Equilibratory Data Augmenting Machine for Speech Emotion Recognition

Jung Hwan Kim, Alwin Poulose, Woo-Sung Son and Dong Seog Han*
School of Electronics and Electrical Engineering, Kyungpook National University
Republic of Korea
jkim267@knu.ac.kr, alwinpoulosepalatty@knu.ac.kr, sonws1230@knu.ac.kr, dshan@knu.ac.kr*

Abstract

The speech emotion recognition (SER) system detects a driver's emotion to prevent road rage related-accidents. The SER system handles the audio signals to recognize the driver's emotion. While building the SER system's foundation, we obtained the publicly available SER's dataset, called the Ryerson audio-visual database of emotion speech and song (RADVESS). The reliable SER system is supposed to detect the driver's emotion even in noisy environments. Nevertheless, we discovered that even simple noise could reduce the SER system's performance. In this paper, we propose an equilibratory data augmenting machine (E-DAM) on the RADVESS dataset. The proposed E-DAM could marginally improve the SER system's performance by adding noise and magnify the amount of RADVESS's samples. The performance of the SER system originally reached 71%, but it increased by approximately 3% after applying E-DAM.

I. Introduction

The speech emotion recognition (SER) system detects a driver's voice and emotion. The benefit of researching the SER dataset could detect the driver emotion more accurately, accompanying our previous research, the facial emotion recognition (FER) system [1-4]. Detecting a driver's emotion via speech is considerably challenging for many beginning researchers since it requires solid knowledge of signal processing. The driver's speech could contain environmental noises such as raindrops, car engines, and passengers' speech. Those noises could disrupt detecting a driver's emotion, although the sound is audible for our ears.

In addition, building the foundation of the SER system requires any SER dataset, so we obtained the Ryerson audio-visual database of emotion speech and song (RADVESS) [5]. The RADVESS contains samples of video, audio, and songs. Our SER system is trained with only 1,440 speech samples to detect a driver's emotion as the infuriated driver does not sing a song to threaten other drivers. Each audio wave file's name represents the properties of the audio file. The third numerical characters of the audio file name represent the speech emotion classification and have 8 different emotions: neutral, calm, happy, sad, angry, fear, disgust, and surprise. The length of each audio is 3 seconds. We merged the neutral and calm audio samples since the SER's classifications must match the FER's emotion classifications, which are 7 emotions: angry, disgust, fear, happiness, neutral, sadness, and surprise. We installed and applied Librosa [6] to load and pre-process the raw audio file. Then, all audio signals are

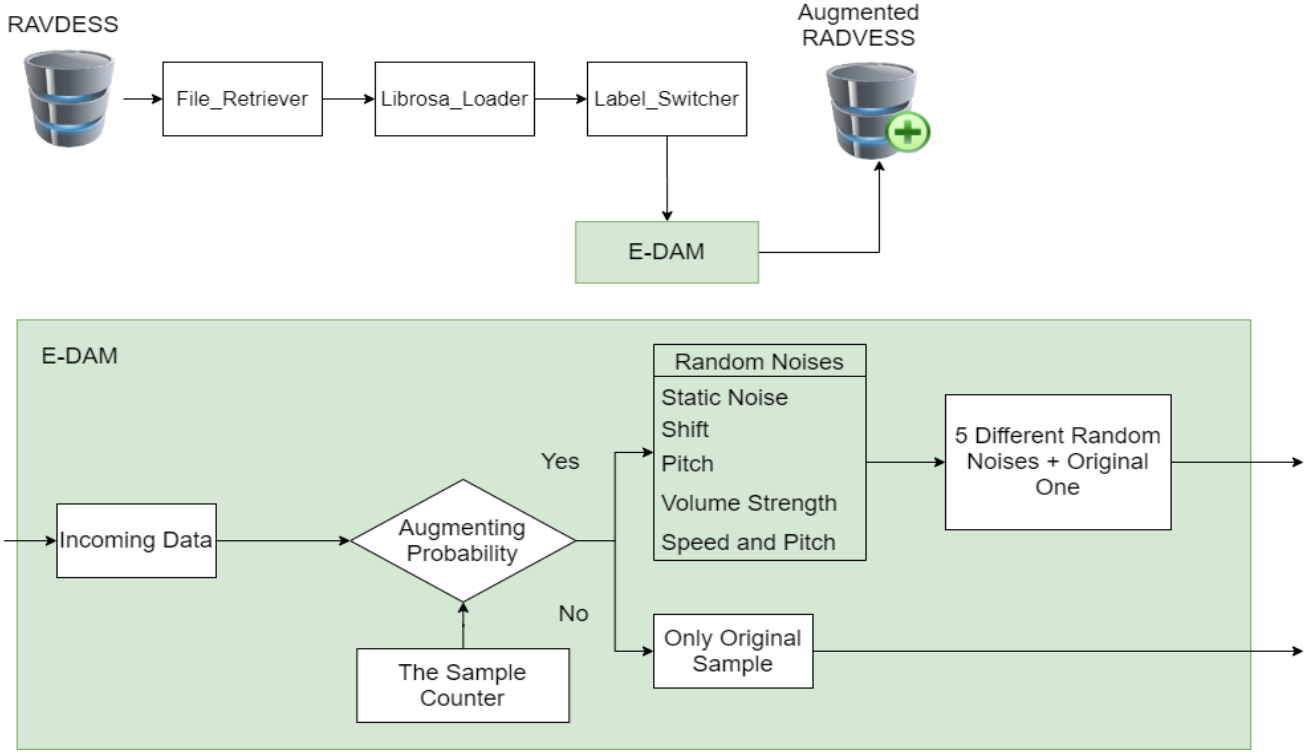
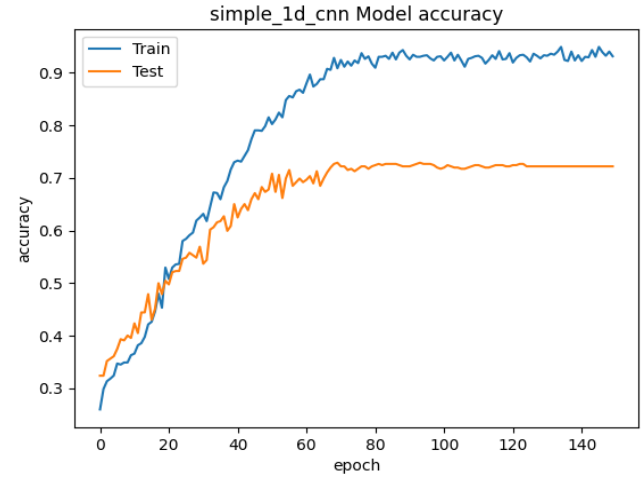


Fig. 1 Diagram of the proposed E-DAM.

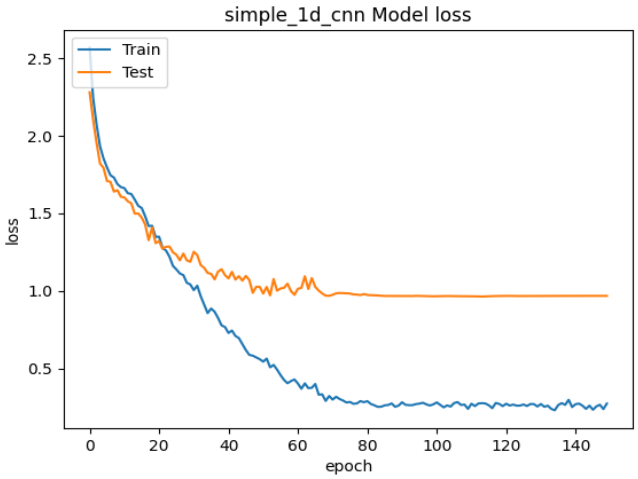
stored in the comma separable values (CSV) file.

In this paper, we propose an equilibratory data augmenting machine (E-DAM) to detect the driver's emotion in a noisy environment. E-DAM generates 5 different random noise additional audio samples per one audio sample from RADVESS. Besides, the proposed E-DAM can generate additional audio samples to balance the distribution of each class's samples.

II. The EDAM



(a)

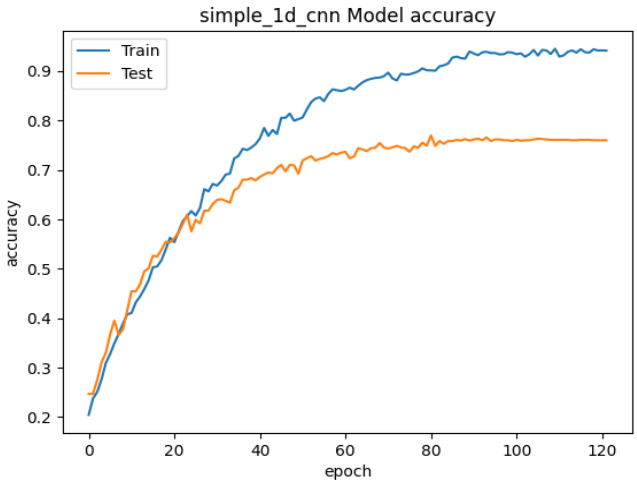


(b)

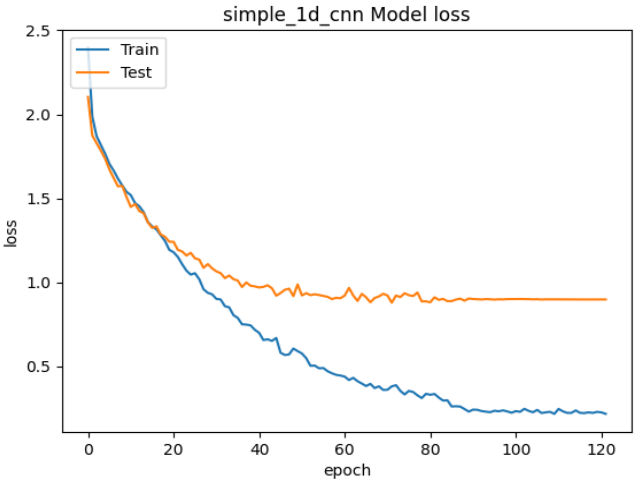
Fig. 2 The performance of the simple 1-dimentional CNN's architecture with RADVESS's training. (a) The validating accuracy. (b) The validating loss.

Fig 1. demonstrates E-DAM's structure. E-DAM consists augmenting probability, the sample counter, and random noises after all audio samples are loaded via the file retriever, Librosa loader, and label switcher.

Before entering E-DAM, the file retriever locates the audio file in the RADVESS. Librosa loader [6], which loads the file, contains tools for pre-processing raw audio signals and converts from time-sequential signals into discrete signal values. The label switcher switches the order of labels matching our previous FER's system. Before the discrete values convert into a Mel spectrogram, E-DAM generates 5 different random noises from the



(a)



(b)

Fig. 3 The performance of the simple 1-dimentional CNN's architecture with E-DAM and RADVESS's training. (a) The validating accuracy. (b) The validating

original discrete values and increases the number of audio samples per each classification. The sample counter handles increasing the samples in each classification in an equilibratory fashion. E-DAM contains 5 different random noises such as white noise, shift, speed, and strength over the original sound. After including the random noises, Librosa [6] can generate the Mel spectrogram, then converted them into decibel values. After pre-processing from the Librosa module, the dimensional channel size is reduced by calculating the decibel average from different channels. Calculating the decibel average of the different channels helps to smooth the model training. Finally, every 3 seconds of the audio file becomes 128 × 1 sized decibels, stored in the CSV file.

III. Test Results

The conditions for our experiment are 10^{-3} of the initial learning, 10^{-7} of Adam's epsilon, 10^{-2} of L2-regularizer, 150 epochs, and 50 of the patients. The software that we use is Python V. 3.8 and TensorFlow-GPU 2.2. Our computer hardware is Intel R CoreTM i5 10600k CPU @ 4.10 GHz, 32GB RAM, and GeForce RTX 2070.

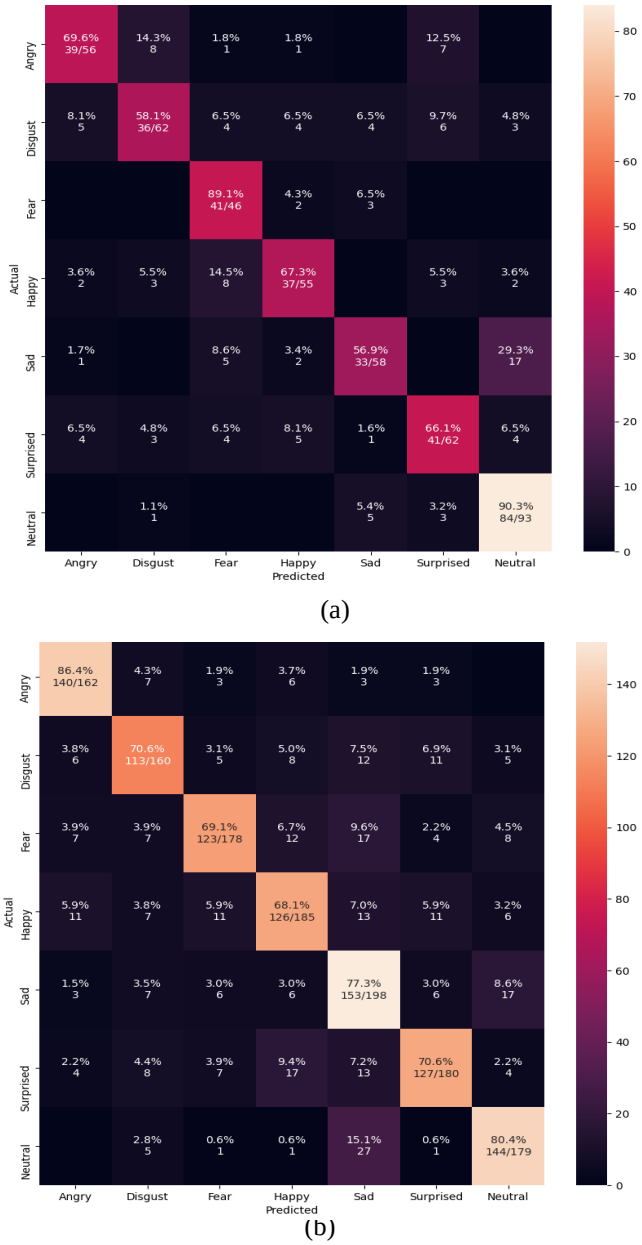


Fig. 4 The confusion matrices of the simple 1-dimentional CNN's architecture with RADVESS's training. (a) Without E-DAM. (b) With E-DAM.

As shown in Figs. 2 and 3, the simple convolution neural network (CNN) [7] for a 1-dimensional dataset reached 71% of the validating accuracy and 0.9639 of the validating loss on the original RADVESS training. From Fig. 3, E-DAM adds the random noises and increases the number of samples on the original RADVESS dataset to alleviate the overfitting problem as we previously applied on the FER dataset [1]. Applying E-DAM achieves 74% of the validating accuracy and 0.8793 of validating loss. Fig. 4 displays how the pre-trained 1-D CNN's architecture more precisely classifies speech emotion samples after E-DAM's approach. From Table I, the precision, recall, and F1-score summarize the experimental results from Fig 4. After applying E-DAM on the RADVESS dataset, the overall performance of the SER's system increased by about 3%.

IV. Conclusion

In this paper, we proposed EDAM to have improved the detection of the driver's emotions. The proposed E-DAM would solve performance problems in many disruptive environments. Increasing the number of the original SER dataset and adding the random noise deliberately for training the SER's model improves the SER system reliability. A reliable SER system precisely detects a driver's emotion even in a disruptive environment and saves more lives from preventable road rage-related accidents.

Table I. The confusion matrix summarizes the performance graph from Fig. 4.

Approaches	Precision	Recall	F1 Score
Without E-DAM	72.088 %	71.9907 %	71.4754 %
With E-DAM	**74.9147 %**	**74.5571 %**	**74.5485 %**

ACKNOWLEDGMENT

This research was supported by MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2020-0-01808) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

References

[1] J. H. Kim, A. Poulose, and D. S. Han, "The Extensive Usage of the Facial Image Threshing Machine for Facial Emotion Recognition Performance," Sensors, vol. 21, p. 2026, 2021.
[2] J. H. Kim and D. S. Han, "Data Augmentation & Merging Dataset for Facial Emotion Recognition," in Proceedings of the Symposium of the 1st Korea Artificial Intelligence Conference, Jeju, Korea, 2020, pp. 12-16.
[3] J. H. Kim, A. Poulose, and D. S. Han, "Facial Image Threshing Machine for Collecting Facial Emotion Recognition Dataset," The Journal of Korean Institute of Communications and Information Sciences, pp. 67-68, 2020.
[4] J. H. Kim and D. S. Han, "Data Augmentation & Merging Dataset for Facial Emotion Recognition," in Proceedings of the Symposium of the 1st Korea Artificial Intelligence Conference, Jeju, Korea, 2020, pp. 12-16.
[5] 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, p. e0196391, 2018.
[6] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, et al., "Librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, 2015, pp. 18-25.
[7] S. Malek, F. Melgani, and Y. Bazi, "One-dimensional convolutional neural networks for spectroscopic signal regression," Journal of Chemometrics, vol. 32, p. e2977, 2018.