Journal of Convergence for Information Technology
Vol. 11. No. 6, pp. 33-39, 2021
e-ISSN 2586-4440
DOI : https://doi.org/10.22156/CS4SMB.2021.11.06.033

CNN(Convolutional Neural Network) 알고리즘을 활용한 음성신호 중 비음성 구간 탐지 모델 연구

이후영
이르테크 기업부설연구소

A Study on a Non-Voice Section Detection Model among Speech Signals using CNN Algorithm

Hoo-Young Lee
IIR-TECH AI Lab.

요 약 음성인식 기술은 딥러닝과 결합되며 빠른 속도로 발전하고 있다. 특히 음성인식 서비스가 인공지능 스피커, 차량용 음성인식, 스마트폰 등의 각종 기기와 연결되며 음성인식 기술이 산업의 특정 분야가 아닌 다양한 곳에 활용되고 있다. 이러한 상황에서 해당 기술에 대한 높은 기대 수준을 맞추기 위한 연구 역시 활발히 진행되고 있다. 그중에서 자연어처리(NLP, Natural Language Processing)분야에서 음성인식 인식률에 많은 영향을 주는 주변의 소음이나 불필요한 음성신호를 제거하는 분야에 연구가 필요한 상황이다. 이미 많은 국내외 기업에서 이러한 연구를 위해 최신의 인공지능 기술을 활용하고 있다. 그중에서 합성곱신경망 알고리즘(CNN)을 활용한 연구가 활발하게 진행되고 있다. 본 연구의 목적은 합성곱 신경망을 통해서 사용자의 발화구간에서 비음성 구간을 판별하는 것으로 5명의 발화자의 음성파일(wav)을 수집하여 학습용 데이터를 생성하고 이를 합성곱신경망을 활용하여 음성 구간과 비음성 구간을 판별하는 분류 모델을 생성하였다. 이후 생성된 모델을 통해 비음성 구간을 탐지하는 실험을 진행한 결과 94%의 정확도를 얻었다.

주제어 : 음성인식, 딥러닝, 합성곱신경망, 인공지능, NLP

Abstract Speech recognition technology is being combined with deep learning and is developing at a rapid pace. In particular, voice recognition services are connected to various devices such as artificial intelligence speakers, vehicle voice recognition, and smartphones, and voice recognition technology is being used in various places, not in specific areas of the industry. In this situation, research to meet high expectations for the technology is also being actively conducted. Among them, in the field of natural language processing (NLP), there is a need for research in the field of removing ambient noise or unnecessary voice signals that have a great influence on the speech recognition recognition rate. Many domestic and foreign companies are already using the latest AI technology for such research. Among them, research using a convolutional neural network algorithm (CNN) is being actively conducted. The purpose of this study is to determine the non-voice section from the user's speech section through the convolutional neural network. It collects the voice files (wav) of 5 speakers to generate learning data, and utilizes the convolutional neural network to determine the speech section and the non-voice section. A classification model for discriminating speech sections was created. Afterwards, an experiment was conducted to detect the non-speech section through the generated model, and as a result, an accuracy of 94% was obtained.

Key Words : Speech Recognition, Deep-Learning, CNN, Artificial-Intelligence, NLP

*Corresponding Author : Hoo-Young Lee(hooyoung.paul.lee@cedartrees.co.kr)

Received March 29, 2021
Revised June 6, 2021
Accepted June 20, 2021
Published June 28, 2021

1. 서론

1.1 연구의 동기

음성 분석 기법은 딥러닝(Deep Learning)기술을 접목하여 과거에 비해 큰 기술 향상을 이뤄냈다. 최근 스마트폰, 인공지능 스피커, 차량 내 음성인식 등 음성을 활용한 다양한 서비스가 이뤄지고 있다. 음성 분석 기술은 과거에 비해 큰 성과를 내고 있지만, 주변의 소음 여부 등에 따라서 인식률의 감소가 불가피한 상황이며 해당 분야에 지속적인 연구와 발전이 필요한 상황이다[1].

현재 해외의 많은 기업에서 음성인식 정확도를 향상하고자 하는 많은 연구가 있고 특히 인공지능 기술을 활용한 연구가 최근 활발하게 진행되고 있다. 그중 IBM 트루노스, 구글 웨이브넷, 아마존 알렉사 등이 대표적이다[2-4].

국내에서는 네이버, 카카오 등에서 인공지능 기술을 활용한 음성인식 기반의 AI 스피커를 출시하였고 음악, 스트리밍, 뉴스 등의 다양한 컨텐츠와 합하여 그 활용도가 높아지고 있다. AI 스피커는 이용자의 대부분이 일반 대중으로 생활 전반의 편의 기능에 집중되어 있다[5]. 이러한 상황에서 이용자의 음성인식 기술 성능개선에 대한 요구는 날로 높아지고 있다. 이미 인공지능을 음성인식에 활용하는 다양한 연구들이 진행되어 왔고 또 일부 좋은 성과를 거두었다. 특히 사용자의 음성 정보를 통해 발화자의 나이, 성별 등을 탐지하여 분류하는 연구는 서비스 제공자 측면에서 큰 이점이 있다[6,7]. 또 딥러닝 알고리즘 중 합성곱신경망(CNN)을 활용하여 음성을 분류하는 연구[8] 역시 활발히 진행되고 있다. 딥러닝은 인공신경망을 이용한 인공지능의 한 분야로 1960년대부터 패턴인식에 응용하려는 연구가 활발히 진행되어 왔으며 또한 오차역전파 학습 알고리즘의 등장으로 인하여 여러 분야에서 응용[9]되고 있다.

합성곱신경망을 통한 신호감지(Signal Detection) 연구는 최근의 연구를 통해서 기존 방법 및 기타 딥러닝 알고리즘 보다 더 좋은 성능을 발휘한다 연구발표[10]가 있고 이러한 기술을 활용해서 광대역 통신망에서 모스 신호를 감지하는 경우 기존의 방법 보다 성능이 우수함[11]을 보여주는 연구결과가 있었다.

1.2 연구 목적 및 방법

본 연구의 목적은 합성곱 신경망을 통해서 사용자 발화 중 음성구간과 비음성 구간을 탐지하고 그 정확도를 측정하고자 하는 연구이다. 이를 위해서 일반 사무실과 가정에서 5명의 여성 발화자가 합 300개의 문장을 발화한 음성파일(wav)을 취합하여 각 음성을 1초 단위로 분할한 후 해당 음성 파일(wav)을 학습 및 정확성 검증 데이터로 사용하였다. 해당 학습은 지도학습의 한 분야로 분할 저장된 각각의 음성 파일이 유효한 발화 구간인지 비음성 구간인지를 레이블링(Labeling)하여 데이터셋을 최종 생성한다.

학습용 데이터를 딥러닝 알고리즘 중에 하나인 합성곱-신경망 알고리즘(CNN, Convolutional Neural Network Algorithm)을 사용하여 해당 파일의 음성/비음성 여부를 판별하는 예측 모델을 생성하여 모델의 정확도를 판단하였고 이후 해당 모델을 통해서 실제 음성 파일에서 비음성 구간을 제거하는 실험을 진행하였다.

2. 관련연구

2.1. 합성곱 신경망

합성곱 신경망은 합성곱 연산을 적용하여 신경망을 연산하는 것으로 이미지 및 자연어 해석, 영상분류 등 다양한 분야에 활용되고 있다[12]. 기존의 신경망 기법은 연산의 결과를 다음 층으로 직접 연결하는 Fully-Connectetd 구조로 모델을 구성한다. 합성신경망 역시 다차원 출력 데이터를 Fully-Connected로 연결하고 마지막 층에 선형층을 통과 하게한 후에 나온 결과를 Softmax 같은 활성함수를 사용한다.

합성곱신경망은 이미지의 일부분에서 특징을 추출하고 이러한 특징을 통하여 예측하는 것으로 인간이 생성한 특징을 통해서 예측하는 것보다 높은 정확도를 가진다[13,14].

Fig.1과 같이 합성곱 신경망의 일반적인 구조는 2개의 부분으로 구성[15]된 것을 확인 할 수 있다.

첫 번째는 Convolution Filter를 통해서 M×N 크기의 매트릭스가 설정한 채널의 수만큼 생성되는 부분이다. 이러한 과정을 수행하며 해당 매트릭스의 특징값을 추출하게 된다.

두 번째는 이렇게 생성된 필터들을 모두 연결하는 Fully Connected 부분이다.

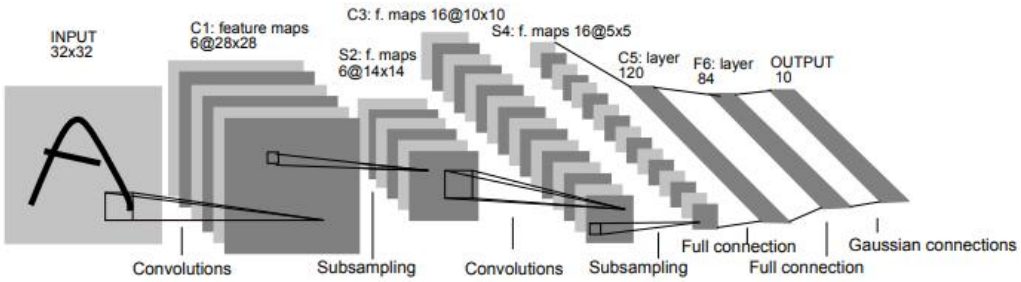


Fig. 1. Structure of a typical Convolutional Neural Network

2.2. WAV(Waveform Audio File Format)

WAV는 파형 오디오 형식(Waveform audio format)의 준말로 개인용 컴퓨터에서 오디오를 재생하는 마이크로소프트와 IBM 오디오 파일 포맷 표준이다. Chunk 데이터를 저장하기 위한 RIFF(Resource Interchange File Format) 비트 스트림 포맷 방식에서 변환한 것으로 가공되지 않은 오디오를 위한 윈도우 시스템에 쓰이는 기본 포맷[16-18]이다.

WAV 파일에는 압축된 오디오가 포함될 수 있지만 가장 일반적인 WAV 오디오 형식은 선형 펄스 코드 변조(LPCM, Linear Pulse Code Modulation) 형식의 비압축 오디오이다. LPCM은 44,100Hz로 샘플링된 2 채널 LPCM 오디오를 샘플 당 16 비트로 저장하는 오디오 CD의 표준 코딩형식이다[19].

3. 제안 모델

3.1 연구 프로세스

음성신호에는 음성 구간 함께 비음성 구간이 존재한다. 비음성 구간은 음성발화 중에 나오는 휴지기(休止期)이며 이 구간은 보통 음성의 발화의 구간과는 신호의 차이가 있다. 본 연구는 음성 신호에 있는 비음성 구간을 딥러닝 알고리즘 중에 하나인 합성곱-신경망 알고리즘(Convolutional Neural Network Algorithm)을 통해서 판별하여 비음성 구간을 추출하는 방법에 대한 것으로 연구의 흐름은 아래의 Fig.2와 같다.

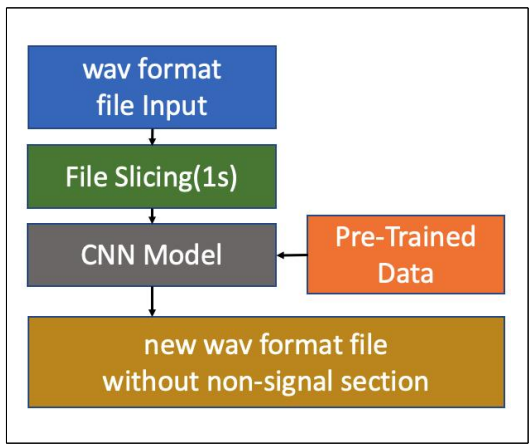


Fig. 2. System Configuration

발화자의 음성은 WAV(Waveform Audio File Format) 신호로 입력된다. 입력된 음성신호는 주파수(Hz)의 기본단위인 1초(Second) 단위로 음성을 분리(Slice)하여 임시 저장소에 음성 파일을 저장한다.

예를 들어 1분의 음성이라면 60개의 초단위 파일로 분리되어 생성된다. 분리된 파일은 정수(float) 형태의 1차원 리스트 형태로 메모리에 적재한다. 메모리에 적재되는 사이즈는 CNN을 활용한 데이터 분석을 위해 고정된 사이즈로 변경해야 한다. 이를 위해 부족한 데이터는 Padding 값을 사용하여 리스트의 오른쪽에 데이터를 추가한다.

학습을 위한 음성 데이터 신호를 생성한 후 CNN 모듈을 생성하고 음성/비음성 구간을 추출하는 모델을 학습한다. 학습을 위한 인공지능 프레임워크로 PyTorch를 사용한다. PyTorch는 오픈소스 소프트웨어로 Tensorflow와 함께 널리 사용되는 Python 기반의 머신러닝 프레임워크이다.

학습을 완료한 모델을 통해 입력된 발화음성을 각

Slice 단위로 해당 신호가 음성구간인지 비음성구간인지를 체크하고 비음성 구간일 경우에는 해당 부분을 제거한후 음성 부분만 합하여 새로운 파일로 생성한다.

3.2 음성/비음성 신호의 특징

학습에 활용할 음성신호의 전처리를 위해서 전체 음성파일을 1초 단위로 분할하고 음성의 초당 샘플링 수(Sample Rate)를 22,050으로 설정하여 해당 음성파일로부터 벡터정보를 추출하였다.

이렇게 생성한 벡터 데이터는 32비트 부동소수점 형태로 표현된다. 즉, 1분의 음성데이터는 전처리를 통해서 [60×22050] 형태의 데이터로 변환 할 수 있다. 변환된 데이터가 음성의 정보를 잘나내고 있는지 확인하기 위해서 음성구간과 비음성 구간을 샘플링하여 파이썬의 시각화 도구인 matplotlib를 활용하여 나타낼 수 있다.

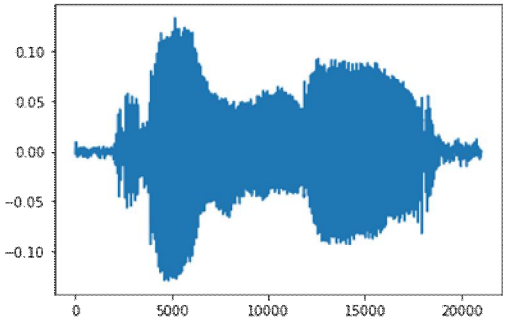


Fig. 3. WAV Signal - Speech Signal

Fig.3는 음성 구간 데이터를 시각화 한 것이다. 해당 파일을 보면 발화자 음성의 높낮이에 따라서 파형이 차이가 있는 것을 확인 할 수 있다. 반면 Fig.4은 녹음 중 발생하는 지속적인 비음성(Noise) 구간으로 소리의 파형이 규칙적이고 지속적임을 확인 할 수 있었다.

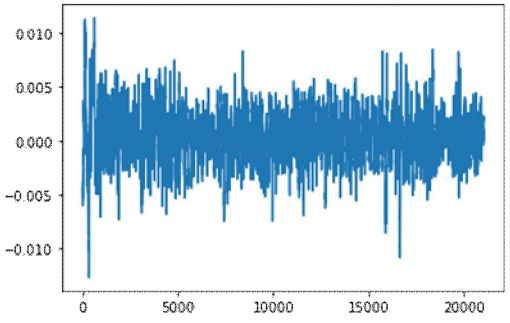


Fig. 4. WAV Signal - Non-Speech Signal

본 연구의 실험 모델 생성을 위해 학습 데이터로 발화자의 음성 샘플을 수집하여 1초 단위로 샘플링하고 샘플링한 데이터를 각각 음성구간과 비음성 구간의 데이터로 분류하여 Label을 생성하였다.

Table 1. Train Dataset

Data	Count	Format
Speech Signal	3,837	wav
Non-Speech Signal	2,915	wav

Table.1은 수집한 음성 데이터의 수집 현황 정보이다. 음성 3,837개, 비음성 2,915개 합 6,752개의 음성 학습용 데이터를 수집하였다.

3.3 신경망 모델 구성

본 연구를 위해서 사용한 신경망은 딥러닝 알고리즘 중 CNN(Convolutional Neural Network) 알고리즘이다. 해당 모델을 구성하기 위해서 앞서 언급했듯이 파이썬 기반의 오픈소스 머신러닝 프레임워크인 파이토치(PyTorch 1.8.0 GPU, CUDA 11.2)를 사용하였다.

다음의 Fig.5는 본 연구모델의 CNN 구조이다. 신경망의 구조는 Conv2d, ReLU, BatchNorm2d를 이어 연결하는 구조를 사용하였다. Conv2d는 합성곱신경망을 구축하는 PyTorch에서 제공하는 함수로 입력채널 정보와 출력채널 정보를 입력하는 것으로 신경망 구조를 만들어 낼 수 있다. ReLU(Rectified Linear Unit) 함수는 학습 중 깊고 넓은(Deep-Wide) 신경망을 구성할 때에 가중치 값이 극히 작은 값으로 셋팅되어 해당 입력 값이 출력에 미치는 영향이 미약해지는 Gradient Vanishing 문제를 해결하기 위해서 사용하는 활성함수로 입력이 0을 넘으면 그 값을 그대로 출력하고 0 이하일 경우에는 0을 출력하는 함수이다.

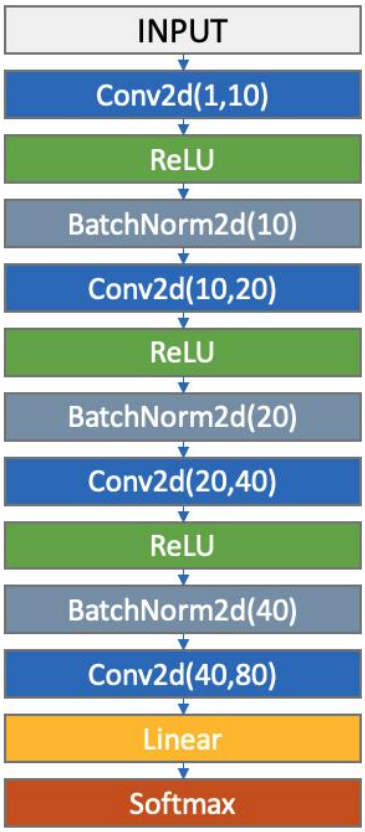
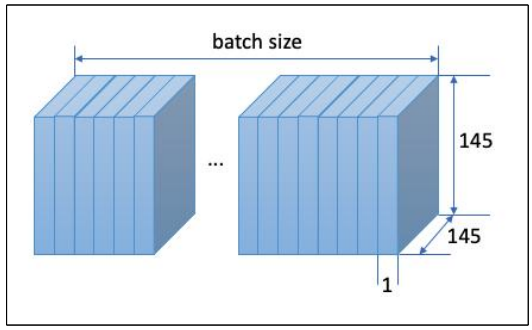


Fig. 5. CNN Moudle

또 정규화를 통해서 학습을 더 빠르게 할 수 있도록 BatchNorm을 사용하였다. Conv2d를 통해 나온 결과는 FC Linear 레이어를 통과하고 최종 결과는 Softmax를 통해 0 음성정보, 1 비음성정보 즉 [0,1] 형태로 출력된다.

CNN에 입력되는 데이터는 4차원 벡터의 형태를 가지기 때문에 1차원 벡터 정보를 4차원 형태로 변경해한다. 본 연구에서는 모듈에 입력되는 1차원 벡터인 22,050를 [n_batch,1,145,145] 형태의 4차원 벡터로 변형(Reshape)하여 사용한다. 이때 n_batch는 입력데이터의 총 개수를 나타낸다. 예를 들어 60초 길이의 음성파일을 1초로 나누게 되면 이 값은 60이 된다. 그 다음 값이 1은 channel 정보로 고정된 값을 가진다.

Fig. 6는 이러한 데이터구조를 나타낸 것이다.



Flg. 6. CNN Train Data Shape

그 다음은 음성의 Sampling Rate인 22,050의 벡터를 CNN 분석을 위한 데이터 형태로 [145×145] 형태로 변경한다. 이렇게 최종 과정을 거친 데이터의 형태는 [batch_size,1,145,145] 형태가 된다.

4. 실험 및 고찰

학습 데이터는 음성 데이터 3,837개와 비음성 데이터 2,915개를 합한 6,752개 데이터로 구성되어 있다. 레이블이 “1”인 경우는 비음성 데이터이고 “0”인 경우는 음성데이터이다. 해당 음성/비음성 데이터를 Shuffle하여 하나의 데이터셋으로 만들고 이 두 데이터셋을 Train/Test으로 분리한다. 해당 데이터셋에 대한 정보는 Table 2의 내용과 같다.

Table 2. Train/Test Dataset

Rate	Train	Test	Total
8:2	5,401	1,351	6,752

Train 데이터는 모델을 학습하는데 활용하고 학습이 완료된 모델을 테스트하기 위해서 Test 데이터를 사용하였다. 최종 모델의 정확도는 훈련 데이터가 제외된 테스트용 데이터를 통해서만 학습률을 계산한다.

Train 데이터는 전체 데이터를 한번에 학습할 경우 전체 파일을 메모리에 올려야 하기 때문에 시스템 성능의 문제가 발생할 수 있어 1,000개 단위로 미니 배치(Mini-Batch)를 생성하고 1,500회 Epoch을 수행하면서 학습을 진행한다. Learning Rate는 0.001로 셋팅하여 학습을 진행하였다.

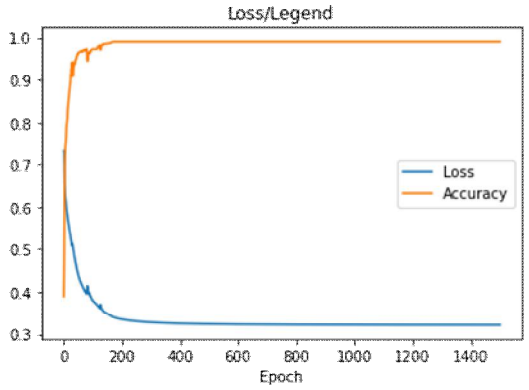


Fig. 7. Loss/Accuracy

최종 1,500회 학습을 수행하며 변화되는 학습곡선 그래프는 Fig.7과 같다. 학습이 진행되면서 Loss가 줄어들고 이에 반하여 Accuracy가 증가하였다. 학습 곡선 중 Epoch 1,000-1,500 구간은 Loss와 Accuracy의 변화가 미미함을 확인하고 최종 학습은 Epoch 1,500회에서 중지하였다. 학습을 완료한 후 학습 모델을 통해서 Test 데이터를 통해 예측을 수행한 결과 94%의 정확도를 보였다. 예측 모델의 정확도를 확인한 후 해당 모델을 활용해서 사용자가 녹음한 임의의 음성 파일을 입력받아 비음성 구간을 제거한 새로운 파일을 생성하여 파일의 음성 길이를 실험하였다. Fig. 8은 사용자가 녹음한 음성파일이며 Fig.9는 동일한 파일에서 해당 탐지 모델을 활용하여 비음성 구간을 제외하고 새롭게 생성한 파일의 파형정보를 출력한 것이다.

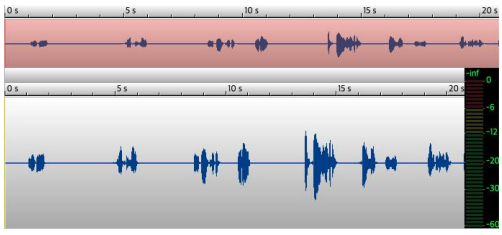


Fig. 8. sample wav file

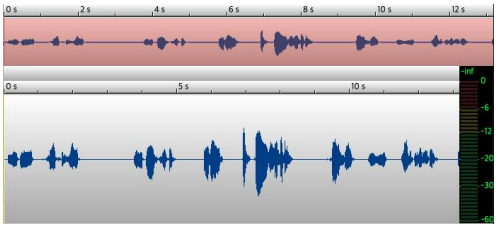


Fig. 9. generated wav file

두 파일의 비교를 통해서 Fig 8의 총 음성 구간이 21초였으나 비음성 구간을 제거한 후 Fig 9에서와 같이 재생 시간이 13초로 단축된 것을 알 수 있었다.

5. 결론 및 향후 연구 방향

본 연구의 특징은 인간의 음성신호가 아닌 다양한 비음성 신호 데이터를 인공신경망이 스스로 학습하여 특징 데이터를 찾아내고 이를 통해 발화자의 음성 구간 중에서 비음성 구간을 탐지하는 방안에 대한 연구이다.

이를 위해 녹음된 5명의 화자가 각각 녹음한 파일을 생성하였고 해당 파일에서 비음성 구간을 분리한 후 이를 딥러닝 알고리즘 중 하나인 합성곱 신경망(CNN)을 통해 비음성 신호의 패턴을 학습하고 실험을 통해 해당 모델이 비음성 구간을 탐지하는데 유효함을 확인하였다.

본 연구에 활용한 음성 데이터는 사무실, 가정에서 녹음한 데이터를 활용하여 학습 데이터를 생성한것으로 향후 이어지는 연구에서 식당, 자동차 실내, 카페, 기타 다중 이용시설 등에서 다양한 환경에서 생성한 파일을 학습에 활용하여 필요한 데이터 수집과 분석 모델 연구를 진행할 것이다.

REFERENCES

[1] D. S. Park. (2018). A Study on the Gender and Age Classification of Speech Data Using CNN. *Journal of KIIT, 16(11),* 11-21.
DOI : 10.14801/jkiit.2018.16.11.11

[2] Filipp Akopyan et all. (2015). TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 34(10), 1537-1557.
DOI : 10.1109/TCAD.2015.2474396

[3] A. Oord et all. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499,* 1-15.

[4] S. S. Jo & Y. G. Kim. (2017). AI (Artificial Intelligence) Voice Assistant Evolving to Platform. *IITP, p1-25,* Feb.

[5] B. S. Kim & H. J. Woo. (2019). A Study on the

Intention to Use AI Speakers: focusing on extended technology acceptance model. *The Korea Contents Association, 19(9),* 1-10.
DOI : 10.5392/JKCA.2019.19.09.001

[6] L. H. Meng & J. S. Han. (2017). The Impact of Relational Benefits on Positive Affect, Perceived Value, and Behavior Intention in Social Commerce : Focused on Chinese Tourist having the Hotel Service of Social Commerce environment. *Journal of tourism and leisure research, 29(10),* 69-88.

[7] J. H. Seo & Y. T. Kim. (2013). Effects of Service Convenience on Customer Satisfaction and Reuse Intention by Korail Talk App Users among Korail Passengers. *Journal of the Korean Society for Railway, 16(5),* 410- 417.

[8] H. Zhou et al. (2017). Using deep convolutional neural network to classify urban sounds. *In TENCON 2017-2017 IEEE Region 10 Conference* (pp. 3089-3092). IEEE.
DOI : 10.1109/TENCON.2017.8228392

[9] J. G. van Velden & G. F. Smoorenburg. (1991). Vowel recognition in noise for male, female and child voices. *In Acoustics, Speech, and Signal Processing, IEEE International Conference on* (pp. 989-992). IEEE Computer Society.
DOI : 10.1109/ICASSP.1991.150507

[10] X. Zha, H. Peng, X. Qin, G. Li & S. Yang. (2019). A deep learning framework for signal detection and modulation classification. *Sensors, 19(18),* 4042.
DOI : 10.3390/s19184042

[11] Y. E. Yuan. (2019). DeepMorse: A deep convolutional learning method for blind morse signal detection in wideband wireless spectrum. *IEEE Access, 7,* 80577-80587.
DOI : 10.1109/ACCESS.2019.2923084

[12] Y. LeCun, Y. Bengio & G. Hinton. (2015). Deep learning. *nature, 521(7553),* 436-444.
DOI : 10.1038/nature14539

[13] B. Theodore et all. (2013, August). Feature extraction with convolutional neural networks for handwritten word recognition. *In 2013 12th International Conference on Document Analysis and Recognition* (pp. 285-289). IEEE.
DOI : 10.1109/ICDAR.2013.64

[14] L. Xiaojun et al. (2017). Feature extraction and fusion using deep convolutional neural networks for face detection. *Mathematical Problems in Engineering.* 1-9.
DOI : 10.1155/2017/1376726

[15] Y. Lecun et all. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86(11),* 2278-2324.
DOI: 10.1109/5.726791

[16] Library of Congress. (2008). WAVE Audio File Format.

[17] Microsoft Corporation. (1998). *WAVE and AVI codec Registries-RFC 2361,* IETF.

[18] IBM & Microsoft. (1991). Multimedia Programming interface and Data Specifications 1.0

[19] R. Branson. (2015). *What Makes WAV Better than MP3,* Online Video Converter.

이 후 영(Hoo-Young Lee) [정회원]

· 2002년 2월 : 우송대학교 컴퓨터과학과(공학사)
· 2017년 2월 : 공주대학교 대학원 멀티미디어공학과(공학석사)
· 2017년 3월 ~ 2020년 2월 : 공주대학교 대학원 컴퓨터공학과 박사
· 2019년 9월 ~ 현재 : 이르테크 기업부설연구소장
· 관심분야 : 인공지능, 빅데이터, 자연어처리, 음성인식
· E-Mail : hooyoung.paul.lee@cedartrees.co.kr