Supplementary Material

A. DETAILS OF EMOTION DATASET

Table A1 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 A2 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 PODCAST 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 [1]. 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.

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 [2], we have included detailed information on the dataset splits in section 4. 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 [6], 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.

For the purpose of our study on SER, we focus exclusively on emotional annotations derived from the voice-only scenario. While many previous SER investigations utilized annotations from the audio-visual scenario as a learning target or failed to specify annotation details altogether, we opted 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 Supplementary Material C.2.

A.4. The MSP-IMPROV

The MSP-IMPROV dataset, also known as IMPROV [3], 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. 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 [42], 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. 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 Supplementary Material C.1 for the sake of clarity and consistency in subsequent analyses.

A.5. The MSP-PODCAST

The PODCAST [43] 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 [42] 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, incorporat-

Table A1: The table summarizes other detailed information about the 6 public emotion databases.

Database	**IMPROV** [3]	**CREMA-D** [6]	**PODCAST** [7]	**B-PODCAST** [9]	**IEMOCAP** [1]	**NNIME** [8]
Citation (Paper)	349	**625**	303	2	**3461**	54
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

Table A2: 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

ing 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 NNIME [8] 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 the 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 Supplementary Material C.3 to ensure reproducibility and facilitate further research using the NNIME dataset.

A.7. The BIIC-PODCAST

The B-PODCAST [9] presents a Mandarin-Chinese variant of the MSP-PODCAST, featuring audio recordings sourced from real-world podcasts under commercial licenses. 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 leverage two mainstream categories of SSLMs, pre-trained using generative loss, DeCoAR 2 [15], Autoregressive Predictive Coding (APC) [10], VQ-APC [11], Non-autoregressive Predictive Coding (NPC) [12], TERA [14], and Mockingjay [13]), and discriminative loss (**XLS-R-1B**) [21], WavLM Large [16], Hubert Large [17], wav2vec 2.0 Large (**W2V2 Large**) [18], wav2vec 2.0 Robustness (**W2V2 R**) [19], VQ wav2vec (**VQ-W2V**) [22], wav2vec (**W2V**) [23], PASE+ [24], and Contrastive Predictive Coding (CPC) (**M CPC**)[25]).

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

Here are the details about data partitions for experiments on the MSP-IMPROV (IMPROV), CREMA-D, and BIIC-NNIME (NNIME) datasets.

C.1. 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 C3. 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.

C.2. 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 C4. This partitioning strategy allows for a 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 4.

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 C4. 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 4.

D. ANALYSIS

Fig. D2 and D3 show the weights across layers, and Fig. C1 shows the performances across 4 models evaluated on 9 testing conditions.

D.1. Layer-wise Analysis

In Fig. D2, 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

Table C3: 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

Table C4: 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

CREMA-D, it is evident that earlier layers receive higher weighting when utilizing Data2Vec-A.

In Fig. D3, our observations align with those reported in the study by [44], indicating a concentration of the model's focus on the shallow layers (1st-5th). However, it is noteworthy that [44] 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 C5: 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

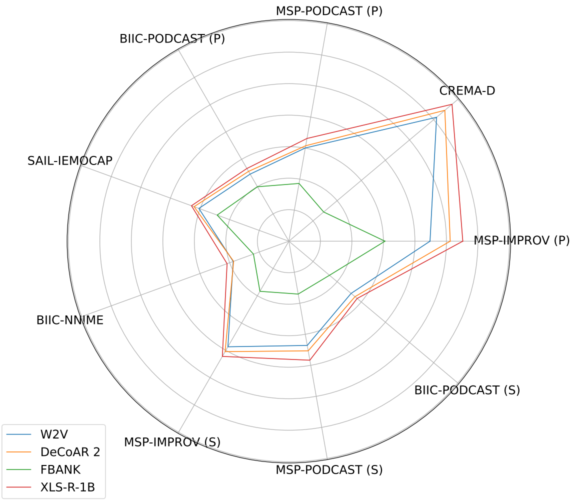
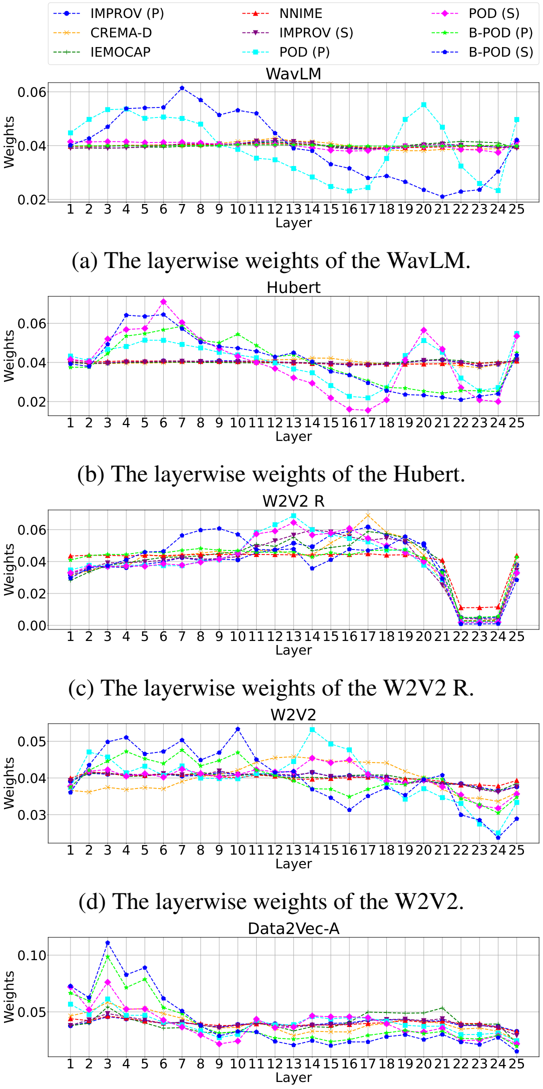


Fig. C1: Demonstration of radar chart to compare four models, W2V, DeCoAR 2, XLS-R-1B and FBANK across 9 conditions.



(a) The layerwise weights of the WavLM.

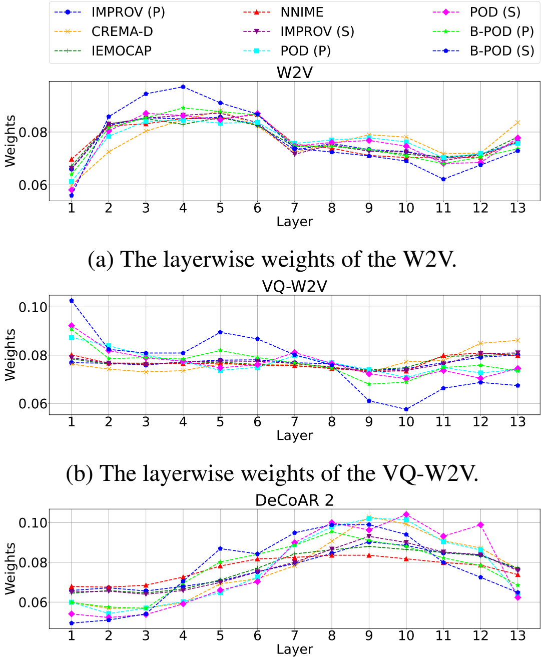
(b) The layerwise weights of the Hubert.

(c) The layerwise weights of the W2V2 R.

(d) The layerwise weights of the W2V2.

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

Fig. D2: 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.

Fig. D3: The layerwise weights analysis across three models.