

Are Multimodal Foundation Models All That Is Needed for Emofake Detection?

Mohd Mujtaba Akhtar^{†‡*}, Girish^{‡§*}, Orchid Chetia Phukan^{‡*}, Swarup Ranjan Behera[¶]
 Pailla Balakrishna Reddy^{||}, Ananda Chandra Nayak^{**}, Sanjib Kumar Nayak^{††}, Arun Balaji Buduru[‡]
[†]V.B.S.P.U, India, [‡]IIIT-Delhi, India, [§]UPES, India, [¶]Independent Researcher, India
^{||}Reliance AI, India, ^{**}KAC, India, ^{††}VSSUT, India
 E-mail: orchidp@iiitd.ac.in

Abstract—In this work, we investigate multimodal foundation models (MFMs) for EmoFake detection (EFD) and hypothesize that they will outperform audio foundation models (AFMs). MFMs due to their cross-modal pre-training, learns emotional patterns from multiple modalities, while AFMs rely only on audio. As such, MFMs can better recognize unnatural emotional shifts and inconsistencies in manipulated audio, making them more effective at distinguishing real from fake emotional expressions. To validate our hypothesis, we conduct a comprehensive comparative analysis of state-of-the-art (SOTA) MFMs (e.g. LanguageBind) alongside AFMs (e.g. WavLM). Our experiments confirm that MFMs surpass AFMs for EFD. Beyond individual foundation models (FMs) performance, we explore FMs fusion, motivated by findings in related research areas such synthetic speech detection and speech emotion recognition. To this end, we propose SCAR, a novel framework for effective fusion. SCAR introduces a nested cross-attention mechanism, where representations from FMs interact at two stages sequentially to refine information exchange. Additionally, a self-attention refinement module further enhances feature representations by reinforcing important cross-FM cues while suppressing noise. Through SCAR with synergistic fusion of MFMs, we achieve SOTA performance, surpassing both standalone FMs and conventional fusion approaches and previous works on EFD.

I. INTRODUCTION

In recent years, significant research has focused on advancing audio deepfake detection techniques to address various categories of deepfakes, including synthetic speech, scene-fake, and singfake. Synthetic speech deepfakes leverage advance text-to-speech (TTS) and voice conversion (VC) models to generate highly realistic voices, often indistinguishable from genuine recordings [1]. Scene-fake manipulations involve altering background acoustic environments to misrepresent situational context [2], while singfake techniques modify vocal attributes to synthesize high-fidelity singing voices [3]. These deepfakes pose serious risks, such as identity fraud and misinformation through synthetic speech, manipulation of forensic evidence via scene-fake techniques, and copyright infringement or reputational harm from sing-fake synthesis. Such threats undermine trust in audio authenticity, highlighting the need for robust detection measures. As such researchers have proposed various techniques for audio deepfake detection [4]–[6]. However, with the wider and easier availability of foundation models (FMs) in recent times, audio deepfake

detection research has seen sufficient advancement [7]–[9]. These FMs are pre-trained on diverse large-scale audio data and provide uplift in performance as well as prevention of training models from scratch and this trend of leveraging FMs is consistent across synthetic speech detection [10], scenefake detection [11], and singfake detection [12]. Despite these significant advancements, an emerging vulnerability remains largely unaddressed i.e. emotionally manipulated speech deepfakes (EmoFake).

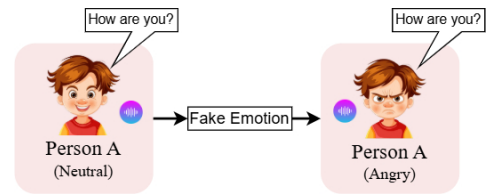


Fig. 1: Demonstration of EmoFake: Speaker A’s happy speech is manipulated to synthesize a sad emotional tone in the audio while maintaining the same spoken content (‘I’m fine, thanks!’).

EmoFake (Figure 1) alters a speaker’s emotional attributes while preserving linguistic content and speaker identity. Unlike conventional synthetic speech, it subtly modifies pitch, tone, and intensity to create realistic emotional manipulations. This emerging threat poses serious risks in security, forensics, and misinformation, enabling malicious actors to exploit emotional deception for fraud and manipulation. As such Zhao et al. [13] proposed the first initial dataset for EmoFake Detection (EFD) and presented initial baselines on it. However, they haven’t explored FMs for EFD which have shown its efficacy in detecting various types of audio deepfakes. In this study, we focus on EFD and explore FMs for the first time for EFD, to the best of our knowledge. *We hypothesize that multimodal FMs (MFMs) such as LanguageBind (LB), ImageBind (IB) will outperform audio FMs (AFMs) (WavLM, wav2vec2, Unispeech-SAT, etc.) for EFD. Unlike AFMs, which rely solely on audio, MFMs due to their cross-modal pre-training, learns emotional patterns across multiple modalities and this strength enables MFMs to more effectively detect unnatural emotional shifts and inconsistencies in manipulated audio, making them superior at distinguishing genuine from fake emotional expressions.* To test our hypothesis, we conduct a comprehensive comparative analysis of state-of-the-art (SOTA) MFMs and AFMs. Our

* Contributed equally as first authors.

experiments confirm that MFMs consistently outperform AFMs in EFD. Further, inspired from research in synthetic speech detection [14], speech emotion recognition [15] where fusion of FMs have shown improved performance. We also explore this direction and we are the primary study to explore fusion of FMs for EFD. To this end, we introduce **SCAR** (NeSted Cross-Attention NetwoRk), a novel framework for effective fusion of FMs. **SCAR** employs a novel nested cross-attention mechanism that enables multi-stage interaction between FMs representational space, refining information exchange. Additionally, a self-attention refinement module enhances feature representations by amplifying critical cross-FMs cues while suppressing noise. By leveraging **SCAR** with the fusion of MFMs, we achieve topmost performance surpassing both standalone FMs, baseline fusion techniques, and thus achieving SOTA in comparison to previous SOTA work for EFD.

Our study offers the following key contributions:

- We present the first comprehensive comparative exploration into various SOTA MFMs and AFMs to investigate the efficacy of MFMs for EFD. We show that MFMs are the best for EFD attributing to their multimodal pre-training.
- We propose, **SCAR** for fusion of FMs. With **SCAR** through the fusion of MFMs, we report the topmost performance against individual FMs, baseline fusion techniques and setting SOTA in comparison to previous works.

The resources i.e., models and code are accessible at ¹ for future research.

II. FOUNDATION MODELS

In this section, we discuss the SOTA AFMs and MFMs used in our study.

Audio Foundation Models: Unispeech-SAT² [16], Wav2vec2³ [17], WavLM⁴ [18], and HuBERT⁵ [19] were pre-trained in self-supervised manner. We use the base versions of Unispeech-SAT, Wav2vec2, WavLM, HuBERT with 94.68M, 95.04M, 94.70M, 94.68M parameters respectively and pre-trained on librispeech 960 hours of english data. Unispeech-SAT and WavLM reports the SOTA performance across different tasks in SUPERB. Unispeech-SAT was pre-trained in multi-task speaker-aware format and WavLM was pre-trained for solving masked speech modeling and speech denoising simultaneously. Wav2vec2 follows contrastive objective during its pre-training while HuBERT solves BERT-like mask prediction objective. We also consider Whisper⁶ [20] that has shown its effectiveness for related synthetic speech deepfake detection [21] and singfake detection [22]. Whisper was pre-trained on 96 languages and follows vanilla transformer encoder-decoder architecture. It was pre-trained in a multi-task learning manner and we use its base version with 74M parameters. We resample the audio samples to

16KHz before feeding to the AFMs. We extract representations from the last hidden layer of the frozen AFMs by application of pooling-average. The extracted representations are of dimension: 768 for Unispeech-SAT, Wav2vec2, WavLM, HuBERT; 512 for Whisper. For Whisper, we use only its encoder for extracting representations by discarding the decoder.

Multimodal Foundation Models: We use IB⁷ [23] and LB⁸ [24] for MFMs. IB is a multimodal learning framework that aligns modalities— images, text, audio, depth, thermal, and IMU—into a unified embedding space with image modality as the binding modality. It empolys a VIT enocder with contrastive learning objective. Without requiring direct supervision for all modality pairs, IB enables zero-shot recognition and retrieval for multimodal tasks. LB aligns multiple modalities such as video, audio, depth, and infrared—directly to the language space, eliminating reliance on image-based alignment. It employs a contrastive learning approach with a frozen 12-layer Transformer-based language encoder. It was pre-trained on VIDAL-10M, a dataset comprising 10 million video-text, infrared-text, depth-text, and audio-text pairs. We resample the audios to 16KHz before passing to the MFMs and extract representations from the last hidden state of the audio encoders of the MFMs by using average pooling. We extract representations of 768, 1024 size for LB and IB respectively.

III. MODELING

In this section, we give detailed explanation of the downstream modeling involved with individual FMs followed by the proposed framework, **SCAR** for fusion of FMs. We use fully connected network (FCN) and CNN as downstreams as used by previous research in related areas such as SER [25] and shout intensity prediction [26]. The CNN (Figure 2 (b)) consists of a convolutional block with 1D CNN layer of 32 kernels with size 3 followed by maxpooling. The features are flattened and passed to a FCN block with two dense layers of 512 and 128 neurons followed by the output layer for binary classification. We use sigmoid as the activation function in the output layer. For FCN, we keep the same architectural details as the FCN block in CNN model. The CNN models trainable parameters ranges from 0.5 to 0.7M and for FCN, it ranges from 0.45 to 0.6M depending on the input representation dimension size.

A. SCAR

The architecture of the proposed framework is illustrated in Figure 2. **SCAR** introduces a novel nested cross-attention mechanism that facilitates multi-stage interactions within the representational space of FMs, refining information exchange. Furthermore, a self-attention refinement module strengthens feature representations by emphasizing key cross-FMs cues while minimizing noise. The workflow of **SCAR** is as follows: First, representations are extracted from FMs and passed through a convolutional block with same modeling paradigm as

¹<https://github.com/Helix-IIIT-Delhi/SCAR-EmoFake>

²<https://huggingface.co/microsoft/unispeech-sat-base>

³<https://huggingface.co/facebook/wav2vec2-base>

⁴<https://huggingface.co/microsoft/wavlm-base>

⁵<https://huggingface.co/facebook/hubert-base-ls960>

⁶<https://huggingface.co/openai/whisper-base>

⁷<https://github.com/facebookresearch/ImageBind/tree/main>

⁸<https://github.com/PKU-YuanGroup/LanguageBind>

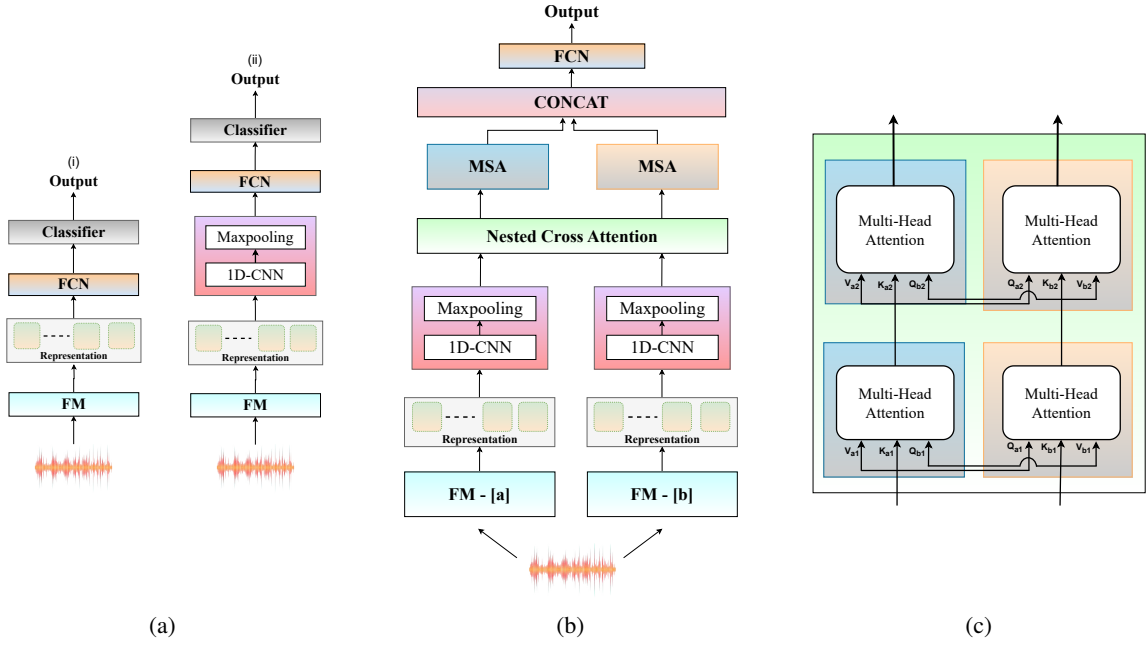


Fig. 2: Modeling Architectures: Subfigure (a) FCN and CNN; Subfigure (b) shows the proposed framework, **SCAR**; Subfigure (c) provides the detailed illustration of the nested cross-attention mechanism

done with individual FMs above. Following this, the flattened features are passed through the proposed nested cross-attention block that consists of two cross-attention layers sequentially. Let the flattened features of two FMs a and b be \mathbf{Z}_a and \mathbf{Z}_b respectively. First, the query, key, and value matrices for the first cross-attention step are computed as follows: $\mathbf{Q}_{a1} = \mathbf{Z}_a \mathbf{W}_{Q1}^a$, $\mathbf{K}_{b1} = \mathbf{Z}_b \mathbf{W}_{K1}^b$, $\mathbf{V}_{b1} = \mathbf{Z}_b \mathbf{W}_{V1}^b$ and $\mathbf{Q}_{b1} = \mathbf{Z}_b \mathbf{W}_{Q1}^b$, $\mathbf{K}_{a1} = \mathbf{Z}_a \mathbf{W}_{K1}^a$, $\mathbf{V}_{a1} = \mathbf{Z}_a \mathbf{W}_{V1}^a$. Using these, the cross-attention outputs for the first cross-attention step are computed as:

$$\mathbf{Z}_a^{(1)} = \text{softmax} \left(\frac{\mathbf{Q}_{a1} \mathbf{K}_{b1}^T}{\sqrt{d_k}} \right) \mathbf{V}_{b1}$$

$$\mathbf{Z}_b^{(1)} = \text{softmax} \left(\frac{\mathbf{Q}_{b1} \mathbf{K}_{a1}^T}{\sqrt{d_k}} \right) \mathbf{V}_{a1}$$

A second cross-attention step refines these representations: $\mathbf{Q}_{a2} = \mathbf{Z}_a^{(1)} \mathbf{W}_{Q2}^a$, $\mathbf{K}_{b2} = \mathbf{Z}_b^{(1)} \mathbf{W}_{K2}^b$, $\mathbf{V}_{b2} = \mathbf{Z}_b^{(1)} \mathbf{W}_{V2}^b$ and $\mathbf{Q}_{b2} = \mathbf{Z}_b^{(1)} \mathbf{W}_{Q2}^b$, $\mathbf{K}_{a2} = \mathbf{Z}_a^{(1)} \mathbf{W}_{K2}^a$, $\mathbf{V}_{a2} = \mathbf{Z}_a^{(1)} \mathbf{W}_{V2}^a$. The cross-attention outputs for the second cross-attention step are computed as follows:

$$\mathbf{Z}_a^{(2)} = \text{softmax} \left(\frac{\mathbf{Q}_{a2} \mathbf{K}_{b2}^T}{\sqrt{d_k}} \right) \mathbf{V}_{b2}$$

$$\mathbf{Z}_b^{(2)} = \text{softmax} \left(\frac{\mathbf{Q}_{b2} \mathbf{K}_{a2}^T}{\sqrt{d_k}} \right) \mathbf{V}_{a2}$$

Finally, a self-attention refinement step is applied to each modality separately:

$$\mathbf{Z}_a^{(3)} = \text{softmax} \left(\frac{\mathbf{Z}_a^{(2)} \mathbf{Z}_a^{(2)T}}{\sqrt{d_k}} \right) \mathbf{Z}_a^{(2)}$$

$$\mathbf{Z}_b^{(3)} = \text{softmax} \left(\frac{\mathbf{Z}_b^{(2)} \mathbf{Z}_b^{(2)T}}{\sqrt{d_k}} \right) \mathbf{Z}_b^{(2)}$$

This refinement module strengthens feature representations by highlighting essential cross-FM cues while reducing noise. The matrices \mathbf{Q} , \mathbf{K} , \mathbf{V} represent the query, key, and value matrices, while \mathbf{W}_Q , \mathbf{W}_K , \mathbf{W}_V are learnable weight matrices. The term d_k denotes the dimensionality of the key vectors used for scaling. The intermediate representations $\mathbf{Z}_a^{(i)}$ and $\mathbf{Z}_b^{(i)}$ denote the outputs at different stages of attention processing. This hierarchical cross-attention structure enables structured and progressive alignment of representations, enhancing the interaction between FMs while refining feature representations. The final fused representation is obtained by concatenating the refined outputs: $\mathbf{Z}_{\text{fused}} = \text{Concat}(\mathbf{Z}_a^{(3)}, \mathbf{Z}_b^{(3)})$. The concatenated features are then passed to a FCN block with two dense layers of 512 and 128 neurons respectively followed by the final output layer for binary classification. We keep the number of attention heads in nested cross attention for both the cross-attention stages as 2. For self-attention refinement stage, we keep the number of heads also as 2. **SCAR** consists of 1.8 to 2.3M trainable parameters depending on the dimension size of the FMs representations.

IV. RESULT AND ANALYSIS

A. Dataset

We utilize the sole open-source dataset for EFD, introduced by Zhao et al. [13]. This dataset includes recordings in both

FM	E		C	
	FCN	CNN	FCN	CNN
U	2.43	2.21	7.68	7.35
W2	2.51	2.24	4.36	3.29
W	5.67	5.25	8.78	4.91
Hu	3.62	2.93	10.94	10.23
Wh	2.31	2.04	3.25	3.11
LB	1.64	1.32	2.19	1.89
IB	1.87	1.36	2.10	1.81

TABLE I: EER scores are given in % for various FMs; E, C stands for English and Chinese subset; FMs Abbreviations are given as follows: Unispeech-SAT – U, Wav2vec2 – W2, WavLM – W, Whisper – Wh, XLS-R – X, x-vector – XV, MMS – MM, HuBERT – Hu, LB – LanguageBind, IB – ImageBind; The intensity represents performance levels, with darker shades signifying lower EER values and lighter shades indicating higher EER values; The abbreviations of the FMs and the intensity scale in this table are maintained consistently for Table II

Combinations	Concatenation		SCAR	
	E	C	E	C
U + W2	3.93	6.63	2.45	4.05
U + W	3.48	7.05	3.61	6.56
U + Hu	3.06	5.21	2.96	4.62
U + Wh	4.11	4.35	2.12	2.84
U + LB	3.42	2.94	1.95	2.45
U + IB	2.90	3.43	1.89	1.97
W2 + W	2.65	5.14	2.13	3.94
W2 + Hu	3.04	5.87	2.96	5.31
W2 + Wh	3.03	6.11	2.63	3.07
W2 + LB	2.07	2.93	1.92	1.10
W2 + IB	2.35	4.94	1.85	2.04
W + Hu	3.01	4.34	2.86	3.57
W + Wh	3.76	4.06	3.63	2.92
W + LB	2.98	3.92	1.95	2.94
W + IB	2.93	3.37	2.13	2.45
Hu + Wh	3.21	4.86	2.65	3.02
Hu + LB	4.04	6.02	3.02	2.93
Hu + IB	3.03	5.07	2.73	4.90
Wh + LB	1.38	1.89	1.24	1.66
Wh + IB	2.67	1.93	1.53	1.81
LB + IB	1.21	1.43	1.15	1.02

TABLE II: Evaluation scores (EER in %) for pairwise combinations of FMs; E, C represents the English and Chinese subsets

English and Chinese, containing samples from multiple speakers across five primary emotions: Neutral, Happy, Angry, Sad, and Surprise. Fake emotional expressions were created using seven open-source Emotional Voice Conversion models, which modify emotional cues while preserving speaker identity. The dataset is organized into separate English and Chinese subsets, each with designated training, development, and test splits. Both the English and Chinese partitions consist of 27,300 training samples, 9,100 development samples, and 17,500 test samples. We use the official split for training, validation, and evaluation

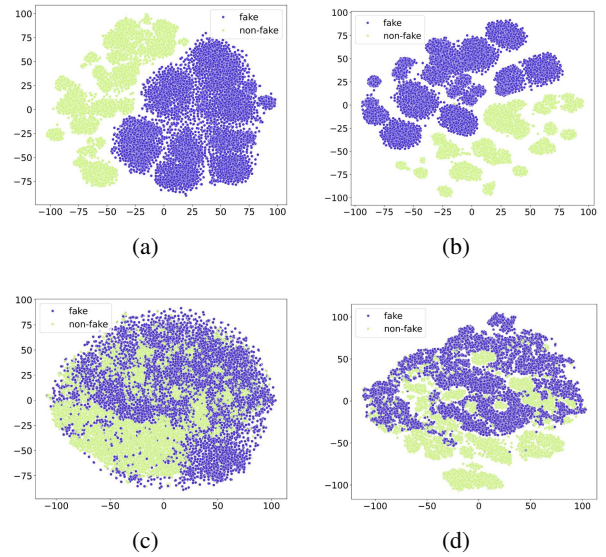


Fig. 3: The subfigures (a) LB (b) IB (c) Wav2vec2 and (d) Whisper represent the t-SNE plot visualization

of the models.

Training Details: We use the Adam as the optimizer to train the models with a learning rate of 1e-3 and a batch size of 32. For the loss function, we use binary cross-entropy. To avoid overfitting, we use dropout and early stopping.

B. Experimental Results

We use Equal Error Rate (EER) as the evaluation metric following previous research on EFD [13]. Table I presents the evaluation scores of downstream models trained on representations from various FMs. We observe that MFMs consistently achieve lower EER values across both language subsets and downstream network architectures in comparison to unimodal AFMs. *This validates our hypothesis that MFMs outperform AFMs in EFD as their cross-modal pretraining allows them to learn emotional patterns from multiple modalities. This enhances their ability to detect unnatural emotional shifts and inconsistencies in manipulated audio.* Among the AFMs, Whisper achieved the highest performance, which can be attributed to its multilingual pretraining. This aligns with previous research on synthetic speech detection [14], which found that multilingual AFMs better capture pitch, tone, and intensity variations, enhancing their ability to detect synthetic or manipulated speech. Overall, CNN-based models generally yield lower EER scores compared to FCNs with all the FMs. We also plot the t-SNE plots of raw representations from the FMs in Figure 3. We can observe clear clusters across the classes for MFMs and thus providing support to our results.

Table II presents the evaluation scores for different combinations of FMs. We use concatenation as baseline fusion technique. For modeling concatenation-based fusion, we follow similar architecture as **SCAR** and discard nested cross-attention block and the self-attention refinement. We follow the same training specifications as used for **SCAR**. We can observe

clear dominance of **SCAR** based fusions of FMs compared to concatenation-based technique and this shows the strength of **SCAR** for effective fusion. The topmost performance is achieved by fusion of MFMs i.e. LB and IB through **SCAR** and this shows the emergence of strong complementary behavior among them.

Comparison to SOTA: We compare our best model **SCAR** with LB and IB with previous SOTA work [13]. They reported the EER values as 3.65% and 8.34% for English and Chinese subsets respectively. However, **SCAR** with LB and IB reported the best EER of 1.15% and 1.02% for English and Chinese respectively and thus setting new SOTA for EFD.

V. CONCLUSION

Our study shows that MFMs outperform AFMs for EFD due to their cross-modal pretraining, which enables better recognition of unnatural emotional shifts in manipulated audio. Further, we also introduce, **SCAR**, a novel framework for fusion of FMs that leverages nested cross-attention and self-attention refinement. **SCAR** with the fusion of MFMs achieves achieves SOTA results, surpassing standalone FMs and baseline fusion method, establishing a new benchmark for EFD. Also, our study serves as a guide for future research exploring FMs for EFD.

REFERENCES

- [1] X. Wang, J. Yamagishi, M. Todisco, *et al.*, “Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Computer Speech & Language*, vol. 64, p. 101 114, 2020.
- [2] J. Yi, C. Wang, J. Tao, *et al.*, “Scenefake: An initial dataset and benchmarks for scene fake audio detection,” *Pattern Recognition*, vol. 152, p. 110 468, 2024.
- [3] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, “Singfake: Singing voice deepfake detection,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 156–12 160.
- [4] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, “Spoofing speech detection using modified relative phase information,” *IEEE Journal of selected topics in signal processing*, vol. 11, no. 4, pp. 660–670, 2017.
- [5] H. M. Tran, D. Guennec, P. Martin, *et al.*, “Spoofed speech detection with a focus on speaker embedding,” in *Interspeech 2024*, 2024, pp. 2080–2084. DOI: 10.21437/Interspeech.2024-481.
- [6] A. Guragain, T. Liu, Z. Pan, H. B. Sailor, and Q. Wang, “Speech foundation model ensembles for the controlled singing voice deepfake detection (ctrsvdd) challenge 2024,” in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024, pp. 774–781.
- [7] Y. Guo, H. Huang, X. Chen, H. Zhao, and Y. Wang, “Audio deepfake detection with self-supervised wavlm and multi-fusion attentive classifier,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 702–12 706.
- [8] O. Pascu, A. Stan, D. Oneata, E. Oneata, and H. Cucu, “Towards generalisable and calibrated audio deepfake detection with self-supervised representations,” in *Interspeech 2024*, 2024, pp. 4828–4832. DOI: 10.21437/Interspeech.2024-1302.
- [9] Y. Yang, H. Qin, H. Zhou, *et al.*, “A robust audio deepfake detection system via multi-view feature,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 13 131–13 135.
- [10] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, “Improved deepfake detection using whisper features,” in *Interspeech 2023*, 2023, pp. 4009–4013. DOI: 10.21437/Interspeech.2023-1537.
- [11] Y. E. Kheir, Y. Samih, S. Maharjan, T. Polzehl, and S. Möller, “Comprehensive layer-wise analysis of ssl models for audio deepfake detection,” *arXiv preprint arXiv:2502.03559*, 2025.
- [12] X. Chen, H. Wu, R. Jang, and H.-y. Lee, “Singing voice graph modeling for singfake detection,” in *Interspeech 2024*, 2024, pp. 4843–4847. DOI: 10.21437/Interspeech.2024-1185.
- [13] Y. Zhao, J. Yi, J. Tao, C. Wang, and Y. Dong, “Emofake: An initial dataset for emotion fake audio detection,” in *China National Conference on Chinese Computational Linguistics*, Springer, 2024, pp. 419–433.
- [14] O. Chetia Phukan, G. Kashyap, A. B. Buduru, and R. Sharma, “Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2496–2506. DOI: 10.18653/v1/2024.findings-naacl.160. [Online]. Available: <https://aclanthology.org/2024.findings-naacl.160/>.
- [15] Y. Wu, P. Yue, C. Cheng, and T. Li, “Investigation of ensemble of self-supervised models for speech emotion recognition,” in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2023, pp. 988–995.
- [16] S. Chen, Y. Wu, C. Wang, *et al.*, “Unispeech-sat: Universal speech representation learning with speaker aware pre-training,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6152–6156.
- [17] A. Baeovski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural infor-*

- mation processing systems, vol. 33, pp. 12 449–12 460, 2020.
- [18] S. Chen, C. Wang, Z. Chen, *et al.*, “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, 2022.
- [19] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.
- [21] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, “Improved deepfake detection using whisper features,” *arXiv preprint arXiv:2306.01428*, 2023.
- [22] F. Sharma and P. Gupta, “Deepfake detection of singing voices with whisper encodings,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10887871.
- [23] R. Girdhar, A. El-Nouby, Z. Liu, *et al.*, “Imagebind: One embedding space to bind them all,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.
- [24] B. Zhu, B. Lin, M. Ning, *et al.*, “Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment,” *arXiv preprint arXiv:2310.01852*, 2023.
- [25] O. Chetia Phukan, A. Balaji Buduru, and R. Sharma, “Transforming the embeddings: A lightweight technique for speech emotion recognition tasks,” in *Interspeech 2023*, 2023, pp. 1903–1907. DOI: 10.21437/Interspeech.2023-2561.
- [26] T. Fukumori, T. Ishida, and Y. Yamashita, “Investigating the effectiveness of speaker embeddings for shout intensity prediction,” in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2023, pp. 1838–1842.