SELECTIVE MULTI-TASK LEARNING FOR SPEECH EMOTION RECOGNITION USING CORPORA OF DIFFERENT STYLES

Heran Zhang, Masato Mimura, Tatsuya Kawahara

Kyoto University
Kyoto, Japan

Kenkichi Ishizuka

Revcomm, Inc.
Tokyo, Japan

ABSTRACT

While speech emotion recognition (SER) has been actively studied, the amount and variations of training data are limited compared with speech recognition and speaker recognition tasks. Therefore, it is promising to combine multiple corpora to train a generalized SER model. However, the manner of emotion expression is different according to the settings, task domains, and languages. In particular, there is a mismatch between acted datasets and spontaneous datasets since the former includes much more rich and explicit emotion expressions than the latter. In this paper, we investigate effective combination methods based on multi-task learning (MTL) considering the style attribute. We also hypothesize the neutral expression, which has the largest number of samples, is not affected by the style, and thus propose a selective MTL method that applies MTL to emotion categories except for the neutral category. Experimental evaluations using the IEMOCAP database and a call center dataset confirm the effect of the combination of the two corpora, MTL, and the proposed selective MTL.

Index Terms— speech emotion recognition, multiple corpora, multi-task learning, self-attention mechanism

1. INTRODUCTION

Recently, speech-based interactive systems have been deployed in many devices such as smartphones, smart speakers, and social robots, but the function and the ability of emotion recognition are still limited. Speech emotion recognition (SER) has much room to be improved [1, 2].

While there is a variety of settings and task domains related to SER, most of the conventional studies focused on acted emotion datasets due to the convenience of collecting data [3, 4, 5]. Studies in the field of psychology have shown that there is a big gap between the spontaneous and acted emotion expressions [6, 7]. Here, the acted emotion expressions include those expressed given a scenario (e.g. IEMOCAP [8] database and movie datasets such as AFEW [9]), and the spontaneous emotion expressions mean those expressed in real life. Apparently, the former has rich and explicit expressions compared with the latter.

There have been a limited number of studies in the real spontaneous SER. In [10], a Gaussian mixture model (GMM) based classifier was used with 3 different features, but the result of the 3-class emotion classification (Positive, Negative, and Neutral) is only slightly better than the random baseline. Recently, [11] showed that combining prior knowledge and classifier fusion improved the performance of SER in spontaneous speech, and [12] proposed a multi-scale deep convolutional LSTM for spontaneous speech.

On the other hand, recent studies on SER which focused on the acted datasets have given many new ideas and models to improve the performance. In [13], an attention pooling-based Convolutional Neural Network (CNN) was proposed for the SER task. Both [14] and [15] showed that the bidirectional LSTM (Bi-LSTM) multi-head self-attention model can bring a large improvement to SER. [16] and [17] showed that adding text features to the SER model can also bring significant improvements. However, it is not easy to automatically transcribe spontaneous and emotional speech in a real-world noisy environment, and a large word error rate would have a negative impact on the SER performance. Therefore, we do not use any text-level features but focus on the audio-based features in this study.

When we focus on the audio-based features, it is easy to combine multiple datasets regardless of the task domain and the language for training a general model. In this study, we explore a combination of an acted dataset and a spontaneous dataset for improving the SER performance on both datasets. It will mitigate the bottleneck of data sparseness in SER tasks, which is serious compared with automatic speech recognition tasks. In this case, we need to take into account the difference in the style of emotion expressions in the SER model. A simple solution is to use multi-task learning (MTL). Lee et al. [18] tried MTL with language and gender recognition, but did not show an improvement. On the other hand, Li et al. [15] showed the effect of MTL with gender recognition in the IEMOCAP database since the emotion expression is different between the genders.

In this work, we investigate MTL with style recognition, where the style is either acted or spontaneous. Moreover, we hypothesize that expressions of neutral emotion are universal and independent of the style and the datasets. Thus, we also propose a novel method of selective MTL, which applies MTL to emotion categories other than the neutral category.

We describe our baseline model and the proposed model in Section 2 and introduce the two corpora that we have used in this research in Section 3. Section 4 gives the details of the experimental settings, results and analysis. Finally, conclusions are given in Section 5.

2. MODEL ARCHITECTURE

2.1. Baseline Model

According to [14, 15, 16, 17], we adopt the Bi-LSTM encoder for SER. We denote the input feature as X, then the Bi-LSTM encodes it to a latent variables H, where $\boldsymbol{h}_t$ is the concatenation of the forward and the backward recurrent hidden state at time-step t.

According to [14, 15], we adopt the self-attention mechanism which computes an attention weight $\boldsymbol{a}_i$ by training two parameters W_1 and $\boldsymbol{\omega}_2$:

$$\boldsymbol{a}_i = softmax(\boldsymbol{\omega}_{2,i} tanh(W_{1,i} H^T)) \quad (1)$$

Finally, we get an output H_{att} by concatenating all i-th attention heads with their corresponding weight $\boldsymbol{a}_i$:

$$H_{att} = Concatenate(\boldsymbol{a}_1 H, ..., \boldsymbol{a}_n H) \quad (2)$$

Fig. 1 without the red rectangle shows our baseline SER model. Both Bi-LSTM layers and the self-attention mechanism play an important role. Finally, a fully connected (FC) layer with the ReLU activation function computes the output probability for each emotion category.

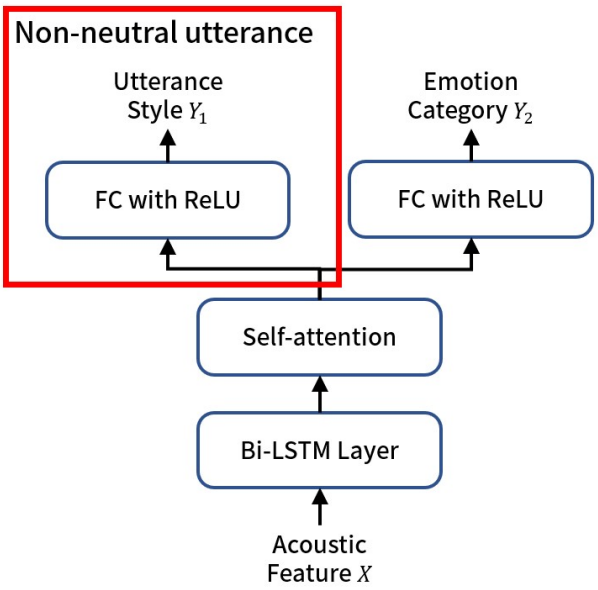

Fig. 1: Proposed Model Architecture.

2.2. Multi-task Learning (MTL) with Style Recognition

Multi-task learning (MTL) is introduced to effectively combine multiple datasets for SER. In this study, the style recognition task is added, where the style is either acted or spontaneous, which largely affects the emotional expression. The red rectangle in Fig. 1 shows the added network of FC for this task.

The entire neural network model is trained with the joint loss of the two tasks as below,

$$\mathcal{L} = \lambda \mathcal{L}_{SER} + (1 - \lambda)\mathcal{L}_{Style} \quad (3)$$

where L_{SER} denotes the loss of SER and L_{Style} denotes the loss of the style recognition. Each loss is defined by the cross-entropy between the predicted output and the ground-truth label.

2.3. Selective Multi-task Learning (MTL)

We also hypothesize that expression of neutral emotion is common among the speaking styles while that of happy, angry, and sad emotions are much different according to the style. Note that the number of neutral samples is usually the largest in real-world datasets, which means we need to build a stable model for the category. However, simply using a very large number of neutral samples in training results in a model biased to the neutral category, which cannot detect other emotion categories effectively.

Therefore, we propose a selective MTL method, where MTL is applied to the emotion categories except for the neutral category by using a similar number of samples for each category. The model architecture is the same as shown in Fig. 1 and Equation (3), but the style recognition loss is not computed for the neutral category. The value of λ becomes 1 for the neutral category.

3. DATASETS

In this study, we use two datasets to evaluate the proposed methods.

3.1. Acted Emotion Datasets (ACT)

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [8] has been widely used in SER studies. Some previous works regarded it as a "spontaneous" emotion corpus in that the subjects were not reading the text, but it is safe to classify it to "acted" datasets because the actors were performing for explicit emotions given a scenario (e.g., loss of a friend, separation). The emotional expressions were prepared and should be different from those observed in daily conversations in real life. It is also shown in [19] that professional actors have different speech patterns from untrained speakers, and thus the way of emotional expression should be different from ordinary persons.

Therefore, we choose IEMOCAP, one of the most widely-used English emotion corpora in the field of SER, as our acted emotion corpus. In this corpus, five male speakers and five female speakers are equally divided into five sessions, and each utterance has its transcription and emotion labels. Ten different kinds of emotion labels (angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other) are evaluated by at least three different annotators.

Following the conventions used in many previous studies using this database, we combine the utterances labeled "excitement" into "happiness", and only use data labeled "happiness", "sadness", "neutral", and "anger". Thus, we have 5,531 utterances in total and the details are given in Table 1.

3.2. Spontaneous Emotion Datasets (SPON)

For the spontaneous emotion dataset, we use recorded conversations in a call center, which was collected by the company of the last author. The telephone conversations between customers and operators are natural and spontaneous. They often involve emotional expressions such as "angry" and "happy" primarily from customers, which need to be detected, and that is the reason why we choose this corpus.

This is a Japanese corpus, but we do not use any linguistic information in this study and instead explore a general audio-based SER model independent of the language, following studies on multilingual speech processing [20]. Furthermore, it is known [21] that vocal tone is more dominant than the linguistic content in Asians' emotional expressions as Asian people tend to avoid using explicit linguistic emotional expressions [22].

This dataset has 109,933 utterances in total, and each utterance has its transcription and emotion labels evaluated by three different annotators. There are five emotion categories: angry, happy, disgust, sad and neutral. In the data cleaning process, we first ignored the samples that three annotators gave completely different labels.

We found that a large majority of samples are labeled "neutral", and many of the neutral utterances are "*hai* (Yes.)", "*ee* (Emm...)", and *noise*. We removed these utterances and randomly selected 2,000 utterances from the remaining neutral utterances to make the

number of samples comparable to other categories and also the IEMOCAP dataset. As a result, a total of 6,597 utterances are used, and their details are given in Table 1.

Table 1: Numbers of utterances used for each emotion class in acted and spontaneous datasets.

Database	Anger	Sadness	Happiness	Neutral
ACT	1103	1084	1636	1708
SPON	3029	692	876	2000

4. EXPERIMENTS

4.1. Data Preprocessing

For both acted utterances and spontaneous datasets, we used a 40-channel log mel-scale filter bank (lmfb) as an input feature vector using $python\_speech\_features$. Following previous work [15, 16], we set the maximal length of the utterance to 7.5 seconds, which means for utterances longer than 7.5 seconds, we only used the first 7.5 seconds, and we padded with zeros to the utterances that are less than 7.5 seconds.

We conduct the 5-fold cross validation in our experiments. We also tried to make SER experiments in a speaker-independent manner. For the acted database, IEMOCAP, we used four sessions for training and the remaining one session for testing to ensure the speaker-independent setting in each fold's validation. The spontaneous dataset did not have speaker labels, but there are so many (about 750) speakers in the dataset. Thus, we randomly divided the whole dataset to five equal parts and use one part for testing in our 5-fold cross validation.

4.2. Configuration and Hyperparameters

Our baseline and proposed models were implemented with Pytorch and trained using the Adam method [23] as an optimizer with a 10^{-4} learning rate and a 10^{-5} decay rate. We trained the model for 100 epochs. We also applied a dropout probability of 0.2 in each Bi-LSTM layer. The details of the model are summarized in Table 2. After preliminary experiments, we set the value of λ for MTL to 0.8, which weighs SER more than the style recognition.

Table 2: Settings of hyperparameters in training.

Notation	Meaning	Value
d_{LSTM}	Number of cells in Bi-LSTM	256
d_{Att}	Number of nodes in attention	512
d_{Den1}	Number of nodes in FC1	4096
d_{Den2}	Number of nodes in FC2	1024
n_{Head}	Number of heads in attention	8
n_{AttL}	Number of attention layers	1
λ	Loss Weights of SER	$0.1 - 0.9$

4.3. Results

Table 3 shows the overall SER results of UA (unweighted accuracy) and WA (weighted accuracy) for all emotion categories averaged over the two datasets. It presents a comparison of the baseline model, a combined model that simply combines the two datasets for training without MTL, the standard MTL model, and the selective MTL model proposed in this paper.

We also conducted an experiment by halving the number of training data amounts after the combination of the two datasets in order to offset the effect of the increased data amount. The results with this setting are indexed with "(half)" in the table.

Table 3: Overall SER results by baseline, combined, MTL, and selective MTL.

Training Data	UA	WA
$Baseline$		
−ACT+SPON (separate)	59.83%	62.08%
$Combined$		
−ACT+SPON	60.13%	63.09%
−ACT+SPON (half)	60.32%	60.06%
MTL		
−ACT+SPON	60.48%	63.21%
−ACT+SPON (half)	61.13%	60.99%
$Selective\ MTL$		
−ACT+SPON	**62.68%**	**65.15%**
−ACT+SPON (half)	61.06%	62.06%

The baseline performance of the acted database is UA of 55.65% and WA of 54.57%, which are comparable to those reported for the audio-only model in previous works [16]. A simple combination of the two datasets brings a slight improvement (UA by 0.30% and WA by 1.01% absolute, respectively). However, when we half the training data size, the WA is worse than the baseline. This is reasonable because the two datasets are different in style.

By introducing MTL, a significant improvement of accuracy is achieved (UA by 0.65% and WA by 1.13% absolute from the baseline). The improvement slightly surpassed that by the simple combination. With the selective MTL, a much larger improvement of accuracy is achieved (UA by 2.85% and WA by 3.07% absolute from the baseline). It significantly outperformed the standard MTL. Moreover, UA with the half training samples gets much better than the baseline (by 1.23% absolute), suggesting the model is generalized with these two different datasets. The results show the proposed selective MTL realizes effective training.

From these results, we can get the following primary conclusions. First, the simple combination of different datasets to increase the amount of training data is somewhat effective even if the style of the datasets is different. However, using MTL is more effective to mitigate the mismatch of the two datasets. Lastly, the proposed selective MTL can consider the mismatch more elaborately and further improves the SER performance.

4.4. Error Analysis

In this section, we analyze the performance in each corpus by looking at the detailed accuracy of each emotion category.

4.4.1. Acted Emotion Corpus

Table 4 shows the detailed accuracy of each emotion category for the acted dataset (IEMOCAP). It is observed that by simply combining spontaneous utterances, the performance of the neutral category has an accuracy improvement of 1% absolute and the performance of the happiness category has a much larger accuracy improvement

Table 4: Detailed results in acted corpus.

Model Name	Training Data	Emotion				UA	WA
		Anger	Sadness	Happiness	Neutral		
Baseline	ACT	**56.82%**	72.14%	48.82%	44.82%	55.65%	54.57%
Combined	ACT+SPON	55.38%	68.44%	52.72%	45.90%	55.61%	54.48%
MTL	ACT+SPON	53.14%	66.82%	53.55%	44.31%	54.45%	53.14%
Selective MTL	ACT+SPON	44.77%	**76.76%**	**64.42%**	**51.92%**	**59.47%**	**56.87%**
Ghosh et al. [24]	ACT (Audio Only)	53.58%	64.27%	36.98%	52.63%	51.86%	50.47%
Feng et al. [16]	ACT (Audio Only)	-	-	-	-	57.0%	55.7%
Chenchah et al. [25]	ACT+SPON (Audio Only)	-	-	-	-	-	50.06%

Table 5: Detailed results in spontaneous corpus.

Model Name	Training Data	Emotion				UA	WA
		Anger	Sadness	Happiness	Neutral		
Baseline	SPON	82.93%	31.39%	**71.46%**	53.65%	59.96%	67.17%
Combined	ACT+SPON	84.35%	30.06%	65.98%	60.30%	60.17%	68.93%
MTL	ACT+SPON	85.08%	**33.96%**	67.01%	61.10%	**61.79%**	70.04%
Selective MTL	ACT+SPON	**89.14%**	20.95%	67.12%	**61.75%**	59.74%	**70.76%**

of about 4%, accompanied by a slight decrease in other categories. By introducing MTL, the accuracy of any category is not significantly improved. The selective MTL brings a large improvement for all categories except for anger. It is primarily because of improved training of the neutral category. These results confirm our hypothesis: there is a big gap in emotional expression between acted and spontaneous conversations, but the expression of neutral emotion is almost shared.

We also compare our result with other SER models using the audio-only features in the IEMOCAP database in the lower part of Table 4. It is confirmed that our results are much better than others.

4.4.2. Spontaneous Emotion Corpus

Table 5 shows the detailed accuracy of each emotion category for the spontaneous dataset. We can see that using the combined datasets, MTL and selective MTL methods have a significant improvement in the accuracy of the anger and neutral categories. We reason that some variety of anger expressions are complemented by the acted dataset. As a result of the selective MTL, WA is significantly improved while UA is not changed so much. This is because of the enhancement of the neutral and anger categories, which are dominant in our spontaneous dataset.

4.5. Analysis of Style Classification

We also examine the results of the other task, the style classification. In all experiments, the accuracy of the style classification has never been lower than 99.95%. It turned to be a very easy task. In fact, there are many other factors to distinguish the two datasets: speakers (actors vs. ordinary people), language (English vs. Japanese), and recording conditions (studio vs. telephone).

It is hard to spot the primary factor, but in terms of SER, we suppose that the major difference is the style of acted vs. spontaneous. In any case, the results show the effectiveness of the selective MTL when combining the different kinds of datasets.

5. CONCLUSION

In this paper, we have investigated the effective use of multiple corpora for training an end-to-end speech emotion recognition model. MTL is introduced to mitigate the difference in the style. Moreover, we propose a novel method of selective MTL which considers the similarity in the neutral category. Experimental evaluations demonstrate the effectiveness of these methods.

This finding can make our research on SER no longer limited to acted speech corpus, and open applicability to many kinds of real-world application scenarios in human-machine or human-robot interactions.

6. REFERENCES

[1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] Björn W Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[3] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[4] Emily Mower, Maja J Matarić, and Shrikanth Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2010.

[5] Siqing Wu, Tiago H Falk, and Wai-Yip Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[6] Rebecca Jürgens, Kurt Hammerschmidt, and Julia Fischer, "Authentic and play-acted vocal emotion expressions reveal

acoustic differences," *Frontiers in psychology*, vol. 2, pp. 180, 2011.

[7] Patrik N Juslin, Petri Laukka, and Tanja Bänziger, "The mirror to our soul? comparisons of spontaneous and posed vocal expression of emotion," *Journal of nonverbal behavior*, vol. 42, no. 1, pp. 1–40, 2018.

[8] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[9] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 03, pp. 34–41, 2012.

[10] Daniel Neiberg, Kjell Elenius, Inger Karlsson, and Kornel Laskowski, "Emotion recognition in spontaneous speech," in *Proceedings of fonetik*. Citeseer, 2006, pp. 101–104.

[11] Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kumar Kopparapu, "Spontaneous speech emotion recognition using prior knowledge," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2866–2871.

[12] Shiqing Zhang, Xiaoming Zhao, and Qi Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional lstm," *IEEE Transactions on Affective Computing*, 2019.

[13] Pengcheng Li, Yan Song, Ian Vince McLoughlin, Wu Guo, and Li-Rong Dai, "An attention pooling based representation learning method for speech emotion recognition," 2018.

[14] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2822–2826.

[15] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning.," in *Interspeech*, 2019, pp. 2803–2807.

[16] Han Feng, Sei Ueno, and Tatsuya Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model.," in *INTERSPEECH*, 2020, pp. 501–505.

[17] Jennifer Santoso, Takeshi Yamada, Shoji Makino, Kenkichi Ishizuka, and Takekatsu Hiramura, "Speech emotion recognition based on attention weight correction using word-level confidence measure," *Proc. Interspeech 2021*, pp. 1947–1951, 2021.

[18] Shi-wook Lee, "The generalization effect for multilingual speech emotion recognition across heterogeneous languages," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5881–5885.

[19] Rebecca Jürgens, Annika Grass, Matthis Drolet, and Julia Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *Journal of nonverbal behavior*, vol. 39, no. 3, pp. 195–214, 2015.

[20] Tanja Schultz and Alex Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*. Citeseer, 1998, pp. 259–262.

[21] Edward T Hall, "Beyond culture. garden city," *NY: Anchor*, 1976.

[22] Marton Szemerey, "Linguistic representation of emotions in japanese and hungarian: Quantity and abstractness," *Acta Linguistica Asiatica*, vol. 2, no. 1, pp. 61–72, 2012.

[23] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, "Representation learning for speech emotion recognition.," in *Interspeech*, 2016, pp. 3603–3607.

[25] Farah Chenchah and Zied Lachiri, "Speech emotion recognition in acted and spontaneous context," *Procedia Computer Science*, vol. 39, pp. 139–145, 2014.