Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 19(3): 148-154, Sep. 2021 **Regular paper**

Speech Emotion Recognition Using 2D-CNN with Mel-Frequency Cepstrum Coefficients

Youngsik Eom[1] and Junseong Bang[2*], ***Member, KIICE***

[1]Department of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
[2]Public Safety Intelligence Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejun 34129, Republic of Korea

Abstract

With the advent of context-aware computing, many attempts were made to understand emotions. Among these various attempts, Speech Emotion Recognition (SER) is a method of recognizing the speaker's emotions through speech information. The SER is successful in selecting distinctive 'features' and 'classifying' them in an appropriate way. In this paper, the performances of SER using neural network models (e.g., fully connected network (FCN), convolutional neural network (CNN)) with Mel-Frequency Cepstral Coefficients (MFCC) are examined in terms of the accuracy and distribution of emotion recognition. For Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, by tuning model parameters, a two-dimensional Convolutional Neural Network (2D-CNN) model with MFCC showed the best performance with an average accuracy of 88.54% for 5 emotions, anger, happiness, calm, fear, and sadness, of men and women. In addition, by examining the distribution of emotion recognition accuracies for neural network models, the 2D-CNN with MFCC can expect an overall accuracy of 75% or more.

Index Terms: Convolutional neural network, Deep learning, Mel-frequency cepstrum coefficients, Speech emotion recognition

I. INTRODUCTION

Recognizing emotions is a key ingredient of communication. The human-computer communication has recently been evolved beyond the simple Human-Computer Interaction (HCI) of the past into context-aware computing. Rather than deriving outputs that depend on the user's input, it must be able to recognize the situation and provide appropriate services. 'Emotion' provides important information during communication with people.

Consequently, recent studies have been conducted to recognize emotions through human facial expressions or utterances. Among them, speech information in utterance has a significant influence on how people perceive emotions. A voice can be detected differently depending on a particular emotion, which is associated with the temporal and spectral features of that voice. Therefore, previous studies attempted to implement speech emotion recognition based on the emotion dependence of the voice spectral features.

Numerous research have long been underway to exploit various features of SER. First, the study was conducted in feature extraction, separated by the temporal and spectral features. Temporal speech features can be accessed from time domain, such as amplitude, tempo, signal energy, zero crossing rate, maximum amplitude, and minimum energy. Some studies utilize tone information with speech signal tempo to perform an emotion recognition that incorporates the temporal feature [1]. Furthermore, another study was conducted considering speech speed together with Mel-Frequency Cepstral Coefficients (MFCC) feature [2]. However,

Received 17 May 2021, **Revised** 21 July 2021, **Accepted** 23 July 2021
***Corresponding Author** Junseong Bang (E-mail: hjbang21pp@etri.re.kr, Tel: +82-42-860-6165)
Public Safety Intelligence Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejun, 34129 Republic of Korea.

Open Access **https://doi.org/10.6109/jicce.2021.19.3.148** print ISSN: 2234-8255 online ISSN: 2234-8883

most studies, including the above, focused on spectral features rather than temporal features. MFCC, one of the spectral features, was extracted from the voice signal to classify happiness, sadness, and anger [3]. Some studies proposed an emotion recognition that classifies emotions by recognizing the spectral differences in emotions with spectrogram as a feature [4].

Subsequent studies have been conducted to enable self-learning of the features of raw waveform through deep learning rather than focusing on feature extraction. This allowed finding meaning and differences in the waveform itself without feature selection and extraction. Based on the Korean speech corpus, emotion recognition with raw waveform without applying feature extraction was proposed using convolutional neural network (CNN), recurrent neural network (RNN), and Attention mechanisms [5]. Furthermore, 1D-CNN with Long Short-Term Memory (LSTM) and 2D-CNN with LSTM models were proposed to self-learn features based on the raw waveform to classify emotions [6, 7]. Recent studies were conducted with transfer learning using VGG-19 [8] models trained with ImageNet [9], noting the image characteristics of spectral features such as MFCC or spectrogram [10].

To perform SER based on a deep learning model combined with spectral features, we introduce a 2D-CNN SER model utilizing MFCC, one of the spectral features. In this experiment, we effectively classified emotions using 2D-CNN with RAVDESS [11] dataset.

II. Speech Emotion Recognition using 2D-CNN with MFCC

The entire SER system is presented in Fig. 1. Target speech signal data enters into the signal processing block. In signal preprocessing, long data based on 3 seconds are trimmed and short data are zero-padded. After processing, the data is divided into segments and entered the feature extractor block. Feature extractor block generates MFCC feature with segments. When the data enter the neural network model, the model categorizes emotions (anger, happiness, calmness, fear, and sadness) based on the generated MFCC feature.

A. MFCC

We used MFCC as a spectral feature of the SER system. MFCC is created by applying cepstral analysis to the log mel-spectrogram. The cepstral analysis is an analyzing technique using log-scaled spectrum with inverse Fourier transforms. It allows us to obtain the fundamental frequency which helps identifying the distinct sound structure of the speaker. After converting raw audio data into spectrogram via Short Time Fourier Transform, a Mel-Frequency filter bank is applied to make spectrogram perceived similar to human cochlea. This is because the human auditory system is not linear, sensitive in low-frequency and insensitive in high-frequency bands. Then, we used log scale to perform cepstral analysis and obtain MFCC by performing Discrete Cosine Transform. The MFCC is used to indicate the spectral envelope that helps understand the tone and sound structure of a human vocal tract; hence, it can be used for SER as a spectral feature, which is varied by emotion states. Fig. 2

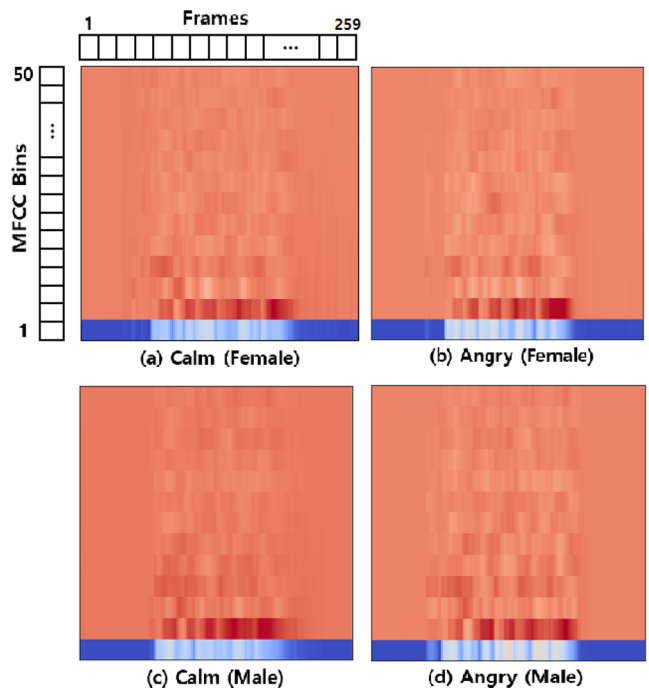

Fig. 2. MFCC on speeches in different emotions.

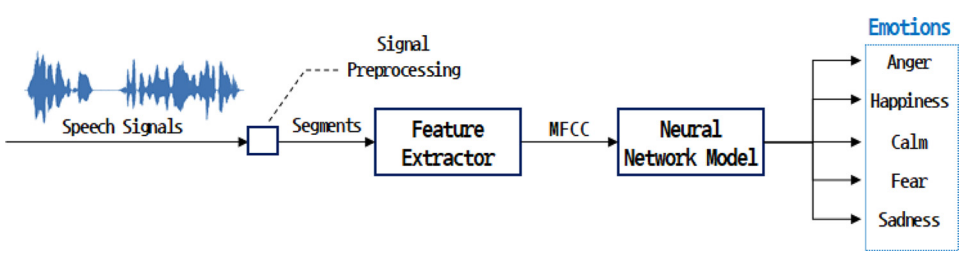

Fig. 1. Block diagram of SER system using MFCC.

shows the results of MFCC when men and women speak in different emotions for the same sentence, “Kids are talking by the door”. Here, x- and y-axes represent time and MFCC number, and color represents intensity. In this experiment, the model used the following MFCC variables: Sampling rate = 44.1 kHz, number of MFCC = 50, window length = 2048 samples, hop length = 512 samples, window = hann.

B. Deep Learning Based Models for Feature Extraction

1) FCN, LSTM, Bi-LSTM Models

A Fully Connected Network (FCN) is a type of neural network with all neurons and nodes in one layer connected to the next. It is sometimes referred to as a general neural network and is used to reduce the number of dimensions by connecting to the end of other networks. In our experiment, we used three dense layers connected, with 1024, 512, and 10 nodes each. The ReLu function was used as the activation function of the layer except for the last SoftMax layer, and batch normalization and dropout were applied respectively.

The LSTM is designed to solve the problem of RNN, learning ability degradation of long time series data due to the gradient vanishing problem. It is a structure that adds a cell state to the hidden state of RNN. In our experiment, we designed a many-to-many LSTM monolayer structure with 50 units to return the sequence. A TimeDistributed layer is also included with 50 nodes dense layer as input. The TimeDistributed layer is required to deal with the many-to-many problem; the cost is calculated for every time steps and back-propagates to the lower step.

A bi-directional LSTM (Bi-LSTM) is designed because both forward and reverse inferences of LSTM can have significant results. In the Bi-LSTM, LSTM is executed in both directions. We designed a many-to-many Bi-LSTM monolayer with 25 units, allowing the sequence to be returned, including the TimeDistributed layer.

2) CNN and Its Variants Models

Convolutional Neural Network (CNN) resolves the problem that loss of spatial and local information due to dimensionality reduction, a common FCN problem, when dealing with data containing spatial information such as images. Generally, it consists of convolutional layers, which perform convolution operation of input features and an FCN at the end.

1D-CNN performs convolution operation only in a one-way direction; in contrast, 2D-CNN's convolution kernel performs in two dimensions. This can also be used for the analysis of time series data [12]. Our experiment consisted of three convolution layers, each containing 128 kernels with a length of 5. It was zero-padded to preserve information on edges and used ReLu activation. In addition, each layer was conducted pooling operation for size 4 and included a dropout of 0.1 and batch normalization.

As previously stated, the 2D-CNN model uses the MFCC feature as input, expressed as a 2D matrix. This model consists of five convolution blocks, including convolution layers and pooling layers. The input size of the model was (N × 50 × 259 × 1), where N is the number of input sample data, 50 is the number of feature dimensions, 259 is the number of frames per sample, and 1 is the depth of the channel per sample. Some studies show that excessive convolution layers negatively effect the SER accuracy when using CNN [13]. The above experiment analyzed how the number of convolution layers affects the implementation of CNN with Bi-LSTM. Above study showed that up to 4 convolution layers, the performance of a neural network improves with the number of layers; however, the performance sharply drops for more than 4 layers. We obtained the optimal number of layers by changing the number of layers in this experiment; we confirmed that 5 convolution layers are suitable. Each convolution layer in a block has a kernel to extract local features. The first layer has 64 convolution kernels, and the second, third, fourth, fifth layers have 128, 256, 512, 512 convolution kernels each. All the kernels, except the fifth layer have the size of (3 × 3); the fifth layer has kernels with a size of (2 × 2). Each layer was zero-padded and the stride was fixed with a size of (1 × 1). In addition, the activation function of each layer used the ReLu function; at the end of the layer, the number of parameters was reduced by pooling that size of (2 × 2). In order to prevent over-fitting, batch normalization and dropout were applied to each block. The dropout ratio was 0.1, with 10 % neurons randomly excluded from the training for every iteration during the process.

As previously mentioned, the method was introduced, which can strengthen the time series characteristics of the extracted feature maps by combining CNN with LSTM [6, 7]. In this experiment, for 1D-CNN with LSTM, we combine a many-to-many LSTM monolayer structure with 128 units that allow sequences’ return, including a TimeDistributed layer. The 1D-CNN Bi-LSTM is the same as LSTM; however, ithas 64 units. For 2D-CNN with LSTM, we combined a many-to-many LSTM monolayer structure with 512 units and, allowing sequences’ return, including a TimeDistributed layer. The 2D-CNN with Bi-LSTM is the same as LSTM; however, it has 256 units.

C. Classification

The feature extraction performed by deep learning models, described earlier, generates feature matrices with enhanced original spectral characteristics. For classification, the features are first converted into one-dimensional feature vectors via the flatten layer to be suitable for the classifier. The converted feature vectors are finally classified through 10 nodes

Table 1. SER accuracies for various neural network models

Model	Accuracy (%)
FCN	72.22
LSTM	70.14
Bi-LSTM	63.54
1D CNN	87.47
1D CNN + LSTM	81.94
1D CNN + Bi-LSTM	86.46
2D CNN	**88.54**
2D CNN + LSTM	84.03
2D CNN + Bi-LSTM	78.47

of the SoftMax classifier; here, 10 represents the number of classes to be classified. (5 emotions for each male and female).

III. RESULTS

A. Dataset

RAVDESS [11] (Ryerson Audio-Visual Database of Emotional Speech and Song) is the dataset of emotional media. It contains 7356 files of English emotional speech, song, audio, and video data: including neutral, anger, happiness, calm, surprise, fear, sadness, and disgust. For the experiment, we separated only 1440 speech files; 12 male and 12 female actors vocalize each of the following two sentences with different emotions: “Kids are talking by the door,” “Dogs are sitting by the door.” This characteristic of the dataset, consisting only of two sentences help model to distinguish the speaker’s voice characteristics, not the linguistic features. This experiment used only 960 speech data, consisting of anger, happiness, calm, fear, sadness classes; the total 1440 only-speech data additionally consist of neutral, surprise, and disgust classes. Besides, we labeled the samples with distinguishing gender in the same class (gender-specific) because the differences of pitch / energy between males and females can influence classifying speech emotion [14]. Most sample data are 2 to 4 seconds long. Therefore, we decided that there would be meaningful information for three seconds from the beginning, so we sampled the data for only three seconds altogether. Thus, information of more than 3 seconds was cut off, and information of less than 3 seconds was zero-padded. The sampling information is as follows: Sampling frequency = 44.1 kHZ, resample type = kaiser_best.

B. Experimental setting

We developed the model and experimented with it using Keras and Tensorflow framework. The operating system was NVIDIA Quadro RTX 6000 on Linux with Docker environment. Various functions included in the LibROSA [15] library were used to process audio data and extract spectral features such as MFCC. In order to train the model, categorical cross entropy was used as a loss function; an RMSprop with a learning rate of 0.00001 and a decay of 1e-6 was used as an optimizer. During the learning, the training was conducted with 'accuracy' as a metric; batch size was considered 128, 2000 epochs.

As mentioned earlier, the number of learning data is 960, classified into five emotions; each was labeled with 10 classes separated by gender. The training and test set were divided by 7:3, and data augmentation was performed only for significantly small training data. Therefore, three data augmentation techniques were applied, corresponding to adding white noise, 15% left shifting, and 15% right shifting. In this case, data augmentation using white noise is randomly applied with a weight of 0.001 using a random function. This allowed the training set to be 4 times larger comparing that of no data augmentation.

C. Results & discussion

After training 2000 epochs, the 2D-CNN model showed 88.54% accuracy for the test set. The training and test sets were separated earlier; hence, the model never used the test set earlier. Therefore, we confirm that the model shows high accuracy for the new data. The accuracy graphs and confusion matrices of all experimented models are shown in Figs. 3 and 4.

Fig. 3 shows SER accuracies according to training epochs for various neural network models. The blue and orange lines represent accuracies for training and test sets. For each, the horizontal and vertical axes refer to the number of epochs and accuracy. Fig. 3 also shows that CNN models, especially a 2D-CNN model, exhibit the highest accuracy for test sets compared to other models.

We trained and tested other network models such as FCN, LSTM, Bi-LSTM, 1D-CNN and 2D-CNN with their variants with the same dataset and environment to compare the results. Consequently, we obtained that FCN shows accuracy of 72.22%; LSTM, Bi-LSTM, 1D CNN, 1D CNN with LSTM, 1D CNN with Bi-LSTM, 2D-CNN with LSTM, and 2D-CNN with Bi-LSTM show accuracies of 70.14%, 63.54%, 87.47%, 81.94%, 86.46%, 84.03%, and 78.47%, respectively.

In addition, 2D-CNN showed high accuracy for each emotion and average accuracy for the entire emotion. Fig. 4 shows SER confusion matrices for various neural network models. The x- and y-axes represent model predicted and real labels. Thus, the diagonal components of the matrix allow us to determine the exact predicted proportions of the model for each label. The graphs below the confusion matrix represent the accuracy of each label that the model predicted

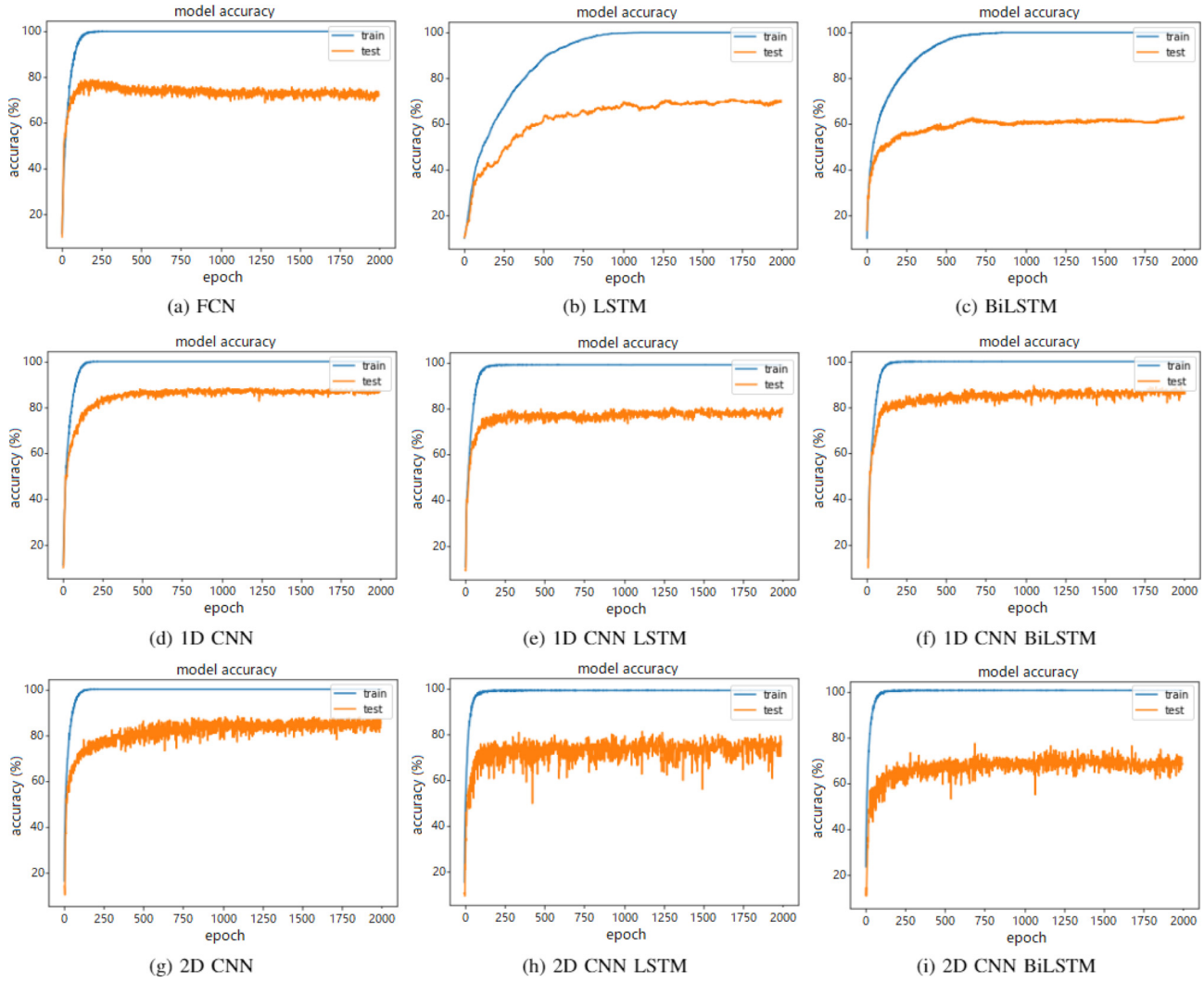

Fig. 3. SER accuracies on training epochs for various neural network models.

correctly. The vertical axes represent accuracies; the horizontal axes represents the following labels: f_a = female angry, f_c = female calm, f_f = female fear, f_h = female happy, f_s = female sad, m_a = male angry, m_c = male calm, m_f = male fear, m_h = male happy, m_s = male sad. Less than 75% accuracy was shown in red; 75% or more accuracy was shown in green. Overall, some labels were not distinguished well by each model, especially for the male happy label. However, we confirmed that the 2D-CNN performs best regardless of the label with high accuracy, including the male happy label.

We confirmed that 2D-CNN exhibits the best performance when using the MFCC feature as a feature vector. The MFCC is an audio feature; however, it appears in the form of a 2D matrix. Hence, the advantage of CNN, such as well-identifying spatial and regional characteristics of data, performs successfully. Furthermore, we confirmed that simple CNN performance is better in our experiments than other existing studies [6, 7] where performance increased when LSTM or Bi-LSTM was combined with CNN. Since, unlike previous studies, where information about time series characteristics was important using raw waveform as input, MFCC feature was used as input in our work; in addition, performance improvement was not achieved due to loss of time series information by repetition of pooling layer.

IV. CONCLUSIONS

This paper examines the accuracy and the performance distribution of SER using models of the FCN, LSTM, Bi-LSTM, 1D-CNN, 1D-CNN with LSTM, 1D-CNN with Bi-LSTM, 2D-CNN, 2D-CNN with LSTM, and 2D-CNN with Bi-LSTM. For the RAVDESS dataset, compared to other

Fig. 4. SER confusion matrices for various neural network models.

neural network models, the 2D-CNN model with the MFCC showed high emotion recognition accuracy for all types of emotions (5 types here) with an average of 88.54% accuracy, showing the best performance. Therefore, we plan to collect large amounts of data and apply various data augmentation techniques for designing better models in a future study of the classifying emotion of Korean corpus.

ACKNOWLEDGEMENTS

This research was supported and funded by the Korean National Police Agency. [Pol-Bot Development for Conversational Police Knowledge Services / PR09-01-000-20]

REFERENCES

[1] S. Byun and S. Lee, "Emotion recognition using tone and tempo based on voice for IoT," *Trans. of the Korean Institute of Electrical Engineers*, vol. 65, no. 1, pp. 116-121, 2016. DOI: 10.5370/kiee.2016.65.1.116.

[2] I. Hong, Y. Ko, Y. Kim, and H. Shin, "A study on the emotional feature composed of the mel-frequency cepstral coefficient and the speech speed," *Journal of Computing Science and Engineering*, vol. 13, no. 4, pp. 131-140, 2019. DOI: 10.5626/JCSE.2019.13.4.131

[3] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in *2017 Int. Conf. on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2257-2260, Mar. 2017. DOI: 10.1109/WiSPNET.2017.8300161.

[4] S. Park, D. Kim, S. Kwon, and N. Park, "Speech emotion recognition based on CNN using spectrogram," in *Information and Control Symposium*, pp. 240-241, Oct. 2018.

[5] J. Lee, H. Ryu, D. Chang, and M. Koo, "End-to-end Korean speech emotion recognition using deep neural networks," in *Korea Computer Congress*, pp. 1000-1002, Jun. 2018.

[6] G. Tangriberganov, T. A. Adesuyi, and B. Kim, "A hybrid approach for speech emotion recognition using 1D-CNN LSTM," in *Korea Computer Congress*, pp. 833-835, July. 2020.

[7] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200-5204, Mar. 2016. DOI: 10.1109/ICASSP.2016.7472669.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv: 1409.1556, 2014.

[9] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 248-255, Jun. 2009. DOI: 10.1109/CVPR.2009.5206848.

[10] J. Lee, U. Yoon, and G. Jo, "CNN-based speech emotion recognition model applying transfer learning and attention mechanism," *Journal of KIISE*, vol. 47, no. 7, pp. 665-673, 2020. DOI: 10.5626/JOK.2020.47.7.665

[11] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English," PLoS ONE, vol. 13, no. 5, pp. e0196391, May. 2018. DOI: 10.1371/journal.pone.0196391.

[12] W. Tang, G. Long, L. Liu, T. Zhou, J. Jiang, and M. Blumenstein, "Rethinking 1D-CNN for time series classification: a stronger baseline," arXiv: 2002.10061, 2020.

[13] L. Huang, J. Dong, D. Zhou, and Q. Zhang, "Speech emotion recognition based on three-channel feature fusion of CNN and BiLSTM," in *2020 the 4th International Conference on Innovation in Artificial Intelligence (ICIAI)*, pp. 52-58, May. 2020. DOI: 10.1145/3390557.3394317

[14] P. Mishra and R. Sharma, "Gender differentiated convolutional neural networks for speech emotion recognition," in *12th Int. Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pp. 142-148, Oct. 2020. DOI: 10.1109/ICUMT51630.2020.9222412.

[15] librosa [Internet]. Available: https://librosa.org/doc/latest/index.html.

Youngsik Eom
received the B.S. degree in Electronic and Electrical Engineering from Sungkyunkwan University, Suwon, Republic of Korea, in 2021. Since 2021, he has been studying for an M.S. degree with the Korea Advanced Institute of Science and Technology, Daejun, Republic of Korea. His research interests include speech recognition and NLU.

Junseong Bang
received the B.S. degree in Computer Science Engineering from Hanyang University, Ansan, Republic of Korea, in 2006; He completed his M.S. and Ph.D. degrees in Information and Communications from the Gwangju Institute of Science and Technology, Gwangju, Republic of Korea, in 2009 and 2013, respectively. Since 2013, he has been with Electronics and Telecommunications Research Institute (ETRI), Daejun, Republic of Korea. In 2016, he joined the University of Science and Technology (UST), Daejun, Republic of Korea. He is currently a Senior Researcher with ETRI and an Associate Professor with UST. His research interests include contextual computing, XR, computer vision, speech recognition, and conversational chatbot.