



Interactive Fusion of Multi-View Speech Embeddings via Pretrained Large-Scale Speech Models for Speech Emotional Attribute Prediction in Naturalistic Conditions

Yuyun Liu^{1,#}, Yujia Gu^{1,#}, Jiahao Luo¹, Wenming Zheng¹, Cheng Lu^{1,*}, Yuan Zong^{1,*}

¹Key Laboratory of Child Development and Learning Science of Ministry of Education, School of Biological Science and Medical Engineering, Southeast University, Nanjing 211189, China

{liyuyun, 213211323, 213222967, wenming-zheng, cheng.lu, xhzongyuan}@seu.edu.cn

Abstract

This paper addresses the Task 2 of Speech Emotion Recognition in Naturalistic Conditions Challenge at INTERSPEECH 2025, i.e., Emotional Attribute Prediction, and presents a simple and effective method named the Interactive Fusion of Multi-View Speech Embeddings (IF-MVSE). In this method, pretrained large-scale speech models are first utilized to extract multi-view speech embeddings, allowing for capturing complementary speech representations from multiple perspectives of speech signals. Subsequently, we design an interactive fusion strategy consisting of dual-feature interactive attention and multi-view self-balancing gated operations to integrate and enhance these speech embeddings from multiple views to predict the dimensional emotion attributes. Our IF-MVSE achieved the average CCC of 0.5955 on the official testing set, securing the *Third Place* in this track.

Index Terms: Speech emotion recognition, emotional attribute prediction, pretrained large-scale model, interactive fusion

1. Introduction

“Speech Emotion Recognition (SER) in Naturalistic Conditions Challenge” hosted by INTERSPEECH 2025 [1] is a global competition aimed at advancing research in emotion recognition under real-world conditions, which consists of two independent tracks: Task 1 - Categorical emotion recognition and Task 2 - Emotional attribute prediction. This challenge invites researchers to participate in one or both for developing the promising SER models. In the challenge, the Training, Validation, and Testing datasets are collected from the MSP-Podcast corpus [2] with eight emotional categories (i.e., anger, happiness, sadness, fear, surprise, contempt, disgust, and neutral) and three dimensional attributes (i.e., arousal, valence, and dominance). Note that this paper reports our solution for Task 2.

This challenge can be traced back to Odyssey 2024 [3], where most previous participating teams constructed multimodal SER models by combining transcribed text information with speech, achieving improvements in recognition performance. For instance, Chen et al. [4] formed an integrated system using seven multimodal models, where each model takes speech and text features as input and is trained independently with different acoustic features, loss functions, and class weights. Harm et al. [5] constructed a fine-tuning framework with speech-based Wav2Vec2-BERT [6] and the text-based LLaMA2-7B [7] for emotion classification and attribute prediction tasks. Bellver-Soler et al. [8] combined the audio encoder of Whisper Large V3 [9] with large language models

(e.g., Phi 1.5 [10] and Gemma 2b [11]) fused audio and text features for speech emotion recognition.

Most of the methods in the aforementioned challenges rely on transcribing text from speech and combining it with speech information to construct multimodal frameworks [12]. However, since the challenge dataset only contains speech data and lacks text information, and given that the data is designed for naturalistic conditions, dialogue information in the speech is easily interfered with by noise and background sounds. This leads to the transcription of noisy text, which in turn affects the SER performance of text and speech based multimodal models. From the speech signal itself, there is a wealth of information embedded, e.g., speaker identity, emotions. Therefore, accurately extracting these pieces of information is crucial for the speech-related tasks. Currently, driven by vast amounts of data, the Pretrained Large-Scale Speech Models, e.g., WavLM [13], Whisper [9], and HuBERT [14] have demonstrated powerful capabilities in general speech representation, achieving state-of-the-art (SOTA) performance in various speech-related tasks. Feng et al. [15] leveraged pretrained speech models such as WavLM and Whisper to achieve label classification for speech emotion recognition tasks. They accomplished this by freezing the main parameters of these models and utilizing the generated speech representations. Inoue et al. [16] optimized self-supervised models by adding three types of adapters (E-adapters, L-adapters, and P-adapter) to large pretrained models, thereby achieving speech emotion recognition. Therefore, this paper explores how to leverage the popular Pretrained Large-Scale Speech Models to capitalize on the high robustness and generalization of speech representations, further improving speech emotional attribute prediction in naturalistic conditions.

Based on the considerations mentioned above, we propose the Interactive Fusion of Multi-View Speech Embeddings (IF-MVSE) method to tackle Task 2 – Emotional Attribute Prediction of Speech Emotion Recognition in Naturalistic Conditions Challenge in INTERSPEECH 2025. To achieve the goal, we first extract robust multi-view speech emotion embeddings using popular Pretrained Large-Scale Speech Models, e.g., WavLM, Whisper, and HuBERT. Next, we perform interactive fusion of these multi-dimensional speech features, enabling precise representation of speech emotional attributes. Our proposed model achieved an average accuracy of 0.5955 on the challenge test set and obtained the third place in Task 2.

2. Method

As shown in Figure 1, our IF-MVSE consists of two core modules: the Multi-view Feature Extraction module and the Multi-View Feature Interaction Fusion module. In the first stage, we utilize Pretrained Large-Scale Speech Models to extract Multi-

Equal Contributions

* Corresponding Authors

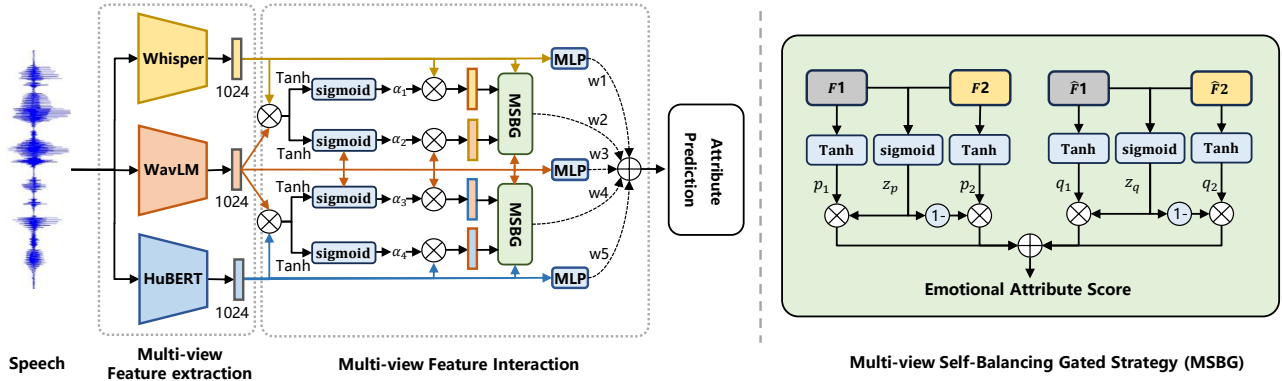


Figure 1: *Interactive Fusion of Multi-View Speech Embeddings (IF-MVSE) for Speech Emotional Attribute Prediction in Naturalistic Conditions*

View Speech Embeddings from the original speech signals, fully capturing the emotional features in the speech. Subsequently, in the Multi-View Feature Interaction Fusion module, we adopt a Multi-View Self-Balancing Gated Strategy with a Dual-Feature Interactive Attention mechanism to achieve the deep interactive fusion of multi-dimensional speech features.

2.1. Multi-View Feature Extraction

Traditional methods face numerous challenges in speech emotion recognition, especially in complex real-world scenarios. In contrast, leveraging language feature vectors extracted from Pretrained Large-Scale Speech Models has significantly enhanced the performance of emotion recognition. Ma et al. [17] conducted comprehensive experiments on well-known emotion recognition datasets such as JL-Corpus [18], RAVDESS [19], and MER2023 [20]. They employed various large-scale speech models for feature extraction, including wav2vec 2.0 [21], HuBERT [14], and WavLM [13]. By adjusting and optimizing the model parameters, they evaluated the performance using multiple metrics such as unweighted average accuracy (UA), weighted average accuracy (WA), and macro F1 score. The experimental results showed that Whisper Large V3 performed optimally on most datasets, and WavLM Large, HuBERT Large, and Data2vec Large [22] also demonstrated powerful capabilities.

In Multi-View Speech Embeddings extraction phase, we selected four representative models, namely Whisper Large V3¹, WavLM Large², HuBERT Large³, and Data2vec Large⁴, to extract features directly from the raw speech signals. Then, these features were fed into a Multi-Layer Perceptron (MLP) for predicting emotional attributes on the competition dataset.

These models perform excellently in speech signal processing, which benefits from their unique architectural characteristics. Whisper Large V3 adopts an encoder-decoder architecture, and only the encoder part is used for feature extraction. The speech signal is downsampled to 16 kHz, and Mel-spectrogram features are extracted. The features output by the encoder have a dimension of $T \times 1280$, which is compressed to 1280×1 through an average pooling layer. WavLM Large, HuBERT Large, and Data2Vec Large are all based on the Transformer

¹<https://huggingface.co/openai/whisper-large-v3>

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

³<https://huggingface.co/facebook/hubert-large-ls960-ft>

⁴<https://huggingface.co/facebook/data2vec-audio-large-960h>

architecture, and the speech signal is also downsampled to 16 kHz. The dimension of the final hidden states is all $T \times 1024$, and after average pooling, it is compressed to 1024×1 .

Based on subsequent experimental analysis, we ultimately selected the features extracted by WavLM, Whisper, and HuBERT for the next step of feature interaction.

2.2. Multi-View Feature Interaction Fusion

This study made targeted improvements to the original GBAN [23] multimodal emotion recognition framework. To achieve effective interaction and fusion between features from different perspectives, we designed the Multi-View Feature Interaction Fusion module. Through the Dual-feature Interactive Attention and Multi-View Self-Balancing Gated Strategy, multi-view emotion attributes of speech can be aggregated and the importance weights of each view feature are automatically adjusts through an adaptive balancing mechanism. This allows for the full utilization of the complementarity of multi-view features.

In terms of feature representation, we adopted multi-view features extracted by two representative large speech models: the first group consists of features extracted by the WavLM and Whisper models, while the second group includes features extracted by the WavLM and HuBERT models. These combinations aim to mine deep semantic and acoustic feature information from speech signals. It is important to note that the acoustic feature representation dimension extracted through Whisper is 1280×1 . During the multi-view feature interaction phase, a linear layer is applied to map it to the dimension 1024×1 for consistency.

To formally describe the multi-view feature interaction process, let F_1 and F_2 represent the acoustic feature vectors obtained after processing the original speech signals with different large speech models.

2.2.1. Dual-Feature Interactive Attention

The proposed Dual-Feature Interactive Attention mechanism adopts a dual-path symmetric architecture to achieve efficient processing and interactive fusion of acoustic features from two different speech models. This innovative mechanism can effectively capture the correlation and potential association between features, thereby significantly enhancing the discriminative ability of fused feature representations. The calculation process can be formally defined as follows.

We define F_1 and F_2 (where $F_1, F_2 \in \mathbb{R}^{d \times 1}$) as the acous-

tic feature vectors extracted from two pre-trained large-scale speech models (e.g., Whisper and WavLM). Specifically, \hat{F}_1 and \hat{F}_2 are derived as follows:

$$F_1 = G_{\text{speech}}(x_i), \quad F_2 = H_{\text{speech}}(x_i), \quad (1)$$

where $G_{\text{speech}}(\cdot)$ and $H_{\text{speech}}(\cdot)$ represent the respective speech models responsible for extracting high-dimensional embeddings from the raw speech signal x_i .

First, the interaction feature F_{inter} is computed via the Hadamard product (element-wise multiplication) of F_1 and F_2 :

$$F_{\text{inter}} = F_1 \odot F_2. \quad (2)$$

This operation captures the element-wise similarity and association strength between the two feature vectors, providing a measure of their shared information content.

Subsequently, the attention weight α is computed by sequentially applying the hyperbolic tangent (tanh) and sigmoid functions:

$$\alpha = \sigma(\tanh(F_{\text{inter}})). \quad (3)$$

Here, the tanh function maps the interaction features to the range $(-1, 1)$, effectively preventing gradient saturation while preserving non-linear relationships. The sigmoid function then normalizes these values to the range $(0, 1)$, yielding the attention weight $\alpha \in [0, 1]$.

Finally, the attention weights are applied to perform a weighted fusion of the original features, resulting in the refined feature representations:

$$\hat{F}_1 = \alpha_1 \odot F_1, \quad \hat{F}_2 = \alpha_2 \odot F_2, \quad (4)$$

where $\hat{F}_1, \hat{F}_2 \in \mathbb{R}^{d \times 1}$. These processed features are then passed to the subsequent Multi-View Self-Balancing Gated Strategy for further processing and fusion.

2.2.2. Multi-View Self-Balancing Gated Strategy

In this work, we propose a novel Multi-View Self-Balancing Gated (MSBG) Strategy to effectively integrate features from diverse sources, thereby enhancing the performance of emotion attribute prediction. This strategy automatically adjusts the contribution of each feature group based on their relevance and discriminative power, ensuring a robust and adaptive fusion process.

The MSBG Strategy operates on two distinct feature groups: (1) Primary Features (F_1, F_2): These are the direct outputs of pre-trained speech models (e.g., Whisper and WavLM) and encapsulate rich semantic and acoustic information inherent in the speech signal. (2) Interaction Features (\hat{F}_1, \hat{F}_2): Derived from the Dual-Feature Interactive Attention module, these features incorporate cross-view information through element-wise interactions, thereby capturing higher-order relationships between different feature representations.

The mathematical formulation of the MSBG is as follows:

$$\begin{aligned} p_1 &= \tanh(W_1^p F_1), \\ p_2 &= \tanh(W_2^p F_2), \\ z_p &= \sigma(W_z^p [F_1, F_2]), \\ q_1 &= \tanh(W_1^q \hat{F}_1), \\ q_2 &= \tanh(W_2^q \hat{F}_2), \\ z_q &= \sigma(W_z^q [\hat{F}_1, \hat{F}_2]), \end{aligned} \quad (5)$$

$$\begin{aligned} S &= z_p \odot p_1 + (1 - z_p) \odot p_2 \\ &\quad + z_q \odot q_1 + (1 - z_q) \odot q_2, \end{aligned} \quad (6)$$

where W_1^p, W_2^p, W_1^q , and W_2^q are learnable weight matrices that apply non-linear transformations (via the tanh function) to the primary and interaction features, respectively. W_z^p and W_z^q are weight matrices used to determine the contribution of each feature group. The σ function represents the sigmoid activation, which maps the gate weights to the range $[0, 1]$. The symbol \odot denotes element-wise multiplication. The final output S is a weighted combination of the gated representations from both feature groups, capturing the most discriminative information for emotion prediction.

Multi-View Feature Interaction Fusion produces five distinct score sets S_i ($i = 1, 2, \dots, 5$), each corresponding to different feature extraction and fusion configurations: S_1, S_3 , and S_5 represent the predicted scores obtained by directly processing features extracted by Whisper, WavLM, and HuBERT, respectively, through a Multi-Layer Perceptron (MLP). S_2 corresponds to the score derived from feature interaction fusion between features extracted by Whisper and WavLM. S_4 corresponds to the score derived from feature interaction fusion between features extracted by HuBERT and WavLM.

Table 1: Performance of Pretrained Large-Scale Models on Emotional Attributes in the Development Set: **Bold Entries** Are the Top Results Per Column

Model	Arousal	Dominance	Valence	Average
Data2vec	0.2466	0.4335	0.2437	0.3079
HuBERT	0.4160	0.4980	0.3893	0.4344
Whisper	0.5590	0.6074	0.5101	0.5588
WavLM	0.6562	0.5993	0.6426	0.6327

2.3. Loss Function

The final prediction result can be obtained through the following formula:

$$Y_{\text{pred}} = w_1 S_1 + w_2 S_2 + w_3 S_3 + w_4 S_4 + w_5 S_5, \quad (7)$$

w_i ($i = 1, 2, 3, 4, 5$) represents the weight corresponding to each distinct score generated by Multi-View Feature Interaction Fusion.

The Concordance Correlation Coefficient (CCC) [1] using the following formula:

$$CCC(Y_{\text{true}}, Y_{\text{pred}}) = \frac{2\rho\sigma_{Y_{\text{true}}}\sigma_{Y_{\text{pred}}}}{\sigma_{Y_{\text{true}}}^2 + \sigma_{Y_{\text{pred}}}^2 + (\mu_{Y_{\text{true}}} - \mu_{Y_{\text{pred}}})^2}.$$

Here, Y_{true} represents the true score of a certain emotional attribute, while Y_{pred} is the score obtained after the model inference. ρ is the Pearson correlation coefficient between Y_{true} and Y_{pred} , $\sigma_{Y_{\text{true}}}$ and $\sigma_{Y_{\text{pred}}}$ are the standard deviations of Y_{true} and Y_{pred} respectively, and $\mu_{Y_{\text{true}}}$ and $\mu_{Y_{\text{pred}}}$ are the means of Y_{true} and Y_{pred} respectively. Thus, the total loss function of our IF-MVSE can be expressed as follows:

$$\text{Loss} = 1 - CCC(Y_{\text{true}}, Y_{\text{pred}}). \quad (8)$$

Table 2: *Model Performance on Recognizing Emotional Attributes on Training and Development Sets When Integrating Multiple Model Features: Optimal Results Are Marked in **Bold**.*

Method	Train			Dev		
	Arousal	Dominance	Valence	Arousal	Dominance	Valence
WavLM	0.6841	0.5948	0.5569	0.6562	0.5993	0.6426
WavLM + Whisper	0.7404	0.6908	0.6791	0.6765	0.6054	0.7119
WavLM + Whisper + HuBERT	0.7560	0.6827	0.6555	0.6882	0.6224	0.7141

Table 3: *Performance of Recognizing Emotional Attributes on the Test Set: The Optimal Results Are Marked in **Bold**.*

Model	Arousal	Dominance	Valence	Average
Baseline [1]	0.6232	0.4775	0.6385	0.5797
Ours	0.6242	0.4914	0.6709	0.5955

3. Experiments

3.1. Database

“Speech Emotion Recognition under Natural Conditions” Challenge in INTERSPEECH 2025 provides a comprehensive benchmark for evaluating emotion recognition methods, including two main tasks: Task 1 - Categorical emotion recognition and Task 2 - Emotional attribute prediction. Herein, we focus on Task 2, i.e., predicting arousal, valence, and dominance.

The challenge dataset is derived from the MSP-Podcast corpus [2] and contains the Training set (Train) with 84260 samples, Development set (Dev) with 31961 samples, and Testing set (Test) with 3200 samples. In this study, we conduct a regression analysis based on the continuous emotion attribute annotations and evaluate our model’s performance using the competition’s proprietary test set.

3.2. Experimental Setup

The experiments are conducted using the PyTorch framework, with the AdamW [24] optimizer. All network training and evaluation are carried out on NVIDIA GeForce GTX 4090D GPUs. We performed comprehensive hyperparameter tuning, and the training parameters for each emotional attribute are set as follows: batch size of 32, learning rate of $1e-5$, and 80 training epochs. For the MLP layer in the network, we adopt a dual-branch structure. Each large-model-encoded speech feature traverses two identical MLP paths with ReLU activation and 0.5 dropout in hidden layers. The number of layers of the MLP is specifically set as (1024, 512, 1). Weight optimization on validation set: coarse search (0-1, step 0.1) then fine search (step 0.01) for optimal hyperparameters. Finally, by using the optimal model parameters, we test the model on the provided test set. The weight settings for the predicted emotional attribute scores are as follows: $\{w_1 : 0.02, w_2 : 0.15, w_3 : 0.01, w_4 : 0.8, w_5 : 0.02\}$.

3.3. Results and Analysis

The experimental results for features extracted by large model encoders and processed through an MLP for score prediction

are presented in Table 1. Due to the inferior performance of Data2vec, only Whisper Large V3, WavLM Large, and HuBERT Large were retained for subsequent feature fusion experiments.

Based on previous performance comparisons, WavLM features exhibited the strongest predictive power, followed by Whisper. Consequently, the initial feature fusion experiments combined WavLM and Whisper features, resulting in notable performance enhancements. Further integration of HuBERT features with WavLM and Whisper led to additional improvements in overall performance.

As shown in Table 2, in the development set (Dev), WavLM achieved scores of arousal (0.6562), dominance (0.5993), and valence (0.6426). The WavLM + Whisper combination demonstrated superior performance across all three metrics, attaining scores of 0.6765, 0.6054, and 0.7119, respectively. The most effective configuration, WavLM + Whisper + HuBERT, yielded the highest scores of 0.6882 for arousal, 0.6224 for dominance, and 0.7141 for valence. These results validate the efficacy of our feature fusion mechanism in improving emotion attribute prediction.

Our proposed model demonstrates significant improvements over the competition baseline when evaluated on the unpublished test set, as illustrated in Table 3. Specifically, our feature fusion-based approach achieves scores of 0.6242, 0.4914 and 0.6709 for arousal, dominance, and valence, respectively, with an average score of 0.5955. This underscores the effectiveness of multi-view language feature fusion in enhancing emotion prediction accuracy.

4. Conclusion

In this paper, we focus on Task 2 (i.e., emotional attribute prediction) of the “Speech Emotion Recognition in Naturalistic Conditions Challenge” in INTERSPEECH 2025, and propose an effective Interactive Fusion of Multi-View Speech Embeddings (IF-MVSE) method. Firstly, IF-MVSE extracts the Multi-View Speech Embeddings of emotional speech through the popular Pretrained Large-Scale Speech models, i.e., HuBERT, WavLM, and Whisper. Subsequently, it captures complementary speech representations from the speech embeddings of multiple views, through a Multi-View Self-Balancing Gated Strategy with Dual-Feature Interactive Attention, so as to improve the accuracy of speech emotion attributes prediction. Our proposed IF-MVSE method achieved an average Concordance Correlation Coefficient (CCC) with 0.5955 on the test set and secured the *third place* in Task 2. Future research will further explore adaptive fusion strategies and extending them to more diverse multi-views features.

5. Acknowledgement

This work was supported in part by the NSFC under the Grant 62476057, in part by the YESS Program by CAST under the Grant 2023QNRC001, in part by the ASFC under the Grant 2023Z071069003, in part by the Zhishan Young Scholar Program of Southeast University, in part by the YESS Program by JSAST under the Grant JSTJ-2023-XH033, in part by the China Postdoctoral Science Foundation under the Grant 2023M740600, and in part by the Jiangsu Province Excellent Postdoctoral Program.

6. References

- [1] A. Reddy Naini, L. Goncalves, A. N. Salman, P. Mote, I. R. Ülgen, T. Thebaud, L. Velazquez, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "The interspeech 2025 challenge on speech emotion recognition in naturalistic conditions," in *Interspeech 2025*, vol. Under submission, Rotterdam, The Netherlands, August 2025.
- [2] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.
- [3] L. Goncalves, A. N. Salman, A. R. Naini, L. M. Velazquez, T. Thebaud, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results," *Development*, vol. 10, no. 9,290, pp. 4–54, 2024.
- [4] M. Chen, H. Zhang, Y. Li, J. Luo, W. Wu, Z. Ma, P. Bell, C. Lai, J. D. Reiss, L. Wang, P. C. Woodland, X. Chen, H. Phan, and T. Hain, "1st place solution to odyssey emotion recognition challenge task1: Tackling class imbalance problem," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 260–265.
- [5] H. Härm and T. Alumäe, "Taltech systems for the odyssey 2024 emotion recognition challenge," in *Proc. odyssey 2024*, 2024, pp. 255–259.
- [6] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenhaler, P.-A. Duquenne, B. Ellis, H. Elshahar, J. Haasheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.
- [7] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
- [8] J. Bellver, I. Martín-Fernández, J. M. Bravo-Pacheco, S. Esteban, F. Fernández-Martínez, and L. F. D'Haro, "Multimodal audio-language model for speech emotion recognition," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 288–295.
- [9] 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.
- [10] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.
- [11] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.
- [12] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multimodal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription," in *Interspeech*, 2019, pp. 3302–3306.
- [13] 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.
- [14] 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.
- [15] T. Feng and S. Narayanan, "Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.
- [16] N. Inoue, S. Otake, T. Hirose, M. Ohi, and R. Kawakami, "Elp-adapters: Parameter efficient adapter tuning for various speech processing tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [17] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," *arXiv preprint arXiv:2406.07162*, 2024.
- [18] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," *Interspeech 2018*, 2018.
- [19] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS one*, vol. 13, no. 5, p. e0196391, 2018.
- [20] Z. Lian, H. Sun, L. Sun, K. Chen, M. Xu, K. Wang, K. Xu, Y. He, Y. Li, J. Zhao *et al.*, "Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9610–9614.
- [21] 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.
- [22] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.
- [23] P. Liu, K. Li, and H. Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," *arXiv preprint arXiv:2201.06309*, 2022.
- [24] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.