Heterogeneity over Homogeneity: Investigating Multilingual Speech Pre-Trained Models for Detecting Audio Deepfake

Orchid Chetia Phukan[1†*], **Gautam Siddharth Kashyap**[1†], **Arun Balaji Buduru**[1]
Rajesh Sharma[1,2]
[1]IIIT-Delhi, India
[2]University of Tartu, Estonia
orchidp@iiitd.ac.in

Abstract

In this work, we investigate multilingual speech Pre-Trained models (PTMs) for Audio deepfake detection (ADD). We hypothesize that multilingual PTMs trained on large-scale diverse multilingual data gain knowledge about diverse pitches, accents, and tones, during their pre-training phase and making them more robust to variations. As a result, they will be more effective for detecting audio deepfakes. To validate our hypothesis, we extract representations from state-of-the-art (SOTA) PTMs including monolingual, multilingual as well as PTMs trained for speaker and emotion recognition, and evaluated them on ASVSpoof 2019 (ASV), In-the-Wild (ITW), and DECRO benchmark databases. We show that representations from multilingual PTMs, with simple downstream networks, attain the best performance for ADD compared to other PTM representations, which validates our hypothesis. We also explore the possibility of fusion of selected PTM representations for further improvements in ADD, and we propose a framework, **MiO** (**M**erge **i**nto **O**ne) for this purpose. With **MiO**, we achieve SOTA performance on ASV and ITW and comparable performance on DECRO with current SOTA works.

1 Introduction

The popularity of audio deepfakes has raised multiple concerns in areas dealing with personal and public security due to its capability to impersonate and share false, often malicious information. Scammers, for example, have utilized audio deepfake to mimic a German executive, successfully convincing a transfer of 220,000 Euros to a Hungarian supplier (Stupp, 2019). Thus, checking and evaluating authenticity of any audio content is important through robust and reliable measures. Motivated by this, in this work, we focus on Audio Deepfake Detection (ADD).

To combat this progressing issue, various detection methods have been proposed (Hanilçi et al., 2015; Qian et al., 2016; Ma et al., 2021; Luo et al., 2021). These works leveraged statistical attributes or the raw audio as input to the models. However, with the wide-scale accessibility of Pre-Trained Models (PTMs), ADD as a task has undergone exponential advancement. Representations from PTMs are used as input features to downstream ADD and it comes with a series of benefits, which include higher accuracy in ADD and saving time as well as resources, in building ADD systems from the ground up. PTMs come in various architectures as well as varied pre-training schemes and are trained on large-scale datasets. They can be either trained in a supervised (Eg. Whisper (Radford et al., 2023)) or in a self-supervised manner (Eg. Wav2vec2 (Baevski et al., 2020)) and also on either single or multiple languages. Despite these PTMs being pre-trained on only real speech data, representations from these PTMs have shown exceptional performance in identifying real content from their fake counterparts (wen Yang et al., 2021).

Our work relies on the hypothesis that *multilingual PTMs trained on large-scale diverse multilingual data, acquire knowledge about diverse pitches, accents, tones, and are more robust to variations, hence will be more effective for identifying audio deepfakes than other PTMs*. So to validate our hypothesis, we extract representations from eight state-of-the-art (SOTA) PTMs including multilingual (XLS-R, Whisper, MMS), monolingual (Unispeech-SAT, WavLM, Wav2vec2), speaker recognition (x-vector), and emotion recognition (XLSR_emo) and evaluate them with two simple probing networks (Fully Connected Network (FCN), Convolution Neural Network (CNN)) on three benchmark datasets ASVSpoof 2019 (ASV),

* Corresponding Author
† Authors contributed equally as first authors

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2496–2506

In-the-Wild (ITW), and DECRO. We also investigate by combining representations from different PTMs as it has been seen in other speech processing tasks such as speech recognition (Arunkumar et al., 2022) that certain representations act as complementary to each other and we propose a framework, **M**erge **i**nto **O**ne (**MiO**) for the same. To the best of our knowledge, this is the first study, to explore fusion of PTM representations for ADD. Our study makes the following contributions:

- Comprehensive empirical study to demonstrate the performance of multilingual PTMs for ADD, which have shown top performance in comparison to its other PTM counterparts across the three datasets.
- A novel approach to fuse representations from different PTMs, namely **MiO**. Our approach shows demonstrable improvement in performances over individual representations. It achieves SOTA in terms of Equal Error Rate (EER) in ASV, ITW, and competitive performance in DECRO.

2 Related Works

In this section, we give an overview of various prolific ADD methods proposed. ADD as a task has caught the attention with the release of the ASVspoof 2015 (Wu et al., 2015) database. Initially, researchers built GMM and SVM-based modeling approaches with statistical audio features as input (Sahidullah et al., 2015). Previous works have also harnessed neural network-based models such as CNN, RNN, etc. for ADD (Tom et al., 2018; Gomez-Alanis et al., 2019; Alzantot et al., 2019).

Researchers have exploited self-supervised learning (SSL) modeling approaches for ADD (Lee et al., 2023; jin Shim et al., 2020; Jiang et al., 2020). Further, different types of PTMs such as Wav2vec, HuBERT, TERA, Mockingjay, etc also been explored for ADD (Eom et al., 2022; wen Yang et al., 2021). Wang et al. 2023 showed that generalization of ADD systems increases with combination of Wav2vec, prosodic, and pronounciation information as input features. In this work, we evaluate eight PTMs to validate our hypothesis that multilingual PTMs trained on extensive and diverse multilingual datasets allow them to capture knowledge related to diverse pitch, accent, tone, and so on. This broad exposure enhances their robustness to different variations in audio signals. As a result, the representations learned by these PTMs are particularly effective for discerning audio deepfakes when compared to representations from other PTMs.

3 Pre-Trained Models

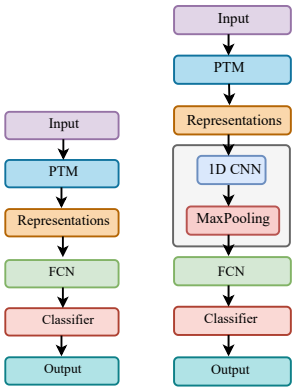
We compile the top-performing PTMs for our experiments. For multilingual PTMs we choose, **XLS-R** (Babu et al., 2022), **Whisper** (Radford et al., 2023), and **Massively Multilingual Speech (MMS)** (Pratap et al., 2023). XLS-R was pretrained on 128 languages while Whisper on 96 and MMS over 1400 languages. Whisper improves over XLS-R in various downstream speech processing tasks while MMS improves over Whisper. We selected the monolingual PTMs (**WavLM**, **Unispeech-SAT**, **Wav2vec2**) based on the SUPERB (wen Yang et al., 2021). WavLM and Unispeech-SAT have shown SOTA performance on SUPERB so we choose them. Wav2vec2 (Baevski et al., 2020) has not shown top performance like WavLM (Chen et al., 2022a) and Unispeech-SAT (Chen et al., 2022b) on SUPERB, however, as previous works have shown its efficacy for ADD (Zhang et al., 2023; Cai et al., 2023), so we selected it.

Additionally, models pre-trained for more specific tasks such as speaker recognition PTM (Ma et al., 2023) and models trained for emotion recognition (Conti et al., 2022) show exceptional performance for ADD, so we included them in our experiments. As speaker recognition PTM, we consider, **x-vector** (Snyder et al., 2018) and as emotion recognition PTM we use, **XLSR_emo** (Cahyawijaya et al., 2023). Additional details regarding these selected PTMs are available in Appendix 9.2.

4 Modeling

As we are evaluating how representations of different PTMs will behave for ADD, we keep the PTM layers frozen and keep the downstream modeling as simple as possible. We experimented with two modeling approaches (see Figure 1a and 1b). For the first approach (FCN), we employ an FCN on the extracted PTM representations and for the second (CNN), we use a 1D-CNN layer on top of representations followed by a Maxpooling layer and FCN. Softmax is used as the activation function in the output layer of the models which gives output as probabilities.

Merge into One: For fusing representations of



(a) FCN (b) CNN

Figure 1: Modeling Approaches

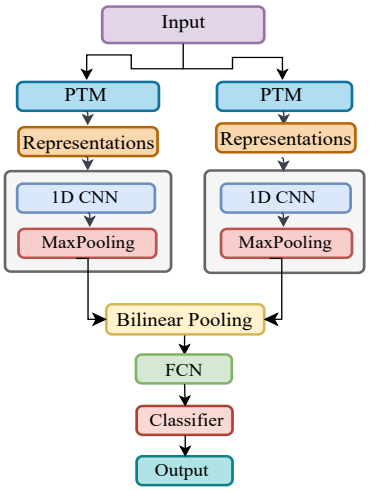


Figure 2: Proposed Modeling Framework for Fusion of PTM Representations, **MiO**

different PTMs, we propose **MiO**. The model architecture is shown in Figure 2. Here, we follow the same modeling pattern for each incoming representation as the second approach mentioned above. Then we apply linear projection to a dimension of size 120 followed by bilinear pooling (BP), which allows effective interaction between the features as shown by (Kumar and Nandakumar, 2022). BP is the outer product of two vectors p and q of dimension (D,1) and (D,1) such that the resultant will be a matrix of dimension (D, D) and it is given as:

$$\mathbf{BP_{D,D}} = \mathbf{p_{D,1}} \otimes \mathbf{q_{D,1}} = \mathbf{pq}^T \quad (1)$$

Linear Projection to lower dimension is carried out for computational resource constraints as the resultant of BP results in a matrix of much bigger shape. The resultant of BP is flattened and passed through an FCN.

We keep the number of epochs at 20 and the batch size of 32 for all the modeling approaches.

PTM	ASV	ITW	D-C	D-E
XLS-R	**1.67**	**0.24**	**1.42**	**0.12**
Whisper	**2.34**	**0.73**	**0.69**	**0.11**
MMS	**3.10**	**0.31**	**1.11**	**0.19**
Unispeech-SAT	9.97	2.36	2.14	0.54
WavLM (Base)	10.46	8.48	4.22	0.57
WavLM (Large)	10.23	8.41	4.20	0.60
Wav2Vec2	12.45	16.54	6.95	1.10
x-vector	9.49	0.98	2.65	0.66
XLSR_emo	9.36	2.70	3.11	0.18

Table 1: EER (%) scores for FCN models with different PTM representations; D-C, D-E represents Chinese and English Set of DECRO respectively

PTM	ASV	ITW	D-C	D-E
XLS-R	**1.03**	**0.17**	**1.42**	**0.07**
Whisper	**2.02**	**0.22**	**0.58**	**0.06**
MMS	**1.50**	**0.20**	**0.70**	**0.09**
Unispeech-SAT	9.76	2.20	1.89	0.42
WavLM (Base)	9.90	8.31	4.21	0.44
WavLM (Large)	9.30	7.60	3.14	0.45
Wav2Vec2	10.33	14.77	6.66	1.05
x-vector	8.61	0.27	2.62	0.92
XLSR_emo	9.20	1.57	2.83	0.10

Table 2: EER (%) scores for CNN models with different PTM representations; D-C, D-E represents Chinese and English Set of DECRO respectively

We use Cross-entropy as the loss function and Adam as the optimizer. We use *Tensorflow* library for our experiments. For reproducing our experiments, we will make our codebase available here[1].

5 Experiments

5.1 Benchmark Datasets

We selected three benchmark datasets for our experiments. They are **ASVSpoof 2019 (ASV)** (Wang et al., 2020), **In-the-Wild Dataset (ITW)** (Müller et al., 2022), and **DECRO** (Ba et al., 2023). Details regarding the datasets, data preprocessing, and experimental setting is provided in Appendix 9.1.

5.2 Experimental Results

For ASV and DECRO, the results presented are on the official testing split, and for ITW, the scores are the average across five splitting seeds as no official split was given. On DECRO, we built and trained models individually on the Chinese and English set of DECRO. We use D-C and D-E as notations for DECRO Chinese and English sets respectively.

Table 1 and 2 presents the EER scores for FCN and CNN models for different PTM representations as input features. Models trained on multilingual PTMs (XLS-R, Whisper, MMS) representations

[1] https://github.com/orchidchetiaphukan/MultilingualPTM_ADD_NAACL24

performed the best with lowest EERs in comparison with other PTMs. Within the multilingual PTMs, XLS-R achieves the lowest EERs in ASV and ITW while Whisper representations report the lowest in DECRO. MMS showed mixed performance achieving the second lowest EER in ITW and D-C with FCN, while with CNN it got the second least EER in ASV, ITW, and D-C. *This validates our hypothesis that multilingual PTMs will be more effective for ADD* due to their pre-training on extensive and varied multilingual datasets. As a result, they acquire information on a wide range of pitch, accents, and tones during the pre-training phase. This acquisition of diverse knowledge enhances their ability to effectively recognize and identify variations. Overall, CNN models showed superior performance to FCN models due to their ability to capture further important features.

We experimented with both WavLM (Base) and WavLM (Large) as WavLM (Large) holds the top position in SUPERB and also to see if the version with more parameters comes with the benefit of increased ADD performance. WavLM (Large) performs better than WavLM (Base) in some instances which might be due to its larger size. However, Unispeech-SAT has superior performance compared to WavLM (Large) while having the same number of parameters as WavLM (Base) and the best among the monolingual PTMs, Unispeech-SAT achieved the lowest EER in most instances with both FCN (9.97%, 2.36%, 2.14%, 0.54% in ASV, ITW, D-C, D-E respectively) and CNN (2.2%, 1.89%, 0.42% in ITW, D-C, D-E respectively). This can be attributed to speaker-aware pre-training of Unispeech-SAT, which leads to capturing various speech attributes such as pitch, accent, etc, more effectively and that helps in identifying deepfakes with more efficacy than its other monolingual counterparts. We can also see that representations of x-vector and XLSR_emo are performing better in some instances than the monolingual PTMs as they are trained on more specific non-semantic tasks leading to capture speech attributes far better than the monolingual PTMs for ADD. Wav2vec2 performed the worst among all the PTMs considered in our study showing its ineffectiveness in capturing attributes important for segregrating fake from real audio. Visualization of the representational space from the PTMs last hidden state are shown in Appendix (See Figure 4, 5, 6, 7 for ASV, ITW, D-C, and D-E respectively). We observe better clustering across the classes (real/fake) for representations from multilingual PTMs.

PTM Combinations	ASV	ITW	D-C	D-E
XLS-R + Whisper	0.95	0.27	1.08	**0.05**
XLS-R + MMS	0.56	0.29	1.62	**0.06**
XLS-R + Unispeech-SAT	**0.45**	**0.11**	1.03	0.13
XLS-R + WavLM (Base)	0.82	0.16	1.36	0.14
XLS-R + WavLM (Large)	0.72	0.14	1.16	0.12
XLS-R + Wav2Vec2	1.06	**0.12**	1.80	0.11
XLS-R + x-vector	**0.41**	**0.07**	1.63	0.46
XLS-R + XLSR_emo	1.35	0.21	1.60	0.12
Whisper + MMS	2.24	0.15	**0.27**	**0.08**
Whisper + Unispeech-SAT	2.16	1.03	0.46	0.28
Whisper + WavLM (Base)	1.90	0.97	2.95	0.15
Whisper + WavLM (Large)	1.95	0.91	2.10	0.13
Whisper + Wav2Vec2	2.19	1.08	0.98	0.65
Whisper + x-vector	3.30	0.22	0.91	0.32
Whisper + XLSR_emo	1.81	0.63	0.88	0.21
MMS + Unispeech-SAT	4.44	0.17	**0.19**	0.24
MMS + WavLM (Base)	0.99	3.50	0.21	0.25
MMS + WavLM (Large)	1.00	3.10	**0.15**	0.22
MMS + Wav2Vec2	**0.50**	0.22	0.39	0.33
MMS + x-vector	5.40	0.14	0.77	0.25
MMS + XLSR_emo	1.80	0.36	0.81	0.32
Unispeech-SAT + WavLM (Base)	10.18	2.79	2.31	0.48
Unispeech-SAT + WavLM (Large)	9.19	2.99	2.11	0.41
Unispeech-SAT + Wav2Vec2	9.74	2.55	2.88	0.59
Unispeech-SAT + x-vector	5.82	0.15	2.56	0.54
Unispeech-SAT + XLSR_emo	6.80	1.70	2.09	0.57
WavLM (Base) + Wav2Vec2	12.46	8.54	1.93	0.51
WavLM (Base) + x-vector	6.03	0.19	3.04	0.61
WavLM (Base) + XLSR_emo	7.91	2.31	2.64	0.21
WavLM (Large) + Wav2Vec2	11.36	7.14	1.44	0.54
WavLM (Large) + x-vector	5.01	0.19	2.21	0.60
WavLM (Large) + XLSR_emo	6.92	2.20	2.01	0.18
Wav2Vec2 + x-vector	7.31	0.26	3.51	0.47
Wav2Vec2 + XLSR_emo	7.50	1.91	2.87	0.32
x-vector + XLSR_emo	7.89	0.43	1.11	0.71

Table 3: EER (%) scores for different PTM representations combinations with **MiO**

Table 3 shows the EER scores with combined representations of different PTMs. With the fusion of XLS-R and x-vector representations, we got the lowest EER score in ASV and ITW which shows that combining speaker-specific informative features leads to further gain in performance and these representations are acting as complementary to each other. In D-C, the fusion of MMS and WavLM (Large), and in D-E, XLS-R and Whisper pair, reported the lowest EER, which shows that these multilingual PTM's representations are showing additive behavior, leading to further lowering of EER. However, in some, instances the fusion of certain PTM representations leads to degradation of performance compared to its individual performance, such as the combination of XLS-R and XLSR_emo gave 1.35% and 0.21% EER in ASV, ITW respectively which is lower than individual EER of XLSR_emo, but higher than XLS-R (1.03% in ASV). This can be depicted as contradictory behavior shown by the representations. As additional experiments, we carried out a cross-corpus evaluation (see Tables 5 and 6 in Appendix). We found that models trained on multilingual PTM representations, generalize better in cross-corpus

evaluation.

Dataset	Model	EER (%)
ASV	CQT-DCT-LCNN (Lavrentyeva et al., 2019)	**1.84**
	MiO(XLS-R + x-vector)	**0.41**
ITW	STATNet (Ranjan et al., 2022)	**0.20**
	MiO(XLS-R + x-vector)	**0.07**
D-E	Res-TSSDNet (Ba et al., 2023)	**0.02**
	MiO(XLS-R + Whisper)	**0.04**

Table 4: Comparison with SOTA on ASV, ITW, and D-E in terms of EER(%); **MiO(XLS-R + x-vector)**, **MiO(XLS-R + Whisper)** represents the proposed methodology **MiO** with combination of XLS-R, x-vector and XLS-R, Whisper representations

5.3 Comparison with State-of-the-art

We compare the proposed approach, **MiO** with previous SOTA works on respective datasets. Table 4, presents the comparison with SOTA studies on ASV, ITW, and D-E respectively. D-C was used as a testing set for evaluating the transferability of ADD systems from English to Chinese by Ba et al. 2023, so previous works trained on D-C and evaluated on D-C are not present. In ASV and ITW, we report the lowest EER compared to existing SOTA works, and for D-E, we report competitive performance in comparison to existing SOTA work.

6 Conclusion

In this work, we validated our hypothesis that multilingual PTMs pre-trained on large diverse multilingual data will be more effective for ADD as they learn diverse pitches, accents, and tones during their pretraining phase and are more robust to variations. We carried out a comprehensive empirical analysis by extracting representations from eight PTMs and our findings show that representations from multilingual showed the lowest EER on three benchmark datasets ASV, ITW, and DECRO. Also, we found that fusion of representations from PTMs lead to a further drop in EER and for this, we proposed, **MiO**. We report SOTA performance in ASV, ITW and competitive performance in DECRO in comparison to previous SOTA works with **MiO**.

7 Limitations

We have considered only eight PTMs and this may limit our findings, so in the future, we will consider more relevant PTMs. Also, results varies with different downstream networks as shown by (Zaiem et al., 2023) and we only experimented with two downstream networks. So, we will extend this by evaluating more downstream networks. Also, we will also look into why certain PTMs representations combinations works better than others.

8 Ethics Statement

Deepfakes' have a significant impact on the privacy and integrity of individuals. It is important to address the ethical implications of research conducted on Deepfakes. This work ensures that privacy and integrity of specific individuals or organizations are not revealed and is not affected. The data used for research in this work is collected from publicly available datasets and are anonymized. The experimental results and interpretations also do not have any ethical implications.

References

Moustafa Alzantot, Ziqi Wang, and Mani B. Srivastava. 2019. Deep Residual Neural Networks for Audio Spoofing Detection. In *Proc. Interspeech 2019*, pages 1078–1082.

A Arunkumar, Vrunda Nileshkumar Sukhadia, and Srinivasan Umesh. 2022. Investigation of Ensemble features of Self-Supervised Pretrained Models for Automatic Speech Recognition. In *Proc. Interspeech 2022*, pages 5145–5149.

Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. 2023. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*, pages 2033–2044.

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 *Proc. 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.

Samuel Cahyawijaya, Holy Lovenia, Willy Chung, Rita Frieske, Zihan Liu, and Pascale Fung. 2023. Cross-Lingual Cross-Age Adaptation for Low-Resource Elderly Speech Emotion Recognition. In *Proc. INTERSPEECH 2023*, pages 3352–3356.

Zexin Cai, Weiqing Wang, and Ming Li. 2023. Waveform boundary detection for partially spoofed audio.

In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li, et al. 2022b. 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)*, pages 6152–6156. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hosler, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C Stamm, and Stefano Tubaro. 2022. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8962–8966. IEEE.

Youngsik Eom, Yeonghyeon Lee, Ji Sub Um, and Hoirin Kim. 2022. Anti-spoofing using transfer learning with variational information bottleneck. *arXiv preprint arXiv:2204.01387*.

Alejandro Gomez-Alanis, Antonio M. Peinado, Jose A. Gonzalez, and Angel M. Gomez. 2019. A Light Convolutional GRU-RNN Deep Feature Extractor for ASV Spoofing Detection. In *Proc. Interspeech 2019*, pages 1068–1072.

Cemal Hanilçi, Tomi Kinnunen, Md. Sahidullah, and Aleksandr Sizov. 2015. Classifiers for synthetic speech detection: a comparison. In *Proc. Interspeech 2015*, pages 2057–2061.

Ziyue Jiang, Hongcheng Zhu, Li Peng, Wenbing Ding, and Yanzhen Ren. 2020. Self-Supervised Spoofing Audio Detection Scheme. In *Proc. Interspeech 2020*, pages 4223–4227.

Hye jin Shim, Hee-Soo Heo, Jee weon Jung, and Ha-Jin Yu. 2020. Self-Supervised Pre-Training with Acoustic Configurations for Replay Spoofing Detection. In *Proc. Interspeech 2020*, pages 1091–1095.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183.

Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. 2019. Stc antispoofing systems for the asvspoof2019 challenge. *arXiv preprint arXiv:1904.05576*.

Yerin Lee, Narin Kim, Jaehong Jeong, and Il-Youp Kwak. 2023. Experimental case study of self-supervised learning for voice spoofing detection. *IEEE Access*, 11:24216–24226.

Anwei Luo, Enlei Li, Yongliang Liu, Xiangui Kang, and Z Jane Wang. 2021. A capsule network based approach for detection of audio spoofing attacks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6359–6363. IEEE.

Xinyue Ma, Tianyu Liang, Shanshan Zhang, Shen Huang, and Liang He. 2021. Improved lightcnn with attention modules for asv spoofing detection. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Xinyue Ma, Shanshan Zhang, Shen Huang, Ji Gao, Ying Hu, and Liang He. 2023. How to boost anti-spoofing with x-vectors. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 593–598. IEEE.

Nicolas Müller, Pavel Czempin, Franziska Diekmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does Audio Deepfake Detection Generalize? In *Proc. Interspeech 2022*, pages 2783–2787.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Yanmin Qian, Nanxin Chen, and Kai Yu. 2016. Deep features for automatic spoofing detection. *Speech Communication*, 85:43–52.

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.

Rishabh Ranjan, Mayank Vatsa, and Richa Singh. 2022. Statnet: Spectral and temporal features based multi-task network for audio spoofing detection. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. 2015. A comparison of features for synthetic speech detection. In *Proc. Interspeech 2015*, pages 2087–2091.

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. IEEE.

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

Francis Tom, Mohit Jain, and Prasenjit Dey. 2018. End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention. In *Proc. Interspeech 2018*, pages 681–685.

Chenglong Wang, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, Shuai Zhang, and Xun Chen. 2023. Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features. In *Proc. INTERSPEECH 2023*, pages 3844–3848.

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.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth annual conference of the international speech communication association*.

Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, and Mirco Ravanelli. 2023. Speech Self-Supervised Representation Benchmarking: Are We Doing it Right? In *Proc. INTERSPEECH 2023*, pages 2873–2877.

Jiachen Zhang, Guoqing Tu, Shubo Liu, and Zhaohui Cai. 2023. Audio anti-spoofing based on audio feature fusion. *Algorithms*, 16(7):317.

9 Appendix

9.1 Dataset

Additional information regarding the benchmark datasets considered in our study is given as follows: **ASVSpoof 2019**[2] : Voice Cloning Toolkit (VCTK), a multispeaker English corpus is used as a base database. It contains audio clips from 107 (46 male, 61 female) speakers. Various spoofing algorithms were employed to create counterfeit versions of the authentic clips. These algorithms include SOTA text-to-speech synthesis techniques as well as different voice conversion methods. The availability of large-scale labeled audio recordings makes ASV a valuable resource for training and testing ML models to identify fake audio. We use the Logical Access (LA) database from ASV. We train, validate, and evaluate the models on the official split given by Wang et al. 2020.

In-the-Wild Dataset[3]: This dataset comprises 37.9 hours of audio content and features English-speaking celebrities and politicians. It encompasses fake audio encountered in various real-life scenarios, presenting a challenge in distinguishing it from genuine recordings. The clips associated with the fake audio associated with a particular celebrity/politician are collected from openly available social media sites and video-sharing platforms. This dataset acts as an important resource for evaluating ADD systems on real-world data. For ITW, there is no official split given so we split the dataset as 70% as training, 10% as validation, and 20% as test set.

DECRO[4]: ADD models trained on one language fail when evaluated in zero-shot format in some other language. So as to make up for this, Ba et al. 2023 proposed the DECRO dataset for the evaluation of ADD systems in a cross-lingual manner. It comprises fake and real audio clips in English (DECRO-E) and Chinese (DECRO-C). We use the official split given by Ba et al. 2023 for training, validation, and evaluation.

The distribution of the audio clips with real and fake labels for each database is shown in Figure 3.

[2] https://www.asvspoof.org/index2019.html

[3] https://deepfake-demo.aisec.fraunhofer.de/in_the_wild

[4] https://zenodo.org/records/7603208

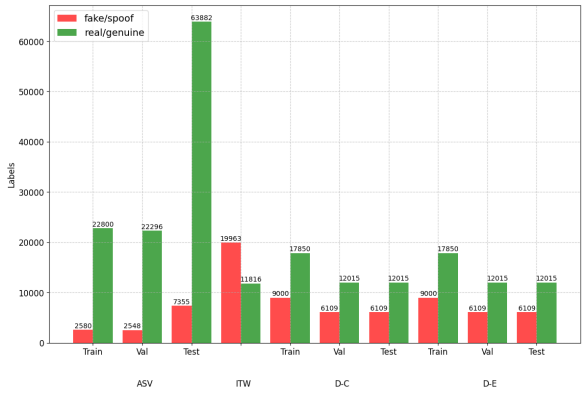


Figure 3: Class-wise data distribution across the datasets; D-C and D-E represents DECRO Chinese and English Set

9.2 Detailed Information of the Pre-Trained Models

Here, we describe various PTMs considered for our work. They are as follows:

XLS-R: It is multilingual representation learning model based on Wav2vec2 architecture. Training is carried out in a self-supervised manner and the objective involves solving a contrastive task over masked feature encoder outputs. It is trained on 436k hours of speech data comprising corpora VoxPopuli, Multilingual Librispeech, CommonVoice, VoxLingua107, and BABEL. XLS-R improves over XLS-R-53 pre-trained on 53 languages for various downstream speech processing tasks. We use the base[5] version comprising of 1 billion parameters for our work.

Whisper: Pre-training is carried out on 680k hours of data, incorporating a multitask format in a weakly-supervised way. It is an encoder-decoder-based modeling architecture. It is trained with the primary purpose of predicting transcriptions of audio content available on the internet. Whisper shows improved performance on speech recognition over XLS-R. We exploit the base[6] version with 74 million parameters for our use case.

Massively Multilingual Speech (MMS): It is pre-trained on over 500k hours of speech data. It is built upon the Wav2vec2 model architecture that consists of a convolutional encoder followed by a BERT-like transformer block. MMS also follows contrastive pre-training as Wav2vec2. Pre-training data consists of various datasets such as MMS-lab, FLEURS, BABEL, etc. We use the 1 billion parameters version[7] openly available.

Unispeech-SAT: It utilizes a contrastive loss model in conjunction with multitask learning. During its pre-training, UniSpeech-SAT follows a speaker-aware format. It is pre-trained on 960 hours of Librispeech English speech data. We make use of the base[8] version consisting of 94.68 million parameters.

WavLM: In its pre-training phase, WavLM simultaneously learns to predict masked speech and perform denoising. This dual process equips WavLM to effectively handle complex aspects of speech data, including speaker identity and spoken content, among others. WavLM (base)[9] and WavLM (Large)[10] versions are used for our experiments with 94.70 million and 316.62 million parameters respectively. WavLM (Base) and WavLM (Large) were pre-trained on 960 hours of Librispeech English data and Mix 94k data respectively.

Wav2vec2: During training, Wav2vec2 masks the speech input in the latent space and completes a contrastive task defined over a quantization of the jointly learned latent representations. It was pre-trained on 960 hours of Librispeech English data. We choose the base[11] version with 95.04 million parameters and containing 12 transformer encoder blocks.

x-vector[12]: We took x-vector from *Speechbrain* (Ravanelli et al., 2021) library. x-vector is a time-delay neural network trained in a supervised fashion for speaker recognition and achieves higher performance in comparison with the previous SOTA speaker recognition system, i-vector. It is trained on the combination of training data of Voxceleb1 and VoxCeleb2 with approx 4.2 million parameters.

XLSR_emo[13]: It is an XLS-R-53 (Conneau et al., 2021) model fine-tuned on training sets of various English and Chinese speech-emotion recognition databases such as CREMA-D, CSED, ElderReact, ESD, IEMOCAP, and TESS.

The input audio is sampled to 16KHz before passing as input to the PTMs. We extract the

[5] https://huggingface.co/facebook/wav2vec2-xls-r-1b

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

[7] https://huggingface.co/facebook/mms-1b

[8] https://huggingface.co/microsoft/unispeech-sat-base

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

[10] https://huggingface.co/microsoft/wavlm-large

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

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

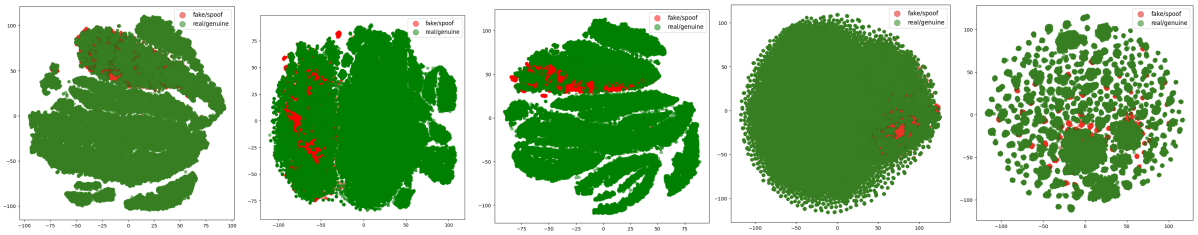
[13] https://huggingface.co/CAiRE/SER-wav2vec2-large-xlsr-53-eng-zho-all-age

last hidden states from XLS-R, MMS, Unispeech-SAT, WavLM (Base), WavLM (Large), Wav2vec2, XLSR_emo and convert the hidden states to vectors of dimensions 1280, 1280, 768, 768, 1024, 768, and 1024, respectively through the application of mean pooling. We discard the decoder for Whisper and extract the hidden representations from the encoder and through average pooling, we convert the representations to a vector of 512-dimension. Similarly, for the x-vector, we extract representations as vector size of 512-dimension.

9.3 Cross-Corpus Evaluation

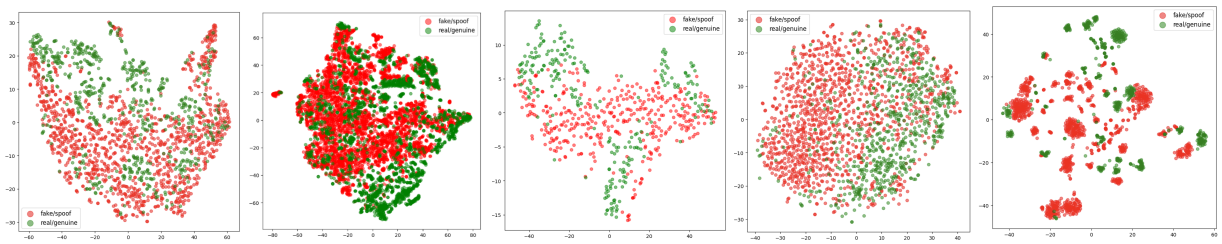
We also investigate the cross-corpus generalization capability of the models trained on multilingual PTM representations as it has been shown in the literature of ADD (Ba et al., 2023; Müller et al., 2022) that models trained in one dataset or a certain language fail to perform in others. We use the same modeling approach as Figure 1b. As the representations from different PTMs are of different dimensions, we use Principal Component Analysis (PCA) to transform the representations to the same dimension. We set the final dimension size after PCA to 120 and 240. We train the models on one training set of one dataset and evaluate on the test set of the others. We keep the training details like number of epochs, batch size, etc same as in Section 4. We compare the multilingual PTMs with a monolingual PTM (Wav2vec2) and also speaker recognition PTM (x-vector) which reported competitive results (Table 2) for better understanding of their generalization abilities.

The results of our experiments are presented in Table 5 and 6. Models trained with multilingual PTM representations performed the best and this shows their cross-corpus generalization abilities. However, the multilingual PTMs shows fluctuating performance among them, in some instances, representations from MMS performed the best, such as in Table 5 when trained on ASV and D-C Training set, whereas we achieve competitive results with XLS-R when trained on ITW and D-E and tested on the others. We notice significant differences in the results obtained across 120 and 240-dimension sizes. This points out that the dimension size of the representations also plays a minor role in the performance achieved in the downstream task. We also present cross-corpus evaluation scores for selected representations pairs with **MiO** in Table 7 and 8.



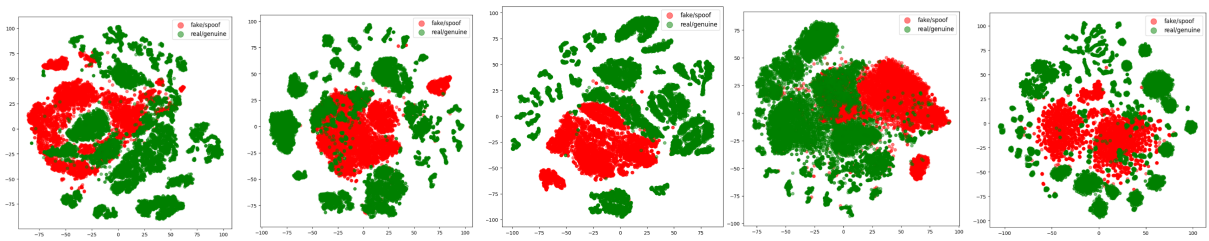
(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 4: t-SNE plots of different PTM representations on ASV



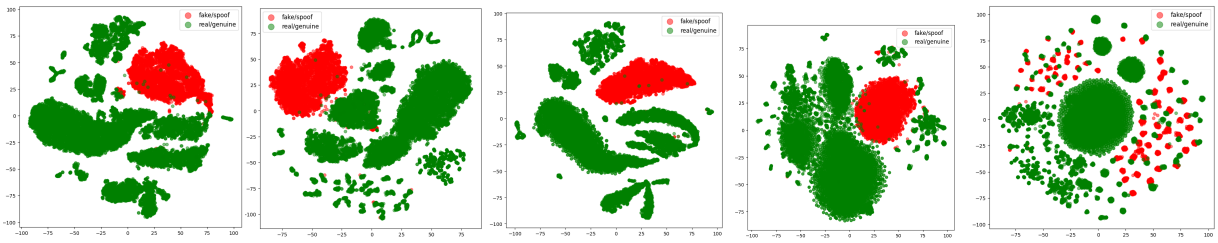
(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 5: t-SNE plots of different PTM representations on ITW



(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 6: t-SNE plots of different PTM representations on D-C



(a) XLS-R (b) Whisper (c) MMS (d) Unispeech-SAT (e) x-vector

Figure 7: t-SNE plots of different PTM representations on D-E

PTM	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLS-R	**22.78**	**12.21**	**28.11**	**10.23**	**19.78**	**15.32**	**10.45**	**18.88**	**21.03**	**11.06**	**39.58**	**14.94**
Whisper	**30.51**	**11.09**	**29.21**	**17.19**	**35.55**	**15.34**	**16.91**	**39.62**	**30.30**	**15.78**	**30.67**	**17.18**
MMS	**19.62**	**9.61**	**18.83**	**4.11**	**19.51**	**8.71**	27.50	50.63	**20.19**	**14.78**	**32.75**	**21.59**
Wav2Vec2	48.23	19.48	43.01	19.82	40.07	18.74	45.37	49.15	49.99	28.80	46.21	35.03
WavLM (Large)	31.10	13.19	32.12	31.23	36.69	15.99	28.13	**43.91**	35.93	16.61	40.60	24.61
x-vector	32.98	13.62	37.43	39.53	38.66	17.69	**25.43**	47.90	37.95	18.62	39.65	28.62

Table 5: Cross-Corpus Evaluation with representations of different PTMs kept at 120-dimension; Scores are in EER(%); ASV Training, D-C Training, ITW Training, D-E Training represents training dataset and evaluated on the other datasets; For ITW, we select a test set for one splitting seed

PTM	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLS-R	**21.33**	**12.78**	**29.28**	**12.14**	**23.22**	**21.88**	**8.21**	**17.69**	**29.88**	**21.99**	**11.93**	**14.65**
Whisper	**25.22**	**21.00**	**34.88**	**13.66**	**29.22**	**16.94**	**17.82**	**33.60**	**17.87**	**24.88**	**11.87**	**21.86**
MMS	**34.61**	**26.14**	**19.28**	**19.88**	**31.40**	**32.97**	**20.74**	42.11	**10.85**	**23.16**	**18.42**	**21.43**
Wav2Vec2	57.66	43.98	46.66	44.63	47.22	41.90	27.34	53.33	44.97	35.76	43.98	46.95
WavLM (Large)	35.19	30.31	42.20	34.99	33.21	34.28	21.11	**41.39**	30.83	29.51	33.47	35.11
x-vector	35.79	32.34	45.04	37.09	39.22	39.02	21.12	43.22	31.00	32.55	32.77	32.98

Table 6: Cross-Corpus Evaluation with representations of different PTMs kept at 240-dimension; Scores are in EER(%); ASV Training, D-C Training, ITW Training, D-E Training represents training dataset and evaluated on the other datasets; For ITW, we select a test set for one splitting seed

Model	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLS-R + x-vector	**21.11**	**12.04**	**24.80**	**16.53**	**23.34**	**16.14**	**58.78**	**49.84**	**69.14**	**21.34**	**35.72**	**48.15**
Whisper + Unispeech-SAT	**44.59**	**7.80**	**38.86**	**26.89**	**47.02**	**15.74**	**55.02**	**29.94**	**47.72**	**12.89**	**44.26**	**27.50**

Table 7: Cross-Corpus Evaluation with combined representations kept at 120-dimension; Scores are in EER(%); ASV Training, D-C Training, ITW Training, D-E Training represents training dataset and evaluated on the other datasets; For ITW, we select a test set for one splitting seed

Model	ASV Training			D-C Training			ITW Training			D-E Training		
	D-C	D-E	ITW	ASV	ITW	D-E	ASV	D-C	D-E	ASV	D-C	ITW
XLS-R + x-vector	**17.21**	**15.14**	**24.53**	**14.84**	**17.97**	**13.53**	**55.85**	**50.91**	**57.04**	**37.68**	**14.23**	**26.91**
Whisper + Unispeech-SAT	**31.21**	**15.70**	**41.56**	**24.61**	**46.94**	**18.52**	**54.80**	**39.57**	**46.75**	**42.57**	**18.34**	**31.09**

Table 8: Cross-Corpus Evaluation with combined representations kept at 240-dimension; Scores are in EER(%); ASV Training, D-C Training, ITW Training, D-E Training represents training dataset and evaluated on the other datasets; For ITW, we select a test set for one splitting seed