

Dual-Feature Ensemble Deep Learning Architecture for Robust Spoofing Speech Detection

1st Haruto Namba
Tokyo Metropolitan University
Tokyo, Japan
namba-haruto@ed.tmu.ac.jp

2nd Sayaka Shiota
Tokyo Metropolitan University
Tokyo, Japan
sayaka@tmu.ac.jp

Abstract—In this paper, we propose a spoofing speech detection method using a deep learning model based on a dual-feature ensemble. The threat of spoofing speech attacks has increased with the advancements in speech synthesis and voice conversion technologies. This highlights the need for highly accurate spoofing speech detection methods. Our proposed method leverages information extracted from waveforms and mel spectrograms to enhance the robustness of spoofing speech detection. By combining a state-of-the-art spoofing detection technique with a high-performing model from sound-related tasks, we aim to develop a robust system for spoofing speech detection. Evaluation experiments using the ASVspoof 2021 dataset show that the proposed model outperforms conventional methods, including the wav2vec2.0 + AASIST baseline. Our experimental results suggest that the proposed method captures various characteristics of speech from multiple angles, thereby improving the generalization performance of the model.

Index Terms—Spoofing speech detection, Feature ensemble, Deep learning

I. INTRODUCTION

In recent years, the rapid advancement of speech synthesis technology has significantly expanded its applications. However, this progress has also led to growing concerns about the malicious use of synthetic speech for spoofing. Spoofed speech can be used to create fake videos by impersonating someone else's voice or bypassing voice-based biometric authentication systems [1]. Such misuse poses serious risks to personal information security and can have substantial social implications. Therefore, developing effective and reliable methods for spoofing speech detection has become increasingly important, particularly as the use of voice-based authentication systems becomes more widespread.

Various efforts have been made to develop countermeasures against spoofed speech, particularly through international challenges such as the ASVspoof challenge [2], which has driven significant advancements in this area. In the logical access (LA) scenario, where attacks involve synthetic speech, a range of approaches have been explored [3]. These include techniques that leverage diverse acoustic features and advanced model architectures to improve detection performance. However, there remain several challenges that hinder the deployment of robust spoofing detection systems in real-world environments. For example, existing models often struggle to maintain high performance when confronted with previously

unseen speech synthesis techniques. Moreover, they are sensitive to environmental factors such as background noise and channel variability, which can significantly degrade detection accuracy [4], [5]. These challenges highlight the need for more robust and generalizable models that can reliably detect spoofed speech across diverse scenarios.

To address these challenges, this paper proposes a deep learning-based approach for spoofing speech detection in the LA scenario. Our method aims to enhance detection performance by integrating complementary information from two different acoustic representations: waveforms and their corresponding mel spectrograms. Specifically, the proposed model combines two architectures: one that processes waveforms using the combination of wav2vec2.0 [6] and audio anti-spoofing using integrated spectro-temporal graph attention networks (AASIST) [7], and another that processes mel spectrograms using the audio spectrogram transformer (AST) [8]. The waveform-based architecture, which we refer to as w2v+AASIST, has demonstrated state-of-the-art performance in the ASVspoof2021 challenge [9], while AST has shown strong performance in sound classification tasks. By combining these two architectures, the proposed approach seeks to capture a wider range of acoustic features, leveraging the complementary strengths of each representation to improve overall detection robustness.

We evaluate the effectiveness of the proposed method on the ASVspoof2021 LA dataset [10]. Our experiments show that the proposed dual-feature ensemble model achieves an equal error rate (EER) of 0.817%, outperforming the baseline w2v+AASIST approach, which achieved an EER of 0.823%. Notably, the proposed method demonstrates improved performance in detecting spoofed speech generated using voice conversion techniques, which are particularly challenging for traditional approaches. These results indicate that integrating information from both waveforms and mel spectrograms can enhance the robustness of spoofing speech detection systems.

II. RELATED WORK

A. Spoofing Speech Detection

Spoofing speech detection is the task of distinguishing between bona fide speech spoken by humans and spoofed speech generated by methods such as speech synthesis, voice conversion, or playback attacks. Since spoofed speech can

be maliciously used to attack voice authentication systems or voice assistants, accurate detection of such spoofed speech is critically important. In this work, we focus on the detection of spoofed speech generated through speech synthesis and voice conversion, which is known as the LA scenario. Spoofing speech detection has been extensively studied through challenges such as ASVspoof [2]–[5], and various approaches have been proposed in the literature [10]–[13].

B. Previous Work on Spoofing Speech Detection

Among previous works, the approach that achieved the highest detection performance in the ASVspoof2021 LA task was a deep learning model that combined wav2vec2.0, which is a self-supervised pre-trained model used as a frontend, with AASIST [9]. AASIST converts the input waveform into feature maps via an encoder and then processes these features using spectral and temporal graph neural networks. The combined wav2vec2.0 and AASIST model extracts features directly from the waveforms and uses them to classify spoofed speech. However, since wav2vec2.0 is pre-trained on data intended for different tasks, and only fine-tuned on spoofing detection data, there may be limitations in its ability to extract features that are critical for spoofing speech detection.

C. Audio Spectrogram Transformer

The AST [8] was originally proposed for tasks such as environmental sound classification and acoustic event detection. AST takes mel spectrograms as input and is based on the vision transformer (ViT) [14] architecture. ViT was designed for image data, dividing inputs into fixed-size patches and treating each patch as a token that is processed by transformer blocks. AST applies this approach to mel spectrograms, converting them into a form suitable for transformer-based processing. While AST was originally developed for sound classification tasks, it has also been effectively applied to various other tasks, such as acoustic event detection and speaker recognition [15], [16].

III. PROPOSED METHOD

A. Overview

We propose an approach that extends the state-of-the-art spoofing speech detection model, w2v+AASIST, by further integrating the AST. The w2v+AASIST has demonstrated high performance for spoofing speech detection tasks, as described in Section 2.2. However, it primarily relies on information extracted from waveforms, which may not fully capture the diverse characteristics of spoofed speech that are critical for robust detection. To address this, our approach incorporates features extracted from mel spectrograms in addition to the waveform-based features. This integration leverages complementary information from different input representations. Recent studies have highlighted the effectiveness of Transformer-based models in spoofing speech detection [17], [18]. However, there have been few studies that apply AST, which has shown strong performance in environmental sound classification and other audio-related tasks, to the domain

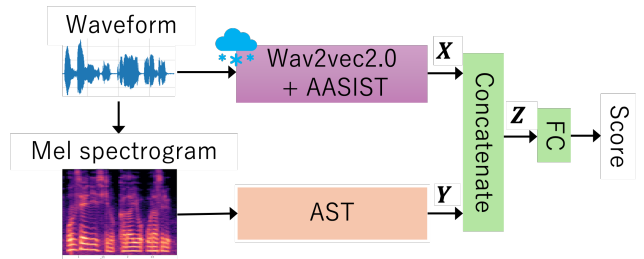


Fig. 1. Overview of the proposed method. X : Feature vector from wav2vec2.0+AASIST, Y : Feature vector from AST, Z : Concatenated feature vector, FC: Fully-connected layer.

of spoofing speech detection. As noted in Section 2.3, AST has achieved promising results in various audio event detection tasks, suggesting its potential for spoofing detection if appropriately adapted. Therefore, we hypothesize that the inclusion of AST as an additional feature extractor will enable the model to learn more diverse and informative representations, leading to improved detection accuracy. In summary, our approach combines the waveform-based features captured by w2v+AASIST with the frequency-domain features extracted by AST, forming a unified model that can capture complementary aspects of spoofed speech. This multi-faceted feature extraction is expected to enhance the robustness and generalization ability of the model, making it more effective at detecting spoofing attacks across various scenarios and conditions.

B. Model Architecture

The architecture of the proposed model is shown in Figure 1. As described earlier, the proposed method integrates the w2v+AASIST with AST. For the w2v+AASIST, we adopt the same architecture as described in [9]. In this setup, the AASIST module processes the features extracted by wav2vec2.0 and generates a feature vector, denoted as X . Specifically, we use the output from the layer immediately before the final classification layer in AASIST, as this representation retains rich and informative features relevant for spoofing detection. For the AST component, we adopt the architecture proposed in [8], where AST processes mel spectrogram inputs and outputs a feature vector corresponding to the class token, denoted as Y . The class token in the AST architecture encapsulates global information about the input spectrogram, providing a holistic representation of the audio characteristics. The final step in our proposed architecture is the concatenation of the feature vectors X and Y along the feature dimension, resulting in a combined feature vector $Z = [X, Y]$. This concatenated vector is then fed into a fully connected (FC) layer, followed by a hidden layer, and finally passed to an output layer that predicts the probability of the input being bona fide speech or spoofed speech. By combining the feature representations obtained from both AASIST and AST, our model can learn to weigh and utilize the most informative aspects of each representation. This enables the model to develop a more nuanced understanding of spoofed speech, thereby improving

its generalization performance and robustness against various spoofing techniques.

IV. EXPERIMENTS

A. Database

For the training and validation of our proposed model, we utilized the ASVspoof2019 database for the LA task, and the evaluation was conducted using the ASVspoof2021 LA task test data. The datasets contain speech samples generated using various synthetic methods, including text-to-speech (TTS), voice conversion (VC), and hybrid techniques that combine these approaches. The test data are labeled to indicate the type of synthesis used for each utterance. All speech signals were sampled at 16 kHz. To improve the generalization capability of the model during training, we employed noise augmentation. Specifically, we followed the procedure described in [9], using the RawBoost toolkit to introduce noise perturbations to the training data. This noise augmentation simulates more realistic environmental conditions, helping to enhance the robustness of the model against unseen noise and distortions.

B. Experimental Conditions

The input segment length for the proposed model was set to 64,600 samples, corresponding to approximately 4 seconds of speech. For mel spectrogram extraction, we used a frame length of 25ms, a frame shift of 10ms, and an FFT size of 512. The number of mel-frequency bins was set to 128. When inputting the mel spectrogram into the AST model, the patch size was set to 16, with stride values of 10 for both the time and frequency axes. To limit computational cost, the maximum number of frames along the time axis was capped at 1024. The model was trained for 100 epochs using the Adam optimizer, with a batch size of 8. We selected the model achieving the best performance on the validation data across the training epochs for the final evaluation.

The proposed model combines w2v+AASIST, which have been proposed in previous research for spoofing detection, and an AST-based classification model that uses mel-spectrograms as input. The pre-trained w2v+AASIST model was used without updating the parameters [19], while the AST parameters were fine-tuned from the pre-trained AST model [20], as fine-tuning the w2v+AASIST model led to a degradation in performance. To explore the impact of feature dimension matching, we conducted two variations of the proposed model. In the first configuration (no length adjustment), the feature vectors from AASIST (\mathbf{X} , 160 dimensions) and AST (\mathbf{Y} , 768 dimensions) were directly concatenated to form a combined feature vector \mathbf{Z} with 928 dimensions. In the second configuration (with length adjustment), the AST feature vector \mathbf{Y} was passed through a FC layer to reduce its dimension to 160. This was then concatenated with \mathbf{X} to create \mathbf{Z} with 320 dimensions. We also experimented with different numbers of hidden layers in the fully connected layers that process \mathbf{Z} . Specifically, we evaluated configurations with either a single hidden layer (250 units) or two hidden layers (250 and 50 units, respectively). The final output layer produces a two-class classification

TABLE I
EER (%) ON THE ASV SPOOF 2021 EVALUATION DATASET.
T: TTS, V: VC, H: HYBRID, HL: HIDDEN LAYER.

	Attack	Baseline w2v+	Proposed			
			Without Adjustment		With Adjustment	
	method	AASIST	1HL	2HLs	1HL	2HLs
T	A07	0.303	0.303	0.303	0.303	0.303
	A08	0.830	0.818	0.822	0.818	0.782
	A09	0.205	0.205	0.205	0.205	0.216
	A10	0.522	0.522	0.513	0.525	0.485
	A11	0.430	0.419	0.419	0.419	0.410
	A12	0.419	0.411	0.411	0.411	0.382
H	A13	0.203	0.203	0.203	0.203	0.195
	A14	0.303	0.303	0.303	0.303	0.295
	A15	0.360	0.360	0.360	0.360	0.332
T	A16	0.377	0.357	0.357	0.357	0.357
V	A17	0.998	0.966	0.966	0.966	1.133
	A18	2.672	2.612	2.612	2.612	2.740
	A19	1.156	1.127	1.127	1.116	1.079
	Pooled	0.823	0.817	0.823	0.817	0.877

score indicating whether the input is a bona fide or spoofed utterance. The final output score for the binary classification (bona fide vs. spoofed speech) is input to a cross-entropy loss function during training. The evaluation metric was equal error rate (EER), when the threshold is adjusted so that the false acceptance rate and false rejection rate of spoofed speech are equal.

C. Experimental Results

Table I shows the EERs obtained on the ASVspoof2021 evaluation set. To compare the performance of the proposed ensemble method with a standalone system, the EERs of the w2v+AASIST single system are also shown as a baseline. The evaluation set consists of spoofed speech generated by different synthesis methods, labelled A07 to A19, and the EERs are reported for each method and the pooled result across all methods. A07 to A12 and A16 were generated using TTS synthesis; A13 to A15 are the hybrid systems of VC and TTS; A17 to A19 were generated using VC [13]. From Table I, it can be seen that two configurations of the proposed methods, namely without length adjustment and with one hidden layer, and with length adjustment and with one hidden layer, both achieved an EER of 0.817, which is lower than the EER of the baseline. However, the proposed method without length adjustment and with two hidden layers achieved the same EER as the baseline. The proposed method with length adjustment and with two hidden layers performed worse, yielding a pooled EER of 0.877, indicating that increasing model complexity does not always translate to improved performance. The contrasting performance between the one-layer and two-layer configurations under length adjustment can be attributed to the information compression inherent in the length adjustment process. The length adjustment compresses the dimensionality of the AST output, which reduces the information content available for spoofing detection. With a single hidden layer, the model effectively leverages this compressed feature rep-

resentation, as evidenced by the improved EER compared to the baseline. However, when a second hidden layer is added, the model may struggle to extract meaningful patterns from the already compressed features. This additional layer could introduce unnecessary complexity, leading to overfitting on the limited information content, or the reduced feature dimensionality may be insufficient to support the increased model capacity. In contrast, for the model without length adjustment, incorporating a second hidden layer did not negatively impact performance, suggesting that the larger feature dimension might provide sufficient capacity to benefit from increased model complexity.

Analyzing the spoofing method-wise performance, we observe that the proposed methods without length adjustment (one or two hidden layers) and with length adjustment (one hidden layer) outperform the baseline for many TTS-based attacks (A07–A12 and A16). Similarly, for VC-generated speech (A17–19), the proposed methods showed some improvements. This indicates that incorporating AST-derived features from mel spectrograms complements the waveform-based representations of w2v+AASIST, effectively capturing spoofing artifacts that are otherwise overlooked, although its effectiveness can vary depending on the specific VC technique. Although not presented in the table, experiments using AST alone yielded a high EER of 7.63, indicating its limited effectiveness as a standalone spoofing detection system. Overall, these findings demonstrate that the proposed combination of waveforms and mel spectrogram-based feature extractors enhances the ability of the model to generalize across a wide variety of spoofing attacks. This suggests that incorporating AST-derived features from mel spectrograms complements the waveform-based representations of w2v+AASIST, effectively capturing spoofing artifacts that are otherwise overlooked.

V. CONCLUSION

In this paper, a dual-feature ensemble model for spoofing speech detection has been proposed, which combines the waveform input model w2v+AASIST and the mel spectrograms input model AST. From the experimental results, it was demonstrated that the dual features extracted from the waveforms and mel spectrograms effectively complemented each other for spoofing speech detection. As future work, we will explore more effective combinations with models that use different features such as CQT, LFCC, and MFCC, as well as improving the structure of the proposed model.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI (Grant Number JP24K14993), SCAT, and ROIS DS-JOINT (026RP2025) to S. Shiota

REFERENCES

- [1] Korshunov, Pavel et al., “Vulnerability assessment and detection of deepfake videos,” *Proc. international conference on biometrics (ICB)*, pp. 1–6, 2019.
- [2] Wu, Zhizheng et al., “ASVspoof: The automatic speaker verification spoofing and countermeasures challenge,” *Trans. on IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.
- [3] Liu, Xuechen et al., “ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild,” *Trans. on IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.
- [4] Yamagishi, Junichi et al., “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection,” *Proc. 2021 Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVspoof 2021 Workshop)*, 2021.
- [5] Massimiliano, Todisco et al., “ASVspoof 2019: Future horizons in spoofed and fake audio Detection,” *Proc. Interspeech*, pp. 1008–1012, 2019.
- [6] Baevski, Alexei et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Proc. Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.
- [7] Jung, Jee-weon et al., “AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” *Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371, 2022.
- [8] Yuan, Gong et al., “AST: Audio spectrogram transformer,” *Proc. Interspeech 2021*, pp. 571–575, 2021.
- [9] Hemlata, Tak et al., “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, pp. 112–119, 2022.
- [10] Delgado, Héctor et al., “ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan,” *arXiv preprint arXiv:2109.00535*, 2021.
- [11] Wu, Zhizheng et al., “Spoofing and countermeasures for speaker verification: A survey,” *Trans. on speech communication*, vol. 66, pp. 130–153, 2015.
- [12] Li, Menglu et al., “A survey on speech deepfake detection,” *Trans. on ACM Comput. Surv.*, 2025 (Accepted).
- [13] Xin, Wang et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Trans. on Comput. Speech Lang.*, vol. 64, p. 101114, 2019.
- [14] Alexey, Dosovitskiy et al., “An image is worth 16x16 words: transformers for image recognition at scale,” *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [15] Li, Kang et al., “AST-SED: An effective sound event detection method based on audio spectrogram transformer,” *Proc. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.
- [16] Ahmed, Sara et al., “ASiT: Local-global audio spectrogram vision transformer for event classification,” *Trans. on IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3684–3693, 2024.
- [17] Yinlin, Guo et al., “Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12702–12706, 2023.
- [18] Khan, Awais et al., “SpotNet: A spoofing-aware transformer network for effective synthetic speech detection,” *Proc. the 2nd ACM International Workshop on Multimedia AI against Disinformation (MAID)*, pp. 10–18, 2023.
- [19] https://github.com/takhemlata/ssl_anti-spoofing.
- [20] <https://huggingface.co/MIT/ast-finetuned-audioset-10-10-0.4593>.