The Speaker and Language Recognition Workshop (Odyssey 2024)
18-21 June 2024, Quebec City, Canada

1st Place Solution to Odyssey Emotion Recognition Challenge Task1: Tackling Class Imbalance Problem

Mingjie Chen[1], *Hezhao Zhang*[1], *Yuanchao Li*[2], *Jiachen Luo*[3], *Wen Wu*[4], *Ziyang Ma*[5], *Peter Bell*[2], *Catherine Lai*[2], *Joshua Reiss*[3], *Lin Wang*[3], *Philip C. Woodland*[4], *Xie Chen*[5], *Huy Phan*[6⋆], *Thomas Hain*[1]

[1] *Department of Computer Science, University of Sheffield, United Kingdom*
[2] *Centre for Speech Technology Research, University of Edinburgh, United Kingdom*
[3] *Centre for Digital Music, Queen Mary University of London, United Kingdom*
[4] *Department of Engineering, University of Cambridge, United Kingdom*
[5] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, China*
[6] *Amazon AGI, Cambridge, United Kingdom*

mingjie.chen@sheffield.ac.uk, t.hain@sheffield.ac.uk

Abstract

Speech emotion recognition is a challenging classification task with natural emotional speech, especially when the distribution of emotion types is imbalanced in the training and test data. In this case, it is more difficult for a model to learn to separate minority classes, resulting in those sometimes being ignored or frequently misclassified. Previous work has utilised class weighted loss for training, but problems remain as it sometimes causes over-fitting for minor classes or under-fitting for major classes. This paper presents the system developed by a multi-site team for the participation in the Odyssey 2024 Emotion Recognition Challenge Track-1. The challenge data has the aforementioned properties and therefore the presented systems aimed to tackle these issues, by introducing focal loss in optimisation when applying class weighted loss. Specifically, the focal loss is further weighted by prior-based class weights. Experimental results show that combining these two approaches brings better overall performance, by sacrificing performance on major classes. The system further employs a majority voting strategy to combine the outputs of an ensemble of 7 models. The models are trained independently, using different acoustic features and loss functions - with the aim to have different properties for different data. Hence these models show different performance preferences on major classes and minor classes. The ensemble system output obtained the best performance in the challenge, ranking top-1 among 68 submissions. It also outperformed all single models in our set. On the Odyssey 2024 Emotion Recognition Challenge Task-1 data the system obtained a Macro-F1 score of 35.69% and an accuracy of 37.32%.

1. Introduction

It is commonly assumed that for progress in conversational AI systems it is essential to enable computers to understand emotions from human speech signals. Speech emotion recognition (SER) is gaining increasing attention due to its wide range of potential application, especially in the context of the recent advancement of large language models [1]. SER has been a research focus for a long time, however it is still a complex task because of the multitude of factors that affect the task, including context information, speaking environments, the personality and the speaking style of speakers, language, cultural aspect, commonsense knowledge etc. [2, 3].

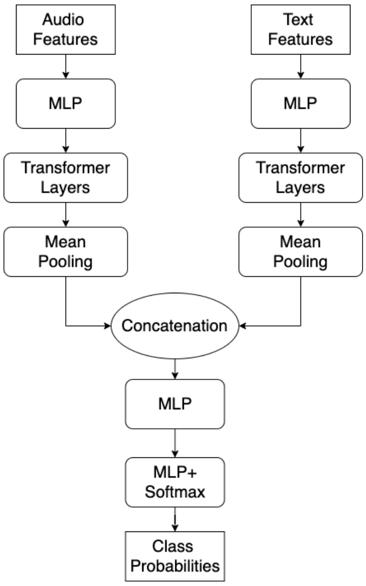

Figure 1: The architecture of each model in the ensemble system. 'MLP' denotes multi-layer perceptron.

Typically there are two types of SER task due to the annotation style used in emotion labelled datasets, namely classification and regression [3]. In SER classification tasks, speech segments are typically annotated with labels from a small set (4–8) of emotion classes. The task is to predict the correct (single) emotion class representing the complete speech segment.

Many datasets [4, 5, 6] have been created for SER. Most of them [6] are created by recording actors portraying the required emotion in their speech. Other type of dataset [4] have been created by prompting speakers to express specific emotions. There are a few datasets [5], usually referred to as natural datasets,

[⋆] The work does not relate to Huy Phan's position at Amazon.

10.21437/odyssey.2024-37

that are directly collected from sources containing spontaneous speech with natural emotional expressions. Previous work [7, 8] has shown that the performance of SER models on these three types of datasets differs significantly. SER classification tasks on natural datasets are still challenging, for many reasons [9].

One of the many challenges of SER classification on natural datasets is the imbalanced class distributions. Discriminative machine learning methods typically also choose decision boundaries on the basis of the prevalence of a class. Classes with low occurrence will not only get a poor representation, but also end up being considered of less relevance. Thus the training of machine learning models is difficult for minor classes. Models can be easily trained on major classes but can tend to ignore minor classes. One simple solution is to re-balance (reweight) the loss of a class by class frequencies. Many other solutions have been proposed to solve this problem, such as data augmentation [10, 11], new sampling strategies [12], and the use of a modified loss function [13]. However, all of the above methods are likely to cause over-fitting problems on minor classes, thus sacrificing performance on major classes. This issue is particularly pertinent here, as the SER system is designed for participation in the Odyssey 2024 Emotion Recognition Challenge Track-1, which has class imbalanced training data, but a class balanced test set.

We have therefore developed a system that makes use of an ensemble system of 7 models. Each model takes in multi-modal features, from both audio and text. In order to obtain text representations an automatic speech recognition (ASR) model [14] is used to generate transcriptions from speech segments. To enhance the quality of the ASR transcriptions, an error correction model is used in post-processing

In the ensemble system all models share the same architecture. They are trained independently with different audio features or different optimisation configurations. The loss function chosen is either the focal loss [15] or the cross entropy loss, weighted by prior-based class weights or uniform class weights.

The prior-based class weights are used to give more preference to minor classes than major classes during training. However, this strategy was found to cause over-fitting for minor classes, and thus a reduction in overall performance. In order to alleviate this issue, the use of focal loss [15] was included, which aims to give higher weights to more difficult samples and lower weights to easier samples.

The model trained with the focal loss together with the prior-based class weights obtained the best overall performance. However, this model performs more poor on major classes than the models trained with uniform class weights. To reach better overall performance the ensemble system is designed to comprise of models with different preferences on major classes and minor classes. Experimental results show that the ensemble system reaches the state-of-the-art performance in the Odyssey 2024 Emotion Recognition Challenge track-1 [16]. The system obtained a Macro-F1 of 35.69%, and has ranked the first among 68 submissions.

2. Related Works

2.1. Fusion Techniques for SER

Over the past decade, research on fusion techniques for SER has made significant progress. Alongside traditional feature-level fusion (i.e., early fusion) and decision-level fusion (i.e., late fusion), there has been widespread exploration of sophisticated tensor-level fusion methods. For instance, [17] combined both modality-invariant and modality-specific features and applied various regularisation functions to reduce the distance between the modalities. [18] proposed a weighted fusion method based on a cross-attention module for encoding inter-modality relations and selectively capturing effective information. [19] developed a dual-branch model, with one time-synchronous branch that combined speech and text modalities, and a time-asynchronous branch integrating sentence text embeddings from context utterances. [20] fused ASR hidden states and ASR transcriptions with audio features in a hierarchical manner.

2.2. ASR Error Correction

As outlined above, the system makes use of ASR for transcript generation, using an off-the-shelf system. To enhance ASR performance ASR Error Correction (AEC) methods can be helpful, by post-processing using some knowledge about the task or target domain. The standard method for addressing language domain mismatch is to train an in-domain language model for direct integration with ASR systems. [21, 22]. However, an alternative is to use AEC sequence to sequence models that correct the output. This is particularly useful in scenarios where the ASR is a black box [23]. More recently, there has been interest in employing generative error correction using large language models [24]. Furthermore, some studies have explored using both speech and ASR hypotheses as input, instead of relying solely on text data, leading to the development of cross-modal AEC methods [24].

3. System Description

As mentioned above, the system yielding the best performance makes use of an ensemble of models. Each model takes as input in frame-level audio features as well as token-level text features. The output of each model are the probabilities for each emotion class. In the majority voting, the prediction of each model is equally used for voting to emotion classes. The most voted emotion class is going to be the final prediction of the ensemble system.

To extract text features, this work used transcriptions generated by the Whisper-large-v2 model [14]. As there are no transcriptions available for the test set, an ASR system is needed for transcriptions, but it will inevitably produce erroneous transcripts. Training an emotion classification model on ground truth transcripts would result in a mismatch between training and test conditions, thus all models make use of ASR output. To enhance the quality of the ASR transcriptions, an ASR error correction model [25] is trained and then used to correct errors in the transcriptions in the test set.

3.1. Model Architecture

Figure 1 illustrates the principal architecture of models. Frame-level audio features and token-level text features are encoded by two Multi-Layer Perceptron (MLP) modules. Then transformer layers [26] are used to process audio and text features to encode dynamic information in features. In order to avoid over-fitting, the number of heads in the transformer layers is set to 1. The transformer layers are followed by a mean pooling layer, then the utterance-level audio features and text features are concatenated. The concatenated features are processed by the two MLP modules. The softmax output of the final MLP produces class probabilities.

Features	Modality	Dim	#Params	Hours
WavLM-large	Audio	1024	300M	94K
Hubert-extra-large	Audio	1280	1B	60K
Whisper-large-v3	Audio	1280	1.5B	680K
Roberta-large	text	1024	355M	-

Table 1: Audio and text features, where ‘Dim’ denotes the dimensionality of frame-level features, ‘#Params’ denotes the number of parameters, ‘Hours’ denotes the amount of speech data used for pretraining.

Index	Loss Function	Class Weights	Audio Features
1	Focal ($\gamma = 2$)	Prior-based	Whisper
2	Focal ($\gamma = 2.5$)	Prior-based	Whisper
3	CE	Prior-based	Whisper
4	Focal ($\gamma = 2$)	Uniform	Whisper
5	CE	Uniform	Whisper
6	Focal ($\gamma = 2$)	Prior-based	WavLM
7	Focal ($\gamma = 3$)	Prior-based	Hubert

Table 2: Configurations of models in the ensemble systems, where ‘CE’ denotes cross-entropy.

3.2. Audio Features and Text Features

The audio and text features used in this work are presented in Table 1. Three types of audio features are used: the final layer representations of WavLM-large [27]; the final layer representations of Hubert-extra-large [28]; and the final layer output of the encoder of Whisper-large-v3 [14]. In terms of text features, this work utilises Roberta-large [29]. In order to enhance the representative capability of the text features, this work uses the average representations of the last 4 layers’ from Roberta-large.

3.3. Loss Functions and Class Weights

This work considers two options for loss functions, the focal loss and the cross entropy loss. The loss function can be weighted by the prior-based class weights or uniform class weights. Combining loss functions with class weights, four types of optimisation configurations can be used.

3.3.1. Uniform Class Weights and Prior-based Class weights

Consider a classification task with C classes, a training dataset $\{(\mathbf{x}_i, y_i)|i = 1, ..., N\}$, $y_i \in \{1, ..., C\}$, $\mathbf{x}_i \in \mathrm{R}^d$. The uniform class weights can be written as follows,

$$\mathbf{w}_{\text{uni}} = \{w_j = 1|j = 1, ..., C\}.$$

Prior-based class weights can be written as:

$$\mathbf{w}_{\text{prior}} = \{w_j = \frac{N}{N_j}|j = 1, ..., C\},$$

where N_j is the number of samples in class j in a training or development set.

3.3.2. Cross-Entropy Loss and Focal Loss

Following the above notations , the cross-entropy loss can be defined as:

$$\mathcal{L}_{\text{ce}} = 1/N \sum_{i=1}^{N} -\log p_i,$$

where $p_i = P(y_i|\mathbf{x}_i)$ is the output probability of the corrected class. The focal loss can be written as follows,

$$\mathcal{L}_{\text{focal}} = 1/N \sum_{i=1}^{N} -(1 - p_i)^{\gamma} \log p_i,$$

where γ is a hyper-parameter.

By combining with the class weights, the class-weighted focal loss can be written as:

$$\mathcal{L}'_{\text{focal}} = 1/N \sum_{i=1}^{N} -w_j (1 - p_i)^{\gamma} \log p_i,$$

where w_j is the class weight for class y_i.

3.4. Ensemble Strategy

Table 2 lists the configurations of the 7 models in the ensemble system. All models use the averaged representations of the last 4 layers of Roberta-large (refer to Table 1) as text features. Apart from the optimisation configurations, the values of the focal loss hyper-parameter was explored with $\gamma = 2$ and $\gamma = 2.5$, as well as three types of audio features.

4. Experimental Setup

4.1. Dataset

The Odyssey 2024 Emotion Recognition Challenge [16] used the MSP-Podcast dataset v1.11 [5]. The dataset is derived from podcasts, and annotated through crowd-sourcing. Different from most datasets containing acted speech [6] or elicited speech [4], this dataset contains spontaneous speech with natural human emotions. The dataset is composed of five subsets: the training set, the development set, the test-1 set, the test-2 set, and the test-3 set. In this challenge, the test-3 set is used to measure the outcome, and the reference labels have not been made public. The speech segments have been annotated with 10 classes, in which 8 classes are used for this challenge. These are Neutral, Happy, Angry, Disgust, Sad, Surprise, Contempt and Fear. There are two remaining labelled classes, Other and No Agreement, are not used. Since the Other and No Agreement classes are not used in the challenge, this work removed associated samples, retaining only samples with the challenge classes. Detailed information of the training set, development set and test set are given in Table 3, including the subset statistics, before and after the sample removal.

4.2. Implementations

4.2.1. Multi-Modal Classifier Model

The Hubert-extra-large model, the WavLM-large model, the Whisper-large-v3 model and the Roberta-large model are imported from the transformers toolkit [30].

In terms of the model architecture, the hidden size of the transformer layers is 512 and the number of the transformer layers is set to 2. The MLP module before the transformer layers has a hidden size of 512. The MLP modules after the concatenation layer has a hidden size of 512 and the output size of the final MLP module is 8.

The models are trained with a batch size of 128, an initial learning rate of 1e-4, with a learning scheduler [31]. The model checkpoint of the epoch with the best Macro-F1 on the development set is chosen for evaluation. The implementations are

Subset	# Samples All	# Speakers All	Duration All	# Samples Used	# Speakers Used	Duration Used
Training	68119	1405	110.2h	53386	1391	86.3h
Dev	19815	454	31.7h	15341	446	24.4h
Test-3	2437	187	3.9h	-	-	-

Table 3: The detailed information of subsets of the MSP-Podcast v1.11, where 'All' denotes subset statistics before the filtering and 'Used' denotes subset statistics after the filtering.

Class	# Training Samples	# Dev Samples
Neutral	25016	5667
Happy	13440	3340
Angry	3053	2413
Sad	3882	1101
Disgust	1426	486
Contempt	2443	1323
Surprise	2897	729
Fear	1139	282

Table 4: Class distributions in the training set and the development set after the filtering.

based on the recipe of the speechbrain toolkit [32]. All models are implemented and trained with the PyTorch toolkit [33]. Training takes about 2 hours and uses a maximum of 20 epochs.

4.2.2. Feature Fusion

A number of methods for feature fusion were considered. Due to limited time, however, only the following ones were implemented and compared.

- *Early fusion*: text and audio features are concatenated at the embedding level.
- *Late fusion*: text and audio features are learned independently and the final decision is determined based on respective outputs.
- *Early fusion + late fusion*.
- *Tensor fusion*: unimodal information and bimodal interactions are learned explicitly and aggregated [34].
- *Low-rank tensor fusion*: multimodal fusion with modality-specific low-rank factors, which scale linearly in the number of modalities [35].

While sophisticated fusion approaches have often outperformed early fusion in various scenarios, this phenomenon was not observed in the experiments. Therefore, early fusion was used in the model, as it was found to outperform all other methods tested.

4.2.3. ASR Error Correction

A pretrained AEC model was used in this work, which has been trained on the English version of Common Voice 13.0 [36] and TED Talk corpus [37] using a publicly available Sequence-to-Sequence (S2S) encoder-decoder architecture [25]. This model was trained to convert ASR transcriptions to human-transcribed transcriptions (i.e., ground-truth text). Considering the disparity between the MSP-Podcast dataset and the two AEC pretraining corpora (e.g., out-of-domain words), this model was fine-tuned to enhance its performance. Specifically, the model was trained on the training set of the provided MSP-Podcast corpus for 10 epochs and then validated on the development set. The best checkpoint was saved to correct errors on the test set. The correction quality was evaluated using WER, BLEU, and GLEU scores for a comprehensive assessment. The results of the best checkpoint on the development set are presented in Table 5, demonstrating the effectiveness of the AEC model. As there is no ground-truth text for the test set, there is no way to further evaluate its effectiveness. Given the same domain of the development and test sets, it is expected an improvement in accuracy on the test set by approximately 1%, as suggested by previous research on the impact of WER on SER performance [38].

Transcription	WER ↓	BLEU ↑	GLEU ↑
Original	17.65	81.32	78.02
Corrected	14.51	83.48	81.19

Table 5: Comparison of the quality between the original and corrected transcriptions on the development set. All values are presented in percentage scale (%). ↓: the lower the better. ↑: the higher the better.

4.3. Evaluation Metrics

According to the challenge evaluation setup, Macro-F1 is used as the primary metric. Macro-F1 is the unweighted average of the F1 score of each class. Apart from Macro-F1, the weighted accuracy (WA) and unweighted accuracy (UA) are used. [1]

5. Results

Results for the 7 models in the ensemble system are presented in Table 6. Generally speaking, different configurations of audio features, loss functions and class weights yield differences in WA and UA results. From analysing the performance difference in WA and UA results, it would be easy to understand how the models perform on the major classes and minor classes. For convenience, in the following discussion, the index of the models in Table 6 are used to denote the models (e.g. model-1).

Among the 7 models, the best Macro-F1 was obtained by the model-1, applying the focal loss ($\gamma = 2$) and prior-based class weights. Comparing model-1 to model-4, increasing $\gamma = 2$ causes an improvement in WA but a drop in UA, causing a reduction in Macro-F1.

Among the 3 types of audio features for model-1, model-2 and model-3, it is clear that the Whisper features yield better results than both the WavLM features and the Hubert features. One possible reason is that the Whisper model were trained supervisedly with text transcriptions, while the other two models were trained without supervision. These are also difference due to the model sizes and the amount of training data used bay these pretrained models.

[1] The accuracy_score function and the balanced_accuracy_score function from the scikit-learn toolkit are used for implementing unweighted accuracy and weighted accuracy, respectively

Index	Audio	Text	Loss	Class Weights	Macro-F1 (%)	WA (%)	UA (%)
1	Whisper	Roberta	Focal ($\gamma = 2$)	Prior-based	**34.2**	35.6	45.7
2	WavLM	Roberta	Focal ($\gamma = 2$)	Prior-based	32.4	35.3	43.6
3	Hubert	Roberta	Focal ($\gamma = 2$)	Prior-based	32.7	34.3	45.5
4	Whisper	Roberta	Focal ($\gamma = 2.5$)	Prior-based	33.8	35.8	45.1
5	Whisper	Roberta	CE	Prior-based	33.8	**36.0**	44.0
6	Whisper	Roberta	Focal ($\gamma = 2$)	Uniform	33.3	32.9	**51.9**
7	Whisper	Roberta	CE	Uniform	32.8	32.6	51.1
Ensemble	-	-	-	-	**35.6**	**36.6**	49.3

Table 6: Development set results of the 7 models, they are trained independently with different audio features, text features, loss and class weights.

The results for model-1, model-5, model-6 and model-7 can help to understand the effect of the two loss functions and the two class weights. Generally speaking, different loss functions and class weights yields models that have different preferences for major and minor classes. Specifically, when training with the uniform class weights, model-6 and the model-7 show good performance on UA but poor performance on WA. Comparing model-7 and model-5, the prior-based class weights give more attention to minor classes, causing a significant improvement in WA but a large drop in UA. This means that the prior-based class weights improve the performance on the minor classes, but sacrificing the performance on the major classes. The performance drop may be due to model-5 over-fitting on major classes. The effect of focal loss can be found through comparing model-1 and model-5, which shows that focal loss helps model-1 reach a better balance between the major classes and the minor classes. Hence, the model-1 reaches the highest overall macro-F1 performance.

Based on the diverse performance of the models, an obvious strategy to build an ensemble is to combine the outputs of these models through a majority voting process. The results show that the ensemble system outperforms all of the 7 models on Macro-F1 and WA, reaching the state-of-the-art performance.

6. Conclusions

This paper introduces an ensemble system that includes 7 multimodal models, constructed for participation in the Odyssey 2024 Emotion Recognition Challenge. The system showed the best performance among a total of 68 submissions to the challenge, in all metrics under consideration. The 7 models were trained independently with different loss functions and class weights. Specifically, the cross entropy loss and the focal loss were used. Uniform class weights and prior-based class weights are studied. The experiment results show that the combinations of loss functions and class weights lead to different preference on the major classes or the minor classes

7. Acknowledgements

This work was primarily conducted at the Voicebase/Liveperson Centre of Speech and Language Technology at the University of Sheffield which is supported by Liveperson, Inc..

8. References

[1] OpenAI, “Chatgpt,” 2022.

[2] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L Schneider, “Speech emotion recognition using machine learning—a systematic review,” *Intelligent systems with applications*, p. 200266, 2023.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[5] 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, 2017.

[6] Sanaul Haq and Philip JB Jackson, “Multimodal emotion recognition,” in *Machine audition: principles, algorithms and systems*, pp. 398–423. IGI global, 2011.

[7] Rosanna Milner, Md Asif Jalal, Raymond WM Ng, and Thomas Hain, “A cross-corpus study on speech emotion recognition,” in *Proc. of ASRU 2019*. IEEE, 2019, pp. 304–311.

[8] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen, “Emotion2vec: Self-supervised pre-training for speech emotion representation,” *arXiv preprint arXiv:2312.15185*, 2023.

[9] Lucía Gómez-Zaragozá, Óscar Valls, Rocío del Amor, María José Castro-Bleda, Valery Naranjo, Mariano Alcañiz Raya, and Javier Marín-Morales, “Speech emotion recognition from voice messages recorded in the wild,” *arXiv preprint arXiv:2403.02167*, 2024.

[10] Tao Meng, Yuntao Shou, Wei Ai, Nan Yin, and Keqin Li, “Deep imbalanced learning for multimodal emotion recognition in conversations,” *arXiv preprint arXiv:2312.06337*, 2023.

[11] Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, and Xie Chen, “Leveraging speech

PTM, text LLM, and emotional TTS for speech emotion recognition," *Proc. ICASSP*, 2024.

[12] Hemin Ibrahim, Chu Kiong Loo, and Fady Alnajjar, "Bidirectional parallel echo state network for speech emotion recognition," *Neural Computing and Applications*, vol. 34, no. 20, pp. 17581–17599, 2022.

[13] Sungrack Yun and Chang D Yoo, "Speech emotion recognition via a max-margin framework incorporating a loss function based on the watson and tellegen's emotion model," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4169–4172.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. of ICML*. PMLR, 2023, pp. 28492–28518.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proc. of ICCV*, 2017, pp. 2980–2988.

[16] L. Goncalves, A. N. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, L.P. 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.

[17] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[18] Jiachen Luo, Huy Phan, Lin Wang, and Joshua Reiss, "Mutual cross-attention in dyadic fusion networks for audio-video emotion recognition," in *Proc. of ACIIW 2023*. IEEE, 2023, pp. 1–7.

[19] Wen Wu, Chao Zhang, and Philip C Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *Proc. of ICASSP 2021*. IEEE, 2021, pp. 6269–6273.

[20] Yuanchao Li, Peter Bell, and Catherine Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in *Proc. of ICASSP 2022*. IEEE, 2022, pp. 7362–7366.

[21] Tomohiro Tanaka, Ryo Masumura, Hirokazu Masataki, and Yushi Aono, "Neural error corrective language models for automatic speech recognition.," in *INTERSPEECH*, 2018, pp. 401–405.

[22] Hirofumi Inaguma, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, "Improving OOV detection and resolution with external language models in acoustic-to-word ASR," in *Proc. of SLT 2018*. IEEE, 2018, pp. 212–218.

[23] Junwei Liao, Sefik Eskimez, Liyang Lu, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng, "Improving readability for automatic speech recognition transcription," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1–23, 2023.

[24] Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, Rohit Kumar, Narsis Kiani, David Gomez-Cabrero, and Jesper Tegnér, "Whispering LLaMA: A cross-modal generative error correction framework for speech recognition," in *Proc. of EMNLP 2023*, 2023, pp. 10007–10016.

[25] Pinzhen Chen and Gerasimos Lampouras, "Exploring data augmentation for code generation tasks," in *Proc. of EACL 2023 Findings*, 2023, pp. 1542–1550.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Proc. of NIPS 2017*, vol. 30, 2017.

[27] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong 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.

[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[30] HuggingFace, "Transformers," 2016.

[31] Hang Su and Haoyu Chen, "Experiments on parallel training of deep neural network using model averaging," *arXiv preprint arXiv:1507.01239*, 2015.

[32] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[34] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. of EMNLP 2017*, 2017, pp. 1103–1114.

[35] Zhun Liu and Ying Shen, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. of ACL 2018*, 2018.

[36] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[37] Mauro Cettolo, Christian Girardi, and Marcello Federico, "Wit3: Web inventory of transcribed and translated talks," in *Proc. of EAMT 2012*, 2012, pp. 261–268.

[38] Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai, "ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition," in *Interspeech 2023*, 2023.