



# EmoSphere-SER: Enhancing Speech Emotion Recognition Through Spherical Representation with Auxiliary Classification

Deok-Hyeon Cho\*, Hyung-Seok Oh\*, Seung-Bin Kim\*, Seong-Whan Lee†

Department of Artificial Intelligence, Korea University, Seoul, Korea

dh\_cho@korea.ac.kr, hs\_oh@korea.ac.kr, sb-kim@korea.ac.kr, sw.lee@korea.ac.kr

## Abstract

Speech emotion recognition predicts a speaker’s emotional state from speech signals using discrete labels or continuous dimensions such as arousal, valence, and dominance (VAD). We propose EmoSphere-SER, a joint model that integrates spherical VAD region classification to guide VAD regression for improved emotion prediction. In our framework, VAD values are transformed into spherical coordinates that are divided into multiple spherical regions, and an auxiliary classification task predicts which spherical region each point belongs to, guiding the regression process. Additionally, we incorporate a dynamic weighting scheme and a style pooling layer with multi-head self-attention to capture spectral and temporal dynamics, further boosting performance. This combined training strategy reinforces structured learning and improves prediction consistency. Experimental results show that our approach exceeds baseline methods, confirming the validity of the proposed framework.

**Index Terms:** Speech emotion recognition, dimensional emotion, affective computing

## 1. Introduction

Speech emotion recognition (SER) predicts a speaker’s emotional state using acoustic features [1]. Understanding the emotional state of speech is important in applications such as human-computer interaction, virtual assistants, mental health monitoring, and conversational artificial intelligence of advances in deep learning [2, 3, 4, 5].

There are two main approaches to representing emotions in SER. The first is categorical emotion classification [6, 7]. Emotion labels correspond very closely to the categories that we use in our daily lives. Paul Ekman [8] derived six primary emotions: happiness, anger, disgust, sadness, anxiety, and surprise based on universally recognized facial expressions. This method is straightforward and easy to implement, as it reduces the complex nature of emotions to a set of clear categories. Traditional machine learning techniques, like support vector machines and hidden Markov models, have been used to develop these systems [9, 10]. They rely on carefully designed features extracted from speech, such as pitch, tone, and rhythm, to classify emotions effectively. However, this approach overlooks the nuanced variations of emotions. The second approach is dimensional emotion prediction [11, 12]. Rather than using fixed categories, this method represents emotions along continuous scales [13]. Common dimensions include valence, arousal, and dominance (VAD). This framework offers a more detailed representation of emotional states, capturing subtle variations that categorical

labels might miss. Most emotional attribute prediction methods employ direct regression models that estimate VAD as separate numerical values [14, 15]. Although straightforward, this method does not fully account for the structured nature of the emotional space. Since emotions exist in a continuous space rather than as isolated points, the lack of structured predictions in current models increases the likelihood of unnatural or conflicting outputs.

Traditional SER systems relied on a variety of extracted features to capture the nuances of speech. The researchers employed methods such as Mel-frequency cepstral coefficients [16], principal component analysis [17], wavelet features [18], and Fourier parameters [19] to describe different aspects of the speech signal, including its spectral, prosodic, and temporal characteristics. Recent advances in SER have boosted prediction accuracy through the use of self-supervised learning models (SSL), multi-modal fusion, and deep neural networks [6]. Models such as WavLM [20], HuBERT [21], and Wav2Vec 2.0 [22] extract rich representations from speech, and multimodal strategies [23] that incorporate transcript text further enhance performance. Although feature extraction methods have advanced the field, many models still struggle to capture the full range of emotional expressions. We hypothesize that incorporating a structured representation of emotion into SER models helps them learn more stable features and produce more consistent predictions.

In psychology, Reisenzein demonstrated that using the angle and length of the vector in polar coordinates is the only possible option for interpreting the relationships between emotions [24, 25]. Building on this idea, we introduce EmoSphere-SER—a novel approach that explicitly models emotional space using a spherical representation inspired by [26, 27] and applies spherical regions as an auxiliary loss. While prior work such as [28] has used an auxiliary classification task to predict discrete emotion categories, our approach focuses on predicting the region on the sphere, thereby emphasizing auxiliary learning for dimensional representations. Additionally, we employ a dynamic weighting scheme to adaptively adjust the contributions of the regression and classification tasks. Furthermore, we integrate a style pooling layer to capture key spectral and temporal details of speech. These components work jointly to enhance the coherence and stability of the learned representations, paving the way for more reliable emotion recognition. As a result, our method outperforms baseline approaches, confirming the validity of the proposed framework.

## 2. EmoSphere-SER

EmoSphere-SER is an emotional attribute prediction system that employs a spherical representation to explicitly model emo-

\*Equal contribution

†Corresponding author

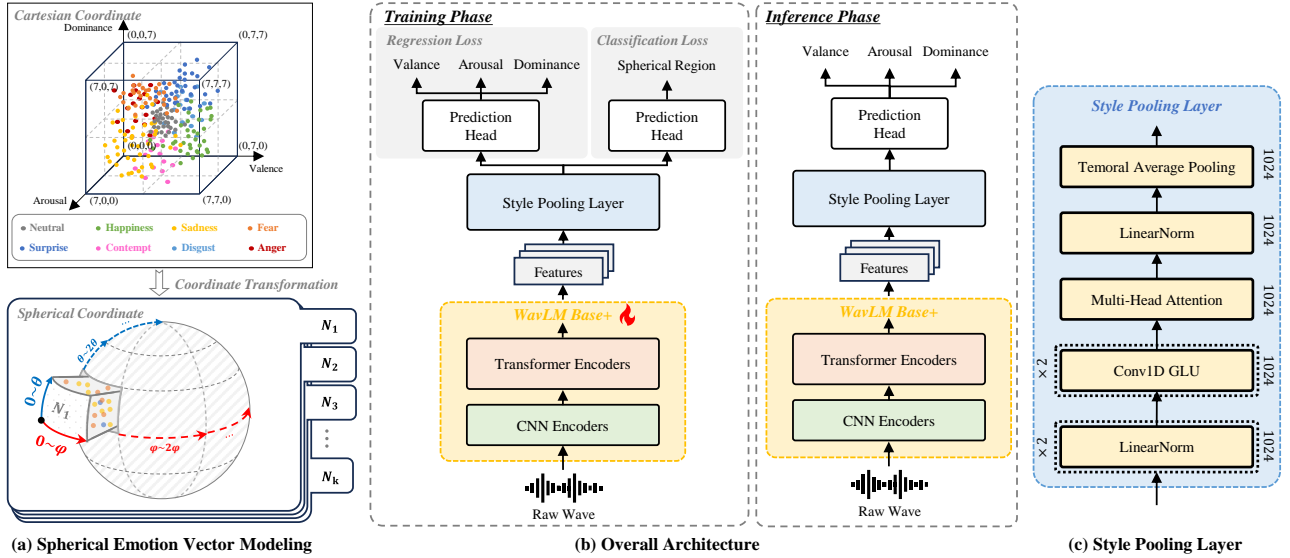


Figure 1: Overall framework of EmoSphere-SER

tional space, enabling auxiliary learning for dimensional representations. The implementation details are described in the following subsections.

## 2.1. Spherical emotion vector modeling

In this section, we present a spherical emotion vector modeling inspired by [26, 27] to enhance VAD prediction. The approach to building the space is structured around two key components: 1) normalization and spherical transformation and 2) spherical region partitioning and label assignment.

### 2.1.1. Normalization and spherical transformation

The original VAD annotations are provided on a  $[1, 7]$  scale. To facilitate coordinate transformation, we first normalize each of the three emotion dimensions to the range  $[-1, 1]$ . Once normalized, the three-dimensional VAD vector is converted from Cartesian coordinates to spherical coordinates. This transformation decomposes the vector into a magnitude of  $r$ , azimuth of  $\phi$ , and elevation of  $\theta$ .

### 2.1.2. Spherical region partitioning and label assignment

Inspired by the modeling of complex emotional characteristics using relative distance and angular vectors, we assume that the angle from the center determines emotional style. To implement this in the spherical coordinate system, we quantify azimuth and elevation angles to partition the space into distinct spherical regions. The azimuth angle is divided into  $N_\phi$  equal intervals, while the elevation angle is divided into  $N_\theta$  equal intervals, resulting in  $N = N_\phi \times N_\theta$  regions. Each region is then assigned a unique categorical label corresponding to a prototypical emotional state within the affective space.

## 2.2. Architecture

In this section, we introduce the overall architecture of emotional attribute prediction based on a pre-trained SSL model. The framework consists of three main components: a pre-trained SSL model, a style pooling layer, and prediction heads, which will be explained in detail in the following subsections.

### 2.2.1. Pre-trained SSL model

We adopt WavLM [20] as our feature encoder for extracting features from audio. WavLM is a pre-trained SSL model trained on large-scale data, and it is capable of extracting frame-level high-dimensional representations. To further enhance the model’s ability to capture emotional nuances in speech, we fine-tuned WavLM specifically for emotion recognition.

### 2.2.2. Style pooling layer

To aggregate frame-level features into an utterance-level representation, we use a style pooling layer inspired by [29]. This layer consists of LayerNorm, Conv1D with GLU, and multi-head attention, allowing the model to attend to different temporal segments simultaneously. Finally, we apply temporal average pooling to summarize the frame-level features.

### 2.2.3. Prediction head

From the utterance-level representations generated by style pooling layer, we employ two distinct prediction layers to forecast both the spherical region and the continuous VAD values. To predict the spherical region, a classification layer maps the learned representations to one of the  $N$  discrete categories. Simultaneously, a regression layer is implemented to output three scalar values, each corresponding to one of the VAD dimensions.

## 2.3. Training strategy

Our training strategy leverages an auxiliary spherical classification loss and the primary VAD regression loss. The spherical region classification loss is designed to support VAD prediction.

### 2.3.1. Spherical region classification loss

To facilitate the alignment of the utterance-level representations with the  $N$  discrete spherical regions, we employ an auxiliary classification loss, denoted as  $\mathcal{L}_{\text{sph}}$ . This loss is computed via a weighted cross-entropy (WCE) loss between the predicted and

Table 1: Performance comparison of our proposed method under different experimental settings. A  $\times$  symbol indicates that the corresponding component was excluded from the model.

Method	Auxiliary Loss	Style Pooling Layer	Data Preprocessing	Valance	Arousal	Dominance	Average
EmoSphere-SER	✓	✓	✓	0.6952	<b>0.7482</b>	<b>0.6220</b>	<b>0.6884</b>
Ablation Study	✓	✓	✗	<b>0.6953</b>	0.7447	0.6166	0.6855
	✓	✗	✗	0.6908	0.7456	0.6192	0.6852
	✗	✗	✗	0.6845	0.7349	0.6210	0.6801

ground-truth spherical region labels, as the following equation:

$$\mathcal{L}_{\text{sph}} = - \sum_{i=1}^N w_i y_i \log(p_i), \quad (1)$$

where  $i$  represents index of each class and  $p_i$  is the predicted probability that the sample belongs to class  $i$ . The variable  $y_i$  is a one-hot vector representing the true class. The  $w_i$  is the weight assigned to class  $i$ , determined based on its inverse frequency to mitigate class imbalance.

### 2.3.2. VAD regression loss

For the primary task of VAD prediction, we adopt the concordance correlation coefficient (CCC) loss, which directly optimizes the agreement between the continuous predicted values and the ground-truth annotations. Given the predicted VAD values  $\hat{y}$  and the corresponding ground-truth  $y$ , the  $\mathcal{L}_{\text{CCC}}$  is defined as:

$$\mathcal{L}_{\text{CCC}} = \frac{2\sigma_{y\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}, \quad (2)$$

where  $\mu_y$  and  $\mu_{\hat{y}}$  denote the means of  $y$  and  $\hat{y}$ , respectively,  $\sigma_y^2$  and  $\sigma_{\hat{y}}^2$  are the variances,  $\sigma_{y\hat{y}}$  represents the covariance between  $y$  and  $\hat{y}$ .

### 2.3.3. Overall loss

The complete training objective combines both the auxiliary spherical classification loss and the VAD regression loss:

$$\mathcal{L} = \mathcal{L}_{\text{CCC}} + \lambda_{\text{sph}} \mathcal{L}_{\text{sph}}, \quad (3)$$

where  $\lambda_{\text{sph}}$  denotes a weighting coefficient that controls the contribution of the auxiliary loss. During the initial training phase,  $\lambda_{\text{sph}}$  decays linearly, and after a certain point, the model is trained solely with  $\mathcal{L}_{\text{CCC}}$ . Specifically, the coefficient  $\lambda_{\text{sph}}$  is defined as follows:

$$\lambda_{\text{sph}} = \begin{cases} 1 - \frac{0.99}{5}e, & \text{if } 0 \leq e < 5, \\ 0, & \text{if } e \geq 5, \end{cases} \quad (4)$$

where  $e$  denotes the number of epoch.

While the SSL model and style pooling layer are shared between both learning tasks, the remaining components are fundamentally separate. Notably, the spherical region classification loss converges rapidly, allowing the model to shift its focus to VAD regression once a certain level of convergence is achieved. This enables more fine-grained predictions of emotional attributes, with empirical results showing that performance peaks at around 5 epochs.

## 3. Experiments and results

### 3.1. Experimental setup

We utilized the MSP-Podcast corpus dataset [30], which comprises approximately 237 hours of speech data annotated with both categorical emotion labels and dimensional VAD values. Each utterance has been perceptually annotated by at least five raters. The dataset includes eight categorical emotion classes: happiness, sadness, fear, surprise, contempt, disgust, and a neutral state. For dimensional emotion labels, raters evaluated arousal, dominance, and valence using a seven-point Likert scale. We adhered to the challenge organizers’ guidelines for data partitioning, utilizing the recommended training split for model development [31]. Additionally, to ensure a balanced representation of categories within the validation set, we further divided the original validation set into a validation set and an in-house test set. Specifically, the validation set was constructed by sampling 300 instances per category, maintaining uniform class distribution, while the remaining data was designated as the in-house test set. The “X” label corresponds to instances where no plurality voting winner could be determined among the emotion categories, reflecting cases of annotator disagreement or ambiguity. To minimize label noise and ensure more reliable supervision, all instances with the “X” label were excluded from the training, validation, and in-house test sets.

We utilize the AdamW optimizer [32] with a learning rate of  $1 \times 10^{-5}$  for training. The model is trained for 20 epochs with a batch size of 32. The model checkpoint corresponding to the best validation performance on the development set is selected and saved based on the lowest validation loss. The training process of the TTS module was conducted over approximately 24 hours on a single NVIDIA RTX A6000 GPU.

### 3.2. Implementation details

The style pooling layer is based on the structure of the style encoder proposed in [33] to embed the reference speech into a latent vector. It consists of three main components: The spectral processing module includes two fully connected layers with 1024 hidden units each. The temporal processing module is composed of two gated 1D convolutional neural networks with residual connections, where the convolutional layers have a filter size of 1024 and a kernel size of 5. Following this, the multi-head self-attention module has a hidden size of 1024 with two attention heads. On top of this module, a fully connected layer with a dimensionality of 1024 is applied, followed by temporal average pooling.

### 3.3. Performance Metrics

To evaluate the model’s performance across different emotion representations, we use metrics designed for each type. For spherical emotion vector representations, F1-score and accuracy are used to evaluate model performance, offering a comprehensive view of both per-class balance and overall predic-

Table 2: Comparison results on auxiliary loss modifications.

Methods	Valance	Arousal	Dominance	Average
w/o Dynamic Weighting	0.6951	0.7405	0.6267	0.6875
w/ Categorical Recognition	0.6912	0.7342	0.6281	0.6845
w/ Cross Entropy	<b>0.6968</b>	0.7352	<b>0.6308</b>	0.6876
EmoSphere-SER (Proposed)	0.6952	<b>0.7482</b>	0.6220	<b>0.6884</b>

tive correctness. For dimensional emotion representations, the mean concordance correlation coefficient (CCC) assesses how well the model predicts continuous emotional states. This combination of metrics provides a comprehensive evaluation of the model’s ability to capture both discrete spherical regions and the continuous nature of emotions.

### 3.4. Model performance

As shown in Table 1, our model improves performance, explained as follows: 1) “**w/o Data Preprocessing**” indicates that the model is trained on the entire dataset without excluding category data with the “X” label. By removing instances with multiple speakers or ambiguous emotions, the model achieved better performance in predicting emotional attributes. 2) “**w/o Style Pooling Layer**” denotes to attentive statistics pooling [34], which employs an attention mechanism to assign different weights to frames. By replacing attentive statistics pooling with the style pooling layer, our model benefits from a more expressive style representation. Unlike traditional pooling methods, the style pooling layer effectively captures both spectral and temporal dynamics while leveraging multi-head self-attention to enhance feature extraction. This allows the model to generate richer and more speaker-adaptive embeddings, ultimately leading to improved performance in predicting emotional attributes. 3) “**w/o Auxiliary Loss**” indicates the absence of an additional auxiliary loss designed to enhance VAD prediction. Specifically, we introduce an auxiliary loss that converts the original VAD values into a spherical emotion vector representation, predicting its corresponding region in spherical coordinates using WCE as a categorical loss. This auxiliary objective aids the model in the early stages of training by providing additional supervision, ultimately leading to improved VAD prediction performance. These results confirm that the proposed method effectively enhances the prediction of emotional attributes and demonstrates the ability of the SER model to adapt to various emotions and speaker styles in the prediction of VAD.

### 3.5. Impact of auxiliary loss modification

To evaluate and compare the effectiveness of our proposed auxiliary loss, we conducted three additional experiments: (1) “**w/o Dynamic Weighting**” refers to training without dynamically adjusting the auxiliary loss weight, instead maintaining a constant weight throughout the entire training process. (2) “**w/ Categorical Recognition**” replaces the spherical region classification with a categorical emotion prediction task, where the model uses categorical emotion labels as the auxiliary loss instead. (3) “**w/ Cross Entropy**” replaces WCE with standard cross-entropy in the spherical region classification task.

As shown in Table 2, the results indicate that our proposed auxiliary loss, which models VAD as a spherical region prediction task, provides more effective supervision than direct categorical emotion prediction. This suggests that learning a latent structured representation in the spherical region better aligns with VAD estimation. Additionally, the dynamic weight-

Table 3: Performance comparison of different angular divisions in the auxiliary loss across varying azimuth angles with identical elevation angles to ensure uniform spherical partitioning. The total number of regions is represented as  $N = N_\phi N_\theta$ , where  $N_\phi$  and  $N_\theta$  denote the divisions along the azimuth and elevation axes, respectively.

N (Angle)	Emotional Attribute				Spherical Region	
	Valance	Arousal	Dominance	Average	Macro F1	Accuracy
32 (45°)	0.6833	0.7385	0.6204	0.6808	4.77	16.95
18 (60°)	<b>0.6962</b>	0.7376	<b>0.6250</b>	0.6863	17.96	31.95
8 (90°)	0.6952	<b>0.7482</b>	0.6220	<b>0.6884</b>	36.27	59.51

ing scheme contributes to refining VAD predictions. By emphasizing auxiliary supervision in the early training stages and gradually shifting the focus toward VAD regression, the model achieves more precise and detailed predictions in later stages. Moreover, the use of WCE outperforms standard cross-entropy by better handling class imbalances, ensuring that each spherical region is adequately represented during training. In summary, the proposed module enhances VAD prediction by leveraging structured auxiliary supervision and dynamic weighting, leading to more precise and stable predictions. Combining both models provides a more comprehensive understanding of emotions, capturing both fundamental patterns and subtle nuances of emotional experiences.

### 3.6. Impact of angular division

As shown in Table 3, we investigated the impact of different angular divisions for the spherical region. We conducted experiments by dividing the total number of regions, represented as  $N = N_\phi N_\theta$ , based on angular thresholds of 90°, 60°, and 45°. The experimental results indicate that the model achieved the best performance in both the prediction of emotional attributes and the classification of the spherical region when the angular division was set to 90° (i.e.,  $N_\phi = 2$ ,  $N_\theta = 4$ , and consequently  $N = 8$ ). For angular divisions smaller than 90°, the increased difficulty in predicting the spherical region made auxiliary learning more challenging, leading to a decline in performance. These findings suggest that 90° provides an optimal balance between categorical supervision and effective auxiliary learning, leading to enhanced VAD prediction.

## 4. Conclusion

In this work, we present EmoSphere-SER, a speech emotion recognition model that improves the prediction of continuous emotional attributes by incorporating a structured spherical representation of emotion. Our model divides the emotional space into distinct regions and employs an auxiliary classification task to identify the region corresponding to each emotional state, thereby guiding the regression process. Furthermore, the dynamic weighting scheme maintains balance in the overall learning process, and a style pooling layer with multi-head self-attention effectively captures both spectral and temporal dynamics. Experimental results demonstrate that our approach outperforms baseline methods, confirming the benefits of our structured framework for enhanced emotional attribute prediction. To the best of our knowledge, this work represents the first attempt to predict emotional attributes using a spherical coordinate-based approach. Future research can further explore refining region partitioning strategies and extending this framework to other affective computing tasks.

## 5. Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (Artificial Intelligence Graduate School Program (Korea University) (No. RS-2019-II190079), Artificial Intelligence Innovation Hub (No. RS-2021-II212068), AI Technology for Interactive Communication of Language Impaired Individuals (No. RS-2024-00336673), and Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents (IITP-2025-RS-2025-02304828)).

## 6. References

- [1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.
- [2] D. Thiripurasundari, K. Bhargale, V. Aashritha, S. Mondreti, and M. Kothandaraman, "Speech emotion recognition for human-computer interaction," *Int. J. Speech Technol.*, vol. 27, no. 3, pp. 817–830, 2024.
- [3] J. Kim, J. Schultz, T. Rohe, C. Wallraven, S.-W. Lee, and H. H. Bühlhoff, "Abstract representations of associated emotions in the human brain," *Journal of Neuroscience*, vol. 35, no. 14, pp. 5655–5663, 2015.
- [4] K. Lee, S.-A. Kim, J. Choi, and S.-W. Lee, "Deep reinforcement learning in continuous action spaces: a case study in the game of simulated curling," in *Int. Conf. Mach. Learn.*, 2018.
- [5] D.-H. Lee, J.-H. Jeong, K. Kim, B.-W. Yu, and S.-W. Lee, "Continuous eeg decoding of pilots' mental states using multiple feature block-based convolutional neural network," *IEEE access*, vol. 8, pp. 121 929–121 941, 2020.
- [6] W. Chen, X. Xing, P. Chen, and X. Xu, "Vesper: A compact and effective pretrained model for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1711–1724, 2024.
- [7] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "Mser: Multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Syst. Appl.*, vol. 245, p. 122946, 2024.
- [8] P. Ekman and W. V. Friesen, "A new pan-cultural facial expression of emotion," *Motiv. Emot.*, vol. 10, pp. 159–168, 1986.
- [9] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *Int. Conf. Mach. Learn. Cybern.*, vol. 8, 2005, pp. 4898–4901.
- [10] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *2011 Int. Conf. Electron. Mech. Eng. Inf. Technol.*, vol. 2, 2011, pp. 621–625.
- [11] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multi-task learning," *APSIPA Trans. Signal Inf. Process.*, vol. 9, p. e17, 2020.
- [12] P. Mote, B. Sisman, and C. Busso, "Unsupervised domain adaptation for speech emotion recognition using k-nearest neighbors voice conversion," in *Interspeech*, 2024, pp. 1045–1049.
- [13] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, 1977.
- [14] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10 745–10 759, 2023.
- [15] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multi-task learning," *APSIPA Trans. Signal Inf. Process.*, vol. 9, p. e17, 2020.
- [16] A. Milton, S. S. Roy, and S. T. Selvi, "Svm scheme for speech emotion recognition using mfcc feature," *Int. J. Comput. Appl.*, vol. 69, no. 9, 2013.
- [17] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process.*, vol. 22, no. 6, pp. 1154–1160, 2012.
- [18] K. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using mfcc and wavelet features," in *IEEE Int. Adv. Comput. Conf. (IACC)*, 2013, pp. 842–847.
- [19] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 69–75, 2015.
- [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 J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [21] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
- [22] A. Baeovski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12 449–12 460.
- [23] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2021, pp. 4275–4279.
- [24] R. Reisenzein, "Pleasure-arousal theory and the intensity of emotions," *J. Pers. Soc. Psychol.*, vol. 67, no. 3, p. 525, 1994.
- [25] R. Jenke and A. Peer, "A cognitive architecture for modeling emotion dynamics: Intensity estimation from physiological signals," *Cogn. Syst. Res.*, vol. 49, pp. 128–141, 2018.
- [26] D.-H. Cho, H.-S. Oh, S.-B. Kim, S.-H. Lee, and S.-W. Lee, "Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech," in *Interspeech*, 2024, pp. 1810–1814.
- [27] D.-H. Cho, H.-S. Oh, S.-B. Kim, and S.-W. Lee, "Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector," *IEEE Trans. Affect. Comput.*, pp. 1–16, 2025.
- [28] J. L. Bautista and H. S. Shin, "Speech emotion recognition model based on joint modeling of discrete and dimensional emotion representation," *Appl. Sci.*, vol. 15, no. 2, 2025.
- [29] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 7748–7759.
- [30] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, pp. 471–483, 2017.
- [31] A. R. 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. To appear, Rotterdam, The Netherlands, August 2025.
- [32] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Int. Conf. Learn. Represent.*, 2019.
- [33] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *Int. Conf. Mach. Learn.*, 2021, pp. 7748–7759.
- [34] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech*, 2018, pp. 2252–2256.