2D-CNN 기반 우울증 감지를 위한 음성데이터 전처리

박준희*, 문남미**
*호서대학교 컴퓨터공학과
**호서대학교 컴퓨터공학부
cach456@gmail.com, mnm@hoseo.edu

Speech data preprocessing for detection of depression based on 2D-CNN

JunHee Park*, NamMee Moon**
*Dept. of Computer Engineering, Hoseo University
**Dept. of Computer Science, Hoseo University

요 약

세계보건기구(WHO)에 따르면 전 세계적으로 우울증 장애를 앓고 있는 사람이 3 억 2,200 만명에 달하며, 매년마다 빠르게 늘어나는 환자로 인해 전세계적으로 문제가 되고 있다. 이에 따라 우울증을 감지하기 위한 시스템에 대한 연구가 진행되어지고 있다. 본 논문에서는 우울증 감지에 있어 높은 정확도를 얻을 수 있는 최적의 음성 세그먼트 길이와 멜 밴드의 수를 확인하고자 한다. DAIC-WOZ(Distress Analysis Interview Corpus Wizard of Oz) 데이터셋을 기반으로 2D-CNN(2Dimension - Convolutional Neural Network)를 사용하여 음성 세그먼트 길이와 멜 밴드의 수에 변화를 주며 테스트를 진행하였다. 최종적으로 12 초 길이의 음성 세그먼트와 512 개의 멜 밴드에서 86.3%의 정확도로 최적의 결과를 확인하였다.

1. 서론

세계보건기구(WHO)에 따르면 우울증 장애를 앓고 있는 사람이 3 억 2,200 만명에 달하며, 매년마다 빠르게 환자의 수가 늘어나고 있다[1]. 우울증 장애는 발병률에 비해 진단이 어려우며, 제대로 된 치료가 이루어지지 않는 경우에는 자살로 이어질 수 있다.

우울증 장애는 환자의 발화 상태나 문장에 영향을 주기 때문에 음성 데이터, 텍스트 등 의사소통에서 특징을 동반한다. 이러한 특징들을 통해 우울증 장애를 감지하기 위한 연구가 지속적으로 이루어지고 있다. 이 중 음성 데이터에서의 우울증 감지는 전처리 방법에 따른 정확도의 변화로 인해 방법론에 대한 연구가 진행되어지고 있다[2,3].

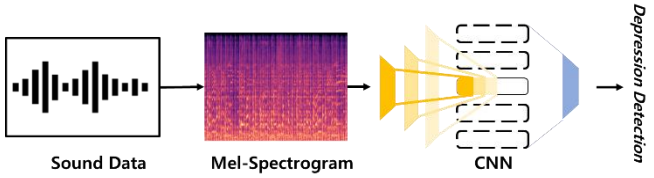
따라서 본 논문에서는 2D-CNN 모델을 사용하여 우울증 감지에서의 최적의 세그먼트 길이와 멜 밴드 수를 확인하고자 한다. 제안된 방법은 DAIC-WOZ 데이터셋[2]으로부터 Mel-Spectrogram 을 추출하는 과정에서 세그먼트 길이와 멜 밴드 수의 변화를 주어 전처리를 진행한다. 그리고 2D-CNN 모델을 통한 분류 학습에서의 정확도를 확인하여 우울증 감지에서의 최적의 세그먼트 길이 및 멜 밴드를 도출한다.

2. 관련연구

음성 데이터를 기반으로 하는 분류에서는 음성 세그먼트의 길이와 멜 밴드의 수에 대한 연구가 지속적으로 이루어지고 있다. Wang 은 오디오 이벤트 분류에서 CNN 모델의 성능에 영향을 미치는 요소에 대한 연구를 진행하였으며, 짧은 시간 동안 발생하는 이벤트 오디오에 대한 과적합을 방지하기 위한 신호 분할 방식을 도출하였다[3]. Alghifar 는 MATLAB R2018B 모델을 사용하여 MFCC 데이터로부터 1~20 초의 음성 세그먼트 특징 추출 시간과 인식률 사이의 균형에서 7 초의 세그먼트 길이가 적합함을 확인하였다[4].

3. 실험 개요

(그림 1) 은 본 논문에서 우울증 감지를 위한 음성 세그먼트 길이와 멜 밴드 수를 확인하기 위한 실험 개요이다. 제안된 시스템은 음성 파일에서 Mel-Spectrogram 을 도출하고, 2D-CNN 모델을 학습을 진행하여 정확도를 확인한다. Mel-Spectrogram 을 도출하는 과정에서 세그먼트 길이와 멜 밴드의 수에 변화를 주어 각각의 정확도를 비교한다.



(그림 1) 실험 개요도

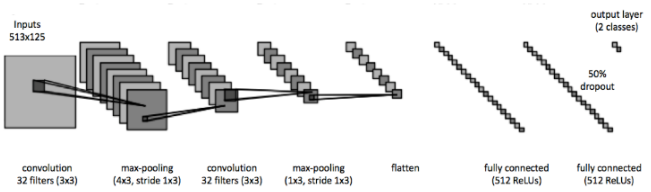
4. 실험

실험을 위해 데이터셋은 DAIC-WOZ 를 사용하며, 전처리 도구로 Librosa 라이브러리를 사용한다. DAIC-WOZ 데이터셋은 AVEC 2017[5]의 우울증 감지 작업을 위해 분할된 데이터셋이며, 모든 참가자에게 동일한 질문 세트를 묻는 가상인간 에이전트가 참가자를 인터뷰한 시나리오에서 수집된 데이터이다. Librosa 라이브러리를 통해 Mel-Spectrogram 을 추출하는 과정에서 멜 밴드(n_mels)를 얼마나 적용할 것인지에 대한 하이퍼파라미터는 512,256,128 로 설정한다.

추출된 Mel-Spectrogram 은 우울증을 감지하기 위한 대상에 대한 전체 음성 데이터이다. 따라서 Mel-Spectrogram 에서 특정한 길이의 음성 세그먼트 특징을 구해야 되며, 세그먼트 길이를 구하는 수식은 (식 1)을 사용한다.

$$\textbf{Width} = \frac{sec}{HanningWidnow/sr} \qquad \text{(식 1)}$$

이를 통해 4sec = 250width, 8sec = 500 width, 12sec = 750 width 의 음성 세그먼트 길이를 도출하였다.



(그림 2) 실험 모델(2D-CNN)

(그림 2)는 전처리된 데이터를 기반으로 학습을 진행하기 위한 2D-CNN 모델이다. 3x3 필터를 가지는 2 개의 Convolution 레이어와 2 개의 max-pooling 레이어, 2 개의 Fully connected 레이어 그리고 2 개의 Batch_Normalization 으로 이루어져 있다.

<표 1>은 세그먼트 길이와 멜 밴드의 수에 차이를 두어 (그림 2)의 모델에 학습시킨 결과이다. 학습결과와 같이 음성 세그먼트의 길이는 12sec(750 width), 멜 밴드는 512 일 때 가장 높은 acc, Test_acc 값을 보였다.

<표 1> 세그먼트 길이와 멜 밴드 수에 따른 학습결과

Sampling Rate(Hz)	Width	n_mels	acc	Val_acc	Test_acc
16000	**750**	**512**	**0.9482517**	**0.5523012**	**0.8632479**
16000	750	256	0.7776224	0.58158994	0.74786323
16000	750	128	0.7874126	0.50627613	0.6752137
16000	500	512	0.7367475	0.56873316	0.5494506
16000	500	256	0.91015273	0.606469	0.7967033
16000	500	128	0.71967655	0.51212937	0.55494505
16000	250	512	0.9204852	0.64016175	0.5618132
16000	250	256	0.772684	0.5903	0.66483516
16000	250	128	0.6257852	0.53986525	0.5631868

5. 결론

본 논문은 2D-CNN 모델에서의 학습과 그에 대한 결과로 우울증 감지를 위한 최적의 음성 세그먼트 길이와 멜 밴드가 (12sec,512)인 것을 확인하였다. 이를 통해 우울증 감지에서의 높은 정확도를 기대할 수 있다.

향후 연구에서는 제안된 방법에서의 2D-CNN 모델 이외에 ResNet50 과 같은 이미지넷에서의 검증을 통해 최적의 세그먼트 길이 및 멜 밴드 수를 확인할 것이다.

본 연구는 2021 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음(20190018340031001)

참고문헌

[1] World Health Organization. “Depression and other common mental disorders: global health estimates”, World Health Organization, No.WHO/MSD/MER/2017.2, 2017

[2] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., ... & Morency, L. P. “The distress analysis interview corpus of human and computer interviews.”, In Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC'14, pp. 3123-3128, 2014.

[3] Wang, H., Chong, D., Huang, D., & Zou, Y. ”What Affects the Performance of Convolutional Neural Networks for Audio Event Classification.”, In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 140-146. IEEE, 2019.

[4] Alghifari, M. F., Gunawan, T. S., Nordin, M. A. W., Kartiwi, M., & Borhan, L., ”On the Optimum Speech Segment Length for Depression Detection.”, In 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application, ICSIMA, pp. 1-5. IEEE.

[5] Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., ... & Pantic, M. “Avec 2017: Real-life depression, and affect recognition workshop and challenge.”, In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge .pp. 3-9. 2017.