Interspeech 2024
1-5 September 2024, Kos, Greece

EmoBox: Multilingual Multi-corpus Speech Emotion Recognition Toolkit and Benchmark

Ziyang Ma[1*], *Mingjie Chen*[2*], *Hezhao Zhang*[2], *Zhisheng Zheng*[1], *Wenxi Chen*[1], *Xiquan Li*[1], *Jiaxin Ye*[3], *Xie Chen*[1†], *Thomas Hain*[2†]

[1]MoE Key Lab of Artificial Intelligence, X-LANCE Lab, Shanghai Jiao Tong University, China
[2]Department of Computer Science, University of Sheffield, United Kingdom
[3]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, China

{zym.22, chenxie95}@sjtu.edu.cn, {mingjie.chen, t.hain}@sheffield.ac.uk

Abstract

Speech emotion recognition (SER) is an important part of human-computer interaction, receiving extensive attention from both industry and academia. However, the current research field of SER has long suffered from the following problems: 1) There are few reasonable and universal splits of the datasets, making comparing different models and methods difficult. 2) No commonly used benchmark covers numerous corpus and languages for researchers to refer to, making reproduction a burden. In this paper, we propose EmoBox[1], an out-of-the-box multilingual multi-corpus speech emotion recognition toolkit, along with a benchmark for both intra-corpus and cross-corpus settings. For intra-corpus settings, we carefully designed the data partitioning for different datasets. For cross-corpus settings, we employ a foundation SER model, emotion2vec, to mitigate annotation errors and obtain a test set that is fully balanced in speakers and emotions distributions. Based on EmoBox, we present the intra-corpus SER results of 10 pre-trained speech models on 32 emotion datasets with 14 languages, and the cross-corpus SER results on 4 datasets with the fully balanced test sets. To the best of our knowledge, this is the largest SER benchmark, across language scopes and quantity scales. We hope that our toolkit and benchmark can facilitate the research of SER in the community.
Index Terms: speech emotion recognition, toolkit, benchmark, cross-corpus

1. Introduction

In the realm of human-computer interaction (HCI), the ability of machines to understand and respond to human emotions through speech has emerged as a pivotal area of research, known as Speech Emotion Recognition (SER). The significance of SER extends across a wide array of applications, from enhancing user experience in virtual assistants [1] to facilitating emotional well-being in healthcare services [2]. Despite its growing importance, the field of SER faces persistent challenges that hinder progress and innovation. Among these challenges are the scarcity of universally accepted dataset splits [3] and the absence of a comprehensive benchmark encompassing a diverse range of corpora and languages [4]. These limitations complicate the comparison of models and methods, as well as the replication of research findings, thus impeding the advancement of SER technology.

Recognizing existing critical gaps, this paper introduces EmoBox, a groundbreaking multilingual multi-corpus speech emotion recognition toolkit designed to streamline research in this field. EmoBox is accompanied by a meticulously curated benchmark tailored for both intra-corpus and cross-corpus evaluation settings. For intra-corpus evaluations, we have devised a systematic approach to data partitioning across various datasets, ensuring that researchers can conduct rigorous and comparable analyses of different SER models. In the cross-corpus context, we leverage a foundational SER model, emotion2vec [5], to address annotation discrepancies and create a test set that achieves a balance in speaker and emotion distribution, a feat previously unattained in SER research.

Table 1: *The datasets involved in EmoBox at a glance.*

Dataset	Source	Lang	Emo	Spk	#Utts	#Hrs
AESDD [6]	Act	Greek	5	5	604	0.7
ASED [7]	Act	Amharic	5	65	2474	2.1
ASVP-ESD [8]	Media	Mix	12	131	13964	18.0
CaFE [9]	Act	French	7	12	936	1.2
CASIA [10]	Act	Mandarin	6	4	1200	0.6
CREMA-D [11]	Act	English	6	91	7442	5.3
EMNS [12]	Act	English	8	1	1181	1.9
EmoDB [13]	Act	German	7	10	535	0.4
EmoV-DB [14]	Act	English	5	4	6887	9.5
EMOVO [15]	Act	Italian	7	6	588	0.5
Emozionalmente [16]	Act	Italian	7	431	6902	6.3
eNTERFACE [17]	Act	English	6	44	1263	1.1
ESD [18]	Act	Mix	5	20	35000	29.1
IEMOCAP [19]	Act	English	5	10	5531	7.0
JL-Corpus [20]	Act	English	5	4	2400	1.4
M3ED [21]	TV	Mandarin	7	626	24437	9.8
MEAD [22]	Act	English	8	48	31729	37.3
MELD [23]	TV	English	7	304	13706	12.1
MER2023 [24]	TV	Mandarin	6	/	5030	5.9
MESD [25]	Act	Spanish	6	11	862	0.2
MSP-Podcast [26]	Podcast	English	8	1273	73042	113.6
Oreau [27]	Act	French	7	32	434	0.3
PAVOQUE [28]	Act	German	5	1	7334	12.2
Polish [29]	Act	Polish	3	5	450	0.1
RAVDESS [30]	Act	English	8	24	1440	1.5
RESD [31]	Act	Russian	7	200	1396	2.3
SAVEE [32]	Act	English	7	4	480	0.5
ShEMO [33]	Act	Persian	6	87	2838	3.3
SUBESCO [34]	Act	Bangla	7	20	7000	7.8
TESS [35]	Act	English	7	2	2800	1.6
TurEV-DB [36]	Act	Turkish	4	6	1735	0.5
URDU [37]	Talk show	Urdu	4	29	400	0.3
Total	–	–	–	3510	262020	294.4

Our contributions are manifold. Not only do we offer the SER community a powerful toolkit to easily conduct experiments on different datasets, but we also establish a new benchmark for the field. We detail our data partitioning, which we believe reduces the burden on researchers for future research. We present comprehensive intra-corpus SER results derived from the application of 10 pre-trained speech models across 32 emotion datasets in 14 languages. To our knowledge, this represents the most extensive SER benchmark to date, spanning the broadest scope of languages and the largest scale of data. Besides, we showcase the cross-corpus SER performance on 4 datasets, utilizing test sets that are fully balanced in terms of speakers and emotions. Through EmoBox, we aim to provide the SER community with a robust toolkit and benchmark that will catalyze further research, enhance model comparability, and ultimately, foster innovation in the field of speech emotion recognition.

Co-first author*. Corresponding author†.
[1]https://github.com/emo-box/EmoBox

10.21437/Interspeech.2024-788

2. Data Preparation and Partitioning

2.1. Datasets

The comprehensive overview of the datasets utilized in this study is delineated in Table 1. There are 32 emotional datasets spanning 14 distinct languages, comprising 12 in English, 3 in Mandarin, and 2 in French, German, and Italian. Additionally, there is 1 dataset each in Amharic, Bangla, Greek, Persian, Polish, Russian, Spanish, Turkish, and Urdu, as well as two datasets featuring a mixture of languages.

For analytical purposes, each dataset is systematically classified according to several criteria: ***Source*** denotes the origin of the samples; ***Lang*** indicates the language of the dataset; ***Emo*** represents the number of emotional categories encompassed; ***Spk*** specifies the number of speakers; ***#Utts*** details the total number of utterances; and ***#Hrs*** quantifies the aggregate hours of the samples.

The speech data extracted from these datasets undergoes a uniform processing protocol, being converted into a monophonic format with a sampling rate of $16,000$ Hz. Each piece of speech data is uniquely annotated with an emotion label, ensuring a precise correlation between the utterance and its emotional categorization.

2.2. Intra-corpus SER

Ensuring proper data partitioning is pivotal for leveraging a corpus efficiently, particularly when dealing with corpora of limited size. Through meticulous observation and analysis of the distribution of speakers and emotions across the 32 datasets, as detailed in Section 2.1, we establish a set of criteria for data partitioning as follows:

1. Each dataset is divided into training and testing sets, with the division possibly encompassing single or multiple folds depending on the data distribution.
2. In cases where datasets come with officially predefined splits, these original partitions are adhered to. For instance, the IEMOCAP dataset is organized into 5 folds, featuring 2 distinct speakers per fold, whereas the MELD dataset is split into train, dev, and test splits.
3. For datasets characterized by a speaker count of fewer than 4, such as the PAVOQUE dataset which includes only a single speaker; or those with 4 or more speakers but exhibit an imbalanced distribution of emotions among speakers, such as the M3ED dataset, a speaker-dependent approach is employed. Here, 25% of data for each emotion is earmarked for testing, with the remainder allocated for training.
4. For datasets whose speaker number is greater than or equal to 4 with a balanced emotion distribution among speakers, the leave-one-out n-fold cross-validation manner is adopted. More specifically, if the number of speakers $\in \{4, 5, 6\}$, n is set the same as the number of speakers; If the number of speakers exceeds 6, n is taken to be 4 and multiple speakers are merged within each fold.

The data partitioning of the aforementioned criteria is conducted in EmoBox and model performance on the benchmark is thoroughly examined in Section 3.2. This structured approach to data partitioning underscores our commitment to fostering robust and replicable research methodologies within the field.

2.3. Cross-corpus SER

In real scenarios, the ability of a SER model to generalize to unseen speakers and unknown recording conditions is essential. To evaluate this capability, cross-corpus zero-shot testing emerges as a profitable strategy to assess the robustness of SER models against variability in speakers and recording contexts. To implement this approach, we meticulously select 4 datasets: IEMOCAP, MELD, RAVDESS, and SAVEE. As shown in Table 2, these datasets cover a diverse range of sources, accents, and recording environments. Such diversity is crucial for the cross-corpus setting as it ensures the testing encompasses a broad spectrum of real-world scenarios, thereby providing a comprehensive assessment of models' adaptability and robustness.

Table 2: *Detailed meta information of different datasets for cross-corpus settings.*

	Source	Accent	Recording
IEMOCAP	Elicitation	English	Crosstalk
MELD	Spontaneousness	English	Noisy
RAVDESS	Act	North American	Clean
SAVEE	Act	British	Clean

Table 3: *Balanced test sets statistical information. Each corpus contains 240 pieces of test data, with 4 identical emotions including angry, happy, neutral, and sad.*

	#Emotion	#Speaker	#Number	#Total
IEMOCAP	4	10	6	240
MELD	4	6	10	240
RAVDESS	4	20	3	240
SAVEE	4	4	15	240

To address potential errors in annotation, we leverage the fine-tuned version[2] of the emotion2vec [5] model, a foundation speech emotion recognition model with iterative fine-tuning on over $10,000$ hours of speech data. Our methodology involves an initial application of emotion2vec to assign pseudo-labels to our datasets. Subsequently, we refine our dataset by retaining only those instances where there is congruence between the original annotations and those generated by emotion2vec. This step ensures the reduction of annotation discrepancies and enhances the reliability for further analysis. To establish a fully balanced test set across each dataset, we select 240 speech-emotion pairs for each dataset. Shared emotions among these datasets contain angry, happy, neutral, and sad, resulting in 60 pieces for each emotion. The composition of the test set, including the number of speakers and the allocation of speech-emotion pairs per speaker for each emotion, is detailed in Figure 3. For the IEMOCAP and SAVEE datasets, all speakers are encompassed, including 5 male and 5 female speakers in IEMOCAP, and 4 male speakers in SAVEE. In the case of the MELD dataset, we focus on the 6 protagonists. For the RAVDESS dataset, we include a selection of 20 speakers, equally divided between 10 male and 10 female speakers. This meticulous composition of the test set is instrumental in our analysis of the models' generalization capabilities across diverse corpora, helping our assessment of the robustness and adaptability of models under cross-corpus settings.

3. Benchmark

3.1. Experiments Setup

10 pre-trained models are employed to establish the benchmark of EmoBox, including self-supervised wav2vec 2.0 [38] base, HuBERT [39] base/large, WavLM [40] base/large, data2vec [41] base/large, data2vec 2.0 [42] base/large, and an additional supervised ASR encoder of Whisper [43] large v3.

[2] https://github.com/ddlBoJack/emotion2vec

Table 4: *Intra-corpus SER results of 10 pre-trained speech models on 32 emotion datasets spanning 14 distinct languages with EmoBox data partitioning. Unweighted Average Accuracy (UA(%)), Weighted Average Accuracy (WA(%)), and Macro F1 Score (F1(%)) are reported, with Top1, Top2, and Top3 scores highlighted.*

Model	UA(%) ↑	WA(%) ↑	F1(%) ↑	UA(%) ↑	WA(%) ↑	F1(%) ↑	UA(%) ↑	WA(%) ↑	F1(%) ↑	UA(%) ↑	WA(%) ↑	F1(%) ↑
	AESDD (el)			**ASED (am)**			**ASVP-ESD (mix)**			**CaFE (fr)**		
wav2vec 2.0 base	67.59	67.65	67.61	88.59	88.63	88.63	48.99	59.12	49.78	42.76	42.47	41.03
HuBERT base	82.30	82.35	82.36	94.17	94.13	94.13	48.69	59.78	49.72	54.16	54.16	53.36
HuBERT large	78.85	78.90	78.88	96.19	96.19	96.17	53.33	63.00	54.14	59.50	58.73	58.22
WavLM base	78.99	79.08	78.78	94.27	94.31	94.29	46.38	58.05	47.35	52.71	52.33	51.66
WavLM large	84.40	84.49	84.19	96.44	96.45	96.42	56.31	65.91	56.83	62.20	61.33	61.14
data2vec base	49.96	50.03	49.30	86.39	86.34	86.34	37.66	50.79	38.26	42.18	42.36	41.69
data2vec large	45.63	45.69	44.65	88.31	88.31	88.30	46.95	56.36	47.33	42.24	42.85	41.11
data2vec 2.0 base	46.55	46.68	45.33	93.99	93.99	93.98	46.00	57.57	46.62	51.83	51.67	50.52
data2vec 2.0 large	72.26	72.34	71.82	94.30	94.28	94.27	52.18	62.35	52.64	59.04	58.02	57.51
Whisper large v3	79.13	79.18	79.13	96.75	96.73	96.74	61.14	71.52	62.08	69.43	68.84	68.06
Model	**CASIA (zh)**			**CREMA-D (en)**			**EMNS (en)**			**EmoDB (de)**		
wav2vec 2.0 base	39.56	39.56	34.86	61.95	61.90	61.75	65.14	65.27	64.80	82.06	83.14	82.21
HuBERT base	47.23	47.23	42.47	71.13	70.98	71.00	75.83	76.25	75.70	87.73	87.73	87.82
HuBERT large	45.30	45.30	39.10	73.83	73.64	73.73	73.94	74.28	73.67	89.81	90.26	89.86
WavLM base	47.25	47.25	41.78	69.64	69.49	69.54	69.46	69.71	69.24	87.03	87.12	86.76
WavLM large	52.12	52.12	46.55	74.50	74.32	74.39	83.97	84.12	83.97	92.58	92.67	92.57
data2vec base	34.72	34.72	30.88	58.03	57.78	57.73	34.33	34.58	33.12	58.12	60.01	58.32
data2vec large	37.65	37.65	33.50	63.80	63.51	63.48	48.52	48.96	48.39	60.96	61.95	61.26
data2vec 2.0 base	43.31	43.31	38.90	65.74	65.48	65.47	47.83	48.39	47.21	75.07	75.86	75.49
data2vec 2.0 large	45.57	45.57	41.46	69.55	69.27	69.25	57.80	58.60	57.15	79.36	80.41	79.96
Whisper large v3	59.58	59.58	56.27	76.75	76.48	76.60	69.73	70.58	69.31	91.26	92.43	91.84
Model	**EmoV-DB (en)**			**EMOVO (it)**			**Emozionalmente (it)**			**eNTERFACE (en)**		
wav2vec 2.0 base	97.85	98.03	97.90	31.07	31.07	27.24	56.69	56.69	56.64	64.19	64.12	63.81
HuBERT base	98.72	98.77	98.71	46.10	46.10	41.28	66.30	66.30	66.26	79.14	79.11	78.97
HuBERT large	99.36	99.40	99.37	45.74	45.74	40.56	69.83	69.83	69.81	88.19	88.17	88.14
WavLM base	98.38	98.49	98.39	42.39	42.39	37.33	63.02	63.02	63.02	88.30	88.27	88.20
WavLM large	99.44	99.47	99.45	48.82	48.82	44.16	75.00	75.00	74.97	92.43	92.42	92.40
data2vec base	93.26	93.61	93.23	32.47	32.47	29.22	48.97	48.97	48.77	91.61	91.62	91.64
data2vec large	94.47	94.93	94.58	35.66	35.66	33.59	53.92	53.92	53.66	89.46	89.46	89.65
data2vec 2.0 base	95.81	96.09	95.80	42.96	42.96	41.01	56.22	56.22	56.00	90.81	90.80	90.83
data2vec 2.0 large	98.17	98.32	98.19	45.88	45.88	43.63	64.10	64.10	63.93	94.02	94.02	94.01
Whisper large v3	99.36	99.37	99.34	57.82	57.82	56.06	76.91	76.91	76.90	97.69	97.68	97.68
Model	**ESD (mix)**			**IEMOCAP (en)**			**JL-Corpus (en)**			**M3ED (zh)**		
wav2vec 2.0 base	69.17	69.27	68.66	58.27	57.73	57.83	45.27	45.27	40.73	23.13	43.20	22.91
HuBERT base	72.41	72.41	72.11	63.87	63.10	63.45	51.11	51.11	50.56	23.80	42.55	24.03
HuBERT large	75.85	75.85	75.38	67.42	66.69	67.24	56.62	56.62	52.77	23.25	44.49	23.28
WavLM base	72.90	72.90	72.55	62.92	63.94	63.40	53.79	53.79	52.36	22.76	42.79	22.03
WavLM large	79.14	79.14	78.87	69.47	69.07	69.29	60.86	60.86	57.34	26.58	44.86	26.98
data2vec base	65.05	65.05	64.55	54.19	53.20	53.76	49.29	49.29	47.86	19.44	37.32	19.24
data2vec large	72.01	72.01	71.77	52.56	51.11	51.71	48.87	48.87	46.92	20.20	38.73	20.26
data2vec 2.0 base	73.40	73.40	73.10	54.40	53.19	53.71	54.04	54.04	52.85	22.82	41.42	22.89
data2vec 2.0 large	76.81	76.81	76.43	57.30	56.23	56.70	65.14	65.14	63.49	23.82	43.02	23.98
Whisper large v3	84.62	84.62	84.33	73.54	72.86	73.11	66.71	66.71	65.19	32.84	49.42	33.76
Model	**MEAD (en)**			**MELD (en)**			**MER2023 (zh)**			**MESD (es)**		
wav2vec 2.0 base	72.17	73.57	72.32	20.06	45.17	20.04	40.40	46.78	40.73	62.93	62.89	62.85
HuBERT base	74.76	75.71	74.92	23.53	45.47	24.29	42.56	49.80	42.77	47.52	47.48	46.33
HuBERT large	76.87	77.84	77.12	24.13	46.37	24.99	43.96	50.49	44.45	53.71	53.67	53.67
WavLM base	71.85	72.86	72.14	23.44	44.71	24.25	41.80	48.71	41.97	43.58	43.52	42.94
WavLM large	81.27	82.03	81.43	28.18	49.31	29.11	48.17	54.77	49.36	62.54	62.49	62.33
data2vec base	65.36	66.05	65.53	23.82	45.57	24.37	37.94	43.06	38.15	34.37	34.35	33.24
data2vec large	67.57	68.44	67.88	23.35	45.74	24.10	35.14	40.27	34.28	36.67	36.61	35.81
data2vec 2.0 base	69.66	70.64	69.90	24.79	46.65	25.28	42.59	46.64	43.22	44.86	44.85	43.6
data2vec 2.0 large	74.13	75.24	74.43	26.33	47.72	27.35	42.05	46.81	42.08	48.46	48.45	46.8
Whisper large v3	76.35	77.34	76.55	31.54	51.89	32.95	61.22	65.23	62.29	69.78	69.67	69.64
Model	**MSP-Podcast (en)**			**Oreau (fr)**			**PAVOQUE (de)**			**Polish (pl)**		
wav2vec 2.0 base	15.5	37.48	13.8	40.14	41.06	39.23	84.95	91.23	86.09	67.35	67.35	66.76
HuBERT base	16.97	38.7	15.89	51.98	52.80	51.89	86.12	92.19	87.06	69.40	69.40	69.08
HuBERT large	18.07	40.02	17.2	63.69	64.35	63.66	87.04	92.76	87.94	70.20	70.20	70.74
WavLM base	17.11	37.38	16.53	57.78	58.06	57.45	83.73	90.84	85.60	69.31	69.31	69.46
WavLM large	18.6	40.47	17.97	65.54	66.29	65.67	87.73	93.40	88.43	79.29	79.29	79.02
data2vec base	16.19	37.15	15.49	54.84	55.67	54.76	74.94	85.11	76.63	71.31	71.31	70.65
data2vec large	17.24	37.45	16.86	48.29	49.12	48.21	73.26	84.89	75.71	68.05	68.05	67.07
data2vec 2.0 base	16.79	39.97	15.33	59.40	59.73	58.13	78.3	87.82	80.92	75.75	75.75	75.52
data2vec 2.0 large	18.35	41.33	17.07	64.64	64.88	64.50	85.38	92.07	86.75	74.00	74.00	74.05
Whisper large v3	22.24	44.1	22.12	84.48	84.79	85.01	87.72	93.17	88.41	83.27	83.27	82.77
Model	**RAVDESS (en)**			**RESD (ru)**			**SAVEE (en)**			**ShEMO (fa)**		
wav2vec 2.0 base	54.33	55.40	53.99	52.82	53.39	52.90	44.89	49.83	42.07	56.34	78.96	57.34
HuBERT base	65.43	66.21	65.31	51.56	52.34	51.65	58.90	59.85	64.05	58.31	81.17	63.15
HuBERT large	70.00	70.29	69.54	50.82	51.51	50.74	71.91	75.05	71.83	64.29	83.35	66.26
WavLM base	61.56	62.10	61.18	44.81	45.09	44.97	63.57	67.05	62.83	60.73	78.76	62.06
WavLM large	72.00	72.22	71.42	55.87	56.47	55.82	66.74	70.80	66.37	71.72	87.13	73.55
data2vec base	51.92	52.22	51.10	37.09	37.90	36.86	75.65	78.25	78.38	47.61	70.07	49.49
data2vec large	59.30	59.50	58.74	30.78	31.57	30.81	68.01	71.45	71.08	56.42	74.09	60.98
data2vec 2.0 base	64.66	64.84	64.18	43.54	44.03	43.00	72.65	75.50	75.59	60.59	79.03	64.03
data2vec 2.0 large	71.15	71.63	70.94	44.08	44.64	44.25	75.75	78.59	78.24	64.09	82.68	68.47
Whisper large v3	75.32	75.87	75.19	54.98	55.54	54.99	74.07	77.24	75.31	80.23	89.55	82.94
Model	**SUBESCO (bn)**			**TESS (en)**			**TurEV-DB (tr)**			**URDU (ur)**		
wav2vec 2.0 base	51.25	51.25	50.91	97.92	97.92	97.92	67.19	68.01	67.19	87.50	87.50	87.57
HuBERT base	57.89	57.89	57.64	99.62	99.62	99.62	72.81	73.26	72.89	88.41	88.41	88.40
HuBERT large	64.53	64.53	64.33	99.86	99.86	99.86	70.93	72.08	70.96	81.75	81.75	81.66
WavLM base	57.06	57.06	56.78	99.10	99.10	99.10	69.97	70.51	70.18	82.82	82.82	82.85
WavLM large	65.33	65.33	65.09	99.78	99.78	99.78	79.5	80.09	79.51	86.61	86.61	86.64
data2vec base	46.22	46.22	45.80	94.67	94.67	94.65	51.62	52.58	51.52	65.42	65.42	65.35
data2vec large	53.06	53.06	52.62	96.89	96.89	96.89	54.69	55.11	54.56	64.81	64.81	64.93
data2vec 2.0 base	57.97	57.97	57.69	98.55	98.55	98.55	63.44	63.69	63.37	69.97	69.97	69.91
data2vec 2.0 large	66.45	66.45	66.25	99.54	99.54	99.54	63.77	64.13	63.00	78.10	78.10	78.10
Whisper large v3	73.05	73.05	72.94	99.96	99.96	99.96	81.32	81.58	81.31	82.52	82.52	82.41

Speaking of the parameters, the base SSL model is about 95M, the large SSL model is about 315M, and the Whisper encoder is the largest, for about 635M. We extract features of the pre-trained models from the last Transformer layer [3] and conduct layer normalization uniformly to speed up convergence. The downstream networks are a simple linear hidden layer and a classification head with a ReLU activation function and a pooling layer sandwiched between them. For each dataset and each model, we sweep the learning rate $\in \{1e-3, 1e-4\}$ and hidden layer dimensions $\in \{128, 256\}$. The optimal settings, determined by the best performance outcomes, are then selected for presentation. In each training, 20% of data from the training set is selected as the validation set. All experiments are trained for 100 epochs, with the first 10 epochs to warm up to the maximum learning rate. Three key performance metrics are reported: Unweighted Average Accuracy (UA), Weighted Average Accuracy (WA), and Macro F1 Score in the evaluation of the benchmark of EmoBox.

3.2. Intra-corpus SER Results

Table 4 presents intra-corpus SER results of selected 10 pre-trained speech models on 32 emotion datasets spanning 14 distinct languages, with EmoBox data partitioning. The experimental results show that the Whisper large v3 encoder performs significantly better than other SSL models, ranking top1 on $23/32$ datasets and top3 on $30/32$ datasets. The possible reason is that whisper large v3 is trained with speech from multiple languages with data and model scaling. Except for Whisper large v3 encoder, WavLM large performs best, ranking top3 on $31/32$ datasets, followed by HuBERT large with $14/32$ and data2vec 2.0 large with $12/32$. All three models are pre-trained on English speech data, while WavLM employs more data as well as introduces noise in SSL training to enhance robustness. This suggests that SSL features trained with more data can enhance the performance of SER. Researchers can explore more about emotion in speech with EmoBox benchmark, such as linguistic correlation and difference.

3.3. Cross-corpus SER Results

Table 5 presents cross-corpus SER results of selected 10 pre-trained speech models on the refined EmoBox test sets. As seen from the table, Whisper large v3 encoder still performs best on the cross-corpus settings, with top1 on $9/12$ train-test pairs, while HuBERT large, WavLM large and data2vec base takes 1 each, respectively.

Figure 1 illustrates the average accuracy of different models with cross-corpus settings. The formula for calculating the average accuracy $\overline{Acc}$ is given as follows:

$$\overline{Acc} = \frac{1}{n^2 - n} \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} Acc_{i,j}, \quad (1)$$

where $Acc_{i,j}$ denotes training with dataset i and testing on dataset j, and $n = 4$ is in our settings. As seen from Figure 1, Except for Whisper large v3 encoder, WavLM performs best among large model sizes and HuBERT base performs best among base model sizes.

4. Conclusion, Limitation and Future Work

EmoBox offers an easy-to-use toolkit for multilingual multi-corpus SER research with data preparation and partitioning, and the largest benchmark for intra-corpus and cross-corpus evaluations to date. We invite the research community to adopt EmoBox toolkit and evaluate EmoBox benchmark for advancing SER methodologies, thereby contributing to the development of more emotionally intelligent human-computer interactions. Due to the huge workload, we only evaluate the last layer of features from the pretrained model. We will design more comprehensive evaluations in the future and give more instructive generalization conclusions to the SER community.

[3] We found that the last layer features from wav2vec 2.0 large fail to conduct SER task on the most corpus, so we did not report them.

Table 5: *Accuracy (%) with features from different pre-training models for cross-corpus settings, where the horizontal direction represents the training sets, while the vertical direction represents the test sets.* ***I***, ***M R***, ***S*** *stand for IEMOCAP, MELD, RAVDESS and SAVEE, respectively.* ***BOLD*** *indicates the best results for each train-test pair among 10 pre-trained models.*

	I	M	R	S	I	M	R	S
	wav2vec 2.0 base				**Whisper large v3**			
I	\	29.78	18.25	28.84	\	46.14	38.24	**46.12**
M	22.50	\	31.39	35.24	**51.42**	\	**47.00**	36.44
R	27.15	23.20	\	33.77	**48.12**	**40.68**	\	**66.91**
S	31.34	29.19	21.36	\	**49.30**	**42.18**	**49.63**	\
	HuBERT base				**HuBERT large**			
I	\	37.32	22.63	35.88	\	44.60	15.03	39.99
M	38.31	\	31.60	32.67	44.69	\	38.22	**43.74**
R	42.00	33.43	\	43.47	36.18	25.02	\	56.96
S	39.39	29.03	38.41	\	42.81	31.54	31.92	\
	WavLM base				**WavLM large**			
I	\	38.25	27.80	39.40	\	**48.59**	34.16	35.53
M	46.30	\	21.38	35.75	39.06	\	23.06	25.74
R	30.78	29.58	\	43.24	44.03	33.90	\	63.35
S	34.00	27.40	27.38	\	43.69	34.10	36.69	\
	data2vec base				**data2vec large**			
I	\	43.86	**40.46**	32.53	\	44.99	39.03	36.88
M	42.57	\	24.16	24.33	40.62	\	29.10	31.57
R	22.28	20.32	\	27.02	26.50	26.73	\	32.54
S	29.41	31.96	34.68	\	26.82	24.05	16.96	\
	data2vec 2.0 base				**data2vec 2.0 large**			
I	\	44.96	30.52	31.40	\	47.43	17.80	29.67
M	42.35	\	33.32	30.42	41.75	\	31.80	29.66
R	30.77	26.80	\	37.31	38.79	34.21	\	35.43
S	29.12	35.29	19.24	\	36.39	37.79	23.58	\

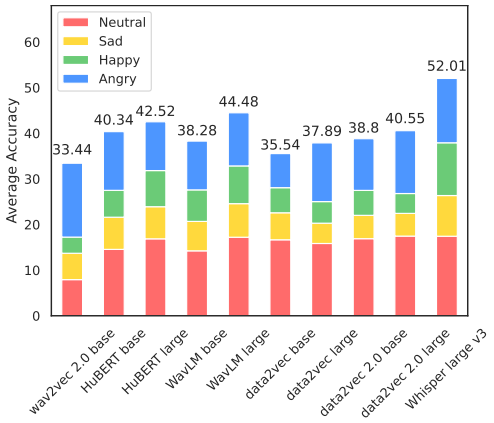

Figure 1: *Average accuracy for cross-corpus settings.*

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62206171 and No. U23B2018), Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102 and the International Cooperation Project of PCL. This work was also supported by Liveperson, Inc. and conducted at the Voicebase/Liveperson Centre of Speech and Language Technology at the University of Sheffield.

6. References

[1] G.-T. Lin, C.-H. Chiang, and H.-y. Lee, "Advancing large language models to capture varied speaking styles and respond properly in spoken conversations," in *Proc. ACL*, 2024.

[2] W. Wu, M. Wu, and K. Yu, "Climate and weather: Inspecting depression detection via emotion recognition," in *Proc. ICASSP*, 2022.

[3] N. Antoniou, A. Katsamanis, T. Giannakopoulos, and S. Narayanan, "Designing and evaluating speech emotion recognition systems: A reality check case study with IEMOCAP," in *Proc. ICASSP*, 2023.

[4] N. Scheidwasser-Clow, M. Kegler, P. Beckmann, and M. Cernak, "Serab: A multi-lingual benchmark for speech emotion recognition," in *Proc. ICASSP*, 2022.

[5] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Proc. ACL Findings*, 2024.

[6] N. Vryzas, R. Kotsakis, A. Liatsou, C. A. Dimoulas, and G. Kalliris, "Speech emotion recognition for performance interaction," in *Proc. AES*, 2018.

[7] E. A. Retta, E. Almekhlafi, R. Sutcliffe, M. Mhamed, H. Ali, and J. Feng, "A new Amharic speech emotion dataset and classification benchmark," in *Proc. TALLIP*, 2023.

[8] T. Landry Dejoli, Q. He, H. Yan, and Y. Li, "ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances," in *Proc. GSJ*, 2020.

[9] P. Gournay, O. Lahaie, and R. Lefebvre, "A Canadian French emotional speech dataset," in *Proc. ACM Multimedia*, 2018.

[10] J. Tao, F. Liu, M. Zhang, and H. Jia, "Design of speech corpus for Mandarin text to speech," in *The Blizzard Challenge Workshop*, 2008.

[11] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," in *Proc. TAC*, 2014.

[12] K. A. Noriy, X. Yang, and J. J. Zhang, "EMNS/Imz/Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels," in *arXiv preprint*, 2023.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of German emotional speech," in *Proc. Interspeech*, 2005.

[14] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," in *arXiv preprint*, 2018.

[15] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in *Proc. LREC*, 2014.

[16] F. Catania, "Speech emotion recognition in Italian using Wav2Vec 2," in *Authorea Preprints*, 2023.

[17] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. ICDE Workshop*, 2006.

[18] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*, 2021.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," in *Proc. LREC*, 2008.

[20] J. James, L. Tian, and C. Watson, "An open source emotional speech corpus for human robot interaction applications," *Proc. Interspeech*, 2018.

[21] J. Zhao, T. Zhang, J. Hu, Y. Liu, Q. Jin, X. Wang, and H. Li, "M3ED: Multi-modal multi-scene multi-label emotional dialogue database," in *Proc. ACL*, 2022.

[22] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. ECCV*, 2020.

[23] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL*, 2019.

[24] Z. Lian, H. Sun, L. Sun, K. Chen, M. Xu, K. Wang, K. Xu, Y. He, Y. Li, J. Zhao *et al.*, "MER 2023: Multi-label learning, modality robustness, and semi-supervised learning," in *Proc. ACM Multimedia*, 2023.

[25] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The Mexican emotional speech database (MESD): elaboration and assessment based on machine learning," in *Proc. EMBC*, 2021.

[26] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Proc. Interspeech*, 2020.

[27] L. KERKENI, C. CLEDER, S.-R. Youssef, and K. RAOOF, "French emotional speech database-oréau," 2020.

[28] I. Steiner, M. Schröder, and A. Klepp, "The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech," in *Proc. Phonetik & Phonologie*, 2013.

[29] M. Miesikowska and D. Swisulski, "Emotions in Polish speech recordings," https://doi.org/10.34808/h46c-hb44, 2020.

[30] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," in *Proc. PloS One*, 2018.

[31] I. Lubenets, N. Davidchuk, and A. Amentes, "Aniemore." [Online]. Available: https://github.com/aniemore/Aniemore

[32] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (SAVEE) database." University of Surrey, 2014.

[33] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection," in *Proc. LREC*, 2019.

[34] S. Sultana, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla," in *Proc. PloS One*, 2021.

[35] K. Dupuis and M. K. Pichora-Fuller, "Toronto emotional speech set (TESS) - younger talker_happy." University of Toronto, 2010.

[36] S. F. Canpolat, Z. Ormanoğlu, and D. Zeyrek, "Turkish Emotion Voice Database (TurEV-DB)," in *Proc. SLTU Workshop*, 2020.

[37] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *Proc. FIT*, 2018.

[38] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[39] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," in *Proc. TASLP*, 2021.

[40] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," in *Proc. JSTSP*, 2022.

[41] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. ICML*, 2022.

[42] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *Proc. IMCL*, 2023.

[43] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.