# EMO-SUPERB: An In-depth Look at Speech Emotion Recognition

**Anonymous ACL submission**

## Abstract

Speech emotion recognition (SER) is a pivotal technology for human-computer interaction systems. However, 80.77% of SER papers yield results that cannot be reproduced (Antoniou et al., 2023). We develop EMO-SUPERB, shorted for **EMO**tion **S**peech Universal **PER**formance **B**enchmark, aims at enhancing open-source initiatives for SER. EMO-SUPERB includes a user-friendly codebase to leverage 15 state-of-the-art speech self-supervised learning models (SSLMs) for exhaustive evaluation across six open-source SER datasets. EMO-SUPERB streamlines result sharing via an online leaderboard, fostering collaboration within a community-driven benchmark and thereby enhancing the development of SER. On average, 2.58% annotations are annotated using natural language. SER relies on classification models and is unable to process natural languages, leading to the discarding of these valuable annotations. We prompt Chat-GPT to mimic annotators, comprehend natural language annotations, and subsequently re-label the data. By utilizing labels generated by ChatGPT, we consistently achieve an average relative gain of 3.08% across all settings. We make all resources open-source to facilitate future developments in SER. The source code and complete analysis are on the project website [1]

## 1 Introduction

Speech Emotion Recognition (SER) aims to discern emotional cues from speech inputs, representing a pivotal technology for human-computer interaction systems. Recent years have witnessed significant advancements in SER. However, there are some unsolved problems in the SER domain:

**Issue 1**: Devoted annotators prefer using natural language rather than traditional emotion labels when annotating data, resulting in **typed descriptions** (e.g., "Slightly Angry, calm" to notify the intensity of emotion). While these descriptions are highly valuable, SER models, designed as classification models, cannot process natural languages and thus discard them. Notably, approximately 2.58% (on average) of the annotations across all datasets use typed descriptions.

**Issue 2**: The author of SAIL-IEMOCAP (Busso et al., 2008), the most renowned SER dataset, has demonstrated that over 80.77% of SER papers produce results that cannot be reproduced (Antoniou et al., 2023) due to the absence of released codes.

**Issue 3**: Official data partitioning guidelines are lacking in most SER datasets. Consequently, different papers adopt varying partitioning strategies, leading to potential **data leakage** problems: Typically, SER datasets comprise dialogues between two participants, denoted as Speaker A and Speaker B. In the process of segmenting these dialogues to isolate individual utterances, it is common to encounter scenarios where Speaker A's segments contain speech from Speaker B. This can cause issues because many studies adopt a straightforward approach to dividing the dataset. They possibly allocate utterances from Speaker A for training and those from Speaker B for testing. However, this approach inadvertently exposes the model to Speaker B's speech during training, leading to **data leakage**. Studies employing this cheating partition role, with data leakage, tend to achieve 4.011% performance improvements than those without it (Antoniou et al., 2023). However, comparing settings with data leakage to those without it is unfair.

We introduce EMO-SUPERB to advance open-source initiatives in SER. Then, we detail how to address the above three issues individually.

- For **Issue 1**, we employ ChatGPT to mimic annotators, comprehend typed descriptions, and re-label the data accordingly, as shown in Section 2. With labels generated by Chat-GPT, we consistently achieve relatively 3.08%
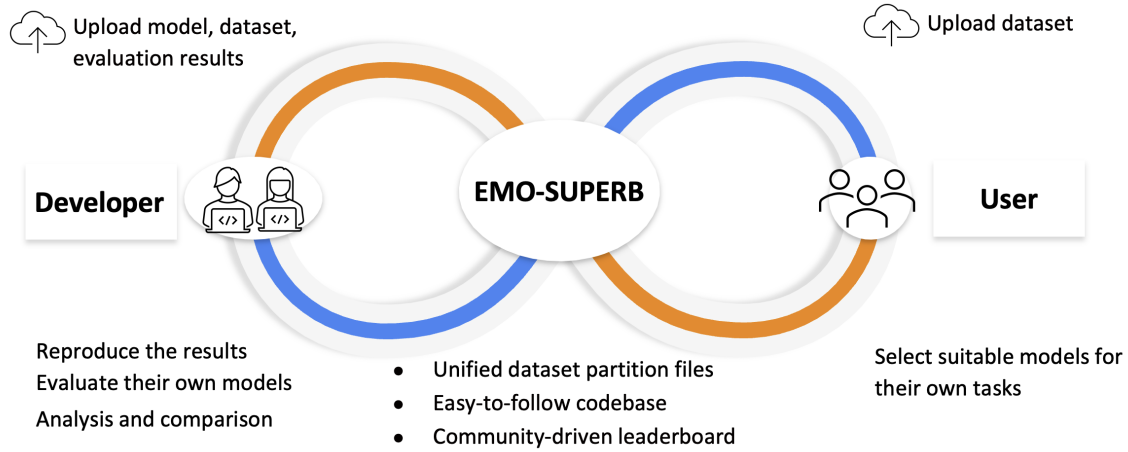
---

Figure 1: Demonstration for the EMO-SUPERB platform: Developers design and evaluate SER models using our standardized dataset partition files and evaluation criteria. Developers then contribute these prediction results to the online leaderboard, enriching the benchmark database and enabling comparative analyses with other SER models. Finally, developers harness the visualization and statistical tools on the website to compare performance, gathering invaluable insights for future works. From the user's standpoint, they can upload datasets and select appropriate models tailored to their individual applications.

performance improvements across all settings.

- For **Issue 2**, we develop a codebase to harness 15 SSLMs, renowned for enhancing state-of-the-art performance in speech emotion recognition, for exhaustive evaluation across all open-source SER datasets in Section 3.2. Developers can utilize a single command line to execute both training and evaluation processes seamlessly, and we will release the easy-to-follow codebase.

- For **Issue 3**, we partition six open-source SER datasets and address potential data leakage issues during the partitioning process, as shown in Section 3.3.

Finally, we make all datasets labeled with Chat-GPT, data partition files, codes, and checkpoints open source to the community.

## 2 Empower by ChatGPT

### 2.1 Importance of typed descriptions

Emotion datasets (Busso et al., 2008; Chou et al., 2017; Lotfian and Busso, 2017) allow annotators to employ natural language to describe their perception of emotion corresponding to the given data if the provided label options are insufficient to capture their emotional perception fully. These descriptions, articulated in natural language, are called typed descriptions. Typed descriptions like

"Slightly Angry, calm" serve to indicate the intensity of emotion, while "Haaapy" is used to emphasize happiness. Appendix D shows more examples of the typed descriptions.

Although typed descriptions account for only approximately 2.58% of annotations across the four emotion databases that include them, they contain valuable information for emotion perception (Lotfian and Busso, 2019; Chou et al., 2022). Because typing down the natural language to describe the emotion perception takes more time than choosing labels from label options. Only motivated annotators will use typed descriptions. In fact, annotators are compensated based on their working hours rather than the volume of data they handle (Lotfian and Busso, 2019). Moreover, exemplary annotators receive additional bonuses. Some annotators invest extra time to ensure a comprehensive description of emotions to get more bonuses. However, SER, primarily based on classification models, cannot process natural language and consequently overlooks these valuable typed descriptions.

### 2.2 Why using ChatGPT for relabeling

ChatGPT (Achiam et al., 2023) exhibits a remarkable ability to comprehend and analyze natural language. Kheiri and Karimi (2023) had used Chat-GPT to do effective sentiment analysis. Hence, we utilize ChatGPT to mimic annotators, summarizing their thoughts to re-label the data with typed
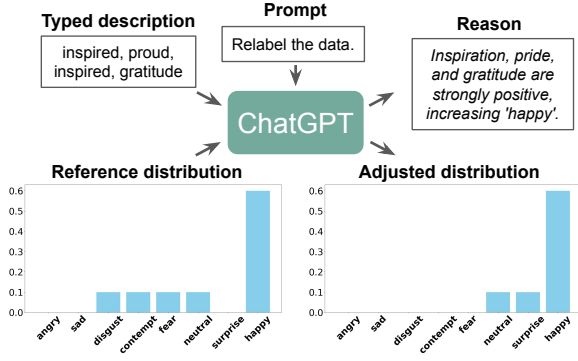
Figure 2: Labeling process using ChatGPT. Three inputs are **Typed description**, **Reference distribution**, and **Prompt**. Two outputs are **Reason** and **Adjusted distribution**. Notice that the reference distribution is calculated by the number of votes for emotion classes. In the raw annotations of an example, there are instances of disgust, contempt, fear, neutrality, and happiness (*6), resulting in values of 0.6 for happiness and 0.1 for each of the remaining appearing emotions.

descriptions. While GPT models have been previously utilized for data labeling tasks, our approach stands out due to its innovative application in generating a distribution of labels instead of assigning a single label. We show that this approach leads to consistent improvements across all experimental settings as shown in Table 3.

## 3 EMO-SUPERB platform

### 3.1 Prompt ChatGPT

We design a carefully crafted prompt to transform the released version of GPT-4 Turbo, a variant of ChatGPT, into a knowledgeable assistant psychologist. Its primary function is to generate a distribution across emotion labels based on the input typed descriptions from annotators.

As shown in Figure 2, three inputs are provided to ChatGPT: the typed descriptions, reference distributions, and a well-designed prompt. When we prompt ChatGPT to refer to the distribution label, it fails to provide the distribution unless we supply the reference distribution. The format of the output emotion label is also a distribution. Guided by the prompt, the ChatGPT can adjust or maintain the reference distribution based on the typed descriptions. In the prompt, we also let ChatGPT explain why it changes or doesn't change the reference distributions. Without this, ChatGPT might default to laziness, consistently avoiding modifying the reference distributions. For detailed information and the final prompt, please refer to Table 11 in

Appendix D.1 due to space limitations.

We choose the MSP-PODCAST (P) dataset to verify the efficacy of our proposed prompt method in utilizing typed descriptions to improve SER, as it is the largest dataset and has the highest percentage (6.08%) of typed descriptions among all other datasets. Figure 5 and the Appendix D.1.3 show the label distributions between original and re-label ones. ChatGPT can understand the typed distribution and output reasonable distributions.

Ultimately, we achieved an average performance improvement of approximately 3.08% across the 16 models on the MSP-PODCAST (P) as shown in Table 3 of Section 5.2. Designing an effective prompt can enhance the accuracy of the re-labeling process. This paper opens the door to utilizing large language models for comprehending typed descriptions. We welcome the community to use our user-friendly codebase to evaluate their datasets relabeled by large language models.

As shown in Figure 1, our platform is designed to empower developers with seamless access to replicate our results, evaluate their custom SER models, compare model characteristics, and foster future SER development. This is facilitated by integrating three essential components: an easy-to-follow codebase, unified dataset partition files, and a community-driven leaderboard website. Users can select SER models for their own usage.

### 3.2 SSLM-based codebase

#### 3.2.1 Framework

Self-supervised learning (SSL) is a promising direction for developing speech models. This approach entails training a large model with large-scale unlabeled data to obtain robust and general representations. Notably, after pre-training, one can achieve nearly SOTA performance on downstream tasks by employing the fixed SSLMs alongside task-specific lightweight prediction heads (Yang et al., 2021; Feng et al., 2023). Furthermore, SSLMs significantly enhance SER and demonstrate SOTA performance, as evidenced in (Wagner et al., 2023).

Inspired by achievements evidenced by (Yang et al., 2021; Tsai et al., 2022; Feng et al., 2023), and in an effort to enhance open-source initiatives, we develop a comprehensive codebase. This codebase leverages 15 speech-supervised learning models as feature extractors and trains lightweight heads for exhaustive evaluation across 6 open-source SER datasets with 9 common settings, as shown
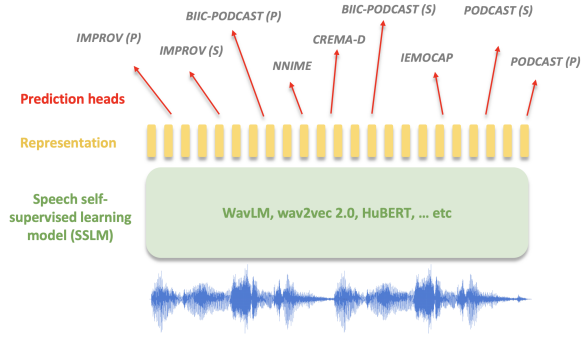
Figure 3: Illustration of SSLM-based SER

in Figure 3. The six datasets adopted are SAIL-IEMOCAP, CREMA-D (Cao et al., 2014), MSP-IMPROV (Busso et al., 2017), MSP-PODCAST, BIIC-NNIME (Chou et al., 2017), and BIIC-PODCAST (Upadhyay et al., 2023).

### 3.2.2 Self-supervised learning models

In our codebase, we include two mainstream categories of SOTA SSLMs, pre-trained using generative losses and discriminative losses. We summarize them in Table 1 and details can be found in Appendix B due to space limitation.

### 3.2.3 Pros of the codebase

The codebase has the following merits:

- **High-performance**: Our choice to utilize SSLMs is based on their ability to consistently achieve SOTA results in speech emotion recognition, aligning with our goal to boost open-source efforts in this domain.

- **Affordability**: The computing barrier is greatly diminished by leveraging pre-trained SSLMs and solely fine-tuning a lightweight head, enhancing affordability for researchers from diverse backgrounds.

| Model | Loss |
|---|---|
| Autoregressive Predictive Coding (APC) (Chung et al., 2019) | Generative loss |
| VQ-APC (Chung et al., 2020) | Generative loss |
| Non-autoregressive Predictive Coding (NPC) (Liu et al., 2020a) | Generative loss |
| Mockingjay (Liu et al., 2020b)) | Generative loss |
| TERA (Liu et al., 2021) | Generative loss |
| DeCoAR 2 (Ling and Liu, 2020) | Generative loss |
| WavLM (Chen et al., 2022) | Discriminative loss |
| Hubert (Hsu et al., 2021) | Discriminative loss |
| wav2vec 2.0 (**W2V2**) (Baevski et al., 2020) | Discriminative loss |
| Data2Vec (Baevski et al., 2022) | Discriminative loss |
| XLS-R (Babu et al., 2021) | Discriminative loss |
| VQ wav2vec (**VQ-W2V**) (Baevski et al., 2019) (Ling and Liu, 2020) | Discriminative loss |
| wav2vec (**W2V**) (Schneider et al., 2019) | Discriminative loss |
| Contrastive Predictive Coding (CPC) (**M CPC**)(Oord et al., 2018)) | Discriminative loss |

Table 1: Summary of SSLMs

- **Reproducibility**: All codes, data partition files, and checkpoints are released, ensuring easy reproducibility of results.

- **Easy-to-follow**: Developers can employ a single command line to execute all training and evaluation processes, making it exceptionally user-friendly.

### 3.3 Unified dataset partition rules

Typically, emotion databases are collected from dialogues (Busso et al., 2008, 2017; Chou et al., 2017; Lotfian and Busso, 2017). These dialogues often involve multiple speakers engaging in intensive turn-taking, overlap, and interruption. The segmented utterances for each speaker commonly include speech from their conversation partners.

For example, consider the SAIL-IEMOCAP corpus, which comprises 5 dyadic interactions (dialogues between two speakers) involving a total of ten speakers. In 50% of previous studies, researchers randomly divide the recordings of these ten speakers into train and test sets (Antoniou et al., 2023). However, due to overlap often present across speaker's segments, this practice can lead to data leakage because speaker B's speech has already been used for the model training, mentioned in section 1 (**Issue 3**).

In this study, we establish partition rules that adhere to speaker-independent criteria to mitigate the risk of leakage. Specifically, we ensure that all utterances from both speakers involved in dialogues are assigned to either the training or testing set. Further details regarding partitioning the six emotion databases can be found in Appendix C. We provide the **standardization** of the training and testing splits and setups across the six public SER datasets.

### 3.4 Community-driven leaderboard

The leaderboard website holds significant importance within EMO-SUPERB, continuously expanding and welcoming submissions worldwide, evolving it into a dynamic benchmark that goes beyond showcasing our own evaluation results. To mitigate the participation barrier, the website accepts submissions with participants' own models, especially when migrating their codes to the codebase in Section 3.2 is not straightforward. Participants simply need to adhere to the data partition files outlined in Section 3.3, evaluate their trained models, and submit the results. The website also offers

4

useful visualization (e.g. radar chart Figure 9 in Appendix E) and statistical tools for comparing detailed characteristics of different models, thereby enhancing future model development.

Additionally, our platform encourages community contributions of prompts and datasets with newly re-labeled typed descriptions. Submitters can conveniently evaluate the quality of their labeled datasets using a single command line on our codebase introduced in Section 3.2.

### 3.5 Artifacts

Modern deep learning models present a reproducibility challenge, even with released codes, due to the potential impacts of minor hyperparameter change or package version disparities on performance. To assist users in debugging their training procedures, we offer Tensorboard files, hyperparameters, and pre-trained weights in our codebase. Furthermore, we provide downstream prediction files for several state-of-the-art models, enabling users to visualize and analyze results easily.

## 4 Experimental setup

### 4.1 Datasets

We include the six public emotion datasets in the work. Some datasets use both primary emotions (denoted as (**P**)) and secondary emotions (marked as (**S**)) to allow annotators to choose single and multiple emotions, respectively. The Appendix A presents detailed information, and Table 5 summarizes statistical data regarding the six emotion databases. Appendix A.1 outlines the license terms and usage issues. We provide details of partitions in Appendix C to avoid issue 2 in Section 1, data leakage. The key information about these datasets is summarized as follows.

#### 4.1.1 The IEMOCAP

The SAIL-IEMOCAP (Busso et al., 2008), referred to as **IEMOCAP**, collects motion capture, audio, and video recordings from five dyadic conversations acted by ten professional actors in English. The recorded sessions were manually segmented into 10,039 utterances. The emotional annotations contain ten emotions and typed descriptions.

#### 4.1.2 The CREMA-D

The CREMA-D (Cao et al., 2014) contains high-quality audio-visual clips from 91 professional actors. There are 43 female and 48 male actors. There are 7,442 clips in English annotated via a crowd-sourcing platform. The process of perceptual annotations has three scenarios: voice-only, face-only, and audio-visual. In this work, we only use voice-only emotional annotations.

#### 4.1.3 The IMPROV

The MSP-IMPROV (Busso et al., 2017), referred to as **IMPROV**, consists of high-quality audio-video sessions acted by 12 actors in English. All sessions are manually segmented into 8,438 clips. The annotation process has two scenarios: primary (**P**) and secondary (**S**) emotions. The corpus collected the typed descriptions.

#### 4.1.4 The POD

The MSP-PODCAST (Lotfian and Busso, 2019), referred to as **POD**, collected spontaneous and diverse emotional speech from various real-world podcast recordings with a commercial license. The labeling setting also contains primary and secondary scenarios. The major difference is the number of emotions in the given options. We use the release version 1.11 of the database, including 84,030 utterances in the train set, 19,815 in the development set, 30,647 in the test1 set, and 14,815 in the test2 set. We combine the test1 and test2 as the test set.

#### 4.1.5 The NNIME

The BIIC-NNIME (Chou et al., 2017), referred to as **NNIME**, consists of video, audio, and physiology recordings of dyadic conversations acted by 43 actors in Mandarin Chinese. All sessions are manually segmented into 5,596 clips. We exclude turns annotated by "other" from all annotators or by less than three annotators. The corpus also collects typed descriptions in Chinese.

#### 4.1.6 The B-POD

The BIIC-PODCAST (Upadhyay et al., 2023), referred to as **B-POD**, is a variant of MSP-PODCAST in Mandarin Chinese. We use the release version 1.01. There are 48,815 utterances in the train set, 10,845 in the development set, and 10,340 in the test set. At least five annotators annotate each utterance, and the emotional annotators contain primary emotions (**P**) and secondary emotions (**S**), which is the same as MSP-PODCAST.

Table 2: The table summarizes the overall performance of SSLMs across the 6 public emotion datasets. **#Par.(M)** means the number of the SSLM parameters (frozen).

| SSLM | #Par. (M) | Average | IMPROV (P) | CREMA-D | POD (P) | B-POD (P) | IEMOCAP | NNIME | IMPROV (S) | POD (S) | B-POD (S) |
|------|-----------|---------|------------|---------|---------|-----------|---------|-------|------------|---------|-----------|
| **XLS-R-1B** | 965 | **0.38352** | 0.552 | **0.676** | 0.331 | **0.266** | 0.329 | 0.209 | 0.422 | **0.384** | **0.283** |
| **WavLM** | 317 | 0.38334 | **0.559** | 0.673 | **0.350** | 0.252 | 0.336 | **0.209** | 0.430 | 0.369 | 0.272 |
| **Hubert** | 317 | 0.38331 | 0.553 | 0.675 | 0.342 | 0.262 | 0.337 | 0.197 | 0.427 | 0.383 | 0.274 |
| **W2V2 R** | 317 | 0.37874 | 0.555 | 0.672 | 0.331 | 0.251 | **0.339** | 0.196 | **0.433** | 0.363 | 0.269 |
| **Data2Vec-A** | 313 | 0.37334 | 0.536 | 0.659 | 0.329 | 0.254 | 0.331 | 0.188 | 0.414 | 0.378 | 0.270 |
| **DeCoAR 2** | 90 | 0.36229 | 0.512 | 0.646 | 0.308 | 0.256 | 0.320 | 0.187 | 0.405 | 0.353 | 0.274 |
| **W2V2** | 317 | 0.35851 | 0.469 | 0.669 | 0.321 | 0.255 | 0.306 | 0.178 | 0.396 | 0.353 | 0.281 |
| **APC** | 4 | 0.34975 | 0.497 | 0.608 | 0.298 | 0.249 | 0.316 | 0.186 | 0.389 | 0.340 | 0.266 |
| **VQ-APC** | 5 | 0.34594 | 0.497 | 0.603 | 0.296 | 0.246 | 0.312 | 0.181 | 0.389 | 0.331 | 0.259 |
| **TERA** | 21 | 0.34547 | 0.493 | 0.596 | 0.295 | 0.253 | 0.308 | 0.193 | 0.385 | 0.337 | 0.249 |
| **W2V** | 33 | 0.34212 | 0.448 | 0.612 | 0.300 | 0.246 | 0.304 | 0.188 | 0.387 | 0.336 | 0.258 |
| **Mockingjay** | 85 | 0.33592 | 0.485 | 0.576 | 0.275 | 0.244 | 0.308 | 0.185 | 0.379 | 0.318 | 0.253 |
| **NPC** | 19 | 0.33150 | 0.470 | 0.570 | 0.274 | 0.240 | 0.304 | 0.172 | 0.364 | 0.333 | 0.256 |
| **VQ-W2V** | 34 | 0.33127 | 0.442 | 0.605 | 0.292 | 0.246 | 0.294 | 0.156 | 0.361 | 0.325 | 0.260 |
| **M CPC** | 2 | 0.31508 | 0.453 | 0.529 | 0.265 | 0.228 | 0.285 | 0.175 | 0.337 | 0.318 | 0.246 |
| **FBANK** | 0 | 0.19099 | 0.305 | 0.144 | 0.186 | 0.199 | 0.242 | 0.120 | 0.184 | 0.170 | 0.168 |

## 4.2 Preprossessing

### 4.2.1 Data Format

We ensure the presence of audio recordings and extract them from video clips if the original datasets lack separate audio files. If the audio is in stereo format, we convert it to a monophonic channel. Furthermore, we resample the audio to 16 kHz as it is the most common sampling rate for speech processing. Prior to passing the speech input into modeling, we normalize it by subtracting the mean and dividing it by the standard deviation of the training set across all our experiments.

### 4.2.2 Selection of emotions

Most SER prior studies (Atmaja and Sasou, 2022; Achiam et al., 2023) only choose anger, happiness, sadness, and neutral state emotions as target emotions. In addition, they regard the excitement/joy annotations as happiness; however, excitement and happiness are not the same emotions (Cowen and Keltner, 2017), though those two emotions have correlations (Mogilner et al., 2011).

In contrast to previous approaches, we retain all original emotion labels and refrain from merging any emotions into others to balance the data (e.g., combining excitement with happiness). This strategy allows us to accurately assess performance and mirror natural emotion perceptions under real-world conditions.

### 4.2.3 Label representation

Inspired by **Semantics Space Theory** (Cowen and Keltner, 2021), we gather numerous annotations and compute a distribution-like (soft label) representation, aiming to more accurately capture the high-dimensional nature of emotion perception. Notice that these distribution-like labels are the same as the **reference distribution** used for ChaptGPT as the reference label in section 3.1.

Here is one example: Let's assume we gather five annotations from five distinct raters for a single sample. These annotations comprise neutral (N), anger (A), anger (A), sadness (S), and sadness (S). Subsequently, we compute the label distributions, which in this instance are represented as (N, A, S, H) = (0.2, 0.4, 0.4, 0.0) for training SER systems. Additionally, in order to enhance SER performance, we employ the label smoothing technique proposed by (Szegedy et al., 2016) to refine the vector, utilizing a smoothing parameter of 0.05. This approach assigns a small probability to emotional classes with zero values.

## 4.3 Evaluation metric

We use the macro-F1 score (Opitz and Burst, 2019) to evaluate the SER performance via the Scikit-learn (Pedregosa et al., 2011), considering recall and precision rates simultaneously. For the distribution-like multi-label training target, we select target classes by applying thresholds on the ground truth. A prediction is deemed successful if the proportion for a class surpasses $1/C$, where $C$ represents the number of emotional classes, aligning with the settings employed in prior research (Riera et al., 2019). The example is in Appendix E.2

## 4.4 Objective function and training details

Inspired by the study (Cui et al., 2019), we adopt the class-balanced cross-entropy loss, as our primary objective function due to the imbalanced label

distributions across the six databases. Please refer to Appendix E.1 for detailed descriptions of class-balanced cross-entropy loss. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a 0.0001 learning rate and the batch size is 32. We choose the best models according to the lowest value of the class-balanced cross-entropy loss on the development set. We use the Nvidia Tesla v100 GPUs with 32 GB memory for all results. The total of GPU hours is around 3,300 hours. According to (Yang et al., 2021; Tsai et al., 2022; Feng et al., 2023), SSLMs usually result in consistent results and consume large computations, all results in the work are single-run. We also verify it by running experiments for small SSLMs, the variance is only 1% on average.

## 5 Results and Analysis

### 5.1 SSLMs for SER

We mainly use SSLMs as our backbone models to train SER systems in the work.

#### 5.1.1 Overall results

Table 2 summarizes macro-F1 scores obtained by 16 SSLMs and FBANK across six datasets under nine conditions. FBANK, the most commonly used speech feature, is the baseline for comparison with SSLMs. We have the following observations: (1) All SSLMs exhibit significantly superior performance compared to FBANK. Notably, XLS-R-1B achieves a remarkable improvement of relatively 100.8% compared to FBANK. (2) The XLS-R-1B model demonstrates the highest average performance, surpassing WavLM, which typically achieves state-of-the-art results in most speech-

processing tasks. Despite this, WavLM still maintains considerable strength, achieving the highest performance in three out of nine conditions. (3) Surprisingly, despite its modest 90 million model parameters, the DeCoAR 2 model outperforms the W2V2 model, which has 317 million parameters. This finding suggests that DeCoAR 2 could be an attractive choice for developers of SER facing computational resource constraints.

#### 5.1.2 Layer analysis

Our training strategy involves extracting features from each layer of the SSLM, multiplying these features with layer-specific weights, and then aggregating the weighted features. These aggregated features are then fed into the downstream model. Only the layer weights and the downstream model are trainable. A large weight assigned to a specific layer suggests that the layer encodes rich emotional information. Additionally, we conduct a layer-wise analysis of the SSLMs. We select SSLMs with top-five performance, each with the same number of layers: WavLM, Hubert, W2V2 R, Data2Vec-A, and W2V2. We extract the layer weights from the best checkpoint of each model and normalize them using the softmax function to ensure values between 0 and 1. If emotion datasets contain multiple partitions (e.g., IEMOCAP and CREMA-D), we average the layer-wise weights. We show main results and additional layer-wise analysis can be found in Appendix F.1.

From the model perspective (Figure 4a), where we sum the layer weights across all datasets for each model and plot the resulting curves. We have the following observations: Different models have higher weights on different layers. For instance, the W2V2 R has the highest weight on the 17th layer, but the Data2Vec-A's is on the third layer. Also, the other three models have similar patterns in emphasizing all layers. Additionally, it's worth noting that the weights of W2V2 R in layers 22 to 24 are considerably lower compared to those in other layers. We observe that the SSLMs act differently and have no clear patterns in the work.

Figure 4b illustrates the layer weights for the state-of-the-art model, XLS-R-1B. Similar to other models, it exhibits a tendency to prioritize the shallow layers. However, there are two notable peak weights observed on the 30th layer, particularly trained on the POD (P) and POD (S) datasets.
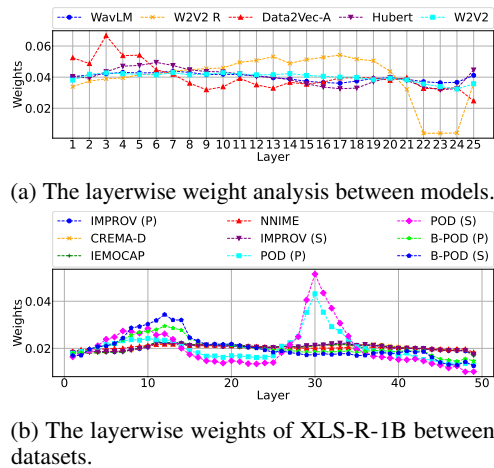


(a) The layerwise weight analysis between models.



(b) The layerwise weights of XLS-R-1B between datasets.

Figure 4: The layerwise weights analysis.

Table 3: The table presents macro-F1 scores using the integration of labels from ChatGPT.

| SSLM | wo ChatGPT labels | w ChatGPT labels | Relative gain |
|---|---|---|---|
| WavLM | 0.350 | 0.353 | 0.77% |
| W2V2 R | 0.331 | 0.335 | 1.08% |
| XLS-R-1B | 0.331 | 0.341 | 3.09% |
| Data2Vec-A | 0.329 | 0.338 | 2.74% |
| Hubert | 0.342 | 0.350 | 2.22% |
| W2V2 | 0.321 | 0.325 | 1.28% |
| VQ-W2V | 0.292 | 0.300 | 2.74% |
| W2V | 0.301 | 0.305 | 1.58% |
| CPC | 0.265 | 0.290 | **9.45%** |
| DeCoAR 2 | 0.308 | 0.317 | 3.14% |
| TERA | 0.295 | 0.306 | 3.52% |
| Mockingjay | 0.275 | 0.298 | 8.49% |
| NPC | 0.275 | 0.290 | 5.75% |
| VQ-APC | 0.296 | 0.310 | 4.94% |
| APC | 0.298 | 0.307 | 3.19% |
| FBANK | 0.186 | 0.186 | 0.00% |
| Average | 0.298 | 0.307 | 3.08% |

## 5.2 Results with re-labeled data

Figure 5 shows a random selected example to compare the original distribution and relabeled distribution by ChatGPT (data sample "POD-CAST_1631_0043_0001.wav"). We observe that ChatGPT effectively comprehends the typed descriptions conveying positive emotions, thereby assigning greater weight to the "happy" emotion category. More examples can be found in Figure 8a to Figure 8d and Table 10 in Appendix D.1.3.

Table 3 presents the macro-F1 scores of the experiment along with the effects of incorporating data labeled by ChatGPT. We denote "w/o Chat-GPT labels" and "w/ ChatGPT labels" to signify
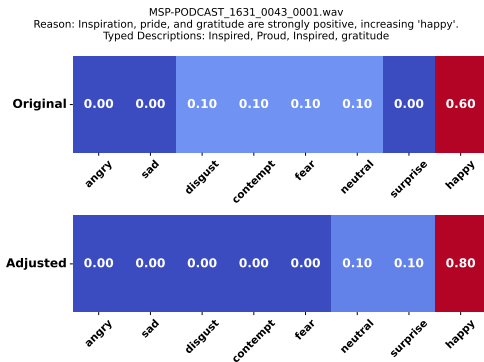


Figure 5: Original and adjusted distributions. The original distribution, determined by tallying the votes for each emotion class, is compared with the adjusted distribution resulting from ChatGPT's re-labeling. In the raw annotations of the example, there are instances of disgust, contempt, fear, neutrality, and happiness (*6), resulting in values of 0.6 for happiness and 0.1 for each of the remaining emotions.

results without and with ChatGPT labels, respectively, while maintaining all other settings the same. We note the following observations: (1) The experiments involved 16 models, resulting in an average relative performance gain of 3.08%. (2) Particularly noteworthy is the case of CPC, which exhibits a substantial 9.45% relative improvement.

## 5.3 Takeaways

Here are some takeaways: (1). DeCoAR 2 achieves considerate results on SER within a modest model size, making it an appealing option for developers constrained by computation resources. (2). SSL models trained for SER exhibit a tendency to assign greater weight to shallow layers, which are used to encode emotional information. (3). Leveraging ChatGPT for relabeling typed descriptions holds significant promise in boosting SER performance.

## 6 Ethical Considerations and Limitations

In terms of limitations, our study solely focused on emotion datasets in English and Chinese, omitting datasets in other languages. Additionally, the absence of recordings featuring elderly and child speech, coupled with unknown annotator details, may hinder the representation of emotional perception across certain demographics. We do not address potential performance biases related to speaker gender within the SER systems. We only utilize ChatGPT for relabeling typed descriptions in the PODCAST dataset, given the considerable expense associated with utilizing the ChatGPT API. The task of designing improved prompts and labeling typed descriptions for other datasets remains for future investigation. We haven't attempted fine-tuning the entire SSLMs due to the significant computational resources required.

## 7 Conclusion and Future Work

We propose EMO-SUPERB, an ecosystem containing user-friendly codebases, ChatGPT re-labeled datasets, pre-trained models, fair data partition files, and a community-driven leaderboard for SER. We effectively address open questions in SER, including (1) boosting reproducibility, (2) addressing data leakage, and (3) leveraging unused typed descriptions. We encourage the community to use EMO-SUPERB to develop and evaluate the SER systems. We plan to expand our investigation in future work by incorporating additional evaluation angles, such as calibration error and gender bias.

8

# References

Josh Achiam et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nikolaos Antoniou, Athanasios Katsamanis, Theodoros Giannakopoulos, and Shrikanth Narayanan. 2023. Designing and Evaluating Speech Emotion Recognition Systems: A Reality Check Case Study with IEMOCAP. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Bagus Tris Atmaja and Akira Sasou. 2022. Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition. *IEEE Access*, 10:124396–124407.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. 2016. Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2017. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 292–298.

Huang-Cheng Chou, Wei-Cheng Lin, Chi-Chun Lee, and Carlos Busso. 2022. Exploiting Annotators' Typed Description of Emotion Perception to Maximize Utilization of Ratings for Speech Emotion Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 7717–7721, Singapore.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An unsupervised autoregressive model for speech representation learning. *arXiv preprint arXiv:1904.03240*.

Yu-An Chung, Hao Tang, and James Glass. 2020. Vector-quantized autoregressive predictive coding. *arXiv preprint arXiv:2005.08392*.

Alan S. Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

Alan S. Cowen and Dacher Keltner. 2021. Semantic Space Theory: A Computational Approach to Emotion. *Trends in Cognitive Sciences*, 25(2):124–136.

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, California, USA.

Tzu-hsun Feng, Annie Dong, Ching-Feng Yeh, Shu-wen Yang, Tzu-Quan Lin, Jiatong Shi, Kai-Wei Chang, Zili Huang, Haibin Wu, Xuankai Chang, et al. 2023. Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1096–1103. IEEE.

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

9

Kiana Kheiri and Hamid Karimi. 2023. SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning.

Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2023. Exploration of a Self-Supervised Speech Model: A Study on Emotional Corpora. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 868–875.

Shaoshi Ling and Yuzong Liu. 2020. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*.

Alexander H Liu, Yu-An Chung, and James Glass. 2020a. Non-autoregressive predictive coding for learning speech representations from local dependencies. *arXiv preprint arXiv:2011.00406*.

Andy T Liu, Shang-Wen Li, and Hung-yi Lee. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020b. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Reza Lotfian and Carlos Busso. 2017. Formulating Emotion Perception as a Probabilistic Model with Application to Categorical Emotion Classification. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, pages 415–420, San Antonio, TX, USA.

Reza Lotfian and Carlos Busso. 2019. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Cassie Mogilner, Sepandar D. Kamvar, and Jennifer Aaker. 2011. The Shifting Meaning of Happiness. *Social Psychological and Personality Science*, 2(4):395–402.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Juri Opitz and Sebastian Burst. 2019. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Fabian Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pablo Riera, Luciana Ferrer, Agustín Gravano, and Lara Gauder. 2019. No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems. In *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, pages 11–15, Graz, Austria.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy T Liu, Cheng-I Jeff Lai, Jiatong Shi, et al. 2022. Superb-sg: Enhanced speech processing universal performance benchmark for semantic and generative capabilities. *arXiv preprint arXiv:2203.06849*.

Shreya G Upadhyay, Woan-Shiuan Chien, Bo-Hao Su, Lucas Goncalves, Ya-Tse Wu, Ali N Salman, Carlos Busso, and Chi-Chun Lee. 2023. An Intelligent Infrastructure Toward Large Scale Naturalistic Affective Speech Corpora Collection. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833

# A  Details of Emotion Dataset

Table 5 summarizes six public emotion datasets, detailing the number of citations for each, the total length of audio recordings (partitioned length), original recording sampling rates, and length statistics. Additionally, it provides statistics on segmented utterances, speaker and annotator information, and collection settings for all datasets.

## A.1  License of Emotion Dataset

Table 4 offers a concise summary of the licensing details for six emotion datasets to facilitate their accessibility within the research community. It's important to note that the majority of these datasets are restricted to academic use. However, the POD-CAST and BIIC-PODCAST datasets stand out as they also offer the option for commercial licensing, albeit for a fee.

## A.2  The SAIL-IEMOCAP

The SAIL-IEMOCAP dataset, referenced in this document as **IEMOCAP**, was meticulously assembled from the motion capture, audio, and video recordings of dyadic conversations involving ten professional English-speaking actors (Busso et al., 2008). This dataset uniquely captures a blend of scripted and spontaneous dialogues, focusing primarily on scenarios that portray lovers in a relationship to elicit a rich spectrum of emotions. Each recording session featured a pair of speakers, one female and one male, engaging in interactions designed to provoke distinct emotional responses.

To ensure a diverse emotional range, the actors were provided with scripts curated to trigger specific feelings. The completed recordings were subsequently divided into 10,039 segments, each meticulously transcribed to facilitate further analysis.

Annotators then reviewed these segments, selecting emotions from a predefined list that encompassed ten distinct states: neutral, happiness,

sadness, anger, surprise, fear, disgust, frustration, excitement, and an "other" category for emotions outside the listed spectrum. Additionally, a provision for typed descriptions allowed for a more nuanced annotation, accommodating the labeling of complex emotions that might not fit neatly into predefined categories. Notably, the annotation process permitted the selection of multiple emotions for a single utterance, with each segment being evaluated by at least three individuals from a pool of twelve annotators, comprised equally of actors and untrained (naive) annotators.

A significant aspect of the IEMOCAP dataset is its focus on both self-perception and observed perception annotations, differentiating it from other emotional databases. This dual approach provides a more comprehensive understanding of the emotional landscape captured within the dataset.

To address challenges related to data reproducibility, highlighted in previous research (Antoniou et al., 2023), we have included detailed information on the dataset splits in Appendix refss:cviemocap. This is especially crucial considering the absence of standard split sets within the original corpus, a gap that our documentation aims to bridge.

## A.3  The CREMA-D

The CREMA-D dataset, introduced by (Cao et al., 2014), is a valuable resource comprising high-quality audio-visual clips featuring performances from 91 professional actors, including 43 females and 48 males. These actors were tasked with recording one of twelve predetermined sentences, expressing six distinct emotions: anger, disgust, fear, happiness, sadness, and a neutral state.

A notable aspect of this dataset is its extensive annotation process, involving 7,442 clips in English, evaluated by 2,443 unique annotators through a crowdsourcing platform. Each clip received feedback from at least six annotators, with each annotator attributing one of the six aforementioned emotions to the utterance. The perceptual annotation process unfolds across three distinct scenarios: voice-only, face-only, and audio-visual. In the voice-only scenario, annotators solely listen to the audio of the clips. Conversely, in the face-only scenario, annotators observe the facial expressions of the actors without accompanying audio. Finally, the audio-visual scenario allows annotators to assess both facial expressions and audio simultaneously.

834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884

Table 4: The table summarizes license information of the six public emotion databases. The license can be accessed by clicking the word, **Agreement** or **Website**.

| Dataset | License | Commercial Purposes |
| --- | --- | --- |
| SAIL-IEMOCAP | Agreement | No |
| CREMA-D | Agreement | No |
| MSP-IMPROV (P) | Agreement | No |
| MSP-PODCAST | Agreement | YES |
| BIIC-NNIME | Agreement | No |
| BIIC-PODCAST | Website | YES |

Table 5: The table summarizes other detailed information of the 6 public emotion databases.

| Database | IMPROV | CREMA-D | PODCAST | B-PODCAST | IEMOCAP | NNIME |
|---|---|---|---|---|---|---|
| Citation (Paper) | 323 | **562** | 271 | 1 | **3181** | 46 |
| Length (hrs) | 9.53 | 5.26 | 235.94 | 147.43 | 12.44 | 3.26 |
| Length (Train, Dev., Test) (hrs) | (6.35, 1.59, 1.59) | (3.15, 1.05, 1.05) | (134.34, 31.72, 69.88) | (102.51, 22.32, 22.61) | (7.46, 2.49, 2.49) | (1.96, 0.65, 0.65) |
| Sampling Rate (K Hz) | 44.1 | 16 | 16 | 16 | 16 | 16 |
| Max. Length (sec.) | 31.91 | 5.01 | 11.94 | 16.02 | 34.14 | 71.81 |
| Avg. Length (Std.) (sec.) | 4.09 (2.89) | 2.54 (0.51) | 5.69 (2.35) | 7.58 (3.31) | 4.46 (3.06) | 2.32 (2.81) |
| Min. Length (sec.) | 0.41 | 0.51 | 1.91 | 0.51 | 0.58 | 0.128 |
| No. of Utt. | 8385 | 7442 | 149307 | 70000 | 10039 | 5028 |
| Excluded Utt. | 53 | 0 | 2347 | 0 | 0 | 568 |
| No. of Speaker | 12 | 91 | 2172+Unknown | Unknown | 10 | 43 |
| Speaker Gender | 6F; 6M | 43F; 48M | 904F;1268M;Unknown | Unknown | 5F; 5M | 24F; 19M |
| Transcriptions | V | | V | V | V | V |
| Other Modality | Video | Video | | | Video | Video, Physiology |
| Labels/Utt. | 7.36 | 9.84 | 5.72 | 3.33 | 3.24 | 5.12 |
| Raters/Utt. (Min.) | 5 | 6 | 5 | 5 | 3 | 3 |
| Raters/Utt. (Mean/Std.) | 7.31 | 9.84 | 5.72 (2.29) | 3.33 (0.75) | 3.24 (0.43) | 4.97 (0.67) |
| Raters/Utt. (Max.) | 50 | 12 | 32 | 9 | 4 | 6 |
| No. of Rators | Unknown | 2443 | 14363 | Unknown | 12 | 6 |
| Perception | Observed | Observed | Observed | Observed | 6 Observed, 6 Self | Observed |
| Stimulus | Audio-visual | Voice-only | Voice-only | Audio-visual | Audio-visual | |
| Emotions (P) | 4 | 6 | 8 | 8 | | |
| Emotions (S) | 10 | | 16 | 16 | 9 | 11 |
| Setting | Scripted+Improvised | Scripted | Real-world | Real-world | Scripted+Improvised | Improvised |
| Context | V | | | | V | V |
| Speakers | Actors | Actors | Real-world | Real-world | Actors | Actors |

For the purpose of our study on Speech Emotion Recognition (SER), we focus exclusively on emotional annotations derived from the voice-only scenario. Notably, while many previous SER investigations utilized annotations from the audio-visual scenario as a learning target, or failed to specify annotation details altogether, we opt to leverage annotations solely from the voice-only setting for our analysis. Furthermore, we provide comprehensive details regarding the dataset splits employed in our paper, as outlined in Appendix C.2.1.

## A.4 The MSP-IMPROV

The MSP-IMPROV dataset, also known as IM-PROV (Busso et al., 2017), comprises high-quality audio-video recordings featuring interactions acted out by 12 actors in English. These sessions encompass four distinct emotions: anger, happiness, sadness, and a neutral state. Notably, each dyadic interaction involves one male and one female speaker and is recorded in four different scenarios: preparation, scripted, improvised, and improvised-scripted scenes.

To ensure comprehensive annotation, all sessions within the MSP-IMPROV dataset are manually segmented into 8,438 clips, each evaluated by at least five annotators via a crowdsourcing platform. Utilizing the quality control method proposed by (Burmania et al., 2016), the dataset employs mechanisms to identify and eliminate unreliable annotators.

The annotation process within MSP-IMPROV presents two scenarios: primary (P) and secondary (S) emotions. In the primary scenario, annotators select one emotion from a set of five options: anger, happiness, sadness, neutral state, or "other." Secondary emotions, on the other hand, encompass a broader range, including frustration, depression, disgust, excitement, fear, and surprise. Notably, 53 utterances annotated as "other" by annotators are excluded from the dataset, although annotators have the option to provide textual descriptions when choosing this category.

Given that the dataset does not include predefined train, development, and test sets, we introduce our proposed split sets in Appendix C.2 for the sake of clarity and consistency in subsequent analyses.

## A.5 The MSP-PODCAST

The MSP-PODCAST dataset, or PODCAST (Lotfian and Busso, 2019), offers a rich collection of spontaneous and diverse emotional speech extracted from real-world podcast recordings obtained under commercial licenses. Initially, the podcast recordings are segmented into individual utterances, which are then annotated via a crowdsourcing platform. Similar to MSP-IMPROV, the dataset implements quality control measures based on the methodology outlined by (Burmania et al.,

2016) to ensure the reliability of annotators.

The annotation framework within MSP-PODCAST encompasses both primary (P) and secondary (S) scenarios. In the primary scenario, annotators select from a set of nine predefined emotions: anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and "other," with the option to provide additional textual descriptions if necessary. The secondary scenario expands upon the primary emotions, incorporating an additional eight classes: amusement, frustration, depression, concern, disappointment, excitement, confusion, and annoyance, totaling 17 options.

Each utterance within the dataset is evaluated by at least five unique annotators, ensuring robustness in the annotation process. The dataset version 1.11 consists of 84,030 utterances in the train set, 19,815 in the development set, 30,647 in the combined test set (test1 and test2), and 2,347 in the test3 set, which is excluded from analysis due to its private nature lacking annotations.

In total, the dataset encompasses contributions from over 2,172 distinct speakers and involves 14,363 annotators, providing a comprehensive resource for studying emotional speech.

### A.6   The BIIC-NNIME

The BIIC-NNIME dataset, also known as NNIME (Chou et al., 2017), is a comprehensive resource featuring video, audio, and physiology recordings of dyadic conversations acted out by 43 actors in Mandarin Chinese. These sessions are characterized by spontaneous, unscripted interactions set in everyday home environments, encompassing six emotional scenes: anger, frustration, happiness, sadness, surprise, and neutral states.

Each session within the NNIME dataset is meticulously segmented into 5,596 clips. To maintain annotation quality, utterances labeled as "other" by annotators or those annotated by fewer than three annotators are excluded from analysis. Notably, NNIME stands out from other emotion datasets due to its annotation of both speech and non-verbal behaviors, such as laughter, sighing, sobbing, and other vocal expressions.

With a total of 43 unique speakers and annotations from six different annotators, the labeling process within NNIME resembles that of the SAIL-IEMOCAP dataset. Annotators view clips sequentially and select emotions from a pool of 12 options: anger, frustration, disappointment, sadness, fear, surprise, excitement, happiness, relaxation, joy, neutral state, and "other." Moreover, annotators have the flexibility to express emotional perceptions using Chinese words.

Given the absence of standard split sets for training deep-learning models within the corpus, we provide details of our proposed split sets in Appendix C.3 to ensure reproducibility and facilitate further research using the NNIME dataset.

### A.7   The BIIC-PODCAST

The BIIC-PODCAST dataset, or B-PODCAST (Upadhyay et al., 2023), presents a Mandarin Chinese variant of the MSP-PODCAST, featuring audio recordings sourced from real-world podcasts under commercial licenses. Notably, the dataset diverges from MSP-PODCAST in its labeling process, employing college students as annotators instead of utilizing a crowdsourcing platform. This approach aims to enhance quality control and ensure the reliability of annotations, a methodology similar to MSP-PODCAST's quality assessment standards.

Version 1.01 of the B-PODCAST dataset includes 48,815 utterances in the train set, 10,845 in the development set, and 10,340 in the test set. Each utterance undergoes evaluation by a minimum of five annotators. The emotional annotations within B-PODCAST encompass both primary (P) and secondary (S) emotions, mirroring the structure of MSP-PODCAST.

The primary emotions (P) include the same set of nine options as MSP-PODCAST, comprising anger, sadness, happiness, surprise, fear, disgust, contempt, neutral, and "other." Similarly, the secondary emotions (S) expand upon the primary emotions with additional classes, maintaining consistency with MSP-PODCAST.

Overall, B-PODCAST serves as a valuable resource for studying emotional speech in Mandarin Chinese, offering a curated dataset with robust annotations and quality control measures.

## B   SSLM introductions

In our codebase, we include two mainstream categories of SSLMs, pre-trained using generative loss, DeCoAR 2 (Ling and Liu, 2020), Autoregressive Predictive Coding (APC) (Chung et al., 2019), VQ-APC (Chung et al., 2020), Non-autoregressive Predictive Coding (NPC) (Liu et al., 2020a), TERA (Liu et al., 2021), and Mockingjay (Liu et al., 2020b)), and discriminative loss (**XLS-R-1B**)

(Babu et al., 2021), WavLM (Chen et al., 2022), Hubert (Hsu et al., 2021), wav2vec 2.0 (**W2V2**) (Baevski et al., 2020), VQ wav2vec (**VQ-W2V**) (Baevski et al., 2019), wav2vec (**W2V**) (Schneider et al., 2019), and Contrastive Predictive Coding (CPC) (**M CPC**)(Oord et al., 2018)).

APC employs a pretraining strategy similar to language models on a sequence of acoustic features (**FBANK**). It utilizes unidirectional RNNs to predict future FBANK frames based on past ones. VQ-APC improves APC's representation by integrating vector-quantization (VQ) layers. NPC boosts APC's efficiency by replacing RNNs with CNNs for faster inference. Mockingjay consists of Transformer encoders. It masks segments of input acoustic features along the time axis and reconstructs them during training. TERA builds upon Mockingjay's architecture by extending the masking strategy to frequency bins.

DeCoAR 2.0 refines Mockingjay's design by incorporating a VQ layer just before final predictions, similar to VQ-APC's approach. Its training involves larger input masks, increased batch sizes, and the utilization of more unlabeled data to improve performance. Wav2vec introduced several architectural enhancements to refine CPC's performance. VQ-wav2vec integrates a VQ module into wav2vec, discretizing speech into tokens post after InfoNCE pretraining. These discrete tokens are used for training a BERT model, to get contextualized representations. Wav2vec 2.0 streamlines the vq-wav2vec pipeline into an end-to-end framework. This involves employing time masking in the latent space and substituting BERT's token prediction with InfoNCE's negative sampling. XLS-R builds upon wav2vec 2.0, expanding its capabilities to encompass multiple languages and augmenting the dataset size. HuBERT enables BERT's token prediction through offline clustering of representations. Predictions are made based on the clustered labels at masked locations. WavLM, based on Hubert, introduces noise during pretraining to enhance the robustness of SSL features.

## C    Partition Setting

In the speaker-independent scenario, where the model is trained on data from certain speakers and tested on data from speakers not seen during training, ensuring fair and robust evaluation is crucial. Here are the details about data partitions for experiments on the SAIL-IEMOCAP, MSP-IMPROV, CREMA-D, and BIIC-NNIME datasets.

### C.1    The IEMOCAP

Table 6 summarizes the partitioning settings for the IEMOCAP corpus. Considering each session, we define five speaker-independent splits (i.e., Dyad 1 to Dyad 5). Each session consists of two speakers engaged in dyadic interactions. In our experiments, we conduct a 5-fold cross-validation as illustrated in Table 6, where each fold includes a unique combination of training, development, and test sets to ensure comprehensive evaluation of the model's performance across different dyadic interactions within the IEMOCAP corpus.

Table 6: IEMOCAP corpus partitions.

| Partition | Training Set | Development Set | Test Set |
|---|---|---|---|
| 1 | Dyad 1,2,3 | Dyad 4 | Dyad 5 |
| 2 | Dyad 2,3,4 | Dyad 5 | Dyad 1 |
| 3 | Dyad 3,4,5 | Dyad 1 | Dyad 2 |
| 4 | Dyad 1,4,5 | Dyad 2 | Dyad 3 |
| 5 | Dyad 1,2,4 | Dyad 3 | Dyad 4 |

### C.2    The IMPROV

In the speaker-independent scenario, the MSP-IMPROV corpus is partitioned into six folds for cross-validation. Each fold consists of a unique combination of training, development, and test sets, as illustrated in Table 7. This partitioning strategy ensures that the model is trained on interactions involving different sets of speakers and evaluated on unseen speaker combinations, facilitating robust evaluation of model generalization across various dyadic conversations within the MSP-IMPROV corpus.

Table 7: MSP-IMPROV corpus partitions.

| Partition | Training Set | Development Set | Test Set |
|---|---|---|---|
| 1 | Dyad 1,2,3,4 | Dyad 5 | Dyad 6 |
| 2 | Dyad 1,2,3,6 | Dyad 4 | Dyad 5 |
| 3 | Dyad 1,2,5,6 | Dyad 3 | Dyad 4 |
| 4 | Dyad 1,4,5,6 | Dyad 2 | Dyad 3 |
| 5 | Dyad 3,4,5,6 | Dyad 1 | Dyad 2 |
| 6 | Dyad 2,3,4,5 | Dyad 6 | Dyad 1 |

#### C.2.1    The CREMA-D

In the speaker-independent scenario, the CREMA-D corpus is divided into five sets based on speaker IDs. Each set consists of a different combination of male and female speakers, as well as a distinct range of speaker IDs, as summarized in Table 8. This partitioning strategy allows for a

14

fair and balanced evaluation of models trained on CREMA-D data by ensuring that the test sets contain speakers not seen during training, thereby assessing the model's ability to generalize across different speaker characteristics and expressions. The standard partitions follow a similar methodology as the one mentioned for the IEMOCAP dataset in section C.1.

Table 8: The CREMA-D sessions. M represents male and F represents female.

| Session | Gender | Speaker ID |
|---------|--------|------------|
| 1 | 7M;11F | 1037-1054 |
| 2 | 12M;6F | 1001-1018 |
| 3 | 13M;6F | 1073-1091 |
| 4 | 9M;9F | 1055-1072 |
| 5 | 15M;3F | 1019-1036 |

### C.3 The NNIME

In the speaker-independent scenario, the NNIME corpus is randomly split into five sets based on speaker IDs. Each set comprises a different combination of male and female speakers, as well as a distinct set of speaker IDs, as summarized in Table 8. This partitioning strategy enables a fair and unbiased evaluation of models trained on NNIME data by ensuring that the test sets include speakers not encountered during training, thereby assessing the model's generalization capabilities across various speakers and expressions. The standard partitions follow a similar methodology as the one mentioned for the IEMOCAP dataset in section C.1.

Table 9: The NNIME sessions. M represents male and F represents female.

| Session | Gender | Speaker ID |
|---------|--------|------------|
| 1 | 6M;3F | 01,02,03,04,22 |
| 2 | 4M;4F | 05,06,07,08 |
| 3 | 1M;7F | 09,10,11,12 |
| 4 | 2M;6F | 13,14,15,16] |
| 5 | 6M;4F | 17,18,19,20,21 |

## D Typed Descriptions

Figure 6 displays a word cloud generated from typed descriptions collected in the POD emotion dataset. Word clouds are visual representations that highlight the frequency of words in a text corpus, with larger font sizes indicating higher frequencies.
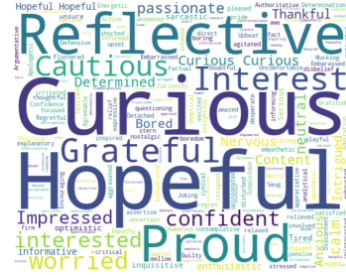


Figure 6: The figures shows typed descriptions of the MSP-PODCAST for an example.

In this context, the word cloud provides insight into the types of words or phrases used by annotators to describe their emotional responses in the dataset.

Analyzing typed descriptions can be valuable for understanding the nuances of human emotion and improving the comprehensiveness of emotion recognition systems. By incorporating natural language processing techniques, researchers can extract valuable insights from typed descriptions to enhance the accuracy and granularity of emotion annotation in datasets like POD.

### D.1 Prompt for ChatGPT

#### D.1.1 Design of Prompt

Table 11 shows the well-designed prompt, which contains five parts: objective, input format, example, output format, and refine and iterate. In the objective part, we clearly describe the goal of task. Then, we define the input format, including descriptions and reference emotion. Afterwards, we provide an example that provide the template the ChatGPT can follow. Finally, we ask the ChatGPT output the file in JSON format. Notice that the current version of prompt is the 14th version. In the refine and iterate part, we show the more rules that can enhance the accuracy of the output of the ChatGPT. We encourage the community to provide the designed prompt.

#### D.1.2 Parameters and Cost of ChatGPT API

Listing shows the simple code of ChatGPT API. We setup the temperature as 0. and seed as 7 in the work to keep the reproducibility. We choose the version, "gpt-4-0125-preview", as the main model.

The average cost per data sample is $0.0045 USD. The target emotion is represented by an 8-dimensional distribution vector. During training and development, ChatGPT examined 34.21% and 33.32% of the data, respectively. Out of the total
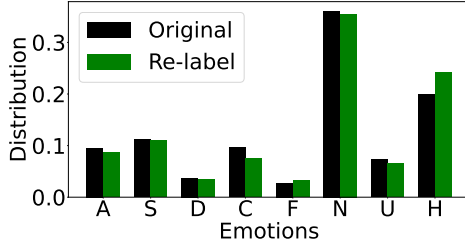
15

Figure 7: The figures shows comparisons of original and re-labeled label distribution. Emotion includes anger (A), sadness (S), happiness (H), surprise (U), fear (F), disgust (D), contempt (C), and neutral (N).

35,352 data samples available, only 12.65% were not modified by ChatGPT. This translates to a total experimental cost of approximately $160 USD.

```
1  response =
       client.chat.completions.create(
2      temperature = 0.,
3      seed = 7,
4      model="gpt-4-0125-preview",
5      response_format={'type':
           'json_object' },
6      messages=[
7          {
8              "role": "system",
9              "content": prompt},
10         {"role": "user",
11             "content": input_data}
12     ])
```

Listing 1: ChatGPT API Python Code

### D.1.3 Output of ChatGPT

Figure 7 and 8 shows the changes of label distributions between original labels and re-label one. The ChatGPT increased the number of fear and happiness and decreased the other emotions. In addtion, the Table 10 shows ten examples including typed descriptions and reasons provided by ChatGPT.

## E Experiments

### E.1 Class-balanced cross-entropy loss

This approach proposed by the study (Cui et al., 2019) helps mitigate the impact of class imbalance by giving more weight to minority classes during the optimization process, leading to improved model performance, especially for datasets with imbalanced class distributions. The main idea is to add a weighting factor to adjust the values of the used loss function based on the inverses of the class frequency considering the training set. The factor is $\frac{1-\beta}{1-\beta^{n_j}}$, where $n_j$ is the number of positive samples in the $j^{th}$ emotion class in the train set,

and $\beta \in (0, 1]$ is a hyperparameter. The number of factors to weigh the loss values equals to the number of target emotions. The CBCE value can be calculated using Eq. 1, where $\mathcal{L}_{CE}^{(j)}$ is the value of cross-entropy loss for the $j^{th}$ emotion.

$$\mathcal{L}_{CBL} = \sum_{j=1}^{K} (\frac{1-\beta}{1-\beta^{n_j}} \cdot \mathcal{L}_{CE}^{(j)}). \quad (1)$$

### E.2 Evaluation Example

For instance, consider a four-class emotion recognition task, and the emotion classes contain neutral, anger, sadness, and happiness. Assume we consider the predictions for three different models: (0.2,0.35,0.35,0.1), (0.1,0.45,0.45,0.0), and (0.45,0.1,0.0,0.45). The three predictions are transformed into (0,1,1,0), (0,1,1,0), and (1,0,0,1), respectively, using the threshold. In these cases, only the first two predictions are fully corrected.

## F Analysis

Figure 10 and 11 show the weights across layers.

### F.1 Layer-wise Analysis

In Figure 10, the patterns of POD (P) and POD (S) diverge from those of other datasets. This variance could be attributed to the incorporation of more real-life data, which likely includes additional background noise and a wider range of speakers. Concerning Chinese datasets, namely B-POD (P), B-POD (S), and NNIME, the layer weights remain relatively stable. This consistency may stem from the fact that the SSLMs were trained using English data. Across all datasets, it appears that the later layers (22nd-24th) of W2V2 are of lesser significance. Examining IEMOCAP, IMPROV (P and S), and CREMA-D, it is evident that earlier layers receive higher weighting when utilizing Data2Vec-A.

In Figure 11, our observations align with those reported in the study by Li et al. (2023), indicating a concentration of the model's focus on the shallow layers (1st-5th). However, it is noteworthy that Li et al. (2023) identified the sixth layer as the one providing the most effective representation for the SER task. Additionally, the DeCOAR model relies on the later layers for the SER task.

Table 10: Relabeled examples of typed descriptions with ChatGPT. ChatGPT also provides the reason to change the reference distribution or not. **Change** denotes whether the reference distribution label is changed or not.

| Change | Index | Typed Descriptions | Reason |
|---|---|---|---|
| Yes | 01 | calm,Slightly Angry,calm | Increased disgust to reflect slight anger. |
| | 02 | Sarcastic ,Sarcastic | Sarcasm can be mistaken for happiness but has underlying sadness or neutrality. |
| | 03 | Curious,Curious | Increased surprise and happiness for curiosity. |
| | 04 | hopeful,hopeful | Increased neutral and happy due to hopeful; decreased disgust. |
| | 05 | Inspired,Proud,Inspired,gratitude | Inspiration, pride, and gratitude are strongly positive, increasing 'happy'. |
| No | 06 | Tranquil | Maintained high neutral for tranquil's peacefulness without strong emotions. |
| | 07 | mellow,Contented | Kept happy at maximum for contentedness and mellowness. |
| | 08 | praising | Kept 'happy' at 1.0 due to 'praising' indicating strong positive feedback. |
| | 09 | Cocky,Mocking | Cockiness and mocking suggest contempt, but also a high degree of happiness. |
| | 10 | pride | Pride is a positive emotion, strongly associated with happiness. |

Table 11: The Prompt for ChatGPT.

**Objective:**

As a knowledgeable assistant psychologist, your role is to analyze the given words and reference labels. You generate emotion label distributions. The emotions to consider are: 'angry,' 'sad,' 'disgust,' 'contempt,' 'fear,' 'neutral,' 'surprise,' and 'happy.' The order of emotions is very important. Please provide 8-dimensional emotion distributions for these 8 emotion classes based on the user input.

**Input format:**

The user input has two parts separated by #: The first part is the description. The second part is 8-dimensional reference emotion distribution, 'angry,' 'sad,' 'disgust,' 'contempt,' 'fear,' 'neutral,' 'surprise,' and 'happy.' The order of reference emotion is very important.
The input has the format "descriptions#reference emotion distribution". Also give the reason for each data point why you want to change the reference emotion distribution.
When given the answer, you should focus 25% on the "descriptions" and 75% on the "reference emotion distributions".

**Example:**

I will give you one example:
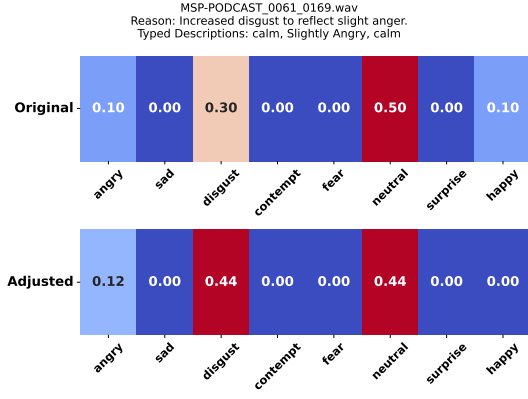User Input: Concerned,Interest#0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0.
Generated Labels: {'angry': 0.1, 'sad': 0.2, 'disgust': 0.2, 'contempt': 0.3, 'fear': 0.0, 'neutral': 0.2, 'surprise': 0.0, 'happy': 0.0, "reason": ""}
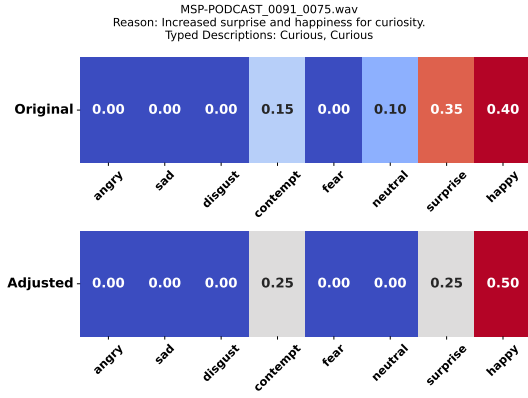
**Output format:**

Reminder for the give data, It's very important to output the JSON format with index.
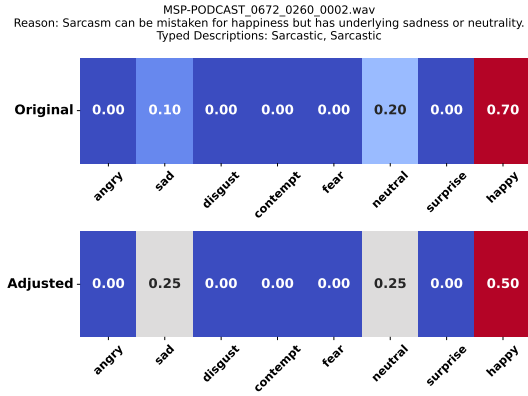
**Refine and Iterate :**

I will give you 30 data points each time. Each data is separated by "|". It's very important. It's very important to make sure that you complete every response for 30 data points each time. Please reminder it. Output the JSON file that contains adjusted emotion label distributions based on reference distributions and detailed reasons why you adjust the reference emotion distributions each word by each word. It's very important. It's very important that the JSON output file must contain the reference distributions and reasons. It's very important that do not contain the reference distributions and words. It's very important that use 15 to 20 words to explain the reason you want to change the reference distributions. It's very important that the sum of label distributions equals 1. It's very important to make sure that you explain the reasons for each word in descriptions.
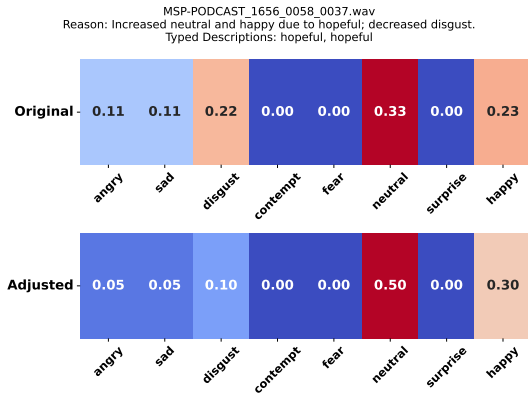
MSP-PODCAST_0061_0169.wav
Reason: Increased disgust to reflect slight anger.
Typed Descriptions: calm, Slightly Angry, calm

(a) The second GPT modified example.

MSP-PODCAST_0091_0075.wav
Reason: Increased surprise and happiness for curiosity.
Typed Descriptions: Curious, Curious

(b) The third GPT modified example.

MSP-PODCAST_0672_0260_0002.wav
Reason: Sarcasm can be mistaken for happiness but has underlying sadness or neutrality.
Typed Descriptions: Sarcastic, Sarcastic

(c) The fourth GPT modified example.

MSP-PODCAST_1656_0058_0037.wav
Reason: Increased neutral and happy due to hopeful; decreased disgust.
Typed Descriptions: hopeful, hopeful

(d) The fifth GPT modified example.

Figure 8: The original distribution and adjusted distribution after ChatGPT re-labeling.
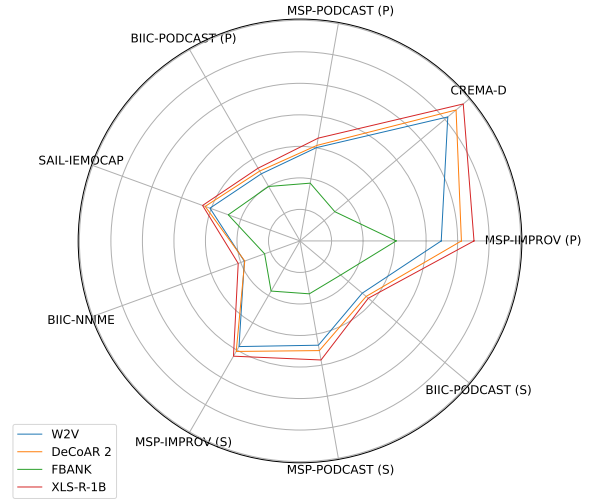


Figure 9: Demonstration of radar chart to compare four models, W2V, DeCoAR 2, XLS-R-1B and FBANK across 9 conditions.
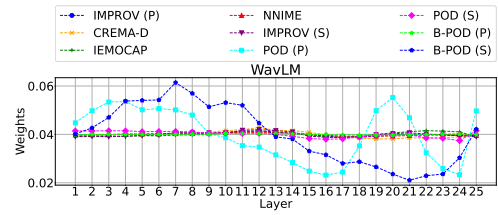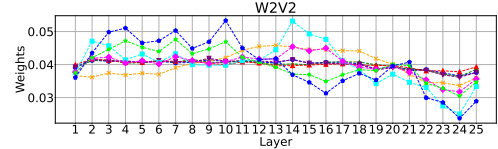


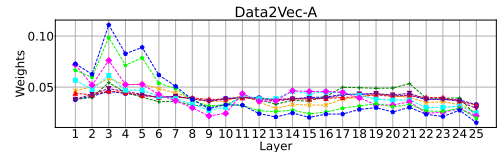(a) The layerwise weights of the WavLM.
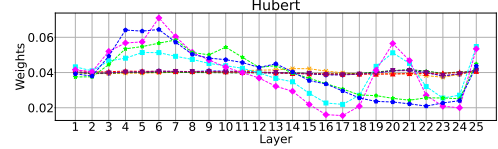
(b) The layerwise weights of the Hubert.
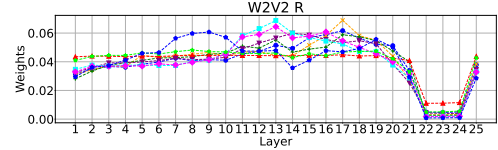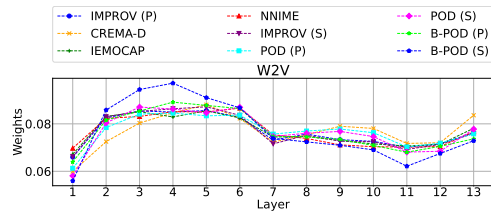
(c) The layerwise weights of the W2V2 R.

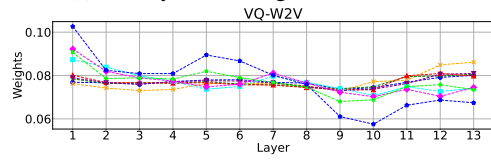(d) The layerwise weights of the W2V2.
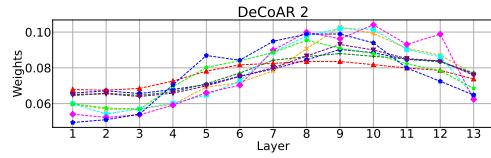
(e) The layerwise weights of the Data2Vec-A.

Figure 10: The layerwise weights analysis across 5 models.

(a) The layerwise weights of the W2V.



(b) The layerwise weights of the VQ-W2V.



(c) The layerwise weights of the DeCoAR 2.

Figure 11: The layerwise weights analysis across three models.