ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447923

AUDIO DEEPFAKE DETECTION WITH SELF-SUPERVISED WAVLM AND MULTI-FUSION ATTENTIVE CLASSIFIER

Yinlin Guo, Haofan Huang, Xi Chen, He Zhao, Yuehai Wang

Zhejiang University
Department of Information and Electronic Engineering
HangZhou, China

ABSTRACT

With the rapid development of speech synthesis and voice conversion technologies, Audio Deepfake has become a serious threat to the Automatic Speaker Verification (ASV) system. Numerous countermeasures are proposed to detect this type of attack. In this paper, we report our efforts to combine the self-supervised WavLM model and Multi-Fusion Attentive classifier for audio deepfake detection. Our method exploits the WavLM model to extract features that are more conducive to spoofing detection for the first time. Then, we propose a novel Multi-Fusion Attentive (MFA) classifier based on the Attentive Statistics Pooling (ASP) layer. The MFA captures the complementary information of audio features at both time and layer levels. Experiments demonstrate that our methods achieve state-of-the-art results on the ASVspoof 2021 DF set and provide competitive results on the ASVspoof 2019 and 2021 LA set.

***Index Terms*—** Audio Deepfake Detection, Speech Self-Supervised Model, WavLM, Attention

1. INTRODUCTION

Automatic Speaker Verification (ASV) system is an important biometric solution widely used in identity authentication applications such as access control systems, telephone banking, and forensic scenarios [1]. It operates by verifying the claimed speaker's identity through specific features derived from speech signals. However, with the development of speech synthesis and voice conversion technologies, a large number of spoofed audio samples have emerged that are difficult to distinguish from real samples [2]. Such samples can easily deceive both humans and ASV systems, posing significant challenges for speech anti-spoofing research.

With the improving interest in developing robust spoofing countermeasures (CMs), self-supervised features have gained increasing attention in recent research. Speech self-supervised models are capable of leveraging large amounts of unlabeled data to extract representations of speech signals. Many researchers have already utilized these models as front-end feature extractors for audio deepfake detection. Xie et al. [3] propose the utilization of Wav2vec2 model [4] to train an embedding siamese neural network to protect anti-spoofing models from attacks. A similar method [5] uses the pre-trained Wav2vec2 [4] and a downstream classifier to detect spoofed audio. Wang et al. [6] explore different pre-trained self-supervised speech models, such as Wav2vec2 and XLS-R [7], as the front-end of spoofing CMs. They find that a self-supervised front-end pre-trained using diverse speech data performed quite well. Recently, Tak et al. [8] also achieve significantly improved performance in the filed of spoofing detection by applying the XLS-R model.

However, features from current speech self-supervised models are inadequate for multi-speaker tasks [9], including audio deepfake detection. The models are usually trained with masking prediction pretext task. During the pre-training stage, a certain percentage of time steps are masked in the latent feature encoder space. The model is learned to identify the quantized latent audio representation for each masked frame. Despite their excellent performance in tasks such as phoneme classification and automatic speech recognition, their effectiveness is constrained when it comes to certain speaker-related tasks. Features retrieved by these models contain the content information of the audio samples, but neglect speaker-related attributes which are more compatible with audio deepfake detection due to the presence of speaker-related artefacts in spoofing speech [1].

In this paper, we draw inspiration from recent advancements in speech self-supervised model, WavLM [9] and introduce the usage of WavLM as a front-end feature extractor for the first time. Attributed to the masked speech prediction and denoising training, WavLM has the potential to learn non-ASR features, including complex acoustic environments and speaker-related information and had demonstrated improved performance on certain non-ASR tasks. Additionally, we propose the Multi-Fusion Attentive (MFA) classifier based on the attentive statistics pooling layer [10]. The MFA aggregates the output representations of WavLM to focus on features at different layers and time steps, thus facilitating the extraction of highly discriminative features.

This paper is organized as follows. Sec.2 elaborates on the

proposed detection system. Sec.3 provides a detailed description of the implementation and training details. The results are discussed in Sec.4. Finally, in Sec.5, we summarize our methods and draw conclusions.

2. PROPOSED METHOD

The overall framework of the model is illustrated in Fig.1. The input raw waveforms are fed into the WavLM model to obtain layers of frames of feature embedding. All the embeddings are then passed through the Multi-Fusion Attentive (MFA) classifier to get the final prediction. This section provides a detailed description of these modules.

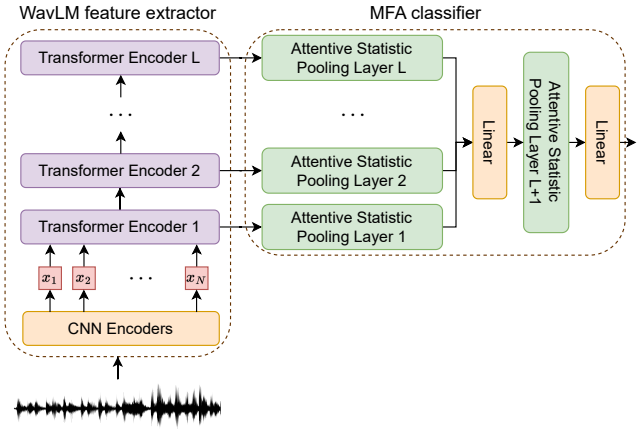

Fig. 1. Pipeline of our proposed model. Left part: the WavLM model; Right part: the MFA Classifier.

2.1. WavLM Model

WavLM is a speech self-supervised model that employs Wav2vec2 as its backbone. It consists of a convolutional feature encoder and Transformer encoders. The convolutional feature encoder converts the raw waveform into a feature sequence $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ of length N, where N is the number of frames. Transformer encoders are composed of several Transformer layers, where the input for the first layer is the output features retrieved by the CNN encoder, and the input for each subsequent layer is derived from the preceding layer. The output representation of the l-th Transformer layer be denoted by $\mathbf{H}^l = \{\mathbf{h}_1^l, \mathbf{h}_2^l, \cdots, \mathbf{h}_N^l\}$. Let $\mathcal{H} = \{\mathbf{H}^l\}_{l=1}^L$ be the set of output representations from all layers, where L is the number of Transformer Encoder layers.

The WavLM utilizes a masked speech denoising and prediction framework, where noise (or overlapping) is added to the input audio before masking. For the masked frames, the model is trained to predict pseudo-labels. The masked speech denoising allows WavLM to learn non-ASR features related to speaker characteristics and acoustic environments, making it more suitable for our audio detection tasks, since fake speech often comprises many speaker-related artefacts.

2.2. ASP-based MFA

2.2.1. *Attentive Statistics Pooling*

Attentive Statistics Pooling (ASP) [10] was originally used for extracting speaker embeddings and achieved excellent performance in speaker verification tasks [11]. ASP is a pooling method that combines the advantages of both attention and statistic pooling. Higher-order statistics can enhance speaker discriminability, and thus facilitate the detection of speaker-related artefacts in spoofed samples.

Given a feature sequence $\mathbf{z}_t, \ t = 1, \cdots, T$, ASP calculates the weighted mean μ and standard deviation σ of the sequence as follows:

$$\begin{aligned} \mu &= \sum_{t=1}^{T} \alpha_t \mathbf{z}_t, \\ \sigma &= \sqrt{\sum_{t}^{T} \alpha_t \mathbf{z}_t \odot \mathbf{z}_t - \mu \odot \mu}, \end{aligned} \tag{1}$$

where α_t is the attention weight of the t-th frame, and $\odot$ denotes the element-wise product. Attention weights are computed for each frame through linear layers that output a scalar, which is then normalized using a softmax function:

$$\begin{aligned} e_t &= \boldsymbol{v}^T f\left(\boldsymbol{W}\mathbf{z}_t + \boldsymbol{b}\right) + k, \\ \alpha_t &= \frac{\exp\left(e_t\right)}{\sum_{\tau}^{T} \exp\left(e_\tau\right)}, \end{aligned} \tag{2}$$

where $\boldsymbol{v}$, $\boldsymbol{W}$, and $\boldsymbol{b}$ are learnable parameters of the modules.

2.2.2. *MFA Classifier*

Previous work [5, 12, 13] has demonstrated that the intermediate feature of self-supervised model contains certain discriminative information. According to [14], the authors have also indicated that by weighting the representations of different layers, self-supervised model has shown great potential in various speech downstream tasks. To learn the optimal weight configuration, we propose the Multi-Fusion Attentive classifier based on the ASP layer. The MFA classifier consists of the stacked time-wise ASP (T-ASP) layers and a single layer-wise ASP (L-ASP) layer. The T-ASP layer is used to extract the time-level features from the hidden representations of different transformer layer. Each T-ASP computes a concatenated representation of the mean and standard deviation of the input sequence to form a single vector. Finally, the L-ASP layer incorporates all the layer-level features to generate the output representation.

More specifically, given the output representations $\mathcal{H}$ from all L layers of WavLM Transformer encoders, there are L independent ASP layers that act on the time channel of the representations, computing the statistical pooling mean and variance via:

$$\begin{aligned}\mathbf{r}^l &= \text{concatenate}(\mu^l, \sigma^l) \\ &= \text{T-ASP}_l(\mathbf{H}^l), \quad l = 1, \cdots, L,\end{aligned} \tag{3}$$

Where the T-ASP_l denotes the T-ASP layer corresponding to the l-th layer of WavLM. All the layer-level statistical pooling representations are then stacked to form a new channel, where an additional L-ASP layer is applied to perform attention calculation as

$$\mathbf{o} = \text{L-ASP}(\text{concatenate}(\mathbf{r}^1, \mathbf{r}^2, \cdots, \mathbf{r}^L)). \tag{4}$$

Output representation $\mathbf{o}$ is then fed into fully connected layers to obtain the final prediction.

3. EXPERIMENTS

3.1. Datasets and metrics

We focused on the logical access (LA) and speech deepfake (DF) partitions of the ASVspoof 2021 challenge [15]. With no new training or development data being released, the ASVspoof 2021 challenge requires the use of the training and development partitions of the ASVspoof 2019 databases [16]. So we use ASVspoof 2019 LA train and dev sets for training, and evaluate our approach on the ASVspoof 2019 LA, ASVspoof 2021 LA, and ASVspoof 2021 DF evaluation set.

The 2021 LA and DF sets are similar to the 2019 LA data, but are intentionally more challenging: the LA evaluation data contains new trials for each speaker and both encoding and transmission artefacts, and the DF evaluation data exhibits audio coding and compression artefacts.

We evaluate our model with two metrics: the minimum normalized tandem detection cost function (min t-DCF) [17] and equal error rate (EER). The min t-DCF assesses the combined (tandem) performance of CMs and ASV whereas the EER reflects the standalone spoofing detection performance.

3.2. Implementation Details

In the pre-processing stage, the audio samples are pre-emphasized with a coefficient of 0.97 and then truncated or concatenated to a fixed length of approximately 4 seconds (64600 sample points). We do not apply voice activity detection or any normalization to the audio samples.

Models are trained using Adam optimizer with $\beta = [0.9, 0.999]$. We employ a step learning-rate decay scheduler to accelerate convergence. For the frontend WavLM fine-tuning, we set the initial learning rate at 3×10^{-6}, which decays every 6000 steps with a decay factor, or gamma, of 0.1. The batch size is 4. When the WavLM is fixed, the batch size is increased to 32, the learning rate is set to 0.003, and the step size and gamma are 3200 and 0.5, respectively.

The entire experiments are conducted on four NVIDIA GeForce RTX 3090 GPUs. For each configuration, the model is trained for about 16,000 steps. The WavLM model and MFA classifier are trained jointly. For the WavLM model, we use the pre-trained weights provided by the authors official repository [1] as the initialization parameters. We use two different WavLM models: WavLM Base and WavLM Large. The dimension of the hidden representation of the MFA classifier is set to match the dimension of the output representation, and the number of T-ASP layers is equal to the number of transformer layers in WavLM.

4. RESULTS AND ANALYSIS

4.1. Results on ASVspoof 2021 LA and DF evaluation set

Table 1 presents a comparison of our results and those of other models that use self-supervised front-ends on the ASVspoof 2021 LA and DF evaluation dataset. As is shown, our approach with WavLM and MFA achieves the best performance on the DF set and a competitive result on the LA set. To the best of our knowledge, it is the lowest reported EER on the DF evaluation set.

We explain the reasons behind our superior performance. On the one hand, most existing approaches using self-supervised models adopt Wav2vec2 or its variations as the pre-trained front-end feature extractor. However, since these models are trained by masked prediction, the contextualized representations from the model contain more information about the audio content than the speaker information. Instead, the WavLM is pre-trained by masking speech denoising and prediction and is capable of learning more speaker-related information and complex acoustic environments, such as speaker characteristics and diverse audio backgrounds. Such information is particularly helpful for audio deepfake detection, as most spoofed speech typically contains speaker-related artefacts. On the other hand, compared with [6] using a complex back-end model based on heterogeneous graph neural networks, our proposed ASP-based MFA classifier is essentially the fully connected layers but can attend features of WavLM at different time and layer levels, leveraging the complementary information of audio features.

4.2. Results on ASVspoof 2019 LA evaluation set

We further test our model on the ASVspoof 2019 LA evaluation set, results are shown in Table 2. Our approach demonstrates a pooled min t-DCF of 0.0126 and an EER of 0.42%.

[1] https://github.com/microsoft/unilm/tree/master/wavlm

Table 1. Comparative Pooled EER (%) results of our proposed method with other anti-spoofing systems based on the self-supervised model in the ASVspoof 2021 LA and DF evaluation set.

System	Front-end	Pooled EER(%)	
		DF	LA
Wang et al. [6]	Wav2vec2-XLSR	5.44	7.18
Doñas et al. [5]	Wav2vec2-XLS128	4.98	3.54
Wang et al. [6]	Wav2vec2-XLSR	4.75	6.53
Tak et al. [8]	Wav2vec2	2.85	**0.82**
Ours	WavLM-Large	**2.56**	5.08

Results show that our model achieve better performance compared with recent anti-spoofing systems, demonstrating the effectiveness and superiority of our proposed method.

Table 2. Comparison with other anti-spoofing systems in the ASVspoof 2019 LA evaluation set, reported in terms of pooled min t-DCF and EER(%).

System	min t-DCF	EER(%)
Hua et al. [18]	0.0481	1.64
Yang et al. [19]	0.0360	1.21
Zhang et al. [20]	0.0368	1.14
Tak et al. [21]	0.0335	1.06
Jung et al. [22]	0.0275	0.83
Huang et al. [23]	0.0176	0.52
Ours	0.0126	0.42

4.3. Ablation Study

Table 3 describes the results of our ablation experiment on each component of the modified architecture. The performance deteriorates significantly with Wav2vec2. It is also worth noting that when using WavLM-L as the front-end, our approach achieves an EER of 5.79% even without fine-tuning. This is a very competitive result compared with fine-tuned Wav2vec2. The features of WavLM exhibit substantial resilience to variations in speech within the DF dataset, resulting in a significant reduction in the EER.

As for the back-end classifier, results show that the ASP-based MFA classifier is beneficial. The GAP classifier only utilizes final layer representations of WavLM, thereby exhibiting a constrained capacity to capture the self-supervised representations. And the TN + FCs classifier applies simple time-wise normalization, which is less effective for higher speaker discriminability. The two classifiers resulted in a relative degradation in performance of 136.6% and 15.0%, respectively.

Table 3. The ablation study intends to demonstrate the effectiveness of each component of the system. *ft*: fine tuning. *-L* and *-S*: large and small. *ASP*: attentive statistics pooling. *GAP*: global average pooling. *TN*: temporal normalization. *FCs*: fully connected layers.

Ablation	Configuration	Pooled EER
WavLM-L(ft) + MFA	-	**2.56**
w/o WavLM-L	Wav2vec2-S	11.68
	Wav2vec2-L	10.03
	WavLM-S	10.06
w/o ft	WavLM-S	10.81
	WavLM-L	5.79
w/o MFA	GAP	6.01
	TN + FCs	2.92

5. CONCLUSION

In this paper, we focus on audio deepfake detection based on the speech self-supervised model. We employ the self-supervised WavLM as the front-end feature extractor for the first time and propose a novel ASP-based Multi-Fusion Attentive classifier. The multi-layer representations of WavLM are aggregated by the time-wise and layer-wise ASP, which can effectively capture the complementary information of audio features. By jointly training the WavLM and MFA, we show that the proposed methods show competitive results on both the ASVspoof 2019 and ASVspoof 2021 datasets.

6. REFERENCES

[1] Z. K. Anjum and R. K. Swamy, "Spoofing and countermeasures for speaker verification: A review," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017, pp. 467–471.

[2] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Voice spoofing countermeasures: Taxonomy, state-of-the-art, experimental analysis of generalizability, open challenges, and the way forward," *ArXiv*, vol. abs/2210.00417, 2022.

[3] Y. Xie, Z. Zhang, and Y. Yang, "Siamese network with wav2vec feature for spoofing speech detection," in *Proc. Interspeech 2021*, 2021, pp. 4269–4273.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 12449–12460.

[5] J. M. Martín-Doñas and A. Álvarez, "The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9241–9245.

[6] X. Wang and J. Yamagishi, "Investigating Self-Supervised Front Ends for Speech Spoofing Countermeasures," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 100–106.

[7] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[8] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, July 2022.

[10] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[12] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6147–6151.

[13] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[14] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[15] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.

[16] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech 2019*, 2019, pp. 1008–1012.

[17] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification," in *Speaker Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.

[18] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards end-to-end synthetic speech detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[19] M. Yang, K. Zheng, X. Wang, Y. Sun, and Z. Chen, "Comparative analysis of asv spoofing countermeasures: Evaluating res2net-based approaches," *IEEE Signal Processing Letters*, 2023.

[20] Y. Zhang12, W. Wang12, and P. Zhang12, "The effect of silence and dual-band fusion in anti-spoofing system," in *Proc. Interspeech*, 2021.

[21] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 1–8.

[22] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[23] B. Huang, S. Cui, J. Huang, and X. Kang, "Discriminative frequency information learning for end-to-end speech anti-spoofing," *IEEE Signal Processing Letters*, vol. 30, pp. 185–189, 2023.