Investigating Prosodic Signatures via Speech Pre-Trained Models for Audio Deepfake Source Attribution

Orchid Chetia Phukan[1*], **Drishti Singh**[1*], **Swarup Ranjan Behera**[2]
Arun Balaji Buduru[1], **Rajesh Sharma**[3,4]
[1]IIIT-Delhi, India, [2]Independent Researcher, India
[3]University of Tartu, Estonia, [4]Plaksha University, India
Correspondence: orchidp@iiitd.ac.in

Abstract

In this work, we investigate various state-of-the-art (SOTA) speech pre-trained models (PTMs) for their capability to capture prosodic signatures of the generative sources for audio deepfake source attribution (ADSD). These prosodic characteristics can be considered one of major signatures for ADSD, which is unique to each source. So better is the PTM at capturing prosodic signs better the ADSD performance. We consider various SOTA PTMs that have shown top performance in different prosodic tasks for our experiments on benchmark datasets, ASVSpoof 2019 and CFAD. x-vector (speaker recognition PTM) attains the highest performance in comparison to all the PTMs considered despite consisting lowest model parameters. This higher performance can be due to its speaker recognition pre-training that enables it for capturing unique prosodic characteristics of the sources in a better way. Further, motivated from tasks such as audio deepfake detection and speech recognition, where fusion of PTMs representations lead to improved performance, we explore the same and propose **FINDER** for effective fusion of such representations. With fusion of Whisper and x-vector representations through **FINDER**, we achieved the topmost performance in comparison to all the individual PTMs as well as baseline fusion techniques and attaining SOTA performance.

1 Introduction

Imagine waking up to find your voice used in a viral audio clip, falsely implicating you in a scandal. This increasingly plausible scenario highlights the growing threat of audio deepfakes. With advancements in text-to-speech (TTS) and voice conversion (VC) technologies, malicious actors can now create synthetic audio that is nearly indistinguishable from authentic recordings. From high-profile frauds (Stupp, 2019) to the viral spread of falsified audio targeting political figures (BBC, 2024), the misuse of synthetic audio for financial scams and misinformation underscores the urgent need for reliable detection methods. As the authenticity of audio content becomes increasingly difficult to verify, the importance of audio deepfake detection (ADD) has never been more pressing.

Despite advancements in ADD, research has primarily focused on binary classification - distinguishing real from fake audio (Wu et al., 2015; Kinnunen et al., 2017; Todisco et al., 2019; Liu et al., 2023; Yamagishi et al., 2021; Yi et al., 2022, 2023; Shaaban et al., 2023; Hamza et al., 2022; Altalahin et al., 2023; Kilinc and Kaledibi, 2023). This approach, while effective in its simplicity, lacks the granularity needed to address a crucial aspect of deepfake detection: source attribution. Audio deepfake source attribution (ADSD) goes beyond merely identifying whether audio is real or fake; it seeks to uncover the specific tool or model responsible for generating the synthetic audio (Yan et al., 2022b; Zhang et al., 2023; Zhu et al., 2022). This capability is vital for improving the explainability of detection systems and enabling targeted countermeasures, especially in high-stakes contexts such as audio forensics and intellectual property protection.

Generative sources such as TTS, VC, etc. systems embed their unique prosodic characteristics, such as pitch, tone, rhythm, and intonation, into their generated audios, reflecting the inherent design and processing patterns of the generative system. These prosodic signatures are vital for accurately identifying the source and can be considered one of the major fingerprint. In this study, we investigate various state-of-the-art (SOTA) speech pre-trained models (PTMs) for capturing these prosodic signatures of source for ADSD. We specifically consider PTMs that have shown efficacy in various prosodic tasks such as speech emotion recog-

*Equal Contribution

Findings of the Association for Computational Linguistics: ACL 2025, pages 4206–4214

nition (SER), depression detection, etc. We evaluate PTMs as they have shown efficacy in previous ADSD and ADD works (Klein et al., 2024a; Chetia Phukan et al., 2024) by providing performance benefit as well as prevention of training models from scratch. Further, motivated by tasks like ADD (Chetia Phukan et al., 2024), speech recognition (Arunkumar et al., 2022), and emotion recognition (Phukan et al., 2025), we also investigate the fusion of different PTM representations and propose, **FINDER** (**F**us**I**on through Re**N**yi **D**iv**ER**gence) for effective fusion. We believe, we are the first work, to the best of our knowledge, for exploring fusion of PTMs representations for ADSD.

To summarize, the main contributions:

- We give a comprehensive comparative study of SOTA speech PTMs that have SOTA performance in various prosodic tasks for investigating their capacity of capturing prosodic signatures for ADSD.
- We show that x-vector, a speaker recognition PTM, achieves the highest performance and this behavior can be attributed to its speaker recognition pre-training that enables it to capture prosodic features better.
- We propose **FINDER**, a novel framework that leverages renyi divergence as a fusion mechanism for fusion of PTMs representations.

We will release the codes here[1].

2 Related Work

Early work on ADSD introduced the problem of identifying attacker signatures, showing that representations from RNN can characterize both seen and unseen attackers with high accuracy (Müller et al., 2022). Subsequent studies focused on detecting vocoder-specific fingerprints, revealing that vocoders leave identifiable traces in generated audio (Yan et al., 2022b,a). Building on this, methods such as t-SNE visualization and ResNet-based architectures further improved fingerprint detection accuracy (Yan et al., 2022a). Deng et al. 2024 proposed VFD-Net, a patch-wise supervised contrastive learning method, which achieved robust performance under cross-set and audio compression conditions. More recent work (Klein et al., 2024b) has shown potential of using PTMs as backbones for improved ADSD. In this study, for the first time, we investigate SOTA PTMs that have shown for assessing their capability of capturing unique prosodic signatures of sources for better ADSD.

3 Pre-Trained Models

Wav2vec2 (Baevski et al., 2020) and WavLM (Chen et al., 2022) are monolingual PTMs. Wav2vec2 trained on the LibriSpeech dataset, masks the input in latent space, it has shown effectiveness in prosodic tasks such as SER (Pepino et al., 2021). WavLM showed SOTA performance on SUPERB including various prosodic tasks. XLS-R (Babu et al., 2022) and Whisper (Radford et al., 2023) are multilingual PTMs. XLS-R was pre-trained on 128 languages for 436k hours of unlabeled speech while Whisper (Radford et al., 2023) on 96 languages for 680k hours of labelled data. XLS-R shows good performance in ML-SUPERB (Shi et al., 2023) that includes prosodic tasks while Whisper shows potential for SER (Feng and Narayanan, 2023). In addition to these PTMs, we consider, x-vector (Snyder et al., 2018), trained for speaker recognition. It excels in various prosodic tasks such as SER (Chetia Phukan et al., 2023), shout intensity prediction (Fukumori et al., 2023), depression detection (Egas-López et al., 2022), and so on. We also consider, Wav2Vec2-emo[2] fine-tuned for SER, as SER is a prosodic task and we think its representations might be helpful for ADSD. Additional details regarding the above PTMs are provided in Appendix A.1.

4 Modeling

We consider two downstream networks i.e. fully connected network (FCN) and CNN with individual PTM representations. The FCN model consists of 3 dense layers with 256, 128, and 64 neurons while CNN model has convolutional blocks followed by three dense layers and finally the output layer. For CNN, both the convolution blocks consists of 1D-CNN and max-pooling layer followed by flattening and FCN with similar configuration as FCN network above. Hyperparameters detail is given in Appendix A.3.

FINDER: We propose **FINDER** for effective fusion of PTMs representations. The architecture is given

[1] https://github.com/orchidchetiaphukan/prosodic_ADSD_ACL2025

[2] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

in Figure 1. The PTMs representations are passed through two convolution blocks with same configuration as CNN model built for modeling with individual PTM representations above.

Features are flattened after the convolution blocks and linearly projected to 120-dimensional size to keep the same dimensions and also for computational constraints. The projected features then passed through the renyi divergence (RD). RD is a measure of divergence or dissimilarity between two probability distributions (Van Erven and Harremos, 2014). Here, we frame RD as a loss function that calculates the divergence between the feature representations of two different PTMs. The lower the RD value, the more aligned the PTMs representations are to each other. We aim to reduce the disimilarity between the feature representations and make it closer to each other. e_{a} and e_{b} be the feature space for two PTMs networks.

RD between the two feature distributions e_{a} and e_{b} is given by:

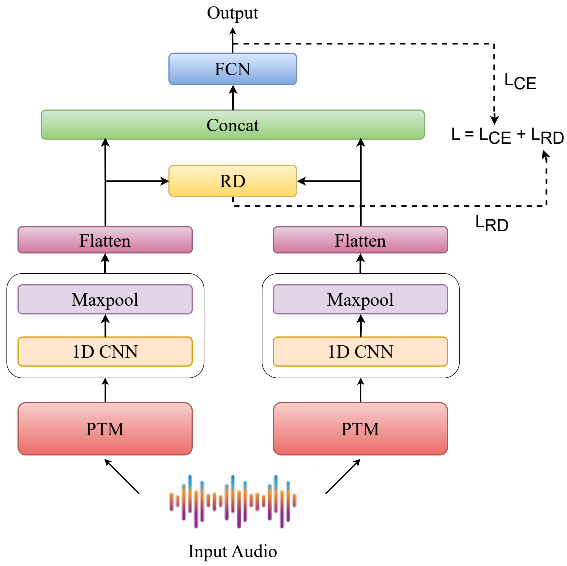


Figure 1: Proposed Framework **FINDER**: RD and FCN stand for renyi divergence and fully connected network, respectively; L, L_{CE}, and L_{RD} represent the total loss, cross-entropy loss, and renyi divergence loss, respectively.

$$\mathcal{L}_{RD} = \frac{1}{\alpha - 1} \log \left(\sum_{i=1}^{D} (e_{\mathrm{a},i} + \epsilon)^{\alpha} (e_{\mathrm{b},i} + \epsilon)^{1-\alpha} \right)$$

where D is the embedding dimension, $\alpha > 1$ controls the order of the divergence, and ϵ is a small constant for numerical stability.

Finally, we add the RD loss L_{RD} to the cross-entropy loss L_{CE} for joint optimization. Total loss is given as:

$$\mathcal{L} = \lambda \mathcal{L}_{CE} + (1 - \lambda) \mathcal{L}_{RD}$$

where λ is a hyperparameter and weightage parameter for the losses.

5 Experiments

Benchmark Datasets: We use two benchmark datasets for our experiments namely ASVSpoof 2019 (ASV) (Wang et al., 2020) and FAD Chinese Dataset (CFAD) (Ma et al., 2024). We combine the train, validation and testing splits for ASV and in resultant we got 19 classes (A01 to A19) as fake audio sources, while CFAD has 9 classes (0 - 8) for source classes. We followed 5-fold cross validation for ASV and followed official split given for CFAD. For more details on the datasets and data preprocessing, refer to Appendix A.2.

Training Details: We use softmax as the activation function in the output layer of all the models that outputs class probabilites. We use Adam as optimizer with learning rate of 10^{-3}. We trained all the models for 40 epochs with a batch size of 32. We use cross-entropy as the loss function for all the models. We use early stopping and dropout for preventing overfitting. For experiments with **FINDER**, we set $\alpha = 2$, $\epsilon = 0.1$, and $\lambda = 0.4$ and keep these values constant throughout the experiments as we got better results with these values through some preliminary exploration.

Experimental Results: We use accuracy and equal error rate (EER) as the evaluation metrics for experiments as used by previous works on ADSD (Klein et al., 2024a) and ADD (Liu et al., 2023). For EER, we present the average scores of one-vs-all.

Table 1 presents the results of downstream models trained on individual PTM representations. x-vector consistently delivers best results across both the datasets, achieving high accuracy and lower EER. This performance can be traced back to its speaker recognition pre-training that equips x-vector to better capture prosodic features also consistent across various prosodic tasks (Chetia Phukan et al., 2023; Fukumori et al., 2023; Egas-López et al., 2022). In contrast, monolingual PTMs like Wav2vec2 and WavLM underperformed due to their limited capacity to capture source-specific prosodic features. The performance of Wav2vec2-emo is a known behavior as it was trained for SER,

Table 1: Performance Comparison of individual PTMs representations on ASV and CFAD; All the scores are average of 5-folds for ASV; All the scores are in %; High Accuracy, Low EER better the model

Representations	ASV				CFAD			
	FCN		CNN		FCN		CNN	
	Accuracy	EER	Accuracy	EER	Accuracy	EER	Accuracy	EER
Wav2vec2	45.25	21.54	61.75	7.76	49.25	29.20	74.50	10.20
WavLM	33.48	15.25	45.46	10.50	32.23	22.23	35.78	27.60
XLS-R	63.98	11.55	79.04	4.01	50.25	19.30	76.90	9.50
Whisper	75.69	9.85	87.03	4.01	70.14	15.85	85.01	8.10
x-vector	**87.48**	**4.42**	**97.60**	**2.03**	**74.58**	**10.02**	**91.39**	**4.40**
Wav2vec2-emo	78.45	8.40	86.35	2.50	65.30	12.12	86.50	8.60

Table 2: Performance Comparison of Fusion Methods on ASV and CFAD; All the scores are average of 5-folds for ASV; All scores are in %; High Accuracy, Low EER better the model

Representations	ASV				CFAD			
	Concatenation		FINDER		Concatenation		FINDER	
	Accuracy	EER	Accuracy	EER	Accuracy	EER	Accuracy	EER
Wav2vec2 + Wav2vec2-emo	96.57	0.67	96.69	0.63	89.80	3.60	93.32	3.51
WavLM + Wav2vec2-emo	93.60	1.10	94.60	1.08	85.65	5.70	88.03	5.55
XLS-R + Wav2vec2-emo	91.23	1.10	94.35	1.04	92.47	2.79	95.72	2.64
Whisper + Wav2vec2-emo	96.03	0.66	96.66	0.63	94.95	1.50	98.47	1.41
x-vector + Wav2vec2-emo	93.12	1.33	98.62	0.37	91.79	2.80	96.71	2.50
x-vector + Wav2vec2	96.25	0.63	98.16	0.33	85.50	5.40	86.77	4.40
WavLM + Wav2vec2	96.38	0.69	96.99	0.65	89.75	2.84	96.26	2.61
Whisper + Wav2vec2	96.87	0.51	97.21	0.50	94.11	2.04	97.59	1.91
XLS-R + Wav2vec2	96.09	0.65	96.54	0.57	94.93	1.45	98.09	1.24
WavLM + XLS-R	86.92	2.30	87.40	2.04	92.63	1.80	98.75	1.17
Whisper + XLS-R	93.99	1.10	94.50	1.01	89.50	3.40	94.98	2.97
x-vector + XLS-R	96.58	1.20	97.84	0.35	90.87	3.30	95.58	1.60
Whisper + WavLM	94.26	0.85	94.69	0.84	88.86	3.84	94.28	3.08
x-vector + WavLM	95.85	0.40	98.37	0.36	64.32	10.10	78.77	8.17
Whisper + x-vector	**97.16**	**0.32**	**98.91**	**0.26**	**95.00**	**1.10**	**99.01**	**1.07**

but not better than x-vector. Whisper and XLS-R also shows better performance than monolingual PTMs as seen in previous research for ADD (Chetia Phukan et al., 2024) that multilingual PTMs capture diverse pitches, tones, etc. prosodic characteristics better than monolingual PTMs. We also plots the t-SNE plots of the PTMs representations in Appendix Figure 2 and 3. The plots complements the performance of x-vector as we can observe better clusters across the source classes. Also, CNN models consistently outperform FCN models.

Table 2 presents the results for fusion of PTMs representations through baseline concatentation fusion technique and **FINDER**. For the baseline, we use the same modeling paradigm except the RD loss. We also kept the training details same as kept for **FINDER**. We observe that results of fusion of PTMs representations through both the fusion techniques are better than the individual PTM representations, thus, showing their complementary nature. Fusion of PTM representations through **FINDER** consistently beat the concatenation based baseline fusion techniques showing its effectiveness. Fusion of Whisper and x-vector through **FINDER** shows the best performance across both the datasets.

Dataset	Model	Accuracy(%)	EER(%)
ASV	**FINDER (Whisper + x-vector)**	**98.91**	**0.26**
	MiO(Whisper + x-vector)	97.75	0.68
	AASIST(Wav2vec2)	63.81	5.68
CFAD	**FINDER**	**99.01**	**1.07**
	MiO(Whisper + x-vector)	97.31	2.15
	AASIST(Wav2vec2)	77.92	9.69

Table 3: Comparison to previous SOTA works

Comparison to Previous Works: As we have considered all the source classes across train, validation, and test split for ASV, so we can't directly compare our results to previous works. For CFAD, we are the first one to perform ADSD, so there is not previous work to compare to. So, we reimplemented some of the SOTA methods for ADSD and ADD and compared it with our results. For ADD, we consider MiO (Chetia Phukan et al., 2024), an SOTA method that proposed combination of PTM representations. We implemented with Whisper and x-vector representations i.e. the best performing pair. For ADSD, we implemeted AASIST (Jung et al., 2022) as downstream with Wav2vec2 representations used by Klein et al. 2024a. Table 3 presents the comparison of the proposed with the SOTA methods. We observe the **FINDER** outperforms both the methods and attains SOTA performance showing its effectiveness for ADSD.

6 Conclusion

In this work, we investigate various SOTA speech PTMs across different prosodic tasks for their ability to capture prosodic signatures of generative sources for ADSD. We evaluate monolingual, multilingual, and speaker recognition PTMs on benchmark datasets (ASV, CFAD), finding that x-vector, a speaker recognition PTM, outperforms others due to its ability to capture better source-specific prosody. Further, we explored fusion of PTM representations for ADSD and propose **FINDER** for the same. With fusion of x-vector and Whisper representations through **FINDER**, we achieve the topmost performance surpassing both individual PTMs and baseline fusion techniques and attains SOTA performance.

7 Limitations

One major limitation of our work is the proposed systems are not built for open-vocabulary ADSD. It can only identify generative systems present in the datasets, considered in our study. In our future work, we will work towards building systems for open-vocabulary ADSD.

Another limitation is the experimentation with limited downstream networks, previous research has shown that the downstream performance changes with the downstream modeling technique (Zaiem et al., 2023). Here, we have experimented with only FCN and CNN. In future, we will explore further varied downstream networks for ADSD.

8 Ethics Statement

ASV and CFAD are publicly available and widely recognized benchmarks for audio deepfake research and have been automatically identified to protect speaker privacy. Our proposed method aims to enhance ADSD, supporting efforts to combat misinformation, fraud, and misuse of generative models. We acknowledge the potential risks associated with generative technologies and emphasize that our work is solely intended for improving detection systems and promoting cybersecurity.

References

2024. Fake Biden robocall tells voters to skip New Hampshire primary election - BBC News. https://www.bbc.com/news/world-us-canada-68064247. Last accessed: 2024-03-05.

I. Altalahin, S. AlZu'bi, A. Alqudah, et al. 2023. Unmasking the truth: A deep learning approach to detecting deepfake audio through mfcc features. In *Proc. of 2023 International Conference on Information Technology (ICIT)*, pages 511–518.

A Arunkumar, Vrunda Nileshkumar Sukhadia, and Srinivasan Umesh. 2022. Investigation of ensemble features of self-supervised pretrained models for automatic speech recognition. In *Interspeech 2022*, pages 5145–5149.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Orchid Chetia Phukan, Arun Balaji Buduru, and Rajesh Sharma. 2023. Transforming the embeddings: A lightweight technique for speech emotion recognition tasks. In *INTERSPEECH 2023*, pages 1903–1907.

Orchid Chetia Phukan, Gautam Kashyap, Arun Balaji Buduru, and Rajesh Sharma. 2024. Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2496–2506, Mexico City, Mexico. Association for Computational Linguistics.

Junlong Deng, Yanzhen Ren, Tong Zhang, Hongcheng Zhu, and Zongkun Sun. 2024. Vfd-net: Vocoder fingerprints detection for fake audio. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12151–12155. IEEE.

José Vicente Egas-López, Gábor Kiss, Dávid Sztahó, and Gábor Gosztolya. 2022. Automatic assessment of the degree of clinical depression from speech using x-vectors. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8502–8506.

Tiantian Feng and Shrikanth Narayanan. 2023. Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Takahiro Fukumori, Taito Ishida, and Yoichi Yamashita. 2023. 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)*, pages 1838–1842. IEEE.

A. Hamza, A. R. R. Javed, F. Iqbal, et al. 2022. Deepfake audio detection via mfcc features using machine learning. *IEEE Access*, 10:134018–134028.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. 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)*, pages 6367–6371. IEEE.

H. H. Kilinc and F. Kaledibi. 2023. Audio deepfake detection by using machine and deep learning. In *Proc. of 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 1–5.

T. Kinnunen, M. Sahidullah, H. Delgado, N. Evans, M. Todisco, et al. 2017. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Proc. of INTERSPEECH*.

Nicholas Klein, Tianxiang Chen, Hemlata Tak, Ricardo Casal, and Elie Khoury. 2024a. Source tracing of audio deepfake systems. In *Interspeech 2024*, pages 1100–1104.

Nicholas Klein, Tianxiang Chen, Hemlata Tak, Ricardo Casal, and Elie Khoury. 2024b. Source tracing of audio deepfake systems. *ArXiv*, abs/2407.08016.

X. Liu, X. Wang, M. Sahidullah, et al. 2023. ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Haoxin Ma, Jiangyan Yi, Chenglong Wang, Xinrui Yan, Jianhua Tao, Tao Wang, Shiming Wang, and Ruibo Fu. 2024. Cfad: A chinese dataset for fake audio detection. *Speech Communication*, 164:103122.

Nicolas Michael Müller, Franziska Dieckmann, and Jennifer Williams. 2022. Attacker attribution of audio deepfakes. *ArXiv*, abs/2203.15563.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Interspeech 2021*, pages 3400–3404.

Orchid Chetia Phukan, Mohd Mujtaba Akhtar, Girish, Swarup Ranjan Behera, Sishir Kalita, Arun Balaji Buduru, Rajesh Sharma, and S.R Mahadeva Prasanna. 2025. Strong alone, stronger together: Synergizing modality-binding foundation models with optimal transport for non-verbal emotion recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

O. A. Shaaban, R. Yildirim, and A. A. Alguttar. 2023. Audio deepfake approaches. *IEEE Access*, 11:132652–132682.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2023. Ml-superb: Multilingual speech universal performance benchmark. In *INTERSPEECH 2023*, pages 884–888.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Catherine Stupp. 2019. Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. *The Wall Street Journal*, 30(08).

M. Todisco, X. Wang, V. Vestman, Md. Sahidullah, and K. Lee. 2019. ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Proc. of INTERSPEECH*.

Tim Van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, et al. 2015. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *Proc. of INTERSPEECH*.

J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, and N. Evans. 2021. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *Proc. of INTERSPEECH*.

Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Zhengkun Tian, and Ruibo Fu. 2022a. System fingerprints detection for deepfake audio: An initial dataset and investigation. *ArXiv*, abs/2208.10489.

Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, and Ruibo Fu. 2022b. An initial investigation for detecting vocoder fingerprints of fake audio. In *Proc. of the 1st International Workshop on Deepfake Detection for Audio Multimedia*.

J. Yi, R. Fu, J. Tao, et al. 2022. ADD 2022: The first audio deep synthesis detection challenge. In *Proc. of ICASSP*, pages 9216–9220.

J. Yi, J. Tao, R. Fu, et al. 2023. ADD 2023: The second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. 2023. Speech self-supervised representation benchmarking: Are we doing it right? In *INTERSPEECH 2023*, pages 2873–2877.

Chu Yuan Zhang, Jiangyan Yi, Jianhua Tao, Chenglong Wang, and Xinrui Yan. 2023. Distinguishing neural speech synthesis models through fingerprints in speech waveforms. *ArXiv*, abs/2309.06780.

Tinglong Zhu, Xingming Wang, Xiaoyi Qin, and Ming Li. 2022. Source tracing: Detecting voice spoofing. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

A Appendix

A.1 Detailed Information of PTMs

In this section, we present detailed information regarding the PTMs utilized in our study.

- **Wav2vec2**[3]: It is trained in self-supervised manner to perform a contrastive task based on the quantization of jointly learned latent representations. It is trained on 960 hrs of audio LibriSpeech data, particularly English language. We are using the Wav2Vec2-base model with approximately 95 million parameters.

- **Wav2vec2-emo**: It is a fine-tuned version of Wav2vec2 for SER on the benchmark IEMOCAP dataset using the general purpose speech toolkit - SpeechBrain. Similar to Wav2Vec2, it has 95 million parameters.

- **XLS-R**[4]: It is a cross-lingual learnng model trained in self-supervised manner, based on the Wav2Vec2 framework. The model is trained on approximately 500k hrs of open source speech audio data, spanning over 128 languages. For our experiments, we are utilizing the version with 300 million parameters.

- **Whisper**[5]: It is a model trained in a multitask manner on internet data of about 680k hours, consisting of multilingual and multitask supervision. Whisper demonstrates ability to generalize on diverse datasets and domains without the need for fine-tuning, in a zero-shot setting. Whisper shows improved performance on speech recognition over XLS-R. We have used the base model with 74 million parameters.

- **WavLM**[6] : It is self-supervised PTM designed to address the challenges of learning universal speech representations for diverse speech processing tasks. WavLM uses masked speech prediction with speech denoising during pre-training, enabling it to model both spoken content and non-ASR tasks effectively. It is trained on 960k hours of Librispeech English data, outperforming models like Wav2vec2 and HuBERT. We have used the base version with 94.70 million parameters.

- **X-vector**[7]: It is a time-delay neural network (TDNN) trained in a supervised manner for speaker recognition. It is trained on Voxceleb 1+ Voxceleb2 training data, using the general purpose speech toolkit - Speechbrain. It has achieved SOTA performance in speaker recognition, outperforming models like i-vector. We are using the Speechbrain model with approx 4.2 million parameters.

A.2 Benchmark Datasets

ASV: It was developed to advance research in detecting audio deepfake and protecting automatic speaker verification systems from manipulation. It encompasses three major spoofing types: synthetic speech, converted speech, and replay attacks, generated using SOTA neural acoustic and waveform models. The dataset comprises bonafide (genuine) and spoofed audio samples, sourced from 19 synthetic systems, with recordings at a 16 kHz sam-

[3] https://huggingface.co/facebook/wav2vec2-base

[4] https://huggingface.co/facebook/wav2vec2-xls-r-300m

[5] https://huggingface.co/openai/whisper-base

[6] https://huggingface.co/microsoft/wavlm-base

[7] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

Category	Frequency
Dev	24,844
Train	25,380
Eval	71,237
Total Samples	**121,461**
Real	12,483
Total Fake Audio Samples	108,978

Table 4: ASV Statistics

Category	Total	Fake
Train	138,400	25,600
Validation	14,400	9,600
Test	42,000	28,000
Total Samples	**194,800**	**63,200**

Table 5: CFAD Statistics

ple rate and 16-bit depth. Bonafide recordings include diverse speakers, capturing a range of accents, speaking styles, and vocal characteristics. Spoofed samples are predominantly produced using SOTA voice conversion (VC) and text-to-speech (TTS) methods, ensuring high clarity and naturalness. For this study, we leverage the 19 spoofed classes for neural generator attribution. Statistics are presented in Table 4.

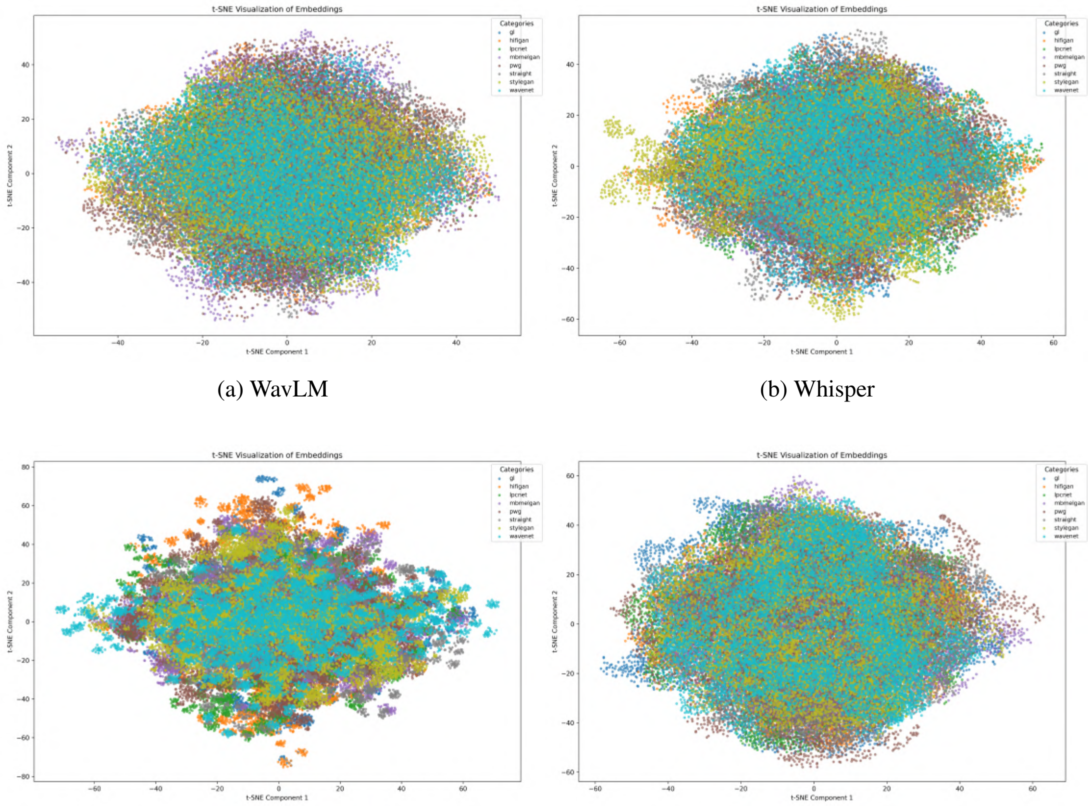
CFAD: It addresses the lack of public Chinese datasets under noisy conditions for fake audio detection. It includes bonafide and fake audio generated by 12 advanced speech generation techniques. To simulate real-world conditions, three noise datasets were added at five different signal-to-noise ratios (SNRs). Spoofed samples were synthesized using 11 SOTA TTS and VC methods, as well as neural network-based speech generation models. The dataset facilitates evaluation using metrics such as Equal Error Rate (EER) and Tandem Detection Cost Function (t-DCF). CFAD contributes significantly to audio forensics, enabling the identification of manipulated content and the attribution of spoofing algorithms.

Data Pre-Processing: We resample the audios to 16 KHz before passing it to the PTMs and extract representations from the last hidden state of the PTMs by average pooling. We extract representations of 768-dimensions for Wav2vec2, Wav2vec2-emo and WavLM. We get representations of 512 for Whisper-encoder and x-vector and 1024 for XLS-R.

A.3 Hyperparameters and System Configurations

The first layer of convolution block of the CNN model has 256 filters and a kernel size of 3, followed by batch normalization, and max pooling (pool size 2). The second layer uses 128 filters and a kernel size of 3, succeeded by followed by batch normalization, and max pooling (pool size 2. The trainable parameters for CNN models with individual PTMs representations ranged from 0.8 to 1.2 million parameters depending on the input representation dimension size. Also, for the fusion experiments, the trainable parameters of the models range from 1.3 to 1.5 million parameters.

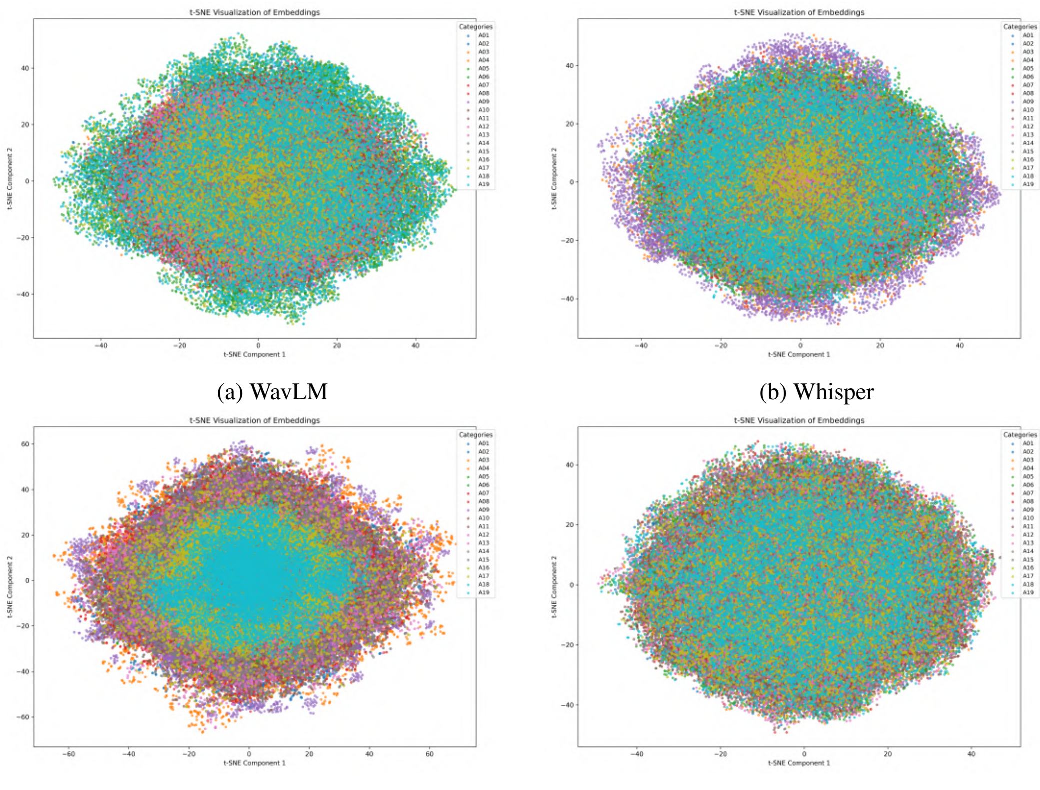
We use *Tensorflow* library for carrying out our experiments. We use A5000 GPU for running our experiments.



(a) WavLM (b) Whisper

(c) x-vector (d) Wav2vec2-emo

Figure 2: Representation Space Visualization of PTMs for CFAD



(a) WavLM (b) Whisper

(c) x-vector (d) Wav2vec2-emo

Figure 3: Representation Space Visualization of PTMs for ASV