Beyond Speech and More: Investigating the Emergent Ability of Speech Pre-Trained Models for Classifying Physiological Time-Series Signals

Orchid Chetia Phukan[†*], Swarup Ranjan Behera[‡*], Girish[†§*], Mohd Mujtaba Akhtar[†¶*]
Arun Balaji Buduru[†], Rajesh Sharma[‖**]
[†]IIIT-Delhi, India, [‡]Independent Researcher, India, [§]UPES, India, [¶]V.B.S.P.U, India
[‖]University of Tartu, Estonia, [**]Plaksha University, India
E-mail: orchidp@iiitd.ac.in

***Abstract*—In this study, we push the boundaries by evaluating speech pre-trained models (SPTMs) for classifying physiological time-series signals (PTSS), an out-of-domain (OOD) task. We test two key hypotheses: first, that SPTMs can generalize to PTSS by capturing shared temporal patterns; second, that multilingual SPTMs will outperform others due to their exposure to greater variability during pre-training, leading to more robust, generalized representations. Our experiments, conducted for stress recognition using ECG (Electrocardiogram), EMG (Electromyography), and EDA (Electrodermal Activity) signals reveal that models trained on SPTM-derived representations outperform those trained on raw PTSS. Among all SPTMs, multilingual SPTMs achieve the highest accuracy, supporting our hypothesis and demonstrating their better OOD capabilities. This initial work positions SPTMs as promising tools for new uncharted domains beyond speech.**

I. Introduction

The dawn of large-scale pre-trained models (PTMs) has revolutionized machine learning across diverse modalities, including speech, audio, text, and image. These PTMs, trained on vast datasets, possess exceptional generalization abilities, enabling them to perform a wide range of tasks both in-domain (ID) as well as out-of-domain (OOD). Large language models (LLMs), despite being pre-trained on text data, have shown excellent OOD performance in predicting protein phase transition [1] and speech-based depression detection [2]. Similarly, researchers have leveraged vision transformers pre-trained on visual data for speech emotion recognition [3]. Speech PTMs (SPTMs) primarily trained on large-scale speech datasets have exhibited cross-task generalization, handling tasks like audio classification [4], [5], bio-acoustics [6], etc. These successes highlight the ability of PTMs to extract robust and transferable features across tasks, even when the data domains significantly differ. This underscores the emergent capabilities of PTMs to extend beyond their original training domains, unlocking new avenues for exploration and application across disciplines.

In this study, we focus on investigating the ability of SPTMs to perform a specific OOD task: the classification of physiological time-series signals (PTSS). The classification of PTSS is particularly challenging due to their inherent noise, inter-subject variability, and low signal-to-noise ratios [7], [8]. Previously, researchers in the NLP and vision communities have explored the usage of text-based PTMs and vision-based PTMs for PTSS applications [9]–[12]. However, to the best of our knowledge, no effort has been made to leverage SPTMs for the classification of PTSS and this study addresses this gap by exploring whether SPTMs, trained solely on speech data, can effectively generalize to PTSS. We hypothesize that: (i) *SPTMs can generalize to classify PTSS by leveraging the shared temporal patterns inherent in both speech and PTSS data*, and (ii) *Multilingual SPTMs will perform better than other SPTMs due to their large-scale pre-training on diverse languages, equipping them with robust, generalized representations that enhance their ability to capture complex temporal patterns, making them more effective.*

To test these hypotheses, we consider three different types of PTSS: ECG (Electrocardiogram), EMG (Electromyography), and EDA (Electrodermal Activity), to better understand the OOD generalizability of SPTMs. We conduct experiments on the WESAD dataset, a benchmark for physiological stress recognition. We extract representations from the frozen SPTMs for better understanding of their intrinsic capabilities for transferability to PTSS classification. We show that downstream models trained on SPTMs representations outperform models trained on raw PTSS data. Among the SPTMs, we demonstrate that multilingual SPTMs achieve the topmost performance. These results validate the potential of SPTMs to generalize effectively to non-speech tasks and highlight their utility for physiological signal applications.

The main contributions are summarized as follows:

- We are, to the best of our knowledge, the first study to evaluate diverse state-of-the-art (SOTA) SPTMs—WavLM, Wav2vec2, Unispeech-SAT, x-vector, HuBERT, MMS, XLS-R, and Whisper—for classifying PTSS such as ECG, EMG, and EDA for stress recognition.
- Our findings show that SPTMs representations outperform models trained on raw physiological data. This mirrors findings in NLP, where LLM representations have shown competitive performance compared to raw data [9].
- We demonstrate that multilingual SPTMs exhibit the best performance among all the SOTA SPTMs. This superior

* Contributed equally as first authors.

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

performance is consistent across the three PTSS (ECG, EMG, and EDA) under consideration.

II. Related Work

In this section, we briefly explore the application of SPTMs for tasks beyond speech, followed by an examination of PTMs from other domains utilized for PTSS tasks. SPTMs have shown exceptional performance in diverse tasks outside speech processing. Sarkar et al. [13] investigated various SOTA SPTMs such as WavLM, Wav2vec2, etc for classifying animal callers. Cauzinille et al. [14] evaluated self-supervised SPTMs for classifying gibbon's vocal signatures. Turian etl al. [15] explored various SPTMs representations for achieving SOTA performance across various acoustic tasks such as environmental sound classification, music tasks, and so on. However, the audio tasks, as well as the speech processing applications, fall under the broader domain of sounds. This leaves a gap in understanding the performance of SPTMs for OOD tasks outside the broader sound domain. In our study, we focus on a challenging OOD task: the classification of PTSS with diverse applications ranging from affective computing to healthcare. Researchers from NLP and CV domains have explored the usage of their domain-specific PTMs for classifying physiological signals. Hu et al. [10] exploited LLMs for mental health assessment with EEG signals and Ishaque et al. [16] leveraged various vision foundation models such as ResNet, VGG-16, etc for stress recognition from ECG signals. However, SPTMs haven't been explored yet for the classification of PTSS, and in this work, we take the first step towards this direction by classifying PTSS for stress recognition.

III. Speech Pre-Trained Models

In this section, we provide an overview of the different SOTA SPTMs considered in our study. We consider diverge range of SPTMs ranging from those trained for monolingual speech processing (WavLM, Wav2vec2, Unispeech-SAT, HuBERT), multilingual (XLS-R, Whisper, MMS) as well as speaker recognition (x-vector). Such consideration will allow us to understand the transferability of SPTMs for PTSS data in a much better manner.

WavLM [17] is a SOTA SPTM on SUPERB that integrates masked speech prediction and denoising, making it effective for various speech-processing tasks. It is trained on only english data. We utilize base[1] (WavLM-base) and large[2] (WavLM-large) versions with parameters of 94.70 and 316.62 millions respectively in our experiments. We included large variant as it shows the topmost performance in SUPERB.

Wav2vec2[3] [18] is a self-supervised learning model trained by solving a contrastive task, designed to learn contextualized speech representations. It was pre-trained on 960 hours of english librispeech and we utilize base version with 95.04 million parameters.

UniSpeech-SAT[4] [19] is also a SOTA SPTM on SUPERB, pre-trained in a speaker-aware format. It solves a contrastive learning objective in a multi-task manner. We use the base version of 94.68 million parameters trained on 960 hours of librispeech.

HuBERT[5] [20] is a self-supervised model that learns from continuous speech data by iteratively refining clusters through a masked prediction loss. We use the base version of 94.68 million parameters trained on 960 hours of libirspeech.

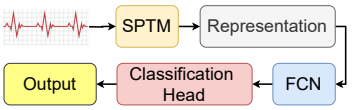
x-vector[6] [21] is a time-delay neural network trained for speaker recognition and it improves over its previous SOTA i-vector. It has 4.2 million parameters and trained on combination of Voxceleb1 and Voxceleb2.

XLS-R[7] [22]is a cross-lingual version of Wav2vec2, trained on 128 languages, with its architecture designed to support multilingual and cross-domain tasks. We make use of 300 million parameters version in our experiments. Its pre-training was carried out on datasets such as Commonvoice, BABEL, Voxpopuli, and so on.

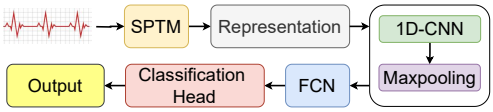
Whisper[8] [23] is an encoder-decoder-based multilingual transformer model trained on 96 languages. It is a multi-task learning model and trained in a weakly-supervised manner. We use the base version with 74 million parameters.

MMS[9] [24] is trained on over 1400 languages and is built upon the Wav2vec2 architecture. The training datasets includes FLEURS, BABEL, MMS-lab and so on. We use the 1 billion parameters variant.

We use the frozen SPTMs and extract representations from the last hidden layer of each SPTM by applying average pooling. The extracted representation dimensions are: 512 (x-vector, Whisper), 768 (WavLM, Wav2vec2, UniSpeech-SAT, HuBERT), 1024 (WavLM-large), and 1280 (MMS, XLS-R). Whisper's representations were specifically obtained from its encoder, with the decoder discarded.



(a) Fully Connected Network (FCN)



(b) Convolutional Neural Network (CNN)

Fig. 1: Modeling

[1]https://huggingface.co/microsoft/wavlm-base
[2]https://huggingface.co/microsoft/wavlm-large
[3]https://huggingface.co/facebook/wav2vec2-base
[4]https://huggingface.co/microsoft/unispeech-sat-base
[5]https://huggingface.co/facebook/hubert-base-ls960
[6]https://huggingface.co/speechbrain/spkrec-xvect-voxceleb
[7]https://huggingface.co/facebook/wav2vec2-xls-r-300m
[8]https://huggingface.co/openai/whisper-base
[9]https://huggingface.co/facebook/mms-1b

	FCN						CNN					
	ECG		EMG		EDA		ECG		EMG		EDA	
	A	F1	A	F1	A	F1	A	F1	A	F1	A	F1
Baseline: Raw Signals												
	58.71	58.68	72.17	54.42	39.38	36.74	75.42	64.18	72.90	62.12	55.26	55.20
SPTM												
WavLM-base	81.60	73.42	78.05	69.99	64.26	43.45	81.90	76.39	78.66	71.36	70.29	63.70
WavLM-large	85.00	80.70	73.78	60.89	67.23	42.02	89.78	88.18	76.69	70.97	69.69	55.76
Wav2vec2	80.45	75.59	75.17	66.99	68.17	49.99	84.21	81.47	77.24	72.23	71.45	61.67
Unispeech-SAT	76.63	65.07	72.45	54.84	64.99	41.20	79.60	75.89	75.73	66.56	69.78	55.23
x-vector	83.42	78.86	75.87	60.33	73.54	55.34	86.78	82.92	75.90	66.24	74.72	63.28
HuBERT	82.93	76.84	75.17	72.57	73.02	55.21	84.51	82.97	79.78	73.57	74.96	63.09
MMS	**86.94**	**83.08**	**80.18**	**72.82**	**75.08**	**64.54**	**91.51**	**89.94**	**81.39**	**77.46**	**79.51**	**72.23**
XLS-R	**91.48**	**89.21**	**79.63**	**73.40**	**74.33**	**63.06**	**91.69**	**90.22**	**80.21**	**76.83**	**78.33**	**71.96**
Whisper	**93.42**	**91.94**	**85.15**	**80.90**	**75.72**	**67.57**	**97.85**	**97.43**	**87.00**	**83.92**	**77.42**	**68.14**

TABLE I: Evaluation Scores: A and F1 represent accuracy and macro-average F1 scores, respectively, for 2-class classification (stress vs. non-stress) with downstream models trained on different SPTMs representations. The scores are average of five-folds in %. Colored cells highlight the top three performances per column: 1st, 2nd, 3rd.

	FCN						CNN					
	ECG		EMG		EDA		ECG		EMG		EDA	
	A	F1	A	F1	A	F1	A	F1	A	F1	A	F1
Baseline: Raw Signals												
	44.26	36.18	40.32	35.51	33.25	21.02	57.78	44.30	58.05	45.72	46.23	33.04
SPTM												
WavLM	52.89	39.08	44.59	39.03	49.86	27.36	67.02	49.27	58.93	47.86	53.14	38.15
WavLM-large	66.75	48.12	56.05	38.14	53.11	35.31	69.41	50.20	64.35	50.35	56.47	46.32
Wav2vec2	65.91	47.56	62.32	44.17	53.44	36.26	70.63	63.43	63.69	54.28	57.53	47.44
Unispeech-SAT	63.23	44.21	51.53	36.72	46.14	24.55	64.47	56.31	66.18	48.05	53.11	39.37
x-vector	69.11	56.02	59.35	40.08	45.32	43.82	76.72	67.60	61.41	53.56	59.62	41.38
HuBERT	66.69	49.63	61.44	46.91	56.08	42.06	67.78	63.35	65.99	53.13	59.84	43.38
MMS	**72.42**	**60.61**	**64.02**	**47.35**	**59.68**	**44.57**	**78.27**	**72.03**	**68.29**	**56.32**	**62.93**	**49.06**
XLS-R	**74.96**	**66.29**	**66.02**	**47.39**	**60.69**	**45.38**	**77.81**	**69.06**	**66.87**	**58.18**	**61.72**	**50.27**
Whisper	**81.45**	**72.43**	**74.84**	**67.10**	**60.08**	**44.89**	**87.27**	**85.40**	**78.60**	**74.69**	**63.56**	**52.38**

TABLE II: Evaluation Scores: A and F1 represent accuracy and macro-average F1 scores, respectively, for 3-class classification (baseline vs. stress vs. amusement) with downstream models (FCN, CNN) trained on different SPTMs representations. The scores are averages of five-fold in %. Colored cells highlight the top three performances per column: 1st, 2nd, 3rd.

IV. Experimental Setup

A. Benchmark Dataset

We use the WESAD (Wearable Stress and Affect Detection) dataset [25], a benchmark dataset widely used for PTSS-based stress recognition tasks. It consists of PTSS from chest and wrist-worn sensors, collected from 15 subjects at a 700 Hz sampling rate. The dataset supports both 2-class (stress vs. non-stress) and 3-class (baseline vs. stress vs. amusement) classification tasks. For our experiments, we leverage ECG, EMG, and EDA signals collected from chest-worn sensors. Following [26], we segment the data into non-overlapping windows of 5 seconds with a shift of 2 seconds. Each signal is resampled to 16 kHz to align with the input requirements of SPTMs, and these resampled signals are directly passed to the SPTMs to extract representations for downstream modeling.

B. Downstream Modeling

We kept the downstream modeling simple to investigate the implicit behavior of SPTMs for classifying PTSS. We employ CNN (Convolutional Neural Network) and FCN (Fully Connected Network) as downstream networks and built models for both 2-class and 3-class classification. The CNN (Figure 1b) model consists of three 1D convolutional layers with kernel size of 3, respectively, and 16, 32, and 64 filters for the layers. Each convolutional layer is followed by max-pooling with a pool size of 2. Then, we flattened the output and passed it through an FCN with a dense layer with 128 neurons. We keep the number of neurons in the output layer depending on the classification task, i.e., 2 neurons for 2-class classification or 3 neurons for 3-class classification, applying softmax activation for output probabilities. For the FCN (Figure 1a), we kept the same FCN modeling as use in CNN modeling above with 128 neurons dense layer. The trainable parameters of the FCN models range from 3 to 5 millions, while for the CNN model, they vary between 4 and 13 millions, depending on the size of the SPTMs representation.We train the models for 50 epochs with a batch size of 32, a learning rate of 1e-3, and use Adam as the optimizer. We use cross-entropy as the loss function. To

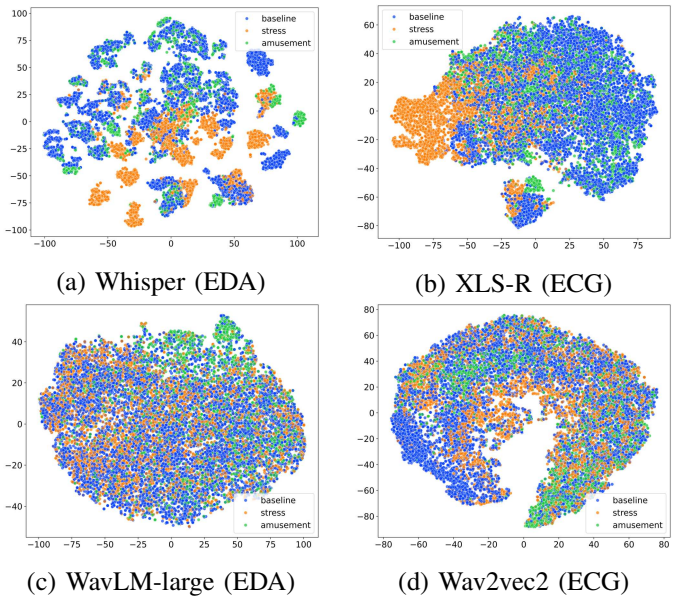
prevent overfitting, we use dropout and early stopping. For all our experiments, we use five-fold cross validation for training and evaluating the models where four folds are utilized for training and one fold for evaluation.

C. Experimental Results

Table I and II Baseline: Raw Signals presents the performance of models trained on raw PTSS for 2-class (stress vs. non-stress) and 3-class classification (baseline vs. stress vs. amusement). Previous research in PTSS has primarily focused on building models on raw signals [26], [27], so we consider this approach as the baseline in our study. We use the same modeling as the downstream modeling with SPTMs for baselines and follow the same training paradigm for the baseline models. In both the 2-class classification and 3-class classification, the CNN models achieve the best scores across ECG, EMG, and EDA compared to FCN.

Tables I and II SPTMs representations presents the evaluation scores for the 2-class classification task and the 3-class classification task respectively with SPTM representations. The scores demonstrate a clear dominance of downstream models trained on SPTMs representations compared to models trained on raw data. This supports our first hypothesis that *SPTMs will generalize for classifying PTSS by utilizing the common temporal patterns inherent to both speech and PTSS data.* This superior performance of SPTMs is consistent across all the considered models, all PTSS, and both 2-class and 3-classification tasks. Among WavLM-base and WavLM-large, we see that WavLM-large outperforms WavLM-base, likely due to differences in model size: the base version has 94.70 million parameters, while the large version has 316.62 million. However, it cannot be conclusively stated that larger models always yield better representations. For example, the base versions of UniSpeech-SAT and HuBERT, which are significantly smaller than WavLM-large, demonstrate superior performance in some cases and comparable performance in others.

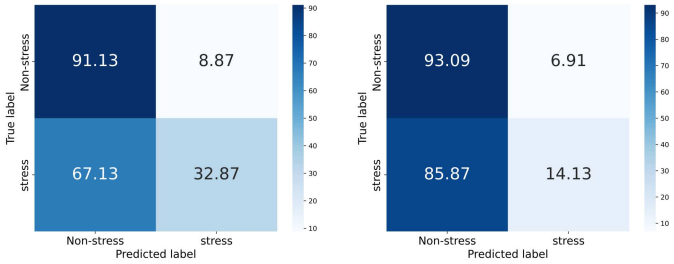
An intriguing observation is the performance of x-vector, which, despite having only approximately 4.2 million parameters, outperforms many larger SPTM counterparts. This highlights that SPTMs designed to capture speaker-specific characteristics can provide robust cross-domain representations, even with fewer parameters. Among all the SPTMs in Tables I and II, multilingual SPTMs consistently outperform other SPTMs, including monolingual SPTMs like UniSpeech-SAT, WavLM-base, WavLM-large, and the speaker recognition SPTM x-vector. This validates our second hypothesis that *multilingual SPTMs will outperform other SPTMs due to their large-scale pre-training across a wide variety of languages, enabling them to adapt more effectively to the nuances of physiological signals and provide more generalized cross-domain representations.* We also analyzed the raw representations extracted from the last hidden states of the SPTMs using t-SNE for visualization. The t-SNE plots revealed that multilingual SPTMs exhibited better clustering of different classes compared to other SPTMs, further showing their superior cross-domain generalizability and providing support to our second hypothesis. We also present the



(a) Whisper (EDA) (b) XLS-R (ECG)

(c) WavLM-large (EDA) (d) Wav2vec2 (ECG)

Fig. 2: t-SNE Plots of raw SPTM representations



(a) Whisper (EDA) (b) XLS-R (ECG)

Fig. 3: Confusion Matrices of multilingual SPTMs for 2-class classification

confusion matrices of models trained on Whisper and XLS-R representations in Figure 3. Among the multilingual SPTMs, performance varies, with certain SPTM excelling in specific instances while others lead in different scenarios. This variation reflects the diverse pretraining paradigms and architectures employed across the models, which contribute to their unique strengths.

V. Limitations and Future Work

In this work, we present a initial exploration towards understanding OOD capabilities of SPTMs for PTSS classification. In our study, we evaluated SPTMs using only two downstream networks. However, prior research in speech processing indicates that SPTMs performance varies across different architectures [28]. In future work, we aim to explore a wider range of downstream models. Additionally, we focused solely on representations from the final layer of the SPTMs. However, previous research evidence suggests that performance can fluctuate based on the extracted features from different layers [29], [30]. We will investigate this dimension in our future

studies as well. Further, we have focused on stress recognition, in our future studies, we will include PTSS datasets from other tasks too. We primarily focus on a single dataset and two tasks; however, this doesn't undermine our contribution, which shows the adaptability of SPTMs for PTSS, as WESAD is a benchmark dataset for evaluating models for PTSS classification. In addition, we have experimented with three PTSS to show the generalizability, and we can observe the complete dominance of SPTMs over baseline raw data models.

VI. Conclusion

In this study, we demonstrate for the first time that SPTMs, despite being trained solely on speech data, exhibit notable effectiveness in the OOD task of classifying PTSS for stress recognition. Our experiments on the WESAD dataset, using various SOTA SPTMs, reveal that downstream models trained on SPTMs representations outperform those trained on raw PTSS. Furthermore, we show that multilingual SPTMs excel compared to other SPTMs counterparts including monolingual and speaker recognition. This performance of multlingual SPTMs can be attributed to its multilingual pre-training that enhances their robustness and cross-domain generalization.
Our findings open new avenues for research on the OOD generalization capabilities of SPTMs and underscore their potential for applications in PTSS and beyond.

References

[1] M. Frank, P. Ni, M. Jensen, and M. B. Gerstein, "Leveraging a large language model to predict protein phase transition: A physical, multiscale, and interpretable approach," *Proceedings of the National Academy of Sciences*, vol. 121, no. 33, e2320510121, 2024.

[2] X. Zhang, H. Liu, K. Xu, *et al.*, "When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection," *arXiv preprint arXiv:2402.13276*, 2024.

[3] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," *Scientific Reports*, vol. 14, no. 1, p. 13 126, 2024.

[4] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *INTERSPEECH 2023*, 2023, pp. 2798–2802. DOI: 10.21437/Interspeech.2023-2193.

[5] R. Ma, A. Liusie, M. Gales, and K. Knill, "Investigating the emergent audio classification ability of ASR foundation models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 4746–4760. DOI: 10.18653/v1/2024.naacl-long.266. [Online]. Available: https://aclanthology.org/2024.naacl-long.266.

[6] M. U. Sheikh, H. Abid, B. S. Shafique, A. Hanif, and M. H. Khan, "Bird whisperer: Leveraging large pre-trained acoustic model for bird call classification," in *Interspeech 2024*, 2024, pp. 5028–5032. DOI: 10.21437/Interspeech.2024-1623.

[7] L. Gonzalez-Carabarin, E. Castellanos-Alvarado, P. Castro-Garcia, and M. Garcia-Ramirez, "Machine learning for personalised stress detection: Inter-individual variability of eeg-ecg markers for acute-stress response," *Computer Methods and Programs in Biomedicine*, vol. 209, p. 106 314, 2021.

[8] Z. Ahmad and N. Khan, "A survey on physiological signal-based emotion recognition," *Bioengineering*, vol. 9, no. 11, p. 688, 2022.

[9] Y. Gao, S. Myers, S. Chen, *et al.*, "When raw data prevails: Are large language model embeddings effective in numerical data representation for medical machine learning applications?" *arXiv preprint arXiv:2408.11854*, 2024.

[10] Y. Hu, S. Zhang, T. Dang, *et al.*, "Exploring large-scale language models to evaluate eeg-based multimodal data for mental health," in *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '24, Melbourne VIC, Australia: Association for Computing Machinery, 2024, pp. 412–417, ISBN: 9798400710582. DOI: 10.1145/3675094.3678494. [Online]. Available: https://doi.org/10.1145/3675094.3678494.

[11] H. Yoon, B. A. Tolera, T. Gong, K. Lee, and S.-J. Lee, *By my eyes: Grounding multimodal large language models with sensor data via visual prompting*, 2024. arXiv: 2407.10385 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2407.10385.

[12] O. C. Phukan, A. Das, A. B. Buduru, and R. Sharma, "Sonic: Synergizing vision foundation models for stress recognition from ecg signals," *arXiv preprint arXiv:2404.00827*, 2024.

[13] E. Sarkar and M. Magimai.-Doss, "Can self-supervised neural representations pre-trained on human speech distinguish animal callers?" In *INTERSPEECH 2023*, 2023, pp. 1189–1193. DOI: 10.21437/Interspeech.2023-1968.

[14] J. Cauzinille, B. Favre, R. Marxer, D. Clink, A. H. Ahmad, and A. Rey, "Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures," in *Interspeech 2024*, 2024, pp. 132–136. DOI: 10.21437/Interspeech.2024-1096.

[15] J. Turian, J. Shier, H. R. Khan, *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*, PMLR, 2022, pp. 125–145.

[16] S. Ishaque, N. Khan, and S. Krishnan, "Detecting stress through 2d ecg images using pretrained models, transfer learning and model compression techniques," *Machine Learning with Applications*, vol. 10, p. 100 395, 2022.

[17] 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.

[18] 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.

[19] S. Chen, Y. Wu, C. Wang, *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6152–6156. DOI: 10.1109/ICASSP43922.2022.9747077.

[20] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[21] 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)*, IEEE, 2018, pp. 5329–5333.

[22] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv*, vol. abs/2111.09296, 2021.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. DOI: 10.48550/ARXIV.2212.04356. [Online]. Available: https://arxiv.org/abs/2212.04356.

[24] 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.

[25] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[26] G. Singh, O. C. Phukan, and R. Kumar, "Stress recognition with multi-modal sensing using bootstrapped ensemble deep learning model," *Expert Systems*, vol. 40, no. 6, e13239, 2023.

[27] R. Tanwar, O. C. Phukan, G. Singh, P. K. Pal, and S. Tiwari, "Attention based hybrid deep learning model for wearable based stress recognition," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107 391, 2024.

[28] S. Zaiem, Y. Kemiche, T. Parcollet, S. Essid, and M. Ravanelli, "Speech self-supervised representation benchmarking: Are we doing it right?" In *INTERSPEECH 2023*, 2023, pp. 2873–2877. DOI: 10.21437/Interspeech.2023-1087.

[29] M. Kodali, S. R. Kadiri, and P. Alku, "Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings," in *INTERSPEECH 2023*, 2023, pp. 4134–4138. DOI: 10.21437/Interspeech.2023-2038.

[30] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 914–921.