emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation

Ziyang Ma[1], Zhisheng Zheng[1], Jiaxin Ye[2], Jinchao Li[3], Zhifu Gao[4], Shiliang Zhang[4], Xie Chen[1]†
[1] Shanghai Jiao Tong University, [2] Fudan University,
[3] The Chinese University of Hong Kong, [4] Alibaba

Abstract

We propose emotion2vec, a universal speech emotion representation model. emotion2vec is pre-trained on open-source unlabeled emotion data through self-supervised online distillation, combining utterance-level loss and frame-level loss during pre-training. emotion2vec outperforms state-of-the-art pre-trained universal models and emotion specialist models by only training linear layers for the speech emotion recognition task on the mainstream IEMOCAP dataset. In addition, emotion2vec shows consistent improvements among 10 different languages of speech emotion recognition datasets. emotion2vec also shows excellent results on other emotion tasks, such as song emotion recognition, emotion prediction in conversation, and sentiment analysis. Comparison experiments, ablation experiments, and visualization comprehensively demonstrate the universal capability of the proposed emotion2vec. To the best of our knowledge, emotion2vec is the first universal representation model in various emotion-related tasks, filling a gap in the field. [1]

1 Introduction

Extracting emotional representation from speech is an essential step of various emotional tasks such as speech emotion recognition (SER) and sentiment analysis. Traditional methods employ Filter Banks (FBanks) or Mel Frequency Cepstrum Coefficients (MFCCs) as speech features. These features are not rich in semantic information, resulting in limited performance on emotional tasks. Popular methods utilize features extracted from speech-based self-supervised learning (SSL) pre-trained models, leading to a significant performance improvement.

One potential challenge blocking further performance improvement is that these SSL models are not entirely suitable for emotional tasks. Wang et al. (2021) explore no fine-tuning, partial fine-tuning, and entire fine-tuning with some SSL models for SER on the IEMOCAP dataset (Busso et al., 2008), and give some empirical conclusions. While this is an ad-hoc solution, on the one hand, fine-tuning SSL models requires a large computational cost, on the other hand, these conclusions may be data-specific or model-constrained. Recently, Chen et al. (2023a) proposed an SER model named Vesper, which is obtained by model distillation from WavLM-large (Chen et al., 2022) with emotion data. Vesper is designed to perform the SER task, whose universal representation capability still needs to be demonstrated. Accordingly, a universal speech-based emotion representation model is urgently needed in the field.

Here we propose emotion2vec, a universal emotion representation model that can be used to extract speech features for diverse emotion tasks. Self-supervised pre-training is performed on 262 hours of open-source emotion data with an online distillation paradigm to obtain emotion2vec. Considering that both whole-play information and local details convey emotion, we propose a pre-training strategy combining utterance-level loss and frame-level loss. On the mainstream IEMOCAP dataset, the downstream linear model trained with features extracted from emotion2vec outperforms all the mainstream SSL models and the latest specialist models. emotion2vec is tested on 13 datasets including 10 languages, and the results show that emotion2vec exhibits language generalization ability. Moreover, in addition to the SER task, we also experimented with emotion2vec features on song emotion recognition, emotion prediction in conversation, and sentiment analysis. The results indicate that emotion2vec has excellent task generalization ability. Extensive ablation experiments and visualization analysis demonstrate the effectiveness of our pre-training methods and the versatility of the proposed emotion2vec model.

†Corresponding author

[1]Code, checkpoints, and extracted features are available at https://github.com/ddlBoJack/emotion2vec

Findings of the Association for Computational Linguistics ACL 2024, pages 15747–15760
August 11-16, 2024

2 Related Work

2.1 Speech-based SSL

Self-supervised learning has achieved remarkable success in the field of representation learning, showcasing its efficacy across natural language processing (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Brown et al., 2020), computer vision (Grill et al., 2020; He et al., 2020; Bao et al., 2021; He et al., 2022), as well as speech processing (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Baevski et al., 2022). For speech representation learning, all SSL models can be classified into two categories according to the self-supervised targets utilized during pre-training (Ma et al., 2023b): **1)** Offline targets. **2)** Online targets. Models employing offline targets often require a well-trained teacher model before the pre-training stage, to extract self-supervised targets. Representative models of this type are HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022) using K-means targets, and PBERT (Wang et al., 2022), MonoBERT&PolyBERT (Ma et al., 2023c) using phoneme-based targets. Models using online targets do not need a pre-trained teacher model in advance, while the teacher models are constantly updated during the pre-training phase, with an online distillation paradigm. Representative models of this type are data2vec (Baevski et al., 2022), data2vec 2.0 (Baevski et al., 2023) using frame-level mask language model (MLM) loss, and CA-DINO (Han et al., 2023) using utterance-level cross-entropy loss. emotion2vec is pre-trained combining both utterance-level loss and frame-level loss, leading to a superior speech emotion representation model.

2.2 Speech Emotion Representation

There have been works on text emotion representation (Felbo et al., 2017; Wang et al., 2020b; Wang and Zong, 2021), however, speech emotion representation is in the night before dawn. We present the first universal speech emotion representation model, while most of the previous works directly employ speech pre-training models (Pepino et al., 2021; Li et al., 2022), or fine-tune the pre-training models on their specific emotional data with specific emotional tasks (mostly SER) (Morais et al., 2022; Chen and Rudnicky, 2023), to extract speech emotion representation. A series of works investigate SER performance of wav2vec 2.0 (Wang et al., 2021), HuBERT (Wang et al., 2021), as well as WavLM (Ioannides et al., 2023), either fine-tuning or not. A recent work (Ma et al., 2023a) found that data2vec features also have a good representation ability for SER task. For speech emotion representation in other emotion tasks, such as multimodal emotion recognition, popular practice (Li et al., 2023a) is similar to what is mentioned above.

3 Methods

Here we mainly introduce the self-supervised pre-training method of the proposed emotion2vec, for which the core is to train the model with **Utterance-level Loss** and **Frame-level Loss** using **Online Distillation** paradigm. Combining utterance-level and frame-level loss is inspired by our observation that both global and local information convey emotion in speech. Moreover, initializing with pre-trained models for the teacher-student network warm-ups the online distillation process, providing a good representation for the subsequent self-supervised bootstrap learning.

3.1 Model Pipeline

As shown in Figure 1, emotion2vec contains two networks in the pre-training phase, which are the teacher network $\mathcal{T}$ and the student network $\mathcal{S}$. Both models share the same model architecture, including a feature extractor $\mathcal{F}$ composed of multi-layer convolutional neural networks and a backbone network $\mathcal{B}$ composed of multi-layer Transformers. These modules can be configured with different architectures, which will be described in Section 4.1. Given a raw audio utterance $X = [x_1, \cdots, x_{N_x}]$, the Teacher $\mathcal{T}$ and the Student $\mathcal{S}$ respectively utilize feature extractors $\mathcal{F}^{\mathcal{T}}$ and $\mathcal{F}^{\mathcal{S}}$ to obtain the downsampled features $Z_0 = [z_1, \cdots, z_{N_z}]$, which can be written as:

$$Z_0^{\mathcal{T}} = \mathcal{F}^{\mathcal{T}}(X), \tag{1}$$

$$Z_0^{\mathcal{S}} = \mathcal{F}^{\mathcal{S}}(X). \tag{2}$$

For the teacher network $\mathcal{T}$, the downsampled features $Z_0^{\mathcal{T}}$ are directly fed into the backbone network $\mathcal{B}^{\mathcal{T}}$. For the student network $\mathcal{S}$, the downsampled features $Z_0^{\mathcal{S}}$ are masked l consecutive frames with probability p for each frame as the start. Then learnable utterance embedding $U = [u_1, \cdots, u_{N_u}]$ is placed in the front before being fed into the backbone network $\mathcal{B}^{\mathcal{S}}$. The formula can be written as follows:

$$Z_i^{\mathcal{T}} = \mathcal{B}_i^{\mathcal{T}}(Z_{i-1}^{\mathcal{T}}), \tag{3}$$

$$Y^{\mathcal{T}} = \frac{1}{k} \sum_{i=n-k+1}^{n} Z_i^{\mathcal{T}}, \tag{4}$$

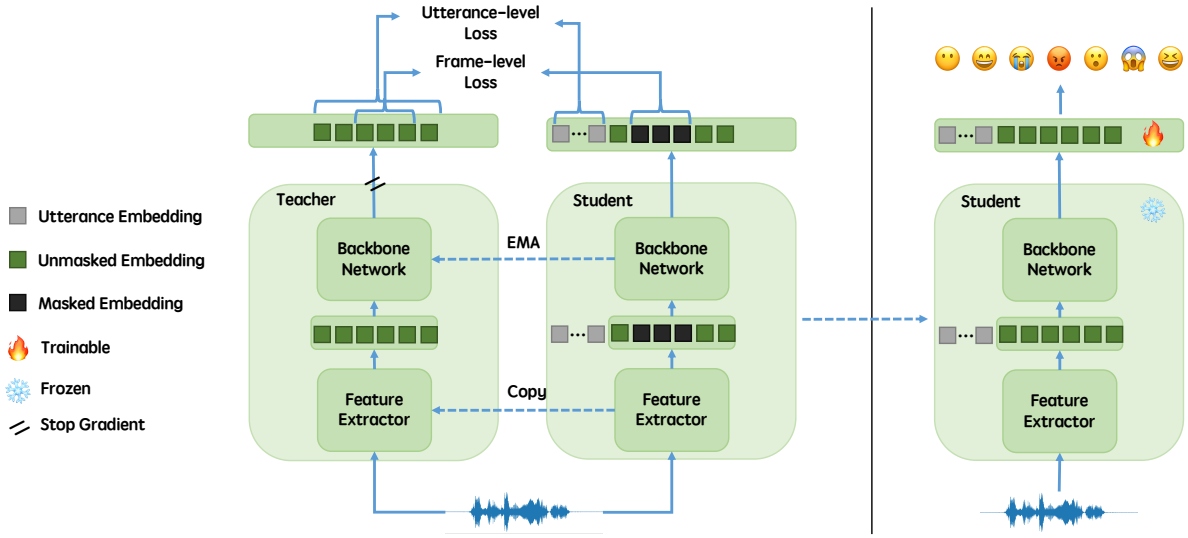


Figure 1: The overall framework of emotion2vec. During the pre-training phase (left), emotion2vec conducts online distillation with a teacher network and a student network, and the parameters of both networks are initialized employing the same pre-trained weights. When a specific downstream task is performed (right), the pre-trained emotion2vec is frozen and a lightweight downstream model is trained.

$$U^{\mathcal{S}}; Y^{\mathcal{S}} = \mathcal{B}^{\mathcal{S}}(U; Mask(Z_0^{\mathcal{S}})), \tag{5}$$

where $Y^{\mathcal{T}}$ is the average of the output embedding of the top k out of n Transformer Block $\mathcal{B}_i^{\mathcal{T}}$. Utterance-level output embedding $U^{\mathcal{S}}$ and frame-level output embedding $Y^{\mathcal{S}}$ are the outputs of the student backbone network $\mathcal{B}^{\mathcal{S}}$. $Mask$ is the applying mask operation. $Y^{\mathcal{T}}$, $Y^{\mathcal{S}}$ and $U^{\mathcal{S}}$ are the same in the hidden layer dimensions, where $Y^{\mathcal{T}}$ and $Y^{\mathcal{S}}$ have the same N_z temporal dimensions, while $U^{\mathcal{S}}$ has N_u temporal dimensions, respectively.

3.2 Utterance-level Loss

Utterance-level loss constructs an utterance-level pretext task to learn the global emotion. We use mean squared error (MSE) to calculate the loss, which can be written as:

$$L_{Utt} = (\bar{Y}^{\mathcal{T}} - \bar{U}^{\mathcal{S}})^2, \tag{6}$$

where

$$\bar{Y}^{\mathcal{T}} = \frac{1}{N_z} \sum_{i=1}^{N_z} Y_i^{\mathcal{T}}, \tag{7}$$

$$\bar{U}^{\mathcal{S}} = \frac{1}{N_u} \sum_{i=1}^{N_u} U_i^{\mathcal{S}}, \tag{8}$$

which means that utterance-level loss L_{Utt} is computed by temporal pooling results of $Y^{\mathcal{T}}$ and $U^{\mathcal{S}}$. Here we propose three ways to compute utterance-level loss, which we call **token embedding**, **chunk embedding**, and **global embedding**, as shown in Figure 2.

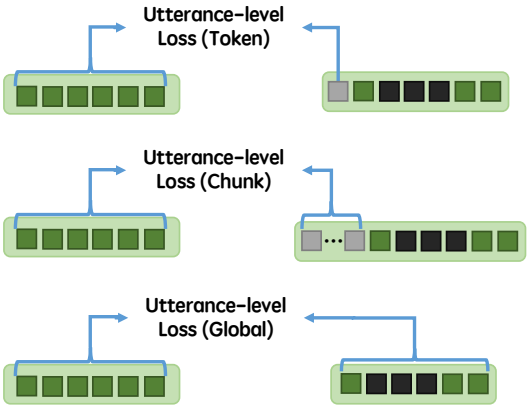


Figure 2: Different ways to compute utterance-level loss between the teacher network outputs (Left) and the student network outputs (Right) in pre-training.

Token Embedding. Token embedding employs a single token to represent global emotion information encoded by the student network $\mathcal{S}$. More explicitly, we set N_u to 1 in the learnable utterance embedding $U = [u_1, \cdots, u_{N_u}]$.

Chunk Embedding. Chunk embedding employs multiple tokens to represent global emotion information. In this case, more global information can be aggregated within the chunk.

Global Embedding. In the case of utilizing global embedding, no additional utterance tokens are added. We use temporal pooling of frame-level output embedding $Y^{\mathcal{S}}$ instead of $U^{\mathcal{S}}$ to compute the loss.

3.3 Frame-level Loss

Frame-level loss constructs a frame-wise pretext task to learn the context emotion. We only compute the loss on the masked part, which is the common practice for a mask language modeling(MLM) pretext task. The frame-level loss L_{Frm} can be expressed as:

$$L_{Frm} = \frac{1}{M}\sum_{i\in\mathbb{M}}(Y_i^{\mathcal{T}} - Y_i^{\mathcal{S}})^2, \tag{9}$$

where $\mathbb{M}$ denotes the index sequence of frame-level output embedding $Y^{\mathcal{S}}$ being masked, and M denotes the total number of tokens being masked.

3.4 Online Distillation

Online distillation is a self-supervised learning strategy for teacher-student learning, where the student network updates parameters by backpropagation and the teacher network updates parameters with an exponentially moving average (EMA) (Grill et al., 2020). For the student network $\mathcal{S}$, the total loss L for backpropagation is a combination of frame-level loss L_{Frm} and utterance-level loss L_{Utt}, donated as:

$$L = L_{Frm} + \alpha L_{Utt}, \tag{10}$$

with a tunable weight α. For the teacher network $\mathcal{T}$, The parameters $\theta_0^{\mathcal{T}}$ are initialized as the same parameters of the student network $\theta_0^{\mathcal{S}}$, and then are updated with EMA within each mini-batch, donated as:

$$\theta_{t+1}^{\mathcal{T}} = \tau\theta_t^{\mathcal{T}} + (1-\tau)\theta_{t+1}^{\mathcal{S}}. \tag{11}$$

where τ is a parameter that increases linearly during pre-training. In practice, within each mini-batch the parameters of teacher feature extractor $\mathcal{F}^{\mathcal{T}}$ are copied directly from $\mathcal{F}^{\mathcal{S}}$, while the parameters of teacher backbone network $\mathcal{B}^{\mathcal{T}}$ are updated with EMA from $\mathcal{B}^{\mathcal{T}}$ and $\mathcal{B}^{\mathcal{S}}$.

4 Experiments Setup

4.1 Initial Model

Different initial models lead to different architectures of feature extractors $\mathcal{F}$, backbone networks $\mathcal{B}$, and initialization parameters θ_0. Here we adopt two models, data2vec [2] and data2vec 2.0 [3], both of which have the same feature extractor design but different backbone network designs. The feature extractor $\mathcal{F}$ is a 7-layer 1-D convolutional neural network with kernel sizes $(5,2,2,2,2,2,2)$ and strides $(10,3,3,3,3,2,2)$, resulting in 320x downsampling. Given the raw audio input X at a 16000 Hz sample rate, the output representations Z are 50 Hz with dimension 512. Then a linear projection for dimension transformation from 512 to 768 is applied, followed by the mask operation to construct the input for the backbone network $\mathcal{B}$. Appendix A briefly introduces different backbone networks in data2vec and data2vec 2.0.

4.2 Training Details

Self-supervised Pre-training. In the pre-training phase, we train emotion2vec with 262 hours of unlabeled emotion data shown in Figure 1 with different initial models. For the training overhead, The pre-training is conducted on 4 NVIDIA A10 Tensor Core GPUs, and we simulate 16 GPUs by setting the update frequency to 4. We train emotion2vec for 100 epochs, each of which takes about 37 minutes. We use a dynamic batchsize, where the maximum number of tokens is 1×10^6. For the optimizing strategy, we use Adam with a learning rate of 7.5×10^{-5} and a weight decay of 1×10^{-2}. We train emotion2vec using a cosine learning rate scheduler, with 5% proportion of linear warm-up. For the student model, each time step of the input has a probability of $p = 0.5$ to be the start index, and the subsequent $l = 5$ time steps are masked. The hyperparameter α that controls the loss weight is set to 1. For the teacher model, we use the average of the top $k = 8$ blocks of the transformer layer outputs for providing the training targets. We apply a linearly increasing strategy for τ from $\tau_s = 0.999$ to $\tau_e = 0.99999$ for the teacher parameters exponentially moving average.

Supervised Fine-tuning. All model architectures of diverse downstream tasks are designed to be as simple as possible, to demonstrate the representation ability of the pretrained model. For the non-sequential task, following the common practice of SUPERB (Yang et al., 2021), we use two linear layers with a ReLU activation function sandwiched between them. For the sequential task, we use two layers of gated recurrent units (GRU) to make predictions.

4.3 Datasets

A summary of the datasets employed in our experiments is presented in Table 1. There are 18 emo-

[2] https://dl.fbaipublicfiles.com/fairseq/data2vec/audio_base_ls.pt

[3] https://dl.fbaipublicfiles.com/fairseq/data2vec2/base_libri.pt

Table 1: The datasets at a glance for emotion2vec pre-training and downstream tasks.

Dataset	Pretrain	Downstream	Source	Emo	Spk	Lang	#Utts	#Hours
IEMOCAP (Busso et al., 2008)	✓	✓	Act	5	10	English	5531	7.0
MELD (Poria et al., 2019)	✓	✓	Friends TV	7	407	English	13847	12.2
CMU-MOSEI (Zadeh et al., 2018)	✓	✓	YouTube	7	1000	English	44977	91.9
MEAD (Wang et al., 2020a)	✓	✗	Act	8	60	English	31792	37.3
MSP-Podcast (V1.8) (Martinez-Lucas et al., 2020)	✓	✗	Podcast	8	10000+	English	72969	113.5
Total	✓	–	–	–	–	English	169053	262.0
CMU-MOSI (Zadeh et al., 2016)	✗	✓	YouTube	7	89	English	2199	2.6
RAVDESS-Speech (Livingstone and Russo, 2018)	✗	✓	Act	8	24	English	1440	1.5
RAVDESS-Song (Livingstone and Russo, 2018)	✗	✓	Act	8	23	English	1012	1.3
SAVEE (Jackson and Haq, 2014)	✗	✓	Act	7	4	English	480	0.5
M3ED (Zhao et al., 2022)	✗	✓	TVs	7	626	Mandarin	24449	9.8
EmoDB (Burkhardt et al., 2005)	✗	✓	Act	7	10	German	535	0.4
EMOVO (Costantini et al., 2014)	✗	✓	Act	7	10	Italian	588	0.5
CaFE (Gournay et al., 2018)	✗	✓	Act	7	12	French	936	1.2
SUBESCO (Sultana et al., 2021)	✗	✓	Act	7	20	Bangla	7000	7.8
ShEMO (Mohamad Nezami et al., 2019)	✗	✓	Act	6	87	Persian	3000	3.4
URDU (Latif et al., 2018)	✗	✓	Talk shows	4	38	Urdu	400	0.3
AESDD (Vryzas et al., 2018)	✗	✓	Act	5	5	Greek	604	0.7
RESD (Lubenets et al.)	✗	✓	Act	7	200	Russian	1396	2.3

tional datasets including 10 different languages: 9 in English, and 1 in Mandarin, Bangla, French, German, Greek, Italian, Persian, Russian, and Urdu. For each dataset, it can be categorized in terms of *Pretrain* (*i.e.*, whether used during the pre-training phase), *Downstream* (*i.e.*, whether tested in the downstream task), *Source* (*i.e.*, where samples collected), *Emo* (*i.e.*, number of emotion categories), *Spk* (*i.e.*, number of speakers), *Lang*, (*i.e.*, Language), *#Utts* (*i.e.*, number of utterances), and *#Hours* (*i.e.*, total duration of samples). Speech data is extracted from these datasets and uniformly processed into a single channel of 16k Hz. Different datasets and configurations are used for pre-training and downstream tasks. If not specified, results are obtained using the random leave-one-out cross-validation (CV) unless the dataset provides a set partition. Refer to Appendix B for more details.

5 Results

5.1 Evaluation Metrics

We apply commonly used evaluation metrics, weighted accuracy (WA), unweighted accuracy (UA), and weighted average F1 (WF1), to evaluate the performance of speech emotion tasks. WA corresponds to the overall accuracy and UA corresponds to the average class-wise accuracy. WF1 is a comprehensive evaluation, especially for the situation of sample imbalance.

5.2 Main Results

The results are shown in Table 2, where we compare different SSL pre-trained models on IEMOCAP dataset, as well as larger-scale pre-trained models, and the latest specialist models designed for SER tasks. We follow the SUPERB (Yang et al., 2021) evaluation, freezing the pre-trained model and training downstream linear layers with the hidden dimensional set to 256. As can be seen from the table, emotion2vec outperforms all existing SSL pre-trained models, across all base models with similar parameters and large models with greater parameters. Compared with Versper-12, an SER model obtained by distillation from WavLM-large, emotion2vec works better with fewer parameters. TIM-NET (Ye et al., 2023), MSTR (Li et al., 2023b), and DST (Chen et al., 2023b) are the latest SER specialist models, respectively, which use different scales of upstream features and downstream networks. The proposed emotion2vec model outperforms or performs on par with these models with only linear layers, while their downstream networks have 2x, 135x, and 114x more parameters than emotion2vec, respectively. We provide the results of leave-one-session-out five-fold cross-validation and leave-one-speaker-out ten-fold cross-validation for reference.

We also conduct experiments on other mainstream English datasets to prove the generalization of emotion2vec in Table 3. MELD is a noisy dataset used to test the SER performance of the model in complex environments. RAVDESS and SAVEE are out-of-domain datasets. Experimental results show that emotion2vec exhibits state-of-the-art performance on different datasets in different environments, delivering powerful scenario generalization ability.

Table 2: SER task performance of different SSL pre-trained models on the IEMOCAP dataset. The setting of the downstream models follows SUPERB (Yang et al., 2021) to use linear layers to test the representation ability of different upstream models. "LS-960" means LibriSpeech 960 hours, "LL-60k" means LibriLight 60k hours, and "Mix-94k" means 94k hours of data including LibriLight, VoxPopuli, and GigaSpeech. For emotion data, "LSSED-206" means LSSED 206 hours, and "Emo-262" refers to the 262 hours of pre-training data in Table 1. Models are tested using leave-one-session-out five-fold cross-validation with 20% from the training set used as the validation set for each session. Models with **underline** are leave-one-speaker-out ten-fold cross-validation with 8 speakers for training, 1 speaker for validation, and 1 speaker for testing within each fold. Models with * imply the same fold for both validation and testing, for a fair comparison as some work uses this principle. We also compare with larger-scale pre-trained models and the latest specialist models designed for SER tasks.

Model	Pre-training Corpus	Upstream	#Upstream Params	Downstream	#Downstream Params	WA(%) ↑
Self-supervised Model						
small size						
wav2vec (Schneider et al., 2019)	LS-960	Proposed	32.54M	Linear	0.13M	59.79
vq-wav2vec (Baevski et al., 2019)	LS-960	Proposed	34.15M	Linear	0.20M	58.24
base size						
wav2vec 2.0 (Baevski et al., 2020)	LS-960	Proposed	95.04M	Linear	0.20M	63.43
HuBERT (Hsu et al., 2021)	LS-960	Proposed	94.68M	Linear	0.20M	64.92
WavLM (Chen et al., 2022)	LS-960	Proposed	94.70M	Linear	0.20M	65.94
WavLM+ (Chen et al., 2022)	Mix-94k	Proposed	94.70M	Linear	0.20M	67.98
data2vec (Baevski et al., 2022)	LS-960	Proposed	93.75M	Linear	0.20M	67.38
data2vec 2.0 (Baevski et al., 2023)	LS-960	Proposed	93.78M	Linear	0.20M	68.58
Vesper-4 (Chen et al., 2023a)	Mix-94k + LSSED-206	Proposed	63.52 M	Linear	0.26M	68.40
Vesper-12 (Chen et al., 2023a)	Mix-94k + LSSED-206	Proposed	164.29 M	Linear	0.26M	70.70
emotion2vec	LS-960 + Emo-262	Proposed	93.79M	Linear	0.20M	**71.79**
emotion2vec*	LS-960 + Emo-262	Proposed	93.79M	Linear	0.20M	**74.48**
emotion2vec	LS-960 + Emo-262	Proposed	93.79M	Linear	0.20M	**72.94**
emotion2vec*	LS-960 + Emo-262	Proposed	93.79M	Linear	0.20M	**77.64**
large size						
wav2vec 2.0 (Baevski et al., 2020)	LL-60k	Proposed	317.38M	Linear	0.26M	65.64
HuBERT (Hsu et al., 2021)	LL-60k	Proposed	316.61M	Linear	0.26M	67.62
WavLM (Chen et al., 2022)	Mix-94k	Proposed	316.62M	Linear	0.26M	70.03
Supervised Model						
TIM-Net (Ye et al., 2023)	-	MFCC	-	CNN(TIM-Net)	0.40M	68.29
MSTR (Li et al., 2023b)	-	HuBERT-large	316.61M	Transformer(MSTR)	27.00M	70.03
DST (Chen et al., 2023b)	-	WavLM-large	316.62M	Transformer(DST)	22.78M	71.80

Table 3: emotion2vec performance on mainstream English datasets.

Model	WA(%) ↑	UA(%) ↑	WF1(%) ↑	WA(%) ↑	UA(%) ↑	WF1(%) ↑	WA(%) ↑	UA(%) ↑	WF1(%) ↑
	MELD			**RAVDESS**			**SAVEE**		
WavLM-base	46.95	16.34	35.16	37.01	37.11	36.08	42.08	38.46	38.93
WavLM-base+	43.78	16.75	34.60	38.89	38.40	37.75	43.54	39.27	42.19
data2vec	45.75	24.98	43.59	69.58	69.70	69.25	82.50	82.26	82.37
data2vec 2.0	48.92	26.10	45.80	81.04	80.80	80.97	83.13	**82.94**	83.03
emotion2vec	**51.88**	**28.03**	**48.70**	**82.43**	**82.86**	**82.39**	**84.38**	82.30	**84.45**

Table 4: emotion2vec performance on datasets of other languages.

Model	WA(%) ↑	UA(%) ↑	WF1(%) ↑	WA(%) ↑	UA(%) ↑	WF1(%) ↑	WF1(%) ↑	UA(%) ↑	WF1(%) ↑
	AESD (Gr)			**CAFE (Fr)**			**RESD (Ru)**		
WavLM-base	55.33	55.50	54.86	31.61	32.02	30.88	56.17	56.17	55.69
WavLM-base+	53.83	54.41	52.48	31.40	33.39	30.40	55.00	55.19	55.08
data2vec	56.67	57.26	56.57	57.10	57.68	57.36	49.42	49.77	48.97
data2vec 2.0	71.33	70.20	70.93	71.51	72.98	71.50	64.08	64.33	64.17
emotion2vec	**72.33**	**72.27**	**71.57**	**74.52**	**75.26**	**74.53**	**64.75**	**65.04**	**64.53**
Model	**EmoDB (De)**			**EMOVO (It)**			**M3ED (Zh)**		
WavLM-base	59.06	55.32	58.96	40.17	40.34	37.36	44.03	18.90	34.50
WavLM-base+	65.66	64.60	64.83	40.34	41.98	40.11	45.09	20.18	36.49
data2vec	67.17	64.81	66.52	51.21	51.97	49.82	44.44	21.10	37.77
data2vec 2.0	83.77	83.07	83.93	60.69	61.27	60.79	47.50	24.12	41.74
emotion2vec	**84.34**	**84.85**	**84.32**	**61.21**	**62.97**	**60.89**	**49.15**	**26.98**	**44.38**
Model	**SUBESCO (Bn)**			**ShEMO (Fa)**			**URDU (Ur)**		
WavLM-base	54.50	54.77	53.96	67.27	46.60	65.63	71.00	70.25	70.82
WavLM-base+	54.73	54.69	54.59	66.73	44.29	65.12	67.25	68.68	67.47
data2vec	78.29	78.25	78.21	70.80	53.96	69.84	71.75	72.67	71.83
data2vec 2.0	87.91	87.95	87.90	77.90	62.03	76.96	77.50	78.42	77.12
emotion2vec	**90.91**	**90.96**	**90.91**	**79.97**	**66.04**	**79.56**	**81.50**	**81.87**	**81.60**

Table 5: emotion2vec performance of the song emotion recognition task on the RAVDESS-Song dataset.

Model	Upstream	Downstream	WA(%) ↑	UA(%) ↑	WF1(%) ↑
Self-supervised Model					
WavLM-base	Freeze	Linear	52.3	52.4	52.1
WavLM-base+	Freeze		54.9	53.9	54.2
data2vec	Freeze		63.8	64.1	63.4
data2vec 2.0	Freeze		73.0	74.6	72.7
L^3-Net (Koh and Dubnov, 2021)	Freeze		71.0	-	-
SpecMAE (Sadok et al., 2023)	Finetune		54.5	-	53.9
VQ-MAE-S (Patch-tf) (Sadok et al., 2023)	Finetune		84.0	-	84.0
VQ-MAE-S (Frame) (Sadok et al., 2023)	Finetune		84.2	-	84.3
emotion2vec	Freeze		**85.0**	**85.2**	**84.8**
Specialist Model					
VQ-MAE-S (Patch-tf) (Sadok et al., 2023)	Finetune	Query2Emo	83.7	-	83.4
VQ-MAE-S (Frame) (Sadok et al., 2023)			85.8	-	85.7

5.3 Language generalization

Given the various languages, the SER datasets exhibit notable domain shifts. The generalization of the model to unseen language is critically important for SER. We validate the generalization of emotion2vec and other baselines on the out-of-domain language SER datasets. We follow the evaluation of SUPERB (Yang et al., 2021), freezing the pre-trained models and training downstream linear layers with the hidden dimensional set to 256, where the WavLM-base, WavLM-base+, data2vec, dat2vec 2.0, and emotion2vec are our implementations following the practice above. As shown in Table 4, emotion2vec outperforms all the SSL baseline methods on the 9 different lingual datasets in terms of WA, UA, and WF1. These results demonstrate that emotion2vec captures the emotion patterns across languages and shows state-of-the-art performance.

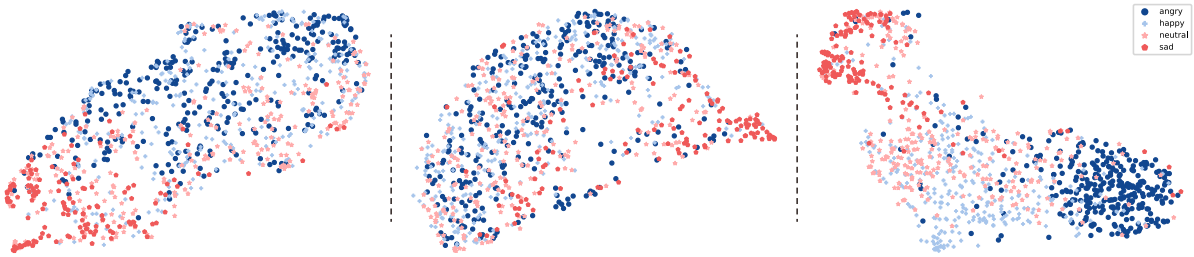
5.4 Task generalization

In order to verify the generalization of the model, in addition to speech emotion recognition, we tested other speech emotion tasks, including song emotion recognition, emotion prediction in conversation, and sentiment analysis.

Song Emotion Recognition. Song emotion recognition is a sub-task of music emotion recognition (MER), which aims to identify the emotion expressed in a singing voice. Following common practice, we perform five-fold cross-validation with randomly shuffled data, and remain one fold unseen during each training, to demonstrate the generalization of the features. WavLM-base, WavLM-base+, data2vec, dat2vec 2.0, and emotion2vec are our implementations, following the practice above. The results for L^3-NET, SpecMAE, and VQ-MAE-S are taken from their papers. As shown in Table 5, emotion2vec outperforms all known SSL models even without finetuning the model in the song emotion recognition task.

Emotion Prediction in Conversation. Emotion prediction in conversation (EPC) refers to predicting the future emotional state of a specific speaker based on historical conversation information. We reproduce Shi et al.'s (2023) method except that the speech features are obtained from our proposed emotion2vec. Briefly, the model employs several GRUs with a hierarchical structure for emotion prediction. Each prediction takes the previous 6 turns of the dialogue, in which one speaker can say multiple utterances in each turn. The network dimensions, hyperparameters, and training strategies are kept the same as the reference implementation with leave-one-speaker-out 10-fold cross-validation. For the speech modality, the input is 768-dimensional emotion2vec features. For the text modality, the input is 378-dimensional BERT (Devlin et al., 2019) features. For the speech + text multimodal, the input is a concatenation of emotion2vec features and BERT features, which also remains the same as the reference implementation. As shown in Table 6, with speech features replaced with emotion2vec in the EPC task, there are performance gains in both speech single modality and speech-text multi-modality.

Sentiment Analysis. Sentiment analysis is the task of analyzing text or speech to determine whether the affective state conveyed is positive, negative, or neutral. Following Lian et al.'s (2023) practice, we eliminate the neutral sentiment and perform the binary classification task



(a) WavLM (b) data2vec (c) emotion2vec

Figure 3: UMAP visualizations of learned features on downstream SER task from WavLM, data2vec, and emotion2vec on the IEMOCAP dataset. Red and Blue tones mean low and high arousal emotional classes, respectively.

Table 6: emotion2vec performance of emotion prediction in conversation on the IEMOCAP dataset.

Modality	Model	UAR(%) ↑	MacroF1(%) ↑
Speech	Noroozi et al.'s (2017)	56.78	55.11
	Shi et al.'s (2020)	61.98	60.21
	Shi et al.'s (2023)	65.01	65.91
	emotion2vec	**77.19**	**76.71**
Text (for reference)	Noroozi et al.'s (2017)	71.19	70.65
	Shi et al.'s (2020)	74.94	74.54
	Shi et al.'s (2023)	77.30	76.67
Speech + Text	Noroozi et al.'s (2017)	74.61	73.62
	Shi et al.'s (2020)	76.31	75.50
	Shi et al.'s (2023)	80.18	80.01
	emotion2vec	**81.68**	**80.75**

on the standard training/validation/test set of CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018), respectively. Also in line with Lian et al.'s (2023) practice, we utilize the mean of the last four layers' features of the pretrained model, to train the downstream linear layers. As shown in Table 7, emotion2vec outperforms pretrained data2vec and WavLM with self-supervised learning and pretrained Whisper Encoder (Radford et al., 2023) with supervised learning utilizing ASR task.

Table 7: emotion2vec performance of sentiment analysis on CMU-MOSI and CMU-MOSEI datasets.

Model	WF1(%) ↑	
	CMU-MOSI	CMU-MOSEI
Base Size		
data2vec	65.06	72.79
WavLM	62.36	73.47
Whisper Encoder	65.41	74.75
emotion2vec	**69.16**	**76.56**
Large Size		
WavLM	68.27	77.66
Whisper Encoder	64.93	76.46

5.5 Visualization

To investigate the intuitive effect of emotion2vec and other SSL baselines on emotion representation learning, we visualize the representations learned by WavLM, data2vec, and emotion2vec through the UMAP technique (McInnes et al., 2018) in Figure 3. We conduct the leave-one-session-out evaluation strategy, and randomly select 10% samples from the training set as the validation set. Specifically, for a fair comparison, the representations from the first linear layer are visualized after an identical training phase for different SSL models.

Figure 3 visualizes different SSL models to represent arousal. In a sense, arousal refers to emotional intensity. Figure 3 (a) and Figure 3 (b) show heavy overlapping between high and low arousal emotion classes. In contrast, Figure 3 (c) shows that the high arousal and low arousal representations are clustered receptively, and the feature distribution exhibits a trend transitioning from high arousal to low arousal, which is more reasonable compared to other methods. The results indicate that emotion2vec provides more emotion-aware representations to support its superior performance. we also visualize discrete emotions with more classes in Appendix D.

6 Conclusion

In this paper, we propose emotion2vec, a universal emotion representation model. emotion2vec is pre-trained on 262 hours of unlabeled emotion data through self-supervised online distillation, leading to universal emotion representation ability. We prove that our strategy of combining utterance-level loss and frame-level loss during emotion pre-training is effective. Extensive experiments demonstrate that the proposed emotion2vec has the ability to extract emotion representation across different tasks, languages, and scenarios. In the future, we will explore the scaling law of emotion representation models, namely how to provide a better representation with more data and larger parameters.

Acknowledge

This work was supported by the Science and Technology Innovation (STI) 2030-Major Projects un-

der Grant 2022ZD0208700 and the National Natural Science Foundation of China (No. 62206171 and No. U23B2018), and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102 and the International Cooperation Project of PCL.

Limitation

While emotion2vec provides a universal emotion representation, it requires training separate downstream models when testing different emotion-related tasks. In addition, whether speaker information is removed exists not explored, which is critical for training emotional TTS systems with the emotion2vec representations.

References

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *Proc. IMCL*.

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 *Proc. ICML*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. In *Proc. ICLR*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. BEiT: BERT pre-training of image Transformers. In *Proc. ICLR*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proc. NeurIPS*.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. 2005. A database of German emotional speech. In *Proc. Interspeech*.

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. In *Proc. LREC*.

Li-Wei Chen and Alexander Rudnicky. 2023. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *Proc. ICASSP*.

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. In *Proc. JSTSP*.

Weidong Chen, Xiaofen Xing, Peihao Chen, and Xiangmin Xu. 2023a. Vesper: A compact and effective pretrained model for speech emotion recognition. In *arXiv preprint*.

Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2023b. DST: Deformable speech Transformer for emotion recognition. In *Proc. ICASSP*.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO corpus: an Italian emotional speech database. In *Proc. LREC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proc. EMNLP*.

Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A Canadian French emotional speech dataset. In *Proc. ACM Multimedia*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: a new approach to self-supervised learning. In *Proc. NeurIPS*.

Bing Han, Zhengyang Chen, and Yanmin Qian. 2023. Self-supervised learning with cluster-aware-DINO for high-performance robust speaker verification. In *arXiv preprint*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proc. CVPR*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*.

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. In *Proc. TASLP*.

George Ioannides, Michael Owen, Andrew Fletcher, Viktor Rozgic, and Chao Wang. 2023. Towards paralinguistic-only speech representations for end-to-end speech emotion recognition. In *Proc. Interspeech*.

Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (SAVEE) database. University of Surrey.

Eunjeong Koh and Shlomo Dubnov. 2021. Comparison and analysis of deep audio embeddings for music emotion recognition. In *Proc. CEUR Workshop*.

Siddique Latif, Adnan Qayyum, Muhammad Usman, and Junaid Qadir. 2018. Cross lingual speech emotion recognition: Urdu vs. western languages. In *Proc. FIT*.

Jinchao Li, Shuai Wang, Yang Chao, Xunying Liu, and Helen Meng. 2023a. Context-aware multimodal fusion for emotion recognition. In *Proc. Interspeech*.

Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2022. Exploration of a self-supervised speech model: A study on emotional corpora. In *Proc. SLT*.

Zhipeng Li, Xiaofen Xing, Yuanbo Fang, Weibin Zhang, and Hengsheng Fan. 2023b. Multi-scale temporal transformer for speech emotion recognition. In *Proc. Interspeech*.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. 2023. MER 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proc. ACM Multimedia*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. In *arXiv preprint*.

Steven R Livingstone and Frank A Russo. 2018. 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*.

Ilya Lubenets, Nikita Davidchuk, and Artem Amentes. Aniemore.

Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, and Xie Chen. 2023a. Leveraging speech PTM, text LLM, and emotional TTS for speech emotion recognition. In *Proc. ICASSP*.

Ziyang Ma, Zhisheng Zheng, Changli Tang, Yujin Wang, and Xie Chen. 2023b. MT4SSL: Boosting self-supervised speech representation learning by integrating multiple targets. In *Proc. Interspeech*.

Ziyang Ma, Zhisheng Zheng, Guanrou Yang, Yu Wang, Chao Zhang, and Xie Chen. 2023c. Pushing the limits of unsupervised unit discovery for SSL speech representation. In *Proc. Interspeech*.

Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The MSP-conversation corpus. In *Proc. Interspeech*.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. In *arXiv preprint*.

Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. 2019. ShEMO: a large-scale validated database for Persian speech emotion detection. In *Proc. LREC*.

Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. 2022. Speech emotion recognition using self-supervised features. In *Proc. ICASSP*.

Fatemeh Noroozi, Neda Akrami, and Gholamreza Anbarjafari. 2017. Speech-based emotion recognition and next reaction prediction. In *Proc. SIU*.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Proc. Interspeech*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. ACL*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. In *OpenAI blog*.

Samir Sadok, Simon Leglaive, and Renaud Séguier. 2023. A vector quantized masked autoencoder for speech emotion recognition. In *arXiv preprint*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Interspeech*.

Xiaohan Shi, Sixia Li, and Jianwu Dang. 2020. Dimensional emotion prediction based on interactive context in conversation. In *Proc. Interspeech*.

Xiaohan Shi, Xingfeng Li, and Tomoki Toda. 2023. Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation. In *Proc. Interspeech*.

Sadia Sultana, M Shahidur Rahman, M Reza Selim, and M Zafar Iqbal. 2021. SUST Bangla emotional speech corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. In *Proc. PloS One*.

Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos A Dimoulas, and George Kalliris. 2018. Speech emotion recognition for performance interaction. In *Proc. AES*.

Chengyi Wang, Yiming Wang, Yu Wu, Sanyuan Chen, Jinyu Li, Shujie Liu, and Furu Wei. 2022. Supervision-guided codebooks for masked prediction in speech pre-training. In *Proc. InterSpeech*.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020a. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *Proc. ECCV*.

Shuo Wang, Aishan Maoliniyazi, Xinle Wu, and Xiaofeng Meng. 2020b. Emo2Vec: Learning emotional embeddings via multi-emotion category. In *Proc. TOIT*.

Xiangyu Wang and Chengqing Zong. 2021. Distributed representations of emotion categories in emotion space. In *Proc. ACL*.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021. A fine-tuned wav2vec 2.0/HuBERT benchmark for speech emotion recognition, speaker verification and spoken language understanding. In *arXiv preprint*.

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. In *Proc. Interspeech*.

Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. 2023. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In *Proc. ICASSP*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. In *arXiv preprint*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proc. ACL*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *Proc. ACL*.

A Initial Model

Figure 4 shows the pre-training pipeline of data2vec and data2vec 2.0, and both employ a teacher-student network for online distillation.

A.1 data2vec

The backbone network contains a 5-layer learnable convolutional positional encoding followed by a 12-layer standard Transformer. Each Transformer block is set to 768 model dimension, 3072 bottleneck dimension, and 12 attention heads. Finally, a linear projection from 768 to 768 is equipped on the student outputs, the results of which are employed to calculate MLM loss with teacher outputs.

A.2 data2vec 2.0

The data2vec 2.0 model shares the same Transformer architecture with data2vec, except for one more CNN decoder. The Transformer encoder only encodes the non-masked parts of downsampled features Z, and then the masked parts are complemented with random Gaussian noise before being passed to the CNN decoder, in a MAE-style fashion, to improve efficiency. The CNN decoder is a 4-layer 1-D convolutional neural network with all kernel sizes set to 7, strides set to 1, and channels set to 384, without downsampling. A linear projection from 384 to 768 is equipped to compute MLM loss, which works the same way as data2vec.

B Experimental Configurations on Different Datasets

B.1 Pre-training

In the pre-training phase, we utilize five large-scale English datasets, including IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), MEAD (Wang et al., 2020a), CMU-MOSEI (Zadeh et al., 2018), and MSP-Podcast (Martinez-Lucas et al., 2020), resulting in a total of 262 hours. The IEMOCAP corpus contains a total of 5 sessions and 10 different speakers, with each session being a conversation of two exclusive speakers. MELD is a multi-party conversational dataset containing about 13,847 utterances from 1,433 dialogues collected from the TV series '*Friends*'. MEAD is a talking-face video corpus featuring 60 actors and actresses talking with 8 different emotions at three different intensity levels. CMU-MOSEI is a multimodal dataset from YouTube for sentiment and emotion analysis in videos. MSP-Podcast is collected from

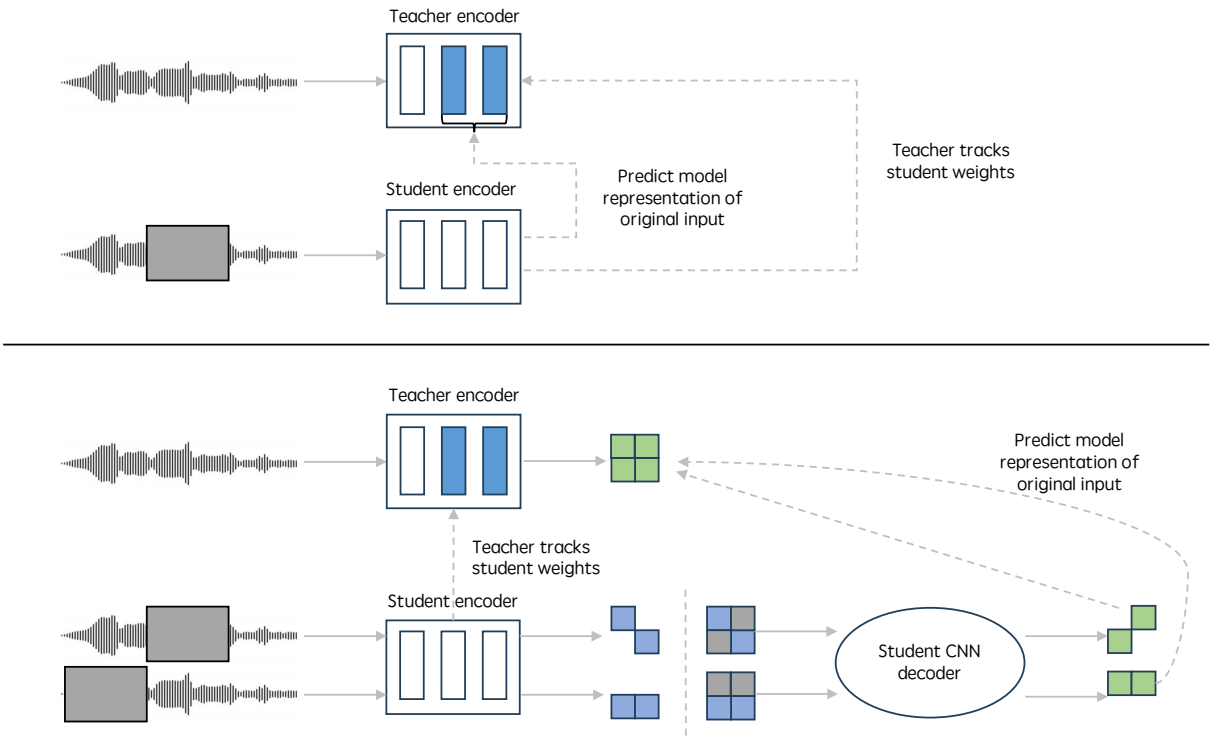


Figure 4: The figure shows the self-supervised pre-training pipeline for data2vec(top) and data2vec 2.0 (bottom).

podcast recordings that discuss a variety of topics like politics, sports, and movies.

B.2 Main Results

Different datasets are used to test different downstream tasks with various languages. For main results in Section 5.2, we report cross-validation (CV) results on the IEMOCAP dataset. The original labels cover five classes, to be consistent and comparable with previous methods (Ye et al., 2023; Chen et al., 2023b), we merge '*excited*' with '*happy*' to better balance the size of each emotion class, resulting in four classes. We conduct both leave-one-session-out 5-fold CV and leave-one-speaker-out 10-fold CV. Moreover, we report results on MELD under its original split setup, and RAVDESS (Livingstone and Russo, 2018), SAVEE (Jackson and Haq, 2014) datasets under a random leave-one-out 10-fold CV setup, which implies at each fold, all samples within the dataset are randomly split into 80%, 10%, and 10% samples in training, validation, and testing sets. Among them, speech in RAVDESS and SAVEE datasets is not seen in the pre-training stage, which demonstrates the generalization of the proposed model on out-of-domain corpora.

B.3 Language Generalization

For language generalization task in Section 5.3, we report CV results for 9 out-of-domain datasets, including 1 in Mandarin (M3ED (Zhao et al., 2022)), Bangla (SUBESCO (Sultana et al., 2021)), French (CaFE (Gournay et al., 2018)), German (EmoDB (Burkhardt et al., 2005)), Greek (AESDD (Vryzas et al., 2018)), Italian (EMOVO (Costantini et al., 2014)), Persian (ShEMO (Mohamad Nezami et al., 2019)), Russain (RESD (Lubenets et al.)), and Urdu (URDU (Latif et al., 2018)). If not specified, language generalization results are obtained using the random leave-one-out 10-fold CV as we mentioned above unless the dataset provides a set partition. Such as the RESD dataset, we follow its original split setup with 280 testing samples and 1116 training samples. Additionally, we allocate 10% from the training samples for validation and others for training.

B.4 Task Generalization

For task generalization task in Section 5.4. We tested other speech emotion tasks, including song emotion recognition, emotion prediction in conversation, and sentiment analysis, on RAVDESS-Song (Livingstone and Russo, 2018), IEMOCAP and CMU-MOSI (Zadeh et al., 2016) & CMU-MOSEI (Zadeh et al., 2018). For song emotion

recognition and emotion prediction in conversation, we report CV results. For sentiment analysis, we report results with its original split setup. To be comparable with previous work, the experimental setup varies according to the specific task.

C Ablation Study

If not specified, results are obtained using the standard leave-one-session-out 5-fold cross-validation on the IEMOCAP dataset.

C.1 Initialization Method

In this experiment, we explore the impact of initialization methods on performance. data2vec and data2vec 2.0 are two representative models trained with online distillation, both of which are pre-trained on Librispeech 960 hours. As shown in Table 8, initializing with a pre-trained model would be better than the cold start method. Model initializing with data2vec 2.0 performs better than the one initializing with data2vec.

Table 8: Ablation study with initialization methods.

Initialization	**WA(%)** ↑	**UA(%)** ↑	**WF1(%)** ↑
Cold Start	61.34	62.71	61.19
data2vec	70.2	70.93	70.11
data2vec 2.0	**71.79**	**72.69**	**71.80**

C.2 Training Loss

In this experiment, we explore the impact on performance of different combinations of loss when pre-training, and different features when training downstream models. As shown in Table 9, if only utterance-level loss is adopted during pre-training, the model almost does not work. If pre-trained with frame-level loss, the model obtains reasonable results whether utterance-level loss exists or not. When pre-trained with both utterance-level loss and frame-level loss, the model achieves good results. We also try to concatenate utterance embeddings and frame embeddings when training downstream models, where we obtain similar results as frame embeddings only.

Table 9: Ablation study with different loss.

Frm Loss	**Utt Loss**	**Downstream**	**WA(%)** ↑	**UA(%)** ↑	**WF1(%)** ↑
✗	✓	Utt	28.96	25.0	13.13
✓	✗	Frm	70.85	71.61	70.71
✓	✓	Utt	62.77	63.53	62.43
✓	✓	Frm	**71.79**	**72.69**	**71.80**
✓	✓	Utt ⊕ Frm	71.37	**72.69**	71.29

C.3 Utterance-level Loss Method

In this experiment, we compare the impact of different types of utterance-level loss proposed in Section 3.2 on performance. As shown in Table 10, chunk embedding chosen to compute utterance-level loss performs the best.

Table 10: Ablation study with the methods for utterance-level loss.

Utt Loss Method	**WA(%)** ↑	**UA(%)** ↑	**WF1(%)** ↑
Token	70.46	71.07	70.33
Chunk	**71.79**	**72.69**	**71.80**
Global	70.30	71.52	70.18

C.4 Utterance-level Loss Weight

In this experiment, we compare the impact of utterance-level loss weight on performance. As shown in Table 11, weighting utterance-level loss and frame-level loss with a ratio of 1:1 works best.

Table 11: Ablation study with the weight for utterance-level loss.

Utt Loss α	**WA(%)** ↑	**UA(%)** ↑	**WF1(%)** ↑
0	70.85	71.61	70.71
0.1	71.06	72.16	71.08
1	**72.14**	**72.84**	**72.13**
10	70.58	71.37	70.61

D Visualization for discrete emotion classes

We conduct the 8:2 hold-out evaluation on SUBESCO,and randomly select 10% samples from the training set as the validation set. The representations from the first linear layer are visualized for different SSL models. Figure 5 shows the ability of different SSL models to represent discrete emotion classes. As Figure 5 (a) and Figure 5 (b) show, WavLM and data2vec suffer from class confusion problems. On the contrary, the features learned by emotion2vec demonstrate a higher intra-class compactness and a much larger inter-class margin. The results indicate that emotion2vec provides more class-discriminative representations.

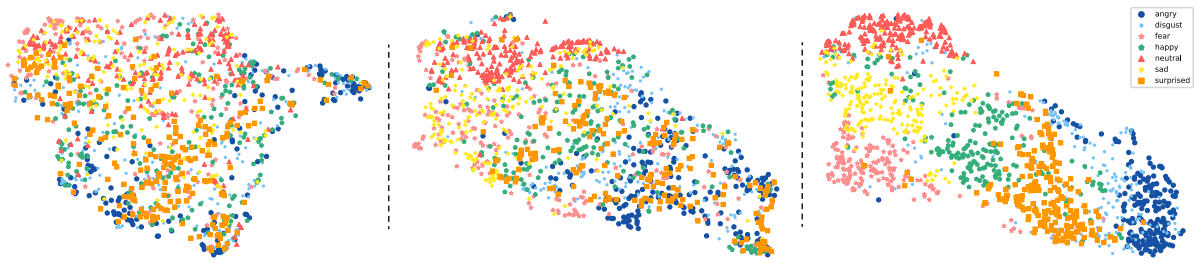


Figure 5: UMAP visualizations of learned features on downstream SER task from WavLM, data2vec, and emotion2vec on the SUBESCO dataset.