OPEN-EMOTION: A REPRODUCIBLE EMO-SUPERB FOR SPEECH EMOTION RECOGNITION SYSTEMS

Conference Paper · August 2024

8 authors, including:

Haibin Wu
National Taiwan University
52 PUBLICATIONS **593** CITATIONS
SEE PROFILE

Huang-Cheng Chou
National Tsing Hua University
25 PUBLICATIONS **218** CITATIONS
SEE PROFILE

Kai-Wei Chang
University of California, Los Angeles
367 PUBLICATIONS **22,685** CITATIONS
SEE PROFILE

Lucas Goncalves
University of Texas at Dallas
16 PUBLICATIONS **94** CITATIONS
SEE PROFILE

OPEN-EMOTION: A REPRODUCIBLE EMO-SUPERB FOR SPEECH EMOTION RECOGNITION SYSTEMS

Haibin Wu[1*] , *Huang-Cheng Chou*[2*], *Kai-Wei Chang*[1†], *Lucas Goncalves*[3†], *Jiawei Du*[1],
Jyh-Shing Roger Jang[1], Chi-Chun Lee[2‡], and Hung-yi Lee[1‡]

[1]National Taiwan University, Taiwan
[2]National Tsing Hua University, Taiwan
[3]The University of Texas at Dallas, USA

ABSTRACT

Speech emotion recognition (SER) is an essential technology for human-computer interaction systems. However, the previous study reveals that 80.77% of SER papers yield results that cannot be reproduced on the well-known IEMOCAP dataset. The main reason for reproducibility challenges is that the database did not provide standard data splits (e.g., train, development, and test sets). Prior papers could define its partition, but they did not provide details of the partition or source code for processing the partition. Therefore, this work aims to make SER open and reproducible to everyone. We develop the EMO-SUPERB, shorted for **EMO**tion **S**peech **U**niversal **PER**formance **B**enchmark, including a user-friendly codebase to leverage 16 state-of-the-art (SOTA) speech self-supervised learning models for exhaustive evaluation plus one SOTA SER model across 6 open-source SER datasets in English and Chinese. We make all resources open-source to facilitate future developments in SER. Researchers can easily upload their systems or datasets to EMO-SUPERB, and we name the project "Open-Emotion".

Index Terms— Speech emotion recognition, Reproducibility, Multi-label classification, Ambiguity of emotion, Subjectivity of emotion perception

1. INTRODUCTION

Speech Emotion Recognition (SER) aims to discern emotional cues from speech inputs, representing a pivotal technology for human-computer interaction systems. In recent years, significant advancements have been witnessed in SER. However, there are two unsolved problems in the SER domain:
Issue 1: The author of IEMOCAP [1], the most renowned SER dataset, has demonstrated that over 80.77% of SER, papers produce results that cannot be reproduced [2] due to the absence of released codes and implementation details.
Issue 2: Official data partitioning guidelines are lacking in most famous SER datasets, such as IEMOCAP and MSP-IMPROV [3]. 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 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 [2]. However, comparing settings with data leakage to those without it is unfair.

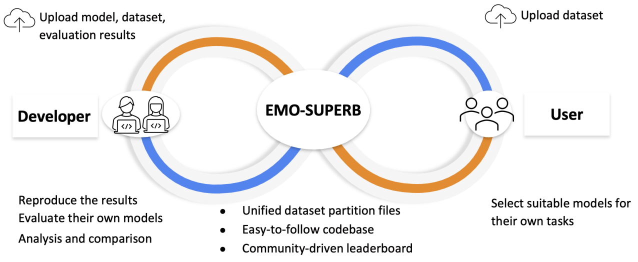

Upload model, dataset,
evaluation results
Upload dataset
Developer
EMO-SUPERB
User
Reproduce the results
Evaluate their own models
Analysis and comparison
• Unified dataset partition files
• Easy-to-follow codebase
• Community-driven leaderboard
Select suitable models for
their own tasks

Fig. 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.

To address the above issues individually, we introduce EMO-SUPERB to advance open-source initiatives in SER. The source code and complete analysis are on the project website [1]. The supplementary material file is available as well [2].

- For **Issue 1**, we develop a codebase to harness 16 SSLMs, renowned for enhancing state-of-the-art (SOTA) performance in SER, for exhaustive evaluation across all open-source SER

[*]equal first contribution,[†]equal second contribution,[‡]equal corresponding author, order is random.

[1]EMO-SUPERB Website
[2]Supplementary Material

datasets in Section 2.2. Developers can utilize a single command line to execute training and evaluation processes seamlessly, and we will release the easy-to-follow codebase.

- For **Issue 2**, we partition six open-source SER datasets and address potential data leakage issues during the partitioning process, as shown in Section 2.1.

Finally, we make processed labels and data partition files of all datasets, codes, and baseline results and their checkpoints open source to the community. We also encourage the emotion recognition community to upload their SER systems or databases to EMO-SUPERB.

2. EMO-SUPERB PLATFORM

As shown in Fig. 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 or upload their own models to the leaderboard.

2.1. Unified Dataset Partition Rules

Take the IEMOCAP corpus as an example; the database includes 5 dyadic interactions (dialogues between two speakers) involving ten speakers. In 50% of previous studies, researchers randomly divide the recordings of these ten speakers into train and test sets [2]. 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 2**).

In this study, we establish partition rules that adhere to speaker-independent criteria to mitigate the risk of leakage, which is closer to the naturalist scenarios because there are numerous speakers in the world. It is hard to do a speaker-dependent scenario in a real-life application. 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 Supplementary Material C. We provide the **standardization** of the training and testing splits and setups across the six public SER datasets.

2.2. SSLM-based Codebase

2.2.1. 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. After pre-training, one can achieve nearly SOTA performance on downstream tasks by employing the fixed SSLMs alongside task-specific lightweight prediction heads [4]. Furthermore, SSLMs significantly enhance SER and demonstrate SOTA performance, as evidenced in [5].

We develop a comprehensive codebase. The codebase depends on S3PRL [3] [4] to leverage 16 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 in Fig. 2. The six datasets adopted are SAIL-IEMOCAP [1], CREMA-D [6], MSP-IMPROV [3], MSP-PODCAST [7], BIIC-NNIME [8], and BIIC-PODCAST [9].

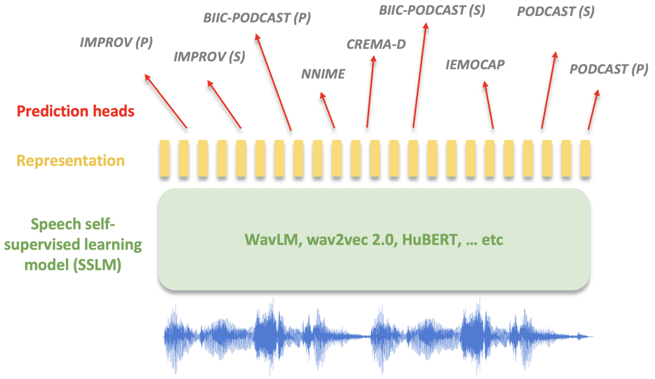


Fig. 2: Illustration of SSLM-based SER.

2.2.2. Self-supervised Learning Models

We leverage two mainstream categories of SOTA SSLMs (in S3PRL), pre-trained using generative losses and discriminative losses. We summarize them in Table 1, and details can be found in Supplementary Material B due to space limitations.

2.2.3. 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,

Model	**Loss**
Autoregressive Predictive Coding (APC) [10]	Generative loss
VQ-APC [11]	Generative loss
Non-autoregressive Predictive Coding (NPC) [12]	Generative loss
Mockingjay [13])	Generative loss
TERA [14]	Generative loss
DeCoAR 2 [15]	Generative loss
WavLM Large [16]	Discriminative loss
Hubert Large [17]	Discriminative loss
wav2vec 2.0 Large (**W2V2 Large**) [18]	Discriminative loss
wav2vec 2.0 Robustness (**W2V2 R**) [19]	Discriminative loss
Data2Vec [20]	Discriminative loss
XLS-R [21]	Discriminative loss
VQ wav2vec (**VQ-W2V**) [22] [15]	Discriminative loss
wav2vec (**W2V**) [23]	Discriminative loss
PASE+ [24]	Discriminative loss
Contrastive Predictive Coding (CPC) (**M CPC**)[25])	Discriminative loss

Table 1: Summary of SSLMs.

[3] https://github.com/s3prl/s3prl

enhancing affordability for researchers from diverse backgrounds.

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.

2.3. Community-driven Leaderboard

The leaderboard website holds significant importance within EMO-SUPERB, continuously expanding and welcoming submissions worldwide, evolving it into a dynamic benchmark 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 2.2 is not straightforward. Participants must adhere to the data partition files outlined in Section 2.1, evaluate their trained models, and submit the results. The website also offers useful visualization (e.g. radar chart Fig. C1 in Supplementary Material) 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 2.2.

2.4. 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 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.

3. EXPERIMENTAL SETUP

3.1. 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 Supplementary Material A presents detailed information, and Table A1 summarizes statistical data regarding the six emotion databases. Supplementary Material A.1 outlines the license terms and usage issues. The MSP-PODCAST and BIIC-PODCAST provide commercial licenses summarized in Table A2 in Supplementary Material. We provide details of partitions in Supplementary Material C to avoid issue 2 in Section 1, data leakage. The key information about these datasets is summarized as follows.

The SAIL-IEMOCAP [1], 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 9 emotions.

The CREMA-D [6] 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 since the paper focuses on the SER task. There are 6 emotions in total.

The MSP-IMPROV [3] 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. IMPROV (P) has 4 emotions; IMPROV (S) has 10 emotions.

The MSP-PODCAST [7] , 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. POD (P) has 8 emotions; POD (S) has 16 emotions.

The BIIC-NNIME [8], 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 utterances annotated by "other" from all annotators or by less than three annotators. All annotators watch clips in order and choose emotions from the given 12 emotion options.

BIIC-PODCAST [9], 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.

4. 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. We take the IEMOCAP as one example. The details of other datasets are in Supplementary Material C. Table 2 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 2, where each fold includes a unique combination of training, development, and test sets to ensure a comprehensive evaluation of the model's performance across different dyadic interactions within the IEMOCAP corpus.

4.1. Preprossessing

4.1.1. 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

Table 2: 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

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.

4.1.2. Selection of Emotions

Most SER prior studies [26, 2] 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 [27], though those two emotions have correlations [28].

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.

4.1.3. Label Representation

Inspired by **Semantics Space Theory** [29], we follow Chou et al. [30] to gather numerous annotations and compute a distribution-like (soft label) representation, aiming to capture the high-dimensional nature of emotion perception more accurately. 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, represented as (N, A, S, H) = (0.2, 0.4, 0.4, 0.0) for training SER systems. Additionally, to enhance SER performance, we employ the label smoothing technique proposed by [31] to refine the vector, utilizing a smoothing parameter of 0.05. This approach assigns a small probability to emotional classes with zero values.

4.2. Evaluation Metric

We use the macro-F1 score [32] to evaluate the SER performance via the Scikit-learn [33], 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 during evaluating stage, aligning with the settings employed in prior research [34, 30]. 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.

4.3. Training Details

We use the AdamW optimizer [35] with a 0.0001 learning rate, a batch size of 32, and an epoch of 200. 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 [4, 36, 37], 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 standard deviation is only less than 1% on average.

4.4. Class-balanced Cross-entropy Loss

Inspired by the study [38], we follow the study [39] to adopt the class-balanced cross-entropy loss as our primary objective function due to the imbalanced label distributions across the six databases. This approach proposed by the study [38] 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.

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

5. RESULTS AND ANALYSIS

5.1. SSLMs for SER

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

5.1.1. Overall Results

Table 3 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**. Also, **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 Large**, which typically achieves state-of-the-art results in most speech-processing tasks. Despite this, **WavLM Large** still maintains considerable strength, achieving the highest performance in three out of nine conditions. (3) To our surprise, despite its modest 90 million model parameters, the **DeCoAR 2** model outperforms the **W2V2 Large** model, which has 317 million parameters. This

Table 3: 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 Large	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 Large	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
PASE+	8	0.31740	0.456	0.521	0.274	0.233	0.292	0.178	0.349	0.306	0.248
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

finding suggests that **DeCoAR 2** could be an attractive choice for developers of SER facing computational resource constraints.

5.2. Comparison between EMO-SUPERB and SOTA

We are interested in the performance gap between the SOTA SER framework and ours, so we follow the study [30] to conduct the experiments using the model proposed by [5], fine-tuning the "Wav2Vec2-Large-Robust". Following [5], our model configuration includes adding two hidden layers, each containing 1,024 nodes, atop the modified "wav2vec2-large-robust" backbone. These layers are activated using the rectified linear unit (ReLU) activation function. A softmax output layer follows these hidden layers, providing a probabilistic distribution over the target emotion classes. Furthermore, we applied average pooling to the resulting representations, feeding it into the classification layers. We applied a dropout function with a probability of 0.5 to the first and second layers of the classification architecture to regularize the model, following the work [5]. The original code of the model is provided in [5] using HuggingFace library [40] implemented on the PyTorch [41][4]. The number of model parameters is around 317 million. Table 4 summarizes macro-F1 scores of the **SER SOTA** and the best results from the EMO-SUPERB. The average performance of the SOTA SER model [5] in macro-F1 score leads to 8.83%, absolutely better than the EMO-SUPERB. We also added the results of the SOTA SER model to the SER leaderboard. Our codebase only costs around 24 GB GPU memory during training, but the SOTA SER model needs more than 34 GB GPU memory. We will also release the results of the SER SOTA and its code files.

5.2.1. Layer analysis

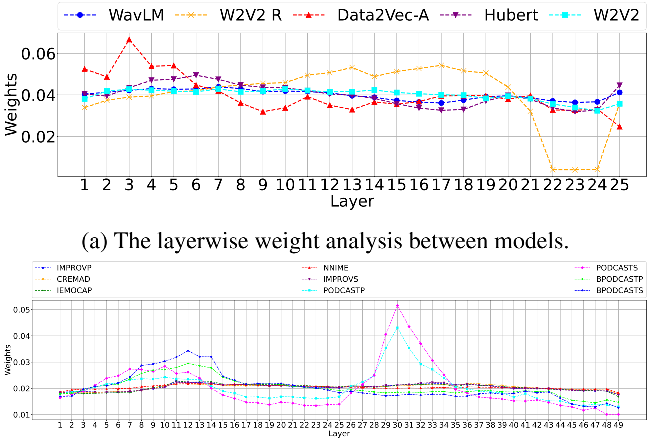
Our training strategy involves extracting features from each layer of the SSLM, multiplying these features with layer-specific weights, and then aggregating the weighted features. These aggregated features are then fed into the downstream model. Only the layer weights and the downstream model are trainable. 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 Large** (WavLM), **Hubert Large** (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. We average the layer-wise weights if emotion datasets contain multiple partitions (e.g., IEMOCAP and CREMA-D). We show the main results and additional layer-wise analysis in Supplementary Material D.1.

From the model perspective (Fig. 3a), 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 that emphasize all layers. Additionally, it's worth noting that the weights

Table 4: The table presents macro-F1 scores of the SER SOTA (denoted "SOTA") and the best results from our platform (denoted "Ours"). The absolute difference is calculated by our results minus the SOTA results. The **Model** denotes the best SSLM of our settings. The **AD** means Absolute Difference.

Dataset	SER SOTA [5]	EMO-SUPERB	AD (%)	Model
IMPROV (P)	0.646	0.559	8.70%	WavLM Large
CREMA-D	0.706	0.676	3.00%	XLS-R-1B
POD(P)	0.457	0.350	10.70%	WavLM Large
B-POD (P)	0.330	0.266	6.40%	XLS-R-1B
IEMOCAP	0.507	0.339	16.80%	W2V2 R
NNIME	0.279	0.209	7.00%	WavLM Large
IMPROV (S)	0.523	0.433	9.00%	W2V2 R
POD (S)	0.491	0.384	10.70%	XLS-R-1B
B-POD (S)	0.355	0.283	7.20%	XLS-R-1B
Average	**0.477**	**0.389**	**8.83**%	

[4] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim



(a) The layerwise weight analysis between models.

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

Fig. 3: The layerwise weights analysis.

of W2V2 R in layers 22 to 24 are considerably lower than those in other layers. We observe that the SSLMs act differently and have no clear patterns in the work.

Fig. 3b 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, two notable peak weights are observed on the 30th layer, particularly trained on the POD (P) and POD (S) datasets.

6. DISCUSSION AND LIMITATIONS

We only focused on emotion datasets in English and Chinese, omitting datasets in other languages. Also, the absence of recordings featuring elderly and child speech and 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.

7. CONCLUSION AND FUTURE WORK

We propose EMO-SUPERB, an ecosystem containing user-friendly codebases, 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 and (2) addressing data leakage. 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. Also, we plan to include more emotion datasets in other languages.

8. ACKNOWLEDGMENTS

This work was supported by NSTC under Grants 110-2634F-002-050 and 110-2221-E-007-067-MY3. We thank the anonymous reviewers and Professor Carlos Busso for their valuable comments. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

9. REFERENCES

[1] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[2] N. Antoniou et al., "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)*, 2023, pp. 1–5.

[3] C. Busso et al., "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[4] S.-W. Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[5] J. Wagner et al., "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, 2023.

[6] H. Cao et al., "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

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

[8] H.-C. Chou et al., "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 292–298.

[9] S. G Upadhyay et al., "An Intelligent Infrastructure Toward Large Scale Naturalistic Affective Speech Corpora Collection," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2023, pp. 1–8.

[10] Y.-A. Chung et al., "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Proc. Interspeech 2019*, 2019, pp. 146–150.

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

[12] Alexander H. Liu et al., "Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies," in *Proc. Interspeech 2021*, 2021, pp. 3730–3734.

[13] Andy T. Liu et al., "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6419–6423.

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

[15] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," 2020.

[16] S. Chen 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.

[17] W.-N. Hsu et al., "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.

[18] A. Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[19] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech 2021*, 2021, pp. 721–725.

[20] A. Baevski et al., "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language," in *Proceedings of the 39th International Conference on Machine Learning*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, Eds. 17–23 Jul 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312, PMLR.

[21] A. Babu et al., "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," *arXiv*, vol. abs/2111.09296, 2021.

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

[23] S. Schneider et al., "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 3465–3469.

[24] Mirco Ravanelli et al., "Multi-Task Self-Supervised Learning for Robust Speech Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6989–6993.

[25] Aaron van den Oord et al., "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] B. T. Atmaja and A. Sasou, "Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition," *IEEE Access*, vol. 10, pp. 124396–124407, 2022.

[27] Alan S. C. and Dacher K., "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proceedings of the National Academy of Sciences*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[28] C. Mogilner et al., "The Shifting Meaning of Happiness," *Social Psychological and Personality Science*, vol. 2, no. 4, pp. 395–402, 2011.

[29] A. S. Cowen and D. Keltner, "Semantic Space Theory: A Computational Approach to Emotion," *Trends in Cognitive Sciences*, vol. 25, no. 2, pp. 124–136, 2021.

[30] H.-C. Chou, L. Goncalves, S.-Gy. Leem, A. N. Salman, C.-Ch. Lee, and C. Busso, "Minority Views Matter: Evaluating Speech Emotion Classifiers with Human Subjective Annotations by an All-Inclusive Aggregation Rule," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024.

[31] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[32] J. Opitz and S. Burst, "Macro f1 and macro f1," *arXiv preprint arXiv:1911.03347*, 2019.

[33] Fabian Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] P. Riera et al., "No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, Graz, Austria, September 2019, pp. 11–15.

[35] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[36] H.-S. Tsai et al., "SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities," *arXiv preprint arXiv:2203.06849*, 2022.

[37] T.-H. Feng et al., "Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1096–1103.

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

[39] Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Chi-Chun Lee, and Carlos Busso, "The Importance of Calibration: Rethinking Confidence and Performance of Speech Multi-label Emotion Classifiers," in *Proc. INTERSPEECH 2023*, 2023, pp. 641–645.

[40] Thomas W. et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Qun Liu and David Schlangen, Eds., Online, Oct. 2020, pp. 38–45, Association for Computational Linguistics.

[41] A. Paszke et al., "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[42] A. Burmania et al., "Increasing the Reliability of Crowdsourcing Evaluations Using Online Quality Assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, 2016.

[43] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[44] Y. Li et al., “Exploration of a Self-Supervised Speech Model: A Study on Emotional Corpora,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 868–875.