INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speech Emotion Recognition based on Attention Weight Correction Using Word-level Confidence Measure

Jennifer Santoso[1], *Takeshi Yamada*[1], *Shoji Makino*[1,2], *Kenkichi Ishizuka*[3], *Takekatsu Hiramura*[3]

[1]University of Tsukuba, Japan
[2]Waseda University, Japan
[3]Revcomm, Inc., Japan

j.santoso@mmlab.cs.tsukuba.ac.jp, takeshi@cs.tsukuba.ac.jp, s.makino@waseda.jp, ishizuka@revcomm.co.jp, hiramura@revcomm.co.jp

Abstract

Emotion recognition is essential for human behavior analysis and possible through various inputs such as speech and images. However, in practical situations, such as in call center analysis, the available information is limited to speech. This leads to the study of speech emotion recognition (SER). Considering the complexity of emotions, SER is a challenging task. Recently, automatic speech recognition (ASR) has played a role in obtaining text information from speech. The combination of speech and ASR results has improved the SER performance. However, ASR results are highly affected by speech recognition errors. Although there is a method to improve ASR performance on emotional speech, it requires the fine-tuning of ASR, which is costly. To mitigate the errors in SER using ASR systems, we propose the use of the combination of a self-attention mechanism and a word-level confidence measure (CM), which indicates the reliability of ASR results, to reduce the importance of words with a high chance of error. Experimental results confirmed that the combination of self-attention mechanism and CM reduced the effects of incorrectly recognized words in ASR results, providing a better focus on words that determine emotion recognition. Our proposed method outperformed the state-of-the-art methods on the IEMOCAP dataset.

Index Terms: speech emotion recognition, confidence measure, automatic speech recognition, self-attention mechanism

1. Introduction

Emotion recognition is a key to improving the quality of human-to-human and human-to-machine interactions. In recent years, technological advancements have enabled emotion recognition systems to receive various inputs, including speech, facial expressions, gestures, and biological signals. In some studies on emotion recognition, the combination of these types of information and their classification using deep neural networks have also been investigated. One study included feature fusion and ensemble learning on both speech and text [1]. Other studies combined both speech and text from transcriptions by aligning them frame-to-frame and learning the contribution weights of both types of information [2] [3]. Another approach incorporates speech, text, and visual information such as video and motion capture, where all types of information are fused at the end of the network for classification[4]. In most previous studies, speech is the most commonly used as it has rich information that reflects the emotion and is most readily available. Moreover, in many practical situations where speech is the only information, such as in call center analysis, other types of information are unusable.

The limitation of information leads to the study of speech emotion recognition (SER) by classifying different emotions from speech. With the advancement of deep neural networks, recurrent neural networks to capture sequence and attention mechanism for importance weighting have helped feature extraction from speech, improving the SER performance [5] [6]. The recent spread of automatic speech recognition (ASR) systems has also enabled the extraction of textual information from speech. ASR opens up the possibility of combining speech and text without the need for human-based transcriptions. In several studies, SER combined with ASR results as the textual information has also been investigated. In one study, SER with ASR was conducted by independently training ASR, SER, and emotion recognition submodels from text and by fine-tuning the combined submodels [7]. In another method, the end-to-end training process was simplified by conducting SER with a joint task to tune the ASR performance on emotional speeches [8]. Despite the reduced ASR error rate, ASR performance remains insufficient. Moreover, the fine-tuning of ASR for emotional speeches is costly.

To solve this problem, we propose an approach to mitigate the effects of speech recognition errors using the information on the confidence measure (CM) [9], an indicator of the reliability of ASR results available upon recognition, as a textual feature in SER. The confidence measure adjusts the importance weights in text information according to the likelihood of a speech recognition error in each word, allowing us to mitigate the speech recognition error effects on SER performance without the need for the fine-tuning of ASR. Moreover, this approach has the prospect of being effective in mitigating the errors in SER using ASR systems on lower-resource languages, in which ASR has a considerably higher error rate. In this study, we investigated three approaches to using CM: as an early-fusion feature, as a late-fusion feature, and as a correction for word-level importance weighting. We conducted experiments on the IEMOCAP dataset [10] to verify the effectiveness of these three approaches. Finally, we compare the performance of our proposed method with the state-of-the-art method.

2. Baseline method

2.1. Overview

Figure 1 shows the classifier structure with acoustic and text information, which is based on many state-of-the-art SER methods as shown in this study [8]. There are three main parts: the acoustic feature extractor, the textual feature extractor, and the final emotion classifier that combines the output of the two previous parts. The flow of SER starts from extracting the acoustic features and textual features of an utterance. Next, to obtain intermediate representations, the acoustic features are fed to bidirectional long short-term memory (BLSTM) [11] and weighted

http://dx.doi.org/10.21437/Interspeech.2021-411

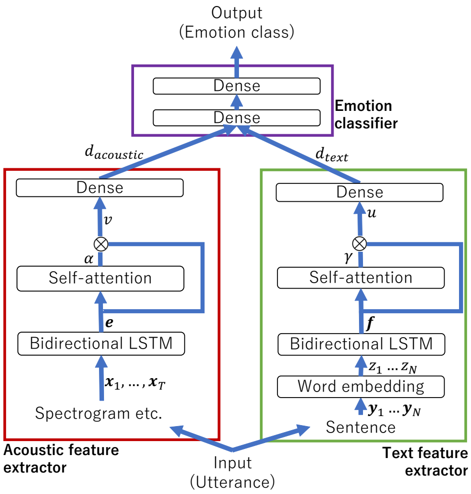


Figure 1: *Base method structure*

with a self-attention mechanism [12] . On the other hand, the textual features are encoded with word embedding and then processed similarly to the acoustic features. Finally, the intermediate representations are concatenated and classified with a fully connected network to obtain the final emotion class as the output. The details are discussed in the next subsections.

2.2. Acoustic feature processing

The first part processes the acoustic features from the utterances. The acoustic features extracted in this study contain a combination of Mel-frequency cepstrum coefficient (MFCC), constant Q-transform (CQT), and fundamental frequency (F0) as adopted from a previous study [13]. These combined features $\mathbf{x}_1 \ldots \mathbf{x}_T$, where $\mathbf{x}_i$ is the acoustic feature vector at the time frame i and T is the number of frames, provide more detailed information on the phoneme information, intonation, and pitches, which are significant in determining the emotion class. The acoustic features are fed to the BLSTM network for sequential feature extraction. The output vector $\mathbf{e}_i$ concatenates the forward and backward hidden states of BLSTM ($\mathbf{g}_i$ and $\mathbf{h}_i$), and is defined as

$$\mathbf{e}_i = \mathbf{g}_i \oplus \mathbf{h}_i, \tag{1}$$

where $\oplus$ represents concatenation. $\mathbf{e}_i$ is then weighed for its importance by the self-attention mechanism defined as

$$\alpha_i = softmax(\mathbf{w}_i tanh(\mathbf{W}\mathbf{e}_i^T)). \tag{2}$$

α_i is the attention weight for each frame, and $\mathbf{w}_i$ and $\mathbf{W}$ are trainable parameters. The self-attention mechanism has been proven to be effective in weighting important parts from sequential features and improving the performance of SER tasks [5][14]. Finally, the weighted sum $\mathbf{v}$ from BLSTM and attention weights defined as

$$\mathbf{v} = \sum_{i=1}^{T} \alpha_i \mathbf{e}_i \tag{3}$$

is calculated and fed to a single fully connected layer to obtain an intermediate representation, $\mathbf{d}_{acoustic}$, from the acoustic features.

2.3. Textual feature processing

The second part processes the texts, which are taken from the ASR result. In this study, the ASR system is a recognizer trained with the Librispeech [15] dataset and Kaldi speech recognition toolkit[16]. The recognition result from ASR $\mathbf{y}_1 \ldots \mathbf{y}_N$, where $\mathbf{y}_j$ is the input word and N is the number of words, is first encoded with a word-embedding model that is trained using BERT [17], a transformer-based [18] word embedding method that models contextual and positional word embeddings. The resulting features $\mathbf{z}_1 \ldots \mathbf{z}_N$ are then fed to textual feature extraction in the same manner as the acoustic feature extraction, using BLSTM to extract the sequence to vector $\mathbf{f}_j$, self-attention mechanism to weight the importance γ_j, and one dense layer to obtain the intermediate representation $\mathbf{d}_{text}$ of textual features. Note that this part uses ASR results directly, which has a high error rate and might decrease SER performance. To reduce the effect of speech recognition error in the ASR result, we proposed a method that is explained in the next section.

3. Proposed method

3.1. Overview

We propose using the combination of a self-attention mechanism and CM to mitigate the effects of errors from ASR results in textual feature extraction. We will explain how to improve the green frame in Fig. 1 and investigate three methods: early fusion, late fusion, and correction for word-level importance weighting.

3.2. Confidence measure (CM)

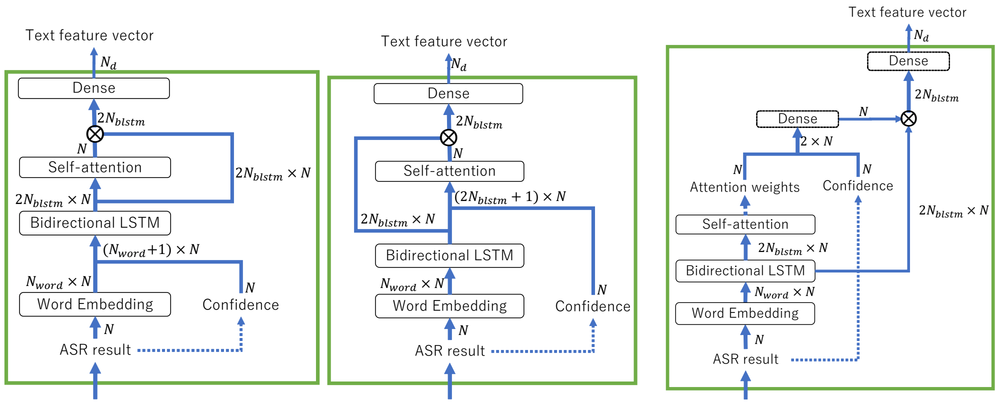
CM [9] is a metric that indicates the reliability of ASR decisions. CM has long been used in ASR systems for the evaluation of word-level and sentence-level recognition results and of how much can the words in the utterance recognized can be trusted on the basis of values between 0 to 1, accurately discriminating parts that contain speech recognition errors. The use of CM in textual features of SER may suppress incorrectly recognized words or emphasize the utterance part with the more correctly recognized words. This study employs the CM in the Kaldi speech recognition toolkit [16], which is defined by the log-likelihood difference between the first and the second best word predictions based on the lattice posterior. Here, CM is only applied to the textual feature extraction part of SER.

3.3. Application of CM in the proposed method

Figures 2(a), 2(b), 2(c) show the application of CM in the proposed method's architecture. In these figures, N, N_{word}, N_{blstm}, and N_d represents the number of words, the dimension number of word embedding, the unit number of BLSTM, and the dimension of the textual intermediate representation respectively. As CM indicates the reliability of ASR results, it is only applied in the textual feature extraction part of the proposed method. There are three mechanisms for the proposed method.

Early fusion (Fig. 2a) CM is treated directly as one of the textual embedding features as they are part of the ASR results and CM is represented by a sequence of weights, which might be suitable for extraction using BLSTM early on. In this method, CM is concatenated after the textual features have gone through BLSTM and before the self-attention mechanism.

Late fusion (Fig. 2b) As CM is small in dimension compared with the textual features, extracting CM sequentially in the early stages might cause early information loss and CM would not have markedly reduce speech recognition errors. Therefore, it would be more effective when used as one feature to consider aside from the extracted sequential text features for the self-attention mechanism. In this method, CM is con-



(a) Early fusion method (b) Late fusion method (c) CM as weighting correction

Figure 2: *Architecture of the proposed method on the textual feature extraction part of SER*

catenated after the textual features have gone through BLSTM and before the self-attention mechanism.

Attention weight correction (Fig. 2c) The previous two mechanisms use CM both directly and indirectly as part of the textual features. These mechanisms would require many training data for the incorrectly recognized words and their weighting. To solve this, we concatenate CM directly to the self-attention mechanism weights and update the weights through a fully connected network. By this method, one can decrease the CM dependence on the textual feature and train with fewer data. As CM indicates how reliable the ASR result is, CM values can act as another weight for ASR results, similarly to what the self-attention mechanism does for the textual feature extraction. The combination of two different weights provides more precise weights for textual information.

In this case, the CM c_j is concatenated with the attention weights α_j calculated using Eq. (2) as a new feature that will be fed to a fully connected layer and normalized using the softmax function to obtain new attention weights. The new attention mechanism is defined as

$$\beta_j = softmax(\mathbf{W}'(\alpha_j \oplus c_j)), \tag{4}$$

$$\mathbf{c'} = \sum_{j=1}^{N} \beta_j \mathbf{e}_j. \tag{5}$$

The fully connected layer represented by the trainable parameter $\mathbf{W}'$ learns and adjusts the attention weights by also considering the confidence measure aligned to the same word position. The resulting self-attention weights β_j are then used to calculate the weighted sum of the BLSTM outputs, producing the new weighted-summed features **c'**, which serve as the intermediate representations from the text.

4. Experiments

The experiment aims to examine the effectiveness of the proposed method in improving SER performance.

4.1. Dataset

In this study, we used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [10], one of the benchmark datasets for emotion recognition, to evaluate the effectiveness of the proposed method. The IEMOCAP dataset consists of both scripted and improvised emotional speeches divided into five sessions, each containing one male and one female speakers. In total, there are 10 speakers (five male and five female)

Table 1: *Dataset specifications*

Dataset	IEMOCAP	
Speakers	5 male and 5 female	
Utterance length	1−19 s	
# of utterances	Happy	1689
	Sad	1084
	Neutral	1708
	Angry	1103

in the IEMOCAP dataset. Each utterance corresponds to the transcriptions and is labeled as one of seven emotions (happy, sad, neutral, angry, excited, frustrated, and other). Similarly to previous works, we included the utterances labeled as excited to the utterances labeled as happy, and we only used four classes (happy, sad, neutral, and angry) for classification. We conducted five-fold cross-validation, in which four sessions were used as the training set and the remaining one session was used as the test set, ensuring speaker independence. The details about the dataset for this study are shown in Table 1.

4.2. Input feature

The features were divided into two parts for acoustic feature extraction and textual feature extraction. For the acoustic feature extraction, we extracted a 33-dimensional feature consisting of 20-dimensional MFCC, 12-dimensional CQT, and one-dimensional F0. For the textual features, first we conducted ASR on the utterances using a recognizer pretrained with the Librispeech [15] dataset and Kaldi speech recognition toolkit[16]. Librispeech consists of approximately 1000 hour of speech sampled at 16 kHz. Next, we encoded the ASR results using pretrained BERT [17], which was trained from lowercase English texts. The pretrained BERT consists of 12-layer, and 110M parameters, resulting in 768-dimensional textual features.

4.3. Classifier specifications

The acoustic extractor for the SER was comprised of two layers of BLSTM with 64 cells. Weighting was carried out using a self-attention mechanism with a single head and 128 cells. For the ASR part, we use the recognition result encoded by pretrained BERT as the input for the text feature extractor. The text feature extractor consist of two layers of BLSTM, a self-attention mechanism with a single head, and 128 cells. To combine both acoustic and ASR features, we used one fully con-

Table 2: *Baseline method comparison*

Method	UA	WA
Speech	61.1	64.3
Text (Transcript)	75.5	75.6
Text (ASR)	71.8	71.9
Speech + Text (Transcript)	78.6	78.4
Speech + Text (ASR)	73.9	74.2
Our proposed method		
Proposed 1	74.3	74.4
Proposed 2	74.9	75.4
Proposed 3	**75.9**	**76.1**

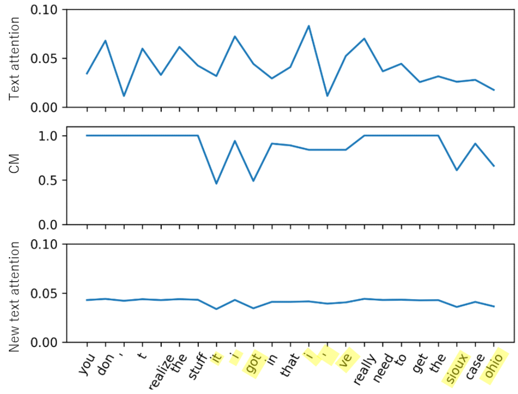


Figure 3: *Attention update visualization*

nected layer with 64 nodes, connecting the output from the self-attention layers. This was followed by a ReLU activation layer and an output layer with softmax to determine the final emotion class. In this experiment, we used Adam [19] as the optimizer with a learning rate of 0.0001 and a weight decay of 0.00001. The dropout was set to 0.3. The batch size was set to 40. The results were taken from the best out of 100 epochs.

4.4. Results

Table 2 shows the results of experiments on the methods with different feature combinations. The word error rate (WER) of the ASR on the IEMOCAP data was 43.5%, which is comparatively high due to the IEMOCAP containing emotions and spontaneous conversational words. As evaluation metrics for SER, we employed unweighted accuracy (UA) and weighted accuracy (WA) following the previous studies. As both metrics have similar tendencies, we analyzed the results primarily using WA. The method using only speech achieved a WA of 64.3%, and the method using transcripts only achieved a WA of 75.6%. The method combining speech and text from transcriptions achieved a WA of 78.4%, which was significantly higher than that achieved by a method using only speech (14.1%) or transcripts (2.8%). On the other hand, the WA of the method using text was degraded by 3.7% when using ASR result instead of transcriptions. Combining speech and ASR results yielded a WA of 74.2%, which indicates a performance degradation of 4.2%. Although the performance degradation was smaller than expected, this may be because the method using ASR result was trained with incorrectly recognized ASR results.

Our proposed methods on incorporating CM into the text feature yielded WA of 74.4%, 75.4%, and 76.1% for early fusion, late fusion, and attention weight correction, respectively. Among all the methods proposed, the attention weight correction showed the most significant performance gain from the result from speech and ASR and the closest performance to the one from the combined speech and transcription model. This also implies that the performance enhancement can be achieved by adjusting the attention weight of the ASR result extraction, in which the CM reduces the importance of the ASR result with a high chance of error. Furthermore, this method shows that the performance enhancement for SER with speech and ASR result is attainable without fine-tuning the ASR to recognize emotional speech.

Table 3: *Comparison of results of present work with those of previous works*

Method	Information source	UA	WA
Mirsamadi et al. [5]	Speech	58.8	63.5
Luo et al. [6]	Speech	63.9	60.3
Li et al. [14]	Speech	72.6	70.5
Chen et al. [1]	Speech + transcription	72.0	71.0
Liu et al. [3]	Speech + transcription	70.1	72.4
Feng et al. [8]	Speech + ASR	69.7	68.6
Heusser et al. [7]	Speech + ASR	71.0	73.5
Proposed method	Speech + ASR + CM	**75.9**	**76.1**

Table 3 shows the results of the present work in comparison with those of the previous studies along with the types of information in the IEMOCAP dataset with similar class settings. Our proposed method achieved the highest UA and WA. The close similarity between UA and WA in the result of our proposed method indicates the stability in conducting SER.

Fig. 3 shows the attention update mechanism in the proposed method of using CM as the correction to text attention weight, taken from a sample "angry" utterance with the correct text *"you don't realize the stuff that I've got in that I really need to get the suitcase okay"*. The top, middle, and bottom graphs show text attention weight, CM, and the updated text attention weights, respectively. The incorrectly recognized words are highlighted in yellow. Here, the incorrectly recognized "I've" is highly emphasized in the initial text attention, whereas some other correctly recognized words such as "realize" and "need" are not as heavily weighted. By applying the weight correction from CM, we found that the weight from "I've" and other incorrectly recognized words are decreased, and the weights of correctly recognized words are increased. This was at first incorrectly recognized as a "happy" utterance but corrected afterward to "angry". Note that as the updated attention weights affect the textual feature intermediate representation, the final emotion class is also determined by the acoustic feature intermediate representation.

5. Conclusions

In this paper, we proposed an SER method that uses ASR results as text information and word-level CM. We investigated the combination of a self-attention mechanism and a word-level CM in reducing the effects of speech recognition errors on the ASR results used as the text feature in SER on the IEMOCAP dataset. Among the three approaches proposed, that using CM as the attention weight correction yields the best performance. The result showed that our proposed method performs better than those in most of the previous studies using the IEMOCAP dataset. Our proposed method using CM can reduce the possible speech recognition errors through correction of importance weight in the ASR results. This opens the possibility of conducting SER with only acoustic features and ASR results without conducting ASR fine-tuning for emotional utterances.

6. References

[1] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition." Proc. Interspeech, pp. 374-378, 2020.

[2] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," Proc. ICASSP, pp. 2822-2826, 2019

[3] P. Liu, K. Li, H. Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," Proc. Interspeech, pp. 379–383, 2020.

[4] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," IEEE Intelligent Systems, vol. 33, no. 6, pp. 17-25, 2018.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," Proc. ICASSP, pp. 2227-2231, 2017.

[6] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," Proc. Interspeech, pp. 152–156, 2018.

[7] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pretrained language models," arXiv preprint arXiv:1912.02610, 2019.

[8] H. Feng, S. Ueno, and T. Kawahara, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR," Proc. Interspeech pp. 501–505, 2020.

[9] J. Hui, "Confidence measures for speech recognition: A survey," Speech Communication, vol. 45, no. 4, pp. 455–470, 2005.

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

[11] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," Proc. ICANN 2005, vol. 2, pp. 799–804, 2005.

[12] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," Proc. ICLR, 2017.

[13] Y. Xu, H. Xu, and J. Zou, "HGFM: A hierarchical grained and feature model for acoustic emotion recognition," Proc. ICASSP, pp. 6499-6503, 2020.

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

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain speech books," in Proc. ICASSP, pp. 5206–5210, 2015.

[16] D. Povey et al., "The Kaldi speech recognition toolkit," Proc. ASRU, pp. 1-4, 2011.

[17] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL, pp. 4171-4186, 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, pp. 5998-6008, 2017.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.