

Adapting WavLM for Speech Emotion Recognition

Daria Diatlova^{1,2,3}, Anton Udalov^{3,4}, Vitalii Shutov^{1,2}, Egor Spirin^{1,2,3}

deepvk¹, VK², VK Lab³, ITMO⁴

d.diatlova@vk.team, a.udalov@niuitmo.ru, vi.shutov@corp.vk.com, egor.spirin@vk.team

Abstract

Recently, the usage of speech self-supervised models (SSL) for downstream tasks has been drawing a lot of attention. While large pre-trained models commonly outperform smaller models trained from scratch, questions regarding the optimal fine-tuning strategies remain prevalent. In this paper, we explore the fine-tuning strategies of the WavLM Large model for the speech emotion recognition task on the MSP Podcast Corpus. More specifically, we perform a series of experiments focusing on using gender and semantic information from utterances. We then sum up our findings and describe the final model we used for submission to Speech Emotion Recognition Challenge 2024.

1. Introduction

Speech Emotion Recognition (SER) is a highly-demanded task in practical applications. These include emotion recognition for analyzing customer satisfaction from voice messages and calls [1], creating a robust Automatic Speech Recognition (ASR) model [2], or providing essential content for generating avatars in video calls [3].

Recently, the interest among machine learning researchers in using self-supervised learning (SSL) models for audio processing has grown [4, 5, 6]. This can be explained by the successful application of transfer learning, which involves leveraging the knowledge of SSL upstream models, initially trained on a large amount of unlabeled data, by a downstream model [7]. The most commonly used method to transfer knowledge to a downstream model involves fine-tuning the upstream model by adding new layers from the downstream model [7].

The application of transfer learning in the audio domain, particularly SER, has received considerable attention of late. In [8], the authors demonstrated that using an SSL model as a feature extractor outperforms SER solutions based on low-level descriptors, even when provided with eight times less labeled data. Following that, works focusing on the specific utilization of SSL to enhance SER performance have emerged [8, 9, 10]. These approaches include feature-wise objective constraints [9], the utilization of text encoders [8, 10], pooling mechanics [11], and more.

In this paper, we share the technical details of creating a SSL-based model for speech emotion recognition using WavLM [6] in the Odyssey 2024 challenge [12], and address several research questions.

- Does applying time-dimensional pooling over WavLM output impact the quality of speech emotion recognition?
- Can incorporating information about the speaker’s gender enhance emotion classification?
- Does leveraging information about the text of an utterance help emotion classification?

Using the MSP Podcast dataset for verification, we obtained a solution which we then used for the Odyssey 2024 challenge.

2. Related Work

In this section, we discuss WavLM [6], a self-supervised audio processing model, and its implications for SER. We explore cutting-edge strategies for information gathering and highlight the importance of advanced pooling techniques. Additionally, we consider the incorporation of auxiliary data such as text and gender embeddings to enhance SER systems.

2.1. WavLM

WavLM [6] is a self-supervised audio processing model that utilizes the Transformer encoder architecture [13]. The model has been enhanced with a gated relative position bias, improving its ability to model sequential information. Specifically, WavLM takes a raw audio signal $X \in \mathbb{R}^n$, where n is the number of samples in the raw signal, and produces an encoded output $Z \in \mathbb{R}^{[l \times m \times h]}$, where:

- l is the number of Transformer layers,
- m is the sequence length,
- h is the hidden dimension.

To accommodate various computational needs and applications, WavLM is available in two configurations: Base, with 12 Transformer layers, and Large, with 24 layers. The WavLM model is trained on extensive audio corpora to encode a wide range of audio applications. It incorporates a speech denoising objective and masked speech prediction during pre-training, enabling efficient extraction of speech and acoustic features.

2.2. Self-Supervised Learning in SER

Self-supervised learning (SSL) has shown promising results in SER [8, 9], offering innovative ways to decode emotional subtleties in speech. It is common for SER models built using SSL to use the Upstream + Downstream model architecture [10]. The upstream block aims to extract features for the downstream model to predict emotion classes.

Before passing the upstream model’s output to the downstream model, it is typically reduced in both the layer and time dimensions. A common approach for layer-wise reduction is:

- to take the $\hat{Z} \in \mathbb{R}^{[l_n \times m \times h]}$ hidden representation, where l_n is the output from the last Transformer layer [14].
- to select the weighted average sum of all layers [14].

For the second case, a vector of trainable weight parameters $w \in \mathbb{R}^l$ is initialized. Then, a layer-wise averaged output $\hat{Z} = Z \cdot w$, where $\hat{Z} \in \mathbb{R}^{[m \times h]}$, is passed through the time dimension reduction block. The simplest implementation of the time-reduction block is to apply average pooling [15]:

$x_{\text{out}} = \sum_{i=1}^m \frac{\hat{Z}_i}{m}$, where $x_{\text{out}} \in \mathbb{R}^{[h]}$. After this step, x_{out} is passed to the downstream block for further classification.

The choice of the downstream architecture, as well as the modification applied to the process of obtaining x_{out} , can enhance the quality of the SER model.

The combination of mean average pooling and sophisticated aggregators, such as ECAPA-TDNN [16], with linear classifiers improved SER performance in [10]. In [11], attentive correlation pooling by adding weights to the statistical estimates was introduced. This technique aims to address the challenge of utilizing emotional information across extended periods of time. Additionally, the authors of [11] demonstrated the benefits of label smoothing for dealing with the emotional ambiguity of speech. State-of-the-art SER performance was achieved by combining attentive correlation pooling with label smoothing and dropout, without the need for fine-tuning the SSL model.

2.3. Information Accumulation

Information accumulation techniques, particularly those involving pooling mechanisms and the integration of auxiliary information such as text and gender embeddings, play a pivotal role in enhancing SER systems.

In [17], introduced a two-stage fine-tuning process for wav2vec 2.0, emphasizing the use of automatic speech recognition (ASR) and gender pretraining to enrich the SER model with additional features. By first embedding these attributes into the model and then focusing on emotion recognition, the approach effectively leverages multitask learning (MTL) and adversarial learning to incorporate auxiliary task information, aiming to mitigate the gradient conflict issue commonly observed in MTL setups. This method highlights the utility of informed feature extraction for boosting SER performance.

Furthermore, the concept of attentive statistics pooling [18] proposes an advanced pooling strategy that goes beyond the traditional average pooling method by including higher-order statistics, such as standard deviation, weighted by attention mechanisms. This approach calculates both the weighted mean and standard deviation of frame-level features, enhancing the model’s ability to capture speaker discriminability through feature variability over time. By applying attention to both means and variances, it introduces a more nuanced understanding of speaker characteristics vital for SER.

3. Model Modifications

In this section, we propose several modifications aimed at addressing our research questions. More specifically, we describe 1) replacing the commonly used average pooling with pooling types, 2) using gender information for predictions, and 3) using text information in conjunction with the pooled output.

3.1. Time-Dimensional Pooling

The output of WavLM is usually reduced over time-dimension using simple average pooling and then passed to the classification head, as discussed in Section 2.2. Instead of average pooling, we suggest using STD and attention pooling.

3.1.1. STD Pooling

In standard deviation (STD) pooling the average pooling output (Eq. 1, where m is the sequence length and Z_i is the layer-wise averaged output of WavLM for a time frame i) is used along with the standard deviation output (Eq. 2). The result-

ing statistics are concatenated over the hidden dimension (Eq. 3). Compared to a simple average pooling, which is generally robust to noisy output, STD pooling focuses on regions with high variability. It can emphasize distinct features that differ from neighboring regions, which can be beneficial for emotion classification.

$$x_{\text{mean}} = \sum_{i=1}^m \frac{\hat{Z}_i}{m} \quad (1)$$

$$x_{\text{std}} = \sqrt{\frac{\sum_{i=1}^m (\hat{Z}_i - x_{\text{mean}})^2}{m}} \quad (2)$$

$$x_{\text{pooled}} = [x_{\text{mean}}, x_{\text{std}}] \quad (3)$$

3.1.2. Attention Pooling

Attention pooling [18] makes it possible for certain frames to become more valuable based on context. Attention weights are computed as shown in Eq. 4, where $p \in \mathbb{R}^h$ is the trainable parameter and \cdot denotes matrix multiplication.

\hat{Z} is multiplied element-wise by the attention weights (Eq. 5), then the mean and variance are computed from the resulting output and concatenated as described in Eq. 3.

$$w_{\text{att}} = \text{Softmax}(p \cdot \tanh(\hat{Z})), w_{\text{att}} \in \mathbb{R}^m \quad (4)$$

$$\hat{Z} = \hat{Z} \cdot w_{\text{att}}, \hat{Z} \in \mathbb{R}^{[m \times h]} \quad (5)$$

Attention pooling provides the opportunity to give certain frames more weight than others when computing the mean and standard deviation.

3.2. Gender Conditioning

The MSP Podcast corpus includes gender labels: male and female. Therefore, we added a lookup table of trainable embeddings.

$$x_{\text{output}} = x_{\text{pooled}} \cdot e_{\text{gender}} \quad (6)$$

For the conditioning mechanism, we used a pointwise multiplication of pooled output and gender embeddings as shown in Eq. 6. In section 5.2, we provide an ablation study on other conditioning mechanisms.

3.3. Text Conditioning

Along with gender information, the MSP Podcast corpus contains textual annotations paired with audio clips. To combine the textual information with the encoded audio representation and gender embeddings, we employed the Sentence Transformer [19], which maps transcriptions to a 384-dimensional dense vector, e_{text} . Conditioning was applied as depicted in Eq. 7, where F represents a sequence of transformations: Linear Layer [20], Layer Normalization [21], ReLU [22], and Dropout [23].

$$x_{\text{out}} = \frac{x_{\text{pooled}} + e_{\text{gender}} + F(e_{\text{text}})}{3} \quad (7)$$

The ablation study on conditioning approaches can be found in Section 5.3.

4. Experimental Setup

In this section, we provide an overview of the dataset, the baseline proposed by the Odyssey Challenge 2024 [12] Organizers, and the setup for training and evaluating the models.

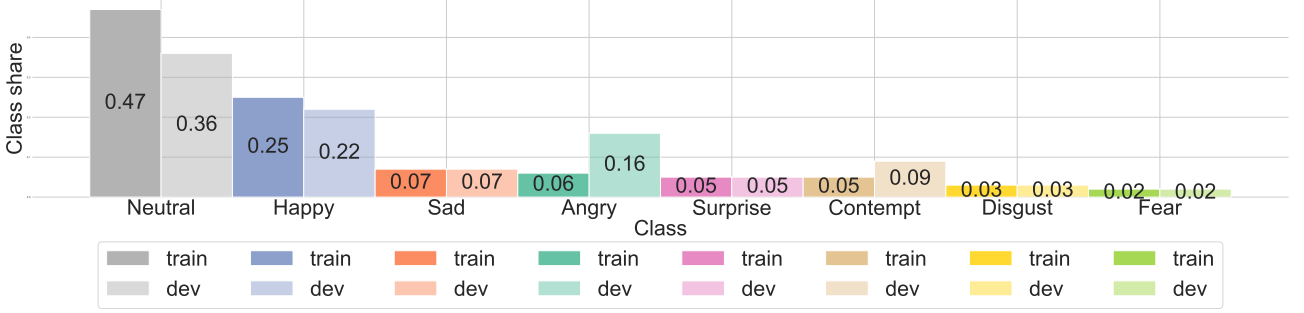


Figure 1: Label distribution in the training and development sets. Note that the dataset is highly imbalanced, and the distribution of samples for each class relative to the total number of classes differs between the training and validation datasets.

4.1. Dataset

The dataset used for training and evaluation is the MSP Podcast Corpus [24]. It consists of 90,522 speaking turns, with the original train, development, and test3 split. The test3 dataset does not have publicly available labels. Therefore, all our experiments, except for the final submission result, are presented using only the development set.

For the train and development set labels, we used the consensus (plurality vote of 3 out of 5 assessors) of the primary emotion, filtering out the classes “Other” and “No agreement”. This resulted in 53,386 samples in the training set and 15,341 samples in the development set. On the early stages of experiments we collected the “confident” training set by taking the consensus (plurality vote of 4 out of 5 assessors) of the primary emotion. This resulted in just 23,509 samples, but did not lead to any improvement of the F1-macro score.

The MSP Podcast dataset exhibits a high level of imbalance, with the class distribution of the training and development sets shown in Figure 1. Note that the order of the number of samples of each class relative to the total number of classes is different in the training and validation datasets. Early on in our experiments, we attempted to balance the dataset by augmenting utterances from classes with fewer samples, but did not observe any significant improvements.

The test3 dataset lacks gender labels, unlike the training and validation datasets. To obtain gender labels, we utilized an internal gender-prediction model. As for text transcriptions, we observed that some examples in the training and validation datasets also lack text transcriptions. We used the Whisper [25] V3 Large model to transcribe these audio recordings, along with the test3 dataset.

4.2. Baseline

An official baseline was provided by the organizers of the Odyssey 2024 challenge [12]. The baseline model is a WavLM Large model, as described in Section 2.1, which produces the output $Z \in \mathbb{R}^{[l \times m \times h]}$, where $l = 24$ (Transformer layers), $h = 768$ (hidden dimension), and m is the dynamically changed sequence length shape.

For classification, the weighted average sum of all layers is selected as discussed in Section 2.2. The pooled output is then passed through the projection layer, which consists of the Linear Layer [20], Layer Normalization [21], ReLU [22], and Dropout [23]. Then, the attention pooling block (Section 3.1.2) is applied to the output and passed through the final Linear Layer to get the logits for each emotion class.

The model is trained with the weighted cross entropy loss [26], the weights are declared as shown in Eq. 8, where w_i is the weight for class i , N_{total} is the size of train dataset, m is the number of classes and N_{class} is the number of samples in class i .

$$w_i = \frac{N_{\text{total}} \cdot N_{\text{class}}}{m} \quad (8)$$

4.3. Training and Evaluation

All our experiments are based on WavLM Large. Unlike the baseline model, we used a single Linear Layer as a projection layer. We did not notice a significant difference in quality from the baseline model. However, our version is less memory-intensive, which is crucial when training large models.

The size of the projector linear layer is set to 256 by default unless specified otherwise. We performed various experiments by adjusting the projection layer size, with the results detailed in Table 1. From the results, it is evident that the optimal choices for the projection layer size are 16, 32, and 256. The default selection of a hidden size of 256 is motivated by the requirement to combine the output of the projection layer with gender and text embeddings, where a larger embedding size makes it possible to encode more information.

In some of our experiments, we apply label smoothing within cross-entropy, which is motivated by emotional data ambiguity. The transition from hard labels to soft labels consistently outperforms cross-entropy loss and leads to faster convergence during training [27]. Label smoothing is defined by parameter γ , where $1 - \gamma$ of the weight is assigned to the true label, and γ is distributed across other classes.

Table 2 presents the results from applying label smoothing. It can be seen that, in our case, the usage of label smoothing leads to F1-macro score degradation.

All models were trained using the AdamW optimizer [28] with a learning rate of $1e-5$ and weight decay of 0.01 for 20 epochs employing a batch size of 32 on a single A100 GPU.

We keep the layers of WavLM CNN feature encoder frozen during all experiments. For model training, we use audio clips of no longer than 5 seconds. For longer utterances, a 5-second segment is randomly selected on each epoch using cropping.

5. Experiments

In this section, we answer the research questions stated in Section 1 by conducting our experiments with the modifications from Section 3. For statistical significance, we run each experiment with several seeds and average the results.

Table 1: *Best F1-macro scores on the development set with varying pooling output size.*

Projection layer size	F1-macro
16	0.31
32	0.31
64	0.28
128	0.30
256	0.31

Table 2: *Best F1-macro scores on the development set with varying portion of label smoothing.*

Label smoothing	F1-macro
0.0	0.31
0.1	0.30
0.2	0.28

5.1. Does applying time-dimensional pooling over WavLM output impact the quality of speech emotion recognition?

We conducted experiments using STD and attention pooling as described in Section 3.1, and compared the results to using standard average pooling.

From the results presented in Table 3, it is evident that the change of pooling type influences overall performance of the SER model based on F1-macro score. STD pooling yielded the best F1-macro score in our experimental setup, while attention pooling performed worse than standard average pooling in our experiment setup.

Note that the baseline model, which uses average pooling, achieves an F1-macro score of 0.31 on the development set. Therefore, it is worthwhile to increase the number of training steps to achieve its best performance. However, we chose to use STD pooling for the rest of our experiments as it converges faster, which is crucial for working within the Challenge.

Table 3: *Best F1-macro scores on the development set with varying temporal pooling type during training.*

Pooling type	F1-macro
average	0.30
std	0.31
attention	0.28

5.2. Can incorporating information about the speaker’s gender enhance emotion classification?

As outlined in Section 3.2, we utilized gender information to condition the pooled output. For this ablation study, we employed STD temporal pooling.

In this setup, “sum” and “multiplication” denote the addition and the multiplication of the pooled output with the gender embedding, “sum / 2” indicates that the output is divided by 2 after addition. In the “stack + linear” configuration, the gender embedding and pooled output are concatenated along the hidden dimension and passed through a linear layer. The term “CLN” refers to conditioning layer normalization [21], as shown in Eq. 9, where g and f represent linear layers.

$$y = f(e_{\text{gender}}) \cdot \frac{x_{\text{out}} - \text{mean}}{\text{var}} + g(e_{\text{gender}}), \quad (9)$$

Table 4: *Best F1-macro scores on the development set with varying gender conditioning types. Note that F1-macro for “sum” and “sum / 2” is the highest.*

Conditioning	F1-macro
no conditioning	0.31
sum	0.32
sum / 2	0.32
multiplication	0.31
stack + linear	0.29
CLN	0.31

As shown by the results in Table 4, the sum and normalized sum reach the best F1-macro scores. CLN and multiplication reach the second best F1-macro score, but it is equal to no conditioning case. Meanwhile, the stack + linear setup leads to F1-macro degradation.

In summary, combining the WavLM output with gender information can enhance the accuracy of the SER model, but the specific strategy of combining them is crucial.

5.3. Does leveraging information about the text of an utterance help emotion classification?

In this experiment, we followed a setup similar to the one in Section 3.3, incorporating STD pooling together with gender and text embeddings. The conditioning is applied using multiplication, normalized sum (“sum / 3”), and CLN.

For CLN, we first concatenated the gender and text embeddings along the hidden dimension and passed the resulting vector through a linear layer to reduce the hidden dimension of a conditioned vector to the initial size of 256.

From the results presented in Table 5, it is evident that the addition of text information led to the degradation in the F1 score compared to using only the gender embedding. The best result was achieved using CLN and sum / 3 conditioning.

Table 5: *Best F1-macro scores on the development set with varying text conditioning types.*

Conditioning	F1-macro
gender conditioning only	0.32
sum / 3	0.31
multiplication	0.30
CLN	0.31

In conclusion, combining the text information with the encoded WavLM output and gender embeddings did not improve the accuracy of the SER model in our specific case. This can be explained by the chosen text encoder’s insensitivity to emotional speech, as it encodes text primarily based on context. Alternatively, it may be due the limited correlation between text and emotion in most classes, or the strategies used to combine this information.

It is important to mention that we included text embedding in the WavLM output along with the information about gender in our experiment setup. This addition helped us enhance the quality of emotion recognition in the previous experiment (Section 3.2). However, combining it with information about the text may negatively impact recognition quality, as it reduces the

Table 6: The configurations of the models used for the submission. Note that out of the 5 models used for prediction fusion, three were conditioned on gender, and two were conditioned on text.

Parameter	Model 1 (Baseline)	Model 2	Model 3	Model 4	Model 5
Temporal pooling	attention	std	std	std	std
Pooled output size	256	32	256	256	256
Gender conditioning	-	-	multiplication	sum / 3	sum / 3
Text conditioning	-	-	-	sum / 3	sum / 3
Label smoothing	0.0	0.0	0.0	0.0	0.1

Table 7: F1 scores for the models used in the challenge submission. Note that each model achieves the highest F1 score for several classes. The F1-macro score received after the fusion is higher than for other models.

Model	F1 Macro	Neutral	Happy	Angry	Contempt	Sad	Surprise	Disgust	Fear
Model 1 (Baseline)	0.31	0.55	0.57	0.40	0.20	0.13	0.42	0.03	0.15
Model 2	0.33	0.58	0.52	0.53	0.15	0.40	0.19	0.21	0.04
Model 3	0.32	0.51	0.58	0.53	0.38	0.21	0.20	0.14	0.02
Model 4	0.31	0.13	0.59	0.40	0.15	0.42	0.21	0.55	0.04
Model 5	0.31	0.29	0.59	0.42	0.15	0.42	0.21	0.12	0.05
Fusion	<u>0.35</u>	0.53	0.57	<u>0.61</u>	<u>0.43</u>	0.17	0.24	0.19	0.05

total information available in the resulting hidden vector about the WavLM output.

6. Results

In this section, we explain how we combined several models for our Odyssey 2024 challenge submission and discuss the final results.

6.1. Fusion Process

Fusing the predictions from multiple models improve the generalization ability and can lead to improved performance [29]. We employed the Constrained Optimization By Linear Approximation algorithm [30] to train the weights matrix $w_{\text{fusion}} \in \mathbb{R}^{n_m \times l}$, where n_m is the number of models and l is the number of classes. The predicted label p from the predictions matrix $M \in \mathbb{R}^{n_m \times l}$ was determined as shown in Eq. 10. The weights matrix w_{fusion} was trained with the following constraints: $\sum_{i=1}^{n_m} w_{\text{fusion}}^i = \sum_{i=1}^l w_{\text{fusion}}^i = 1$.

$$p = \operatorname{argmax}(\sum_{i=1}^{n_m} M \cdot w_{\text{fusion}}) \quad (10)$$

6.2. Model Configuration

We aimed to combine the best of our models based on the F1-macro score and the models with the most diverse configurations. As a result, we choose 5 models with configurations described in Table 6. The F1 score for each emotion class and F1-macro for each model are reported in Table 7.

Please note that the results in Table 7 were obtained during our work on the challenge, and there are some differences in the F1-macro compared to the results reported in Section 5. For instance, model 3 in Table 7 has the same configuration as the model in the fourth row of Table 4, but it shows an F1-macro of 0.32, as opposed to 0.31. This difference is due to running additional experiments after submission to provide statistically significant results for the research questions in Section 5.

6.3. Result Analysis

Table 7 demonstrates that each of the models selected for fusion achieves the highest F1 score for two or three emotional classes

compared to the others. The idea behind fusion is to leverage the strengths of all models. The last row in Table 7 displays the F1 scores for the fused model predictions. By combining predictions from the 5 models, we achieved an F1-macro score of 0.35, which is 0.02 higher than the best individual model’s F1 score. Notably, the F1 scores for the “Angry” and “Contempt” classes after fusion are also higher than the maximum among the 5 models. For the other classes, the F1 score lies between the lowest and highest values in the column.

7. Conclusion

In this paper, we discussed the modifications made to the WavLM model during its fine-tuning for speech emotion recognition as part of our participation in the Odyssey Challenge 2024. We posed several research questions to investigate the importance of time-dimensional pooling, gender, and text accumulation in training the SER model. Our experiments on the MSP Podcast Corpus led us to the conclusion that incorporating STD pooling with accumulated gender information could enhance the performance of SER models based on WavLM.

8. References

- [1] Christophe Vaudable and Laurence Devillers, “Negative emotions detection as an indicator of dialogs quality in call centers,” in *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 5109–5112.
- [2] Vishnu Vidyadhara Raju Vegesna, Krishna Gurugubelli, and Anil Kumar Vuppala, “Application of emotion recognition and modification for emotional telugu speech recognition,” *Mobile Networks and Applications*, vol. 24, pp. 193–201, 2019.
- [3] Ye Pan, Shuai Tan, Shengran Cheng, Qunfen Lin, Zijiao Zeng, and Kenny Mitchell, “Expressive talking avatars,” *IEEE Transactions on Visualization and Computer Graphics*, 2024.
- [4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in

Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

- [5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, “data2vec: A general framework for self-supervised learning in speech, vision and language,” in *Proceedings of the 39th International Conference on Machine Learning*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, Eds. 17–23 Jul 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312, PMLR.
- [6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Y. Qian, Yao Qian, Micheal Zeng, and Furu Wei, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal on Selected Topics in Signal Processing*, 2021.
- [7] Mohammadreza Iman, Hamid Reza Arabnia, and Khaled Rasheed, “A review of deep transfer learning and recent advancements,” *Technologies*, vol. 11, no. 2, pp. 40, 2023.
- [8] Jonathan Boigne, Biman Liyanage, and Ted Östrem, “Recognizing more emotions with less data using self-supervised transfer learning,” 08 2020.
- [9] Yangyang Xia, Li-Wei Chen, Alexander Rudnicky, and Richard M. Stern, “Temporal Context in Speech Emotion Recognition,” in *Proc. Interspeech 2021*, 2021, pp. 3370–3374.
- [10] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz, “Speech emotion recognition using self-supervised features,” *arXiv preprint arXiv: 2202.03896*, 2022.
- [11] Sofoklis Kakouros, Themis Stafylakis, Ladislav Mošner, and L. Burget, “Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022.
- [12] L. Goncalves, A. N. Salman, A. Reddy Naini, L. Morovelazquez, T. Thebaud, L. Paola Garcia, N. Dehak, B. Sisman, and C. Busso, “Odyssey2024 - speech emotion recognition challenge: Dataset, baseline framework, and results,” in *Odyssey 2024: The Speaker and Language Recognition Workshop*, Quebec, Canada, June 2024, vol. To appear.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [14] Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu, “Vesper: A compact and effective pretrained model for speech emotion recognition,” *IEEE Transactions on Affective Computing*, 2024.
- [15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [16] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *INTERSPEECH*, 2020.
- [17] Yuan Gao, Chenhui Chu, and Tatsuya Kawahara, “Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining,” in *Proc. INTERSPEECH 2023*, 2023, pp. 3637–3641.
- [18] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, “Attentive Statistics Pooling for Deep Speaker Embedding,” in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.
- [19] Nils Reimers and Iryna Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 11 2020, Association for Computational Linguistics.
- [20] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, “Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986,” *Biometrika*, vol. 71, pp. 599–607, 1986.
- [21] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, “Layer normalization,” 2016.
- [22] Abien Fred Agarap, “Deep learning using rectified linear units (relu),” *arXiv preprint arXiv: 1803.08375*, 2018.
- [23] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv preprint arXiv: 1207.0580*, 2012.
- [24] Reza Lotfian and Carlos 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, 2019.
- [25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” *International Conference on Machine Learning*, 2022.
- [26] David R Cox, “The regression analysis of binary sequences,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.
- [27] Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu, “Are all losses created equal: A neural collapse perspective,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 31697–31710, 2022.
- [28] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net.
- [29] Thomas G Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.
- [30] Giacomo Nannicini, “Performance of hybrid quantum-classical variational heuristics for combinatorial optimization,” *Physical Review E*, vol. 99, no. 1, pp. 013304, 2019.