Journal of KIIT. Vol. 20, No. 1, pp. 1-10, Jan. 31, 2022. pISSN 1598-8619, eISSN 2093-7571 1
http://dx.doi.org/10.14801/jkiit.2022.20.1.1

음성신호를 이용한 양방향 LSTM 기반 우울증 진단

조아현*, 곽근창**

Bi-Directional LSTM-based Diagnosis of Depression using Speech Signals

A-Hyeon Jo*, Keun-Chang Kwak**

본 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신산업진흥원의 지원을 받아 수행된 헬스케어 AI융합연구개발 사업임(NO.S1601-20-1041). 또한, 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NO.2017R1A6A1A03015496).

요 약

우울증은 개인의 전반적인 삶에 영향을 주고 감정, 생각, 신체 상태, 그리고 행동 등에 변화를 일으키는 심각한 정신질환이다. 따라서 우울증을 초기에 정확하게 진단하는 것이 중요하고 이를 위해 딥러닝을 이용한 여러 가지 시스템 개발되고 있다. 본 논문에서는 음성신호를 이용하여 Bi-LSTM(Bi-directional Long Short-Term Memory) 기반 우울증 진단을 위한 딥러닝 모델을 제안하고 성능을 비교하고 분석한다. 학습에는 딥러닝 모델 중 하나인 Bi-LSTM 모델을 사용했다. 음성신호의 특징은 MFCC(Mel-Frequency Cepstral Coefficient), GTCC (Gammatone Cepstral Coefficients)를 이용해 추출했고, 클래스 불균형을 해결하기 위해 데이터 확장을 수행했다. 데이터 확장 전 Bi-LSTM 모델의 성능은 약 56.58%였고, 확장 후의 성능은 약 98.49%로 학습데이터 확장을 통해 클래스 불균형 문제를 해결함으로써 모델의 분류 성능을 개선했다.

Abstract

Depression is a serious mental illness that affects an individual's overall life and causes changes in emotions, thoughts, physical condition, and behavior. Therefore, it is important to accurately diagnose depression in the early stages, and to this end, several systems are being developed using deep learning. In this paper, we propose a deep learning model for bi-directional LSTM-based diagnosis of depression using speech signals and compare and analyze performance. For training is used bi-directional LSTM model, one of the deep learning models. Features of speech signals were extracted using Mel-Frequency Cepstral Coefficient (MFCC) and Gammatone Cepstral Coefficient (GTCC), and data augmentation was performed to resolve imbalance of class. The performance of the Bi-LSTM model before data augmentation was about 56.58% and after data augmentation was about 98.49%, which improved the classification performance of the model by solving the imbalance of class through learning data augmentation.

Keywords
deep learning, depression diagnostic system, speech signal, bi-directional LSTM, data augmentation

* 조선대학교 전자공학과 IT-Bio 융합시스템 전공
- ORCID: http://orcid.org/0000-0003-1909-5655
** 조선대학교 전자공학과 교수(교신저자)
- ORCID: http://orcid.org/0000-0002-3821-0711

· Received: Sep. 07, 2021, Revised: Oct. 12, 2021, Accepted: Oct. 15, 2021
· Corresponding Author: Keun-Chan Kwak
Dept. of Electronics, Chosun University, 309, Pilmun-daero, Dong-gu, Gwangju, Korea
Tel.: +82-2-4703-1666, Email: kwak@chosun.ac.kr

Ⅰ. 서 론

코로나 팬데믹이 시작된 후 세계 각국에서 우울증 발생이 2배 이상 증가했고, 이는 전 세계의 수백만 명의 사람들에게 부정적인 영향을 미치고 있다. 우울증은 감기만큼이나 흔하게 발생할 수 있는 정신질환이다. 하지만 우울증의 치료를 계속해서 미루고 방치하게 된다면, 불안, 수면장애, 사회생활과 대인관계 기피, 심할 경우 자살 등의 문제를 초래할 수도 있다. 이러한 문제를 예방하기 위해 우울증을 조기에 진단하여 빠른 치료에 도움을 줄 수 있는 다양한 연구들이 진행되고 있다. 지금까지 보편적으로 사용되던 우울증 진단 방법은 의사의 주관적인 판단, 환자 스스로 증상을 설명하거나 자가 진단설문지를 이용하는 것이었다. 그러나 이 경우에는 주관적인 의견에만 의존하는 경향이 있어 우울증 진단의 객관성이 떨어뜨리는 문제가 있다[1]. 이러한 문제를 해결하기 위해 인공지능 기술을 이용하여 우울증 조기 진단에 도움을 줄 수 있는 시스템 개발에 대한 연구가 활발하게 진행되고 있다.

우울증을 진단하는 기준이 되는 신호들로는 영상, 음성, 텍스트, ECG, EEG와 같은 생체신호 등이 있다. 그중 음성신호는 우울증을 겪고 있는 사람은 평소 말했을 때와 말하는 방식, 억양 등이 달라진다는 사실 때문에 우울증 진단에 많이 사용되고 있는 신호이다. 이러한 음성신호를 기반으로 하여 우울증을 진단하는 연구는 보통 2가지 방법으로 진행되고 있다. 첫 번째로, 1차원 음성신호를 스펙트로그램/멜 스펙트로그램 특징을 이용해 시간-주파수 표현인 2차원 이미지로 변화하여 CNN 모델을 통해 분류하는 방법이 있다.

Lin[2]의 연구를 살펴보면 음성신호와 텍스트 신호를 앙상블 하여 우울증을 자동으로 진단하는 연구를 진행되었다. 여기서 음성신호의 특징은 멜 스케일 스펙트로그램을 이용하여 추출되었고 1D-CNN 모델을 통해 분류했을 때 꽤 높은 정확도를 보였다. Srimadhur[3]는 스펙트로그램을 기반으로 학습하는 CNN과 End-to-End CNN 모델을 제안했고, 이 중 End-to-End CNN 모델이 우울증 진단에 있어 더 높은 성능을 가지는 것을 증명했다. 두 번째로, 1차원 음성신호를 기반으로 하여 LSTM(Long Short-Term Memory)을 이용해 우울증은 진단하는 방법이 있다. Alhanai[4]의 연구에서 1차원 음성신호는 3개의 Bi-LSTM(Bi-directional LSTM) 계층이 있는 모델을, 텍스트는 2개의 Bi-LSTM 계층이 있는 모델을 통해 학습시키고 이를 앙상블하여 우울증 진단 시스템을 제안하였다. Rejaibi[5]의 연구에서는 음성신호의 특징을 추출하기 위해 MFCC(Mel Frequency Cepstral Coefficient)를 이용하였고, 추출된 시퀀스 특징을 이용하여 LSTM을 기반으로 우울증을 진단하는 연구를 진행했다.

본 논문에서는 1차원 음성신호를 이용하여 딥러닝 모델 중 하나인 RNN(Recurrent Neural Network) 모델의 일부인 Bi-LSTM을 기반으로 하여 우울증을 진단하는 시스템을 제안하고 분류 성능을 비교 분석한다. 사용된 음성 데이터베이스는 현재 우울증 진단에 많이 사용되고 있는 DAIC-WOZ 우울증 데이터베이스로 총 189명의 데이터를 포함한다. 본 데이터베이스는 가상 인터뷰 진행자 ‘Elie’와 환자가 대화하는 인터뷰 형식으로 녹음된 음성 데이터로 데이터 사용 전 가상 인터뷰 진행자 음성과 묵음을 제거하는 전처리 작업이 필요하다. 전처리가 완료된 1차원 음성신호를 MFCC(Mel Frequency Cepstral Coefficient)와 GTCC(Gammatone Cepstral Coefficients)를 사용하여 학습에 필요한 시퀀스 특징을 추출한다. 추출된 특징을 Bi-LSTM 모델을 2층으로 쌓아 설계한 알고리즘 모델을 통해 학습을 진행하고, 마지막으로 검증데이터를 사용해 알고리즘의 분류 성능을 확인하고, Confusion Matrix를 통해 예측성능과 실제 성능을 비교 및 분석한다.

본 논문은 다음과 같이 구성되어있다. 2장에서는 제안된 음성신호 기반 우울증 진단 딥러닝 모델의 학습에 사용된 DAIC-WOZ라는 우울증 데이터베이스를 소개한다. 또한 인터뷰 형식의 데이터에서 묵음과 가상 인터뷰어의 소리를 제거하는 데이터 전처리 방식에 관해 기술한다. 3장에서는 1차원 음성신호에서 특징을 추출하는 방법들을 제시하고, 사용된 딥러닝 모델 중 하나인 LSTM, Bi-LSTM에 대한 개념과 구조 등을 자세하게 기술한다. 4장에서는 제안된 우울증 진단 모델의 실험 진행 과정 및 결과, 성능을 비교 분석한다. 마지막으로 5장에서는 결론 및 향후 연구 계획에 관해 기술하고 마무리된다.

II. 데이터베이스 및 전처리 과정

2.1 DAIC-WOZ 우울증 데이터베이스

DAIC-WOZ 우울증 데이터베이스는 더 큰 데이터베이스인 DAIC(Distress Analysis Interview Corpus)의 일부이다. 이는 불안, 우울증, 외상 후 스트레스 장애 등 심리적인 상태를 진단하는 데 도움을 주기 위해 고안된 임상 인터뷰를 포함한다. 인터뷰는 그림 1과 같이 애니메이션 가상 인터뷰 진행자 'Elie'와 그를 통제라는 사람, 환자를 통해 진행되고 환자는 우울증 증상이 있는 사람과 증상이 없는 사람 모두를 포함한다[6][7]. 우울증의 여부를 판단하는 기준은 PHQ-8(Patient Health Questionnaire) 사전 설문조사의 점수를 기준으로 한다. PHQ-8은 8개의 질문으로 구성된 우울증의 척도로 대규모 임상 연구에서 우울증에 대한 유효한 진단 및 중증도를 기준을 확립된 설문조사 항목이다[8].

본 데이터베이스는 참가자에 대한 얼굴 영상, 음성, 텍스트 정보를 포함하고, 본 연구에서는 음성 파일만을 사용하여 실험이 진행되었다. 참가자는 총 189명이고 한 참가자당 7분에서 33분(평균 16분) 사이의 길이를 가지는 1개의 음성 파일을 포함한다. 모든 오디오 파일(웨이브 형식)은 16kHz로 녹음되었다. 참가자들은 인터뷰를 하기 전에 사전 설문지인 PHQ-8를 작성하였고, 작성한 PHQ-8의 점수가 10점 이상이면 우울증 증상이 있는 사람, 10점 미만이면 우울증이 없는 사람으로 구분하였다.

그림 1. 가상 인터뷰 진행자 '엘리'와의 인터뷰[7]
Fig. 1. Interview with the virtual interviewer 'Ellie'

표 1을 보면 남녀 참가자 수를 파악할 수 있다. 여기서 남성이 약 20명 정도 더 참여했지만, 우울증 증상을 가지고 있는 참가자는 여성 참가자가 많은 것을 확인할 수 있다. 본 데이터는 실험에 사용할 수 있도록 표 2와 같이 학습 107명, 검증 35명, 테스트 47명으로 데이터가 분리되어있고 우울증 데이터와 비교해 볼 때, 비우울증 데이터가 더 많은 것을 알 수 있다.

표 1. 남녀 참가자 수
Table 1. Number of male and female participants

Gender	Participants	Depression	Non-depression
Male	102	25	77
Female	87	31	56
Total	189	56	133

표 2. 실험 데이터 세트 - 데이터 분리
Table 2. Experimental data sets - data separation

Class	Train	Validation	Test
Depression	30 (Male:13, Female:17)	12 (Male:5, Female:7)	14 (Male:7, Female:7)
Non-depression	77 (Male:50, Female:27)	23 (Male:11, Female:12)	33 (Male:16, Female:17)
Total	107	35	47

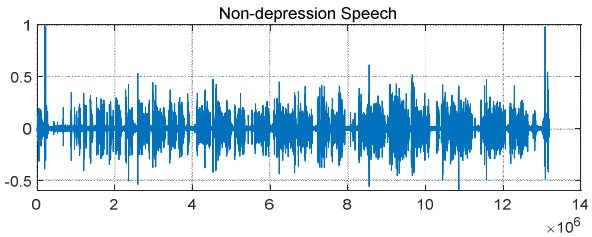
2.2 전처리 과정

음성 데이터는 묵음, 소음 등 다양한 노이즈들이 포함되어 있어 이를 제거하는 데이터 전처리 과정이 필요하다. DAIC-WOZ 데이터베이스를 구축하기 위해 참가했던 참가자들은 인터뷰를 진행할 때, 소음이 최대한 없는 환경 속에서 품질이 좋은 근접 마이크를 착용하고 있어 노이즈가 많이 발생하지 않았다. 따라서 가상 인터뷰어 'Elie'의 말소리와 묵음은 파이썬 라이브러리 중 하나인 pyAudioAnalysis의 세분화 모듈 사용하여 제거할 수 있었다.

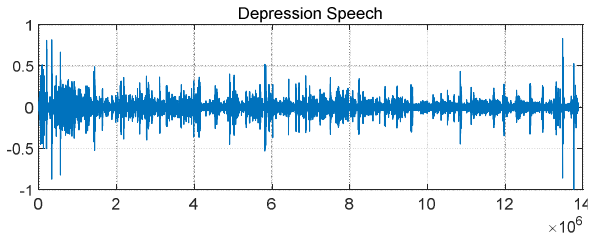
pyAudio-Analysis는 공개 소스 파이썬 라이브러리로, 특징 추출, 오디오 신호 분류, 분할, 콘텐츠 시각화 등 광범위한 오디오 분석 과정을 제공한다. 이는 이미 오디오 이벤트 감지, 음성 감정 인식, 시청

각 특징에 기반한 우울증 분류, 멀티모달 콘텐츠 기반 영화 추천 및 식습관 모니터링 애플리케이션 등 여러 오디오 분석 연구에서 사용되었다[9].

그림 2는 전처리 전의 우울증, 비우울증 음성 데이터를 각각 시각화한 그림이다. 그림 3에서는 전처리한 후의 음성 데이터를 확인할 수 있다. 이는 그림 2와 비교했을 때, 묵음과 다른 화자의 말소리가 제거된 것을 시각적으로 확인할 수 있다.

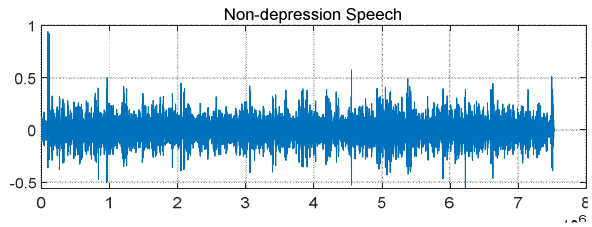


(a) 비우울증 음성
(a) Non-depression speech

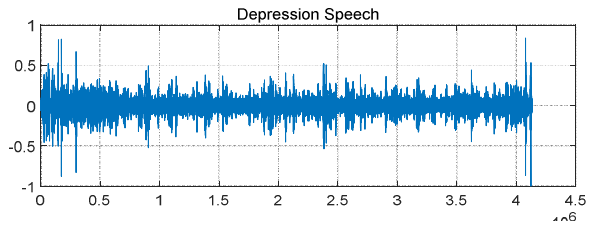


(b) 우울증 음성
(b) Depression speech

그림 2. 전처리 전 음성 데이터 시각화
Fig. 2. Visualize speech data before pre-processing



(a) 비우울증 음성
(a) Non-depression speech



(b) 우울증 음성
(b) Depression speech

그림 3. 전처리 후 음성 데이터 시각화
Fig. 3. Visualize speech data after pre-processing

가상 인터뷰어 'Elie'의 말소리 및 묵음을 제거하는 데이터 전처리 과정은 구글 코랩 환경에서 Python을 사용하여 이루어졌고, 이 외 데이터 증강, 특징추출, 학습, 검증 등 모든 실험은 MATLAB R2021a 환경에서 구현되었다.

III. 특징추출 및 딥러닝 모델

3.1 MFCC

MFCC는 오디오 신호에서 추출할 수 있는 특징으로, 음성의 고유한 특징을 나타내는 수치이다. 이는 사람의 달팽이관 특성을 모방한 Mel-scale을 사용하기 때문에 주로 음성인식, 화자 인식, 음성 합성, 음악 장르 분류 등 오디오 도메인의 문제를 해결하는 데 사용된다. MFCC는 주파수 도메인 특징으로 간주되며, 시간 도메인 특징보다 더 정확하다는 장점이 있다. 이러한 MFCC를 기술적인 측면에서 말하면, 멜 스펙트럼에서 켑스트럴 분석을 통해 추출된 특징값이다.

MFCC는 그림 4와 같은 방법으로 계산할 수 있다. 첫째, 오디로 신호를 프레임별로 나누어 FFT (Fast Fourier Transform)를 적용해 스펙트럼을 구한다. 그 후 앞서 구한 스펙트럼에 Mel Filter Bank를 적용해 멜 스펙트럼을 구한다. 마지막으로 이렇게 구한 멜 스펙트럼에 켑스트럴 분석을 적용하면 MFCC 값을 얻을 수 있다[10].

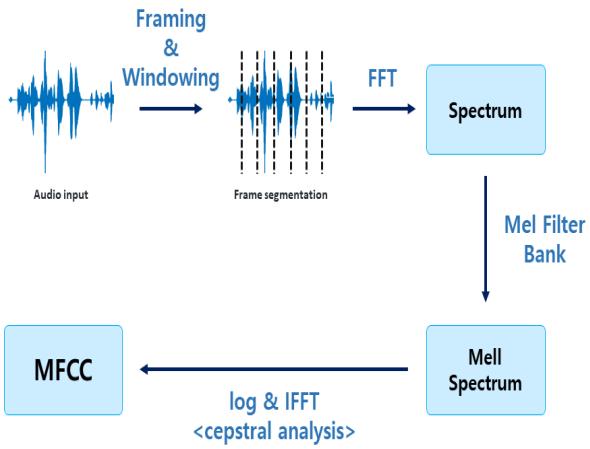


그림 4. MFCC의 특징추출 방법
Fig. 4. Features extract method of MFCC

3.2 GTCC

감마톤 켑스트럴 계수는 감마톤 필터 뱅크를 기반으로 생성된 계수이며, 이를 사용하면 MFCC 추출 방식과 유사한 방법으로 GTCC를 계산할 수 있다. 첫째, 원시 음성신호는 10~50ms 정도의 짧은 프레임으로 윈도우가 설정되는데, 본 연구에서 윈도우는 30ms로 설정했다. 그 후 각 프레임에 대해 FFT를 적용하여 프레임 스펙트럼을 분석한다. 다음으로 필터의 개수, 대역폭 등을 고려하여 GT 필터 뱅크를 생성하여 에너지를 계산한다. 마지막으로 신호가 로그 함수를 통해 전달되고 이산 코사인 변환을 적용하면 GTCC를 얻을 수 있다. GTCC는 식 (1)과 같이 계산할 수 있다.

$$GTCC_m = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} \log(X_n) \cos\left[\frac{\pi n}{N}\left(m - \frac{1}{2}\right)\right] \quad (1)$$

$$(1 \le m \le M)$$

여기서 X_n은 n번째 sub-band의 에너지이고, N은 감마톤 필터 개수, M은 GTCC의 개수를 나타낸다[11].

3.3 LSTM

LSTM은 기울기가 사라져버리는 장기의존성 문제를 가지고 있는 RNN 모델의 문제를 극복하기 위해 제안된 딥러닝 모델이다. 이는 메모리를 가지고 있는 셀(Cell) 상태를 유지하고 그림 5에서 보는 것과 같이 입력, 출력, 삭제 게이트를 통해 제어되는 구조이다.

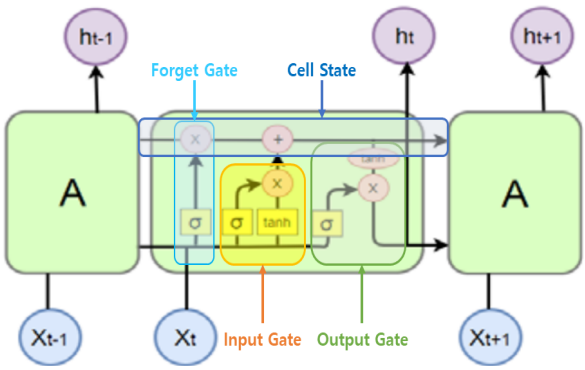


그림 5. LSTM 구조
Fig. 5. Structure of LSTM

이를 통해 기울기가 사라지는 것을 방지하고 오래전의 정보도 기억할 수 있게 하여 시계열/자연어 처리, 음성인식 등 다양한 분야에서 사용되고 있다.

그림 5에서 보면 LSTM은 RNN과 같이 체인 구조로 되어 있지만, 각 반복 모듈은 다른 구조로 되어 있다. 단순히 단일 신경망 layer가 아닌, 4개의 layer가 특별한 방식으로 서로의 정보를 주고받을 수 있도록 설계되어 있다[12].

3.4 Bi-LSTM

일반적인 LSTM 모델은 시계열 개념이 추가된 것으로 은닉층에서 과거의 정보를 기억할 수 있다. 하지만 입력 신호를 시간순으로 처리하기 때문에 직전 패턴을 기반으로 출력이 나오는 한계가 있다[13]. 이러한 문제를 해결하기 위해 Bi-LSTM은 두 개의 독립적인 LSTM 아키텍처를 함께 사용하는 구조이다. 그림 6에서 구조를 살펴보면 순방향으로 학습하는 LSTM에 역방향으로 학습할 수 있는 은닉층을 추가해 시퀀스 특징을 양방향으로 학습할 수 있도록 하여 의미적인 정보를 반영하고 그 정보를 충분히 이해할 수 있도록 설계되어 있다. 이는 추가된 역방향 은닉층을 통해 LSTM 모델보다 더 좋은 성능을 얻을 수 있다.

본 논문에서는 인터뷰 형식의 음성신호 이용하여 우울증을 진단하는 모델의 성능을 비교하기 위해 딥러닝 모델 중 하나인 Bi-LSTM 모델을 간단하게 설계하였다.

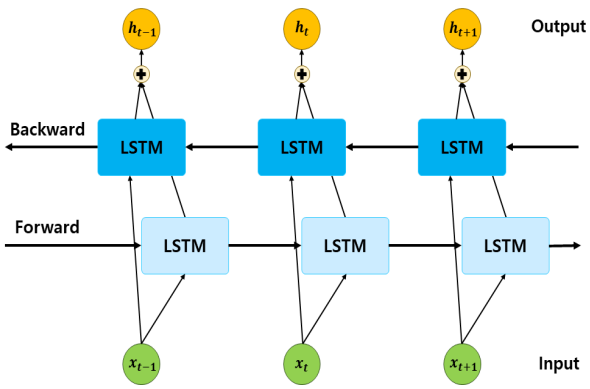


그림 6. Bi-LSTM 구조
Fig. 6. Structure of bi-directional LSTM

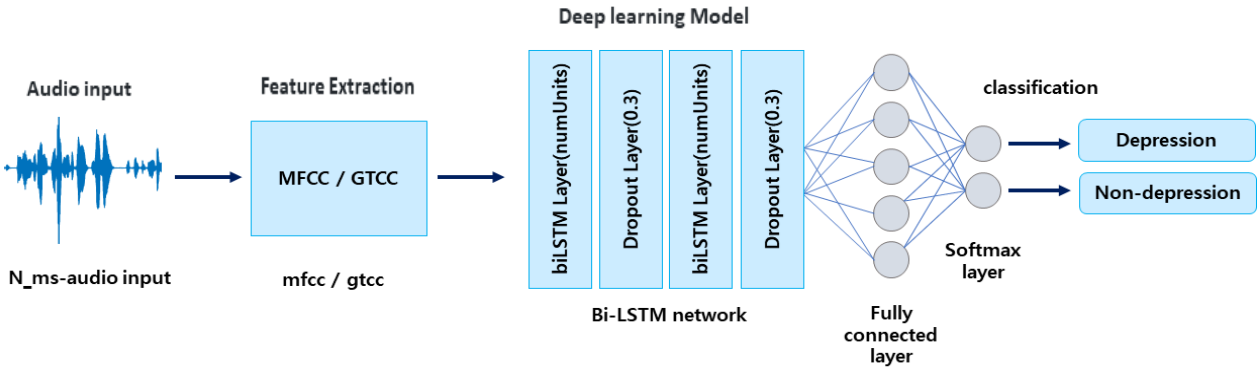


그림 7. 제안된 알고리즘 구조
Fig. 7. Structure of the proposed algorithm

제안된 알고리즘의 구조를 그림 7을 통해 살펴보면, MFCC, GTCC 방법을 이용해 음성의 특징을 추출하였고 Bi-LSTM 계층 2개를 쌓아 네트워크를 구성했다. Bi-LSTM 계층 중간에 30%의 확률을 가지는 드롭아웃 계층을 쌓아 데이터의 과적합 문제를 방지한다. 이렇게 구성된 네트워크를 통해 학습을 진행하고, 완전 연결계층과 소프트맥스 계층을 통과하여 이를 이진 분류한다.

Ⅳ. 실험 진행 과정 및 결과

4.1 실험 환경

본 논문에서 음성기반 우울증 데이터로 DAIC-WOZ 우울증 데이터베이스를 사용하였으며, 누락된 1개의 데이터를 제외한 188명의 음성 데이터를 사용하였다. 데이터의 80%는 학습데이터로, 나머지 20%는 검증 데이터로 사용하여 실험을 진행하였다. 본래 데이터베이스는 우울증 데이터 55개, 비우울증 데이터 133개로 비우울증 데이터가 우울증 데이터와 비교했을 때 약 3배 더 많아 클래스 불균형 문제가 발생한다. 이러한 문제를 해결하기 위해 추가로 데이터 확장 후에 Bi-LSTM의 검증 성능을 비교하는 실험을 진행하였다. 우울증 데이터를 pitch shifting, time shifting, adding noise 등을 이용하여 각 파일마다 3개씩 데이터 확장을 진행하였고, 이를 통해 두 데이터에 대한 클래스 불균형을 해결해주었다. 우울증 진단 시스템의 성능을 확인하기 위해 제안된 모델에 대한 혼동 행렬을 시각화하여 실제 성능과 예측성능을 비교 및 분석하였다. 이 모든 실험은 MATLAB R2021a 환경에서 수행되었다.

4.2 성능평가 및 결과

본 논문에서는 Bi-LSTM의 은닉유닛 개수를 10개로 지정하고, MFCC와 GTCC를 같이 이용하여 음성신호의 특징을 추출하는 방법을 제안한다. 제안된 방법의 성능을 비교 및 평가하기 위해 음성신호의 특징추출 방법과 은닉유닛 개수를 변경해가며 실험을 진행하였다. 사용된 특징추출 방법에는 MFCC, GTCC, MFCC+GTCC가 있고, 은닉유닛 개수는 임의로 10개, 50개, 100개로 지정하였다. 본 실험은 음성데이터 확장 전과 후로 나누어 Bi-LSTM의 검증 성능을 비교하고 분석하였다.

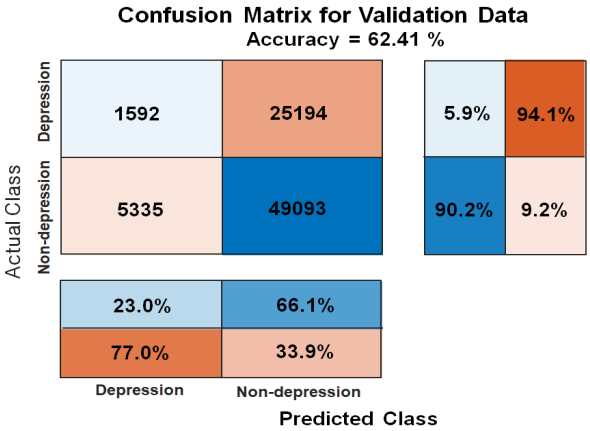
표 3에서는 데이터 확장 전, 특징추출 방법, 은닉유닛의 개수를 바꿔가며 실험을 진행한 Bi- LSTM의 검증 성능을 확인할 수 있다. 이 실험에서는 특징추출 방법은 GTCC를 사용하고 Bi-LSTM의 은닉유닛 개수가 10개일 때 검증 성능이 62.41%로 가장 높았다.

표 3. 데이터 확장 전, Bi-LSTM 검증 성능 비교
Table 3. Comparison of bi-directional LSTM verification performance before data augmentation

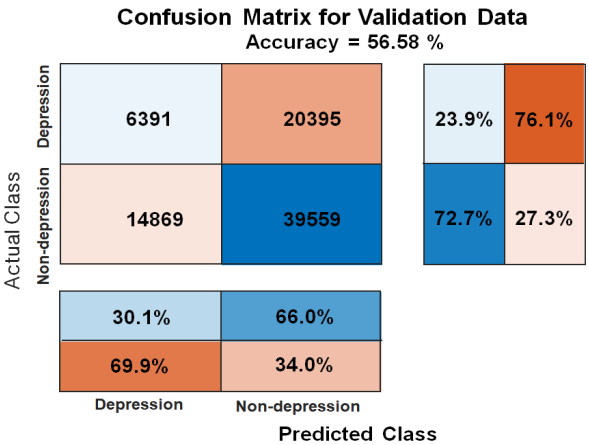
Feature extract \ Hidden units	10	50	100
MFCC	62.01%	59.49%	60.15%
GTCC	62.41%	59.76%	60.25%
MFCC+GTCC (Proposed method)	56.58%	59.02%	59.21%

반면에 본 논문에서 제안한 방법인 은닉유닛 개수가 10개이고 MFCC와 GTCC를 같이 이용하여 특징을 추출했을 때, 약 56.58%로 가장 낮은 성능을 보인다. 그림 8에서 이에 대한 혼동 행렬을 볼 수 있다.

그림 8(a)는 은닉유닛 개수는 10개, GTCC 방법을 이용했을 때의 혼동 행렬을 (b)는 은닉유닛 개수는 10개, MFCC+GTCC 방법을 이용했을 때의 혼동 행렬을 나타내고 있다. GTCC 방법만을 이용했을 때, 비우울증 데이터는 대체적으로 잘 분류하고 있지만, 우울증 데이터는 거의 분류하지 못하는 것을 확인할 수 있다. 반면 MFCC+GTCC 방법을 이용했을 때, GTCC 방법만 이용한 것에 비해 우울증 데이터를 조금 더 잘 분류하는 것을 볼 수 있다.



(a) 은닉유닛 10개, GTCC 이용
(a) 10 hidden units, using GTCC



(b) 은닉유닛 10개, MFCC+GTCC 이용
(b) 10 hidden units, using MFCC+GTCC

그림 8. 데이터 확장 전, Bi-LSTM 모델의 혼동 행렬
Fig. 8. Confusion matrix of bi-directional LSTM models before data augmentation

데이터 확장 전에는 클래스 불균형 문제로 데이터가 많은 비우울증 데이터는 우울증 데이터에 비해 잘 분류하지만, 우울증 데이터는 제대로 분류하지 못한 것을 확인할 수 있다.

표 4에서는 우울증 데이터를 각 파일마다 3개씩 확장 후, 특징추출 방법, 은닉유닛의 개수에 따른 Bi-LSTM의 검증 성능을 확인할 수 있다.

표 4. 데이터 확장 후, Bi-LSTM 검증 성능 비교
Table 4. Comparison of bi-directional LSTM verification performance after data augmentation

Feature extract \ Hidden units	10	50	100
MFCC	98.27%	98.11%	97.99%
GTCC	96.99%	96.66%	97.52%
MFCC+GTCC (Proposed method)	**98.49%**	97.90%	97.86%

이 실험에서는 본 논문에서 제안한 방법인 MFCC와 GTCC를 같이 사용하여 특징을 추출하는 방법을 사용하고 은닉유닛의 개수가 10개일 때 Bi-LSTM의 성능이 98.49%로 가장 높았다. 임의로 설정한 은닉유닛의 개수는 반복적인 실험을 통해 은닉유닛 개수가 10개였을 때 가장 적합하다는 결론을 도출했다.

그림 9에서 이에 대한 혼동 행렬을 살펴보면, 예측 클래스와 실제 클래스에서 우울증과 비우울증을 대체적으로 잘 분류하고 있는 것을 확인할 수 있다.

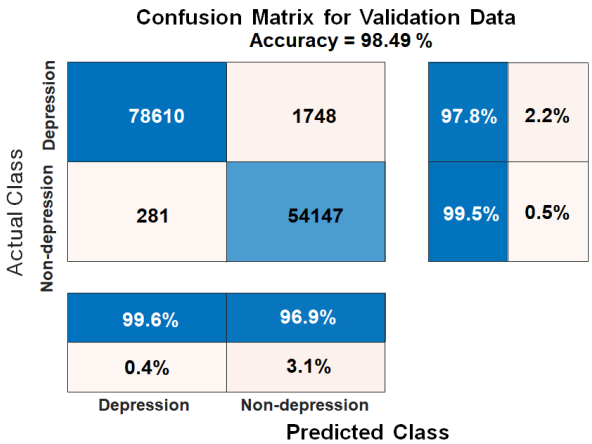
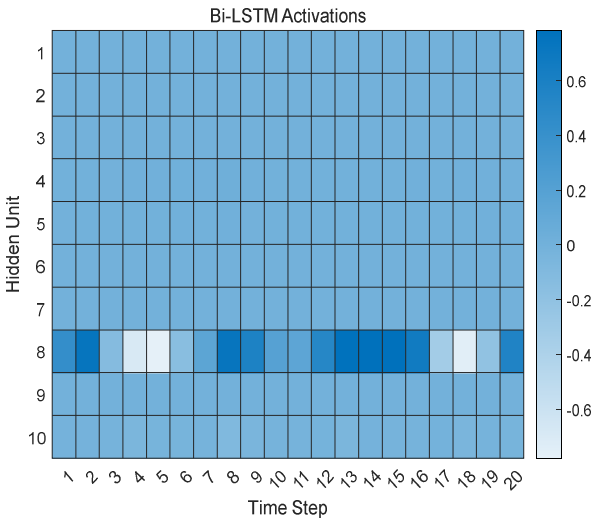
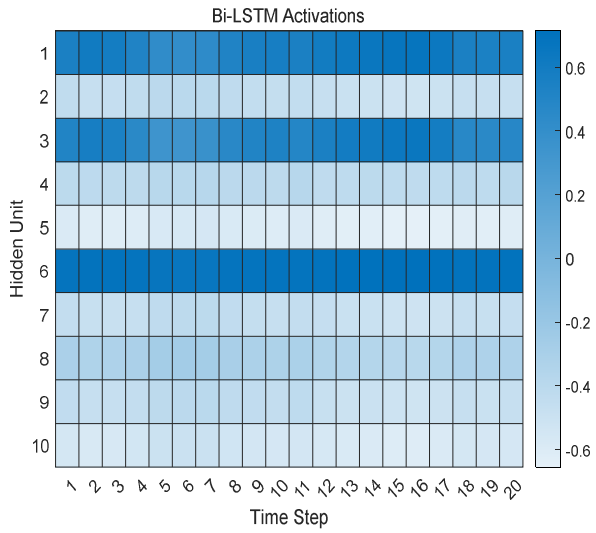


그림 9. 데이터 확장 후, Bi-LSTM 모델의 혼동 행렬
Fig. 9. Confusion matrix of bi-directional LSTM models after data augmentation

그림 10은 히트맵을 통해 시간 지남에 따른 Bi-LSTM의 은닉유닛의 활성화 영역을 보여주고 있다. 히트맵은 각 유닛이 얼마나 강력하게 활성화되는지 보여주고 시간이 지남에 따라 활성화가 어떻게 변하는지 강조한다. 그림 10(b)를 통해 두 번째 Bi-LSTM 계층 은닉유닛의 활성화 영역을 보면, 6번째 은닉유닛이 가장 강하게 활성화되었고, 이는 다른 은닉유닛들과 비교했을 때 최종 출력에 가장 많은 영향을 미친다는 것을 알 수 있다.



(a) 첫 번째 Bi-LSTM 계층
(a) First Bi-LSTM layer



(b) 두 번째 Bi-LSTM 계층
(b) Second Bi-LSTM layer

그림 10. Bi-LSTM Layers 활성화 영역
Fig. 10. Visualize activations of Bi-LSTM network

데이터 확장 전과 확장 후 Bi-LSTM 모델의 우울증 진단 성능을 비교해보았을 때, 데이터 확장 전은 약 56.58%, 확장 후는 98.49%로 데이터 확장을 통해 성능이 약 42% 정도 개선되었다. 클래스 불균형 문제가 해결되기 전에는 본 논문에서 제안한 방법이 가장 낮은 성능을 보였다. 그러나 이러한 클래스 불균형 문제가 해결되었을 때, 본 논문에서 제안한 은닉유닛의 개수를 10개로 지정하고 MFCC와 GTCC를 같이 사용해 특징을 추출하여 학습을 진행하는 방법이 가장 효율적인 것을 증명했다.

Ⅴ. 결론 및 향후 연구 계획

본 논문에서는 1차원 음성신호를 이용하여 Bi-LSTM 기반으로 우울증을 진단하는 모델을 제안하고 그 성능을 비교 분석했다. 우울증 진단 모델을 설계하기 위해 사용된 음성 데이터베이스는 DAIC-WOZ 우울증 데이터베이스로 총 189명의 인터뷰 형식으로 된 음성 파일을 포함한다. 본 실험에서는 1개의 누락 된 파일을 제외한 188명에 대한 데이터를 사용하였고, 클래스 불균형을 해결하기 위해 데이터 확장을 수행하였다. 본 논문에서는 Bi-LSTM의 은닉유닛 개수를 10개로 지정하고, MFCC와 GTCC를 같이 이용하여 음성신호의 시퀀스 특징을 추출하는 방법을 제안하였다.

제안된 방법의 성능평가를 위해 실험은 데이터 확장 전과 후 두 가지로 나누어 진행하였고, Bi-LSTM 모델에 대한 검증 성능을 확인하였다. 데이터 확장 전과 후 제안된 방법에 대한 성능을 비교했을 때, 데이터 확장 전에는 가장 낮은 성능을 보였지만 데이터 확장을 통해 클래스 불균형 문제가 해결됨으로써 분류 성능이 개선되어 가장 높은 성능을 갖는다는 것을 증명했다.

향후에는 1차원 음성신호가 아닌, 이를 시간-주파수 표현의 2차원 이미지로 변환하여 CNN 기반 전이 학습 모델을 통해 우울증을 진단하는 연구를 진행할 예정이다. 또한 LSTM 모델과 CNN 모델을 앙상블하여 시스템을 구현하는 연구를 진행하게 된다면 더 정확하게 우울증을 진단할 수 있는 시스템으로 발전할 가능성이 높아질 것으로 예상된다.

References

[1] Choi B, Shim G, Jeong B, and Jo S., "Data-driven analysis using multiple self-report questionnaires to identify college students at high risk of depressive disorder", Scientific Reports, Vol. 10, No. 1, pp. 7867, May 2020. https://doi.org/10.1038/s41598-020-64709-7.

[2] Lin Lin, Xuri Chen, Ying Shen, and Lin Zhang, "Towards automatic depression detection: A BiLSTM/1D CNN-based model", Ap- plied Sciences, Vol. 10, No. 23, pp. 1-20, Dec. 2020. https://doi.org/10.3390/app10238701.

[3] N. S. Srimadhur and S. Lalitha, "An End-to-End model for detection and assessment of depression levels using speech", Procedia Computer Science, Vol. 171, pp. 12-21, Jan. 2020. https://doi.org/10.1016/j.procs.2020.04.003.

[4] Tuka Al Hanai, Mohammad Ghassemi, and James Glass, "Detecting depression with audio/text sequence modeling of interviews", Interspeech, pp. 1716-1720, Sep. 2018. https://doi.org/10.21437/Interspeech.2018-2522.

[5] E. Rejaibi et al., "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech", Preprint submitted to Signal Processing: Image Communication, pp. 1-14, Mar. 2020. https://doi.org/10.1016/j.bspc.2021.103107.

[6] Karol Chlasta, Krzysztof Wołk, and Izabela Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks", Procedia Computer Science, pp. 618-628, Dec. 2019. https://doi.org/10.1016/j.procs.2019.12.228.

[7] G. Jonathan et al., "The Distress Analysis Interview Corpus of human and computer interviews", Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, pp. 3123-3128, May 2014.

[8] Kroenke K. and Strine TW et al., "The PHQ-8 as a measure of current depression in the general population", Journal of Affective Disorders, Vol. 144, pp. 163-173, Apr. 2009. https://doi.org/10.1016/j.jad.2008.06.026.

[9] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis", PLOS ONE, Vol. 10, No. 12, pp. 1-17, Dec. 2015. http://doi.org/10.1371/journal.pone.0144610.

[10] D. Prabakaran and S. Sriuppili, "Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation", Journal of Physics: Conference Series, Tamil Nadu, India, Vol. 1717, No. 1, pp. 1-8, Jan. 2021. http://dx.doi.org/10.1088/1742-6596/1717/1/012009.

[11] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non -Speech Audio Classification", IEEE Transac tions on Multimedia, Vol. 14, No. 6, pp. 1684-1689, May 2012. https://doi.org/10.1109/TMM.2012.2199972.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural Computation, Vol. 9, No. 8, pp. 1735-1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[13] T. Robinson, "An application of recurrent nets to phone probability estimation", IEEE Transactions on Neural Networks, Vol. 5 No. 2, pp. 298-305, Mar. 1994. https://doi.org/10.1109/72.279192.

저자소개

조 아 현 (A-Hyeon Jo)

2021년 2월 : 조선대학교 전자공학과(공학사)
2021년 3월 ~ 현재 : 조선대학교 전자공학과 IT-Bio 융합시스템 전공 (석사 과정)
관심분야 : 음성신호 처리, 딥러닝

곽 근 창 (Keun-Chang Kwak)

2002년 : 충북대학교 전기공학과 박사 졸업
2003년 ~ 2005년 : 캐나다 앨버타대학교 전기 및 컴퓨터 공학과, 박사후과정
2005년 ~ 2007년 : 한국전자통신연구원 지능형로봇단 선임연구원
2014년 ~ 2015년 : 미국 캘리포니아주립대학교 플러튼, 방문교수
2007년 ~ 현재 : 조선대학교 전자공학부 교수
관심분야 : 계산지능, 인간-로봇상호작용, 바이오인식