Research Proposal:
Towards Image Transformers Architectures for Detecting Audio Deepfakes

Min Park, Gordon Lim, Qirui Jin, William Zheng, Bill Shao
{phyunmin, gbtc, qiruijin, willizhe, szx}@umich.edu

Key Research Problem

The core research question of this project is: Can a Vision Transformer (ViT) effectively capture and classify the information embedded within an audio spectrogram to distinguish AI-synthesized audio from human-generated audio? This project builds on the work presented in the Audio Spectrogram Transformer (AST) (Gong, Chung, and Glass 2021), which introduces a convolution-free, attention-based approach, leveraging the Vision Transformer (ViT) model to capture long-range global context in spectrograms. AST demonstrates that audio spectrograms, when treated as 2D images, can be effectively classified using an attention mechanism.

While the AST model has achieved state-of-the-art performance on general audio classification tasks, this project aims to explore a novel application: determining whether AST can distinguish between AI-synthesized audio and human-generated audio. The challenge lies in whether the AST model, optimized for general audio classification, can adapt to the nuances and subtle differences present in synthetic audio. By extending AST's framework, this project seeks to push the boundaries of transfer learning and attention-based models in the context of audio analysis, offering a fresh perspective on how machine learning can tackle the growing complexity of AI-generated content. The project will utilize the **ASVspoof 2021 dataset** (Liu et al. 2023), a benchmark dataset specifically designed for detecting synthesized speech, including voice conversion (VC) and text-to-speech (TTS) attacks. There is a Github repository available for employing AST with pretrained ImageNet (Deng et al. 2009) and AudioSet (Gemmeke et al. 2017) weights. We will begin by benchmarking these pretrained models. To do so, we will have to preprocess the ASVspoof dataset for the AST which includes converting the data into the correct input size. Based on our initial results, we will investigating methods to optimize the attention blocks to specifically target AI-sythensized data.

Applying Transformers to detect AI-synthesized audio data presents several anticipated challenges. First, our literature review indicates that this approach has not been investigated, suggesting potential difficulties in adapting Transformer models to effectively capture the nuances of deepfakes beyond general sound classification tasks. Another significant challenge is the imbalance between *spoofed* (AI-synthesized) and *bonafide* (human) audio samples. There are virtually limitless ways to spoof audio samples, resulting in a vast range of potential synthetic outputs. In contrast, bonafide speech with clear context is inherently limited, as seen in datasets like ASVspoof, WaveFake (Frank and Schönherr 2021), and FakeAVCeleb (Khalid, Tariq, and Woo 2021). This imbalance can hinder the model's ability to learn effectively, prompting us to investigate the generation of our "in-the-wild" dataset for reliable training and evaluation. Additionally, as we extend our attention model to focus on specific features of the spectrogram, we face performance verification concerns. It is crucial to ensure that the model captures relevant features while maintaining or improving detection accuracy. The interplay between model complexity, feature extraction, and performance metrics will require careful evaluation and optimization to address these complications successfully.

The audio files are provided in 16 kHz format, which is a standard sampling rate for speech processing tasks and aligns with the pretrained AST model, as the AudioSet dataset used to pretrain AST also uses the same sampling rate. This consistency in audio format ensures compatibility and allows for efficient transfer learning moreover ASVspoof is well established dataset with a lot of associated papers for coming up with models with remarkable performance so we anticipate that we will be able to complete it within the given time.

Key Contributions

Significance

AI-synthesized tools capable of generating convincing voices have gained prominence in recent years. Initially developed to enhance communication and creativity, these technologies are widely used in various applications, including entertainment content on platforms like YouTube and AI-assisted services that support individuals with disabilities. However, the malicious use of these tools has led to significant concerns over the phenomenon known as *Audio Deepfake*. The vast amount of audio data available on the internet, combined with open-source models, enables users to create sophisticated voice manipulations using basic mobile devices or personal computers, targeting individuals and organizations alike. A notable incident involved cybersecurity researcher Kyle Wilhoit on 2024, whose family was targeted by a convincing voice-clone scam. In this case, a caller posed as Wilhoit's daughter, claiming she had been in a car accident and urgently needed money for damages. The caller then connected Wilhoit to a person on the phone who sounded exactly like his daughter, creating a sense of urgency and distress. The scammers had previously called his daughter multiple times to gather audio samples, which they used to produce the convincing deepfake of her voice . Another significant incident occurred in 2019 when fraudsters employed AI-based software to impersonate a CEO's voice, successfully swindling over $243,000 via a telephone call. This incident highlighted the urgent need for effective authentication of audio recordings to prevent the spread of disinformation and financial fraud. Given the urgency of addressing the challenges posed by audio deepfakes, leveraging advanced machine learning techniques becomes critical. Transformers have demonstrated state-of-the-art performance in various classification tasks, particularly when trained on extensive datasets. The use of transfer learning, as shown in the Audio Spectrogram Transformer (AST) paper, allows models to adapt efficiently to new tasks with high accuracy, even when limited data is available. By applying Transformer models to the problem of detecting AI-synthesized audio, we can anticipate significant performance improvements in distinguishing between authentic and manipulated audio recordings .

Replication

The authors of the Audio Spectrogram Transformer (AST) paper achieved several key results that highlight the model's effectiveness and versatility in audio classification tasks. Here are the primary results and areas of interest for replication: 1. Performance Comparison on CNN Networks and AST: The AST model demonstrated superior performance in sound source classification tasks compared to traditional Convolutional Neural Networks (CNNs). This finding is significant as it challenges the conventional reliance on CNNs for audio machine learning, indicating that the Transformer architecture may offer a more effective approach for certain classification problems . 2. Ablation Study: The paper includes an ablation study that assesses various design choices, such as the effectiveness of ImageNet pretraining, optimal positional embedding adaptation, and patch overlap and size. Understanding these factors is crucial for optimizing models, especially for tasks like audio deepfake classification. This section provides insights into how specific modifications can enhance model performance . 3. Comparison Against State-of-the-Art Models on ESC-50 (Piczak 2015) and Speech Commands V2 (Warden 2018): The AST model was evaluated against existing SOTA models on the ESC-50 and Speech Commands V2 datasets, establishing a strong foundation for the claim that Transformers can outperform traditional models in audio classification tasks. This comparison underscores the potential of the AST approach in real-world applications, providing compelling evidence for its efficacy. In particular, our replication projects will consist of the following milestones:

- Training the AST Model: We plan to replicate the training of the AST model using transferred weights from the model pretrained on ImageNet and AudioSet, specifically on the ESC-50 dataset. This will help verify the performance results reported in the paper and understand how these pretrained weights influence the model's effectiveness.
- Evaluating Performance on ASVspoof: We will also employ the same pretrained weights (ImageNet and AudioSet) to train the AST model on the ASVspoof dataset. This evaluation will allow us to assess the model's performance against established benchmarks and SOTA models, providing insights into its applicability in audio deepfake detection .
- Conducting an Ablation Study: Following the methodology outlined in the AST paper, we will conduct an ablation study to analyze the impact of various design choices on model performance. This will include examining the effects of different pretraining strategies and architectural modifications, contributing to our understanding of how to optimize the AST model for specific audio classification tasks.

Through these replication efforts, we aim to gain a deeper understanding of the AST model's design choices and performance characteristics, which will inform future work and potential extensions in the realm of audio classification.

Extension(s)

We propose two main ideas for extending this project. First, in our initial exploration of the the ASVspoof dataset, we noticed there was a large imbalance in the number between real audio data and synthetic audio data. Specifically, there was much more synthetically generated audio data than there was real audio data. The potential consequence on having more synthetic audio data in the dataset means that models trained on this dataset could potentially be detecting more false negatives (negatives being synthetic data). This motivated us to investigate if it is possible to augment other sources of data to the existing dataset. Currently, the real audio data is pulled from the VCTK (Yamagishi, Veaux, and MacDonald 2019) audio dataset. This dataset contains the audio of 110 English speakers that each read 400 sentences, which were pulled from the Herald Glasgow newspapers. In addition, there was an algorithm that selected sentences that would yield the most contextual and phonetic coverage from the newspapers (Yamagishi, Veaux, and MacDonald 2019). Presumably, they filtered the possible sentences spoken due to the constraint on needing volunteer participants to read sentences. This filtering could limit the performance of the model in detecting genuine audio from fake audio. So instead, we propose a new method of collecting genuine audio data. Specifically, YouTube provides a much larger collection of real audio data. Documentaries, video essays, podcasts, etc. provide potential for a much broader range of contextual and phonetic coverage. These media can come with hand-written captions and transcripts for much of the dialogue within these videos. YouTube does provide an API for researchers to to access the data they have collected. However, in order to use the API, we would need to submit an application to do so. If we cannot access the data that YouTube has collected, we could potentially source real audio from other media, such as audio books. In either case, significant time is necessary to get familiar with new code libraries and APIs. The initial estimation is that it will take 2 weeks for the on-boarding process. The cost of audio books is none if we are using audio books from the University of Michigan library. They have an extensive collection of audio books that we can directly access now. In contrast, we are uncertain how long the approval process would take to use YouTube's API, possibly over a week. However, once approved, there is no monetary cost for using the API. To measure the quality of our data generated, we could use multiple different metrics. One possibility would be to train a model on our new dataset, and see the models performance on current benchmarks. Another possibility would be to randomly select a sample from the dataset for insight on the quality of the dataset. In either case, the process can potentially be very time consuming.

Second, We plan to enhance our model's embeddings by incorporating background sounds from audio recordings, as these often include significant noise alongside human voices. Since AI-synthesized audio primarily focuses on generating convincing speech, we hypothesize that deepfake models may perform poorly on generating background sounds, which are typically not accounted for in their design. By specifically targeting the higher frequency components of background noise found in natural voice data, we believe we can improve the model's performance. Our approach will involve identifying and extracting these high-frequency background sounds via feature masking, then either integrating them into the model's embeddings or emphasizing their importance during training.

Individual Contributions

Our assigned managerial roles are as follows:

- Project Management: **Gordon** will organize and facilitate meetings, as well as keep track of everyone's progress.
- Strategy: **Min** will keep an eye on the big picture and ensure that everyone's work is aligned with our project goals.
- Devil's Advocate: **Qirui** will put on the hat of a skeptical "peer-reviewer" and offer critical feedback.
- Scribe: **William** will write meeting minutes.
- Course Logistics: **Bill** will ensure that all assignments are submitted on time.

As for our technical roles, we have divided into teams as follows:

- Detection: **Min** and **Bill** will lead the research on our proposed detection model.
- Generation: **Gordon** and **Qirui** will lead the research on deepfake generation techniques.
- Data Mining: **Will** will mine additional audio samples to evaluate the detection models. This includes ensuring the permissions and rights to use the data.

All of these teams will work closely together. The *Data Mining* team will provide the *Generation* team the bonafide speech samples to create bonafide and spoofed pairs to evaluate our proposed detection model. The insights gained by the *Generation* team will inform the *Detection* team's feature engineering efforts.

Next Steps

1. **Employ the Model and Train on the ESC-50 Dataset (By November 1)**
   - **Task:** Begin training the AST model on the ESC-50 dataset to verify its performance. Members A and B will set up the training environment, configure hyperparameters, and ensure data preprocessing aligns with model requirements. They will document the training process and results for analysis.
2. **Preprocess ASVspoof Data (By November 3)**
   - **Task:** Work on preprocessing the ASVspoof dataset to match the input size expected by the AST model. This includes normalizing audio files, trimming or padding them to the appropriate length, and converting them into the required spectrogram format. The goal is to have a clean, well-structured dataset ready for model input.

Week 2: Dataset Generation and Further Training

1. **Start Generating the New Dataset (By November 5)**
   - **Task:** Initiate the process of scraping audio data from the internet to create an ”in-the-wild” dataset. Focus on collecting various audio clips that include both human voices and background sounds, ensuring compliance with copyright and privacy regulations. Organize the data collection process for easy integration into the training pipeline.
2. **Train the Unchanged Model on ASVspoof (By November 7)**
   - **Task:** Train the AST model on the ASVspoof dataset without modifications to establish a baseline performance. After training, compare the results against state-of-the-art models to evaluate the effectiveness of the AST in detecting audio deepfakes. Analyze the performance metrics and document observations regarding the model's strengths and weaknesses.
3. **Work on Adding Background Information into Embeddings (Ongoing, By November 7)**
   - **Task:** Investigate methods to incorporate background sound information into the model's embeddings. This involves identifying high-frequency background sounds in the audio recordings and exploring techniques to extract and integrate these features into the model's architecture. Analyze how these modifications impact model performance and potentially improve the detection of manipulated audio.

Timeline

Please refer to Table 1 for a detailed timeline.

Target Venue

Our group has identified **Interspeech**[1] as a target publication venue. Interspeech is the world's leading conference focused on the science and technology of spoken language with several papers published on the detection and generation of audio deepfakes each year. Interspeech 2025 will be held from August 17 - 21, 2025. The paper submission deadline is **12 February 2025** giving us plenty of time past the course deadlines to finalize our draft for publication.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Frank, J.; and Schönherr, L. 2021. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. AST: Audio Spectrogram Transformer. In *Interspeech 2021*, 571–575.

Khalid, H.; Tariq, S.; and Woo, S. S. 2021. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. arXiv:2108.05080.

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.

Piczak, K. J. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, 1015–1018. New York, NY, USA: Association for Computing Machinery. ISBN 9781450334594.

Warden, P. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR).*

[1] www.interspeech2025.org

Deadline	**Task**
October 30	Gordon will read Qirui's paper on Frequency Masking/Representation and dive deeper into Deepfake generators as well. Min will preprocess ASVspoof 2021 instead of the 2019 version. And think of way to focus on background noise as well. Min should also put extensions: 2 directions 1. get more data in wild 2. deep dive into generators Qirui will look into what the Deepfake generators are? (run some?) Bill will be pretty busy but will try to load pretrained weights and finetune on ASVspoof William will investigate how audio datasets are generated. Then look into creating a new dataset with more bona fide audio using scraping.
November 6	Employ the Model and Train on the ESC-50 Dataset: Begin training the AST model on the ESC-50 dataset to verify its performance Preprocess ASVspoof data: Work on preprocessing the ASVspoof dataset to match the input size expected by the AST model. Start Generating the New Dataset: Initiate the process of scraping audio data from the internet to create an ”in-the-wild” dataset.
November 13	Train the Unchanged Model on ASVspoof: rain the AST model on the ASVspoof dataset without modifications to establish a baseline perfor- mance Work on Adding Background Information into Embeddings: Investigate methods to incorporate background sound information into the model's embeddings
November 20	Finish or start to wrap up testing and experiments Begin paper write up
November 27	Finalize paper write up

Table 1: Timeline