INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Diverse feature mapping and fusion via multitask learning for multilingual speech emotion recognition

Shi-wook Lee

National Institute of Advanced Industrial Science and Technology, Japan

s.lee@aist.go.jp

Abstract

In addition to linguistic information, speech contains nonlexical information, such as emotion, gender, and speaker identity. Recent self-supervised learning methods for speech representation can provide powerful initial feature spaces. However, a few training samples in speech emotion recognition cannot fully utilize the vast pretrained feature space. Herein, we propose an effective use of the feature space. First, to obtain more complementary information, diverse features are extracted by mapping the same utterance to different clusters via multitask learning. Thereafter, fusion methods are investigated according to the correlation among the diversely mapped features. The proposed methods are evaluated on two emotional speech corpora. The experimental results show that the proposed methods can effectively utilize the vast pretrained feature space and achieve state-of-the-art performance, with an unweighted average recall of 78.45% on the benchmark IEMOCAP corpus.

Index Terms: speech emotion recognition, self-supervised learning, fusion, multitask learning, multilingual

1. Introduction

Speech emotion recognition (SER) is a process for recognizing the emotional states of humans using speech signals. Deep learning has achieved excellent results in speech processing, and SER has gained popularity for empathic human-machine communication. Recently, large-scale self-supervised learning (SSL)-based pretrained speech models, such as wav2vec [1][2], HuBERT [3], and WavLM [4][5], have been developed. As these SSL-based models are trained on large-scale unlabeled speech corpora, they do not rely on any linguistic information [6]. Thus, they have been adopted as initial models or feature extractors for various downstream speech-processing tasks (e.g., SER), and their use has resulted in significant performance improvements [6][7].

Multitask learning (MTL) is an effective method for improving the performance of the main task by adding auxiliary tasks [8]. Using deep learning, MTL provides an effective approach for obtaining universal and task-invariant information from multiple tasks. MTL is applied to integrate utterance-wise contrastive loss with the SSL objective function for extracting unsupervised speaker information [5]. Furthermore, efforts, such as the development of domain-adversarial neural networks (DANNs), have been devoted to aggressively removing domain information and thereby improving generalization capability [9]. DANNs have been adopted for domain generalization with the purpose of learning common features from multiple domains such that class-discriminative information is emphasized and domain-specific information is removed [10]. However, a comparison study [11] reported that deep models trained on large datasets learn a speaker-invariant representation in automatic speech recognition (ASR), and the effect on the acoustic model is minor, regardless of whether MTL or DANN is used. Furthermore, removing domain-specific information using a DANN may cause the loss of class-discriminative information [12].

Fusion or ensemble methods combine several individual systems to improve performance. Because emotions can be expressed in different ways, such as facial expressions, body gestures, and speech, numerous studies have focused on fusion methods for multimodal emotion recognition [13][14]. The main achievement of fusion methods is based on the diversity of individual systems. Therefore, various individual systems lead to further improvement. Herein, we use a single modality, namely speech, to utilize SSL-based pretrained speech models and map an utterance into diverse features via MTL. As low correlation and similarly high accuracy among individual systems are required to maximize fusion effectiveness, we investigate the performance according to whether fusion is adopted in the intermediate or late stage.

Based on that the SSL-based pretrained speech model contains extensive information, our objective is to extract various heterogeneous features using MTL and integrate them, thereby improving performance. The most important considerations in improving the effectiveness of the fusion are comparatively high accuracy and low correlation among individual systems. The simplest fusion method is to combine multiple systems of different initial parameters. However, this method is implicit and heuristic and thus not guaranteed to obtain a low correlation, which results in low improvement by fusion. Our method of diverse mapping is explicit and supervised learning to obtain low correlation through domain-biased (MTL), -unbiased (DANN), and only emotion-based (Vanilla) features. The three main contributions of this study are as follows: (1) we leverage the pretrained HuBERT model for diverse feature mapping and fusion via MTL that achieves state-of-the-art SER results on the benchmark interactive emotional dyadic motion capture (IEMOCAP) [15] dataset, (2) we investigate two fusion methods, namely, intermediate concatenation and late combination and (3) we observe similar trends in performance improvement over two heterogeneous languages, English and Japanese, which confirms the effectiveness of the proposed method.

2. Method

2.1. Diverse feature mapping via multitask learning

This section briefly introduces MTL and its variant, adversarial learning. Fig. 1 shows the conceptual block diagram of the learning procedure in SER. MTL is typically used to extract common features across an emotion classifier (EC) and a

10.21437/Interspeech.2023-1425

domain classifier (DC). Thus, the features extracted from a feature extractor (FE) have both emotion and domain information. For adversarial learning in a DANN [9], the auxiliary task involves learning the DC that attempts to predict domains and simultaneously learning the FE to remove domain information by deceiving the DC.

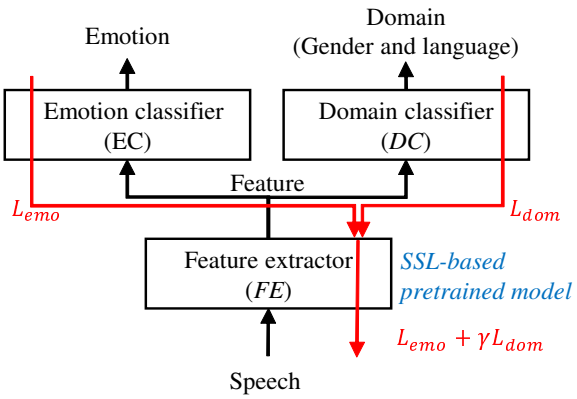

Figure 1: *Conceptual block diagram of the learning procedure. The red line is the backpropagation of losses (L_{emo}, L_{dom}) in learning. Here, $\gamma < 0$ for the DANN, $\gamma > 0$ for MTL, and $\gamma = 0$ for vanilla*

When training a model on multiple speech corpora, we can map an utterance into diverse features according to domain-variant in MTL, domain-invariant in the DANN, or neglecting domain information in vanilla. These diversely mapped features are expected to improve emotion classification through fusion methods. For simplicity, we use three values of γ: the vanilla system without a domain classifier ($\gamma = 0$), DANN ($\gamma = -1$), and MTL ($\gamma = 1$).

2.2. Additional diverse feature mapping

As the SSL-based pretrained speech model provides a large feature space, mapping features to different clusters is expected to reduce the correlation of intermediate features or results between individual systems. In addition to the three types of feature mapping (vanilla, MTL, and DANN) described in Section 2.1., we evaluate an additional feature mapping method for comparative evaluation. First, the three modules, namely FE, EC, and DC, are learned by MTL with both correct emotion and domain labels. Thereafter, the EC is fine-tuned again by correct emotion, the FE is fine-tuned again by correct emotion and fake domain label, and the DC is not fine-tuned to retain the correct domain information. Consequently, the features, that is, the output of the FE, are mapped into a different domain space and thus have correct emotions and incorrect domain information. Because we set four domains from two speech emotion corpora according to gender and language (English male and female speakers, Japanese male, and female speakers), we have four ways to remap features. Fig. 2 illustrates the process of forcing the features into one domain.

2.3. Fusion: intermediate vs. late

In this section, we discuss two fusion methods. Fusion is typically classified into three categories according to the implementation time: early data integration, intermediate feature concatenation, and late decision combination [14]. Herein, we investigate two fusion methods with the exception of the fusion of raw speech data. The first method is *intermediate concatenation*, which concatenates the output features of the FEs, SSL-based pretrained models, and the other method is *late combination*, which sums the output logits of the ECs. Fig. 3 illustrates these two fusion methods.

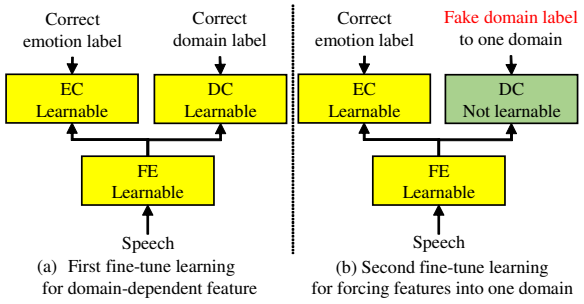

Figure 2: *Two-step fine-tune learning procedure for comparative evaluation*

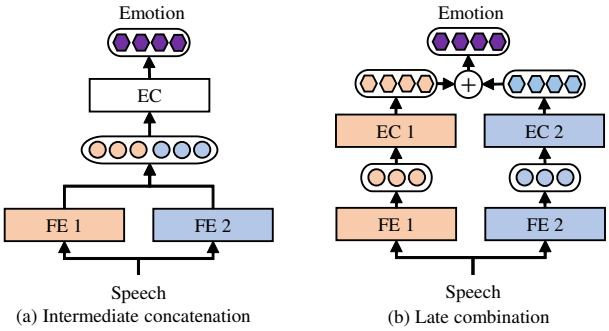

Figure 3: *Fusion methods: intermediate feature concatenation and late combination of summing logits*

3. Experimental setup

We conduct an experimental evaluation using two emotional speech corpora: the IEMOCAP corpus for English [15] and the Japanese Twitter-based emotional speech (JTES) corpus [16]. The IEMOCAP corpus comprises five sessions with five female and five male speakers. As each session is composed of one male and one female speaker, there are a total of five sessions. For classification experiments, we use 5531 samples from four emotional classes: angry, happy (excited), neutral, and sad. In the JTES corpus, 50 spoken sentences for each emotion are acted by 50 female and 50 male participants. Table 1 summarizes the corpora used in this study.

For experimental evaluation, all experiments are conducted in a leave-one-session-out five-fold cross-validation manner; one session is used as the test set, and the other four are used as the training set. This configuration is consistent with those of the IEMOCAP and JTES corpora. In the IEMOCAP, utterances from eight speakers are used as the training set, and those from the remaining two speakers are used as the test set. In the JTES, 100 speakers are equally divided into five sessions, with each session comprising ten female and ten male speak-

Table 1: *Number of utterances in total and by emotion class, speakers, and emotional expression in each of two emotional speech corpora*

Corpus	IEMOCAP	JTES
Language	English	Japanese
Angry	1,103 (20%)	5,000
Happy	1,636 (29%)	5,000
Neutral	1,708 (31%)	5,000
Sad	1,084 (20%)	5,000
No. of utterances	**5,531**	**20,000**
No. of speakers	10 (f:5, m:5)	100 (f:50, m:50)
Emotional expression	scripted & improvised	accurately acted

ers. Utterances from 80 speakers are used as the training set, and those from 20 speakers are used as the test set. This five-fold cross-validation method ensures that the experiments are speaker-independent and establishes four domains according to gender and language: English male speakers, English female speakers, Japanese male speakers, and Japanese female speakers. Owing to the data imbalance between classes, unweighted average recall (UAR) is basically used as an evaluation measure. As an SSL-based pretrained speech model for feature extraction, we adopt HuBERT (huber-large-ll60k) [3][17]. There are 317 million parameters in the HuBERT large model. Due to the small number of training samples compared to the large pretrained model size, all training samples are included in the training with no validation splits, and the test set is evaluated every 100 steps to find the best model [7]. For all experiments, we use a single NVIDIA Tesla A100 GPU.

4. Experimental results

4.1. Baseline results of single-corpus SER

For the baseline single-corpus SER, we evaluate ten trials of the vanilla system using the IEMOCAP and JTES. Table 2 lists the statistics, mean, standard deviation, maximum, and minimum of the ten UARs, where the ten individual systems are fine-tuned by different initial parameters for single-task SER.

Table 2: *Statistics of ten UARs on single-corpus single-task SER and the maximum UAR of late combination*

Test	IEMOCAP	JTES
Mean	73.11	99.91
Stdev	± 0.69	± 0.02
Max	**74.41**	**99.94**
Min	72.13	99.89
Late comb. (Max $_{10}C_k$)	**77.63**	**100**

The last row shows the maximum UAR among the combinations of the ten trials using the late combination, summing logits. The late combination is a combination of n trials taking k at a time without repetition, and the number is denoted by $_nC_k$, where $n \geq k \geq 0$. Compared with those obtained using the CNN-RNN-based system in the previous study [18], the baseline performances obtained using the SSL-based pretrained speech model, HuBERT, are significantly improved. The accuracy of JTES considerably exceeds that of IEMOCAP because the emotions of JTES are accurately acted upon the same 50 sentences; this condition is called the sentence closed condition [19]. Here, the model tends to memorize the sentence structure itself, besides emotion.

4.2. Results of multi-corpus SER

For the experimental evaluation of multi-corpus SER, we train multi-corpus SER models with two corpora but keep the same five-session and five-fold cross-validation configurations used for single-corpus SER. Furthermore, we set three types of learning in the experiments: vanilla, MTL, and DANN. Table 3 summarizes the results of ten trials of multi-corpus SER.

Table 3: *Statistics of ten UARs on multi-corpus SER and the maximum UAR of late combination*

Test	IEMOCAP		
Module	Vanilla	MTL	DANN
Mean	73.47	72.95	72.69
Stdev	± 0.55	± 0.74	± 0.57
Max	74.26	**74.46**	73.36
Min	72.79	71.98	71.60
Late comb. (Max $_{10}C_k$)	**77.34**	76.94	76.83
Test	JTES		
Module	Vanilla	MTL	DANN
Mean	99.78	99.79	99.75
Stdev	± 0.04	± 0.07	± 0.07
Max	99.84	**99.86**	99.84
Min	99.70	99.64	99.60
Late comb. (Max $_{10}C_k$)	**100**	**100**	**100**

As single-corpus SER is corpus-dependent, its performance (see Table 2) is slightly better than multi-corpus SER (see Table 3). The performance of DANN is inferior to those of vanilla and MTL. The removal of domain-specific information using a DANN may result in the loss of emotion-discriminative information, which may eventually degrade the classification performance. Notably, after the late combination, the performances are significantly improved in both single-corpus and multi-corpus SER and in the corpora, IEMOCAP, and JTES. These significant improvements by the late combination experimentally confirm that the SSL-based pretrained speech model has a vast feature space.

4.3. Results of additional diverse feature mapping

Table 4 summarizes the UARs of the diverse feature mapping using the fake domain labels described in Section 2.2.

Table 4: *Statistics of ten UARs of diversely mapped features using the fake domain label*

Test	IEMOCAP			
Fake	EM	EF	JM	JF
Mean	72.60	72.46	72.37	72.67
Max	73.55	73.24	73.56	73.41
Late comb. (Max $_{10}C_k$)	75.80	76.16	76.27	76.59
Test	JTES			
Fake	EM	EF	JM	JF
Mean	99.83	99.84	99.80	99.80
Max	99.88	99.94	99.90	99.89
Late comb. (Max $_{10}C_k$)	99.995	100	100	100

Because the domain is known in the learning stage, the domain can be assigned in a correct or fake way and thus the feature can be optionally mapped. Applying fake domain information, such as forcing features into one domain, may be considered a domain transformation. We set up four mapping methods according to gender and language, namely, English male speakers (EM), English female speakers (EF), Japanese male speakers (JM), and Japanese female speakers (JF). Because the feature has correct emotion information by fine-tuning with the correct emotion label but incorrect domain information caused by the fake domain label, the overall UAR is close to that of the DANN.

4.4. Intermediate and late fusion

This section compares intermediate and late fusions. Table 5 shows the UARs compared between intermediate concatenation and late combining.

Table 5: *Full integration of intermediate concatenation and late combination on the benchmark IEMOCAP dataset*

Full integration	${}_{30}C_{30}$	${}_{40}C_{40}$	${}_{70}C_{70}$
Intermediate concatenation	75.82	74.81	75.88
Late combination	77.31	76.76	77.51

Owing to the computational amount of the combination, we perform a full integration of 30 trials of vanilla, MTL, and DANN (denoted as ${}_{30}C_{30}$), 40 trials of additional diverse feature mapping, EM, EF, JM, and JF (denoted as ${}_{40}C_{40}$), and their sum of 70 trials (denoted as ${}_{70}C_{70}$). We omit the results of JTES from the table because all results reach a UAR of 100%. The model for JTES appears to have dominantly memorized the sentence structure itself rather than emotion; therefore, its generalization capacity for speech emotion recognition is hard to evaluate. The late combination of summing logits is superior to the intermediate concatenation. It is considered that more generalized features are obtained by intermediate concatenation, which is known as the bottleneck feature effect [20]. These generalized features are potentially helpful under domain-independent conditions but are inevitably poor under domain-dependent conditions. Experimental evaluation of domain generalization remains a topic for future work.

4.5. Best accuracy of the late combination

To verify the effectiveness of the late combination, we determine the best UAR from a combination of 30 trials taking k (${}_{30}C_k$) without repetition. The 78.45% of UAR and 77.69% of weighted average recall (WAR) is achieved from ${}_{30}C_9$, combining five vanilla, three MTL, and one DANN. The 78.45% UAR is better than the UARs in Table 3 achieved by the combination in each module. It shows that diverse feature mapping via MTL is effective.

Table 6: *Comparison of our method with previous state-of-the-art approaches on the benchmark IEMOCAP dataset.*

Metric	UAR (%)	WAR (%)
Co-attention [21]	72.70	71.64
MLT-Dnet [22]	73.01	-
GLAM [23]	73.90	73.70
Spk.-norm. [24]	-	74.20
Late comb. (our) (Max ${}_{30}C_9$)	**78.45**	**77.69**

We compare the best results with those obtained using other recent advanced methods; the results are listed in Table 6. Our best UAR outperforms the state-of-the-art method [23][24] by an absolute improvement of 4.55% UAR and 3.49% WAR. The comparison shows the effectiveness of the late combination of diversely mapped features via multitask learning, where the gradients of the auxiliary task are reversed and not. Figure 4 shows the confusion matrix of the best UAR (78.45%) on the benchmark IEMOCAP dataset.

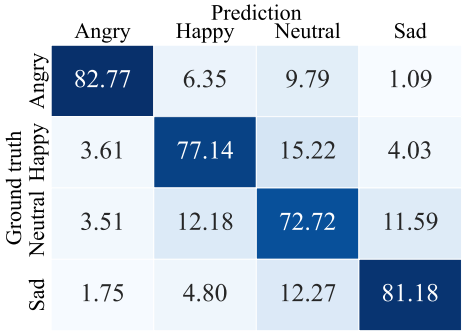

Figure 4: *Confusion matrix of the best UAR (78.45%) on the benchmark IEMOCAP dataset.*

5. Conclusion

We investigate the effectiveness of diverse feature mapping techniques using multitask learning and fusion methods. From the experimental evaluation of the single- and multi-corpus SER, we achieve state-of-the-art performance on the benchmark IEMOCAP dataset:78.45% for UAR and 77.69% for WAR. We confirm that, based on the large-scale feature space of the SSL-based pretrained speech model HuBERT, diverse feature mapping via multitask learning and fusion is effective for multi-corpus SER. More heterogeneous and similarly high performance among individual systems can be achieved from the SSL-based pretrained speech model HuBERT, resulting in an excellent fusion effect. In future studies, we must evaluate the generalization ability of intermediate concatenation and reduce the scale of fused systems via transfer learning, such as teacher-student networks. We are considering the sentence-independent settings of the JTES dataset, as well as the use of the other emotional speech datasets.

6. Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22K12105.

7. References

[1] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. of Interspeech*, 2019, pp. 3465–3469.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NIPS*, 2020, pp. 12 449–12 460.

[3] W.-n. Hsu,B. Bolte, Y.-h. h. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,"

IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.

[4] S. 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.

[5] S. Chen *et al.*, "Unispeech-Sat: Universal speech representation learning with speaker aware pre-training," in *Proc. of ICASSP*, 2022, pp. 6152–6156.

[6] A. Mohamed et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[7] S. w. Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. of Interspeech*, 2021, pp. 1194–1198.

[8] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2016.

[10] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. of CVPR*, 2019, pp. 5400–5409.

[11] Y. Adi, N. Zeghidour, R. Collobert, N. Usunier, V. Liptchinsky, and G. Synnaeve, "To reverse the gradient or not: an empirical comparison of adversarial and multi-task learning in speech recognition," in *Proc. of ICASSP*, 2019, pp. 3742–3746.

[12] A. Sicilia, X. Zhao, and S. J. Hwang, "Domain adversarial neural networks for domain generalization: when it works and how to improve," in *arXiv:2102.03924*, 2021.

[13] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[14] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Implicit fusion by joint audiovisual training for emotion recognition in mono modality," in *Proc. of ICASSP*, 2019, pp. 5861–5865.

[15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

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

[17] "HuBERT," https://github.com/pytorch/fairseq/tree/master/examples/hubert.

[18] S. -w. Lee, "Ensemble of domain adversarial neural networks for speech emotion recognition," in *Proc. of IEEE ASRU*, 2021, pp. 374–379.

[19] Y. Chiba, T. Nose, and A. Ito, "Multi-stream attention-based BLSTM with feature segmentation for speech emotion recognition," in *Proc. of Interspeech*, 2020, pp. 3301–3305.

[20] Y. Du, J. Xu, H. Xiong, Q. Qiu, X. Zhen, C. G. Snoek, and L. Shao, "Learning to learn with variational information bottleneck for domain generalization," in *Proc. of ECCV*, 2020, pp. 200–216.

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

[22] Mustaqeem, and S. Kwon, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Systems with Applications*, vol. 167, 2021.

[23] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *Proc. of ICASSP2022*, 2022, pp. 6437–6441.

[24] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *Proc. of ICASSP*, 2022, pp. 7342–7346.