

# PARROT: Synergizing Mamba and Attention-based SSL Pre-Trained Models via Parallel Branch Hadamard Optimal Transport for Speech Emotion Recognition

*Orchid Chetia Phukan<sup>\*1</sup>, Mohd Mujtaba Akhtar<sup>\*1,2</sup>, Girish<sup>\* 1,3</sup>, Swarup Ranjan Behera<sup>4</sup>, Jaya Sai Kiran Patibandla<sup>4</sup>, Arun Balaji Buduru<sup>1</sup>, Rajesh Sharma<sup>5,6</sup>*

<sup>1</sup>IIT-Delhi, India, <sup>2</sup>V.B.S.P.U, India, <sup>3</sup>UPES, India, <sup>4</sup>Independent Researcher, India, <sup>5</sup>University of Tartu, Estonia, <sup>6</sup>Plaksha University, India

*Correspondence:* orchidp@iitd.ac.in

## Abstract

The emergence of Mamba as an alternative to attention-based architectures has led to the development of Mamba-based self-supervised learning (SSL) pre-trained models (PTMs) for speech and audio processing. Recent studies suggest that these models achieve comparable or superior performance to state-of-the-art (SOTA) attention-based PTMs for speech emotion recognition (SER). Motivated by prior work demonstrating the benefits of PTM fusion across different speech processing tasks, we hypothesize that leveraging the complementary strengths of Mamba-based and attention-based PTMs will enhance SER performance beyond the fusion of homogenous attention-based PTMs. To this end, we introduce a novel framework, **PARROT** that integrates parallel branch fusion with Optimal Transport and Hadamard Product. Our approach achieves SOTA results against individual PTMs, homogeneous PTMs fusion, and baseline fusion techniques, thus, highlighting the potential of heterogeneous PTM fusion for SER.

**Index Terms:** Speech Emotion Recognition, Pre-Trained Models, Mamba-based Models, Attention-based Models

## 1. Introduction

Speech Emotion Recognition (SER) bridges human-computer interaction, finds applications in mental health monitoring as well as in empathetic AI systems [1, 2]. It enables machines to understand and respond to human emotions, fostering more natural and intuitive interactions. Traditional SER research often employs handcrafted features such as MFCCs, which capture the spectral properties of speech and have proven effective in representing emotional cues. These features were initially modelled with classical ML techniques such as SVM [3], tree-based methods [4]. This was followed by the use of deep learning techniques [5, 6].

By the end of last decade, the landscape of SER research has changed for the better through the use and the wide scale availability of self-supervised learning (SSL) pre-trained models (PTMs). These PTMs trained on large-scale diverse data provides performance benefits as well as discard the necessity of training models from scratch. These PTMs have led to significant development in SER. As such, researchers have explored various state-of-the-art (SOTA) PTMs [7, 8, 9, 10]. Pepino et al. [11] used wav2vec2 with LSTM and CNN downstreams and showed its effectiveness in comparison to conventional features such as eGeMAPS and spectrogram. Morais et al. [12] gave a comprehensive comparison of various SSL PTMs such as wav2vec2, HuBERT with different downstream networks. These PTMs have predominantly utilized attention-based architectures,

which excel at capturing contextual dependencies in speech signals.

In recent times, an alternative architecture type of PTMs has captured attention in the community: mamba-based PTMs [13, 14, 15]. These mamba-based PTMs are based on top of mamba architecture which is structured state-space model (SSM) and has emerged as a promising alternative due to its ability to efficiently model long-range dependencies with linear complexity [16]. These mamba-based PTMs have set benchmarks in sequence modeling tasks and initial research suggest that mamba-based PTMs perform competitively with, or even surpass, attention-based PTMs in SER [13]. Prior research has also shown the benefits of PTM fusion for SER [17] due to their complementary behavior. This is also observed across various speech processing tasks such as speech recognition [18] and speech deepfake detection [19]. However, previous works have integrated only attention-based models.

In this study, we fuse mamba and attention-based PTMs for SER and hypothesize that the fusion of these heterogeneous PTMs will yield richer and more robust representations for improved SER as attention-based PTMs will capture intricate global dependencies while mamba-based PTMs excels at efficient long-range processing. We are the first study to the best of our knowledge to explore fusion of such heterogeneous PTMs for SER. To our end, we propose, **PARROT** (PARallel BRanch Hadamard Optimal Transport), a novel framework to align and integrate heterogeneous mamba and attention-based PTMs. It employs parallel branch fusion, incorporating the hadamard product and optimal transport. Hadamard product captures local interactions by performing element-wise operations between the representations, preserving fine-grained details. Meanwhile, Optimal Transport operates at a global scale, aligning the distributions of features across the two PTMs, ensuring that the fused representations are coherent and effectively integrated for improved performance. The key contributions of our study are as follows:

- We introduce **PARROT**, a novel framework that encompasses parallel branch fusion with Hadamard Product and Optimal Transport that inherently captures local interactions and global interaction for improved SER performance.
- With **PARROT** through the synergy of mamba and attention-based PTMs, we achieve the topmost most performance across different SER datasets (CREMA-D (*English*), emo-DB (*German*), MESD (*Mexican Spanish*)) than individual PTMs and homogeneous fusion of attention-based PTMs. It also reports better performance in comparison to baseline fusion methods. These PTMs are SOTA SSL PTMs for SER, thus, **PARROT** achieves SOTA performance for SER with its heterogeneous fusion.

<sup>\*</sup> Contributed equally as a first authors.

All code and models used in this study are accessible at: <https://>

## 2. Pre-Trained Models

In this section, we discuss the PTMs considered in our study. **Audio-MAMBA** [13]<sup>1</sup>: Audio Mamba is a selective state space model that is trained in a self-supervised fusion to learn general-purpose representations from randomly masked spectrogram patches. Trained on the AudioSet dataset, it outperforms its attention-based counterparts baselines across diverse speech and audio tasks including SER. We use the tiny, small, and base versions of 4.8M, 17.9M, and 69.3M parameters.

**WavLM** [20]<sup>2</sup>: It is a SOTA attention-based PTM on SUPERB that integrates masked speech modeling with denoising objectives during its pre-training, effectively learning robust representations from noisy and clean speech alike. We have used the base version with 94.70M parameters and trained on librispeech 960 hours of english speech.

**UniSpeech-SAT** [21]<sup>3</sup>: It is also a SOTA attention-based PTM on SUPERB and trained in a self-supervised fashion with speaker-aware multi-task learning. We utilize the base version of 94.68M parameters and pre-trained on 960 hours of librispeech english data.

**Wav2vec2** [22]<sup>4</sup>: This contrastive SSL attention-based PTM that learns speech representations by masking segments of latent features. We use its base version trained on librispeech 960 hours english data with 95.04M parameters. Wav2vec2 improves previous SOTA methods in speech recognition.

**HuBERT** [23]<sup>5</sup>: HuBERT employs a SSL framework that iteratively refines its representations using k-means clustering and solves a BERT-like masked prediction objective. HuBERT improves over Wav2vec2 in speech recognition. We use the base version trained on english 960 hours librispeech data with 94.68M parameters.

**Massively Multilingual Speech (MMS)** [24]<sup>6</sup>: It is attention-based SSL PTM built on top of Wav2vec2 architecture. It extends pre-training to almost 1400 languages. It improves over XLS-R and Whisper in various multilingual speech processing. We use the 1B parameters version in our experiments.

We extract representations from the last hidden state of the frozen mamba and attention-based PTMs by pooling average. We representations are of dimensions: 768 for WavLM, Unispeech-SAT, Wav2vec2 and HuBERT; 1280 for MMS; 960 for tiny, 1920 for small, 3840 for base versions of Audio-MAMBA.

## 3. Modeling Pipeline

In this section, we discuss the downstream modeling networks to be employed with individual PTMs and the proposed framework for aligning PTMs, **PARROT**. We make use of SVM, Fully Connected Network (FCN), and CNN as the downstreams modeling with individual PTMs. For SVM, we kept the default hyperparameters. For CNN, we make use of two 1D convolutional layers with 64 and 128 filters, respectively, and a kernel size of 3 with ReLU activation function. After each 1D convolutional layer we attach a maxpooling. The output is then flattened and passed through a FCN containing a dense layer with 128 neurons and

ReLU activation, followed by a softmax layer for multi-class classification. For FCN, we keep the modeling same as FCN used in CNN downstream.

### 3.1. PARROT

We propose **PARROT**, a novel framework for integrating PTMs. The architecture of the proposed framework is presented in Figure 1. First, the extracted representations from both PTMs are processed through two 1D convolutional blocks with the same number of filters as used in Individual representation modeling above. Also, the rest modeling remains same. After flattening, the outputs are linearly projected into a 120-dimensional latent space, ensuring computational efficiency while retaining expressive information. Then, **PARROT** employs a parallel branch fusion strategy that combines local feature interactions via Hadamard product (HP) and global distribution alignment via Optimal Transport (OT). HP performs element-wise multiplication between the representations  $\mathbf{R}_p$  and  $\mathbf{R}_q$  of the PTMs and is given by  $\mathbf{HP} = \mathbf{R}_p \odot \mathbf{R}_q$ , where  $\odot$  denotes element-wise multiplication. While HP fusion retains structural details, it does not ensure global coherence between the PTMs. To address this, we employ OT to align the distributions of feature representations from PTMs. First, we compute the cost matrix  $C$  using the normalized Euclidean distance between the PTMs feature matrices  $\mathbf{R}_p$  and  $\mathbf{R}_q$ :

$$C = \frac{\|\mathbf{R}_p - \mathbf{R}_q\|_2}{\max(\|\mathbf{R}_p - \mathbf{R}_q\|_2)} \quad (1)$$

To efficiently compute the transport plan, we apply the Sinkhorn algorithm. The OT plan  $\Gamma$  is then derived as:  $\Gamma = \text{Sinkhorn}(C)$ . Using  $\Gamma$ , we transport features between PTMs to enforce distributional alignment, mapping  $\mathbf{R}_p$  into  $\mathbf{R}_q$ 's space and vice versa:  $\mathbf{R}_p \rightarrow \mathbf{R}_q = \Gamma \cdot \mathbf{R}_p$ ,  $\mathbf{R}_q \rightarrow \mathbf{R}_p = \Gamma^T \cdot \mathbf{R}_q$ . These transported features are then concatenated with the original representations to form the final fused representations  $\mathbf{F} = \text{Concat}(\mathbf{F}_q, \mathbf{F}_p)$ :  $\mathbf{F}_q = \text{Concat}(\mathbf{R}_P \rightarrow \mathbf{R}_q, \mathbf{R}_q)$ ,  $\mathbf{F}_p = \text{Concat}(\mathbf{R}_q \rightarrow \mathbf{R}_q, \mathbf{R}_p)$ . Finally, the fused features from both branches are concatenated and passed through a FCN with a dense layer of 128 neurons with softmax activation function for the final prediction. By preserving both local feature interactions and enforcing global alignment, **PARROT** enables robust fusion of PTMs. The trainable parameters of **PARROT** vary between 3.2M and 13M.

## 4. Experiments

### 4.1. Benchmark Datasets

**Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)** [25] serves as a widely recognized benchmark for SER and comprising 7,442 utterances from 48 male and 43 female actors, it spans a diverse range of speaker ages and ethnicities. This dataset includes six distinct emotional categories: anger, happiness, sadness, fear, disgust, and neutral, with each actor contributing 12 unique sentences. **German Emotional Speech Database (Emo-DB)** [26] is German benchmark SER database and containing 535 utterances from ten actors (five male and five female). The dataset features seven emotional states: anger, anxiety/fear, boredom, disgust, happiness, neutral, and sadness. **The Mexican Emotional Speech Database (MESD)** [27] is Mexican-spanish database containing 864 utterances representing six emotional states: anger, disgust, fear, happiness, neutral, and sadness.

<sup>1</sup><https://github.com/SarthakYadav/audio-mamba-official?tab=readme-ov-file>

<sup>2</sup><https://huggingface.co/microsoft/wavlm-base>

<sup>3</sup><https://huggingface.co/microsoft/unispeech-sat-base>

<sup>4</sup><https://huggingface.co/facebook/wav2vec2-base>

<sup>5</sup><https://huggingface.co/facebook/hubert-base-1s960>

<sup>6</sup><https://huggingface.co/facebook/mms-1b>

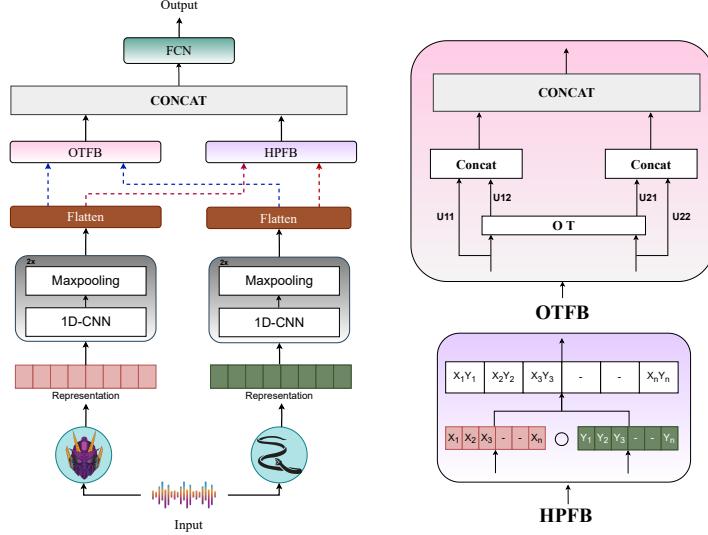


Figure 1: *Proposed Framework: PARROT; OTFB, HPFB stands for Optimal Transport Fusion Block and Hadamard Product Fusion Block respectively;  $U_{11}$ ,  $U_{22}$  represents the representational space of PTM 1 and PTM2;  $U_{12}$ ,  $U_{21}$  represents the transported space of PTM 2 to PTM 1 and vice versa*

**Training Details:** All models are trained using the Adam optimizer with cross-entropy loss. The learning rate is 1e-3, batch size 32, and training runs for 50 epochs. To mitigate overfitting, we use dropout and early stopping. We follow five-fold cross-validation, with four folds for training and one for testing.

PTM's	CREMA-D		Emo-DB		MESD	
	Acc	F1	Acc	F1	Acc	F1
<b>SVM</b>						
A(T)	60.96	59.85	68.96	68.10	70.96	69.63
A(S)	68.96	67.10	78.96	77.65	71.63	70.85
A(B)	66.99	65.25	81.65	80.94	75.69	74.62
W	61.92	60.36	85.09	84.63	45.96	44.12
H	64.25	63.98	86.16	85.31	60.94	59.76
W2	59.86	58.23	86.31	85.93	61.37	60.88
U	62.93	61.09	77.69	76.38	33.95	32.16
M	66.94	65.28	70.11	69.89	81.03	80.07
<b>FCN</b>						
A(T)	62.96	61.85	70.96	69.36	72.96	71.85
A(S)	69.52	68.20	80.66	79.63	76.93	75.11
A(B)	68.41	67.11	83.63	82.89	77.92	76.13
W	62.96	61.08	87.85	86.36	47.98	46.21
H	66.29	65.85	87.33	86.21	61.20	60.96
W2	61.01	60.36	88.63	87.03	62.78	61.96
U	64.82	63.33	79.65	78.51	36.96	35.21
M	67.52	66.36	71.39	70.88	82.66	81.39
<b>CNN</b>						
A(T)	63.60	63.60	71.03	70.14	73.99	73.93
A(S)	69.51	69.49	81.31	79.27	78.61	78.53
A(B)	69.91	69.90	84.11	82.90	78.03	77.96
W	67.96	66.25	89.14	88.69	48.55	48.16
H	69.95	68.23	88.26	87.11	62.43	62.43
W2	61.18	60.88	90.01	89.62	63.01	63.24
U	65.28	64.11	81.36	80.96	38.15	37.91
M	68.25	67.96	72.90	66.24	83.24	83.10

Table 1: *Evaluation Scores; Scores are in % and average of five-folds: Abbreviations used are: Audio-mamba (Tiny A(T), Small A(S), Base A(B)), WavLM (W), HuBERT (H), Wav2vec2 (W2), Unispeech-SAT (U), and MMS (M); Acc, F1 stands for Accuracy and macro average F1 score; Abbreviations used in this Table 1 are kept same for Table 2*

## 4.2. Experimental Results

Table 1 presents the results for individual PTMs with different downstream networks. CNN models generally outperforms SVM and FCN across most models and datasets, with the highest accuracy and F1 scores. FCN performs better than SVM, showing that neural models are better as downstreams with individual PTMs. Among Mamba PTM variants, the base version consistently outperforms the tiny and small versions across all datasets. Its superior performance likely stems from its larger size, enabling better capture of contextual dependencies essential for SER. Also, we can see that PTMs with different downstreams shows variations in results. Such behavior is also reported by previous research [28]. Among the attention-based PTMs, we observe mixed behavior with some PTMs leading in one dataset and some PTMs in other. This brings out limelight the effect of downstream data distribution on the performance of the downstream task. The top performance of MMS in MESD can be traced back to its multilingual pre-training as most of the other PTMs are trained on only English data. However, that's not the case in every scenario, as MMS due to its multilingual pre-training should have good performance in Emo-DB, but it reports one of the lowest performances in comparison to other PTMs. Also, excluding MMS the other attention-based PTMs reported low performance in MESD. In contrast, the some of the attention-based PTMs showed better results than its mamba counterparts in Emo-DB. Overall, there is no clear champion, that mamba or attention-based PTMs are best for SER reinforcing the importance of dataset-specific characteristics in determining the effectiveness of a PTM.

Table 2 shows the results of combinations of different PTMs. We use concatenation-based fusion as the baseline fusion technique. For modeling the concatenation-based fusion, we removed the optimal transport and hadamard product parallel branches from **PARROT** (Figure 1). We kept the rest modeling same and also the training details for fair comparison. Combinations of PTMs through **PARROT** generally shows better performance in comparison to baseline concatenation-based fusion technique, thus, showing its strength for effective fusion. Fur-

	CREMA-D				Emo-DB				MESD			
	Concatenation		PARROT		Concatenation		PARROT		Concatenation		PARROT	
Fusion	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
A(T)+W	63.37	62.89	64.38	63.38	76.96	75.61	78.69	77.25	44.38	43.81	46.85	45.28
A(T)+H	65.54	64.86	66.82	65.76	76.28	75.58	77.62	76.28	44.85	43.38	45.28	44.39
A(T)+W2	60.03	59.94	61.76	60.58	75.14	74.61	76.93	75.58	45.03	44.97	46.92	45.22
A(T)+U	62.56	61.14	63.94	62.28	74.94	73.39	75.58	75.16	44.34	43.08	45.34	44.28
A(T)+M	62.95	61.59	63.86	62.47	76.85	75.28	77.78	76.94	45.37	44.65	46.39	45.80
A(S)+W	63.48	62.94	64.45	63.94	77.26	76.18	79.54	78.51	46.88	45.82	47.58	46.64
A(S)+H	62.14	61.19	63.68	62.24	77.36	76.52	78.82	77.34	45.28	44.28	46.62	45.82
A(S)+W2	61.94	60.68	62.34	61.58	76.95	75.28	77.98	76.25	46.82	45.25	47.68	46.62
A(S)+U	61.64	60.17	62.84	61.34	75.82	74.36	76.52	75.94	44.29	43.57	46.94	45.82
A(S)+M	63.13	62.29	64.86	63.34	74.22	73.58	75.36	74.15	46.82	45.28	48.96	47.28
A(B)+W	63.96	62.10	67.56	67.44	79.41	78.64	80.37	78.56	47.31	46.85	49.13	48.87
<b>A(B)+H</b>	<b>70.63</b>	<b>69.79</b>	<b>73.68</b>	<b>72.90</b>	<b>89.92</b>	<b>88.24</b>	<b>92.24</b>	<b>91.53</b>	<b>58.31</b>	<b>56.08</b>	<b>59.54</b>	<b>58.94</b>
A(B)+W2	69.96	69.21	71.98	70.94	88.94	87.64	89.44	88.57	56.37	55.39	57.23	56.64
A(B)+U	61.27	60.96	62.53	62.75	73.46	72.68	74.77	73.17	37.91	36.49	38.15	36.77
A(B)+M	61.25	60.97	63.26	63.24	65.96	64.05	66.36	57.29	<b>66.37</b>	<b>65.28</b>	<b>69.05</b>	<b>68.72</b>
W+H	68.99	67.64	69.91	69.85	80.96	79.64	81.31	79.90	53.94	52.17	54.34	54.12
W+W2	67.96	66.21	68.30	68.23	86.94	85.61	87.85	87.70	54.96	53.29	55.49	55.14
W+U	66.93	65.07	67.90	67.71	76.68	75.61	77.57	77.52	48.64	47.34	49.71	49.59
W+M	65.39	64.07	67.90	67.84	87.96	86.09	88.29	87.54	53.94	52.18	54.34	54.20
H+W2	69.37	68.13	70.52	69.32	77.91	76.63	78.50	76.97	53.94	52.28	54.34	53.47
H+U	69.58	68.37	70.61	69.51	73.49	72.94	74.77	74.13	57.49	56.19	58.96	58.96
H+M	69.37	68.96	71.52	70.22	87.91	86.34	88.69	87.33	64.94	63.28	65.32	64.74
W2+U	67.94	66.39	68.96	67.16	75.39	74.17	76.64	76.37	57.19	56.41	58.38	58.51
W2+M	64.19	63.61	65.21	65.01	80.91	79.31	81.52	80.96	72.14	71.39	71.10	70.96
U+M	68.34	67.36	69.63	67.45	73.91	72.33	74.63	73.93	49.68	48.31	50.29	49.50

Table 2: Evaluation Scores of different PTM combinations; Scores are in % and average of five-folds

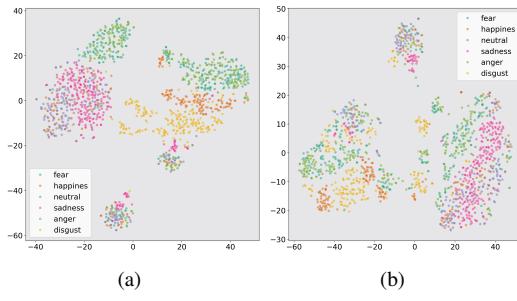


Figure 2: t-SNE plots for CREMA-D:(a) PARROT with Audio-MAMBA(base) and HuBERT (b) PARROT with Audio-MAMBA(base) and Wav2vec2

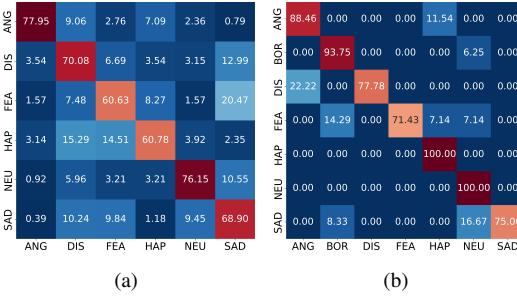


Figure 3: Confusion matrices for PARROT with Audio-MAMBA(base) and HuBERT : (a) CREMA-D (b) EMO-DB; x-axis and y-axis represents predicted and true, respectively

ther, fusion of different PTMs through **PARROT** achieves better performance than the individual PTMs across all the datasets. In contrast, fusion of PTMs using concatenation-based fusion overall shows comparable or less performance than individual PTMs except a few specific cases where the fusion of mamba

and attention-based PTMs brings strong complementary behavior. This behavior is observed across all the datasets. With **PARROT**, we observe the emergence of such complementary behavior amongst mamba-and attention-based PTMs in a much better way due to the capability of **PARROT** to bring out such behavior than baseline concatenation-based fusion technique. For example, fusion of Audio-MAMBA (base) with HuBERT through **PARROT** reported the topmost performance in CREMA-D and Emo-DB. Also, fusion of Audio-MAMBA (base) with MMS through **PARROT** reported the best performance in MESD. These results verifies *our hypothesis that heterogeneous fusion of mamba and attention-based PTMs will lead to more improved SER due to the effective emergence of complementary strengths with attention-based PTMs capturing complex global dependencies and mamba-based PTMs excelling in efficient long-range processing*. These results demonstrate that **PARROT**, by fusing Mamba- and attention-based PTMs, surpasses individual PTMs that previously achieved SOTA performance in SER [29, 13]. Additionally, our approach outperforms most homogeneous PTM fusions and baseline fusion techniques and further establishing its effectiveness in achieving SOTA in SER. We plot the t-SNE plot visualizations of representations from the last penultimate layer in Figure 2. We also plot the confusion matrices of **PARROT** with fusion of Audio-MAMBA(base) and HuBERT in Figure 3.

## 5. Conclusion

In this study, we explore the heterogeneous fusion of mamba and attention-based SSL PTMs for SER. To this end, we propose, **PARROT**, a novel framework that synergizes PTMs via parallel branch fusion of Optimal Transport and Hadamard Product. With **PARROT**, through the fusion of mamba and attention-based PTMs, we report SOTA performance in comparison to individual PTMs, homogeneous fusion of PTMs, and baseline fusion techniques. Our study will act as a reference for future research towards heterogeneous fusion of PTMs for SER.

## 6. References

- [1] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 16–28, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:8157773>
- [2] G. Souganciouglu, O. Verkholyak, H. Kaya, D. Fedotov, T. Cadee, A. A. Salah, and A. Karrov, "Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition," in *Interspeech*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:221535008>
- [3] 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.
- [4] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 239–246, 2017.
- [5] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghriby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.
- [6] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.
- [7] J. Yang, "Ensemble deep learning with hubert for speech emotion recognition," in *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, 2023, pp. 153–154.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Interspeech 2021*, 2021, pp. 3400–3404.
- [12] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.
- [13] S. Yadav and Z.-H. Tan, "Audio mamba: Selective state spaces for self-supervised audio representations," in *Interspeech 2024*, 2024, pp. 552–556.
- [14] M. H. Erol, A. Senocak, J. Feng, and J. S. Chung, "Audio mamba: Bidirectional state space model for audio representation learning," *IEEE Signal Processing Letters*, vol. 31, pp. 2975–2979, 2024.
- [15] S. Shams, S. S. Dindar, X. Jiang, and N. Mesgarani, "Ssamba: Self-supervised audio representation learning with mamba state space model," *arXiv preprint arXiv:2405.11831*, 2024.
- [16] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
- [17] 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.
- [18] 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.
- [19] 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/>
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [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] F. Burkhardt, A. Paeschke, M. Rolfs, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.
- [27] 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.
- [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.
- [29] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.