PUNSER: Large-Scale Pre-Trained and Unified Model for Practical Speech Emotion Recognition

Yu Hayashizaki*, Takashi Nose*, Sumiharu Kobayashi*, Satoru Fukayama† and Akinori Ito*
* Tohoku University, Japan
E-mail: {hayashizaki.yu.t5@dc., takashi.nose.b7@, kobayashi.sumiharu.r4@dc., akinori.ito.a2@} tohoku.ac.jp
† National Institute of Advanced Industrial Science and Technology (AIST), Japan
E-mail: s.fukayama@aist.go.jp

***Abstract*—In this study, we propose PUNSER, a pre-trained speech emotion recognition (SER) model constructed via large-scale supervised learning by integrating multiple speech emotion corpora with multiple label types. Our proposed method employs a novel masked multi-task learning (MTL) framework to leverage both discrete categorical and continuous dimensional emotions, utilizing 150k utterances (226 hours) from 10 corpora. This approach enables a robust training methodology applicable to realistic scenarios with missing labels. Evaluation on 6 corpora with diverse speaking styles and languages shows that PUNSER achieves state-of-the-art performance, surpassing previous methods on both categorical and dimensional emotion recognition tasks. The model also demonstrates strong performance in low-resource and out-of-domain conditions, confirming its high generalization capability and practical utility.**

I. Introduction

Speech Emotion Recognition (SER) is attracting attention as a key technology for enabling more natural and intuitive Human-Computer Interaction. For instance, there is a growing demand for practical applications such as in call centers and emotional text-to-speech [1], [2]. However, previous SER methods are known to suffer significant performance degradation on speech from unseen domains [3], [4]. Therefore, applying an SER model to a specific target domain requires collecting new data from that domain and performing additional training. Furthermore, the annotation of emotion labels is highly subjective and is both costly and labor-intensive [5].

To build robust SER models with limited data, leveraging pre-trained models has become mainstream. Self-Supervised Learning (SSL) models, such as WavLM [6] and XEUS [7], are trained on large-scale, diverse, unlabeled speech, enabling robust, general-purpose feature extraction [8], [9]. However, these SSL models are trained to acquire general acoustic features, and their ability to capture emotional acoustic characteristics is not optimized for SER. To address this issue, emotion2vec [10] built a pre-trained model specialized for SER by performing large-scale SSL on 262 hours of speech emotion data from public corpora. However, emotion2vec cannot explicitly represent the types or degrees of emotional expression since it is built using SSL without emotion labels. Supervised learning that directly utilizes emotion labels is considered more effective for accurately characterizing subjective human emotional expressions.

In this study, we propose PUNSER (large-scale Pre-trained and UNified model for practical Speech Emotion Recognition), a pre-trained model developed through large-scale supervised learning on multiple speech emotion corpora with multiple label types. PUNSER explicitly learns from emotion labels via large-scale supervised pre-training, using 150k utterances (226 hours) from 10 emotion corpora. During this process, we employ masked MTL to leverage both discrete categorical and continuous dimensional emotion labels. Previous MTL studies for SER have mainly focused on integrating tasks such as automatic speech recognition, gender classification, corpus classification, and language identification [4], [11], [12]. However, few have explored simultaneous prediction of multiple emotion representations [13], [14]. Crucially, these studies only address ideal cases where all labels are available, and do not consider realistic scenarios with missing label types [13]–[15]. The masked MTL in PUNSER introduces a mask that specifies which emotion labels are available for each training sample, allowing training even when some label types are missing. When fine-tuning on target domain speech, we use the weights from large-scale supervised pre-training as initialization to transfer acquired emotional knowledge and improve recognition performance.

II. Large-Scale Supervised Pre-Training on Multi-Type Labeled Corpora

A. Overview of PUNSER

PUNSER is a pre-trained model trained on multiple speech emotion corpora with multiple label types. As illustrated in Fig. 1, we develop PUNSER by performing masked MTL to predict both categorical and dimensional emotion labels, using 150k utterances (226 hours of speech) from 10 distinct speech emotion corpora. Through MTL, the model can leverage information from both label types in a mutually beneficial way, enabling efficient acquisition of emotional representations. A challenge arises when integrating multiple corpora for MTL, as some may lack one of the two emotion label types. To address this, we introduce a mask indicating label presence, enabling training even when a certain label type is missing. Subsequently, we perform supervised fine-tuning on target domain speech, using the weights from large-scale supervised pre-training as initialization. This transfer of emotional knowledge is expected to improve recognition performance.

979-8-3315-7206-8/25/$31.00

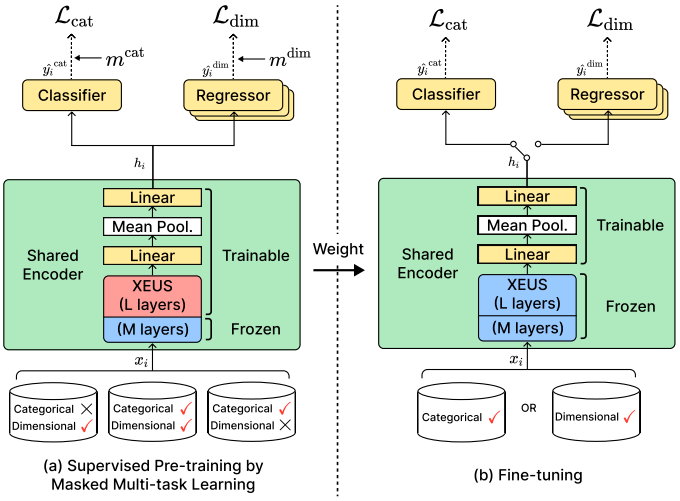


Fig. 1: Overview of PUNSER

B. SER Model with Unified Multi-Type Emotion Labels

The proposed pre-trained model consists of a shared encoder and task-specific networks. We consider a speech emotion recognition dataset $\mathcal{D} = \{(x_i, y_i^{\text{cat}}, y_i^{\text{dim}})\}_{i=1}^{N}$, where x_i is the speech feature of the i-th utterance. $y_i^{\text{cat}} = (y_{i,1}^{\text{cat}}, \ldots, y_{i,C}^{\text{cat}}) \in \{0,1\}^C \cup \varnothing$ is a C-dimensional one-hot vector representing the categorical emotion label, or $\varnothing$ if missing. $y_i^{\text{dim}} = (y_{i,V}^{\text{dim}}, y_{i,A}^{\text{dim}}, y_{i,D}^{\text{dim}}) \in \mathbb{R}^3 \cup \varnothing$ is a 3-dimensional vector representing the dimensional emotion labels: valence (V), arousal (A), and dominance (D).

First, a latent representation is extracted from the speech features by a shared encoder f_{enc}:

$$h_i = f_{\text{enc}}(x_i; \theta_{\text{enc}}) \tag{1}$$

where θ_{enc} is a parameter set of f_{enc}. For the shared encoder, we adopted a structure that appends a linear layer to XEUS [7], which gave the highest average emotion recognition performance in [16] compared with other speech SSL models such as emotion2vec base [10] and WavLM large [6]. In this study, we use the hidden states from the final layer of XEUS as input. We compute a weighted sum of these states, apply mean pooling over the temporal axis, and feed the result into a linear layer to extract a 256-dimensional emotional representation vector, which serves as the shared encoder output h_i. Next, using this shared representation, we construct a classification model for categorical emotion recognition and three regression models for dimensional emotion recognition, one for each dimension $d \in \{\text{V}, \text{A}, \text{D}\}$:

$$\hat{y}_i^{\text{cat}} = f_{\text{cat}}(h_i; \theta_{\text{cat}}), \quad \hat{y}_{i,d}^{\text{dim}} = f_{\text{dim},d}(h_i; \theta_{\text{dim},d}) \tag{2}$$

where f_{cat} is the classification model for categorical emotion recognition, $f_{\text{dim},d}$ is the regression model for dimensional emotion recognition, and θ_{cat} and $\theta_{\text{dim},d}$ are their respective trainable parameters. Following previous work, both the classification and regression models consist of two linear layers with a hidden dimension of 256 and a ReLU activation function [9].

To ensure expressive power for learning from multiple corpora while stabilizing training, we update the weights from the last $L = 6$ layers of XEUS during the masked MTL. During fine-tuning, XEUS weights are frozen, and only the linear layer in the shared encoder and the classification and regression model weights are updated.

C. Masked Multi-Task Learning for Different Type Emotion Labels

While MTL can improve performance by leveraging different emotion label types, standard approaches typically fail when one of the label types is missing [13], [14]. To address this issue, our method introduces a mask to ignore the loss of samples of missing label type. Specifically, we introduce binary masks $m_i^{\text{cat}}, m_i^{\text{dim}} \in \{0,1\}$ to indicate the presence of categorical and dimensional emotion labels for the i-th utterance, where a value of 1 signifies that the label exists.

First, we calculate the respective losses for the classification and regression. To handle batches with missing labels, we first define the set of indices for samples with categorical labels, $I_{\text{cat}} = \{i | m_i^{\text{cat}} = 1\}$. The loss is then calculated by averaging the cross-entropy over only these samples, aligning its scale with the dimensional loss:

$$\mathcal{L}_{\text{cat}} = -\frac{1}{|I_{\text{cat}}|} \sum_{i \in I_{\text{cat}}} \sum_{j \in C} y_{i,j}^{\text{cat}} \log(\hat{y}_{i,j}^{\text{cat}}) \tag{3}$$

where $|I_{\text{cat}}|$ is the number of samples with categorical emotion labels in the batch.

For the regression models for dimensional emotion recognition, we use a loss based on the Concordance Correlation Coefficient (CCC), following previous work [8]. Let $I_{\text{dim}} = \{i | m_i^{\text{dim}} = 1\}$ be the set of indices for samples with dimensional labels in the batch. The vectors of predicted and true values, $\hat{y}_d^{\text{dim}}$ and y_d^{dim}, are constructed from the samples corresponding to these indices. The loss is then defined as:

$$\mathcal{L}_{\text{dim}} = \sum_{d \in \{\text{V,A,D}\}} (1 - CCC_d(\hat{y}_d^{\text{dim}}, y_d^{\text{dim}})) \tag{4}$$

$$\hat{y}_d^{\text{dim}} = \{\hat{y}_{i,d}^{\text{dim}} | i \in I_{\text{dim}}\}, \quad y_d^{\text{dim}} = \{y_{i,d}^{\text{dim}} | i \in I_{\text{dim}}\} \tag{5}$$

For each dimensional emotion label d, CCC loss is calculated as follows:

$$CCC_d(\hat{y}_d^{\text{dim}}, y_d^{\text{dim}}) = \frac{2\rho_d \sigma_{\hat{y}_d^{\text{dim}}} \sigma_{y_d^{\text{dim}}}}{\sigma_{\hat{y}_d^{\text{dim}}}^2 + \sigma_{y_d^{\text{dim}}}^2 + (\mu_{\hat{y}_d^{\text{dim}}} - \mu_{y_d^{\text{dim}}})^2} \tag{6}$$

where ρ_d is the Pearson correlation coefficient between the predicted and true values, $\sigma_{\hat{y}_d^{\text{dim}}}$ and $\sigma_{y_d^{\text{dim}}}$ are the standard deviations of the predicted and true values, and $\mu_{\hat{y}_d^{\text{dim}}}$ and $\mu_{y_d^{\text{dim}}}$ are the means of the predicted and true values, respectively.

The final loss $\mathcal{L}$ is a weighted sum of the two task-specific losses:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{cat}} + (1 - \alpha) \mathcal{L}_{\text{dim}} \tag{7}$$

Here, α is a hyperparameter that controls the importance of each task, we set $\alpha = 0.25$ based on preliminary experiments in Section III-B. By calculating the loss with this masked

TABLE I: List of corpora used. ACT: acted, IMPRO: improvised, SPNT: spontaneous.

Corpus	Pre-training	Fine-tuning	Type	Language	Speaker	Sentence	Categoly	Dimension	Scale	Utterance	Hours
MSP-Podcast [17]	✓	-	SPNT	ENG	1,409	-	7	V,A,D	1-7	84,030	134.3
CMU-MOSEI [18]	✓	-	SPNT	ENG	1,000	-	6	-	-	8,638	18.2
ASVP-ESD [19]	✓	-	SPNT	CHN+	128+	-	8	-	-	11,679	15.5
WHiSER [20]	✓	-	SPNT	ENG	-	-	7	V,A,D	1-7	5,427	6.4
ESCorpus-PE [21]	✓	-	SPNT	SPA	-	-	-	V,A,D	0-5	3,764	5.0
CREMA-D [22]	✓	-	ACT	ENG	91	12	6	-	-	7,442	5.3
RAVDESS [23]	✓	-	ACT	ENG	24	2	7	-	-	1,248	1.3
CaFE [24]	✓	-	ACT	FRA	12	6	7	-	-	936	1.2
QuechuaSER [25]	✓	-	ACT	QUE	7	2,070	6	V,A,D	1-5	12,419	15.6
JTES [26]	✓	-	ACT	JPN	100	50	4	-	-	20,000	23.6
Total (training)	✓	-								**155,583**	**226.4**
IEMOCAP [27]	-	✓	ACT/IMPRO	ENG	10	-	8	V,A,D	1-5	10,039	12.4
MSP-Improv [28]	-	✓	SPNT/ACT/IMPRO	ENG	12	20	4	V,A,D	1-5	8,386	9.5
MELD [29]	-	✓	SPNT	ENG	407	-	7	-	-	13,706	12.1
UUDB [30]	-	✓	SPNT	JPN	14	-	-	V,A,D	1-7	4,840	1.9
SAVEE [31]	-	✓	ACT	ENG	4	15	7	-	-	480	0.5
EmoDB [32]	-	✓	ACT	GER	10	10	6	-	-	454	0.4
Total (evaluation)	**-**	✓								**37,905**	**36.8**

approach, only samples with existing labels contribute to the loss calculation, enabling training even when one of the label types is missing.

D. Large-Scale Corpora for Supervised Training

We hypothesize that training on an integration of multiple speech emotion corpora can improve generalization performance by increasing domain coverage. Therefore, for pre-training with masked MTL, we used 150k utterances (equivalent to 226 hours of speech with emotion labels) from the 10 corpora listed in Table I. The training corpora feature diverse languages such as Chinese, Spanish, and Japanese, as well as a range of speech types from acted speech to spontaneous speech. MSP-Podcast v1.11 is divided into training, validation, and test sets by its authors. We used only the audio from the training set. For CMU-MOSEI, categorical emotion labels are assigned by multiple annotators for a single utterance, and there are cases where two or more labels are tied for the most frequent. We excluded such utterances from our training data.

For categorical emotion labels, we used eight categories: neutral, happy, sad, angry, surprised, fear, disgust, and excited. During inference with the trained classification model, only the classifier outputs corresponding to the specific set of labels in an evaluation corpus are used. For example, when evaluating on a 4-class categorical emotion recognition task (neutral, happy, sad, and angry), the predicted label is determined by selecting the class with the highest value among the four corresponding outputs of the classification model $\hat{y}_i^{\mathrm{cat}}$, while the outputs for all other classes are ignored. For dimensional emotion labels, we used the three dimensions of Valence, Arousal, and Dominance. Since the scales of dimensional emotion labels differ across corpora, we applied min-max scaling to normalize them to the range [0, 1], following previous work [33].

III. Experiments

A. Experimental Conditions

For pre-training with masked MTL, we randomly split the entire training dataset into 90% for training and 10% for validation. For fine-tuning and evaluation, we used 6 corpora with diverse speaking styles and languages: IEMOCAP [27], MSP-Improv [28], MELD [29], UUDB [30], SAVEE [31], and EmoDB [32]. When evaluating on IEMOCAP, we followed previous work and performed leave-one-session-out cross-validation [10]. On IEMOCAP only, we also conducted an evaluation for 4-class categorical emotion recognition (neutral, happy, sad, and angry), as this setting is commonly used in previous studies[9], [10]. MELD provides predefined training, validation, and test splits, we adhered to this partitioning. For the other corpora (MSP-Improv, UUDB, SAVEE, and EmoDB), we performed leave-one-speaker-out cross-validation.

Regarding hyperparameters and evaluation metrics, we used a common experimental setup for both the masked MTL pre-training and the fine-tuning phases. The input audio was resampled to 16 kHz. We used the AdamW optimizer with a learning rate of 0.00001. The weight decay was set to 0.01. Following previous work, the batch size was dynamically adjusted according to the frame length of the audio, with a maximum frame length of 800,000 [10]. The maximum number of epochs was set to 100, and we employed early stopping to terminate training if the validation loss did not improve for 3 consecutive epochs. For evaluation, we report the results from the model that achieved the lowest validation loss. For categorical emotion recognition, we used Weighted Accuracy (WA) and Unweighted Accuracy (UA) as evaluation metrics. For dimensional emotion recognition, we used the CCC.

As comparison methods, we included both self-supervised learning (SSL) models and supervised SER models. For the SSL models, we adopted emotion2vec base [10], emotion2vec+ large[1], WavLM large [6], and XEUS [7]. These models were evaluated by freezing their parameters and training a simple appended classification or regression model. Specifically, similar to the SUPERB benchmark [9], we used the hidden states of the final layer of each model as input, took a weighted sum of these states, applied mean pooling over the temporal axis, and then fed the result into a linear layer to extract a 256-dimensional emotional representation vector. Subsequently, we attached a head model consisting of

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

TABLE II: Relationship between hyperparameter α and recognition performance on validation data

α	Categorical				Dimensional			Mean
	WA_4	UA_4	WA_8	UA_8	CCC_V	CCC_A	CCC_D	
0	0.258	0.265	0.098	0.083	0.580	0.687	0.673	0.378
0.25	**0.761**	0.746	**0.700**	0.608	**0.598**	**0.693**	**0.683**	**0.684**
0.5	0.743	**0.754**	0.694	0.592	0.579	0.686	0.656	0.673
0.75	0.757	0.737	0.698	**0.619**	0.554	0.666	0.645	0.668
1	0.750	0.730	0.691	0.606	-0.033	0.030	0.004	0.397

TABLE III: Relationship between the number of trainable layers and recognition performance on validation data

L	Categorical				Dimensional			Mean
	WA_4	UA_4	WA_8	UA_8	CCC_V	CCC_A	CCC_D	
0	0.687	0.679	0.620	0.509	0.500	0.605	0.566	0.595
3	0.759	0.741	0.693	0.587	0.587	0.674	0.665	0.673
6	**0.761**	**0.746**	**0.700**	**0.608**	0.598	0.693	0.683	**0.684**
9	0.760	0.737	0.696	0.601	**0.603**	**0.699**	**0.684**	0.683

a single linear layer for each task. For the supervised SER models, we included DST [34] and ShiftFormer [35]. For DST and ShiftFormer, we followed their respective original papers, setting the maximum input lengths to 326 and 375, respectively [34], [35]. Other hyperparameters were the same as in our proposed method.

B. Hyper-Parameter Tuning in the Multi-Task Learning Stage

We first examined the effect of the hyperparameter α in Eq. (7) during masked MTL pre-training, as its value determines the balance between learning from categorical and dimensional emotion labels. For this evaluation, we used the validation data from the pre-training phase and also evaluated performance on the 4-class categorical emotion recognition task (neutral, happy, sad, and angry). The results are shown in Table II. As α approaches 0, more emphasis is placed on dimensional emotion recognition, and the case of $\alpha = 0.0$ is equivalent to performing standard dimensional emotion recognition. Similarly, as α approaches 1, more emphasis is placed on categorical emotion recognition. For both dimensional and categorical emotion recognition, performance tended to be higher in the range of $0.25 \leq \alpha \leq 0.75$, where masked MTL is utilized. This indicates that through masked MTL, information from each label type is being learned in a mutually supplementary manner. Based on these results, we set $\alpha = 0.25$ in our study.

Next, we investigated the impact of the number of trainable layers L in the proposed method, as this parameter controls the capacity of the model to learn from multiple corpora. To determine the optimal value of L, we conducted a similar validation process by varying L. The results are shown in Table III. While performance was notably low for $L = 0$, it tended to be higher for $L = 3, 6, 9$. This suggests that the linear layer alone cannot adequately learn from the large volume of audio data. Based on the overall average performance, we set $L = 6$ for our main experiments.

C. The Effect of Unified Supervised Pre-Training

We compared the performance of the proposed method and previous methods after fine-tuning. The results are shown in Tables IV and V. In these tables, the boldface indicates the best performance, and the underline indicates the second-best performance. As a result of the large-scale supervised pre-training, the proposed method surpassed previous methods and achieved state-of-the-art performance in both categorical and dimensional emotion recognition under the fine-tuning condition. When using the pre-trained model directly for inference without any fine-tuning, the accuracy for dimensional emotion recognition was limited, while for categorical emotion recognition, it still demonstrated performance comparable to previous methods. The fact that a single model achieves performance improvements on both recognition tasks suggests that our proposed method successfully acquires superior emotional representations. Furthermore, its strong performance on IEMOCAP, MSP-Improv, and MELD, which are corpora that include spontaneous speech and are close to real-world scenarios, indicates that it is more suitable for practical applications compared to previous methods. Moreover, its superior performance compared to emotion2vec and XEUS on the Japanese corpus UUDB and the German corpus EmoDB indicates the possibility that our model learns more generalizable, cross-lingual emotional representations.

D. Low-Resource and Out-of-Domain Conditions

Since previous SER models generalize poorly to new domains, requiring costly data collection for each adaptation, developing models that perform well with limited data is crucial [3]–[5]. Therefore, using the experimental setup described in Section III-C, we evaluated the performance of our proposed method and the baseline XEUS when the amount of data used for fine-tuning was reduced. Specifically, we used random subsets of the fine-tuning data, ranging from 1% to 50% of the original training set. For the proposed method, we also conducted an evaluation in an out-of-domain setting, where the pre-trained model was used for inference directly without any fine-tuning.

The results of this experiment are shown in Fig. 2. In both categorical and dimensional emotion recognition, the proposed method consistently outperforms XEUS, widening the performance gap significantly in low-resource fine-tuning conditions. Remarkably, for categorical emotion recognition, the model achieves 0.426 in the out-of-domain setting without any fine-tuning, a result that is comparable to the 0.448 achieved by XEUS when fine-tuned with 50% of the data. These results demonstrate that the proposed method possesses high generalization capability and can achieve strong performance with very small amounts of data.

IV. Conclusions

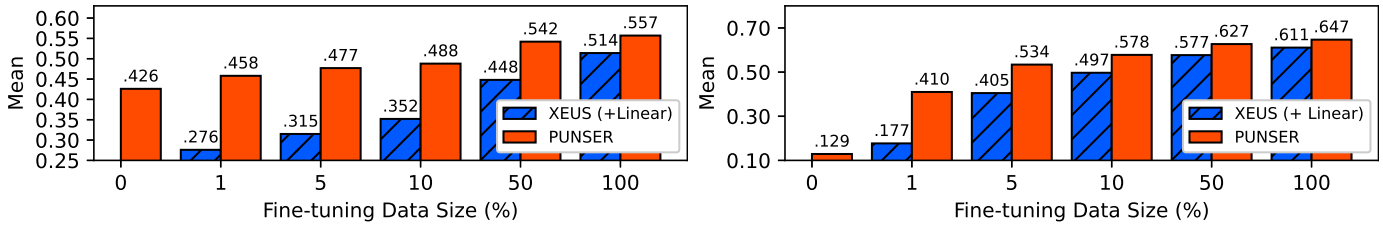
In this study, we proposed PUNSER, a pre-trained model built via large-scale supervised learning that integrates multiple speech emotion corpora with multiple label types. In evaluations on 6 corpora with diverse speaking styles and languages, the proposed model surpassed previous methods and achieved state-of-the-art performance in both categorical and dimensional emotion recognition. Furthermore, the model

TABLE IV: Comparison of the proposed method and previous methods for categorical emotion recognition. Subscripts of each metric indicate the number of categories. * denotes acted speech corpus.

Method	IEMOCAP				MSPImprov		MELD		SAVEE*		EmoDB*		Mean
	WA_4	UA_4	WA_8	UA_8	WA_4	UA_4	WA_7	UA_7	WA_7	UA_7	WA_6	UA_6	
emotion2vec base [10]	0.587	0.521	0.409	0.253	0.574	0.394	0.484	0.161	0.307	0.216	0.588	0.538	0.419
emotion2vec+ large[1]	0.452	0.377	0.362	0.209	0.494	0.286	0.475	0.143	0.314	0.234	0.305	0.260	0.326
WavLM large [6]	0.667	0.592	0.558	0.362	0.620	0.472	0.494	0.179	0.333	0.240	0.674	0.603	0.483
DST [34]	0.533	0.451	0.378	0.236	0.461	0.269	0.475	0.143	0.286	0.224	0.361	0.290	0.342
ShiftFormer [35]	0.654	0.588	0.532	0.340	0.618	0.488	0.497	0.202	0.474	0.415	**0.886**	**0.855**	0.546
XEUS [7]	**0.672**	0.604	0.555	0.359	0.645	0.518	**0.511**	0.220	**0.485**	**0.416**	0.617	0.570	0.514
PUNSER (Pre-trained)	0.579	0.555	0.417	0.328	0.535	0.488	0.286	0.231	0.325	0.266	0.572	0.530	0.426
PUNSER (Fine-tuned)	**0.672**	**0.618**	**0.584**	**0.401**	**0.652**	**0.550**	**0.511**	**0.235**	0.465	0.400	0.811	0.780	**0.557**

TABLE V: Comparison of the proposed method and previous methods for dimensional emotion recognition. V, A, and D denote Valence, Arousal, and Dominance, respectively.

Method	IEMOCAP			MSPImprov			UUDB			Mean
	CCC_V	CCC_A	CCC_D	CCC_V	CCC_A	CCC_D	CCC_V	CCC_A	CCC_D	
emotion2vec base [10]	0.561	0.426	0.453	0.441	0.544	0.466	0.675	0.404	0.791	0.529
emotion2vec+ large[1]	0.413	0.128	0.289	0.247	0.306	0.266	0.533	0.309	0.622	0.346
WavLM large [6]	0.682	0.527	0.527	0.527	0.647	0.517	0.786	0.531	**0.837**	0.620
XEUS [7]	0.682	0.540	0.531	0.531	**0.672**	0.542	0.737	0.439	0.820	0.611
PUNSER (Pre-trained)	0.167	0.027	0.231	0.181	0.346	0.194	0.002	0.004	0.007	0.129
PUNSER (Fine-tuned)	**0.685**	**0.581**	**0.542**	**0.615**	**0.672**	**0.551**	**0.791**	**0.554**	0.832	**0.647**



(a) Categorical (b) Dimensional

Fig. 2: Mean recognition performance as the fine-tuning data size decreases. Note that 0% represents an out-of-domain setting.

demonstrated superior performance in low-resource and out-of-domain conditions, confirming its high generalization capability and practical utility.

Acknowledgment

Part of this work was supported by JSPS KAKENHI (23K20725, 23K21945, and 25K00471), and Tohoku University and AIST Matching Research Support Program 2025.

References

[1] J. M. Martín-Doñas, A. L. Zorrilla, M. deVelasco, *et al.*, "Speech emotion recognition for call centers using self-supervised models: A complete pipeline for industrial applications," in *Proc. ICNLSP*, 2024, pp. 119–128.

[2] D.-H. Cho, H.-S. Oh, S.-B. Kim, S.-H. Lee, and S.-W. Lee, "EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech," in *Proc. INTERSPEECH*, 2024, pp. 1810–1814.

[3] E. Goron, L. Asai, E. Rut, and M. Dinov, "Improving domain generalization in speech emotion recognition with whisper," in *Proc. ICASSP*, 2024, pp. 11 631–11 635.

[4] J. Parry, E. DeMattos, A. Klementiev, *et al.*, "Speech emotion recognition in the wild using multi-task and adversarial learning," in *Proc. INTERSPEECH*, 2022, pp. 1158–1162.

[5] J. Santoso, K. Ishizuka, and T. Hashimoto, "Large language model-based emotional speech annotation using context and acoustic feature for speech emotion recognition," in *Proc. ICASSP 2024*, 2024, pp. 11 026–11 030.

[6] S. Chen, C. Wang, Z. Chen, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[7] W. Chen, W. Zhang, Y. Peng, *et al.*, "Towards robust speech representation learning for thousands of languages," in *Proc. EMNLP*, Association for Computational Linguistics, 2024, pp. 10 205–10 224.

[8] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 745–10 759, 2023.

[9] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "SUPERB: Speech Processing universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.

[10] Z. Ma, Z. Zheng, J. Ye, *et al.*, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *Findings of the Association for Computational Linguistics*, 2024, pp. 15 747–15 760.

[11] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. INTERSPEECH*, 2019, pp. 2803–2807.

[12] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Proc. INTERSPEECH*, 2021, pp. 4508–4512.

[13] D. Tompkins, D. Emmanouilidou, S. Deshmukh, and B. Elizalde, "Multi-view learning for speech emotion recognition with categorical emotion, categorical sentiment, and dimensional scores," in *Proc. ICASSP 2023*, 2023, pp. 1–5.

[14] J. L. Bautista and H. Shin, "Speech emotion recognition model based on joint modeling of discrete and dimensional emotion representation," *Applied Sciences (Switzerland)*, vol. 15, no. 2, p. 623, 2025.

[15] A.-R. Ispas, T. Deschamps-Berger, and L. Devillers, "A multi-task, multi-modal approach for predicting categorical and dimensional emotions," in *Proc. ICMI*, Paris, France, 2023, pp. 311–317.

[16] Y. Hayashizaki, T. Nose, S. Kobayashi, and A. Ito, "Robust speech emotion recognition using recent self-supervised learning with data augmentation," in *2025 IEEE 12th Global Conference on Consumer Electronics (GCCE)*, 2025, pp. 1–2.

[17] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[18] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. ACL*, 2018, pp. 2236–2246.

[19] D. T. T. Landry, Q. He, H. Yan, and Y. Li, "ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances," *Global Scientific Journal*, vol. 8, no. 5, pp. 1793–1798, 2020.

[20] A. R. Naini, L. Goncalves, M. A. Kohler, D. Robinson, E. Richerson, and C. Busso, "WHiSER: White house tapes speech emotion recognition corpus," in *Proc. INTERSPEECH*, 2024, pp. 1595–1599.

[21] A. C. Parari, A. D. Mattos, W. R. Lovón, and H. B. Córdova, *ESCorpus-PE: A speech emotional dataset in spanish with peruvian accent*, 2021.

[22] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[23] 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," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 2018.

[24] P. Gournay, O. Lahaie, and R. Lefebvre, "A canadian french emotional speech dataset," in *Proc. MMSys*, 2018, pp. 399–402.

[25] R. Y. G. Paccotacya-Yanque, C. A. Huanca-Anquise, J. Escalante-Calcina, W. R. Ramos-Lovón, and Á. E. Cuno-Parari, "A speech corpus of quechua collao for automatic dimensional emotion recognition," *Scientific Data*, vol. 9, no. 1, p. 778, 2022.

[26] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *Proc. O-COCOSDA*, 2016, pp. 16–21.

[27] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[28] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[29] 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, pp. 527–536.

[30] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.

[31] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[32] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[33] A. Derington, H. Wierstorf, A. Özkil, F. Eyben, F. Burkhardt, and B. W. Schuller, "Testing speech emotion recognition machine learning models," *arXiv preprint arXiv:2312.06270*, 2023.

[34] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "DST: Deformable speech transformer for emotion recognition," in *Proc. ICASSP*, 2023, pp. 1–5.

[35] S. Shen, F. Liu, and A. Zhou, "Mingling or misalignment? temporal shift for speech emotion recognition with pre-trained representations," in *Proc. ICASSP*, 2023, pp. 1–5.