Audio Engineering Society

Convention Express

Paper 288

Presented at the AES 157th Convention
2024 October 8-10, New York, NY, USA

Real-time Speech Emotion Recognition for Human-robot Interaction

Jimin Jun[1] and Hong Kook Kim[1,2]

[1] *School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Korea*

[2] *AI Graduate School, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Korea*

Correspondence should be addressed to Hong Kook Kim (hongkook@gist.ac.kr)

ABSTRACT

In this paper, we propose a novel method for real-time speech emotion recognition (SER) tailored for human-robot interaction. Traditional SER techniques, which analyze entire utterances, often struggle in real-time scenarios due to their high latency. To overcome this challenge, the proposed method breaks down speech into short, overlapping segments and uses a soft voting mechanism to aggregate emotion probabilities in real time. The proposed real-time method is applied to an SER model comprising the pre-trained wav2vec 2.0 and a convolutional network for feature extraction and emotion classification, respectively. The performance of the proposed method was evaluated on the KEMDy19 dataset, a Korean emotion dataset focusing on four key emotions: anger, happiness, neutrality, and sadness. Consequently, applying the real-time method, which processed each segment with a duration of 0.5 or 3.0 seconds, resulted in relative reduction of unweighted accuracy by 10.61% or 5.08%, respectively, compared to the method that processed entire utterances. However, the real-time factor (RTF) was significantly improved.

1 Introduction

Emotion recognition, which quantitatively assesses human emotional states, plays a crucial role in enabling effective interaction between humans and artificial intelligence [1]. Among the various modalities used to detect human emotions, speech stands out for its efficiency, lower costs, and reduced privacy concerns compared to image- or video-based approaches, making it particularly advantageous for emotional interaction with robots [2]. Speech emotion recognition (SER) focuses on analyzing speech to identify emotions, offering a valuable tool for robots and AI to understand and respond to human emotions in spoken language [3]. Recent advancements in deep learning have significantly enhanced the SER capabilities. These methods use sophisticated neural networks to analyze speech signals and accurately determine the speaker's emotional state.

Traditionally, SER systems process single utterances—often several seconds long—to identify the corresponding emotional state [4-7]. While most SER systems are effective in batch processing scenarios, they are inadequate for real-time applications, where immediate response is essential.

Real-time processing is critical for seamless human-robot interaction [8]. As shown in Figure 1, when a conversation begins, a robot must quickly and accurately interpret both the spoken content and the

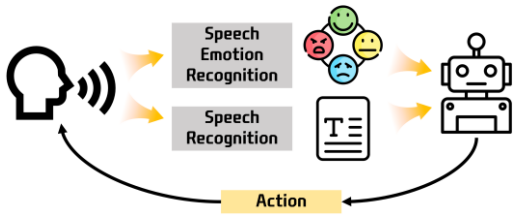

Figure 1. Illustration of real-time interaction between human and robot through speech and speech emotion recognition.

associated emotional cues using SER and speech recognition. This immediate responsiveness is essential for achieving dynamic and engaging interactions. Without real-time processing, the robot might perform well in controlled laboratory settings. However, in practical scenarios, it would likely struggle with delayed speech recognition and emotion interpretation, resulting in responses that fail to align with the user's immediate intentions. Efforts to improve SER have included applying the technology to short segments of each utterance [4-7]. Despite these attempts, segment-level models still struggle to capture the full context of an utterance in real-time, which limits their effectiveness in dynamic interactions.

To overcome these challenges, we propose a new approach for real-time SER. Instead of processing entire utterances, the proposed method divides each utterance into short, overlapping segments and estimates emotion probabilities for each segment. These probabilities are then aggregated using a soft voting mechanism, which progressively averages the vectors from subsequent frames to classify the overall emotion of the utterance. The soft voting method involves progressively averaging the probability vectors and determining the emotion class based on the highest-valued element in the averaged vector. This approach enables continuous real-time SER processing, making it suitable for dynamic human-robot interactions.

This study uses the SER model in [7] as the baseline and modifies it for real-time applications. The baseline model extracts acoustic features from a pre-trained model and processes them through a deep

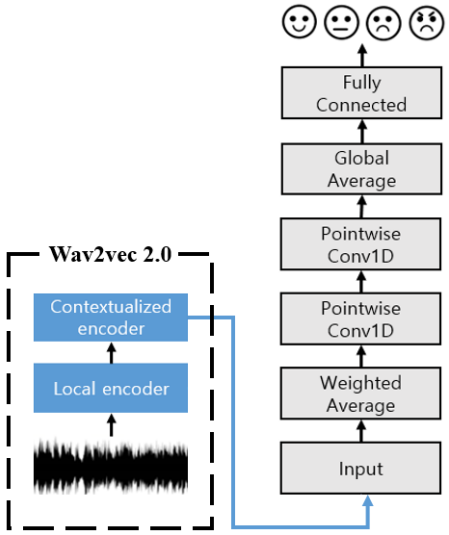

Figure 2. Network architecture of the baseline SER model.

neural network with a time-wise pooling layer, facilitating efficient real-time emotion classification.

Following this introduction, Section 2 outlines the architecture of the baseline model, and Section 3 proposes the real-time SER method. Then, Section 4 evaluates the performance of the proposed real-time SER model and compares it with that of the baseline model. Finally, Section 5 concludes this paper.

2 Baseline SER Model

In this study, the baseline SER model utilizes the pre-trained wav2vec 2.0 [9] to represent speech features. The embeddings generated by wav2vec 2.0 are then fed to a classifier, as shown in Figure 2. Wav2vec 2.0 is a self-supervised learning framework designed for extracting audio features, consisting of three main components: a local encoder, a contextualized encoder, and a quantization module. The local encoder processes the input audio through a series of convolution blocks. This representation is then refined by the contextualized encoder to create context-aware representations. Finally, the quantization module uses multiple codebooks to produce quantized representations based on the output of the local encoder.

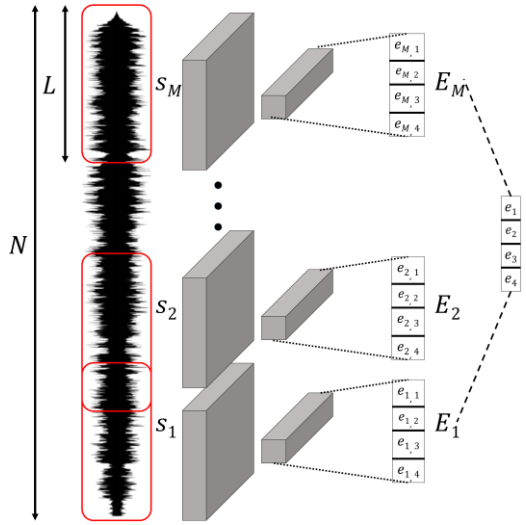

Figure 3. Procedure of the proposed real-time SER using soft voting.

We employ one of the downstream classification models used in [7] as the classifier. In other words, the features extracted from the contextualized encoder of wav2vec 2.0 are fed into a neural network consisting of two 1D pointwise convolutional layers with 128 hidden units each, followed by a rectified linear unit activation and dropout. Temporal information is then aggregated using a global average pooling layer, resulting in a 128-dimensional vector. Finally, emotion classification is performed by a fully connected layer with a softmax activation function.

3 Proposed Real-time SER Model

Although traditional SER models are effective in offline or controlled environments, they are not well-suited for real-time applications due to their inherent processing delays. Real-time SER is crucial for scenarios such as human-robot interaction, where delayed responses can disrupt the natural flow of communication. The challenge lies in the need to process emotional cues as they occur, without analyzing entire utterances.

Figure 3 illustrates the architecture of the proposed real-time SER model, comprising three main components: feature extraction, a downstream classifier, and a voting mechanism. As depicted in the figure, to achieve real-time emotion recognition, the voting method segments each utterance of length N into overlapping segments of length L, resulting in M segments, denoted as $\{s_i, i = 1, 2, \cdots, M\}$. The first segment, s_1, is processed by the SER model to generate an emotion probability vector, E_1. Each E_i is composed of probabilities $e_{i,1}$ to $e_{i,J}$, where J is the number of emotion classes (J = 4 in this study). This process is repeated for each subsequent segment from s_2 to s_M, producing corresponding probability vectors E_2 to E_M. These vectors are progressively averaged to predict the emotion class for the accumulated time up to $L \times M$ samples. This method ensures continuous real-time prediction of the speech emotion class.

Item		Statistics
No. of Speakers		40 (20 Males, 20 Females)
No. of utterances	Angry	1628
	Happy	1121
	Neutral	2859
	Sad	694
	Total	6302

Table 1. Distribution of the Korean speech emotion dataset (KEMDy19).

4 Performance Evaluation

4.1 Experimental Setup

To evaluate the model, we utilized the 2019 Korean Emotional Multi-modal Dataset (KEMDy19) [10]. Table 1 describes this dataset that includes a balanced gender distribution with 40 participants who simulated seven emotions: happiness, surprise, anger, neutrality, disgust, fear, and sadness. For consistency with prior studies in SER [4-7,11], we focused on four key emotions: anger, happiness, neutrality, and sadness. The KEMDy19 dataset, containing 6302 utterances, was divided into training, evaluation, and test sets with 5011, 665, and 626 samples, respectively. To ensure robust results, we employed an 8:1:1 split ratio for the data and performed 10-fold cross-validation, averaging the results across all folds.

Model	Segment-level		Utterance-level	
	UA (%)	WA (%)	UA (%)	WA (%)
MPGLN [9]	59.20	64.20	-	-
Conventional model (entire sentence)	-	-	77.84	76.43
Proposed model (L = 0.5 s)	68.47	67.18	69.58	68.32
Proposed model (L = 1.0 s)	73.21	72.03	71.22	69.89
Proposed model (L = 3.0 s)	75.16	74.35	73.89	73.12

Table 2. Performance comparison of unweighted and weighted accuracies of segment-level and utterance-level processing between the baseline and the proposed models according to different segment lengths.

The model was trained on a single Nvidia Tesla V100 GPU with the following settings: a learning rate of 0.001, a batch size of 32, and the Adam optimization algorithm. Performance was evaluated based on two metrics: unweighted accuracy (UA) and weighted accuracy (WA), where the former treated all classes equally but the latter considered class distribution.

4.2 Discussion

In this study, we investigated the impact of different segment lengths (L) on the performance of real-time SER, where the overlap length between two segments was set to 0.5 s. Table 2 compares the UAs and WAs of segment-level and utterance-level processing between the baseline and the proposed models according to different segment lengths. We first implemented the multi-path and group-loss-based network (MPGLN) [11], which is one of the state-of-the-art models on the KEMDy19 dataset. Note that the MPGLN was operated via the segment-level processing. In parallel, the baseline SER model shown in Figure 2 was evaluated via utterance-level processing. As shown in the table, the baseline SER model operating in the utterance-level processing provided UA and WA of 77.84% and 76.43%, respectively, which were much higher than those of the MPGLN.

Next, we evaluated the proposed real-time SER model by setting L = 0.5 s. Then, both UA and WA decreased compared with those of the baseline model. However, as L was increased from 0.5 to 3.0 s, both UA and WA improved at both the segment-level and the utterance-level processing. These results suggest that longer segment lengths provide richer contextual information, which enhances the model's ability to accurately classify emotions. Specifically, the proposed SER model with L = 3.0 s achieved UAs and WAs of 75.16% and 74.35% at the segment level and 73.89% and 73.12% at the utterance level, respectively. Although the UA and WA of the proposed SER model with L = 3.0 s were lower than those of the baseline, the proposed SER model was advantageous in real-time applications because an appropriate segment length can be selected. In other words, a shorter segment length may be preferred despite a slight reduction in accuracy for scenarios where real-time interaction is critical.

Finally, we calculated the processing delay and real-time factor (RTF) for the proposed SER models. By applying the soft voting mechanism only to the first segment, the processing delays were 0.519, 1.019, and 3.020 s, corresponding to RTFs of 0.173, 0.340, and 1.006 for L = 0.5, 1.0, and 3.0 s, respectively. This indicates that the choice of segment length should be traded off between the SER accuracy and processing time.

5 Conclusion

This paper proposed a segment-based real-time SER model designed to improve human-robot interaction by providing low-latency emotion detection. By leveraging the pre-trained wav2vec 2.0 model for feature extraction and incorporating a voting mechanism for real-time processing, the proposed SER model effectively addressed the limitations of traditional SER models that processed entire utterances. Overall, our findings highlight the importance of segment length for real-time SER systems. The results demonstrate the feasibility of implementing accurate, real-time emotion recognition in real-world applications, setting a new benchmark for future research in the field.

6 Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of robot and service contents supporting children's reading activities based on artificial intelligence, Project Number: R2022060001, Contribution Rate: 100%).

References

[1] Cowie R., et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, January (2001).

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, (2011).

[3] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, nos. 9-10, pp. 1062-1087, Nov. (2011).

[4] S. Padi, S. Gupta, A. Kothapalli, and A. Ganapathiraju, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proc. of 2021 International Conference on Multimodal Interaction*, Montreal, Canada, pp. 645-652, Oct. (2021).

[5] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on Spectrograms," in *Proc. of Interspeech*, pp. 1089-1093, (2017).

[6] M. Neumann, N. T. Vu, and A. Waibel, "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," in *Proc. of Interspeech*, pp. 1263-1267, (2017).

[7] L. Pepino, C. Lalanne, and M. Auli, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. of Interspeech*, Brno, Czech Republic, pp. 3400-3404, Sept. (2021).

[8] P. Foggia, A. Greco, A. Roberto, A. Saggese, and M. Vento, "A Social Robot Architecture for Personalized Real-Time Human–Robot Interaction," *IEEE Internet of Things Journal*, vol. 10, no. 24, pp. 22427-22439, (2023).

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NeurIPS*, Vancouver, Canada, pp. 12449-12460, Dec. (2020).

[10] K. J. Noh and H. Jeong, "KEMDy19," https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR

[11] K. J. Noh, H. Jeong, S. Kim, J. Lee, and Y. Lim, "Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets," *Sensors*, vol. 21, no. 5, p. 1579, Mar. (2021).