the spectrogram. Base on this finding, we propose high-low frequency noise fusion. By applying a weight boost of 0.5 for low and high frequency areas, the model can learn

more from rich-info regions.

The datasets used for model training and evaluation on both vallina AST and FUSEAST come from the ASVspoof 2019 and 2021 challenges, with a particular focus on the Deep Fake (DF) and Logical Access (LA) datasets. Our experiments show overall success. The vanilla AST, trained on the ASVspoof 2019 dataset, achieves an accuracy of 80.8% with an AUC of 0.914 on a balanced ASVspoof 2021 dataset. The AST with focus on high-low frequency bias, trained on the ASVspoof 2019 dataset, achieves an accuracy of 85.7% with an AUC of 0.922 on the balanced dataset of ASVspoof 2021. Although the AST with multi-spectrogram fusion shows an improved accuracy of 90%, the AUC value is reduced to 0.898. This may be due to the fact that we use pretrained weights from AST, which only uses Log-Mel spectrograms as input.

Background

Spectrogram is a visual representation of the frequency spectrum of a signal over time, with time on one axis, frequency on the other, and amplitude indicated by intensity or color. It is widely used in audio analysis, including speech and spoofing detection.

ASVspoof 2021 dataset (Liu et al. 2023) serves as a benchmark for the development of countermeasures (CMs) against spoofing attacks targeting Automatic Speaker Verification (ASV) systems. The challenge consists of 3 different tasks: Physical Access (PA), Logical Access (LA), and DeepFake (DF). In this paper, we focus on the LA task. It includes bonafide and fake audio generated using 13 state-of-the-art voice conversion (VC), text-to-speech (TTS), and hybrid algorithms, as well as different transmission conditions such as VoIP and PSTN+VoIP. The dataset is derived from the ASVspoof 2019 LA evaluation database, which in turn is based on the VCTK database.

Performance metrics include:

- **Accuracy:** Measures overall classification accuracy.
- **Equal Error Rate (EER):** Balances false acceptance and false rejection rates.
- **Area Under the Curve (AUC):** Evaluates classification performance.
- **Precision:** Evaluates the proportion of true positives.
- **Recall:** Measures the percentage of correctly identified positive cases.

These metrics provide a comprehensive framework for evaluating system robustness in real-world scenarios.

Methodology

In this section, we discuss two types of spectrogram feature engineering efforts that we have explored in this thesis.

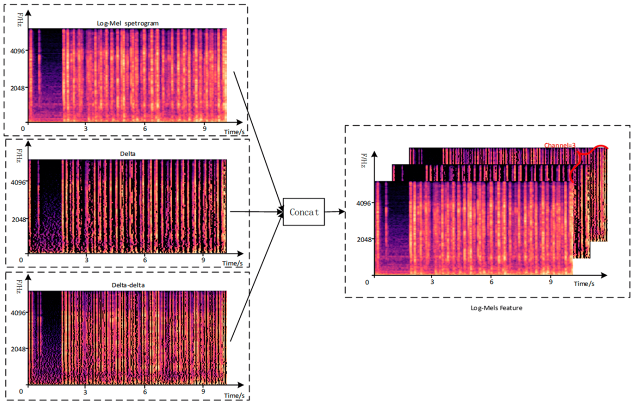

Figure 1: This figure, taken from Liu and Fang (2023), illustrates multi-spectrogram fusion using Log-Mel spectrogram, delta, and delta-delta features.

Multi-Spectrogram Fusion AST traditionally takes Log-Mel spectrograms as input. Log-Mel captures acoustic features such as frequency and amplitude in one form. However, while Log-Mel spectrograms are widely used, they may not comprehensively capture temporal dynamics. To address this, we incorporate delta and delta-delta features, representing the first and second temporal derivatives, respectively (Liu and Fang 2023). By concatenating these features along the channel dimension, we construct a 3D representation that enhances temporal information and improves the model's ability to learn nuanced patterns (Liu and Fang 2023). While previous work has successfully applied this technique to audio classification (Liu and Fang 2023), its application to the more challenging task of audio deepfake detection remains unexplored.

Adapting AST to multi-channel spectrograms, however, requires modifications. Traditionally, AST processes single-channel spectrograms, requiring changes such as the introduction of a new projection layer in the PatchEmbed module to handle multi-channel input. This adjustment introduces additional weights for which pretrained parameters from AudioSet or ImageNet are not available. Due to computational constraints, training the model from scratch is not feasible. Instead, we adjust the pre-trained weights for the Log-Mel spectrogram channel to initialize the delta and delta-delta channels. These weights are finetuned during training, allowing the model to learn and optimize for the additional temporal features.

Noise Fusion via Weight Masking The Vision Transformer (ViT), the foundational architecture for the AST model, is designed to capture both local and global features of input data through its transformer-based framework using patch embeddings and positional embeddings. For example, in the context of image analysis, ViT captures local features, such as the structure of a nose, while modeling global relationships, such as the spatial arrangement of the eyes, nose, and mouth. These relationships are reflected in the embeddings that the model focuses on during processing.

When applied to spectrograms, the model's ability to identify and emphasize specific patches allows it to capture information corresponding to specific frequency bands. The fundamental frequencies (pitch) of human speech typically range from 85 Hz to 255 Hz, while the harmonics and overtones critical to speech intelligibility (e.g., vowels and consonants) are primarily between 255 Hz and 4

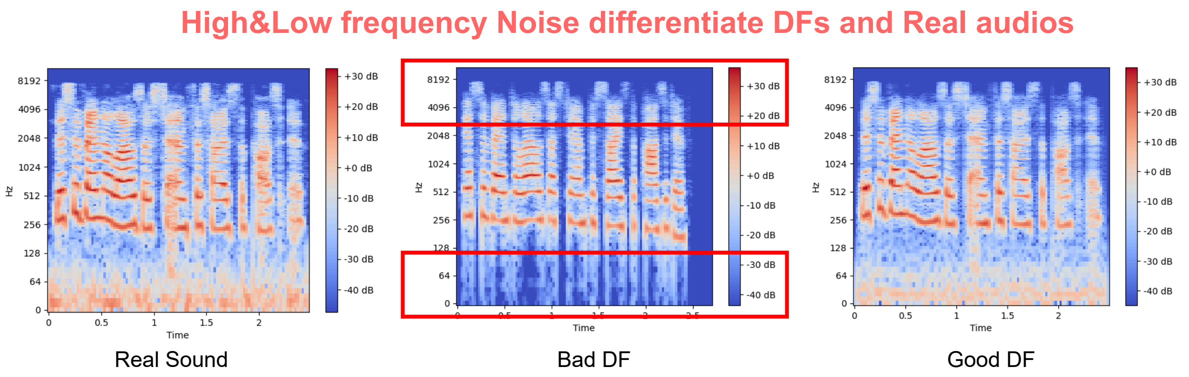

Figure 2: This figure shows an example of deepfake and ground-truth audio spectrograms. *Bad DF* indicates deepfake audio that is distinguishable using human ears, *Good DF* indicates indistinguishable deepfake, and the boxed region indicates the weight-biased region.

kHz. Since most speech audio is sampled at 16 kHz, the Nyquist theorem ensures coverage up to 8 kHz. However, frequencies above the vocal band often contain features unrelated to speech, such as background noise and environmental sounds.

We hypothesize that AI-based voice conversion models prioritize synthesizing natural human voice components, potentially resulting in less accurate synthesis of background noise and other non-human sound characteristics. We also try to find evidence that such a pattern exits, as Figure 2 presents. To address this, we fuse noise in the feature by implementing a weighted embedding mask that targets specific patches corresponding to high and low-frequency regions in the spectrogram. Specifically, weight biases of 0.5 were applied to the patch embeddings of the top 20% and bottom 20% frequency bins to encourage the model to pay more attention to these regions. Note that the 20% ratio was determined by the proportion of the background frequencies in the log-mel spectrograms, and the bias of 0.5 is a tuned hyperparameter. This approach aims to increase the model's sensitivity to critical frequency bands associated with noise and reduce the influence of less relevant human-related regions.

Experiments, Results, and Discussion

We trained both the standard AST model and FuseAST (AST models incorporating feature engineering) using pre-trained weights from AudioSet and ImageNet, followed by fine-tuning. The models were fine-tuned on the ASVspoof2021 training dataset provided in the challenge, derived from the ASVspoof2019 LA dataset. The evaluation was conducted on a balanced subset of the ASVspoof2021 LA evaluation set, comprising 15,000 bonafide and 15,000 spoofed samples. This balanced evaluation set was employed to mitigate the effects of the substantial class imbalance in the training data, which had a 10:1 ratio of spoofed to bonafide samples. By using a balanced evaluation set, we ensured that the performance metrics more accurately reflect the model's generalization capability across both bonafide and spoofed classes, avoiding biases introduced by class imbalance. A comparative analysis was conducted to assess the impact of feature engineering on the models' performance.

Multi-spectrogram fusion

Table 1 shows a comparison between the Vanilla AST model and the FuseAST (Spectrogram Fusion). It is observed that FuseAST achieves improved performance and reliability, with higher accuracy (0.9 vs. 0.81) and recall (0.82 vs. 0.81). However, we also observe a decrease in AUC (0.9 vs. 0.91), Precision (0.79 vs. 0.90), and an unchanged Equal Error Rate (EER). This result could possibly be attributed to improper weight initialization, where the pretrained weights for the Log-Mel spectrogram were reused for the delta and delta-delta features. Alternatively, it may indicate that the vanilla model in its current form is not equipped to handle the increased complexity introduced by multi-channel input.

Weight Masking

Table 1 presents a performance comparison between the Vanilla AST model and the FuseAST (Weight Masking) model and demonstrates the effectiveness of feature engineering in improving spoof detection. The FuseAST model achieves a higher AUC (0.92 vs. 0.91) and accuracy (0.86 vs. 0.81), indicating improved overall classification performance and reliability. Precision is marginally higher for FuseAST (0.91 vs. 0.90), showing its ability to correctly identify spoofed audio with slightly greater consistency. Recall also improves notably (0.85 vs. 0.81), reflecting the model's enhanced capability to detect bonafide audio without missing cases. The Equal Error Rate (EER), a critical metric in spoof detection, is significantly reduced in FuseAST (0.128 vs. 0.154), emphasizing its robustness and superior generalization across unseen scenarios. Collectively, these results highlight the impact of the proposed feature engineering on enhancing the AST model's performance across multiple evaluation metrics.

Table 1: The evaluation results of Vallina AST and FuseAST on ASVspoof 2021

Models	Vallina AST	**FuseAST (Spectrogram Fusion)**	**FuseAST (Weight Masking)**
AUC	0.91	0.9	**0.92**
Accuracy	0.81	**0.9**	0.86
Precision	0.9	0.79	**0.91**
Recall	0.81	0.82	**0.85**
EER	0.154	0.154	**0.128**

Table 2: This is the result of the LFCC-GMM baseline evaluated on the ASVspoof 2021 LA eval balanced subset (only 30000 samples, 1:1 bona-fide: spoof). The EER numbers listed below are in percentages. This table was generated from code found in the ASVspoof challenge GitHub repository: https://github.com/asvspoof-challenge/2021

	C1	C2	C3	C4	C5	C6	C7	Pooled
A07	20.31	31.20	36.71	29.62	32.24	29.63	18.55	28.76
A08	0.89	4.50	13.10	2.28	5.06	7.15	0.08	6.39
A09	0.00	1.61	7.68	0.89	0.08	3.03	0.00	3.45
A10	24.47	35.93	32.09	33.04	34.00	31.19	25.00	31.95
A11	7.65	18.75	33.82	9.39	20.27	26.52	8.99	18.51
A12	2.25	39.55	36.82	9.64	21.19	37.08	4.46	23.22
A13	3.37	13.28	10.10	6.63	8.27	12.09	3.33	8.70
A14	12.09	18.68	29.36	14.08	21.85	14.80	10.28	17.34
A15	7.87	15.13	27.79	11.28	20.87	16.75	8.31	17.40
A16	12.09	34.12	31.26	14.58	26.33	27.84	10.34	23.11
A17	11.85	18.76	23.91	13.52	16.11	21.71	18.56	17.98
A18	11.64	12.21	33.32	12.86	12.73	22.66	9.72	16.54
A19	27.28	27.14	35.28	25.63	29.04	29.49	29.48	30.47
Pooled	14.00	22.52	28.14	16.03	20.64	23.64	13.46	20.15

Table 3: This is the result of the LFCC-GMM baseline evaluated on the Full ASVspoof 2021 LA evaluation dataset. The EER numbers listed below are in percentages. This table was generated from code found in the ASVspoof challenge GitHub repository: https://github.com/asvspoof-challenge/2021

	C1	C2	C3	C4	C5	C6	C7	Pooled
A07	21.29	29.30	33.80	26.16	28.87	29.29	19.73	27.74
A08	0.36	5.41	11.93	1.33	5.87	7.09	0.73	5.97
A09	0.00	1.10	8.10	0.15	0.55	2.59	0.05	3.35
A10	27.44	36.81	33.17	32.06	35.01	31.10	24.08	32.33
A11	6.67	22.51	32.97	11.04	21.30	24.96	10.15	19.29
A12	3.28	42.25	34.98	11.96	18.68	36.25	3.52	23.06
A13	3.52	9.40	11.41	6.39	9.63	11.76	4.58	8.97
A14	11.54	21.46	29.32	14.70	21.31	16.59	10.32	18.53
A15	9.86	20.22	30.53	11.61	19.92	15.91	8.73	17.75
A16	8.91	31.08	31.38	18.21	25.05	29.89	10.95	23.16
A17	13.04	14.78	22.24	12.81	17.57	19.48	16.46	17.31
A18	10.66	12.48	30.51	12.45	11.98	20.85	9.58	15.98
A19	27.34	26.78	34.15	26.79	28.11	28.86	29.59	29.85
Pooled	13.55	22.56	27.65	16.11	20.16	22.62	13.41	19.99

ASVspoof LFCC-GMM Baseline

Tables 2 and 3 present the performance results of the LFCC-GMM baseline model evaluated on the same dataset as the AST models. While the ASVspoof 2021 paper provides four benchmark models, we focused on the LFCC-GMM baseline (referred to as B02 in the original paper (Liu et al. 2023)) due to resource and time constraints. To ensure comparability, we modified the training process of the baseline model to match the training setup of the AST models.

The results in Tables 2 and 3 show that FuseAST significantly outperforms the baseline model, with a pooled EER reduction from 0.201 to 0.128. This highlights the superior performance of FUSEAST in detecting spoofed audio under the same evaluation conditions.

The purpose of the LA ASVspoof Challenge was to address the gap between laboratory and real-world settings in the detection of TTS and VC. In real-world settings, there is potential noise that could be introduced by packet loss and other artifacts caused by transmission infrastructures, transmission rates, etc. Specifically, the paper tested spoofed speech audio transmission through real telephony systems, such as the Voice-over-Internet-Protocol (VoIP) system and the Public Switched Telephone Network (PSTN) system (Liu et al. 2023). These audio transmission modification conditions are referred to as C1 through C7. C1 stands for no transmission artifacts, C2 and C4 to C7 are transmitted via VoIP, and C3 is transmitted via PSTN. For more information on the codec used, sample rate, and bit rate, see (Liu et al. 2023). In the ASVspoof 2021 challenge, 13 different algorithms were tested to generate the TTS and VC spoofing attacks. These algorithms are shown as A07 through A19. For more details on what exactly these algorithms are, see (Liu et al. 2023).

Related Work

In recent years, numerous efforts have been made to detect deepfakes in two main areas: traditional classification and deep learning classification. Traditional classifiers, although easy and widely used, have their performance limited either by the relatively small sample size compared to the massive amount of spoofed audio, while deep learning classifiers, including convolutional neural networks, graph neural networks, and transformers (Yi et al. 2023).

The generalization ability to fool countermeasure systems in real-world unseen scenarios remains a significant challenge, especially for deep learning models. To address this issue, many studies have focused on the development of novel data augmentation techniques. For example, the DKU-CMRI system evaluates data augmentation methods to improve performance in spoofing detection tasks. This approach improves the quality of training data by using the Opus codec and the SoX toolkit to simulate different coding and transmission environments using codecs such as G. 711-law, G.722, GSM-FR, and G.729. For the LA task, baseline models such as RawNet2 and LFCC-LCNN are used for spoofed/bonafide audio classification, achieving a min-tDCF of 0.3310 and an EER of 8.23% (Wang et al. 2021)

In contrast, UR Channel uses Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Networks (ECAPA-TDNN) as the primary model, achieving an EER of 20.33% on the DF task and an EER of 5.46% with a min-tDCF of 0.3094 on the LA task. To enhance robustness, the system integrates one-class learning and channel-robust training strategies for channel-invariant speech representations. In addition, task-specific data augmentation is used, incorporating MP3 and AAC codecs for the DF task and transmission codecs (landline, cellular, VoIP) with device impulse responses for the LA task to address data set biases (Chen et al. 2021).

Unlike these approaches, FUSEAST focuses on feature engineering rather than data augmentation. While methods such as transfer codecs could potentially improve performance by addressing robustness and dataset imbalances, FUSEAST emphasizes the comparison between plain input data and feature-engineered inputs. Our experiments demonstrate notable improvements in performance metrics with feature engineering, and future work will explore the integration of data augmentation techniques to achieve further gains.

Extensions

The original AST is designed to classify audio files. Based on AST, we have made three innovations. First, since AST has proved to be state-of-the-art in audio classification, we propose that AST is capable of classifying deepfake audios. To test our hypothesis, we fine-tuned AST with ASVspoof 2019 and achieved 80.8% accuracy and 0.922 AUC. To further improve the performance, we examined the audio spectrograms of the deepfake audios. We found that the non-human audible frequencies of the fake audio were not as complicated as the real audio. Therefore, we hypothesized that our model should emphasize the non-human-audible frequency ranges. Then, we fine-tuned an AST with the ASVspoof 2019 dataset with embedding weight masking and achieved an accuracy of 85.7% with an AUC of 0.922. This increase in accuracy indicates that focusing on the non-human audible frequency ranges is effective in identifying deepfakes. In addition, in order to capture more details of the audio clip in its audio spectrogram, we included the delta spectrogram and the delta-delta spectrogram in addition to the log-mel spectrogram as input to our model. We believe that including more details of the original audio is helpful for the performance of the model. Unfortunately, our model did not guarantee a performance increase in all metrics. One reason may be that the pretrained weights of AST are based only on the Log-Mel spectrogram, so the weights of our model are incorrectly initialized, resulting in relatively low performance. Alternatively, it may indicate that the AST model in its current form is not equipped to handle the increased complexity introduced by multi-channel input.

Conclusions

To conclude, we finetuned AST with ASVspoof 2019 and successfully transformed it into a deepfake audio classifier. We then modified the embedding layer of AST to increase the weights of non-human audible frequencies, which resulted in an increase in performance. This suggests that focusing on the extremely low and high frequencies may be a shortcut to identifying deepfake audio. We also tried using multi-channel audio spectrograms, where each channel is a different type of spectrogram, but our model does not boost the performance in every metric. Future work can be done on feature engineering based on audio frequencies and corresponding embedding layers, and fine-tuning ViT directly on multi-spectrogram fusion.

Societal Impact

The increasing capability of AI models to generate indistinguishable deep fake audio presents a significant threat to society, as evidenced by the numerous victims of spoofing attacks that have already been documented. Consequently, the development of effective countermeasures is imperative. Our research into Audio Spectrogram Transformer (AST) models, combined with novel feature engineering strategies, has led to the identification of a promising new method for detecting deep fake audio. We argue that the ongoing development and application of AST models for deep fake audio detection will play a vital role in enhancing AI security.

References

Chen, X.; Zhang, Y.; Zhu, G.; and Duan, Z. 2021. UR Channel-Robust Synthetic Speech Detection System for ASVspoof 2021. arXiv:2107.12018.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Firc, A.; Malinka, K.; and Hanáček, P. 2024. Deepfake Speech Detection: A Spectrogram Analysis. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, SAC '24, 1312–1320. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702433.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. AST: Audio Spectrogram Transformer. arXiv:2104.01778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Liu, F.; and Fang, J. 2023. Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition. *Future Internet*, 15(2).

Liu, X.; Wang, X.; Sahidullah, M.; Patino, J.; Delgado, H.; Kinnunen, T.; Todisco, M.; Yamagishi, J.; Evans, N.; Nautsch, A.; and Lee, K. A. 2023. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2507–2522.

TrendMicro. 2019. Unusual CEO Fraud via Deepfake Audio Steals US$243,000 FromUKCompany.Accessed : 2024-12-02$.

Wang, X.; Qin, X.; Zhu, T.; Wang, C.; Zhang, S.; and Li, M. 2021. The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder based Replay Channel Response Estimation. In *Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 16–21.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11): 2884–2896.

Yi, J.; Wang, C.; Tao, J.; Zhang, X.; Zhang, C. Y.; and Zhao, Y. 2023. Audio Deepfake Detection: A Survey. arXiv:2308.14970.

Zhang, Q.; Wen, S.; and Hu, T. 2024. Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 6765–6773. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Zheng, W.; Yi, J.; Xing, X.; Liu, X.; and Peng, S.-H. 2017. Acoustic Scene Classification Using Deep Convolutional Neural Network and Multiple Spectrograms Fusion. In *Workshop on Detection and Classification of Acoustic Scenes and Events*.

Individual Contributions

Gordon My contributions to the project include facilitating meetings, creating slides for updates, booking meeting venues, and tracking course deliverables. I also wrote meeting minutes and provided tutorials on using Overleaf and GitHub. I researched the datasets we intended to use, though we ultimately didn't have time to test them all. Additionally, I set up our GitHub repository by reorganizing our fork to better align with the project's focus and wrote the initial draft of the README documentation. On a technical level, I developed the multi-fusion spectrogram techniques and ran our experiments to evaluate it.

Hyunmin I proposed utilizing transformer models for audio deepfake detection, incorporating weight masking on specific frequency regions to enhance model focus and performance. I was responsible for training and evaluating AST models, optimizing their performance while ensuring seamless integration with the implemented feature engineering methods. Additionally, I refined the preprocessing pipeline to align with the dataset requirements and created custom-balanced evaluation sets from the ASVspoof challenge dataset to enable more robust and fair performance comparisons.

William In terms of group contributions, I wrote the meeting notes for our group meetings in the last month of the replication project. On the technical side, I was in charge of researching potential datasets and ways to create our own in-the-wild dataset. In that pursuit, I made a basic script for scrapping audio data from YouTube in addition to video captions/transcripts. Due to time constraints, I mainly focused on research and completion of baselines for our paper. From this, I have contributed Table 2, Table 3, and the ASVspoof LFCC-GMM Baseline section of this paper.

Bill For logistics aspects, I participate in group meeting, create meeting minute slides, and do reaction notes with my teem. For the replication project, I set up three datasets for evaluation and read relevant papers. I engineered the patch embedment layer and the position embedment layer, which led to our weight masking idea and implementation.

Qirui My contributions to this project were literature research, idea proposal, extension implementation, and paper writing. During weekly meetings, I read relevant papers on audio deepfake detection methods and summarized the core idea for the group. During the idea proposal phase, I proposed several ideas: verifying the robustness of AST using

CROWN, multi-view spectrogram fusion, and also changing the positional embeddings for noise feature fusion. In particular, I manually examined the raw spectrogram (Figure 2, which revealed that deepfake tools are unable to generate high-quality background noise, which inspired and proved our idea to focus more on background noise. In code implementation, I implemented multi-view spectrogram fusion, and also the position embedding-based method background noise feature fusion. In paper writing, I was mainly responsible for writing part of the methodology and experiment. I also revise other members' writings.