



Developing a High-performance Framework for Speech Emotion Recognition in Naturalistic Conditions Challenge for Emotional Attribute Prediction

Thanathai Lertpetchpun*, Tiantian Feng*, Dani Byrd, Shrikanth Narayanan

University of Southern California, Los Angeles, USA

lertpetc@usc.edu, tiantiaf@usc.edu

Abstract

Speech emotion recognition (SER) in naturalistic conditions presents a significant challenge for the speech processing community. Challenges include disagreement in labeling among annotators and imbalanced data distributions. This paper presents a reproducible framework that achieves superior (top 1) performance in the Emotion Recognition in Naturalistic Conditions Challenge (IS25-SER Challenge) - Task 2, evaluated on the MSP-Podcast dataset. Our system is designed to tackle the aforementioned challenges through multimodal learning, multi-task learning, and imbalanced data handling. Specifically, our best system is trained by adding text embeddings, predicting gender, and including “Other” (O) and “No Agreement” (X) samples in the training set. Our system’s results secured both first and second places in the IS25-SER Challenge, and the top performance was achieved by a simple two-system ensemble.

Index Terms: speech emotion recognition, speech foundation model, multi-task learning, representation learning

1. Introduction

Recent advances in foundation models have opened new frontiers in speech processing, leading to more accurate and robust solutions in automatic speech recognition [1, 2], speaker recognition [3], and spoken language understanding [4]. Despite the success of these foundation models, accurately predicting emotions expressed in speech in natural interaction contexts still encounters several challenges, including perception (annotator) subjectivity [5, 6], imbalanced distribution in emotion labels, and mismatch in speaker and environment settings [7]. The Speech Emotion Recognition in Naturalistic Conditions Challenge (IS25-SER Challenge) offers a platform to investigate some of these limitations within a research community “challenge” setting.

In this paper, we focus on Task 2 IS25-SER Challenge¹ [8]. Task 2 involves predicting the level of three-dimensional motion attributes: arousal, valence, and dominance [9]. Arousal reflects the intensity of the emotion, while valence indicates its polarity. Dominance represents the level of control a person feels in an emotional state. These emotion attributes have proved challenging to approach computationally; this was reflected in the previous Odyssey-Speech Emotion Challenge [10], in which none of the participating teams scored higher than 0.5327 in concordance correlation coefficient (CCC) [11]. Consequently, in this study, we follow that baseline to extract speech embeddings, namely a pre-trained speech model (WavLM Large [12]) with a series of linear classification layers.

Aside from the architecture choices, we aim for a simple and reproducible system design. Although ensembling a set of trained models is a common practice that yields competitive performance in several domains [13, 14, 15], here we demonstrate that a combination of just two systems can achieve promising performance—top-tier in the challenge, outscoring the baselines by a considerable margin. Specifically, our best system reaches an average CCC of 0.6076, 4.8% higher than the baseline. While this improvement may seem small overall, it is 35.43% higher than the first runner-up. While our current system yields state-of-the-art performance, our exploration during the challenge reveals many design considerations that lead to compromises in performance. Therefore, we share both positive findings, and the limitations that remain, as insights for future system design.

While previous work by [16] suggests that speaker information is beneficial in predicting emotions, most systems in the recent Odyssey-SER challenge do not integrate speaker-related information, such as gender or speaker identity. This motivates us to explore joint modeling of the emotional attributes and speaker information. Specifically, we experiment with speaker representation learning and show that not all speaker-related information is helpful for emotion attribute prediction. Our results suggest that learning gender information improves the model on the SER task, whereas learning speaker representations leads to decreased performance. Hence, we perform a detailed analysis to explain why speaker representation learning fails to improve SER performance in this challenge.

Furthermore, we address the challenge of imbalanced emotion classes in the training dataset by undersampling, so that the model identifies emotion attributes more uniformly across each primary emotion. Our results show that this simple method is effective in improving SER performance. Building upon this, we further investigate the impact of including or excluding speech samples with mixed annotations, labeled as “Other” (O) and “No Agreement” (X). Most studies choose to discard these labels without an explicit rationale. Instead, our findings indicate that these labels contain valuable information that helps the model better learn emotion attributes.

Our model is at <https://github.com/tiantiaf0627/vox-profile-release>. We believe that our system can serve as a new baseline candidate for future SER challenges. The key contributions and findings are summarized below:

- Incorporating text representations alongside speech embeddings significantly improves model performance.
- Leveraging auxiliary labels such as gender information during training enhances the model’s ability to generalize.
- Addressing data imbalance through undersampling improves model performance, and accelerates training speed.

*These authors contributed equally.

¹https://lab-msp.com/MSP-Podcast_Comppetition/IS2025/

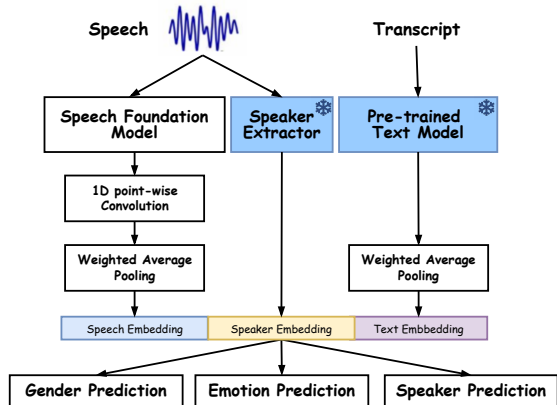


Figure 1: Our proposed multimodal SER framework utilizes a speech foundation model, speaker extractor, and text model. The pre-trained speaker extractor and pre-trained text model are frozen in every setting of our framework.

- Including “Other” and “No Agreement” labels in training, unlike most previous work, improves SER performance.

2. Method

2.1. Foundation Models

Foundation models such as WavLM [12] and Whisper [17] have been shown to improve the performance in SER [18, 19]. In the previous Odyssey-SER Challenge, the baseline system [10], which used the foundation model following downstream MLPs, outperformed every submission. Thus, following the baseline system, our system is built upon speech foundation models, including WavLM and Whisper. For WavLM, we computed the encoder output using a weighted average pooling from all hidden outputs from encoder layers. On the other hand, for Whisper, we used the hidden output from the last layer as the encoder output. Specifically, we process the encoder outputs using 1D point-wise convolutions.

Furthermore, previous studies find that text embeddings from a pre-trained text model [19, 20] and speaker embeddings from a speaker feature extractor [16] are beneficial to improve SER. We concatenated the speech embedding, text embedding, and speaker embedding to obtain the final representation. Our proposed framework is illustrated in Figure 1.

2.2. Attribute Representation Learning

The final representations are fed into three distinct 2-layer MLPs to predict arousal, valence, and dominance in a regression task. Since the challenge uses the concordance correlation coefficient (CCC) as the evaluation metric [11], we apply CCC loss as our learning objective. Moreover, we perform multi-task learning to simultaneously learn emotion attributes alongside speaker-specific information such as gender and speaker identity. To achieve this, we pass the final representation through a separate 2-layer downstream model to predict the gender information, optimizing with cross-entropy loss. Subsequently, we feed the same representation into another downstream model to learn the speaker identity representation. When training the model for speaker prediction, we use additive angular margin loss [21], a widely used loss for speaker verification tasks, and

exclude the speaker embedding from the final representation. Our final loss function is as Eq. 1.

$$L = L_{emotion} + \alpha L_{gender} + \beta L_{speaker} \quad (1)$$

where $L_{emotion}$ represents summation of the CCC loss for each attribute, L_{gender} indicates the cross-entropy loss between for gender prediction, $L_{speaker}$ represents the AAM-Softmax loss for speaker prediction, and α and β denote weights of each loss.

2.3. ‘Other’ and ‘No Agreement’ Label Inclusion

One debated topic in SER is how to handle No Agreement (X) samples and Other (O) samples (we will refer to these two labels as “O/X”). No Agreement (X) samples represent speech samples that lack consensus in emotion annotation, while the Other (O) category indicates that the emotion category falls outside the eight primary emotions described in [10]. Many previous studies simply excluded these two labels from training, development, and test [22, 23]. However, the speech samples of these two labels account for 22.46% of the training set. Instead of discarding these samples, we include them in the training set and demonstrate their effectiveness in improving SER modeling.

2.4. Engineering Design Choices to Improve Performance

Despite our efforts to tackle the challenges described in the previous section, many hurdles still need to be addressed. As announced by the organizer, the test set is balanced in primary emotion labels. Therefore, the proportion of each primary emotion in the test set differs substantially from that in the training set. Training the model on the training set can lead the model to overfit to the majority class, such as Neutral. Moreover, validating our model to choose the best model based on the development may not reflect the actual score on the test set due to the distribution mismatch between the development and test set.

Validation Selection We develop another auxiliary metric to choose our best model. We aim to establish another CCC to reflect on the balanced dataset using existing imbalanced datasets. Specifically, we perform n trials to compute the average CCC in this manner. In each trial, we randomly select k samples from each primary emotion and compute CCC using those samples. Although the metrics may exhibit noise due to randomness in the sample selection, this approach ensures that the CCC is evaluated on the balanced set following the same primary emotion distribution as the test set.

Undersampling We apply the undersampling to ensure that each emotion sample is fed into the model with equal probability. We find $m = \min_i(|emotion_i|)$ where $|\cdot|$ denotes the number of samples in that emotion. Then, we randomly select m samples from each class into the training set.

Target Normalization We apply the sigmoid function to the model output to map the emotion attribute prediction to the range $[0, 1]$. These values are subsequently mapped back to the range $[1, 7]$ using the min-max normalization. Specifically, the output can be computed as:

$$\hat{x}_i = x_i \times [max_j(y_{ij}) - min_j(y_{ij})] + min_j(y_{ij}) \quad (2)$$

where x_i represents the prediction of each attribute and y_{ij} denotes the value of each sample (j) from each attribute (i).

3. Experimental Setup

3.1. MSP-Podcast Dataset

The MSP-Podcast dataset v1.12 [24] was used as the dataset for the IS25-SER Challenge. The dataset comprises podcast data from the Internet which are spontaneous speech and contain human emotions expressed in a naturalistic context. The dataset was annotated with different emotion categories and attributes through crowd-sourcing. Moreover, there are at least five annotators per sample to ensure the reliable labeling. However, as annotators sometimes cannot converge on a single emotion, such samples will be labeled as “No Agreement” (X) and are excluded from the test set. The dataset contains five subsets: the training set, the development set, and the three unique test sets. In the challenge, we were given the training and development set, while, the test-3 set, for which annotation labels were not made publicly available, was used for test evaluation. The annotators also labeled each sample with regard to three different attributes: arousal (negative to positive), valence (calm to active), and dominance (weak to strong), with ratings on a scale from 1 to 7. Our experiment used the entire training set, consisting of ten emotion classes: Neutral, Happy, Sad, Disgust, Angry, Contempt, Fear, Surprise, Other, and No Agreement. Detailed information about the training and development sets are reported in Table 1

Since the challenge rules allowed the training and validation set to be used arbitrarily, we chose to include some parts of the validation set in the training set to improve the system. Specifically, the samples with label “Other” (O) and “No Agreement” (X) were included in the training set and excluded from the development set.

Table 1: Dataset statistics of IS25-SER challenge.

	Training	Development	Test
Neutral	29,243	7,423	400
Happy	16,717	6,344	400
Sad	6,306	2,341	400
Disgust	1,432	542	400
Angry	6,731	5,836	400
Contempt	2,495	1,459	400
Fear	1,120	326	400
Surprise	1,120	987	400
Other	2,948	642	0
No Agreement	15,932	6,061	0

3.2. Experimental Details

Although we acknowledge that hyperparameter tuning for each different system can help improve the aggregate models significantly, we did not focus on this aspect for this challenge: —our goal in this work was to design a reproducible and simple state-of-the-art pipeline. Consequently, all the systems were trained with the same hyperparameters. We used a learning rate of 5e-5 and trained the model for a total of 50 epochs. The filter size in the downstream model is 256 across all experiments. The maximum input speech duration is 15 seconds. The best model was selected based on the performance on the validation set. For the multitask learning, we used $\alpha = 1$ and $\beta = 0.1$.

The systems are trained using the undersampling method as the default method in most experiments. We use WavLM Large and Whisper Large-V3 for the speech foundation model, and RoBERTa-Large as the pre-trained text model. Following the challenge baseline, we fine-tune the pre-trained WavLM

Table 2: Comparison between models trained without text embeddings and with text embeddings.

	A	V	D	Average
WavLM	0.6530	0.7000	0.5651	0.6394
MMWavLM	0.6694	0.7334	0.5735	0.6588

Large along with the downstream models. However, we froze the model for the Whisper Large-V3 and RoBERTa-Large. We will add the prefix “MM (multimodal)” to the model to indicate that the model uses text embedding as an auxiliary input. As we did not have access to the test set and we can only submit the score to the challenge once a week, it was challenging for us to do a systematic evaluation on the test set. Therefore, the performances except for Table 6 are reported on the validation set.

4. Results and Discussion

4.1. Pre-trained Text Model

First, we investigated whether textual information impacts the SER performance by adding text embedding to the final representation. We designed an experiment to compare WavLM Large and WavLM Large + RoBERTa-Large (MMWavLM). Table 2 shows that the MMWavLM performs better in every attribute prediction.

4.2. Attributes Learning

As gender prediction has proven effective in classifying emotions [25], we investigate whether it can also improve the regression task. Table 3 shows that multi-task learning with gender information increases the CCC scores in arousal, valence, and dominance predictions. However, adding the speaker embedding or predicting the speaker label worsens the model performance. We analyzed the speaker label to understand the reasons behind this. First, approximately 80% of the shared dataset were represented by just 26% of the total speakers, causing the model to be overfit to a limited subset of seen speakers. Second, the pre-trained speaker extractor does not generalize well on the dataset. We randomly selected 100 utterances from 10 speakers and extracted speaker embedding from them. As shown in Fig 2, speaker embeddings failed to cluster together, making it difficult for the model to learn.

Table 3: Comparison of models trained with speaker embeddings, speaker prediction, and gender prediction.

	A	V	D	Average
MMWavLM	0.6694	0.7334	0.5735	0.6588
+ S Emb	0.6483	0.6962	0.5643	0.6363
+ S Pred	0.6428	0.7141	0.5684	0.6418
+ G Pred	0.6662	0.7367	0.5787	0.6605

4.3. Undersampling

As shown in Table 4, undersampling significantly helps the model learn the true distribution of each emotion and prevents the model from overfitting only on certain emotions. Specifically, the CCC on valence increases significantly up to 31.57%, leading the average CCC increases up to almost 10%. Furthermore, our experiments show that training with undersampling helps the model converge faster, consequently reducing the training time.

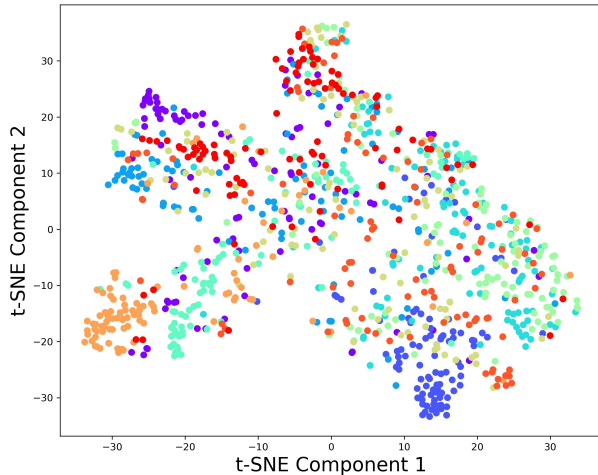


Figure 2: The *t-sne* plot of speaker embeddings from the pre-trained speaker extractor. Each color represents different speakers.

Table 4: Comparison between models trained with balanced and imbalanced data samples

	A	V	D	Average
Whole dataset	0.6620	0.5574	0.5795	0.5996
Undersampling	0.6694	0.7334	0.5735	0.6588

4.4. Other and No Agreement labels

Although O/X samples are not in test sets, adding or removing them substantially impacts performance. Table 5 shows that removing data from the training set, whether it belongs to O or X, degrades performance. The results also show that adding O/X (V) samples to the training set improves the performance. Moreover, adding O/X (V) samples and removing non-O/X (T) samples give the best performance for the SER system. This indicates that, despite the annotators’ lack of agreement in classifying emotions for these samples, they still provide valuable information to the SER system. However, due to submission limitations, we did not submit the system trained with the whole training set + O/X (V) - Non-O/X (T) to the leaderboard, even though it achieved the best performance on the validation set.

Table 5: Comparison between models trained with different amount of data. “O/X” denotes Other (O) and No Agreement (X), “+” denotes including the data, “-” denotes removing the data, and (T) and (V) indicate whether the samples are from the training or validation set

	A	V	D	Average
Training Set	0.6694	0.7334	0.5735	0.6588
- O/X(T)	0.6622	0.7342	0.5774	0.6579
- Non-O/X (T)	0.6692	0.7332	0.5720	0.6581
+ O/X(V)	0.6655	0.7125	0.6034	0.6605
- Non-O/X (T)	0.6830	0.7582	0.6145	0.6852

4.5. Results of the IS25-SER Challenge

The results from the IS25 SER Challenge leaderboard show that our two systems achieved first and second place in the Challenge. “+ S Emb” denotes that the model uses speaker embedding from the pre-trained speaker extractor in the final repre-

Table 6: The results from the IS25-SER Challenge on Task 2. “A”, “V”, and “D” denote arousal, valence, and dominance, respectively.

	A	V	D	Average
Baseline	0.6385	0.6232	0.4775	0.5797
WavLM	0.5339	0.5992	0.4132	0.5155
MMWavLM				
+ Undersampling	0.6099	0.6189	0.4736	0.5675
+ S Emb	0.6079	0.5934	0.4479	0.5497
3-system (2nd)	0.6794	0.634	0.4878	0.6004
2-system (1st)	0.6829	0.642	0.4980	0.6076

sentation while “+ Undersampling” denotes that the model is training with the undersampling method.

As shown in Table 6, MMWavLM performs significantly better than WavLM, indicating that textual embeddings also improve the performance in the test set. Moreover, adding speaker embedding causes the model to overfit on the speaker information, leading to degraded performance, as suggested by our experiments on the validation set.

Our best system consists of only two models: WavLM with gender prediction, undersampling technique and O/X (V) inclusion, and MMWhisper with only undersampling. We ensembled the system by averaging the CCC prediction from each one. The results indicate that ensembling only a few systems can exhibit strong performance. Moreover, we test a 3-system ensemble: WavLM with gender prediction, undersampling technique and O/X (V) inclusion; WavLM with gender prediction and undersampling technique; and MMWhisper with gender prediction, undersampling technique and O/X (V) inclusion. However, this 3-system underperforms the 2-system on the test set. Lastly, it is worth noting that since the number of submissions is limited, it is challenging to demonstrate which aspects of our methods contribute the most to improving scores.

4.6. On Further Improvement

While our current system achieves competitive results in the Challenge, we identify several promising directions for developing next-generation, state-of-the-art SER systems. As suggested by the previous baseline in the Odyssey-SER Challenge [10], training on a separate system for each attribute demonstrated strong performance. Moreover, instead of freezing the pre-trained Whisper Large-V3 and RoBERTa-Large, we could apply LoRa as proposed by [22, 26] to further improve the performance. We believe that integrating these two methods would not only improve the performance of our system but also maintain its simplicity.

5. Conclusion

We proposed a simple and reproducible state-of-the-art model for the IS25 SER Challenge. Our system consists of the ensemble between pre-trained MMWavLM Large fine-tuned with gender prediction, inclusion of O/X label from both the training and validation sets, and using undersampling, with MMWhisper Large-V3 fine-tuned via undersampling. The ensemble system can gain performance up to 0.6076 in average CCC and achieved 1st place in the Challenge leaderboard. Lastly, our analysis shows why using learning speaker representation does not improve the performance of the SER system in this dataset.

6. Acknowledgment

We gratefully acknowledge support from IARPA ARTS (award number 140D0424C0067, JHU subcontract) from the Office of the Director of National Intelligence. NSF (SCH 2204942).

7. References

- [1] J. Shi, D. Berrebbi, W. Chen, E.-P. Hu, W.-P. Huang, H.-L. Chung, X. Chang, S.-W. Li, A. Mohamed, H. yi Lee, and S. Watanabe, "ML-SUPERB: Multilingual Speech Universal PERFORMANCE Benchmark," in *Interspeech 2023*, 2023, pp. 884–888.
- [2] J. Shi, S.-H. Wang, W. Chen, M. Bartelds, V. Bannihatti Kumar, J. Tian, X. Chang, D. Jurafsky, K. Livescu, H. yi Lee, and S. Watanabe, "ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets," in *Interspeech 2024*, 2024, pp. 1230–1234.
- [3] S. Peng, W. Guo, H. Wu, Z. Li, and J. Zhang, "Fine-tune Pre-Trained Models with Multi-Level Feature Fusion for Speaker Verification," in *Interspeech 2024*, 2024, pp. 2110–2114.
- [4] H. Futami, S. Arora, Y. Kashiwagi, E. Tsunoo, and S. Watanabe, "Finding Task-specific Subnetworks in Multi-task Spoken Language Understanding Model," in *Interspeech 2024*, 2024, pp. 802–806.
- [5] W.-S. Chien, S. G. Upadhyay, and C.-C. Lee, "Balancing Speaker-Rater Fairness for Gender-Neutral Speech Emotion Recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 861–11 865.
- [6] B. M. Booth and S. S. Narayanan, "People make mistakes: Obtaining accurate ground truth from continuous annotations of subjective constructs," *Behavior Research Methods*, vol. 56, no. 8, pp. 8784–8800, 2024.
- [7] C.-C. Lee, T. Chaspari, E. M. Provost, and S. S. Narayanan, "An Engineering View on Emotions and Speech: From Analysis and Predictive Models to Responsible Human-Centered Applications," *Proceedings of the IEEE*, pp. 1–17, 2023.
- [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*, vol. To appear, Rotterdam, The Netherlands, August 2025.
- [9] M. Grimm, K. Kroschel, E. Mower, and S. S. Narayanan, "Primitives-based evaluation and estimations of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, nov 2007.
- [10] 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.
- [11] B. T. Atmaja and M. Akagi, "Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition," in *Journal of Physics: Conference Series*, vol. 1896, no. 1. IOP Publishing, 2021, p. 012004.
- [12] 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.
- [13] T. Tran, T. D. Bui, and P. Simatis, "ParallelChain Lab's anti-spoofing systems for ASVspoof 5," in *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, 2024, pp. 9–15.
- [14] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "STC Antispoofing Systems for the ASVspoof2021 Challenge," in *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 61–67.
- [15] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating Frequency Translational Invariance in TDNNs and Frequency Positional Information in 2D ResNets to Enhance Speaker Verification," in *Interspeech 2021*, 2021, pp. 2302–2306.
- [16] T. Lertetchpun and E. Chuangsuwanich, "Instance-based Temporal Normalization for Speaker Verification," in *Proc. INTERSPEECH 2023*, 2023, pp. 3172–3176.
- [17] 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.
- [18] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Interspeech 2021*, 2021, pp. 3400–3404.
- [19] T. Feng and S. Narayanan, "Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 116–12 120.
- [20] Y. Li, P. Bell, and C. Lai, "Fusing asr outputs in joint training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.
- [21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.
- [22] H. Härm and T. Alumäe, "TalTech Systems for the Odyssey 2024 Emotion Recognition Challenge," in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 255–259.
- [23] 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.
- [24] 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.
- [25] Y. Gao, C. Chu, and T. Kawahara, "Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining," in *Interspeech 2023*, 2023, pp. 3637–3641.
- [26] 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.