



# From Pretraining to Performance: Benchmarking Self-Supervised Speech Models for Interspeech-25 SER Challenge

Drishya Uniyal, Vinayak Abrol

<sup>1</sup>CSE Department & Infosys Center for AI, IIT-Delhi, India

{drishyau, abrol}@iiitd.ac.in

## Abstract

Speech Emotion Recognition (SER) in naturalistic conditions remains a challenging task due to the variability of emotional expression and class imbalances in the real world. As part of the Interspeech-25 SER challenge, we benchmark state-of-the-art large-scale self-supervised speech models on the MSP-Podcast corpus. To extract rich and expressive representations, we systematically investigate fine-tuning strategies, loss functions tailored to mitigate class imbalance, and pre-trained encoder layer freezing techniques to optimize performance. Our findings highlight the impact of these design choices on model robustness and generalization, offering practical guidance for developing SER systems that excel in real-world scenarios.

**Index Terms:** speech recognition, human-computer interaction, computational paralinguistics

## 1. Introduction

Speech Emotion Recognition (SER) has emerged as a critical technology in affective computing and human-computer interaction (HCI), enabling machines to detect and interpret human emotions from voice signals. By analyzing vocal attributes such as tone, pitch, and rhythm, SER systems aim to infer emotional states like happiness, sadness, anger, and neutrality, thereby bridging the gap between words and emotional context. These capabilities have transformative potential across domains, from healthcare and education to customer service and entertainment [1]. However, real-world deployment of SER systems remains challenging due to speaker variability, cultural differences, and the inherent complexity of emotional expressions [2, 3]. The need for robust and generalizable SER models, particularly in naturalistic conditions, underscores the importance of advancing methodologies to tackle these challenges.

Deep learning has enabled automatic feature extraction and hierarchical representation learning, significantly improving SER performance [4, 5]. CNNs model spatial features, while RNNs, particularly LSTMs, capture temporal dynamics [6]. Recent approaches explore multimodal architectures [7], integrating auditory and visual cues. Additionally, self-supervised learning (SSL) leverages unlabeled data to enhance speech representations and boost emotion recognition accuracy.

The Speech Emotion Recognition in Naturalistic Conditions Challenge [8](Task 1) at Interspeech-25 focuses on addressing these issues by leveraging the MSP-Podcast corpus—a dataset comprising spontaneous speech recordings from diverse podcasts. Unlike traditional emotional datasets such as the Berlin Emotional Database (EmoDB) [9] and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [10] with controlled or acted emotions, the MSP-Podcast corpus offers a more realistic representation of emotional speech, includ-

ing variations in speaker styles, topics, and background noise. These features make it an ideal benchmark for developing SER systems that can adapt to real-world complexities. The challenge emphasizes creating robust algorithms capable of generalizing across diverse speaking styles, linguistic contexts, and environmental noise, further pushing the boundaries of SER research and practical applications.

In this paper <sup>1</sup>, we present our approach to the Interspeech 2025 SER Challenge, where we benchmark state-of-the-art self-supervised learning (SSL) speech models to extract rich representations for emotion recognition. Our methodology includes exploring fine-tuning strategies, addressing class imbalances with focal loss [11], and investigating encoder layer freezing techniques to optimize performance. Using WavLM-Large as a feature extractor, we demonstrate how selective fine-tuning and an attentive statistics pooling layer enhance representation learning. By integrating a deep feedforward classification model and employing focal loss to handle underrepresented emotional classes, our approach significantly improves robustness and generalization. This study not only advances SER systems for the challenge but also provides actionable insights for designing real-world emotion recognition systems.

## 2. Related Works

Speech Emotion Recognition (SER) has evolved significantly over the past decades, driven by advancements in feature extraction, model architectures, and dataset development. Early SER methods relied on hand-crafted acoustic features such as MFCCs, pitch, energy, formants, entropy, and zero-crossing rate [12]. These features were typically classified using traditional machine learning algorithms like Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) [13]. While foundational, these methods struggled with generalization due to speaker variability, environmental noise, and the inherent subtlety of emotional expressions. Moreover, they relied on controlled, acted datasets, which failed to capture the complexity of real-world emotions.

The rise of deep learning marked a paradigm shift in SER research. Neural networks evolved from simple architectures to more sophisticated models, including attention-based Transformers [14] [15] and multimodal frameworks. These advancements enabled the integration of acoustic, linguistic, and visual features for improved performance. Fusion techniques, such as early, late, and hybrid fusion, further enhanced the contextual understanding of emotions by effectively combining multiple modalities. For instance, [16] introduced a hybrid fusion approach with attention mechanisms to dynamically bal-

<sup>1</sup>This work is supported by Nebius Research Grant & in part by the Infosys Foundation via the Infosys Centre for AI, IIITD.

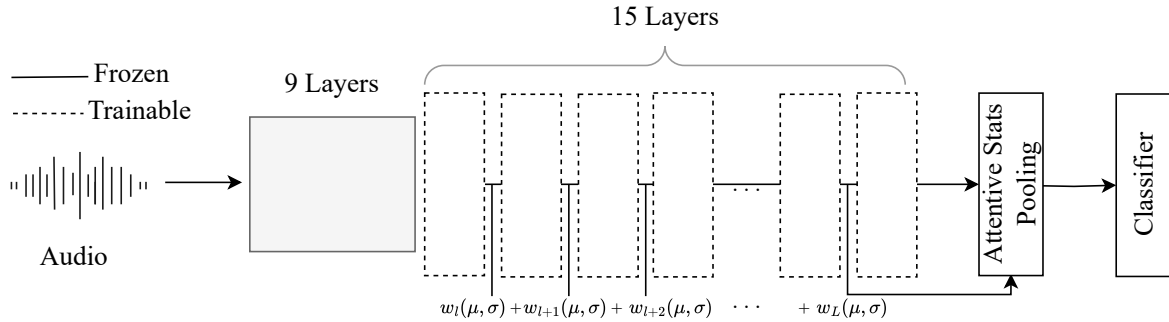


Figure 1: Model architecture of the proposed system for Interspeech-25 SER challenge.

ance modality importance, while [17] employed self-attention for acoustic and textual embeddings, achieving significant improvements in emotion classification accuracy.

To address the challenges of real-world emotion recognition, researchers have focused on robust datasets and data augmentation strategies. [18] introduced the MSP-Podcast dataset, which offers spontaneous and naturalistic emotional expressions, providing a benchmark for evaluating SER models under real-world conditions. [19] and [20] explored data augmentation techniques, such as room impulse response (RIR) simulations and noise injection, to improve model robustness against environmental variability. Additionally, self-supervised learning (SSL) has emerged as a game-changer in recent years. Sheffield-MINI (2024) achieved state-of-the-art performance by leveraging SSL-based features with self-attention [21]. Similarly, TalTech (2024) [22] demonstrated the effectiveness of combining Fbank representations with logistic regression-based fusion, while CONILIUM (2024) [23] employed raw audio inputs with weighted binary cross-entropy loss to address class imbalances and improve classification accuracy. Unlike existing approaches, our study focuses solely on the speech modality. This unimodal approach leverages state-of-the-art self-supervised learning (SSL) models to extract rich and expressive speech representations without relying on auxiliary data.

### 3. Proposed System

Our Speech Emotion Recognition (SER) framework leverages SSL models for acoustic feature extraction and a deep neural classifier for effective emotion classification directly from speech inputs. The proposed SER system is based on the WavLM-large model (also used in the challenge baseline). We incorporated the proposed multi-layer feature fusion method to fine-tune the model for emotion recognition on the train split of the challenge dataset. In addition to extracting rich and context-aware acoustic representations, we adopt a structured layer-freezing strategy to balance general speech feature retention and task-specific adaptation. Empirical experiments reveal that freezing the first nine transformer layers provides a good trade-off between preserving pre-trained knowledge while allowing upper layers to fine-tune for emotion-specific representations. The extracted feature embeddings from WavLM are passed through a feedforward classifier that consists of a projection layer and multiple fully connected (FC) layers using layer normalization, ReLU activation, and dropout regularization to ensure stable training and mitigate overfitting.

#### 3.1. Multi-layer Attentive Statistics Pooling (MASP) based Feature Fusion

As a standard practice in transfer learning, segment-level embeddings for downstream tasks are typically extracted from a pretrained deep acoustic neural model’s encoder (feature extractor) by applying average pooling over the output of the hidden layer. Recent works have shown that incorporating second-order statistics, such as standard deviation, along with the mean embedding can significantly improve the performance of the system [24]. Furthermore, while embeddings are typically extracted from the last or penultimate encoder layer, earlier layers can also capture useful acoustic cues in a hierarchical manner. We combine these two strategies to effectively model temporal features of speaker traits, such as emotion.

In this work, we focus on transformer models, which form the basis of our systems. To extract multi-layer acoustic features, we assign a learnable weight  $w_l$  to each transformer hidden layer  $h_l$ , and the final multi-layer acoustic feature  $F$  is computed as:

$$F = \frac{\sum_{l=1}^L e^{w_l} f_l}{\sum_{l=1}^L e^{w_l}}, \quad \text{s.t.} \quad f_l = \mu_l \oplus \sigma_l \quad (1)$$

$$\mu_l = \frac{\sum_{i=1}^T h_l^i}{T}, \quad \sigma_l = \sqrt{\frac{\sum_{i=1}^T (h_l^i)^2}{T} - \mu_l^2},$$

where  $\mu_l$  and  $\sigma_l$  represent the temporal mean and standard deviation over  $T$  timesteps of each layer’s output, and  $\oplus$  denotes the concatenation operator. The weighted sum of the layer-wise features is computed using the softmax function to ensure that the weights are non-negative and sum to 1. This can be interpreted as a form of attention over layers. However, traditional attention mechanisms typically employ a separate neural network for computing the weights, which is an interesting avenue for future work. Another promising direction is the application of attentive statistics pooling to temporal frame-level features within each hidden layer output.

## 4. Experimental Setup

#### 4.1. Dataset

The MSP-Podcast dataset is split into five subsets: a training set, a development set, and three test sets (test-1, test-2, and test-3), with test-3 serving as the primary benchmark for model evaluation. As reported in Table 1 The dataset includes speech samples annotated with 10 emotional categories, of which 8 are

used in the challenge: Neutral, Happy, Angry, Disgust, Sad, Surprise, Contempt, and Fear. The categories "Other" and "No Agreement (X)" are excluded to focus on well-defined emotions with high inter-annotator agreement. Each segment contains metadata such as emotion label, transcription, speaker ID, and gender, but in the test sets, only speech samples are provided to avoid bias.

Table 1: *Dataset Statistics by Emotion Class for MSP-Podcast dataset*

Emotion	Dev Set	Train Set
Neutral	7,423	29,243
Happy	6,344	16,717
Angry	5,836	6,731
Sad	2,341	6,306
Contempt	1,459	2,495
Surprise	987	2,948
Disgust	542	1,432
Fear	326	1,120

## 4.2. Model Training

Our overall pipeline is implemented in PyTorch [25] and is architecture-agnostic for constituent building blocks. Our model components are trained on a single NVIDIA A100 GPU (40 GB). Model optimization is achieved for 50 epochs with the AdamW optimizer, batchsize of 32 and a learning rate of  $10^{-4}$  with a cosine annealing scheduler. Dropout regularization of 0.3 and weight decay are applied to prevent classifier overfitting. These hyperparameters are selected after empirical evaluation over the development set. In addition to standard cross-entropy (CE) loss, we experimented with Focal loss (FL) [11] to improve imbalance performance on underrepresented emotions. For a fair comparison, we fixed the system and training setup for each individual model configuration.

## 5. Experimental Results

To systematically evaluate the impact of different design choices in our Speech Emotion Recognition (SER) model, we conducted an extensive series of experiments. Our investigations focus on key architectural and training decisions, including the choice of self-supervised learning (SSL) models, classifier variations, layer freezing strategies, loss functions, and data balancing techniques. Each study aims to isolate the effect of specific modifications, allowing us to derive actionable insights for optimizing SER performance on the MSP-Podcast dataset.

### 5.1. Comparison with existing methods

To assess the efficacy of our best model, we compare its performance with the top performers of the Odyssey 2024 SER Challenge [26] on the unseen testset of the same dataset. This comparison provides insights into how our approach measures against state-of-the-art methodologies previously validated on a similar challenge. As reported in Table 2 our best model outperforms both the baselines from Odyssey-24 and Interspeech-25. The baseline system in Odyssey-24 is based on fully finetuned WavLM with MLP classifier head. Further, our model surpasses the previous best 3 models, namely Sheffield-MINI, TalTech (ensemble with multimodal inputs) and CONILIUM. In contrast to full encoder fine-tuning, we leverage insights from Sheffield-MINI on handling class imbalance using focal loss

Table 2: *SER performance comparison using F1 Scores on the unseen challenge testset.*

Model	Input	F1-Ma	F1-Mi
Sheffield-MINI [21]	Audio/Text	0.3569	0.3732
TalTech [22]	Audio/Text	0.3543	0.3703
CONILIUM [23]	Audio	0.3350	0.3473
Baseline Interspeech-25	Audio	0.3293	0.3556
Baseline Odyssey-24 [26]	Audio	0.3113	0.3272
<b>WavLM-large + MASP (Ours)</b>	Audio	<b>0.3661</b>	<b>0.3794</b>

combined with selective layer fine-tuning/encoder-layer freezing and attentive layerwise feature pooling.

### 5.2. Comparative analysis of SSL Models for SER

Since the primary goal of this study is to improve SER performance from audio modality only, we experimented with SOTA pretrained models in order to choose suitable acoustic feature encoders. In particular, we compared HuBERT [27], Wav2Vec2 [28], WavLM [29], Whisper [30] (a supervised model) and ensemble architectures. These models excel at extracting rich audio representations and have shown significant promise across various speech-related tasks. Their ability to capture hierarchical acoustic features makes them well-suited for SER from the audio modality.

Table 3 presents the results of these experiments where we have only shown among many configurations that performed well. The performance is compared in terms of train/dev accuracy & precision/recall/F1 score on the development set. For the classifier, we have primarily experimented with the fully connected layers (FC) and XGBoost or graph-based AASIST [31] model trained with cross-entropy (CE) or generalized-end-to-end (GE2E) loss [32]. It can be observed that out-of-the-box pretrained SSL models (Wav2Vec, WavLM & HuBERT) demonstrate promising results. WavLM, which is jointly trained using masked prediction and denoising objectives, clearly outperform other models. Notice the suboptimal performance of Whisper models that are primarily trained for ASR tasks and are unable to capture the emotional traits well. In addition, an ensemble of the top two performing models (WavLM-large and Wav2Vec2-large) didn't yield any favourable results. Next, we repeated the experiments with full model finetuning, where we observed that, for most models, on average, there was minimal improvement or degraded performance. In an attempt to balance model size and performance, we combined the WavLM-base with alternative XGBoost/AASIST classifiers; however, the performance degraded significantly. As expected, the WavLM-large finetuned turned out to be the best model with the highest train accuracy and F1 score on the development set.

### 5.3. Ablation study with the proposed model

Results in the previous section established the suitability of WavLM model as a potentially good encoder for SER task. The proposed MASP further enhances the performance of WavLM by combining multi-layer acoustic cues. We believe that finetuning WavLM using specific strategies, such as freezing the bottom layers of the encoder presented, may limit the model's adaptability to the required emotion features. Existing literature has demonstrated that freezing the lower layers, which

Table 3: Comparison of SSL models, including pre-trained, fully fine-tuned, ensembles, and SSL with classifiers.

Type	Model	Classifier	Loss	Train Acc	Dev Acc	Precision	Recall	F1
Pre-Trained	WavLM-large	FC	CE	64.61	51.02	0.5021	0.512	0.512
	Wav2Vec2-large	FC	CE	64.36	50.08	0.5057	0.507	0.5015
	HuBERT-large	FC	CE	61.96	52	0.4685	0.4576	0.4985
	Whisper-small	FC	CE	56	44	0.43	0.44	0.435
	Whisper-large	FC	CE	56.24	42.34	0.41	0.42	0.415
	WavLM+Wav2Vec2	FC	CE	56.35	42.1	0.3988	0.4106	0.3912
Full Finetune	Wav2Vec2-large	FC	CE	58.25	45.03	0.4343	0.4544	0.4371
	HuBERT-large	FC	CE	64.36	<b>55.26</b>	0.5135	0.5141	0.5063
	Whisper-small	FC	GE2E	55.68	43.28	0.42	0.43	0.425
	WavLM	XGBoost	CE	50.32	39.02	0.38	0.39	0.385
	WavLM	AASIST	CE	54.36	42.37	0.3456	0.4215	0.3941
	WavLM-large (Baseline)	FC	CE	<b>66.36</b>	46.18	0.5214	0.5214	<b>0.518</b>

Table 4: Ablation study on WavLM fine-tuning, comparing full fine-tuning and partial fine-tuning. Here, the number in brackets denotes the number of trainable layers.

Model	Type	Loss	Train Acc	Dev Acc	Precision	Recall	F1
WavLM-large (Baseline)	Full Finetune	CE	66.36	46.18	0.5214	0.5214	0.518
WavLM-large + MASP (4)	Partial FineTune	CE	70.26	59.36	0.5546	0.5648	0.5432
WavLM-large + MASP (4)		FL	72.36	61.98	0.5412	0.5542	0.5431
WavLM-large + MASP (12)		CE	71.56	58.64	0.516	0.5864	0.5421
WavLM-large + MASP (12)		FL	72.26	63.7	0.6256	0.637	0.6283
WavLM-large + MASP (15)		FL	<b>73.49</b>	<b>66.25</b>	<b>0.6478</b>	<b>0.6625</b>	<b>0.6524</b>

are responsible for capturing universal acoustic features, helps in minimizing the risk of overfitting. At the same time, fine-tuning the higher layers, which require adjustments specific to the task, resulted in further performance improvements. To determine which layers to freeze, we employed the popular Centered Kernel Alignment (CKA) [33] approach to compare layer representations. CKA calculates similarity based on the normalized Frobenius norm of the dot product between layer output matrices. We performed CKA analysis on a subset of the development set using the fully finetuned WavLM model. As shown in Fig 2, we observed 12 to 16 penultimate layers exhibit lower similarity scores for models with and without fine-tuning. This highlights the regions where the model adapted the most, suggesting these layers were crucial for capturing the distinctive acoustic cues and thus require selective fine-tuning. We demonstrate the finding from our CKA analysis empirically using an ablation study, the results of which are reported in Table 4. It can be observed that the optimal performance is achieved using 15 trainable layers with MASP and Focal Loss, yielding the highest train accuracy (73.49%), development accuracy (66.25%), and F1 score (0.6524). Note that compared to CE loss, focal loss results in 5-8% relative improvement in SER accuracy on the development set. This demonstrates the effectiveness of deeper partial fine-tuning combined with advanced pooling mechanisms and loss function.

## 6. Conclusion

This work presents our approach to the Interspeech 2025 SER Challenge, leveraging self-supervised learning (SSL) models for robust emotion recognition. Based on WavLM-Large with a structured layer-freezing strategy (freezing nine transformer

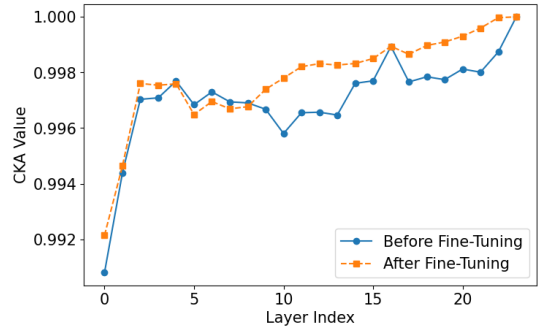


Figure 2: CKA Values Before and After Fine-Tuning across Layers.

layers and fine-tuning rest), our system effectively balances pre-trained feature retention with task-specific adaptation. Integrating an attentive statistics pooling layer and focal loss further enhanced performance, addressing class imbalances and improving model generalization. Our method achieved an F1-Macro score of 0.3661 and an F1-Micro score of 0.3794, outperforming the existing audio-only models and even competing with some audio-text approaches. These results highlight the effectiveness of selective fine-tuning and deep feedforward classification in SER. This study contributes valuable insights into designing more robust emotion recognition systems for real-world applications.

## 7. References

- [1] B. W. Schuller, “Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends,” *Communications of the ACM*, p. 90–99, 2018.
- [2] C. Busso, M. Bulut, and S. Narayanan, *Toward Effective Automatic Recognition Systems of Emotion in Speech*, 2013.
- [3] P. Sharma, V. Abrol, A. Sachdev, and A. D. Dileep, “Speech emotion recognition using kernel sparse representation based classifier,” in *European Signal Processing Conference (EUSIPCO)*, 2016, pp. 374–377.
- [4] B. W. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, “The interspeech 2011 speaker state challenge,” in *Interspeech*, 2011, pp. 3201–3204.
- [5] S. R. Viksit and V. Abrol, “Multi-Layer Acoustic & Linguistic Feature Fusion for ComParE-23 Emotion and Requests Challenge,” in *ACM International Conference on Multimedia*, 2023, p. 9492–9495.
- [6] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, “Cnn+lstm architecture for speech emotion recognition with data augmentation,” in *Workshop on Speech, Music and Mind (SMM)*, 2018, pp. 21–25.
- [7] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “End-to-end multimodal emotion recognition using deep neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 1301–1309, 2017.
- [8] 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*, 2025.
- [9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Interspeech*, 2005, pp. 1517–1520.
- [10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, pp. 335–359, 2008.
- [11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.
- [12] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Communication*, pp. 1162–1181, 2006.
- [13] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, and S. L. Schneider, “Speech emotion recognition using machine learning — a systematic review,” *Intelligent Systems with Applications*, p. 200266, 2023.
- [14] Y. Li, T. Zhao, and T. Kawahara, “Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning,” in *Interspeech*, 2019, pp. 2803–2807.
- [15] Mustaqeem and S. Kwon, “Att-net: Enhanced emotion recognition system using lightweight self-attention module,” *Applied Soft Computing*, pp. 107 101–107 111, 2021.
- [16] W. Wu, C. Zhang, and P. C. Woodland, “Emotion recognition by fusing time synchronous and time asynchronous representations,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6269–6273.
- [17] F. Costa, M. India, and J. Hernando, “Double multi-head attention multimodal system for odyssey 2024 speech emotion recognition challenge,” in *ISCA The Speaker and Language Recognition Workshop (Odyssey)*, 2024, pp. 266–273.
- [18] R. Lotfian and C. Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” *IEEE Transactions on Affective Computing*, pp. 471–483, 2019.
- [19] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.
- [20] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” pp. 1–4, 2015. [Online]. Available: <https://arxiv.org/abs/1510.08484>
- [21] M. Chen, H. Zhang, Y. Li, J. Luo, W. Wu, Z. Ma, P. Bell, C. Lai, J. Reiss, L. Wang, P. Woodland, X. Chen, H. Phan, and T. Hain, “1st place solution to odyssey emotion recognition challenge task1: Tackling class imbalance problem,” in *ISCA The Speaker and Language Recognition Workshop (Odyssey)*, 2024, pp. 260–265.
- [22] H. Härmä and T. Alumäe, “Taltech systems for the odyssey 2024 emotion recognition challenge,” in *ISCA The Speaker and Language Recognition Workshop (Odyssey)*, 2024, pp. 255–259.
- [23] M. Shamsi, L. Gauder, and M. Tahon, “The conilium proposition for odyssey emotion challenge: Leveraging major class with complex annotations,” in *ISCA The Speaker and Language Recognition Workshop (Odyssey)*, 2024, pp. 281–287.
- [24] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” in *Interspeech*, 2018, pp. 2252–2256.
- [25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: an imperative style, high-performance deep learning library*, 2019, pp. 1–12.
- [26] L. Goncalves, A. N. Salman, A. R. Naini, L. Moro-Velázquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso, “Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results,” in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 247–254.
- [27] 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*, p. 3451–3460, 2021.
- [28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *International Conference on Neural Information Processing Systems (NeurIPS)*, 2020, pp. 12 449–12 460. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf)
- [29] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, pp. 1505–1518, 2022.
- [30] 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 (ICML)*, 2023, pp. 1–28.
- [31] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.
- [32] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, p. 4879–4883.
- [33] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, “Similarity of neural network representations revisited,” in *International Conference on Machine Learning (ICML)*, 2019, pp. 3519–3529. [Online]. Available: <https://proceedings.mlr.press/v97/kornblith19a.html>