AN END-TO-END MULTITASK LEARNING MODEL TO IMPROVE SPEECH EMOTION RECOGNITION

Changzeng Fu[1,2], Chaoran Liu[2], Carlos Toshinori Ishi[2], Hiroshi Ishiguro[1,2]

[1]Graduate School of Engineering Science, Osaka University, Japan
[2]Advanced Telecommunications Research Institute International, Japan

ABSTRACT

In this paper, we propose an attention-based CNN-BLSTM model with the end-to-end (E2E) learning method. We first extract Mel-spectrogram from wav file instead of using handcrafted features. Then we adopt two types of attention mechanisms to let the model focuses on salient periods of speech emotions over the temporal dimension. Considering that there are many individual differences among people in expressing emotions, we incorporate speaker recognition as an auxiliary task. Moreover, since the training data set has a small sample size, we include data from another language as data augmentation. We evaluated the proposed method on SAVEE dataset by training it with single task, multitask, and cross-language. The evaluation shows that our proposed model achieves 73.62% for weighted accuracy and 71.11% for unweighted accuracy in the task of speech emotion recognition, which outperforms the baseline with 11.13 points.

Index Terms— speech emotion recognition, multitask learning, speaker recognition

1. INTRODUCTION

In recent years, SER has received increasing interest. SER focuses on using linguistic and acoustic attributes as input features and machine learning models as classifiers to recognize the emotions of speakers [1]. To achieve a good performance for SER is always a challenge because of the variability in signals of speech emotion and speaker-dependent features. A lot of works aimed to extract typical features for speech emotions, such as INTERSPEECH 2009 Emotion Challenge [2], the INTERSPEECH 2013 computational paralinguistics challenge [3], and AVEC challenge [4]. However, directly learning the mapping from speech spectrogram has emerged as a trend and this approach proved better in representing emotion in some cases [5, 6]. Although SER models can achieve a good performance no matter which type of input is used, they still treat these features as general representations of emotions and ignoring personalized differences. These differences, such as language, culture, gender, and age, could affect the emotion expressions [7]. In order to let the SER models notice the personalized differences, Sagha et al. incorporate individual factors (age, gender, personality) as a model selection strategy in valence recognition task [8]. Li et al. proposed a personalized attribute-aware attention network to capture personalized attributions in expressing emotions [9].

In this paper, we propose an attention-based CNN-BLSTM model using the end-to-end (E2E) learning method. We trained the model with single task (emotion recognition), multitask (emotion recognition and speaker recognition), cross-language with single task, and cross-language with multitask respectively. The proposed method was evaluated on SAVEE [10]. Jiang et al. [11] proposed a parallelized convolutional recurrent neural network (PCRN) with spectral features for SER. In our work we treat their work as the baseline. The major contributions of this work are summarized as follows.

1) Classifying emotions using Mel-spectrogram based self-attention CNN-BLSTM plus local attention in an E2E manner.

2) Combining emotion classification and speaker classification using multitask learning.

3) Training the model with cross-language data to improve SER.

The paper is structured in the following way: Some related works are presented in Section 2. In Section 3, we describe the details of our proposed method. The experiments along with the evaluation results are described in Section 4. Some discussion and future works are presented in Section 5. Section 6 is a brief summary of our work.

2. RELATED WORKS

With the recent introduction of deep neural networks to the domain of emotion recognition, there are more DNN-based models emerged. Since the emotion-related information lies in the time sequence of textual/auditory/visual contents, modeling these temporal contexts effectively is a key for the emotion recognition task. The most common approach is using a recurrent neural network (RNN) and its variants. Long short

This work was supported by JST, ERATO, Grant Number JPMJER1401.

term memory (LSTM) has been used in several works to improve the model's ability to catch the long term dependency in a time series [12, 13]. Connectionist temporal classification (CTC) style cost function is used along with LSTM to align the within-utterance emotional expressions with the labels [14]. On the other hand, convolutional neural networks (CNN) have achieved the state of the art performances regarding several time series modeling tasks such as machine translation [15] and language modeling [16]. There are only a few works using CNN for SER [17].

Multitask learning recently has arisen as an approach to improve models' performance in the SER field. Some researchers consider to use multitask learning to share multiple related corpora [18] or jointly analyzing the arousal, valence or dominance [19]. Besides that, for catching the interesting part in extracted features, several types of attention mechanisms are also added to the SER model and showed its effectiveness in emotion prediction tasks [9, 19].

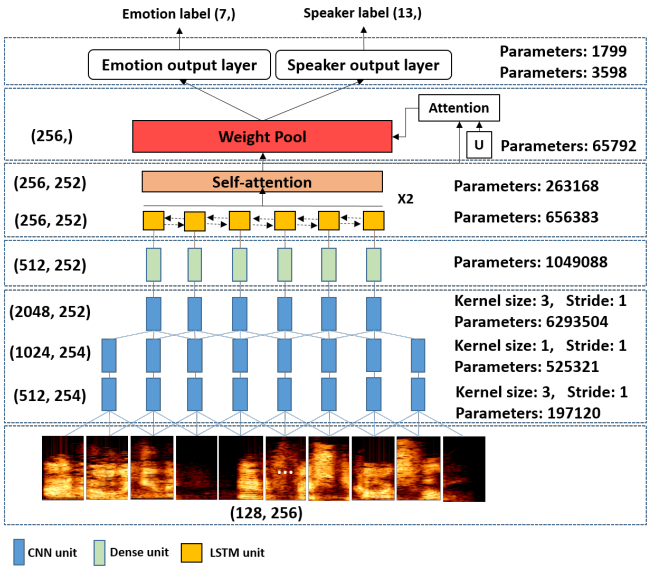

Fig. 1. Architecture of Proposed Model

3. PROPOSED METHOD

3.1. Dataset

We used two speech databases which are commonly used in speech emotion recognition studies: Surrey Audio-Visual Expressed Emotion database (SAVEE) [10] and Berlin Database of Emotional Speech (EMO-DB) [20]. SAVEE database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total with labels as anger, disgust, fear, happiness, neutral, sadness and surprise. EMO-DB is a German speech emotion recognition dataset with about 500 utterances, 10 actors (5 males and 5 females) and 7 emotions of anger, boredom, disgust, fear, happiness, sadness and neutral. We adopted leave-one-speaker-out (speaker-independent) strategy as previous research did.

3.2. Spectrogram Extraction

For the data pre-processing, we extracted spectrogram from each wav file by using a python package called librosa[1]. The sampling rate was set to 16000Hz. A Fourier transform of length 800 with hop length 400 was adopted. The computed spectrogram was mapped into Mel scale, which approximates the mapping of frequencies to patches of nerves in the cochlea to mimic the non-liner human ear perception of sound. A padding processing was implemented on extracted data, where the time length longer than 256 timestamps was cut to 256, and padded to 256 with -1 when the length was shorter.

[1] https://librosa.github.io/librosa/index.html

3.3. Model

As shown in Figure 1, the proposed model mainly consists of 3 parts, namely a 3-layer CNN, a 2-layer self-attention based bidirectional LSTM, and a weighted-pooling layer with local attention.

Denoting that the spectrogram of each input as $X = \{x_1, x_2, ..., x_L\}$, where $x_i \in \mathbf{R^{d_{spec}}}$, L is the temporal length of spectrogram which is equal to 256, and d_{spec} is the dimension of spectrogram vector which is equal to 128 according to the method used for pre-processing.

The convolutional layers adopted in the proposed model are 1-dimension convolutional layer. The kernel size is set as 3 for the first and last convolutional layer to convolve with the layer input over the temporal dimension, while setting the kernel size as 1 for the middle convolutional layer to increase nonlinearity of features. The number of filters in each layer are 512, 1024, 2048 respectively. The final output of convolutional layers $H^{cnn} = \{h_1^{cnn}, h_2^{cnn}, ..., h_{fm}^{cnn}\}$ is passed to a Dense layer that serves as a bottleneck, where $h_i^{cnn} \in \mathbf{R^{n_f}}$, n_f is the number of filters, and fm is the length of feature map in the final convolutional layer. Then, the tensor $H^{bn} = \{h_1^{bn}, h_2^{bn}, ..., h_{fm}^{bn}\}$ is passed to the bidirectional LSTM, where $h_i^{bn} \in \mathbf{R^{bn}}$. The outputs of the bidirectional LSTM are $H^{blstm} = \{h_1^{blstm}, h_2^{blstm}, ..., h_{fm}^{blstm}\}$, where $h_i^{blstm} \in \mathbf{R^{2d_{lstm}}}$ is the concatenation of forward and backward hidden states, and d_{lstm} is 128. Given the hidden states H^{blstm} as input, the self-attention layer yields a sequence of hidden states after computing with attention weights.

$$h_i = Attention(H^{blstm}), i = 1, 2, 3, ...N_{head} \quad (1)$$

Table 1. Dimension details

Notation	Description	Value
d_{spec}	number of spectrogram features	128
n_f	number of filters in final CNN layer	2048
fm	feature map output by final CNN layer	252
bn	size of bottleneck	512
N_{head}	attention heads	16
d_{lstm}	LSTM hidden units	128

$$H^{att} = Concat(h_1, h_2, ...h_{N_{head}})W, \\ (i = 1, 2, 3, ...N_{head}, W \in \mathbf{R^{N_{head}fm \times 2d_{lstm}}}) \quad (2)$$

where $Attention$ is a scaled-dot attention applied in parallel on the input of this layer with N_{head} set to 16. Before passing the tensors to Dense layers and yield the final output, the proposed model gives a local attention [5] operation to emphasize the most contributed hidden states. Given the external trainable parameters $U \in \mathbf{R^{2d_{lstm}}}$, the weights and output are calculated as:

$$\alpha_i = e^{UH_i^{att}} / \sum_{i=1} e^{UH_i^{att}} \quad (3)$$

The obtained weights are used in a weighted average in temporal dimension to get the utterance-level representation:

$$O_{loc} = \sum_{i=1} \alpha_i H_i^{att} \quad (4)$$

Finally, the tensors are passed to two separated Dense layer in order to output the emotion label and speaker label, the loss function for this multitask learning is:

$$L = \gamma \times L_{emotion} + \theta \times L_{speaker} \quad (5)$$

3.4. Training with Cross-language

Humans have an innate set of emotions recognized universally. However, due to some subtle nuances among different languages, there would be an in-group advantage for more accurate recognition in native language. As multicultural communication becomes more intense, people can learn these nuances from a variety of social media, which may help people to better perceive the emotions of others. In our work, we propose to include some datasets in another language to serve as data augmentation when training the model. According to the similarity of the label, we pick EMO-DB as augmented data, which is a German dataset. However, there are some differences on the label set, so we only select the labels that SAVEE has, which are anger, disgust, fear, happiness, neutral and sadness.

Table 2. Comparison results

	WA	UA
PCRN [11]	62.49	59.40
– *ours*		
single task	65.51	64.70
multitask	69.96	69.79
cross-language training	67.39	66.63
multitask cross-language training	73.62	71.11

4. EXPERIMENT SETUP AND RESULTS

4.1. Experimental Setting

Our proposed model was built with Keras and set the optimizer as *rmsprop*. Weights for loss functions in each task were equal to 1. The batch and epochs were 64 and 300 respectively and the best model is saved after each iteration. The dropout rates were set to 0.35. We evaluated the performance of the speaker-independent SER task using both weighted accuracy (WA) and unweighted accuracy (UA). We compared the results of our proposed model with the latest baseline [11] can be found in the year 2019 (see the upper part of Table 2).

4.2. Results

Compared to the baseline PCRN system, our proposed method outperforms their best reported results by a large margin. We improved WA from 62.49% to 65.51% and UA from 59.40% to 64.70% by training the model with a single task and single language. The accuracy achieved 69.96% for WA and 69.79% for UA when trained the model with multitask and a single language. The cross-language training method had an advantage over WA and UA of about 1.88% and 1.93% for single task. The accuracy had a big jump when training the model with multitask and cross-language data, which achieved 73.62% for WA and 71.11% for UA, the F1-score also increased from 0.69 to 0.71. From the confusion matrices shown Table 3 and Table 4 we can find more details for the improvements. After adding the augmented data, the accuracy in anger, fear, neutral and surprise are improved, especially in anger. But there is some decrease in disgust, happiness and sadness. Meanwhile, the WA and UA of the accuracy on EMO-DB with 5 picked labels were 89.84% and 89.69% respectively. The accuracy for speaker recognition achieved 94%.

5. DISCUSSION

From the results, we can see that multitask learning combining speaker recognition could improve the performance of SER, especially in cross-language training. We consider it is

Table 3. Confusion matrix for multitask learning with single language

	Prediction (UA)						
	anger	disgust	fear	happiness	neutral	sadness	surprise
anger	58.33	0	25	16.67	0	0	0
disgust	0	88.89	0	0	11.11	0	0
fear	16.67	0	41.67	8.33	8.33	8.33	16.67
happiness	33.33	0	0	58.33	0	0	8.33
neutral	0	0	0	0	92.00	8.00	0
sadness	0	0	0	0	20.00	80.00	0
surprise	27.27	9.09	0	18.18	0	0	45.45

Table 4. Confusion matrix for multitask learning with cross-language

	Prediction (UA)						
	anger	disgust	fear	happiness	neutral	sadness	surprise
anger	83.33	0	0	8.33	0	0	8.33
disgust	9.09	54.54	0	9.09	18.18	9.09	0
fear	0	0	68.75	6.25	18.75	0	6.25
happiness	33.33	0	0	44.44	11.11	0	11.11
neutral	0	0	0	0	100	0	0
sadness	0	0	0	0	30.00	70.00	0
surprise	6.67	0	33.33	0	0	0	60.00

because the speaker recognition task helps the model to distinguish individual characteristics of each person in expressing emotions, and the differences are more evident across languages. Not only that, we found that incorporating speaker recognition task while training the model could prevent the model from over-fitting to some extent. We will attempt to figure out what mechanism in the interaction of speech emotion recognition and speaker recognition influence features picking and parameter updating.

The results in Table 2 indicate that cross-language training could improve the performance of SER. Regarding multitask learning with cross-language data, it brings benefits to overall performance especially in anger, fear, neural and surprise. This result may suggest that adding samples from another language could increase the diversity of the training data, and allows the model to learn features that are easily overlooked when training with a single language. Also, this result reveals that there would be some differences in expressing happiness and sadness between English and German since the accuracy of these two labels became lower than training with a single language. Additionally, we also consider the proportion of different languages included in the training data might influence the results. In our work, the training data from German and English are almost balanced, the accuracies for both languages was parallelly increasing during training. An idea suddenly came to our minds when we obtained these results: would the model emerge the in-group advantage on SER task if we adjust the proportion of different languages in the training data? Regarding this inference, one of our future works is to investigate the effects of different proportions of languages included in the training set in more detail.

6. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we proposed a CNN-BLSTM model with self-attention and local attention mechanism, and trained it with multitask learning in an E2E manner, operating on the Mel-spectrogram. Our best results achieved an increase of overall weighted accuracy by 11.13 points comparing to the baseline. In future work, we plan to delve deeper into the relationship between speaker recognition and SER, and figure out how speaker recognition and different proportions of languages can improve performance of SER.

7. ACKNOWLEDGMENT

This work was partly supported by Grant-in-Aid for Scientific Research on Innovative Areas JP20H05576, and JST, ERATO, Grant Number JPMJER1401.

8. REFERENCES

[1] Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid

Vogt, Vered Aharonson, and Noam Amir, "The automatic recognition of emotions in speech," in *Emotion-oriented systems*, pp. 71–99. Springer, 2011.

[2] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[3] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[4] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 3–8.

[5] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[6] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa, "Speech emotion recognition using spectrogram & phoneme embedding.," in *Interspeech*, 2018, pp. 3688–3692.

[7] David Matsumoto, "Are cultural differences in emotion regulation mediated by personality traits?," *Journal of Cross-Cultural Psychology*, vol. 37, no. 4, pp. 421–437, 2006.

[8] Hesam Sagha, Jun Deng, and Björn Schuller, "The effect of personality trait, age, and gender on the performance of automatic speech valence recognition," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 86–91.

[9] Jeng-Lin Li and Chi-Chun Lee, "Attentive to individual: A multimodal emotion recognition network with personalized attention profile," *Proc. Interspeech 2019*, pp. 211–215, 2019.

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

[11] Pengxu Jiang, Hongliang Fu, Huawei Tao, Peizhi Lei, and Li Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019.

[12] Changzeng Fu, Thilina Dissanayake, Kazufumi Hosoda, Takuya Maekawa, and Hiroshi Ishiguro, "Similarity of speech emotion in different languages revealed by a neural network with attention," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, 2020, pp. 381–386.

[13] Gaurav Sahu, "Multimodal speech emotion recognition and ambiguity resolution," *arXiv preprint arXiv:1904.06022*, 2019.

[14] Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, and Björn W Schuller, "Towards temporal modelling of categorical speech emotion recognition.," in *Interspeech*, 2018, pp. 932–936.

[15] Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin, "A convolutional encoder model for neural machine translation," *arXiv preprint arXiv:1611.02344*, 2016.

[16] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 933–941.

[17] Abdul Malik Badshah, Nasir Rahim, Noor Ullah, Jamil Ahmad, Khan Muhammad, Mi Young Lee, Soonil Kwon, and Sung Wook Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5571–5589, 2019.

[18] Huijuan Zhao, Zhijie Han, and Ruchuan Wang, "Speech emotion recognition based on multi-task learning," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity)*. IEEE, 2019, pp. 186–188.

[19] Zixing Zhang, Bingwen Wu, and Björn Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6705–6709.

[20] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.