ARE MAMBA-BASED AUDIO FOUNDATION MODELS THE BEST FIT FOR NON-VERBAL EMOTION RECOGNITION?

Mohd Mujtaba Akhtar*†, Orchid Chetia Phukan††, Girish†‡† Swarup Ranjan Behera§
Ananda Chandra Nayak¶, Sanjib Kumar Nayak‖, Arun Balaji Buduru*, Rajesh Sharma**††
*V.B.S.P.U, India, †IIIT-Delhi, India, ‡UPES, India, §Independent Researcher, India
¶KAC, India, ‖VSSUT, India, **University of Tartu, Estonia, ††Plaksha University,India
Correspondence:mmakhtar.research@gmail.com,orchidp@iiitd.ac.in

***Abstract*—In this work, we focus on non-verbal vocal sounds emotion recognition (NVER). We investigate mamba-based audio foundation models (MAFMs) for the first time for NVER and hypothesize that MAFMs will outperform attention-based audio foundation models (AAFMs) for NVER by leveraging its state-space modeling to capture intrinsic emotional structures more effectively. Unlike AAFMs, which may amplify irrelevant patterns due to their attention mechanisms, MAFMs will extract more stable and context-aware representations, enabling better differentiation of subtle non-verbal emotional cues. Our experiments with state-of-the-art (SOTA) AAFMs and MAFMs validates our hypothesis. Further, motivated from related research such as speech emotion recognition, synthetic speech detection, where fusion of foundation models (FMs) have showed improved performance, we also explore fusion of FMs for NVER. To this end, we propose, `RENO`, that uses renyi-divergence as a novel loss function for effective alignment of the FMs. It also makes use of self-attention for better intra-representation interaction of the FMs. With `RENO`, through the heterogeneous fusion of MAFMs and AAFMs, we show the topmost performance in comparison to individual FMs, its fusion and also setting SOTA in comparison to previous SOTA work.**

***Index Terms*—Non-Verbal Emotion Recognition, Mamba-based Audio Foundation Models, Attention-based Audio Foundation Models**

I. INTRODUCTION & RELATED WORK

Emotions play a fundamental role in human communication, shaping how we express ourselves and connect with others. They influence not only verbal interactions but also non-verbal cues, such as facial expressions, gestures, physiological signals, and vocalizations, which are crucial in conveying underlying feelings and intentions. However, non-verbal vocalizations offer a unique and often underexplored perspective. Sounds like laughter, cries, and sighs convey a broad spectrum of emotions that play a crucial role in communication, enhancing human interactions in daily life. Recognizing emotions from these non-verbal vocal cues has diverse applications in areas such as healthcare, human-computer interaction, customer service, and security, where understanding emotional context is vital for decision-making and improving user experience. Unlike speech emotion recognition (SER), which relies on language, these vocalizations bypass the need for verbal content and instead communicate emotions directly through their acoustic features. However, SER has been extensively studied in comparison to non-verbal vocal sounds emotion recognition (NVER).

Initial works on SER focused on the usage of handcrafted spectral features such as MFCC with classical ML algorithms such as SVM, HMM, GMM, decision tree [1], [2], [3]. Researchers have also leveraged handcrafted features with neural network-based approaches such as RNN, CNN, LSTM, Transformer [4], [5], [6]. However, with the advent of foundation models (FMs) in recent years, SER research has significantly shifted towards the use of FMs [7], [8], [9]. These models, pre-trained on large datasets, offer superior performance and the ability to capture complex acoustic patterns, reducing the need for manually engineered features and enabling more robust emotion recognition in diverse contexts. These FMs are either attention-based audio FMs (AAFMs) and mamba-based audio FMs (MAFMs). MAFMs built on state-space models (SSMs), which offer a computationally efficient alternative to traditional attention-based architectures. Unlike attention mechanisms that dynamically assign weights to input features, SSMs model sequences through structured recurrence, enabling long-range dependency capture while maintaining scalability. As such benefits, researchers have explored building audio FMs with mamba-based modeling architectures and they have shown superior or comparative performance in SER in comparison to AAFMs [10].

Despite much advancement in SER, research into NVER haven't seen much limelight except few prolific works [11], [12] despite carrying sufficient potential for emotion recognition. Audio FMs such as Wav2vec2, Whisper have also shown its efficacy for NVER [13]. However, the Audio FMs used in previous research are mostly AAFMs and previous works haven't investigated MAFMs for NVER and this leaves a gap towards understanding the potential of MAFMs for NVER. In this work, we focus on NVER and explore MAFMs for NVER, to the best of knowledge. *We*

† Contributed equally as a first authors

hypothesize that MAFMs will outperform attention-based AAFMs in NVER by leveraging their state-space modeling capabilities for better capture of intrinsic emotional structures. In contrast to AAFMs, which may amplify irrelevant patterns due to their attention mechanisms, MAFMs offer more stable and contextually aware representations, facilitating the differentiation of subtle non-verbal emotional cues. Our experiments, comparing state-of-the-art (SOTA) AAFMs and MAFMs, confirm the validity of our hypothesis. Furthermore, drawing inspiration from related fields such as speech emotion recognition [14] and synthetic speech detection [15], where the fusion of FMs has led to performance improvements, we investigate the fusion of FMs for NVER. To achieve this, we propose **RENO**, (**REN**yi Attenti**O**n Network), a novel framework to to effectively align the FMs. **RENO** incorporates self-attention mechanisms to enhance intra-representation interactions across the FMs representational spacne and utilizes Renyi-divergence as a novel loss function for inter-FM interaction. Through **RENO** with the heterogeneous fusion of MAFMs and AAFMs, we demonstrate superior performance outperforming both individual MAFMs, AAFMs, baseline fusion methods as well as homogeneous fusion of AAFMs, thus setting a new SOTA for NVER.

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

- For investigating the effectiveness of MAFMs for NVER, we present the first comprehensive comparative study of MAFMs and AAFMs. Our experiments results shows that MAFMs outperforms its attention-based counterparts.
- We propose, **RENO**, a novel framework that leverages self-attention for intra-FM interaction followed by the usage of Renyi Divergence loss for inter-FMs alignment. **RENO** with the heterogenous fusion of MAFMs and AAFMs shows the topmost performance in comparison to individual FMs, baseline fusion techniques, homogeneous fusion of AAFMs, and thus achieving SOTA across benchmark NVER datasets such as ASVP-ESD, JNV, and VIVAE.

The code and models developed in this study can be accessed at [1].

II. Foundation Models

In this section, we provide an overview of the FMs used in our study.

Audio-MAMBA[2] [10]: It is a selective SSM designed to learn general-purpose audio representations through self-supervised learning. It extracts information from randomly masked spectrogram patches. Pre-trained on the AudioSet dataset, it consistently shows comparable and sometimes better performance than AAFMs across various tasks, including SER. In our study, we use three versions of Audio-MAMBA: tiny (4.8M parameters), small (17.9M parameters), and base (69.3M parameters).

WavLM[3] [16]: It is a SOTA AAFM, ranked highly on the SUPERB benchmark. It uses masked speech modeling with denoising objectives during its pre-training. We employ the base version, which consists of 94.70M parameters and is trained on 960 hours of English speech from the LibriSpeech dataset.

UniSpeech-SAT[4] [17]: It is another SOTA AAFM on the SUPERB leaderboard. It follows a self-supervised pre-training approach with speaker-aware multi-task learning. We utilize the base version, which has 94.68M parameters and is trained on 960 hours of English speech from LibriSpeech.

Wav2vec2[5] [18]: It is an AAFM that employs contrastive self-supervised learning to learn speech representations. It masks segments of latent features and optimizes them through contrastive loss. We use its base version, which contains 95.04M parameters and is pre-trained on 960 hours of English speech from the LibriSpeech dataset.

HuBERT[6] [19]: It follows a self-supervised learning framework that iteratively refines its representations using k-means clustering while training on a BERT-style masked prediction objective. We employ the base version, which has 94.68M parameters and is trained on 960 hours of English speech from LibriSpeech.

Resampling to 16 KHz is done for the audio samples before passing it to the FMs. We obtain representations from the last hidden state of the frozen FMs including MAFMs and AAFMs using average pooling. The extracted representations have the following dimensions: 768 for WavLM, UniSpeech-SAT, Wav2Vec 2.0, and HuBERT; 1280 for MMS; and for Audio-MAMBA, 960 for the tiny version, 1920 for the small version, and 3840 for the base version.

III. Modeling Methodology

In this section, we discuss the downstream modeling used with the FMs followed by the discussion of the proposed framework, **RENO** for the fusion of FMs. We utilize FCN (Fully connected network) and CNN as downstream classifiers. The CNN model consists of two convolutional blocks consisting of 1D CNN layers with 32 and 64 filters, respectively with filter size of 3. Max-pooling is applied after each convolutional layer. The extracted features are then flattened and passed through a FCN block with two dense layers of 512 and 128 neurons followed by the output layer with softmax activation function. The output layer outputs probabilities for the emotion classes. The FCN follows the same modeling as the FCN block of the CNN model. CNN models training parameters varies between 0.9 to 1.1M while FCN models parameters between 0.7 to 1M depending on the dimension size of input representations.

[1] https://github.com/Helix-IIIT-Delhi/RENO-Non-Verbal

[2] https://github.com/SarthakYadav/audio-mamba-official?tab=readme-ov-file

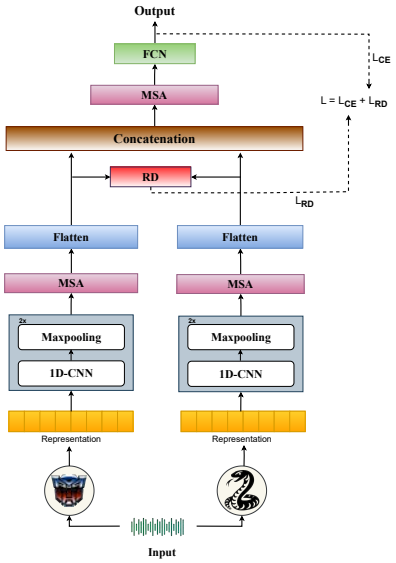


Fig. 1: Novel Framework: **RENO**; MSA stands for Multi-head Self-attention

FM	ASVP_ESD		JNV		VIVAE	
	A ↑	F1 ↑	A ↑	F1 ↑	A ↑	F1 ↑
FCN						
A (T)	66.51	58.71	59.82	58.14	52.21	51.95
A (S)	71.15	68.74	61.17	60.84	57.48	56.31
A (B)	72.64	69.47	62.28	60.97	58.99	58.57
W	52.69	42.86	57.46	56.21	32.54	31.82
W2	61.18	54.32	56.96	55.14	47.85	46.81
U	54.96	53.21	57.52	56.62	34.28	33.68
H	55.67	54.05	58.41	57.96	42.56	41.94
CNN						
A (T)	67.59	65.84	62.77	61.14	53.92	53.57
A (S)	72.90	70.64	63.10	61.42	60.41	59.93
A (B)	**73.96**	**71.98**	**64.29**	**63.81**	**62.42**	**61.29**
W	53.94	50.84	59.85	58.89	33.18	32.76
W2	62.96	60.84	59.27	57.14	48.39	48.18
U	55.41	54.16	60.71	59.20	36.69	35.91
H	59.43	58.21	59.26	58.35	49.04	48.36

TABLE I: Evaluation Scores are in %; A and F1 stands for Accuracy and macro-average F1 score; Audio-mamba (Tiny: A(T), Small: A(S), Base: A(B)), WavLM (W), Wav2vec2 (W2), and Unispeech-SAT (U); Evaluation scores are given in average across five folds; The abbreviations used are kept same for Table II

A. *RENO*

We propose **RENO**, a novel framework designed to align feature representations from distinct FMs. The architecture is given in Figure 1. **RENO** leverages self-attention mechanisms to strengthen intra-representation interactions within the feature space of FMs followed by employing Rényi divergence (RD) as a novel loss function to facilitate inter-FM interaction. Detailed walkthrough of the proposed framework is given as follows: The extracted representations from different FMs are first flattened after passing through convolutional block as used with individual FM representations above. We then passed the flattened features through a self-attention mechanism for better intra-representation interaction. Self-attention is calculated as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (1)$$

where d_k is the scaling factor. $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{Q}$ stands for query, key, value where $\mathbf{Q} = \mathbf{W}_Q\mathbf{X}$, $\mathbf{K} = \mathbf{W}_K\mathbf{X}$, $\mathbf{V} = \mathbf{W}_V\mathbf{X}$. $\mathbf{X}$ represents the input feature matrix and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable weight matrices. Then the features are passed through to RD loss. RD quantifies the difference or dissimilarity between two probability distributions [20]. Lower RD higher the similarity. Here in our study, we introduce RD as novel loss function for measuring the divergence between feature distributions of two FMs. Given two feature spaces e_a and e_b corresponding to two FMs, RD is computed as:

$$\mathcal{L}_{RD} = \frac{1}{\beta - 1}\log\left(\sum_{j=1}^{M}(z_{x,j} + \delta)^{\beta}(z_{y,j} + \delta)^{1-\beta}\right) \quad (2)$$

[3]https://huggingface.co/microsoft/wavlm-base
[4]https://huggingface.co/microsoft/unispeech-sat-base
[5]https://huggingface.co/facebook/wav2vec2-base
[6]https://huggingface.co/facebook/hubert-base-ls960

where M is the feature dimension, $\beta > 1$ controls the divergence order, and δ ensures numerical stability. RD will align the representation space of the FMs to a joint feature space and following this, the aligned features are then concatenated and passed through a self attention block. Self-attention will lead to further refinement of the fused features. Finally, the features are passed through a FCN block with two dense layers with each layer consisting of 512 and 128 neurons. The output layer leverages softmax as activation and outputs probabilities for the emotion classes. For joint optimization with cross-entropy loss $\mathcal{L}_{CE}$, we integrate RD loss $\mathcal{L}_{RD}$ with the $\mathcal{L}_{CE}$:

$$\mathcal{L} = \lambda\mathcal{L}_{CE} + (1 - \lambda)\mathcal{L}_{RD} \quad (3)$$

where λ is a hyperparameter controlling the trade-off between $\mathcal{L}_{CE}$ and $\mathcal{L}_{RD}$. The training parameters varies between 1.3 to 1.5M depending on the dimensional-size of the input representations. For self-attention blocks for each individual repesentation network, we keep the number of heads as 2 and for the self-attention block after concatenation, we set the number of heads also to 2.

IV. EXPERIMENTS & RESULTS

A. Benchmark Dataset

ASVP_ESD: [21]: It comprises thousands of high-quality audio recordings labeled with 12 distinct emotions, along with an additional "breath" category. These recordings were captured in real-world environments. In our study, we specifically utilize only the non-speech component. The audio samples were sourced from a variety of media, including movies, television shows, YouTube channels, and other online platforms.

JNV [22]: The dataset comprises 420 high-quality nonverbal vocalization samples, recorded from four native Japanese

Combinations	ASVP_ESD				JNV				VIVAE			
	Concat		RENO		Concat		RENO		Concat		RENO	
	A ↑	F1 ↑	A ↑	F1 ↑	A ↑	F1 ↑	A ↑	F1 ↑	A ↑	F1 ↑	A ↑	F1 ↑
AAFMs + MAFMs												
A (T)+W	70.32	68.91	75.42	74.23	66.14	65.43	71.45	70.32	54.29	53.21	61.65	60.45
A (T)+W2	68.54	67.74	74.23	73.56	59.42	58.92	64.54	63.76	56.61	55.03	62.34	61.32
A (T)+U	71.89	70.45	77.34	76.12	65.45	64.41	72.45	71.87	56.87	55.91	63.76	62.43
A (T)+H	74.41	73.71	78.71	77.89	**76.56**	**75.43**	**79.09**	**78.41**	56.53	55.72	61.45	61.03
A (S)+W	75.08	74.78	81.65	80.23	66.31	65.79	71.67	70.23	64.43	62.99	73.65	72.49
A (S)+W2	73.06	72.32	79.34	78.31	60.56	59.23	67.29	66.31	64.21	63.78	70.31	69.63
A (S)+U	73.67	72.65	82.44	81.34	63.98	62.54	68.43	67.60	64.42	63.02	71.34	70.56
A (S)+H	75.23	74.62	76.20	75.45	74.37	73.11	77.56	76.83	57.83	56.72	62.47	61.43
A (B)+W	75.21	74.47	83.56	82.54	66.74	65.27	72.56	71.09	63.70	62.65	72.76	71.65
A (B)+W2	75.54	74.78	82.65	81.34	67.62	66.09	72.56	71.50	61.03	60.34	68.93	67.32
A (B)+U	**75.67**	**74.84**	**83.56**	**82.51**	66.72	65.87	69.70	69.04	61.94	60.13	68.53	67.08
A (B)+H	75.09	74.21	77.98	76.48	67.88	66.60	71.90	71.62	**67.69**	**66.51**	**72.51**	**71.82**
AAFMs + AAFMs												
W+W2	63.93	62.32	71.45	70.59	62.01	61.97	69.39	68.51	48.39	47.93	59.31	58.47
W+U	56.73	55.62	67.31	66.10	64.34	63.97	73.41	72.89	38.56	37.54	47.99	46.81
W2+U	65.32	64.75	74.78	73.01	64.53	63.78	71.43	70.89	52.83	51.93	63.42	62.90
H+U	62.57	61.71	66.29	65.72	61.54	60.78	68.21	67.45	54.50	53.02	56.06	55.22
H+W2	64.72	63.90	67.53	66.01	63.65	62.73	69.78	68.65	52.41	51.63	59.71	58.64
H+W	59.72	58.34	63.09	62.21	67.72	66.49	74.51	72.57	55.60	52.71	64.63	62.20

TABLE II: Evaluation scores are in % and average of five folds

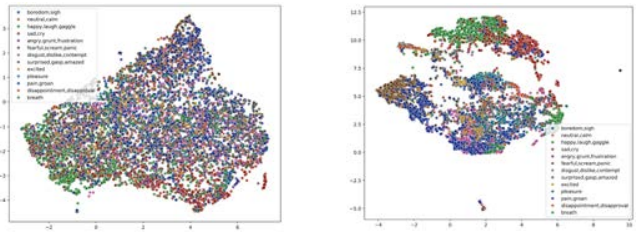
speakers (two male, two female). The dataset covers six distinct emotions—anger, disgust, fear, happiness, sadness, and surprise—and includes 87 unique phrases.

VIVAE [23]: The dataset comprises 1085 high-quality audio samples of non-speech vocalizations, recorded from eleven female speakers in a controlled studio environment. Each sample captures one of six distinct emotional states—achievement/triumph, sexual pleasure, surprise, anger, fear, and physical pain—expressed at four intensity levels: low, moderate, strong, and peak.

Training Details: The models were trained using Adam optimizer with learning rate of 1e-3, a batch size of 32, and 20 epochs. It uses cross-entropy as the classification loss. We use dropout and early stopping for preventing overfitting. For experiments with **RENO**, we fix $\beta = 2$, $\delta = 0.2$, and $\lambda = 0.4$ for all experiments, as preliminary exploration indicated that these values yielded the best results. We follow five fold cross validation for training our models where four folds are used for training and one fold for testing.

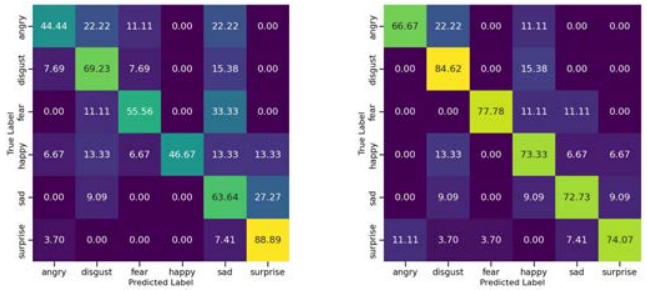
B. Results and Discussion

The evaluation scores for downstream models trained on SOTA MAFMs and AAFMs are given in Table I. Our results reveals that MAFMs consistently outperform AAFMs for NVER achieving the highest accuracy and F1-score across all datasets. The superior performance of MAFMs is attributed to their structured state-space modeling, which efficiently captures long-range dependencies and provides more stable and context-aware emotional representations and thus proving our hypothesis. Among the MAFMs, the Audio-mamba (Base) showed the best performance and this can be due to its larger size in comparison to small and tiny. One interesting observation is that despite Audio-mamba (Tiny) is of 4.8M, it is able to beat large AAFMs. This further amplifies our hypothesis that MAFMs will be the most effective for NVER. Overall, the CNN models showed better performance than FCN models. Among the AAFMs, we observe mix performance, with some AAFMs performing better in one dataset and some in other dataset. We also plot the t-SNE plots of raw representations of Audio-mamba (Base) and Wav2vec2 in Figure 2. We observe better cluster for Audio-mamba (Base), thus amplifying our results.



(a) Wav2vec2 (b) Audio-mamba (Base)

Fig. 2: t-SNE plots for ASVP-ESD



(a) (b)

Fig. 3: Confusion Matrix for JNV dataset; Subfigures (a) Audio-Mamba (Base) with CNN (b) Fusion of Audio-mamba (Tiny) and HuBERT through **RENO**

Table II presents the evaluation scores for different combinations of FMs. We refrain from combining the MAFMs, as it the same model except slight difference in architecture and pre-training data. We use concatenation-based fusion as the baseline fusion technique. We follow the same architectural details as **RENO** except the renyi rivergence loss and the self-attention blocks. We also keep the training details same as **RENO**. Our results shows that fusion with **RENO** leads to better performance in comparison to concatenation-based fusion technique as well as individual MAFMs and AAFMs. We also observe that heterogeneous fusion of MAFMs and AAFMs generally leads to improved performance in comparison to homogeneous fusion of AAFMs. This points out towards observable emergence of complementary behavior as both of them have their own unique strengths. With this heterogeneous fusion of MAFMs and AAFMs through **RENO**, we set the new SOTA for NVER. However, there is no clear winner among which is the best pair for all the NVER datasets considered, as a particular pair shows the topmost in one dataset and some other pair in another dataset. For example, combination of Audio-Mamba (Base) and Unispeech-SAT with **RENO** is leading in ASVP-ESD but Audio-Mamba (Base) and HuBERT is top in VIVAE. This behavior is most possibly to the dependence on downstream data distribution variability. We plot the confusion matrices of CNN model built on Audio-mamba (Base) and fusion of Audio-mamba (Tiny) with HuBERT through **RENO** in Figure 3.

V. CONCLUSION

In this work, we establish the potential of MAFMs for NVER, demonstrating their superiority over AAFMs in capturing intrinsic emotional structures through state-space modeling. Unlike AAFMs, which may amplify irrelevant patterns, MAFMs provide stable, context-aware representations, enhancing the recognition of subtle non-verbal emotional cues. Our experiments validate this hypothesis, showing MAFMs outperforming SOTA AAFMs. Additionally, inspired by advances in SER and synthetic speech detection, we explore FM fusion for NVER. To this end, we introduce **RENO** for effective fusion of FMs. Through the heterogeneous fusion of MAFMs and AAFMs, **RENO** achieves the best performance, surpassing individual FMs, fusion baselines, and sets SOTA for NVER. Our study will act as a baseline for future research exploring FMs for NVER.

REFERENCES

[1] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *2005 international conference on machine learning and cybernetics*, vol. 8. IEEE, 2005, pp. 4898–4901.

[2] K. V. K. Kishore and P. K. Satish, "Emotion recognition in speech using mfcc and wavelet features," *2013 3rd IEEE International Advance Computing Conference (IACC)*, pp. 842–847, 2013.

[3] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[4] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6474–6478.

[5] M. D. Pawar and R. D. Kokate, "Convolution neural network based automatic speech emotion recognition using mel-frequency cepstrum coefficients," *Multimedia Tools and Applications*, vol. 80, pp. 15 563–15 587, 2021.

[6] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[7] L. Pepino, P. E. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *ArXiv*, vol. abs/2104.03502, pp. 3400–3404, 2021.

[8] O. C. Phukan, A. B. Buduru, and R. Sharma, "Transforming the embeddings: A lightweight technique for speech emotion recognition tasks," *arXiv preprint arXiv:2305.18640*, 2023.

[9] D. Diatlova, A. Udalov, V. Shutov, and E. Spirin, "Adapting wavlm for speech emotion recognition," in *Proc. odyssey 2024*, 2024, pp. 303–308.

[10] S. Yadav and Z.-H. Tan, "Audio mamba: Selective state spaces for self-supervised audio representations," in *Interspeech 2024*, 2024, pp. 552–556.

[11] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1675–1686, 2021.

[12] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, vol. 12, pp. 19 752–19 764, 2024.

[13] P. Tzirakis, A. Baird, J. Brooks, C. Gagne, L. Kim, M. Opara, C. Gregory, J. Metrick, G. Boseck, V. Tiruvadi *et al.*, "Large-scale nonverbal vocalization detection using transformers," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] Y. Wu, P. Yue, C. Cheng, and T. Li, "Investigation of ensemble of self-supervised models for speech emotion recognition," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2023, pp. 988–995.

[15] O. Chetia Phukan, G. Kashyap, A. B. Buduru, and R. Sharma, "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 2496–2506. [Online]. Available: https://aclanthology.org/2024.findings-naacl.160/

[16] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *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, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "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.

[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] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[21] D. Landry, Q. He, H. Yan, and Y. Li, "Asvp-esd: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances," *Global Scientific Journals*, vol. 8, pp. 1793–1798, 2020.

[22] D. Xin, S. Takamichi, and H. Saruwatari, "Jnv corpus: A corpus of japanese nonverbal vocalizations with diverse phrases and emotions," *Speech Commun.*, vol. 156, p. 103004, 2023.

[23] N. Holz, P. Larrouy-Maestri, and D. Poeppel, "The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception." *Emotion*, vol. 22, no. 1, p. 213, 2022.