Investigating Polyglot Speech Foundation Models for Learning Collective Emotion from Crowds

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

***Abstract*—This paper investigates the polyglot (multilingual) speech foundation models (SFMs) for Crowd Emotion Recognition (CER). We hypothesize that polyglot SFMs, pre-trained on diverse languages, accents, and speech patterns, are particularly adept at navigating the noisy and complex acoustic environments characteristic of crowd settings, thereby offering a significant advantage for CER. To substantiate this, we perform a comprehensive analysis, comparing polyglot, monolingual, and speaker recognition SFMs through extensive experiments on a benchmark CER dataset across varying audio durations (1 sec, 500 ms, and 250 ms). The results consistently demonstrate the superiority of polyglot SFMs, outperforming their counterparts across all audio lengths and excelling even with extremely short-duration inputs. These findings pave the way for adaptation of SFMs in setting up new benchmarks for CER.**

I. Introduction

Crowd Emotion Recognition (CER) involves the intricate task of predicting collective emotional states in large groups, where emotions are conveyed vocally through cheering, booing, and clapping, and visually through gestures such as waving and synchronized movements. These emotional expressions are prevalent at events like sports matches, concerts, political rallies, and social movements, where they profoundly shape the overall atmosphere and influence both participants and performers. Accurately capturing and interpreting these emotions is crucial across fields such as public safety, event management, and social research. Research in CER has predominantly centered around visual modalities, such as facial expressions and body movements [1]–[6]. However, audio-based approaches remain relatively unexamined, highlighting a significant gap in the field.

CER through audio poses distinct challenges, largely due to the noisy, spontaneous, and overlapping nature of crowd expressions, particularly in audio streams. Crowds express emotions collectively, but the complexity and dynamism of these environments have left CER underexplored compared to speech emotion recognition (SER), that focuses on recognizing a single individual emotions. Additionally, the limited availability of labeled data also hinders the research into CER through audio cues. Overcoming these obstacles necessitates innovative approaches capable of managing both the linguistic diversity and the noisy, dynamic environments typical of large crowds. To solve this gap, Franzoni et al. [7] presented the first audio-based CER dataset. Adding on this, Faisal et al. [8] used MFCC with Random Foreset classifier and Vision-based foundation models like MobileNetV2 with spectrograms for CER. Anand et al. [9] used CLAP representations with a downstream neural network-based approach for predicting crowd excitement score. In this work, we focus on CER through audio.

Unlike CER, SER has sufficient development particularly due to recent advancements in speech foundation models (SFMs), such as Wav2Vec, HuBERT, pre-trained on large-scale diverse speech datasets, offer promising solutions for performance benefit, data scarcity as well as prevention of training models from scratch [10]–[12]. These models have also excelent in various other speech processing tasks such as speech recognition [13], speaker segmentation [14], shout intensity prediction [15], and so on. Applying such SFMs to CER holds the potential to significantly improve data efficiency and enhance performance in CER, especially within the noisy and diverse environments that characterize real-world crowd scenarios.

To bridge this gap, we investigate the potential of state-of-the-art (SOTA) SFMs for recognizing crowd emotions and we hypothesize that *polyglot (multilingual) SFMs will prove highly effective for CER by navigating the noisy and complex environments of crowded environments. This can be attributed to their capacity to capture a wide range of pitches, tones, and emotional nuances owing to their pre-training on diverse speech data that encompasses various languages, accents, and speaking styles.* To validate, our hypothesis, we carry out a large-scale comparison of various SOTA SFMs with different downstreams such as SVM, Random Foreset, Fully Connected Network (FCN), and CNN. To the best of our knowledge, we are the first study to explore SFMs for CER.

To summarize, the major contributions of this study are as follows:

- We conduct comprehensive comparative analysis of SOTA polyglot, monolingual, and speaker recognition SFMs to investigate the efficacy of polyglot SFMs (XLS-R, Whisper, MMS) for CER. Such comparison of SOTA SFMs for CER is the first one to the best of knowledge. Work flow of the paper is given in Figure 1.
- We perform extensive experiments on a benchmark CER corpus and investigate the influence of varying audio durations (1 second, 500 ms, and 250 ms) on CER

* Contributed equally as first authors.

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

performance, aiming to identify optimal conditions for emotion recognition and evaluate the potential of SFMs for CER in extremely short duration audios.
- Our results reveal that polyglot SFMs consistently outperform their counterparts across all audio lengths, underscoring their superiority for CER tasks, even in extremely short-duration audio segments.

Class	#Clips	Duration (s)	#Blocks
Approval	39	518	1787
Disapproval	15	118	388
Neutral	15	1874	7340
Total	69	2510	9515

TABLE I: Dataset statistics: Number of clips per class (#Clips), total duration (in seconds), and total 1-sec blocks (#Blocks).

II. Speech Foundation Models

We leverage SOTA SFMs that excels across various tasks in speech processing across different benchmarks. Each SFM with their distinct technical strengths, provides critical capabilities aligned with our investigation into CER. By integrating these advanced SFMs, our approach harnesses their scalability, robustness, and versatility, making them invaluable for effective CER. Detailed explanation of the Detailed explanation regarding the SFM considered in our study are given below:

WavLM[1] [16] It is a SOTA SFM that outperforms all the other SFMs in various speech procesing tasks in SUPERB benchmark. It learns both speech denoising and masked prediction during training and trained on 960 hours of english librispeech data. We use the base version of 94.70 million parameters.

UniSpeech-SAT[2] [17] employs a contrastive objective alongside multitask learning. Its pre-training follows a speaker-aware approach and is conducted using 960 hours of Librispeech English speech data. We utilize the base variant, which comprises 94.68 million parameters.

Wav2vec2[3] [18] employs a self-supervised learning approach, transforming raw audio into latent speech representations through a convolutional feature encoder and Transformer blocks. We consider the base version of 95.04 million parameters trained on 960 hours of Librispeech data in english.

HuBERT[4] [19] employs a BERT-like masked prediction framework to learn both acoustic and linguistic features effectively. It shows SOTA performance on multiple speech recognition benchmarks in comparison to previous wav2vec2. We use the base version of 94.68 million parameters trained on librispeech 960 hour english data.

XLS-R[5] [20] is a multi-lingual representation learning model based on wav2vec2 architecture and trained on 128 languages. The training datasets comprises of BABEL, Voxlingua107, Commonvoice, MLS, and VoxPopuli. We use the 300 million parameters variant in our work.

Whisper[6] [21] is a multi-task learning SFM based on vanilla transformer encoder-decoder architecture. Pre-training is carried out in 680k hours of multilingual data in 96 languages and in a weakly supervised manner. Whisper shows improve performance than XLS-R for multilingual speech recognition. We utilize the base version of 74 millions parameters.

Massively Multilingual Speech (MMS)[7] [22] is based on the wav2vec2 architecture and pre-trained on apporximately 1400 languages. It uses around 500k hours of data for its pre-training including FLEURS, MMS-lab, BABEL and solves constrastive learning objective. We use the openly available 1 billion parameters variant.

x-vector [23] is specifically designed for speaker recognition and it is a time-delay neural network. However, x-vector has shown its effectiveness in related applications such as SER [11], shout intensity prediction [15], so we thought it might be helpful for CER and so, we include it in our experiments. It consists of 4.2 million parameters.

We use frozen SFMs as we want to understand their implicit capacity of understanding CER. We extract representations from the last hidden state through the use of mean pooling with dimensional size of 768 for WavlM, Unispeech-SAT, wav2vec2, HuBERT and 1280 for XLS-R, MMS. For Whisper and x-vector, the dimensions are 512, however, for Whisper, we extract it from the last hidden state of the encoder and discard the decoder.

III. Experiments

A. Benchmark Dataset

We utilize the only openly accesible dataset for CER, to the best of our knowledge given by Franzoni et al. [7] in our study. It was meticulously curated to capture the rich emotional expressions of crowds during high-attendance events, such as sports matches, concerts, political rallies, and public gatherings. We segment high-quality audio clips into 1-second blocks with a 0.25 second overlap, yielding a total of 9515 blocks from 69 original clips following Franzoni et al. [7]. Each block was assigned with emotional labels assigned to three distinct categories: *Approval* (cheering, clapping), *Disapproval* (booing, hissing), and *Neutral* (background chatter) as the original audio-clip label. More detailed statistics are presented in Table I. Further, we segment these 1-sec blocks into 500 milliseconds (ms), and 250 ms. However, before splitting into ms duration audios, we remove the silence as otherwise some silent audio might be present. We segment into short segments to understand the capability of SFMs for short-segment CER. All the audios were resampled to 16 KHz before passing it through the SFMs.

B. Downstream Modeling

We include both classical ML and DL networks as downstream networks with the SFMs. This includes SVM, Random Forest Classifier (RFC), Fully Connected Network (FCN), and

[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/hubert-base-ls960
[5]https://huggingface.co/facebook/wav2vec2-xls-r-300m
[6]https://huggingface.co/openai/whisper-base
[7]https://huggingface.co/facebook/mms-1b

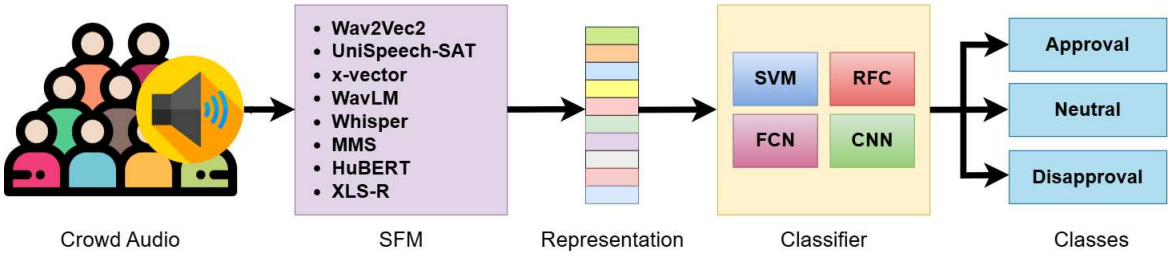
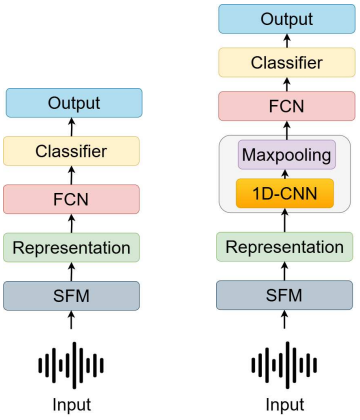


Fig. 1: Overview of the proposed system architecture for CER, illustrating the SFM, representation extraction, and classification processes

SFM	Wav2vec2		UNI		WavLM		HuBERT		x-vector		Whisper		MMS		XLS-R	
	A	F1	A	F1	A	F1	A	F1	A	F1	A	F1	A	F1	A	F1
1 sec																
SVM	96.94	90.14	95.34	84.83	93.08	73.28	97.87	92.03	95.80	84.55	98.61	95.88	98.75	95.59	98.35	94.04
RFC	95.35	79.80	95.24	80.67	93.38	76.03	94.07	77.55	95.63	85.24	98.71	94.96	97.45	92.51	96.48	88.81
FCN	97.29	91.62	96.00	85.75	94.19	78.60	97.87	92.03	97.64	92.80	98.93	96.10	99.06	96.81	98.35	94.04
CNN	97.51	92.73	96.40	87.17	95.19	84.32	98.09	93.29	98.23	94.61	98.97	96.64	**99.11**	**96.85**	98.71	95.39
500 ms																
SVM	96.93	89.97	95.69	85.71	92.67	66.04	97.28	90.26	95.56	86.69	98.99	96.47	99.00	97.60	98.41	94.85
RFC	96.32	86.32	95.71	82.19	92.85	71.57	94.17	75.63	94.57	76.80	97.69	92.49	96.99	90.78	97.00	90.13
FCN	97.17	90.02	98.61	94.43	93.45	77.08	98.07	92.52	96.48	88.17	98.99	95.81	99.10	96.61	98.68	94.94
CNN	97.44	91.67	95.52	84.87	94.22	81.00	98.23	94.71	97.58	91.67	98.90	96.17	**99.19**	**96.66**	98.90	95.63
250 ms																
SVM	95.38	82.83	94.96	83.01	92.48	58.18	97.57	91.86	93.62	81.57	98.81	95.61	98.84	96.34	97.91	92.92
RFC	92.93	72.87	92.22	70.74	91.16	64.46	95.24	80.68	93.44	72.37	97.98	93.65	97.57	91.86	98.20	93.55
FCN	95.66	84.86	94.83	81.77	93.25	74.70	97.69	92.05	94.08	79.66	98.95	96.21	98.87	96.47	98.44	94.63
CNN	96.20	87.25	95.20	83.14	94.08	79.66	97.88	92.38	95.02	83.17	98.97	96.58	**99.20**	**96.65**	98.44	94.63

TABLE II: Evaluation scores of models trained on different SFMs representations; A and F1 stand for accuracy and macro F1-score, respectively. All scores are averaged over five folds and presented as %; UNI stands for Unispeech-SAT. Light green opacity represents performance levels.



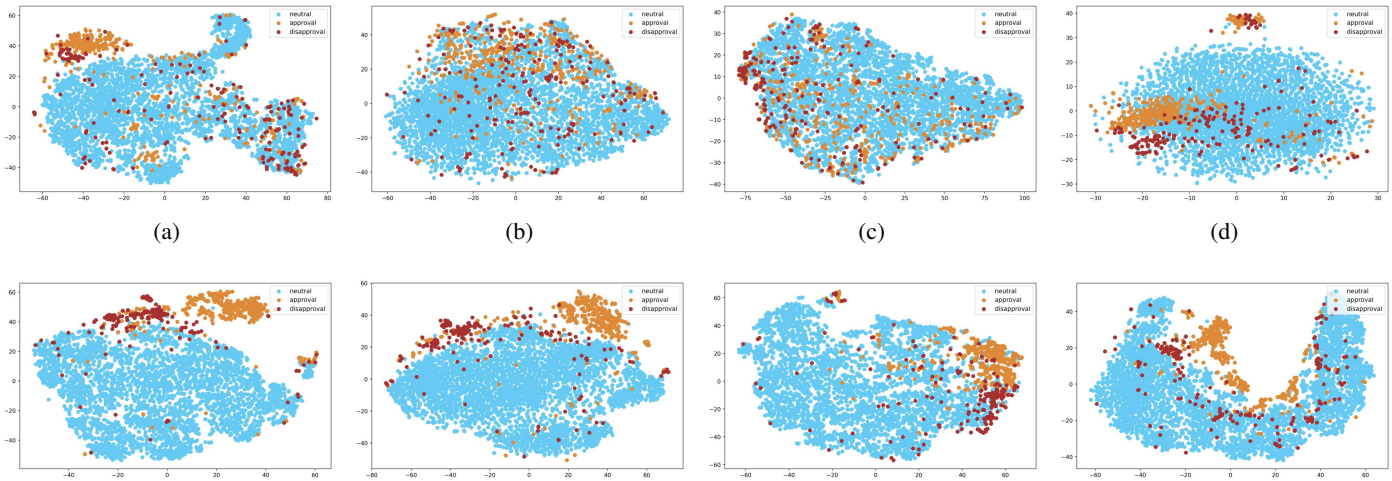
(a) FCN (b) CNN

Fig. 2: Model Architecture

CNN. For SVM and RFC, we kept the default hyperparameters. For CNN, it consists of three 1D-CNN layers: the first with 32 filters, the second with 64 filters, and the third with 128 filters, each with a filter size of 3x3, ReLU activation, and a stride of 1. After each convolutional layer, batch normalization and max-pooling with a pool size of 2x2 were applied. The output was flattened and passed through two fully connected layers, the first with 512 neurons and the second with 128 neurons, both using ReLU activation. A dropout layer with a rate of 0.5 was added between the layers to prevent overfitting. The final output layer used softmax activation with 3 neurons for classification. Similarly, the FCN model followed a similar structure, consisting of fully connected layers with 512, 256, and 128 neurons, each using ReLU activation, and a dropout rate of 0.5 was applied to prevent overfitting. The output layer in the FCN model also had 3 neurons with softmax activation. The design of our methodology is depicted in Figure 1. Detailed architectural details of FCN and CNN are given in Figure 2a and 2b. FCN models trainable parameters are from 1 to 3 millions followed by the CNN models with 2.5 to 4 millions.

C. Training Details

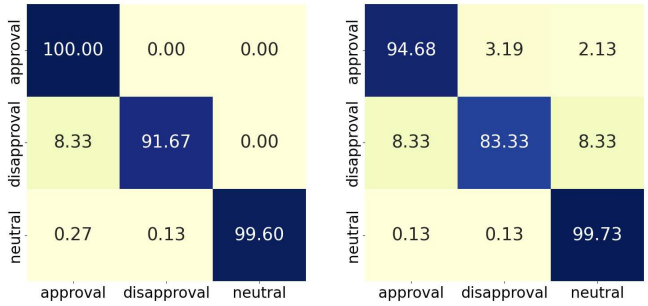
We use Adam as the optimizer and learning rate of 0.001. We employ categorical cross-entropy as loss. Training was conducted over 50 epochs with a batch size of 32. We use 5-fold cross-validation for training and testing. Here, 4 folds are used for training and one fold for testing. Additionally, we make use of early stopping for preventing overfitting.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3: t-SNE plot visualization of different SFMs: (a) Wav2vec2, (b) Unispeech-SAT, (c) WavLM, (d) x-vector, (e) Whisper, (f) MMS, (g) HuBERT, and (h) XLS-R



(a) (b)

Fig. 4: Confusion matrices for 1 sec duration: (a) MMS (b) WavLM; The y-axis represents True Values, while the x-axis denotes Predicted Values.

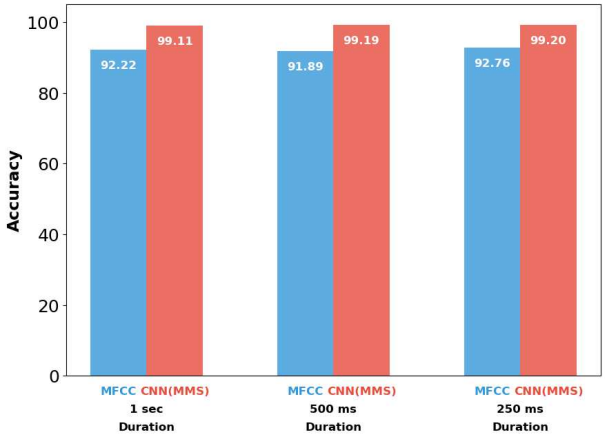


Fig. 5: Comparison of the best models with baseline MFCC; CNN (MMS) represents CNN model trained with MMS representations; Scores presented are in %

D. Experimental Results

Table II provides the evaluation scores of different downstream models trained on top of SFMs representations. For all the audio durations (1 sec, 500 ms and 250 ms), we can see that polyglot SFMs (XLS-R, Whisper, MMS) completely dominate other SFMs including monolingual and speaker recognition SFMs for CER. This validates our hypothesis that *polyglot SFMs will excel in CER due to their ability to capture diverse pitches, tones, and emotional variations. This stems from their pre-training on a broad spectrum of speech data, covering multiple languages, accents, and speaking styles.* Overall CNN models shows superior performance in comparison to other downstreams (SVM, RFC, FCN) with different SFMs in all the audio durations. For the 1-sec audio segments, MMS demonstrated the topmost performance, achieving an accuracy of 99.11% and an F1 score of 96.85% with CNN. Among the polyglot SFMs also, MMS is top and this can be attributed to its larger size of 1 billion parameters allowing it capture the acoustic characteristics required for CER in a much better manner. For 500 ms and 250 ms also, we observe clear dominance of MMS over other SFMs including polyglot SFMs. MMS reported accuracy of 99.19%, 99.20% and F1 score of 96.66%, 96.65% for 500 ms and 250 ms respectively with CNN.

Among the monolingual SFMs (Wav2vec2, Unispeech-SAT, WavLM, HuBERT), HuBERT showed the best results for all the audio durations. It reported accuracy of 98.09%, 98.23%, 97.88% and F1 score of 93.29%, 94.71%, 92.38% for 1 sec, 500 ms, and 250 ms respectively. This could point towards its ability in capturing the diverse crowd sounds in a superior manner. One interesting observation is the performance of x-vector. Despite being a very small SFM comprising of only 4.2 million paramters, it shows comparative performance in comparison with monolingual PTMs in some instances. This

behavior can be traced to its speaker recognition providing ability to capture diverse pitches, tones, and so on speech characteristics for effective CER. We visualize t-SNE plots of the raw representations from the last hidden state of the SFMs in Figure 3. These plots reveal clearer and more distinct clustering for polyglot SFMs across different emotional categories, amplifying the superior performance observed in the results. We also plot the confusion matrices of MMS and HuBERT in Figure 4 for 1 sec duration. These findings underscore the potential of leveraging polyglot SFMs to significantly enhance the performance of CER systems and establishing a solid foundation for future research.

Comparison with MFCC Baseline: As MFCC is one of the most used input representation for speech and audio processing and used by research works as a baseline for evaluating their proposed methods performance [24], [25], we also give a comparison of the best models with baseline MFCC features to understand the effectiveness of polyglot SFMs. We keep the modeling and training details same for experiments with MFCC as set for experiments with SFMs representations. The comparison is presented in Figure 5. The best model CNN(MMS) shows superior performance in comparison to baseline MFCC feature.

IV. Conclusion

In this study, we investigated the effectiveness of polyglot SFMs for CER. We hypothesized that polyglot SFMs, pre-trained on diverse languages, accents, and speech patterns, would be particularly adept at handling the noisy and complex acoustic environments of crowds, offering a distinct advantage for CER. Through extensive experiments on a benchmark CER dataset with varying audio durations (1 sec, 500 ms, and 250 ms), our analysis confirmed that polyglot SFMs consistently outperformed monolingual and speaker recognition SFMs, demonstrating superior performance even with short-duration inputs. These findings reinforce the potential of polyglot SFMs in CER and set a foundation for future research. Our study will also act as a guide for selection of SFMs for CER and related applications.

References

[1] M. W. Baig, E. I. Barakova, L. Marcenaro, M. Rauterberg, and C. S. Regazzoni, "Crowd emotion detection using dynamic probabilistic models," in *From Animals to Animats 13: 13th International Conference on Simulation of Adaptive Behavior, SAB 2014, Castellón, Spain, July 22-25, 2014. Proceedings 13*, Springer, 2014, pp. 328–337.

[2] K. G. Quach, N. Le, C. N. Duong, I. Jalata, K. Roy, and K. Luu, "Non-volume preserving-based fusion to group-level emotion recognition on crowd videos," *Pattern Recognition*, vol. 128, p. 108 646, 2022.

[3] X. Wang, D. Zhang, H.-Z. Tan, and D.-J. Lee, "A self-fusion network based on contrastive learning for group emotion recognition," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 2, pp. 458–469, 2022.

[4] J. Zhang, X. Wang, D. Zhang, and D.-J. Lee, "Semi-supervised group emotion recognition based on contrastive learning," *Electronics*, vol. 11, no. 23, p. 3990, 2022.

[5] Q. Zhu, Q. Mao, J. Zhang, X. Huang, and W. Zheng, "Towards a robust group-level emotion recognition via uncertainty-aware learning," *arXiv preprint arXiv:2310.04306*, 2023.

[6] M. Wu, L. Wang, and G. Li, "Crowd emotion recognition based on causal spatiotemporal structure," in *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 2022, pp. 368–374.

[7] V. Franzoni, G. Biondi, and A. Milani, "Emotional sounds of crowds: Spectrogram-based analysis using deep learning," *Multimedia tools and applications*, vol. 79, no. 47, pp. 36 063–36 075, 2020.

[8] M. A. A. Faisal, M. U. Ahmed, and M. A. R. Ahad, "Eslce: A dataset of emotional sounds from large crowd events," in *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, 2021, pp. 1–7.

[9] A. Anand, S. Jain, S. Sharma, *et al.*, "Pulse of the crowd: Quantifying crowd energy through audio and video analysis," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, 2024, pp. 308–314.

[10] 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.

[11] O. Chetia Phukan, A. Balaji Buduru, and R. Sharma, "Transforming the embeddings: A lightweight technique for speech emotion recognition tasks," in *Interspeech 2023*, 2023, pp. 1903–1907. DOI: 10.21437/Interspeech.2023-2561.

[12] 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.

[13] A. Arunkumar, V. Nileshkumar Sukhadia, and S. Umesh, "Investigation of ensemble features of self-supervised pretrained models for automatic speech recognition," in *Interspeech 2022*, 2022, pp. 5145–5149. DOI: 10.21437/Interspeech.2022-11376.

[14] S. Baroudi, T. Pellegrini, and H. Bredin, "Specializing self-supervised speech representations for speaker segmentation," in *Proc. Interspeech 2024*, 2024, pp. 3769–3773.

[15] T. Fukumori, T. Ishida, and Y. Yamashita, "Investigating the effectiveness of speaker embeddings for shout intensity prediction," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2023, pp. 1838–1842.

[16] 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.

[17] 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.

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

[20] A. Babu, C. Wang, A. Tjandra, *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143.

[21] 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.

[22] 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.

[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)*, IEEE, 2018, pp. 5329–5333.

[24] S. Rathod, M. Charola, A. Vora, Y. Jogi, and H. A. Patil, "Whisper features for dysarthric severity-level classification," in *Interspeech 2023*, 2023, pp. 1523–1527. DOI: 10.21437/Interspeech.2023-1891.

[25] 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.