Neutral/Emotional Speech Classification Using Autoencoder and Output of Intermediate Layer in Emotion Recognizer *

☆ Jennifer Santoso[1], Takeshi Yamada[1], Kenkichi Ishizuka[2],
Taiichi Hashimoto[2], Shoji Makino[1,3]
[1]University of Tsukuba, [2]RevComm, [3]Waseda University

1 Introduction

Emotion recognition has been gaining attention from researchers due to the growing trends in the development of communication-based applications, such as conversation analysis and human-machine dialog systems. In recent years, there have been several deep-learning-based methods[1][2][3] to improve the performance of speech emotion recognition (SER). However, in most of these studies, [2][3], the recognition of neutral speeches, which is the most common type of speech in practical settings, tend to have poor performance. One of the reasons is the wide data distribution for neutral speeches, which is harder to generalize than emotional speeches.

In several practical settings, such as business conversation analysis, most conversations do not contain emotions. Emotional speeches, therefore, are considered an unusual occurrence and might be an indicator of trouble or unanticipated events in the conversation. Therefore, by taking advantage of a large number of neutral speeches available, it is possible to tackle problems in the SER from the anomaly detection approach, where neutral speeches are considered normal and emotional speeches are considered anomalous.

The anomaly detection approach has been investigated in the acoustics domain, namely the anomalous sound detection in machines [4]. The method uses raw spectrograms as input and an autoencoder as the reconstructor. This method successfully detected anomalies in faulty machines from sounds with high performance. However, reconstructing raw spectrograms is difficult, even in the audio domain. It is even more complicated when textual information is considered, such as those in SER tasks.

This study aims to improve the performance of the neutral speech classification in SER tasks. For this purpose, we propose a neutral/emotional speech classification method using an autoencoder and the output of an intermediate layer in a pretrained speech emotion recognizer. Instead of reconstructing a speech spectrogram, the proposed method reconstructs the intermediate layer representation extracted by a pretrained speech emotion recognizer. This enables the utilization of richer information compared to the speech spectrogram by using a speech emotion recognizer that extracts text features in addition to acoustic features. The reconstruction of intermediate layer representation has been studied in the field of image anomaly detection [5] and proven to be effective in improving the anomaly detection performance. Furthermore, the proposed method has the advantage that it does not depend on the length of the input speech, unlike the reconstruction of the spectrogram. In this paper, we investigate the performance of the proposed method by comparing the state-of-the-art SER methods.

2 Proposed method

2.1 Overview of the proposed method

The process flow of the proposed method is illustrated in Fig. 1. The proposed method consists of two parts: feature extractor and reconstructor. First, we extract the feature vectors from the input speech and its ASR text. These feature vectors are fed to the autoencoder-based reconstructor. The reconstructed feature vectors are then compared to the initially extracted feature vectors and have the reconstruction loss calculated as the anomaly score. When the anomaly score exceeds the decision threshold value, the input speech is classified as emotional (anomalous). Otherwise, it is classified as neutral (normal).

2.2 Pretrained feature extractor

The architecture of the pretrained feature extractor is illustrated in Fig. 2. In the proposed method, the feature extractor is based on the pretrained

[*] 感情認識器の中間層出力とオートエンコーダを用いた平静/感情音声の分類 by サントソ ジェニファー, 山田 武志（筑波大）, 石塚 賢吉, 橋本 泰一（RevComm）, 牧野 昭二（早大/筑波大）

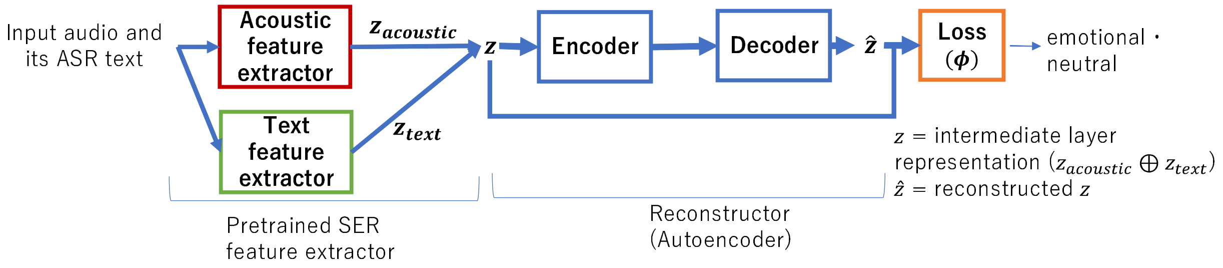
Input audio and its ASR text
Acoustic feature extractor
Text feature extractor
$z_{acoustic}$
z_{text}
z
Encoder
Decoder
$\hat{z}$
Loss (ϕ)
emotional · neutral
z = intermediate layer representation ($z_{acoustic} \oplus z_{text}$)
$\hat{z}$ = reconstructed z
Pretrained SER feature extractor
Reconstructor (Autoencoder)

Fig. 1 Proposed method flow

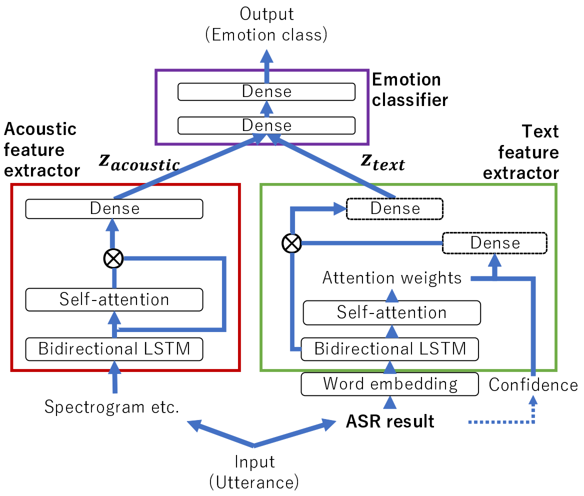

Fig. 2 Pretrained SER feature extractor structure

speech emotion recognizer [6] using acoustic features and its automatic speech recognition (ASR) text as the input. The method uses BLSTM [7] and self-attention mechanism [8] to extract acoustic features. In contrast, the text features are extracted through the attention weight correction to enhance the SER performance. The BLSTM enables the SER to handle various speech lengths. At the same time, the self-attention mechanism focuses on the parts containing important features for SER, resulting in a fixed-length feature vector for SER.

2.3 Autoencoder [9]

Autoencoder is a deep-learning architecture primarily used to represent higher-dimensional data, typically for dimensionality reduction efficiently. Autoencoder consists of an encoder and decoder to capture the most important parts of the input. The autoencoder is trained to minimize the average reconstruction loss of the normal data in the context of anomalous detection. In the proposed method, autoencoder is used to learn the representation of neutral speech through the intermediate layer classification $\mathbf{z} = \mathbf{z}_{acoustic} \oplus \mathbf{z}_{text}$ of the speech emotion recognizer.

Table 1 Speech data specifications

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

3 Experiments

In this section, we conduct experiments to prove the effectiveness of the proposed method as opposed to the state-of-the-art speech emotion recognition methods. Here, the speech emotion recognition is conducted on a 4-class (‘happy’, ‘sad’, ‘neutral’, and ‘angry’) setting. In accordance to the previous studies [1][3][6], the class ‘happy’ and ‘excited’ is merged into class ‘happy’.

3.1 Speech data

Table 1 shows the specifications of the speech data used in this experiment. 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 scripted and improvised emotional speeches divided into five sessions, each containing one male and one female speaker. There are ten speakers (five male and five female) in the IEMOCAP dataset. Each speech corresponds to the transcriptions and is labeled as one of seven emotions (‘happy’, ‘sad’, ‘neutral’, ‘angry’,

Table 2 Speech emotion recognizer and reconstructor specifications

Speech emotion recognizer	Acoustic: BLSTM (33-128), Attention unit: 128 Text : Pretrained BERT, BLSTM (768, 128), Attention unit: 128 Classifier: 256 - 64 - 4
Reconstructor	Autoencoder (256–256–256–256–16–256–256–256–256)
Optimizer (learning rate)	Adam (0.0001)
Dropout	0.3
Loss function	Mean Squared Error
Epoch	1, 2, ..., 200
Decision threshold	0.8

'excited', 'frustrated', and 'other'). We included the speeches labeled as 'excited' to the speeches labeled as 'happy'. 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.

For the pretrained feature extractor, we used the 4-class speech emotion recognizer that classifies 'neutral', 'happy', 'sad', and 'angry', which was trained using the data balanced among neutral and each of the respective emotion classes. We use only the neutral speeches as the training data for the reconstruction part.

3.2 Feature extraction

The features inputted to the pretrained feature extractor were divided into two parts for acoustic feature extraction and textual feature extraction [6]. For the acoustic feature extraction, we extracted a 33-dimensional feature consisting of 20-dimensional Mel-Frequency Cepstral Coefficients (MFCC), 12-dimensional Constant Q-transform (CQT), and one-dimensional fundamental frequency (F0). For the textual features, first, we conducted ASR on the input speeches using a recognizer pretrained with the Librispeech [11] dataset and Kaldi speech recognition toolkit[12]. Librispeech consists of approximately 1000 hours of speech sampled at 16 kHz. Next, we encoded the ASR results using pretrained BERT [13], which was trained from lower-case English texts. The pretrained BERT consists of 12-layer and 110M parameters, resulting in 768-dimensional textual features.

3.3 Speech emotion recognizer and reconstructor specifications

The pretrained speech emotion recognizer consists of two parts: feature extractor (acoustic feature extractor and textual feature extractor) and the emotion classifier. It has the same architecture as the speech emotion recognizer in [6]. In the proposed method, this is used as the pretrained feature extractor.

The reconstructor receives the resulting intermediate layer output vector from the pretrained speech emotion recognizer, which corresponds to the output before the final classification layer. The reconstructor is an autoencoder consisting of 7 layers, with its specifications shown in Table 2. The optimizer, learning rate, and dropout are the same as the speech emotion recognizer. However, the results were taken from the best out of 200 epochs. The reconstruction loss is calculated using Mean Square Error (MSE). The anomaly score is calculated from the Gamma distribution of the reconstruction loss of normal speeches for the reconstructor. The decision threshold is taken from the distributions' percentile point function with a value of 0.8. The speech will be deemed anomalous if the anomaly score exceeds the decision threshold. Otherwise, it will be deemed normal. The results were evaluated using F-score on the neutral class only. The evaluation metrics for F-score is defined in Eqs. (1)–(3):

$$F\text{-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \tag{2}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{3}$$

3.4 Experiment result and discussion

Fig. 3 shows the F-score of our proposed method compared to the state-of-the-art speech emotion recognizers. The methods except for the proposed method conducted SER on 4-class setting, and their

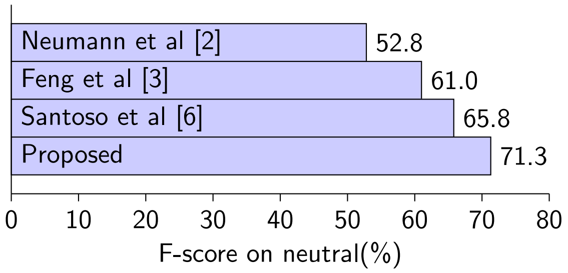

Fig. 3 Comparison of F-score on neutral class

F-scores on only the neutral class are shown in Fig. 3. Our proposed method obtains a higher F-score for neutral classes at 71.3%, at least 5.5% higher than these state-of-the-art speech emotion recognizers. The higher F-score is because the autoencoder is trained to learn the distribution of the neutral class through the intermediate feature vector representation, resulting in a better reconstruction and classification performance on neutral speech.

4 Conclusions

In this study, we proposed a neutral/emotional speech classification method using an autoencoder and the output of an intermediate layer in emotion recognizer and investigated its effectiveness. The use of the intermediate layer representation of speech emotion recognizer helps the autoencoder-based reconstructor obtain better vector representation and its distribution of neutral speeches, improving the classification performance of neutral speech.

Acknowledgements Part of this work was supported by JST SPRING, Grant Number JPMJSP2124.

References

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

[2] M. Neumann, N.T. Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” in Proc. Interspeech, pp. 1263-–1267, 2017.

[3] H. Feng, S. Ueno, T. Kawahara, “End-to-end speech emotion recognition combined with acoustic-to-word ASR,” Proc. Interspeech pp. 501–505, 2020.

[4] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, N. Harada, “Unsupervised detection of anomalous sound based on deep learning and the Neyman-Pearson lemma,” IEEE/ACM Transactions on Audio Speech and Language Processing, pp. 212—224, 2019.

[5] H. Okamoto, M. Suzuki, Y. Matsuo, “Out-of-distribution detection using joint probability between class and geometric transformation,” IPSJ Journal, vol. 62, no. 7, pp. 1382–1392, 2021.

[6] J. Santoso, T. Yamada, S. Makino, K. Ishizuka, T. Hiramura,“Speech emotion recognition based on attention weight correction using word-level confidence measure,” Proc. Interspeech, pp. 1947-–1951, 2021.

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

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

[9] D. E. Rumelhart, G. E. Hinton, R. J. Williams. “Learning internal representations by error propagation,” Parallel Distributed Processing. Vol 1: Foundations. MIT Press, Cambridge, MA, 1986.

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

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

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

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