ENHANCING TWO-STAGE FINETUNING FOR SPEECH EMOTION RECOGNITION USING ADAPTERS

Yuan Gao, Hao Shi, Chenhui Chu, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Kyoto, Japan

{gao,shi,kawahara}@sap.ist.i.kyoto-u.ac.jp chu@nlp.ist.i.kyoto-u.ac.jp

ABSTRACT

This study investigates the effective finetuning of a pretrained model using adapters for speech emotion recognition (SER). Since emotion is related with linguistic and prosodic information and also other attributes such as gender and speaking style, a framework of multi-task learning (MTL) has been shown to be effective for SER. However, the learning targets of automatic speech recognition (ASR) and other attribute recognition are apparently in conflict. Therefore, we propose to employ different adaptation methods for different tasks in multiple finetuning stages. Since ASR is the most challenging and also influential for SER, in the first stage, we finetune all parameters of the pretrained model for ASR and SER. In the second stage, we incorporate adapters to finetune the model for gender and style recognition in addition to SER by freezing the parameters of the main Transformer model tuned for ASR. Experimental evaluations which extensively compare different adaptation methods using the IEMOCAP dataset demonstrate that the proposed approach achieves a significant improvement from the simple MTL.

Index Terms— Speech emotion recognition (SER), pretrained model, multi-task learning (MTL), adapters

1. INTRODUCTION

With the continuous development of artificial intelligence, the utilization of speech in human-computer interaction (HCI) is becoming increasingly prevalent. With its potential to promote natural and user-friendly HCI, speech emotion recognition (SER) has emerged as a significant research area, enhancing the overall user experience across various applications. Consequently, this research has garnered increasing attention.

In earlier studies, researchers have employed convolutional neural networks (CNNs) [1, 2] and recurrent neural networks (RNNs) [3] to derive emotional features from input speech or spectrogram. However, the available training data in existing emotional speech datasets are not sufficient enough to train large models using supervised learning. Transfer learning poses a promising solution to address the data scarcity challenge. In recent years, researchers have explored the utilization of large pretrained models based on self-supervised learning (SSL) such as wav2vec 2.0 [4] for speech processing tasks, including low-resource automatic speech recognition (ASR) [5] and speaker recognition [6]. The model is pretrained on huge amounts of unlabeled data in self-supervised manner to capture both acoustic and linguistic information, and can subsequently be finetuned on limited labeled data for the target task. In the initial work of using wav2vec 2.0 for SER, Pepino et al. [7] showed the superior performance of the pretrained model compared with traditional deep learning approaches.

Since the linguistic and acoustic information have a great impact on emotion expression, multi-task learning (MTL) has been widely used by jointly training the model with auxiliary emotional-related tasks [8, 9]. The shared information across tasks can enhance the model to capture relevant emotional features and boost SER performance. Parthasarathy et al. [10] proposed to jointly learn the arousal, dominance, and valence information. They found that joint training of the model with multiple emotional attributes can largely enhance the performance. Cai et al. [11] proposed an MTL approach of ASR and SER based on the wav2vec 2.0 model. They showed that the joint training with ASR to learn the linguistic information led to better performance of SER.

We hypothesize that one of the main optimization challenges of traditional MTL framework arises from conflicting gradients [12], such as when the learning targets of SER and auxiliary tasks have high positive curvature. For example, it is expected that ASR requires high-level features independent of gender and speaking style information, which are in conflict with classification of these attributes. To address the problem, we have proposed a two-stage finetuning method [13], which finetunes the model for different tasks stage by stage. However, this method shifts all parameters of the model for different tasks in the end, thus does not completely solve the gradient-conflict problem. In this study, therefore, we propose to adopt different adaptation methods for different tasks in multiple stages. Specifically, we introduce adapters for gender and speaking style recognition (acted vs. spontaneous) in addition to SER in the second stage after we finetune all parameters for ASR and SER. We conduct comprehensive evaluations of the adaptation methods and show the superiority of the proposed method.

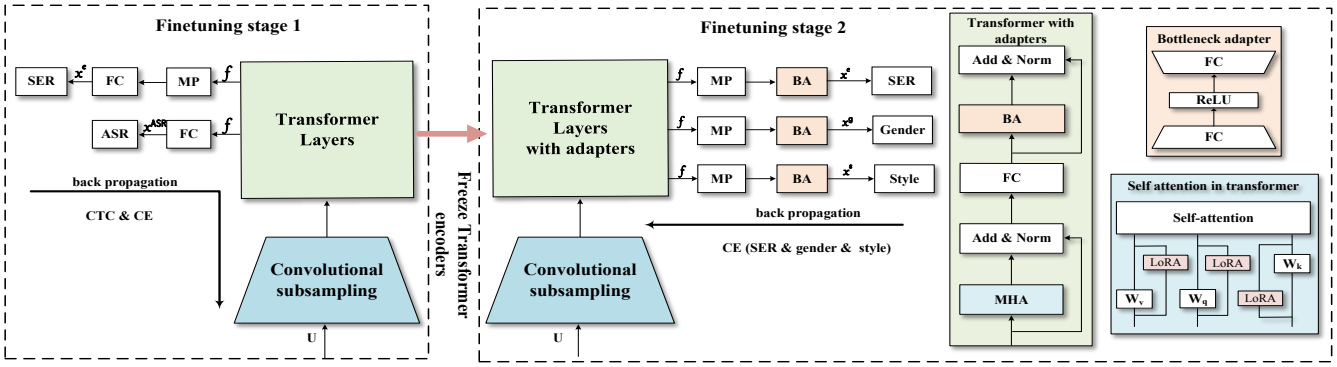

Fig. 1. Network structure of the proposed two-stage finetuning method. The XLS-R model is finetuned in the first stage, then we freeze the Transformer encoder and train adapters in the second stage.

2. FINETUNING PRETRAINED MODEL FOR SER

2.1. Single-stage finetuning with MTL

We first finetune a pretrained model for domain adaptation. To enrich the feature learning, we combine the auxiliary tasks and adopt an MTL framework. For an input waveform U, we extract the output of the last Transformer layer $f \in \mathbb{R}^{t \times d}$ as the latent feature, where t is the length of the utterance, and d denotes the hidden dimension of the Transformer layer. We learn the latent feature x^{ASR} with a fully-connected layer (FC layer) and apply the connectionist temporal classification (CTC) loss function for ASR. For emotion, gender, and speaking style recognition, we apply mean-pooling to the time dimension of f followed by a fully-connected layer (FC layer) for each specific task. And the output features $x^{<e,g,s>} \in \mathbb{R}^{d \times n}$ of each FC layer are used for classification using the cross-entropy (CE) loss function

$$L_{ASR} = CTC(x^{ASR}, y^{ASR}) \tag{1}$$

$$L_{<emotion,gender,style>} = CE(< x^e, x^g, x^s >, < y^e, y^g, y^s >) \tag{2}$$

where y^{ASR} is the text transcription, and y^e, y^g, and y^s are the ground truth labels of emotion, gender, and style recognition tasks. In the conventional MTL, all the tasks are trained simultaneously, and thus the overall objective function L of the model is defined as:

$$L = (1 - 3\alpha)L_{emotion} + \alpha L_{ASR} + \alpha L_{gender} + \alpha L_{style} \tag{3}$$

where α is a weight parameter for each auxiliary task.

2.2. Adapters

Adapters have been introduced for parameter-efficient finetuning of large pretrained models first in natural language processing and then speech processing. It provides a powerful mechanism in transfer learning and model adaptability. This structure allow neural networks to acquire additional task-specific information without the need for extensive finetuning of the entire model parameters of Transformers.

In this work, after finetuning of the Transformer encoder, adapter structures are incorporated to extract emotion and acoustic information for other attributes. As shown in the right part of Figure 1, to learn the shared emotional representation, we implement low-rank adaptation (LoRA) with rank 64 within each Transformer layer, enhancing the multi-head attention (MHA) mechanism, and a bottleneck adapter is incorporated after the fully-connected (FC) layer. Individual bottleneck adapters are inserted after the Transformer encoder for each task of emotion, gender, and style.

2.3. Proposed two-stage finetuning

Applying MTL to facilitate rich transcription from speech can benefit SER. Nevertheless, the learning objectives and the gradient magnitudes across different tasks pose challenges. In the large pretrained model (which has 24 Transformer layers), the Transformer encoder tend to focus on the learning of linguistic information, which is independent of gender and styles. Therefore, the learning objective of ASR is apparently in conflict with gender and style recognition. To address this problem, we introduce a two-stage finetuning approach, which conducts finetuning for different tasks step by step. As shown in the left part of Figure 1, the pretrained model is first finetuned with ASR and SER to embed the linguistic and emotion information into feature extractor:

$$L^{(1)} = (1 - \lambda)L_{emotion} + \lambda L_{ASR} \tag{4}$$

Then in the second stage, we freeze the pretrained model, and only the adapters are trained for emotion, gender, and style recognition. The objective function for the second finetuning stage is expressed as:

$$L^{(2)} = (1 - 2\beta)L_{emotion} + \beta L_{gender} + \beta L_{style} \tag{5}$$

where β denotes a weight parameter.

Table 1. Comparasion of finetuning strategies and auxiliary tasks in single-stage finetuning.

Finetuning strategies		Task				SER		ASR	Gender	Style
Adapters	All parameters	SER	ASR	Gender	Style	UA	WA	WER	UA	UA
✓		✓				68.51	67.83	-	-	-
✓		✓	✓			71.70	71.28	23.68	-	-
✓		✓		✓	✓	70.19	69.85	-	85.82	80.34
	✓	✓				70.88	71.57	-	-	-
	✓	✓	✓			75.40	75.17	**19.99**	-	-
	✓	✓		✓	✓	71.05	70.27	-	86.37	82.16
	✓	✓	✓	✓	✓	**76.19**	**75.49**	20.14	**97.52**	**89.29**

2.4. Feature fusion and self-contrastive loss

To leverage the emotional-related acoustic information for SER, latent features x^g and x^s derived from gender and style recognition modules are fed into a shared emotion classifier. This allows the model for using the gender and style information for SER. Additionally, we introduce a self-contrastive loss (SCL) that aligns the embedding space of x^e, x^g, and x^s, fostering the extraction of more discriminative SER features. Given that the input features are learned from the same utterance, which evidently share same emotion category. The proposed SCL reduces intra-class feature distances. We conduct a comparasion of two variations of the SCL. The first form, denoted as SCL_{norm}, is defined as:

$$SCL_{norm} = -||x^e - x^g|| - ||x^e - x^s|| - ||x^g - x^s|| \quad (6)$$

The second variant SCL_{cos}, is defined as:

$$SCL_{cos} = -cos(x^e - x^g) - cos(x^e - x^s) - cos(x^g - x^s) \quad (7)$$

One of these losses is augmented to MTL of Equ (5).

3. EXPERIMENTAL SETUP

3.1. Database

In this study, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [14] to evaluate the proposed methods. This dataset contains emotional data of approximately 12 hours. Ten American English speakers were engaged in recording five speaker-independent dyadic sessions, and each session is performed by two speakers (one male and one female) with a series of either scripted or improvisational scenarios (speaking style). For each speech utterance, three annotators assigned the emotional categorical labels. We adopt the common practice of merging "happy" and "excited" into one emotion class named "happy" [15, 11], resulting in 5,531 utterances with four emotion classes: happy (1,636), sad (1,084), angry (1,103), and neutral (1,708).

3.2. Implementations

We implemented our proposed models with the PyTorch framework and the Huggingface Transformers repository [16]. In this work, we employed XLS-R (wav2vec2-xls-r-300m) [17], a wav2vec 2.0-based pretrained model designed for speech tasks. This model encompasses 7 Convolutional Neural Network (CNN) layers for transmuting raw audio data into a latent representation, coupled with 24 layers of Transformer designed to capture contextual information. To accommodate varying input lengths, we applied sentence padding within each mini-batch. During training, we finetuned the pretrained model for 200 epochs. The learning rate was 10e-5, and the mini-batch size was 16. In the MTL experiments, we assigned auxiliary task weight parameters (α and β) of 0.1. In accordance with established studies [15, 18], we conducted 5-fold speaker-independent cross-validation on the IEMOCAP dataset.

4. RESULTS AND ANALYSIS

4.1. Comparasion of adaptation methods in single-stage finetuning

We first compare two finetuning strategies. 1) All parameters (of the Transformer): we finetune all parameters of XLS-R. 2) Adapters (only): we freeze the XLS-R and train adapters.

The results are presented in Table 1. With simple finetuning experiments, the performance of SER exhibits a 4.52% improvement in UA when ASR is incorporated. This observation underscores the significance of training the Transformer encoder for ASR, facilitating the effective embedding of linguistic information. On the other hand, when employing the adapter tuning strategy, the best performance is also achieved by simultaneously training SER and ASR, but both tasks yield lower performance compared to finetuning the Transformers.

4.2. Results of two-stage finetuning with adapters

Within the framework of the proposed two-stage finetuning, we systematically evaluate different training strategies. These

Table 2. Comparasion of different finetuning strategies in two-stage finetuning (SER performance).

Setting	Finetuning strategies		UA	WA
	Stage 1	Stage 2		
1	Adapters	Adapters	71.89	71.46
2	All parameters	All parameters	74.92	73.81
3	All parameters	Adapters	**76.54**	**76.07**

Table 3. Results of adapter tuning in second finetuning stage.

Method	SER		Gender	Style
	UA	WA	UA	UA
-	77.13	76.30	98.48	**91.33**
Concatenation	76.89	76.06	**98.77**	89.29
SCL_norm	77.91	77.24	98.31	90.23
SCL_cos	**78.49**	**77.85**	98.64	90.57

experimental settings are defined as follows: setting 1 uses adapter tuning in both the first and second stages, while setting 2 centers around finetuning of the XLS-R model in both stages. In setting 3, we finetune all parameters of XLS-R in the first stage, and use adapter tuning in the second stage. In this section, we train SER and ASR in the first stage since these two tasks are most important in the previous section. Then we exclusively train the model for SER in the second stage to validate two-stage finetuning strategies.

The results are shown in Table 2. Using only adapter tuning in setting 1 results in poor performance, since finetuning Transformers are crucial for ASR. Also, setting 2 does not gain improvement over the single-stage finetuning approach. This outcome can be attributed to the potential loss of linguistically-rich embeddings acquired during the first finetuning stage. In setting 3, freezing the Transformer encoder and finetuning the adapters bring a significant effect for SER. Compared with single-stage finetuning, we achieved a notable 1.04% improvement on UA.

4.3. Evaluation of the proposed feature fusion and SCL

Then we evaluate the proposed feature fusion and SCL for SER. In this experiment, we finetune the Transformer encoder for SER and ASR in the first stage. In the second stage, we use adapter tuning for gender and style recognition as auxiliary tasks of SER.[1]

As shown in Table 3, conducting MTL in the second stage using adapter tuning (first line in Table 3) outperformed the single-task model. Incorporating x^e, x^g, and x^s for SER by simple concatenation achieved similar performance of using only x^e. However, the incorporation of SCL can effectively enhance the discriminative feature learning for SER. Using SCL_cos is better than SCL_norm and achieved the best performance among the comparative experiments.

[1]We systematically assessed various task combinations and their orders in two-stage fine-tuning. This configuration achieved the best performance.

Table 4. Comparsion with previous works using SSL pretrained models (SER performance).

Approach	Year	UA	WA
Pepino et al. [7]	2021	67.20	-
Li et al. [18]	2022	-	63.40
Zou et al. [19]	2022	71.05	69.80
Ioannide et al. [20]	2023	-	74.32
Chen et al. [21]	2023	74.30	-
Fang et al. [22]	2023	74.03	74.95
Gao et al. [13]	2023	76.10	74.94
Proposed	-	**78.49**	**77.85**

4.4. Comparasion with state-of-the-art approaches

In this section, we benchmark the performance of the proposed approach against recent studies employing SSL pretrained models under same experimental conditions. Table 4 provides a comparative summary. Comparing to our earlier work involving skip-connections [13], the proposed approach, which integrates adapters and SCL, achieves a superior performance. As shown in Table 4, our approach outperforms recent studies by more than 2.39% and 2.91% on UA and WA, respectively.

5. CONCLUSION AND FUTURE WORK

We have proposed two-stage finetuning with different adaptation methods to effectively leverage emotional-related information for SER using a pretrained model. The experimental results demonstrate that upon the finetuned model, incorporating adapters in the second stage for additional auxiliary tasks can effectively addresses the gradient conflict problem. The proposed method has achieved UA score of 78.49%. This represents a substantial improvement, surpassing simple MTL by an absolute margin of 2.30% and the single-task learning baseline by an absolute margin of 6.28%. Our future research will explore the potential of adapter tuning to enhance emotional ASR.

6. ACKNOWLEDGMENT

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123, Grant-in-Aid for Scientific Research KAKENHI (JP19H05691, 20H00602 and 23H03454), and JST Moonshot R&D (JPMJPS2011).

7. REFERENCES

[1] H. Kumbhar and S. Bhandari, "Speech emotion recognition using mfcc features and lstm network," in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*. IEEE, 2019, pp. 1–3.

[2] Yang Liu, Haoqin Sun, Wenbo Guan, Yuqi Xia, and Zhen Zhao, "Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework," *Speech Communication*, vol. 139, pp. 1–9, 2022.

[3] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech emotion classification using attention-based lstm," *IEEE TASLP*, vol. 27, no. 11, pp. 1675–1685, 2019.

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[5] A. Rouhe, A. Virkkunen, J. Leinonen, M. Kurimo, et al., "Low resource comparison of attention-based and hybrid asr exploiting wav2vec 2.0," pp. 3543–3547, 2022.

[6] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *Proc. ICASSP 2022*. IEEE, pp. 6147–6151.

[7] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[8] H. Zhang, M. Mimura, T. Kawahara, and K. Ishizuka, "Selective multi-task learning for speech emotion recognition using corpora of different styles," in *Proc. ICASSP 2022*. IEEE, pp. 7707–7711.

[9] Hao Shi, Masato Mimura, Longbiao Wang, Jianwu Dang, and Tatsuya Kawahara, "Time-domain speech enhancement assisted by multi-resolution frequency encoder and decoder," in *Proc. ICASSP 2023*. IEEE, 2023, pp. 1–5.

[10] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning.," in *Proc. Interspeech 2017*, 2017, vol. 2017, pp. 1103–1107.

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

[12] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[13] Y. Gao, C. Chu, and T. Kawahara, "Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining," in *Proc. Interspeech 2023*, pp. 3637–3641.

[14] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms.," in *Proc. Interspeech 2017*, 2017, pp. 1089–1093.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP*, 2020, pp. 38–45.

[17] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[18] Y. Li, P. Bell, and C. Lai, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition," in *Proc. ICASSP 2022*. IEEE, pp. 7362–7366.

[19] H. Zou, Y. Si, C. Chen, D. Rajan, and E. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *Proc. ICASSP 2022*. IEEE, pp. 7367–7371.

[20] G. Ioannides, M. Owen, A. Fletcher, V. Rozgic, and C. Wang, "Towards Paralinguistic-Only Speech Representations for End-to-End Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 1853–1857.

[21] L. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *Proc. ICASSP 2023*. IEEE, pp. 1–5.

[22] Y. Fang, X. Xing, X. Xu, and W. Zhang, "Exploring Downstream Transfer of Self-Supervised Features for Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 3627–3631.