Whisper+AASIST for DeepFake Audio Detection

Qian Luo(✉) and Kalyani Vinayagam Sivasundari

The George Washington University, Washington, DC, USA
{qluo,kalyani.vinayagamsivasundari}@gwu.edu

Abstract. This study introduces a novel approach, combining the Whisper model with the AASIST architecture to enhance the detection performance of deepfake audio. Termed Whisper+AASIST, our investigation demonstrates its competitive edge over existing models on the 2021 ASVspoof DF subset. Notably, it surpasses these models in the challenging In-the-Wild datasets. This innovative fusion signifies a significant leap in deepfake audio detection, showcasing the effectiveness of synergistic model architectures. Through a comprehensive analysis, we unravel the potential of this hybrid approach in addressing evolving challenges within the domain of deceptive audio content. The findings underscore the significance of such inventive model combinations, providing a foundation for further advancements in deepfake audio detection methodologies.

Keywords: Deepfake Detection · AASIST · Whisper

1 Introduction

1.1 Motivation

The rapid advancements in generative artificial intelligence have significantly transformed text-to-speech (TTS) and voice conversion (VC) technologies. These cutting-edge technologies now enable the synthesis of speech so authentic that distinguishing it from real human vocalizations has become a formidable challenge. While these advancements undoubtedly offer increased convenience in various sectors, they also raise critical concerns about societal stability and security.

Recently, two notable incidents involving deepfake audios resulted in significant financial losses for corporate entities [2,15]. These cases gained extensive media attention and sparked widespread public debate. Concurrently, a report by the U.S. Department of Defense highlighted the escalating threats posed by advanced AI-generated content [20]. This report recommended two main countermeasures: implementing proactive authentication methods during content creation, and developing passive detection strategies for analyzing content post-production.

Considering the challenges in achieving broad implementation of authentication protocols in the near term, the importance of post-hoc detection techniques

A. Moallem (Ed.): HCII 2024, LNCS 14729, pp. 121–133, 2024.
https://doi.org/10.1007/978-3-031-61382-1_8

becomes increasingly evident. These methods are crucial for combating the rise in fraudulent activities, creating a continuous dynamic between the creators of generative AI and detection experts.

1.2 Research Question

– How effectively can the integration of OpenAI's Whisper [14] pretrained-transformer model enhance the deepfake audio detection capabilities of the Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks (AASIST) architecture [5] so users can recognize the fake audios?

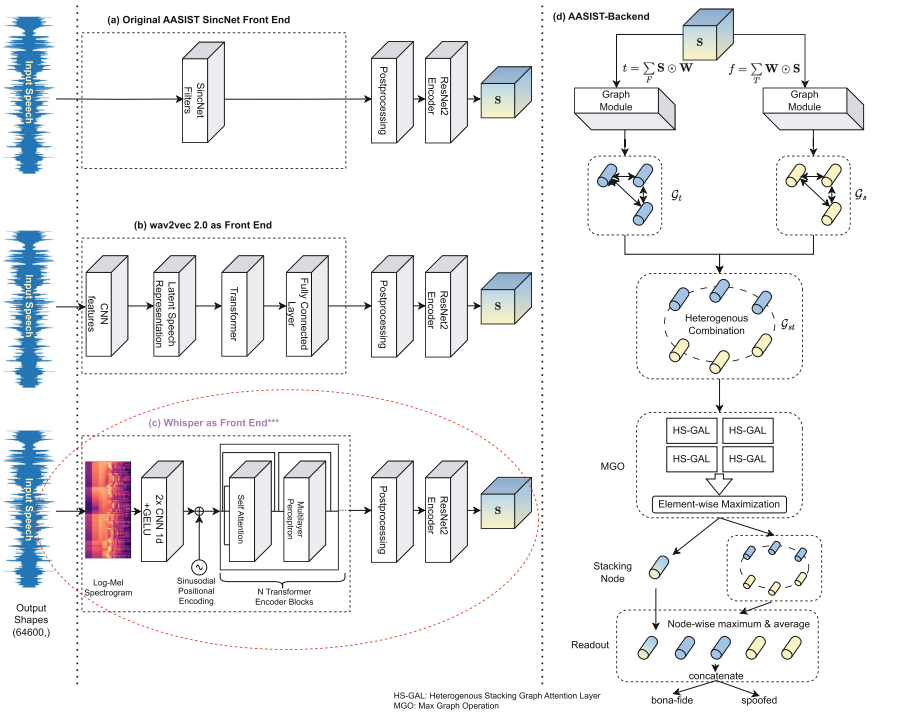

Fig. 1. The AASIST Architecture with Various Front Ends
Note: Panel (a) and (d) are adapted from June et al. [5], panel (b) is adapted from Baevski et al. [1] and Tak et al. [18], and panel (d) is adapted from Radford et al. [14].

2 Related Works

2.1 Overview

Yi et al. [24] conducted an extensive review on deepfake audio detection technologies, categorizing them into three primary approaches:

a) Traditional classification techniques, which include algorithms such as support vector machines [8] and Gaussian Mixture Models (GMM) [3].
b) Deep learning-based classifications, encompassing models like the Res2Net [9] (a modified residual network), graph neural network (GNN) based GAT [16], PC-DARTS [4] (a differentiable architecture search approach), and transformer-based Rawformer.
c) Comprehensive end-to-end architectures, with prominent examples being the graph attention network (GAN)-based AASIST [5] and transformer-based SE-Rawformer [10].

Historically, advancements in audio analysis relied heavily on hand-crafted or learnable features, primarily driven by traditional machine learning classifiers. With the advent of advanced end-to-end architectures like AASIST, there has been a paradigm shift towards integrating feature extraction and classification into a cohesive system.

Despite their reliance on more basic feature extraction techniques, traditional classifiers like GMM have maintained relevance due to their resilience, especially under the constraint of limited training samples. GMM continues to be a benchmark model in the field, with an Equal Error Rate (EER) of 25.25% using linear frequency cepstral coefficients (LFCC) features with ASVspoof 2021 DF data, outperforming most deep learning methods [24]. Conversely, deep learning classifiers have shown variable effectiveness, largely dependent on the specificity of their feature extraction methods.

In contrast, when leveraging XLS-R features extracted via wav2vec2.0 [1], deep learning models generally surpass the performance of GMM. For instance, the Attentive Filtering Network (AFN) utilizing XLS-R achieved an ERR of 14.15%, significantly lower than the GMM's ERR of 28.49% with the same features [24] with ASVspoof 2021 DF data.

End-to-end architectures like AASIST, while not outperforming the top deep learning methods with XLS-R, still demonstrate significant effectiveness with an ERR of 19.77% with ASVspoof 2021 DF data, positioning them competitively among deep learning models using XLS-R features. However, prior work did not focus on the interaction with end users.

2.2 State of the Art

The ASVspoof2021 DF subset [23] saw an initial 15.64% Equal Error Rate (EER) by T23 [11]. Tak et al. later reached a 2.85% EER with AASIST using wav2vec 2.0 (Fig. 1 Panel b) [18], surpassing Martín-Doñas and Álvarez's 4.98% EER [12]. Current SOTA EERs for In-the-Wild data are 24.73% for ASSERT+LPS and 34.81% for AASIST [24].

The field of deepfake audio detection, particularly in the context of the ASVspoof 2021 DF dataset, has witnessed remarkable advancements in recent years. This dataset, crucial for benchmarking, has enabled a clear evaluation of various models' capabilities in detecting fraudulent audio samples.

One of the earliest benchmarks on this dataset was set by T23, who achieved an EER of 15.64% [11]. This performance served as a significant indicator of the challenges inherent in accurately identifying deepfake audios. However, subsequent developments have dramatically improved detection capabilities.

A pivotal advancement was achieved by Tak et al., who utilized the AASIST architecture in conjunction with wav2vec 2.0 features. Their approach culminated in a substantially lower EER of 2.85%, as illustrated in Fig. 1 Panel b [18]. This marked a significant leap from the previous benchmarks, underscoring the effectiveness of integrating sophisticated neural network architectures with advanced feature extraction techniques.

In comparison, Martín-Doñas and Álvarez achieved a 4.98% EER [12], demonstrating the rapid progression of technological advancements in this domain. Their work, while being surpassed by Tak et al., contributed valuable insights into the optimization of deep learning methods for audio spoofing detection.

When considering 'In-the-Wild' data [13], a more challenging and variable dataset, the current state-of-the-art performances indicate a more complex scenario. ASSERT+LPS, a notable method in this category, achieved an EER of 24.73%, while AASIST recorded an EER of 34.81% [24]. These figures highlight the ongoing challenges faced when dealing with more diverse and less controlled audio samples, which more closely resemble real-world scenarios.

Overall, the continuous evolution in deepfake audio detection, as evidenced by these performances on the ASVspoof 2021 DF dataset, indicates both the progress made and the challenges that remain. As deepfake technologies grow more sophisticated, the need for advanced detection methods, capable of handling diverse and evolving threats, becomes increasingly crucial.

2.3 The AASIST Architecture

AASIST, introduced by Jung et al. [5], represents a significant advancement in end-to-end audio processing models. Its primary innovation lies in the integration of sinc convolution and a RawNet2-based encoder, coupled with sophisticated graph learning techniques. This unique combination allows AASIST to excel in both feature extraction and classification tasks, setting it apart from conventional architectures. The model's approach to handling waveform inputs, particularly its advanced representation processing components, has been pivotal in enhancing its overall performance. These developments have been instrumental in AASIST's achievements in various benchmarks, as detailed in Sect. 2.2.

2.4 Whisper

Whisper, developed by OpenAI [14], is a state-of-the-art transformer-based speech recognition model. It is distinguished by its extensive pretraining on an enormous dataset comprising 680,000 h of diverse audio samples. This extensive pretraining regime enables Whisper to offer robust performance across a wide

range of languages and accents, significantly reducing the necessity for task-specific fine-tuning. Its versatility and scalability make it an ideal candidate for integration into various speech-related applications, ranging from transcription to more complex tasks like speech synthesis and deepfake detection. Whisper's design embodies the cutting-edge advancements in the field of transformer-based models, harnessing the power of large-scale datasets to achieve unprecedented levels of accuracy and reliability in speech recognition.

3 Methods and Data

3.1 Methods

This study aims to enhance the AASIST architecture by incorporating elements of OpenAI's Whisper model [14] at the front end, as depicted in Fig. 1-(c). Initially, the AASIST framework employed SincNet filters on raw waveforms for feature extraction (Fig. 1 Panel a), which later evolved to utilize wav2vec 2.0 as the feature extractor (Fig. 1-(b)). Model performance will be evaluated using EER.

Whisper Model

The Whisper model, as outlined by OpenAI [14], is adept at processing raw waveform inputs $x_{1:L}$ to generate a sequence of spectro-temporal representations $o_{1:N}$, where L represents the number of samples in the waveform. As depicted in Fig. 1-(c), the Whisper-large-v2 variant, a specific configuration of the model, includes a series of carefully designed layers to optimize audio processing. Initially, a preprocessing unit transforms the raw waveform into a log-Mel spectrogram. This transformation is crucial for capturing the essential frequency and time characteristics of the audio input.

Following the preprocessing stage, the model incorporates two convolution layers. These layers are instrumental in enhancing the local features of the spectrogram, making the subsequent processing by the transformer encoder more effective. After these convolution layers, a position embedding layer is introduced to encode the temporal information of the audio sequence, a critical aspect for understanding the context in speech.

The core of the Whisper model comprises 32 transformer encoder layers. These layers are the backbone of the model, responsible for capturing complex patterns and relationships within the audio data. The transformer architecture, known for its effectiveness in handling sequential data, allows the Whisper model to process audio with remarkable accuracy and detail.

In our implementation, we omitted the decoder layers that are present in the original Whisper model. These layers are primarily used for speech recognition tasks, and since our focus is on utilizing the encoder-generated embeddings for spoofing detection, they were not required. The length of the spectro-temporal representation sequence, denoted as N, is set to 1500 for all Whisper model variants, determined by the position of the embedding layers.

Müller et al. [13] observed that the performance in spoofing detection is notably enhanced when the waveforms are not truncated to 4-s subsamples. In line with this finding, we chose to use a 30-s input sample for our experiments, which is consistent with the default setting of the Whisper model.

Given that Whisper was initially pretrained exclusively on bona fide (genuine) data [14], we hypothesized that its performance in spoofing detection could be further enhanced by fine-tuning it with a mix of bona fide and spoofed in-domain training data. Additionally, recognizing the limitations of our training dataset, which consists of samples from the ASVSpoof2019 LA subset, we implemented data augmentation techniques to account for potential deepfaking methods not present in the training data but likely encountered in the test data.

To address this, we introduced data augmentation for the training set, aiming to enhance the model's robustness and generalization capability. This strategy is in line with the findings presented in the referenced papers [5,17]. We experimented with three distinct variants of data augmentation:

No Data Augmentation (No DA): In this baseline approach, we trained the model using the original dataset without any augmentation. This setup serves as a control to gauge the effectiveness of the other augmentation techniques.

Coloured Additive Noise: Here, we introduced colored additive noise to the training data. This type of noise is known to simulate a variety of real-world acoustic environments, thereby preparing the model to handle diverse and challenging audio conditions.

Convolutive Noise + Impulsive Noise + Coloured Additive Noise: In this comprehensive approach, we combined convolutive noise, impulsive noise, and coloured additive noise. The inclusion of convolutive noise aims to replicate the effects of different transmission channels and acoustic environments, while impulsive noise mimics sudden, non-continuous disturbances. This combination of noises presents a more rigorous training scenario, aiming to significantly enhance the model's ability to generalize across various audio spoofing techniques.

To explore the potential of fine-tuning and data augmentation, we implemented two versions of the Whisper model: one using the original pretrained version (**Our Method 1**), and another that underwent additional fine-tuning with the mixed dataset (**Our Method 2**), incorporating the described data augmentation strategies. These implementations are designed to assess how each approach influences the model's proficiency in detecting audio spoofing in varied and potentially unseen scenarios.

Postprocessing

In the postprocessing stage, the spectro-temporal representations $o_{1:N}$ undergo a sophisticated transformation via a RawNet2-based encoder. This encoder is pivotal in extracting higher-level features from the input representations, which are essential for the subsequent stages of audio analysis.

The RawNet2 encoder, as described in the paper by Jung et al. [6], is an advanced deep neural network specifically designed to handle raw waveforms

for robust speaker verification. It is characterized by its ability to process raw audio data directly, bypassing the need for traditional handcrafted features. This approach allows for a more nuanced extraction of speaker-specific characteristics, directly from the waveform.

As the spectro-temporal representations pass through the RawNet2-based encoder, they are transformed into a more refined feature map $\mathbf{S} \in \mathbb{R}^{C \times F \times T}$. Here, C represents the number of channels, F denotes the number of spectral bins, and T stands for the number of time frames. Each of these dimensions plays a critical role:

Channels (C): This dimension captures various aspects of the audio signal, allowing the network to analyze different features in parallel.

Spectral Bins (F): The spectral bins represent the frequency components of the audio signal. By capturing a wide range of frequencies, the network can discern subtle nuances in the audio.

Time Frames (T): The time dimension ensures that the temporal dynamics of the speech signal are adequately represented. This is crucial for understanding the context and progression of spoken words or sounds.

The RawNet2-based encoder employs several layers, including gated convolutional layers and residual blocks, to effectively capture these dimensions. The gated convolutions control the flow of information through the network, allowing it to focus on the most relevant features. The residual blocks help in preserving the integrity of the input signal while enabling deeper layers in the network to learn complex patterns.

AASIST Backend

The feature map $\mathbf{S}$ derived from the RawNet2 encoder is fed into the AASIST backend, as depicted in Fig. 1-(d). This backend is intricately designed to further analyze the high-level features obtained from $\mathbf{S}$. The initial phase in the AASIST backend involves constructing two types of graphs: spectral ($\mathcal{G}_s \in \mathbb{R}^{N_s \times d_s}$) and temporal ($\mathcal{G}_t \in \mathbb{R}^{N_t \times d_t}$). In these expressions, N_s and N_t denote the number of nodes in the spectral and temporal graphs, respectively, while d_s and d_t represent the dimensionality of each node within these graphs.

Subsequent to their creation, these spectral and temporal graphs are fused to generate a unified spectro-temporal graph, denoted as $\mathcal{G}_{st}$. This fusion is a pivotal step in integrating the frequency and time-related information of the audio signal, essential for effective spoofing detection.

The combined graph $\mathcal{G}_{st}$ then undergoes advanced processing through a heterogeneous stacking graph attention mechanism. This technique is crucial for highlighting the most relevant features in the graph by weighting the connections between nodes based on their importance. Following this, a max graph operation is applied, which serves to further refine the feature representations by aggregating information from across the graph.

The final step in the AASIST backend is the application of a readout scheme. This step is responsible for interpreting the processed graph data and categorizing the input audio as either bona fide or spoofed. The readout scheme plays a

critical role in translating the complex graph-based representations into a final decision, leveraging the rich information encoded within the spectro-temporal graph.

Table 1. Results Comparison

Model	ASVspoof 2021 DF		In the Wild	
	EER%	Related Work	EER%	Related Work
GMM+LFCC	25.25%	a	37.49%	a
LCNN	25.26%	a	35.14%	a
ASSERT	21.58%	a	**24.73%**	a
Res2Net	19.47%	a	36.62%	a
RawNet2	20.55%	b	49.00%	
ASSIST	19.77%	a	34.81%	a
Wav2Vec 2.0 + ASSIST + SL + DA	**2.85%**		**10.49**%	
Our Method 1 (Base) + No DA	12.25%		36.79%	
Our Method 1 (Base) + Colored DA	12.18%		37.52%	
Our Method 1 (Base) + All DAs	12.51%		35.94%	
Our Method 1 (Large V2) + No DA	9.62%		**20.91%**	
Our Method 1 (Large V2) + Colored DA	**8.67%**		23.54%	
Our Method 1 (Large V2) + All DA	10.60%		24.81%	
Our Method 2 (Base) + Fine Tune + No DA	10.78%		25.98%	
Our Method 2 (Base) + Fine Tune + Colored DA	9.62%		25.11%	
Our Method 2 (Base) + Fine Tune + All DAs	10.68%		24.09%	

Note: a. Yi et al. [24], b. Liu et al. [11], c.Tak et al. [18], and d. Martín-Doñas and Álvarez et al. [12]. Those without related work are implemented by the authors.

3.2 Data

The model's training leveraged the ASVspoof 2019 LA subsets [19,22], which are part of the larger ASVspoof challenge, a benchmark for assessing the robustness of automatic speaker verification systems against spoofing attacks. The ASVspoof 2019 LA dataset, derived from the VCTK database [21], includes a diverse set of bona fide and spoofed speech samples. It features 2,580 bona fide and 22,800 spoofed speech samples from 20 speakers, providing a solid environment for training robust spoofing detection models.

For testing, the model was evaluated against two datasets: the ASVspoof 2021 DF dataset [23] and the In-the-Wild dataset [13]. The ASVspoof 2021 DF dataset, an extension of the ASVspoof 2019 database, includes additional challenges such as compression changes and deepfake samples, reflecting the evolving landscape of audio spoofing techniques. It comprises 22,617 bona fide and 589,212 spoofed samples, recorded by 48 speakers.

In contrast, the In-the-Wild dataset, amassed in 2022, offers a different perspective by featuring real-world audio clips sourced from various online platforms that have confused many social media users. This dataset emphasizes diversity

and real-world applicability by including clips of public figures and other widely circulated audio content. It consists of numerous samples, each representing a unique instance of real-world audio, thus providing a realistic testbed for evaluating the model's performance in practical scenarios. It comprised of 19,963 boni fide and 11,816 fake samples from 58 speakers.

The combination of these datasets presents a rigorous and comprehensive testing ground. The ASVspoof sets, with their controlled yet diverse spoofing techniques, offer a structured environment to assess the model's detection capabilities. Meanwhile, the In-the-Wild data introduces the complexity and variability of real-world scenarios, testing the model's generalizability and robustness against unseen and potentially more sophisticated spoofing attacks.

3.3 Implementation Details

In our setup, audio data were standardized to approximately 30 s, equivalent to 480,000 samples, through cropping or concatenation. This uniformity is crucial for maintaining consistency in input data.

For the training process, we utilized the Adam optimizer [7] with a fixed learning rate of 0.000001 to avoid overfitting with the pre-trained Whisper front-end. Considering the substantial computational requirements of the Whisper-large-v2 model, our experiments were conducted in two computational environments: one with 4 X NVidia V100 GPUs and another with 1 X NVidia A100 GPU. This approach allowed us to balance resource availability with the model's demands.

All models underwent training for 100 epochs, ensuring comprehensive learning and adaptation. To facilitate reproducibility in the research community, all source codes used in our experiments have been made available as open-source resources.

4 Results

Table 1 reveals a comprehensive comparison between our proposed methods and several established baselines and state-of-the-art models in the field. The evaluation is conducted on two datasets: ASVspoof 2021 DF and In the Wild, each posing distinct challenges for deepfake audio detection.

In the context of the ASVspoof 2021 DF dataset, traditional methods such as GMM+LFCC and LCNN exhibit EERs of 25.25% and 25.26%, respectively, while more advanced models like ASSERT and Res2Net achieve improved performance with EERs of 21.58% and 19.47%. Notably, our baseline method, denoted as Our Method 1 (Base), without any data augmentation, demonstrates competitive performance with an EER of 12.25%. The introduction of colored data augmentation (DA) slightly improves results, yielding an EER of 12.18%, while the combination of all augmentation techniques results in an EER of 12.51%. The Whisper large V2 model with colored data augmentation stands out as particularly promising with an EER of 8.67%.

In contrast, within the In-the-Wild dataset, the performance metrics vary. The GMM+LFCC baseline achieves an EER of 37.49%, while LCNN and ASSERT exhibit EERs of 35.14% and 24.73%, respectively. Our Method 1 (Base) demonstrates an EER of 36.79%, with colored data augmentation providing a marginal improvement to 37.52%. The combination of all data augmentation techniques leads to an EER of 35.94%. The Whisper large V2 model, again without data augmentation, performs exceptionally well with an EER of 20.91%.

Further improvements are observed with Our Method 2 (Base), incorporating fine-tuning and various data augmentation strategies. This model achieves competitive results with an EER of 10.78% for ASVspoof 2021 DF and 25.98% for In the Wild. Colored data augmentation and the combination of all augmentation techniques yield EERs of 9.62% and 10.68% for ASVspoof 2021 DF, and 25.11% and 24.09% for In the Wild, respectively.

Despite achieving competitive performance, our proposed method did not surpass the initial expectations, highlighting the complexity of deepfake audio detection and the need for further exploration and refinement in future research endeavors.

5 Discussion

Despite achieving competitive results, our efforts to optimize the Whisper model through fine-tuning encountered a significant challenge due to limitations in our existing GPU infrastructure. The NVidia V100 and A100 GPUs faced constraints in allocating sufficient memory for the larger variants, namely the Whisper-Large-V2 and even the Whisper-Medium model. Faced with this constraint, we explored alternative approaches while still aiming to enhance the model's performance. In response, we conducted a fine-tuning experiment using the less memory-intensive Whisper-base variant front-end in conjunction with AASIST. This specific fine-tuning process aimed to investigate the potential viability of the Whisper+AASIST architecture, serving as a proof of concept for a more tailored and resource-efficient approach within the given constraints.

The decision to fine-tune the Whisper-base variant with AASIST was motivated by the need to overcome memory limitations and optimize model performance. The results of this fine-tuning experiment, as reflected in Our Method 2, reveal promising outcomes. The Whisper model, with fine-tuning and different data augmentation strategies, achieved competitive Equal Error Rates (EERs) such as 10.78%, 9.62%, and 10.68% for various augmentation scenarios.

Although the integration of the Whisper+AASIST architecture led to an enhancement in the performance of the original AASIST architecture, the magnitude of this improvement, as observed in the results, falls short when compared to the impact observed with the incorporation of the Wav2Vec 2.0 architecture.

Upon closer examination of the architectural variances between the Whisper and Wav2Vec 2.0 encoders, it became evident that both models utilize the transformer architecture. However, a potential factor contributing to the lackluster performance of the Whisper model could be its reliance on log-mel spectrograms,

compared to the more versatile CNN filters used by Wav2Vec 2.0. This deviation in approach may influence the comparative effectiveness of the two architectures. Future research could explore whether including CNN as the first feature extractor, rather than the log-mel spectrogram used by Whisper, would lead to more competitive results.

Unfortunately, resource constraints and time limitations prevented the fine-tuning of the large-v2 variants of the Whisper model. This leaves open the possibility that a fine-tuned Whisper-largev2 model might yield substantial performance improvement. Regrettably, our attempts were hindered by a shortage of VRAM, affecting both NVidia 2 X A100 and/or 4 X V100. Despite these challenges, a more extensive exploration of the Whisper architecture could uncover its true potential.

It is noteworthy to mention that the current performance, while not reaching the desired level, is on par with the performance of the Wav2Vec 2.0 model prior to fine-tuning, as documented by [18]. This comparison highlights the nuanced nature of model performance and underscores the significance of considering various factors, such as architectural differences and fine-tuning opportunities, when evaluating and optimizing speech processing models. The ongoing evolution of the Whisper+AASIST architecture positions it as a potential avenue for further exploration and refinement in the dynamic landscape of deepfake audio detection.

6 Conclusion

While our attempts to incorporate the Whisper model into AASIST architecture did not pan out as we would like, this exercise has provided valuable insights into the capabilities and limitations of different speech processing architectures. Our research demonstrates that even with resource constraints, innovative approaches like fine-tuning less memory-intensive models can offer new avenues for performance enhancement. The comparative analysis between the Whisper and Wav2Vec 2.0 models highlights the importance of architectural choices and their impact on model efficiency and effectiveness.

The findings of our study suggest that while the Whisper model shows promise, its full potential may be unlocked only with the availability of more powerful computing resources. This limitation underscores the need for ongoing research and development in the field of speech processing, particularly in optimizing models to function efficiently within the constraints of available hardware.

In conclusion, our work contributes to the evolving narrative of speech processing technology, emphasizing the significance of architectural decisions, the balancing act between resource availability and model performance, and the continuous quest for optimization in a rapidly advancing field. Future research should focus on further exploring the capabilities of the Whisper model, particularly its large-v2 variant, and investigate alternative architectures and fine-tuning strategies that could offer a more resource-efficient path to enhanced performance in speech processing applications.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460 (2020)
2. Brewster, T.: Fraudsters cloned company director's voice in $35 million heist, police find (2021). https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions. Accessed 30 Sept 2023
3. De Leon, P.L., Stewart, B., Yamagishi, J.: Synthetic speech discrimination using pitch pattern statistics derived from image analysis. In: Interspeech, pp. 370–373 (2012)
4. Ge, W., Panariello, M., Patino, J., Todisco, M., Evans, N.: Partially-connected differentiable architecture search for deepfake and spoofing detection. arXiv preprint arXiv:2104.03123 (2021)
5. Jung, J., et al.: AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2022, pp. 6367–6371. IEEE (2022)
6. Jung, J., Kim, S., Shim, H., Kim, J., Yu, H.J.: Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms. arXiv preprint arXiv:2004.00526 (2020)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Klontz, J.C., Klare, B.F., Klum, S., Jain, A.K., Burge, M.J.: Open source biometric recognition. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2013)
9. Li, X., et al.: Replay and synthetic speech detection with Res2Net architecture. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2021, pp. 6354–6358. IEEE (2021)
10. Liu, X., Liu, M., Wang, L., Lee, K.A., Zhang, H., Dang, J.: Leveraging positional-related local-global dependency for synthetic speech detection. In: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2023, pp. 1–5. IEEE (2023)
11. Liu, X., et al.: ASVspoof 2021: towards spoofed and deepfake speech detection in the wild. IEEE/ACM Trans. Audio Speech Lang. Process. **31**, 2507–2522 (2023)
12. Martín-Doñas, J.M., Álvarez, A.: The vicomtech audio deepfake detection system based on Wav2vec2 for the 2022 add challenge. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2022, pp. 9241–9245. IEEE (2022)
13. Müller, N.M., Czempin, P., Dieckmann, F., Froghyar, A., Böttinger, K.: Does audio deepfake detection generalize? arXiv preprint arXiv:2203.16263 (2022)
14. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518. PMLR (2023)
15. Stupp, C.: Fraudsters use AI to mimic CEO's voice in unusual cybercrime case. Wall Street J. (2019). https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
16. Tak, H., Jung, J., Patino, J., Todisco, M., Evans, N.: Graph attention networks for anti-spoofing. arXiv preprint arXiv:2104.03654 (2021)

17. Tak, H., Kamble, M., Patino, J., Todisco, M., Evans, N.: RawBoost: a raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2022, pp. 6382–6386. IEEE (2022)
18. Tak, H., Todisco, M., Wang, X., Jung, J., Yamagishi, J., Evans, N.: Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation. arXiv preprint arXiv:2202.12233 (2022)
19. Todisco, M., et al.: ASVspoof 2019: future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441 (2019)
20. U.S. Department of Defense, Federal Bureau of Investigation, Cybersecurity and Infrastructure Security Agency: Contextualizing Deepfake Threats to Organizations (2023). https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF. Accessed 30 Sept 2023
21. Veaux, C., Yamagishi, J., MacDonald, K.: CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit [sound] (2017). https://doi.org/10.7488/ds/1994
22. Wang, X., et al.: ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech. Comput. Speech Lang. **64**, 101114 (2020)
23. Yamagishi, J., et al.: ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. arXiv preprint arXiv:2109.00537 (2021)
24. Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C.Y., Zhao, Y.: Audio deepfake detection: a survey. arXiv preprint arXiv:2308.14970 (2023)