Rethinking Cross-Corpus Speech Emotion Recognition Benchmarking: Are Paralinguistic Pre-Trained Representations Sufficient?

Orchid Chetia Phukan[†*], Mohd Mujtaba Akhtar[†‡*], Girish[†§*], Swarup Ranjan Behera[¶*]
Parabattina Bhagath[‖], Pailla Balakrishna Reddy[**], Arun Balaji Buduru[†]
[†]IIIT-Delhi, India, [‡]V.B.S.P.U, India, [§]UPES, India, [¶]Independent Researcher, India
[‖]L.B.R College of Engineering, India, [**]Reliance AI, India
E-mail: orchidp@iiitd.ac.in

***Abstract*—Recent benchmarks evaluating pre-trained models (PTMs) for cross-corpus speech emotion recognition (SER) have overlooked PTM pre-trained for paralinguistic speech processing (PSP), raising concerns about their reliability, since SER is inherently a paralinguistic task. We hypothesize that PSP-focused PTM will perform better in cross-corpus SER settings. To test this, we analyze state-of-the-art PTMs representations including paralinguistic, monolingual, multilingual, and speaker recognition. Our results confirm that TRILLsson (a paralinguistic PTM) outperforms others, reinforcing the need to consider PSP-focused PTMs in cross-corpus SER benchmarks. This study enhances benchmark trustworthiness and guides PTMs evaluations for reliable cross-corpus SER.**

I. Introduction

Emotions significantly influence human behavior and interactions, and Speech Emotion Recognition (SER) identifies these cues by analyzing speech features such as pitch, tone, and intensity. This capability is valuable across many fields: it enables empathetic responses in human-computer interaction, supports mental health monitoring, enhances customer service through personalized interactions, and adapts learning environments in education. Additionally, in entertainment and security, SER improves user experience and safety by identifying emotional distress or threats. Thus, accurately recognizing emotions in speech is key for building intelligent, responsive systems.

Early research in SER predominantly relied on handcrafted acoustic features, such as MFCCs, paired with traditional machine learning models [1]–[3]. With the advent of speech pre-trained models (PTMs), SER has seen significant advancements, as these models provide powerful, generalizable representations learned from large-scale, diverse datasets [4]–[6]. This shift has alleviated both performance benefit and need for training models from scratch. PTMs, however, differ in various aspects, including their training data, which spans across diverse data distributions, and whether they are trained on monolingual or multilingual datasets. Additionally, these models vary in terms of architecture and the pre-training strategies used, such as self-supervised or supervised learning approaches. These differences in PTM nature and pre-training methods have direct implications on their downstream performance for SER. As a result, the choice of PTM can significantly affect the accuracy and robustness of SER and this variability in performance underscores the need for a deeper understanding of these PTMs. Several benchmarks have been proposed such as SUPERB [7], EMO-SUPERB [8], OPEN-EMOTION [9] and so on to better understand the performance of different PTMs for SER in monolingual or multilingual settings. These benchmarks also act as a reference for future research for selection of PTMs depending on their use case. However, these benchmarks evaluates PTMs for training and testing on the same corpus. As such there is a recent ongoing interest in the research community to access the cross-corpus SER capability of various SOTA speech PTMs [10], [11]. Here, "cross-corpus" encompasses two key scenarios: (1) the same language with varying data distributions and (2) cross-lingual settings. Cross-corpus SER presents greater challenges compared to same-corpus SER due to the domain shift that occurs between the training and test data. In such scenarios, the models must generalize across different data distributions, speaker demographics, recording environments, or even languages, which can lead to significant performance degradation. The variations in emotional expression, speaking styles, and acoustic conditions across corpora further complicate the task, making cross-corpus SER more demanding.

In response, various benchmarks have been proposed in recent couple of years such as SER-evals [12] and EmoBox [13]. However, these benchmarks haven't considered representations from PTM pre-trained primarily for paralinguistic speech processing (PSP). Such oversight raises concerns about the trustworthiness of the benchmarks as comprehensive references for future research, especially since SER fundamentally is a PSP task. Also, Phukan et al. [14] which has shown the topmost performance of paralinguistic PTM representations for SER in multiple languages, haven't evaluated paralinguistic PTM representations for cross-corpus SER. So, to solve this research gap and also to get better understanding, we explore paralinguistic PTM representations for cross-corpus SER. *We hypothesize that representations from paralinguistic PTM representations are better suited for cross-corpus SER. Unlike*

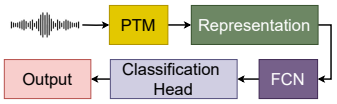
* Contributed equally as first authors.

979-8-3315-7206-8/25/$31.00

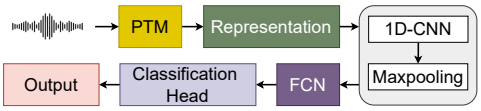
general-purpose PTM representations, paralinguistic PTM representations captures speech characteristics cues such as intonation, pitch, rhythm, and prosody—features that are directly relevant for SER. These paralinguistic representations transcend linguistic boundaries and can generalize more effectively across different languages and data distributions. To validate our hypothesis, we perform a comprehensive comparative study of representations from SOTA PTMs comprising paralinguistic, monolingual, multilingual as well as speaker recognition. These PTMs are SOTA in their respective benchmarks. For example, Whisper reported the topmost performance for cross-corpus SER in SER-evals [12] and EmoBox [13]. So, we are presenting the comparison of best of the best PTMs in our study to validate the performance and show the capability of paralinguistic PTM for cross-corpus SER.

Key contributions of the paper are as follows:

- We present a comprehensive comparative study of representations from SOTA PTMs including paralinguistic (TRILLsson), monolingual (WavLM, Unispeech-SAT, Wav2vec2), multilingual (XLS-R, Whisper, MMS) and speaker recognition (x-vector) for cross-corpus SER. We experiment with CREMA-D (*English*), RAVDESS (*English*), emo-DB (*English*), MESD (*Mexican Spanish*), and AESDD (*Greek*) benchmark SER datasets.
- Our findings demonstrate TRILLsson (paralinguistic PTM) representations superior performance in cross-corpus SER across all the datasets.
- Our study will act as benchmark for future researchers for selection of PTM representations for cross-corpus SER. Also, our study calls for the inclusion of paralinguistic PTM representations in the previous benchmarks for cross-corpus SER. This will ensure reliability and trustworthiness of cross-corpus SER benchmarks, ensuring that future work can build upon more accurate and generalizable results.



(a) Fully Connected Network (FCN)



(b) Convolutional Neural Network (CNN)

Fig. 1: Modeling

II. Pre-Trained Models

In this section, we discuss the PTMs considered in our study. As monolingual PTMs, we use **WavLM**[1] [15], **Unispeech-SAT**[2] [16], and **Wav2vec2**[3] [17]. WavLM is a SOTA PTM in SUPERB and shows top performance compared to other PTMs in various speech processing tasks and trained to perform speech denoising and masked modeling together. Unispeech-SAT also shows SOTA performance in SUPERB and was trained in speaker-aware format. Wav2vec2 is not SOTA like WavLM and Unispeech-SAT in SUPERB, however, we consider it, as it shows relatably good performance in SER as shown by previous research [18]. We use the base versions for WavLM, Unispeech-SAT, Wav2vec2 trained on librispeech 960 hours data with 94.70M, 94.70M, and 95.04M parameters respectively. As multilingual PTMs, we consider, **XLS-R**[4] [19], **Whisper**[5] [20], and **MMS**[6] [21]. XLS-R was trained on 128 languages while Whisper was trained on 96 languages. XLS-R is based on Wav2vec2 architecture and trained in a self-supervised fashion and in contrast, Whisper is based on vanilla transformer encoder-decoder architecture. We use XLS-R 300M parameters version and for whisper,we use the base variant with 74M. MMS is built on top of Wav2vec2 architecture and improves over Whisper in multilingual speech processing applications. It extends its pre-training to almost over 1400 languages. We use the 1B variant of MMS in our experiments. For paralinguistic PTM, we consider **TRILLsson**[7] [22] and it is derived from SOTA paralinguistic Conformer (CAP12) through knowledge distillation. TRILLsson is open-sourced, however, CAP12 is not. It has demonstrated SOTA performance in different paralingusitic tasks such as SER, speaker identification, and so on in NOSS benchmark. It was trained on the speech samples of Audioset and Libri-light datasets during its distillation phase. We use the 63.4 million parameters variant of TRILLsson. We also consider **x-vector**[8] [23], a time delay-neural network trained for speaker recognition. It is trained on the combination of Voxceleb1 + Voxceleb2. x-vector has shown its potential for SER [24], so we included it in our study.

For each frozen PTM, we extract the last hidden states through average pooling. We get representations of 768 from WavLM, Unispeech-SAT, wav2vec2, 1024 from TRILLsson, 1280 from XLS-R, MMS. For whisper, we discard the decoder and extract representations from its encoder with size 512 same as x-vector. We resample the audios to 16 KHz before passing it to the PTMs.

III. Experiments

A. Benchmark Datasets

Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [25]: It serves as a widely-used benchmark for SER and it is gender-balanced. It includes recordings in english from 48 male and 43 female actors, totaling 7,442 utterances. The dataset is valuable due to its representation of a variety of

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

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

[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/facebook/mms-1b

[7]https://tfhub.dev/google/nonsemantic-speech-benchmark/trillsson4/1

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

P	C(Train) - C(Test)		M(Train) - M(Test)		E(Train) - E(Test)		A(Train) - A(Test)		R(Train) - R(Test)	
	FCN	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**
MM	76.00 | 75.82	76.48 | 76.37	76.39 | 76.19	87.50 | 87.55	86.95 | 83.45	93.33 | 92.76	24.15 | 14.58	25.62 | 20.54	61.98 | 60.84	77.08 | 76.79
Wh	71.05 | 70.79	71.13 | 70.80	63.89 | 64.40	68.06 | 68.14	86.67 | 86.69	89.33 | 88.74	53.22 | 48.33	54.55 | 49.62	75.52 | 75.69	79.69 | 79.71
W	70.02 | 68.97	71.36 | 71.21	50.00 | 50.12	51.39 | 51.15	89.33 | 86.68	94.67 | 94.88	81.65 | 75.16	82.64 | 82.26	72.92 | 73.08	80.73 | 80.72
X	73.09 | 72.99	74.67 | 74.57	74.31 | 74.43	77.78 | 77.30	78.67 | 75.66	92.00 | 92.37	39.74 | 35.21	44.63 | 41.80	64.58 | 63.67	82.81 | 82.64
W2	67.19 | 67.11	69.16 | 68.86	62.50 | 62.50	63.19 | 63.57	86.67 | 87.57	96.00 | 95.00	68.35 | 61.25	70.25 | 68.64	71.88 | 71.93	71.88 | 72.09
U	69.08 | 68.74	69.24 | 69.00	36.11 | 35.17	40.97 | 40.51	76.00 | 77.71	80.00 | 70.60	53.72 | 52.48	64.46 | 64.21	68.75 | 68.69	76.56 | 76.47
T	**79.26 | 79.23**	**79.70 | 79.63**	**85.42 | 85.30**	**88.89 | 88.85**	**97.56 | 94.65**	**98.67 | 98.13**	**91.74 | 91.57**	**93.39 | 93.35**	**95.33 | 95.31**	**96.36 | 96.35**
XV	67.66 | 67.44	68.06 | 68.04	77.08 | 77.00	81.94 | 81.92	91.00 | 90.67	92.36 | 92.00	77.54 | 71.65	81.82 | 80.20	81.25 | 81.19	87.50 | 87.48
	C(Train) - M(Test)		**M(Train) - C(Test)**		**E(Train) - M(Test)**		**A(Train) - M(Test)**		**R(Train) - M(Test)**	
	FCN	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**
MM	30.56 | 25.54	32.64 | 28.51	19.43 | 9.33	20.77 | 10.48	30.56 | 25.06	31.94 | 29.57	17.89 | 6.51	19.44 | 13.56	22.92 | 11.22	29.17 | 16.49
Wh	40.28 | 38.94	42.36 | 41.08	18.93 | 12.94	21.24 | 13.54	25.69 | 20.69	38.61 | 32.53	31.25 | 22.47	31.94 | 24.67	24.31 | 13.46	28.47 | 17.39
W	25.69 | 19.79	30.56 | 24.90	20.40 | 9.55	26.59 | 21.13	28.47 | 25.98	34.72 | 33.24	29.17 | 26.93	36.81 | 35.88	30.56 | 23.33	31.25 | 29.94
X	35.42 | 31.89	40.28 | 39.68	27.14 | 19.51	30.45 | 24.66	21.53 | 16.71	28.47 | 28.86	18.74 | 10.26	20.83 | 12.69	22.22 | 12.54	23.61 | 16.29
W2	27.78 | 23.38	29.17 | 27.65	13.38 | 12.18	17.55 | 13.00	27.72 | 22.42	31.25 | 30.97	25.00 | 20.22	27.78 | 23.57	29.17 | 27.10	30.56 | 27.40
U	25.69 | 20.63	25.69 | 20.63	21.09 | 10.75	22.42 | 12.95	22.22 | 19.32	27.78 | 26.93	15.28 | 15.15	34.72 | 34.63	22.92 | 14.34	22.92 | 16.25
T	**45.58 | 44.12**	**51.67 | 47.55**	**44.78 | 37.54**	**49.66 | 41.74**	**50.14 | 44.92**	**55.56 | 46.56**	**49.86 | 43.19**	**53.33 | 49.72**	**49.34 | 42.84**	**54.03 | 49.67**
XV	22.22 | 21.55	30.56 | 30.19	32.10 | 27.75	35.09 | 31.52	27.78 | 25.31	29.17 | 27.89	25.14 | 22.37	27.08 | 24.88	34.72 | 34.17	38.89 | 38.12
	C(Train) - E(Test)		**M(Train) - E(Test)**		**E(Train) - C(Test)**		**A(Train) - E(Test)**		**R(Train) - E(Test)**	
	FCN	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**
MM	33.33 | 10.42	60.00 | 55.17	25.33 | 17.19	49.33 | 49.33	22.82 | 11.67	23.84 | 13.67	29.85 | 19.11	38.67 | 24.69	45.33 | 44.67	52.32 | 52.00
Wh	50.67 | 44.29	62.05 | 60.00	37.33 | 15.76	40.00 | 19.90	29.11 | 21.17	34.54 | 22.27	40.00 | 18.64	76.00 | 71.80	50.67 | 47.97	64.00 | 55.28
W2	34.71 | 26.40	49.33 | 45.10	50.67 | 41.19	52.00 | 47.72	33.52 | 28.31	37.45 | 36.37	57.33 | 58.58	72.00 | 69.57	17.33 | 15.84	20.00 | 16.18
X	72.00 | 68.11	74.67 | 69.71	38.67 | 26.19	40.00 | 26.20	25.10 | 15.27	26.07 | 15.27	38.57 | 25.69	42.78 | 30.97	35.74 | 38.67	40.00 | 41.04
W2	36.00 | 26.30	40.00 | 34.17	49.33 | 35.29	50.67 | 35.69	18.02 | 9.78	20.06 | 12.41	61.33 | 59.75	77.33 | 74.43	52.00 | 40.11	52.00 | 46.80
U	21.33 | 19.52	32.00 | 32.28	38.67 | 31.32	49.33 | 41.37	18.80 | 11.56	21.64 | 13.99	33.33 | 31.02	41.33 | 33.87	14.07 | 17.33	22.66 | 25.33
T	**75.60 | 73.61**	**79.33 | 74.16**	**55.33 | 47.03**	**58.00 | 56.73**	**48.54 | 46.15**	**54.84 | 52.75**	**86.14 | 85.33**	**91.08 | 90.67**	**78.67 | 76.38**	**80.00 | 78.70**
XV	56.00 | 44.84	64.00 | 51.60	52.00 | 46.90	54.67 | 49.27	28.09 | 20.44	29.58 | 21.38	47.25 | 41.25	52.19 | 49.28	68.80 | 65.99	72.00 | 68.27
	C(Train) - A(Test)		**M(Train) - A(Test)**		**E(Train) - A(Test)**		**A(Train) - C(Test)**		**R(Train) - A(Test)**	
	FCN	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**
MM	35.54 | 26.33	45.45 | 35.56	19.83 | 6.62	23.14 | 11.80	47.11 | 41.94	65.29 | 62.80	18.65 | 8.60	20.61 | 15.45	36.36 | 32.85	45.45 | 45.09
Wh	42.15 | 38.31	44.63 | 44.10	23.14 | 14.80	31.40 | 22.18	51.24 | 46.89	52.89 | 46.93	28.32 | 16.33	34.62 | 26.10	29.75 | 24.61	31.40 | 25.79
W2	23.61 | 19.69	45.45 | 40.09	38.84 | 29.54	40.50 | 32.51	44.63 | 39.60	55.37 | 50.97	31.24 | 27.97	32.49 | 28.93	33.88 | 29.61	39.67 | 33.52
X	48.76 | 42.90	55.37 | 49.50	19.18 | 6.62	24.79 | 17.53	36.36 | 29.04	47.93 | 44.72	25.85 | 17.40	27.69 | 23.15	23.97 | 17.46	31.40 | 26.11
W2	27.27 | 17.64	37.19 | 30.99	25.62 | 15.54	41.32 | 37.63	51.24 | 47.91	52.07 | 49.55	22.90 | 15.28	25.96 | 20.07	39.67 | 36.58	40.50 | 39.28
U	29.75 | 24.79	32.23 | 26.49	28.10 | 24.59	38.02 | 32.76	38.02 | 32.38	42.98 | 38.41	25.18| 20.75	25.89 | 23.12	22.31 | 15.51	28.93 | 26.55
T	**62.98 | 55.09**	**67.93 | 61.78**	**52.23 | 44.48**	**62.98 | 54.73**	**66.12 | 65.36**	**73.90 | 73.55**	**50.67 | 45.94**	**53.19 | 49.29**	**51.40 | 44.32**	**58.84 | 55.14**
XV	37.19 | 29.27	47.11 | 38.54	38.02 | 33.50	50.41 | 43.70	61.16 | 60.27	67.77 | 67.23	31.25 | 26.15	32.65 | 29.65	46.28 | 38.97	53.72 | 49.69
	C(Train)-R(Test)		**M(Train) - R(Test)**		**E(train) - R(Test)**		**A(Train) - R(Test)**		**R(train) - C(test)**	
	FCN	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**	**FCN**	**CNN**
MM	31.77| 18.93	38.54 | 28.60	20.31 | 7.67	27.08 | 19.64	20.31 | 6.75	40.62 | 40.85	17.65 | 11.90	19.79 | 15.69	26.28 | 18.75	35.25 | 30.87
Wh	62.50 | 60.84	64.58 | 61.40	20.31 | 8.69	21.30 | 8.75	25.00 | 15.80	37.50 | 34.46	22.40 | 10.57	29.17 | 27.07	36.90 | 32.64	38.79 | 36.73
W	29.17 | 23.17	32.29 | 27.46	30.73 | 23.06	35.94 | 27.42	32.29 | 21.84	32.29 | 25.27	26.56 | 19.30	28.80 | 23.38	25.89 | 17.06	33.28 | 23.24
X	40.62 | 28.50	42.19 | 29.18	20.31 | 7.67	25.52 | 18.09	22.92 | 11.65	40.62 | 28.50	24.98 | 18.64	26.04 | 21.66	31.00 | 24.70	36.66 | 30.48
W2	21.35 | 8.93	26.56 | 18.47	29.17 | 21.53	30.73 | 26.00	34.90 | 27.62	34.90 | 29.65	29.17 | 27.07	32.29 | 32.24	19.75 | 10.54	20.61 | 17.97
U	41.15 | 32.45	43.75 | 37.60	19.79 | 6.61	19.79 | 6.61	21.35 | 8.81	21.35 | 8.81	27.08 | 19.87	31.25 | 26.72	28.87 | 26.77	32.42 | 31.13
T	**71.88 | 71.33**	**72.40 | 72.29**	**45.52 | 34.70**	**48.65 | 38.98**	**66.15 | 66.12**	**72.40 | 72.29**	**64.06 | 63.17**	**69.52 | 69.27**	**57.20 | 56.57**	**62.23 | 61.59**
XV	35.52 | 27.30	41.15 | 32.68	34.38 | 29.53	39.58 | 35.03	37.84 | 31.91	38.54 | 36.40	34.89 | 31.45	37.50 | 34.43	37.84 | 34.80	37.92 | 36.40

TABLE I: Evaluation scores of different models trained with different PTM representations for both same corpus and cross-corpus SER; Scores are presented in Accuracy | Macro F1 format; P, C, M, E, A, R stands for PTM, CREMA-D, MESD, Emo-DB, AESDD, and RAVDESS; X(Train) - X(Test) represents the training and evaluation dataset where X = C, M, E, A, R. For example, C(Train) - M(Test) represents training on CREMA-D and testing on MESD; MMS (MM), Whisper (Wh), WavLM (W), XLSR (X), Wav2Vec2 (W2), UniSpeech (U), TRILLsson (T), and X-vector (XV) are PTMs; The intensity highlights the performance levels, where darker shades indicate higher values and lighter shades indicate lower values. The scores are highlighted block-wise i.e. X(Train) - X(Test) (where X = C, M, E, A, R) block based on its relative performance.

speaker ages and ethnic backgrounds. It covers six emotions: angry, happiness, sadness, fear, disgust, and a neutral state, with each actor delivering 12 distinct sentences.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [26]: It contains 7,356 english clips covering eight emotions: neutral, calm, happiness, sadness, angry, fearful, disgust, and surprise for speech, and six emotions for song. The database provides high emotional validity, as validated by 319 raters who assessed recordings on emotional category, intensity, and genuineness.

German Emotional Speech Database (Emo-DB) [27]: It is recorded in German, contains 535 utterances from five male and five female actors. Each actor was assigned one of ten predefined scripts for the recordings. Emo-DB includes seven

emotional categories: angry, fear, boredom, disgust, happiness, neutral, and sadness.

Mexican Emotional Speech Database (MESD) [28]: It is a culturally tailored emotional speech dataset designed for Mexican Spanish speakers. It includes 864 utterances featuring six emotional states: angry, disgust, fear, happiness, neutral, and sadness. The recordings are categorized into three demographic groups: male adults, female adults, and children, with non-professional actors delivering carefully selected single-word utterances.

Acted Emotional Speech Dynamic Database (AESDD) [29]: It is a Greek-language speech emotion dataset consisting of approximately 600 utterances recorded by five actors. It includes five emotional states: angry, disgust, fear, happiness, and sadness. In our study, we consider the common emotions: happiness, fear, sadness, angry, and disgust as we are primarily interested in cross-corpus SER.

B. Downstream Modeling

We use FCN and CNN as our downstream models as used by previous research in SER [14] and related applications [30]. As we are interested in understanding the PTMs implicit capability for cross-corpus SER, so we kept the downstream modeling as simple as possible. For CNN (Figure 1b), we use 1D-CNN layer with a kernel size of 3 followed by max-pooling with a pool size of 2. The extracted features are then flattened and passed through a FCN block containing a dense layer of 90 neurons. Finally, a classification that contains the output layer with softmax activation function. For FCN (Figure 1a), we use the same modeling as used by the FCN block in CNN modeling. The FCN models trainable parameters are between 0.4 to 0.8M and for CNN, it is between 1 to 2M.

Training Details: We use Adam optimizer with batch size of 32 and cross-entropy as loss function. We train the all the models for 20 epochs. We use dropout and early stopping to prevent overfitting. We follow a five-fold cross-validation for training and evaluation of the models for both same-corpus and cross-corpus experiments.

C. Results

Table I presents the evaluation results for the downstream models trained with different PTMs. For same-corpus evaluation, we present the average performance across five folds. For cross-corpus evaluation, we follow a similar five-fold approach: in each fold, we train on four folds of the training dataset and test on one fold of the testing dataset. This process is repeated five times, and we report the average performance across these folds to ensure robustness.

For same-corpus experiments : TRILLsson consistently outperforms monolingual, multilingual, and speaker recognition PTMs across CREMA-D, MESD, Emo-DB, AESDD, and RAVDESS. This aligns with prior studies on evaluating TRILLsson for SER [14], reinforcing its strength in capturing critical paralinguistic features like pitch, intonation, and rhythm. Among monolingual PTMs, no clear winner emerges, as performance varies across datasets. WavLM outperforms the others

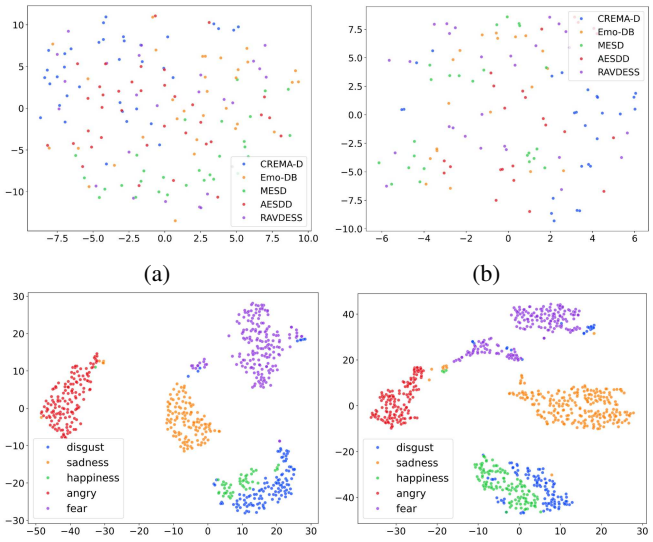


Fig. 2: t-SNE plots of PTMs raw representations: (a) Anger (b) Sadness (c) RAVDESS (d) CREMA-D

Fig. 3: Confusion matrices: (a) Trained and Tested on CREMAD- with TRILLsson representations and CNN downstream (b) Trained on CREMA-D and tested on Emo-DB with TRILLsson representations and CNN downstream; x-axis: Predicted, y-axis: True

in some cases, but Wav2vec2 achieves better results in MESD, indicating sensitivity to data distribution. Similarly, multilingual PTMs exhibit performance variations, with MMS leading for Emo-DB with CNN downstreams, while WavLM (monolingual PTM) surpasses multilingual PTMs in AESDD. Interestingly, x-vector shows mixed performance, occasionally outperforming monolingual and multilingual PTMs but underperforming in other cases, further highlighting the influence of dataset-specific factors. To further validate TRILLsson representations effectiveness, we visualize its raw representations using t-SNE for RAVDESS and CREMA-D in Figure 2 (c) and (d), revealing well-separated emotion clusters. Additionally, we present the confusion matrix of CNN downstream trained on TRILLsson representations for CREMA-D for a particular fold in Figure 3 (a), confirming its superior performance.

For cross-corpus experiments : TRILLsson outperforms other

PTMs—including monolingual, multilingual, and speaker recognition PTMs across all datasets. In some instances, it reports top performance by a very large margin in comparison to other PTMs. This validates our *hypothesis that paralinguistic PTM representations are inherently better suited for cross-corpus SER. Unlike other PTMs, TRILLsson focuses on capturing essential speech characteristics—such as intonation, pitch, rhythm, and prosody-that are fundamental for SER in a much better way*. Because these features are largely independent of linguistic content, TRILLsson's representations exhibit greater robustness across diverse datasets, enabling superior generalization in cross-corpus settings. Among monolingual PTMs, performance varies across datasets. WavLM excels in some cases, while Wav2vec2 achieves better in some, indicating that monolingual PTMs struggle with consistent cross-corpus generalization. Similarly, no single multilingual PTM dominates across all datasets among the multilingual PTMs. These results emphasize that PTM performance in cross-corpus settings is strongly influenced by the underlying data distribution. Despite these multilingual PTMs are pre-trained on multiple languages, they show poor cross-corpus generalization and thus, amplifying the dependence of SER on paralinguistic features and in which TRILLsson excels in. Interestingly, x-vector demonstrates mixed performance, occasionally surpassing both monolingual and multilingual PTMs in some datasets while underperforming in others. This suggests that paralinguistic characteristics embedded in x-vector representations can be beneficial for cross-corpus SER in certain contexts but are not universally effective. We also plot the t-SNE plot visualizations of TRILLsson representations for anger and sadness emotion across all the datasets in Figure 2 (a) and (b) respectively. We segmented the anger and sadness emotions for all the datasets, extracted representations from TRILLsson and visualized it using t-SNE plot. We observe that no clear clusters are present and through this, we can say that TRILLsson converts the same emotion samples to a joint representational space irrespective of linguistic difference. These plots further amplifies our obtained results. We also plot the confusion matrix of CNN model trained on CREMA-D with TRILLsson representations and evaluated on Emo-DB in Figure 3 (b).

IV. Conclusion

In this study, we investigate paralinguistic PTM (TRILLsson) for cross-corpus SER, which have been largely overlooked by previous benchmarks on evaluating PTMs for cross-corpus SER. We hypothesize that paralinguistic PTM would outperform other PTMs. By analyzing SOTA PTM representations, including paralinguistic, monolingual, multilingual, and speaker recognition, we show that TRILLsson, consistently outperforms other PTMs and validating our hypothesis. These results highlight the importance of our work and it will act as benchmark for future works on cross-corpus SER. It also calls for incorporating paralinguistic PTM in previous cross-corpus SER benchmarks, ultimately enhancing their trustworthiness and offering valuable insights for future PTM evaluations in SER tasks.

References

[1] A. Milton, S. S. Roy, and S. T. Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[2] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using mfcc and support vector machine," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, IEEE, 2016, pp. 1080–1084.

[3] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 4749–4753.

[4] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[5] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6907–6911.

[6] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[7] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "Superb: Speech processing universal performance benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198. DOI: 10.21437/Interspeech.2021-1775.

[8] H. Wu, H.-C. Chou, K.-W. Chang, *et al.*, "Emo-superb: An in-depth look at speech emotion recognition," *arXiv preprint arXiv:2402.13018*, 2024.

[9] H. Wu, H.-C. Chou, K.-W. Chang, *et al.*, "Open-emotion: A reproducible emo-superb for speech emotion recognition systems," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024, pp. 510–517.

[10] A. R. Naini, M. A. Kohler, E. Richerson, D. Robinson, and C. Busso, "Generalization of self-supervised learning-based representations for cross-domain speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 031–12 035.

[11] A. Ibrahim, S. Shehata, A. Kulkarni, M. Mohamed, and M. Abdul-Mageed, "What does it take to generalize ser model across datasets? a comprehensive benchmark," in *Interspeech 2024*, 2024, pp. 1590–1594. DOI: 10.21437/Interspeech.2024-1983.

[12] M. Osman, D. Z. Kaplan, and T. Nadeem, "Ser evals: In-domain and out-of-domain benchmarking for speech emotion recognition," in *Interspeech 2024*, 2024, pp. 1395–1399. DOI: 10.21437/Interspeech.2024-2440.

[13] Z. Ma, M. Chen, H. Zhang, *et al.*, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and

benchmark," in *Interspeech 2024*, 2024, pp. 1580–1584. DOI: 10.21437/Interspeech.2024-788.

[14] O. C. Phukan, G. S. Kashyap, A. B. Buduru, and R. Sharma, "Are paralinguistic representations all that is needed for speech emotion recognition?" In *Interspeech 2024*, 2024, pp. 4698–4702. DOI: 10.21437/Interspeech.2024-2233.

[15] 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.

[16] S. Chen, Y. Wu, C. Wang, *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6152–6156, 2021.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[18] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404. DOI: 10.21437/Interspeech.2021-703.

[19] A. Babu, C. Wang, A. Tjandra, *et al.*, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143.

[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] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[22] J. Shor and S. Venugopalan, "TRILLsson: Distilled Universal Paralinguistic Speech Representations," in *Proc. Interspeech 2022*, 2022, pp. 356–360. DOI: 10.21437/Interspeech.2022-118.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.

[24] O. Chetia Phukan, A. Balaji Buduru, and R. Sharma, "Transforming the Embeddings: A Lightweight Technique for Speech Emotion Recognition Tasks," in *Proc. INTERSPEECH 2023*, 2023, pp. 1903–1907. DOI: 10.21437/Interspeech.2023-2561.

[25] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, e0196391, 2018.

[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, *et al.*, "A database of german emotional speech.," in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[28] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The mexican emotional speech database (mesd): Elaboration and assessment based on machine learning," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 1644–1647.

[29] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," *Journal of the Audio Engineering Society*, vol. 66, no. 6, pp. 457–467, 2018.

[30] M. Charola, A. Kachhi, and H. A. Patil, "Whisper encoder features for infant cry classification," in *Interspeech 2023*, 2023, pp. 1773–1777. DOI: 10.21437/Interspeech.2023-1916.