JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 9, pp. 79-87, September 2024
https://doi.org/10.9708/jksci.2024.29.09.079

Design and Development of Open-Source-Based Artificial Intelligence for Emotion Extraction from Voice

Seong-Gun Yun*, Hyeok-Chan Kwon*, Eunju Park**, Young-Bok Cho***

*Student, Dept. of Software Convergence, Andong National University, Andong, Korea
**Professor, Dept. of Software Convergence Center, Andong National University, Andong, Korea
***Professor, Dept. of Computer Education, Andong National University, Andong, Korea

[Abstract]

This study aims to improve communication for people with hearing impairments by developing artificial intelligence models that recognize and classify emotions from voice data. To achieve this, we utilized three major AI models: CNN-Transformer, HuBERT-Transformer, and Wav2Vec 2.0, to analyze users' voices in real-time and classify their emotions. To effectively extract features from voice data, we applied transformation techniques such as Mel-Frequency Cepstral Coefficient (MFCC), aiming to accurately capture the complex characteristics and subtle changes in emotions within the voice. Experimental results showed that the HuBERT-Transformer model demonstrated the highest accuracy, proving the effectiveness of combining pre-trained models and complex learning structures in the field of voice-based emotion recognition. This research presents the potential for advancements in emotion recognition technology using voice data and seeks new ways to improve communication and interaction for individuals with hearing impairments, marking its significance.

▸ **Key words**: Emotion Recognition, Transformer, HuBERT, Mel-Frequency Cepstral Coefficient

[요 약]

본 연구는 청각 장애인의 의사소통 개선을 목표로, 음성 데이터에서 감정을 인식하고 분류하는 인공지능 모델을 개발하였다. 이를 위해 CNN-Transformer, HuBERT-Transformer, 그리고 Wav2Vec 2.0 모델을 포함하는 세 가지 주요 인공지능 모델을 활용하여, 사용자의 음성을 실시간으로 분석하고 감정을 분류한다. 음성 데이터의 특징을 효과적으로 추출하기 위해 Mel-Frequency Cepstral Coefficient(MFCC)와 같은 변환 방식을 적용, 음성의 복잡한 특성과 미묘한 감정 변화를 정확하게 포착하고자 하였다. 실험 결과, HuBERT-Transformer 모델이 가장 높은 정확도를 보임으로써, 음성 기반 감정 인식 분야에서의 사전 학습된 모델과 복잡한 학습 구조의 융합이 효과적임을 입증하였다. 본 연구는 음성 데이터를 통한 감정 인식 기술의 발전 가능성을 제시하며, 청각 장애인의 의사소통과 상호작용 개선에 기여할 수 있는 새로운 방안을 모색한다는 점에서 의의를 가진다.

▸ **주제어**: 인공지능, 감정인식, 트랜스포머, HuBERT, Mel-Frequency Cepstral Coefficient

• First Author: Seong-Gun Yun, Corresponding Author: Young-Bok Cho
*Seong-Gun Yun (canyun0460@naver.com), Dept. of Software Convergence, Andong National University
*Hyeok-Chan Kwon (hc226@naver.com), Dept. of Software Convergence, Andong National University
**Eunju Park (eunju@anu.ac.kr), Dept. of Software Convergence Center, Andong National University
***Young-Bok Cho (ybcho@anu.ac.kr), Dept. of Computer Education, Andong National University
• Received: 2024. 08. 12, Revised: 2024. 09. 19, Accepted: 2024. 09. 21.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

IT기술의 발전으로 실시간 통역기와 같은 언어의 장벽을 넘어 효율적 의사소통이 가능한 기술들이 개발되고 있다[1]. 그러나 대다수 일반인들이 소리로 소통하는 세계에서 수화를 제 1언어로 사용하는 청각장애인들은 차별받는 소수집단 구성원으로 의사소통 시 음성 통화를 사용하지 못하고, 기존 문자언어 기반 통신 서비스에서도 의사소통의 한계를 가진다[2]. 감정은 사람의 마음을 표출하는 중요한 요소로 음성은 감정을 대외적으로 나타내는 중요한 수단의 하나이다[3]. 청각장애인들과 비장애인들은 동일한 사회에서 살아가지만 각기 다른 고유한 문화들을 가져 동일한 상황에 대하여 다른 감정으로 수용하는 경우들이 있으며 이에 따른 의사소통에 문제를 가질 수 있다[4]. 이렇듯 청각 장애인들은 청각의 결함으로 인해 의사소통의 불편함과 긴급상황 인지의 어려움 등 공동체 사회에서 단절 및 소외감을 가지고 있다[1].

음성 감정 인식(SER: Speech Emotion Recognition)은 감정 감지 인식의 종류 중 하나로 사용자의 목소리에서 나타나는 떨림, 어조, 크기 등의 음성 패턴을 분석하여 감정의 상태를 파악하는 기술이다[5]. 코로나19 비대면 환경 전환에 따라 음성기반의 다양한 기술개발 수요가 증가함에 따라 인공지능, 기계 학습, 클라우드 기술 및 자동화 기술과 함께 성장하고 있다[6]. 또한, 음성 기반 연구는 인간과 기계의 상호작용 증가에 따라 스마트폰의 음성 대화 애플리케이션의 문자 메시지 작성, 정보 검색, 자동차의 내비게이션의 음성 명령 등의 간단한 사용에서 ChatGPT, AI 스피커, 음상비서 사용으로 확장되고 있다[7-8]. 이는 사람의 업무를 다양한 방면에서 보조하고 효과적 정보를 제공받을 수 있어 상호작용의 중요성이 강조되는 서비스로 발전하고 있는 것이다[9].

본 연구에서는 청각 장애인과의 의사소통 및 상호작용 향상을 위한 음성 감정 추출 방법을 제안하였다, 본 연구에서는 오디오 파일을 원시 음성 데이터로 자기 지도 학습 방법을 사용하여 감정을 인식하고 분류한다. 음성 처리 작업을 위해 오픈소스 모델인 HuBERT 모델과 오디오 파일을 원시 음성 파형으로 받아 특성을 추출하는 wav2vec모델을 활용하여 화자의 음성으로부터 감정을 인식 및 분류하여 추출하는 방법을 설계하고 구현하였다. 본 논문에서는 데이콘의 음성 감정 인식 데이터셋을 활용하여 두려움, 슬픔, 싫어함, 평범함, 행복함, 화남, 총 6가지의 감정을 추출하였다. 음성 감정 추출에 필요한 높은 성능을 가진 오픈소스 모델을 기반으로 음성에서 감정을 추출함으로 음성을 제대로 인지하지 못하는 청각 장애인의 의사소통 상황을 개선하고 의사소통 시 상호작용에 도움이 될 것이라 사료된다.

II. Related works

음성신호의 특징을 추출·분석 기술은 음성신호에서 정보를 얻는 기술로 최근에는 딥러닝을 활용한 기술들이 많이 활용되고 있다[10].

1. Mel-Frequency Cepstral Coefficient(MFCC)

음성 신호의 특징 추출을 목적으로 사용되는 알고리즘으로 Mel-Frequency Cepstral Coeffiicient(MFCC)가 있다. MFCC는 음성 신호를 짧은 프레임으로 분할하고 각 프레임에 대한 파워 스펙트럼(Power Spectrum)의 주기도 추정값을 계산한다. 멜 필터뱅크(Mel Fillterbank)를 파워 스펙트럼에 적용하고, 각 필터의 에너지를 모두 더한다. 모든 필터뱅크 에너지에 대해 로그를 취하고 로그 필터뱅크 에너지의 DCT(Discrete Cosine Transform)를 취한다. DCT(Discrete Cosine Transform) 계수에서 2~13을 제외한 나머지는 버리고 일반적으로 Deltas and Delta-Deltas(미분 및 가속도 계수) 기능을 추가하거나 프레임 에너지를 각 특징 벡터에 추가한다[11]. 이러한 방식으로 음성 신호로부터 MFCC를 사용하여 음성 신호 특징 추출 작업을 진행한다. Python의 라이브러리인 Librosa를 통해서 간편하게 사용할 수 있다. 이렇게 음성 신호 특징 추출에 특화된 MFCC는 화자를 구분하거나 음악의 장르를 분류하는 등 여러 음성 인식 작업에서 유용하게 사용되며 EEG(Electro Encephalo graphy), ECG(Electro Cardio Gram) 및 산업 신호와 같은 응용 분야에서 유망하다[12]. 본 논문에서는 MFCC 알고리즘을 활용하여 음성 데이터를 전처리하고 전처리된 데이터를 바탕으로 학습을 진행하였다.

2. Transformer Model

Transformer 모델은 주로 자연어 처리(NLP) 분야에서 활용되며 인코더와 디코더라는 모듈을 가지고 있는 seq2seq(sequence-to-sequence)모델로부터 발전되었다. Transformer 모델은 RNN을 사용하지 않고 크게 인코더와 디코더로 각각 여러 층으로 구성되어 있으며 Attention으로는 Scaled Dot-Product Attention과 Multi-Head Attention을 가진다. Attention 하위 레이어

외에도 인코더와 디코더의 각 층(layer)에 완전 연결된 피드 포워드 네트워크가 포함된다. Transformer 모델의 핵심은 Attention 메커니즘에 있으며 세 가지 방식으로 사용된다, 인코더 내의 Self-Attention, 디코더 내의 Self-Attention 그리고 인코더 - 디코더 Attention으로 위치별로 위치를 참조, 다음 토큰 예측, 다음 토큰 생성을 할 수 있다. Self-Attention은 모델의 핵심요소로 다양한 위치간의 글로벌 의존성을 모델링 하여 시퀀스의 순차적 처리 문제를 해결하고 계산을 병렬화하여 학습 속도를 향상한다[13]. 본 논문에서는 음성의 복잡한 특징들을 이해하고 학습하기 위해 Transformer 모델을 활용하여 인공지능 학습을 진행하였다.

3. HuBERT Model

HuBERT 모델은 마스킹된 예측을 통해 강력한 음성 표현을 학습할 수 있는 모델로 자기지도 학습 방식을 사용했다. HuBERT 모델에는 Hidden Unit이라는 음성 데이터에서 정의되지 않고 발견되어야 하는 숨겨진 음향 단위를 학습의 타겟으로 사용한다. 데이터 자체에서 특징을 추출하여 Hidden Unit을 발견하고 음성 데이터의 패턴과 구조를 모델이 자동으로 인식하고 학습한다. HuBERT 모델의 자기지도 음성 표현 학습은 다음과 같다. 학습 데이터로는 LibriSpeech와 Libri-light 데이터셋을 사용하고 k-means 알고리즘과 GMMs(가우시안 혼합 모델)을 사용하여 음성 데이터로부터 Hidden Unit을 추론한다. 하나 이상의 k-means 클러스터링 반복으로 생성된 마스킹된 프레임의 숨겨진 클러스터 할당을 예측한다. 위 방식으로 반복적으로 학습된 모델은 최첨단 기술과 비교하였을 때 우수한 성능을 보였고 모델의 입력으로 파형을 사용하고 반복적 개선이 더 나은 유닛을 학습한다[14]. 본 논문에서는 HuBERT 모델의 반복적인 학습과 Fine-Tuning을 사용하여 음성 인식의 성능을 향상하여 음성 데이터로부터 감정을 추출하는 방법을 제안하였다.

4. wav2vec 2.0 Model

음성 인식 시스템의 학습은 대량의 레이블 작업이 되어 있는 학습데이터를 요구하지만 실제로 이러한 데이터는 구하기 어렵다. wav2vec 2.0 모델은 자기지도 학습과 소량의 레이블 작업이 되어있는 음성 데이터를 통해 Fine-Tuning되어 우수한 성능을 달성할 수 있다[15]. wav2vec 2.0 모델은 원시 오디오 데이터로부터 잠재 음성 표현을 학습한다. 멀티-레이터 컨볼루전 특징 인코더로 구성되어 오디오 데이터를 인코딩하고, Transformer 네트워크로 컨텍스트화된 표현을 구축하고 Self-Attention은 전체 시퀀스의 잠재 표현을 포착한다. 이는 각 단어 또는 토큰이 문장 전체의 다른 단어나 토큰에 얼마나 의존하는지 파악하는 것이다. 모델의 훈련은 잠재 기능 인코더 공간에서 특정 부분을 컨텍스트 네트워크에 공급하기 전에 마스킹하고 모든 마스킹 타임 스탭 사이에 공유되는 훈련된 특징 벡터로 교체한다. 이렇게 마스킹된 표현을 식별하며 최종 모델은 레이블 작업이 완료된 데이터에 대해 Fine-Tuning된다. HuBERT 모델과 마찬가지로 같은 데이터셋인 LibriSpeech와 Libri-light 데이터셋을 사용하여 모델을 평가하였으며 레이블 지정이 없는 오디오 데이터에서 사전훈련되고 레이블 작업이 진행된 데이터로 Fine-Tuning을 하여 wav2vec 2.0 실험에서는 언어 모델을 사용하여 디코딩의 성능을 개선하는 방법을 사용하였다. 결과적으로 wav2vec 2.0모델은 레이블 작업이 매우 적게 되어 있는 데이터를 사용한 작업에서 강력한 성능을 보이며 이는 음성 인식 분야에서 자기지도 학습 방법의 잠재력이 있다는 것을 보여준다. 본 연구에서 활용한 오픈소스 wav2vec 2.0모델은 데이터셋의 감정차원에 맞춰 Fine-Tuning 전에 Transformer 레이어를 24개에서 12개로 정리한 모델이다. 본 연구에서는 wav2vec 2.0모델에 분류기를 추가하여 감정을 분류하여 추출하였다.

5. CNN(Convolution Neural Network)

CNN은 인간의 시신경 구조를 모방하여 이미지나 영상 데이터 인식 및 분류 작업에 특화된 인공신경망이다[16]. CNN은 계층적 구조를 가지고 있으며 입력 이미지를 여러 계층으로 나누어 처리한다. 이 구조는 이미지의 로컬 특징을 효과적으로 학습하여 복잡한 패턴 및 객체를 인식할 수 있다. 주로 컨볼루션(Convolution)레이어, 풀링(Pooling)레이어, 완전 연결(Fully-Connected)레이어 등으로 구성되어 있다[17]. 컨볼루션(Convolution)레이어는 로컬 정보를 활용하여 입력된 이미지로부터 패턴 및 특징을 감지하고 추출하고 풀링 레이어는 특징 맵에 다운 샘플링 작업을 통해 모델이 더 빠르게 학습하고 계산량을 감소시킨다. 완전 연결(Fully-Connected)레이어는 2차원의 배열 형태의 이미지를 1차원 배열로 평탄화하는데 사용된다. CNN은 계층적 구조의 특징이외에도 변환 불변성, 특징 추출을 자동으로 진행하는 특징이 있다[18]. 이러한 특징을 가진 CNN은 컴퓨터 비전 능력에 뛰어난 성능을 보여 많은 다양한 분야에 활용이 되고 음성 파형 이미지로부터 특징을 추출하여 감정 분류 작업에 유용하다고 판단된다.

6. Model Comparison and Analysis

Table 1. Summary of AI Model Features, Adantages and Disadvantages

Model	Key Features	Advantages	Disadvantages
MFCC	Algorithm for extracting features from audio signals	Simple and fast feature extraction	Limited in recognizing complex patterns
Transformer	Sequence learning using the Attention mechanism	Fast learning through parallel processing	Requires large amounts of data and resources
HuBERT	Self-supervised learning model for audio signal pattern recognition	High performance even with unlabeled data	Slow learning speed
Wav2Vec 2.0	Model that maximizes learning performance with small amounts of labeled data	High performance with small data	Requires complex computation and fine-tuning
CNN	Model specialized in image data processing	Capable of learning by converting audio waveforms into images	Image preprocessing is required

이처럼 MFCC, Transformer, Hubert, Wav2Vec 2.0, CNN 등 다양한 모델과 알고리즘을 사용하여 음성 신호의 특징을 추출하고 감정을 분류할 수 있다. 각 모델은 고유한 장점과 단점을 가지고 있으며, 그에 따라 특정 작업에 적합한 모델이 결정된다.

III. The Proposed Scheme

1. Architecture Design

본 논문에서는 인공지능 구축 설계 및 구현 방법으로 파이프필터 패턴을 사용하였으며 이 패턴에 대한 다이어그램은 그림1과 같다.

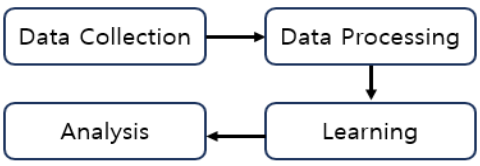


Fig. 1. Architecture Diagram

파이프 필터 패턴이란 소프트웨어 아키텍처의 설계 패턴 중 하나로, 데이터 스트림을 처리하는 방식에 대한 모델을 제공한다. 이 패턴은 일련의 독립적인 처리단계로 구성되어 있으며, 각 처리 단계는 입력 데이터를 받아 처리한 후 그 결과를 다음 단계로 전달한다. 본 연구는 처리단계를 데이터 수집, 데이터 전처리, 인공지능 모델 학습, 검증데이터의 정확도 확인의 순서로 설계하였다.

2. Data Collection

음성 감정 추출을 위한 인공지능 학습을 위한 데이터로는 데이콘의 음성 감정 인식 데이터셋을 활용하였다[19]. 그림 2는 데이콘의 음성 감정 인식 데이터셋이다.

Fig. 2. Dacon Voice Dataset

데이터셋의 구성은 훈련 데이터(Train) 5001개와 테스트 데이터(Test) 1881개로 이루어져 있다. 각 음성 파일은 해당 음성이 표현하는 감정 상태에 대한 라벨을 CSV 파일 형태로 제공된다. 이러한 구성은 모델이 학습 과정에서 각 음성의 특성을 감정 라벨과 연관짓게 하여, 새로운 음성 데이터에 대해 감정을 정확히 예측할 수 있도록 한다. 데이터셋 음성 WAV파일의 라벨링은 0: angry, 1: fear, 2: sad, 3: disgust, 4: neutral, 5: happy으로 총 6가지의 감정으로 분류되어 있다. 해당 데이터셋을 활용하여 음성에서 감정을 인식하고 분류하는 인공지능을 구현하였다.

3. Data Preprocessing

데이터의 전처리는 CNN-Transformer 학습을 위하여 음성의 특징을 파악하는 MFCC 방법과 음성 자체를 인식하는 휴버트 모델 및 wav2vector모델의 재학습을 위한 음성 패딩 방법 2가지를 사용하였다.

첫 번째로 MFCC를 활용하여 전처리를 진행하였다. MFCC는 인간의 귀가 주파수영역을 파악하는 방식을 토대로 음성의 특징을 추출하는 방식이다. 본 연구에서는 python의 라이브러리인 librosa를 활용하여 MFCC의 음성 특징을 추출하였다. 추출된 특성들은 시간 길이에 따라 패딩 또는 슬라이싱 작업을 통해 통일된 형태로 조정되는 과정을 거친 후에 정규화 작업을 진행한다. 그림 3은 음성

을 MFCC를 시각화 한 이미지이다.

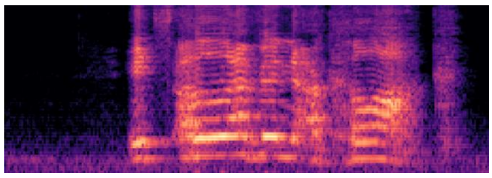


Fig. 3. MFCC Image

MFCC 알고리즘을 활용하여 변환된 이미지에서 가로축은 시간 축이며 세로 축은 주파수 축을 의미한다. 또한 색상의 색은 주파수 성분의 강도를 의미한다. 따라서 이미지의 왼쪽 부분은 거의 검게 보이며 음성의 초반 구간에서는 음성 신호의 에너지가 거의 없거나 낮음을 나타내며 중앙의 밝은 띠들은 해당 시간대의 특정 주파수 대역의 에너지가 강하다는 것을 의미한다.

두 번째 전처리는 휴버트(HuBERT) 모델 및 wav2vector 재학습을 위하여 음성 패딩을 진행하였다. 휴버트 모델은 음성 신호에서 의미 있는 패턴을 학습하여 분류하거나 특징을 추출하는 데 사용된다. 음성 데이터의 길이는 다양하기 때문에 모델이 일관된 크기의 입력을 처리할 수 있도록 모든 음성 데이터를 동일한 길이로 조정하는 것이 필요하다. 이를 위해 짧은 음성 데이터에는 무음 구간을 추가하여 길이를 동일하게 맞추고, 긴 음성 데이터는 필요한 부분만을 잘라내어 사용한다. 이 과정을 음성 패딩이라 한다.

Algorithm 1: Preprocessing and Padding Audio Files

```
Input: batch of audio files
Output: processed and padded audio files
Function preprocess_and_pad_audio(batch):
    1. Initialize an empty list: processed_audios;
    2. Initialize an empty list: lengths;
    3. foreach audio_path in batch['audio'] do
        4. audio = load audio file from audio_path at 16000 Hz;
        5. input_values = feature extraction from audio using
           processor;
        6. Append input_values to processed_audios;
        7. Append the size of input_values to lengths;
    end
    8. max_length = maximum value in lengths;
    9. Initialize padded_audios as a zero tensor with shape (batch size,
       max_length);
    10. foreach audio_idx, audio in enumerate(processed_audios) do
        11. Assign audio to
            padded_audios[audio_idx][:lengths[audio_idx]];
    end
    12. padded_audios = move to GPU (if available);
    13. return padded_audios
```

Fig. 4. HuBERT wav-file preprocessing code

그림 4는 HuBERT-Transformer 모델의 전처리 함수 이미지이다. 입력받는 모든 오디오 파일을 MFCC 방식을 활용하여 데이터를 추출하고 음성파일중 가장 긴 샘플의 길이에 맞춰 패딩을 추가하는 음성 패딩 작업을 진행하여 모든 데이터의 길이를 통일 시킨다. 이후 데이터를 GPU로 이동시킨다.

4. Learning

인공지능 모델의 학습 과정은 본 연구의 핵심 단계로, 전처리를 거친 정제된 데이터를 바탕으로 감정을 인식하고 분류한다. 모델은 NVIDIA A100 GPU와 Intel Xeon CPU를 탑재한 서버에서 학습을 수행하였다. PyTorch 프레임워크를 사용하였으며 본 연구에서는 음성 감정 인식에 CNN-Transformer 모델과 HuBERT-Transformer 모델, wav2vector 모델의 세 가지 방식을 사용하여 구현하고 결과를 비교·분석하였다. 세 모델 모두 음성 데이터의 복잡한 특성과 감정의 미묘한 변화를 인식할 수 있지만 각각의 장단점을 가지는 모델이다.

4.1 CNN-Transformer

CNN-Transformer 모델은 음성 데이터의 특징 추출을 위해 CNN을 사용하고, 이후 추출된 특징 간의 관계를 파악하기 위해 Transformer를 적용한 구조이다. CNN은 음성 신호에서 주요 특징을 효과적으로 추출할 수 있는 강력한 도구로 본 논문에서는 MFCC에서 추출한 특징을 입력으로 사용한다. 이후 Transformer 모델은 CNN을 통해 추출된 특징들을 입력받아 특징 사이의 시퀀스들을 학습한다. Transformer 블록 내에서는 Multi-Head Self-Attention 메커니즘을 사용하여 다양한 특성 간의 관계를 폭넓게 탐색한다. 학습 및 검증 과정에서는 크로스 엔트로피 손실 함수와 Adam 최적화 알고리즘을 사용하여 주어진 음성 데이터셋에 대한 감정 분류 정확도를 최대화한다. 학습과정은 Tensorboard를 통해 모니터링되며 이를 통해 학습 진행상황, 손실 및 정확도와 같은 중요 지표들을 시각적으로 분석하였다. 그림 5는 모델의 학습을 위한 기본 설정 코드 이미지이다.

Algorithm 2: Model Training with TensorBoard Logging

```
Input: Features and labels for model training
Output: Trained model with TensorBoard logging
TrainModel(features, labels)
1. model = build_complex_model((1050, 152), 6);
2. model.compile( loss = "sparse_categorical_crossentropy",
   optimizer = Adam(),
   metrics = ["accuracy"]);
3. log_dir = "logs/fit/" + current_datetime();
4. tensorboard_callback = TensorBoard(log_dir=log_dir,
   histogram_freq=1);
5. model.fit( features, labels,
   epochs = 300,
   validation_split = 0.2,
   callbacks = [tensorboard_callback]);
```

Fig. 5. CNN-Transformer Optimizer function code

모델의 입력 shape,와 출력 shape , 모델의 손실함수, 옵티마이저, 평가지표를 설정한다. 이후 TensorBoard의 CallBack 함수를 선언하여 학습 진행시에 로그를 확인 할 수 있도록 한다. 이후 학습횟수, 배치사이즈, 검증 데이터의 비율 등을 선언하여 학습을 진행하였다.

4.2 Hubert-Transformer

HuBERT-Transformer 모델은 음성 신호에서 중요한 특성을 추출하는 HuBERT 모델과 추출된 특성 간의 관계를 학습하기 위한 Transformer 모델을 통합한 구조로 설계되었다. HuBERT-Transformer 모델의 구현은 초기 단계에서 사전 학습된 HuBERT 모델을 활용하여 음성 데이터로부터 필수적인 특성을 추출한다. 추출된 특성은 이후 Transformer 모델로 전달되어 감정 분류를 위한 고차원적인 학습이 이루어진다. Transformer 모델 내에서는 Multi-Head Self-Attention 메커니즘을 사용해 다양한 특성 간의 복잡한 상호작용을 학습한다. 최종적으로, EmotionClassifierPyTorch 모듈을 통해 각 음성 샘플에 대한 감정 분류를 수행한다. 그림 6은 모델의 학습을 위한 CustomDataset 클래스에 대한 코드 이미지이다.

Algorithm 3: Custom Model with HuBERT, Transformer, and Emotion Classification

Input: Input values for model
Output: Logits after emotion classification
Function *Initialize model and optimizer*:
1. model = SimpleTransformerPyTorch();
2. classifier = EmotionClassifierPyTorch();
3. optimizer = `torch.optim.Adam`(*list(model.parameters()) + list(classifier.parameters()), lr=1e-3*);

Function *CustomModel(*`hubert_model()`, `transformer_model()`, `emotion_classifier()`*)*:
1. **Data:** `hubert_model()` processes input values, extract hidden states
;
2. **Data:** `transformer_model()` transforms the extracted hidden states into a feature output
;
3. **Data:** `emotion_classifier()` classifies the transformed output into emotion logits
;

Function *Forward(input_values)*:
1. features = `hubert_model`(*input_values*).last_hidden_state Process input through HuBERT model;
2. transformer_output = `transformer_model`(*features*) Pass features through Transformer model;
3. logits = `emotion_classifier`(*transformer_output*) Classify features into emotion logits;

return logits;

Function *Instantiate model*:
1. custom_model = `CustomModel`(*hubert_model, SimpleTransformerPyTorch(), EmotionClassifierPyTorch()*);

Fig. 6. HuBERT-Transformer Model code

CustomDataset 클래스를 이용해 구성된 데이터셋을 DataLoader를 통해 배치 단위로 모델에 공급하는 방식으로 진행된다. 학습 동안 모델은 크게 두 가지 작업을 수행한다. HuBERT 모델에 의한 특성 추출과 Transformer 모델을 통한 감정 분류 학습. 이 과정에서는 AdamW 최적화 알고리즘과 크로스 엔트로피 손실 함수를 사용하여 모델의 성능을 최적화하였다. 모델의 성능 평가는 학습이 완료된 후 검증 데이터셋을 사용하여 수행된다. 이 때, 감정 분류의 정확도 측정을 통해 모델의 감정 인식 능력을 검증한다. 평가 과정에서 손실과 정확도는 모델의 성능 지표로 사용되며, 이러한 지표들은 TensorBoard를 통해 시각적으로 모니터링되어 모델 학습의 진행 상황을 실시간으로 확인 및 분석하였다.

4.3 Wav2Vec 2.0

페이스북에서 개발한 사전 학습된 모델인 Wav2Vec 2.0 모델을 활용 하여 학습을 진행하였다. 그림 7은 Wav2Vec2 Model을 활용하여 음성 데이터에서 음향 및 음성 특성을 추출하는 Emotion RecognitionModel 클래스를 사용하여 감정을 분류하는 이미지이다.

Algorithm 4: Emotion Recognition Model with Wav2Vec2

Input: Input audio features and labels
Output: Emotion classification logits
Function *EmotionRecognitionModel(num_labels)*:
1. Initialize 2Vec"facebook/wav2vec2-large-960h";
2. Initialize `nn.Dropout`(*0.3*) as dropout1;
3. Initialize `nn.Linear`(*self.wav2vec2.config.hidden_size, 512*) as fc1;
4. Initialize `nn.Dropout`(*0.3*) as dropout2;
5. Initialize `nn.Linear`(*512, 256*) as fc2;
6. Initialize `nn.Linear`(*256, num_labels*) as fc3;

Function *Forward(input_values)*:
1. Extract outputs from Wav2Vec2: outputs = self.wav2vec2(input_values).last_hidden_state;
2. Select first token: output = outputs[:, 0, :];
3. Apply relu and dropout: output = self.dropout1(F.relu(self.fc1(output)));
4. Apply relu and dropout: output = self.dropout2(F.relu(self.fc2(output)));
5. Get logits: logits = self.fc3(output);

return logits;

Function *TrainEmotionRecognitionModel*:
1. Set device = "cuda" if available else "cpu";
2. Print device information for computation;
3. Load train.csv as df;
4. Split data into train_df and val_df (80/20 split);
5. Initialize AudioDataset for train and validation sets;
6. Initialize DataLoader for train and validation sets (batch_size=4);
7. Instantiate EmotionRecognitionModel(num_labels=6) and move to device;
8. Initialize criterion as CrossEntropyLoss;
9. Initialize optimizer as `torch.optim.AdamW`(*model.parameters(), lr=5e-5*);
10. Initialize scheduler as `StepLR`(*optimizer, step_size=5, gamma=0.1*);

Fig. 7. Wav2Vector 2.0 Model Code

이 과정에서 모델은 먼저 Wav2Vec2Model로 부터 추출된 특성을 사용하며, 이후 여러층의 완전연결층(Fully Connected Layer)와 드롭아웃(Dropout)을 거치며 최종적으로 감정 분류를 위한 출력을 생성한다. 본 모델의 학습을 위해 AudioDataset 클래스를 통해 음성 파일과 해당 감정 라벨을 포함하는 데이터셋을 구성하였다. 각 음성 파일은 16000으로 설정한 TARGET_LENGTH에 맞추어 조정되며, 모델 학습에 사용될 통일된 길이의 입력데이터를 제공한다. DataLoader를 통해 배치 단위로 학습데이

터를 모델에 공급하며, 학습과정은 CrossEntropyLoss를 손실함수로, AdamW를 최적화 알고리즘으로 사용하여 진행 하였다. 모델 성능 평가는 앞선 2가지와 마찬가지로 검증 데이터셋을 사용하며 이루어지며 평가지표로는 정확도와 검증 손실이 사용되었다. 각 에포크마다 검증 데이터셋에 대한 모델의 성능을 평가하며, 최적의 모델을 선정하고 저장한다. 모든 지표들은 TensorBoard를 통해 시각적으로 모니터링되어 모델 학습의 진행 상황을 실시간으로 확인 및 분석하였다.

5. Analysis

본 연구의 감정 인식을 위한 모델 비교 분석에서 CNN-Transformer 모델, HuBERT-Transformer 모델, 그리고 Wav2Vec 2.0 모델은 각각의 성능 평가를 검증 데이터셋을 통해 수행하였다. 각 모델의 성능은 정확도를 기준으로 측정되었으며, 이러한 비교 분석은 음성 데이터에서의 감정 인식 능력을 평가하는 데 중요한 지표로 활용될 수 있다. 그림 8은 CNN-Transformer 모델의 성능 결과를 나타낸 그림이다.

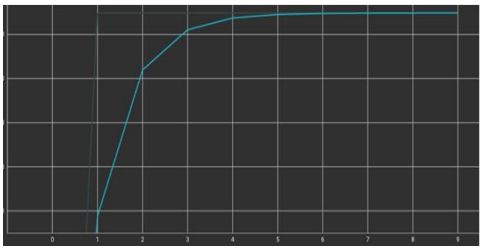


Fig. 8. CNN-Transformer Validation Accuracy Graph

CNN-Transformer 모델은 약 17%의 정확도를 보였다. 이 모델의 낮은 성능은 음성 데이터 내의 복잡한 감성 패턴과 미묘한 변화를 효과적으로 학습하는 데 있어서의 한계로 분석된다. 이는 특히, MFCC를 기반으로 한 특성 추출 방식과 Transformer의 시퀀스 학습 능력이 복잡한 감정 인식 작업에 충분히 학습되지 않았음을 의미한다.

그림 9는 Wav2Vec 2.0 모델의 정확도 결과이다.

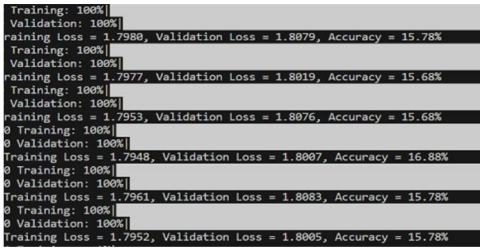


Fig. 9. Wav2Vector 2.0 Model Validation Accuracy

Wav2Vec 2.0 모델은 약 15%의 정확도를 가진다. 사전 학습된 모델을 바탕으로 높은 성능을 기대했으나, 음성 데이터의 감정 인식과 같은 특화된 작업에 대한 추가적인 튜닝 없이는 기대한 성능을 달성하기 어려운 것으로 나타났다. 이 결과는 음성 데이터의 특성을 더 깊이 이해하고, 모델이 이를 효과적으로 학습할 수 있도록 사전 학습된 모델의 재학습 및 세밀한 조정이 필수적임을 알 수 있다.

그림 10은 HuBERT-Transformer 모델의 정확도를 나타낸 그림이다.

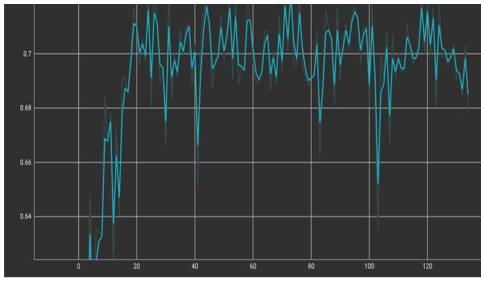


Fig. 10. Hubert-Transformer Validation Accuracy

HuBERT-Transformer 모델은 약 71%의 정확도를 가지며 본 논문에서 구현한 세 가지 모델 중 가장 우수한 성능을 가지는 것으로 나타났다. 이는 HuBERT 모델이 음성 데이터로부터 감정과 관련된 중요한 특성을 효과적으로 추출할 뿐만 아니라, Transformer 모델이 감정의 특성들 사이의 관계를 성공적으로 학습했음을 의미한다. 또한, 이러한 결과는 사전 학습된 모델과 고급 학습 구조의 결합이 음성 데이터에 기반한 감정 인식 분야에서 높은 성능을 달성할 수 있는 효과적인 접근 방식임을 알 수 있었다. Table2는 모델 3가지의 정확도를 비교한 표이다.

Table 2. AI Model Accuracy

AI Model	Accuracy
CNN-Transformer	17%
Wav2Vec 2.0	15%
Hubert-Transformer	71%

결과 분석을 통해 음성 데이터에서 감정을 인식하는 작업에 있어서 모델 선택과 학습 방식이 중요한 역할을 한다는 것을 확인할 수 있었다.

IV. Conclusions

본 연구는 청각 장애인의 의사소통 향상을 목표로 음성 데이터에서 감정을 인식하고 분류하는 인공지능 모델을 설계하고 구현하였다. CNN-Transformer 모델, HuBERT-Transformer 모델, 그리고 Wav2Vec 2.0 모델을 통한 감정 인식 능력의 비교 분석을 수행하였고, HuBERT-Transformer 모델이 상대적으로 높은 정확도와 가장 우수한 성능을 나타내었다.

본 논문에서 제시한 인공지능 모델은 음성 데이터의 복잡한 특성과 감정의 미묘한 변화를 효과적으로 인식하고 분류할 수 있다. 이는 음성 기반 상호작용이 중요한 다양한 분야에서 응용될 수 있음을 보여준다. 특히, HuBERT-Transformer 모델은 사전 학습된 모델과 복잡한 학습 구조의 결합으로 높은 성능을 가지는 모델임을 알 수 있었다.

그러나 본 연구는 몇 가지 한계점을 가지고 있다. 첫째, 다양한 언어와 문화적 배경을 포함하는 데이터셋에 대한 연구가 부족하여, 모델의 일반화 능력에 대한 평가가 제한적이다. 둘째, 멀티 모달 데이터를 활용한 감정 인식에 대한 연구가 이루어지지 않아, 음성 데이터만을 사용할 때보다 더 풍부한 감정 정보를 포착할 수 있는 가능성을 탐색하지 못하였다. 셋째, 실시간 감정 인식 시스템의 개발에 관한 연구가 부족하여, 실제 의사소통 상황에서의 모델 적용 가능성을 충분히 탐구하지 못하였다.

이러한 한계점 극복을 위하여 향후 연구에서는 다양한 언어와 문화적 배경을 포함하는 데이터셋을 활용한 연구를 진행할 예정이다. 또한, 멀티 모달 데이터를 활용하여 감정 인식의 정확도를 더욱 향상시킬 수 있는 방법을 탐색하고, 실시간 감정 인식 시스템 개발에 주력하여 모델의 실제 응용 가능성을 더욱 확장해 나갈 것이다. 이를 통해 음성 데이터에서 감정을 인식하는 기술의 발전만이 아닌, 청각 장애인의 의사소통 향상에 기여할 수 있는 구체적인 방안을 모색할 예정이다.

REFERENCES

[1] S. J. Lee, J. E. Nam, J. Y. Choi, K. H. Kim, E. S. Kim, “Glove-type Sign Language Translator for Communication with the Hearing-impaired Person”, Journal of Rehabilitation Welfare Engineering & Assistive Technology, vol. 17, no. 1, pp. 35-40, 2023. DOI: https://doi.org/10.21288/resko.2023.17.1.35

[2] Eunbyul Jung, Junghwa Bahng, “A Qualitative Study on the Communication Experience of Hearing Impairment Listeners through COVID-19 Pandemic: Using Interpretative Phenomenological Analysis”, Journal of Rehabilitation Research), Vol.27, No.2, 2023. DOI: 10.16884/JRR.2023.27.2.27

[3] Hyun-sam Shin, Joon-ki Hong, “Deep Learning-Based Speech Emotion Recognition Technology Using Voice Feature Filters”, Journal of the Korean Society of Big Data, vol.8, no.2, pp. 223-231, 2023. DOI : 10.36498/kbigdt.2023.8.2.223

[4] Ji-Eun Park, Eun-Ye Kim, Un-Jung Jang, E-Nae Cheong, Young-Ji Eum, Jin-Hun Sohn, “Experiencing and Expression of Deaf Adolescents”, KOSES, vol.19, no.3, pp. 51-58, 2016. DOI: 10.14695/KJSOS.2016.19.3.51

[5] Guiyoung Son, Soonil Kwon, “Spontaneous Speech Emotion Recognition Based On Spectrogram With Convolutional Neural Network”, The Transactions of the Korea Information Processing Society, Vol.13, No.6, pp.284~290, 2024. DOI: https://doi.org/10.3745/TKIPS.2024.13.6.284

[6] Eunji Lee, ASTI MARKET INSIGHT 67: Speech recognition service, 2022.

[7] Semiconductor Network, “Human-machine interaction (HMI), present and future”, https://www.seminet.co.kr/ms_pdf/325_20180509095430_201805_infineon.pdf

[8] Jo Eun-kyoung, “A Study on Dialog Processing for Man-Machine Interaction”, Hangul, Vol. 306, pp. 101-130, 2014.

[9] Soeun Park, Daehee Kim, Soonil Kwon, Neungsoo Park, “Speech Emotion Recognition based on CNN using Spectrogram”, Journal of Information and Control Symposium, Vol.2018 No.10, pp. 240-241, 2018.

[10] Seunghan Ha1, Sangdo Lee, “Feature Extraction and Analysis of AI-Based Speech Signals for Auditory Rehabilitation”, Asia-pacific Journal of Convergent Research Interchange, vol. 9, no.5, pp. 105-116, 2023. DOI: http://dx.doi.org/10.47116/apjcri.2023.05.09

[11] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad, Arpit Bansal, "Feature extraction using MFCC", Signal & Image

Processing: An International Journal, Vol. 4, No. 4, pp. 101-108, 2013.

[12] Abdul, Zrar Kh, Abdulbasit K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review", IEEE Access, Vol. 10, pp. 122136-122158, 2022.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", Advnces in neural information processing systems, 2017.a

[14] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units" , IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, pp. 3451-3460, 2021.

[15] HeeJin Jang, "Diagnosis of Parkinson's disease based on audio voice using wav2vec", Journal of Digital Convergence, Vol. 19. No. 12, pp. 353-358, 2021. DOI: https://doi.org/10.14400/JDC.2021.19.12.353

[16] Seong-Bong Yang, Soo-Jin Lee, "Improved CNN Algorithm for Object Detection in Large Images", Journal of The Korea Society of Computer and Information, Vol. 25, No. 1, pp. 45-53, 2020.

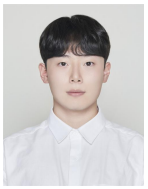
[17] Ji-Seon Park, So-Yeon Kim, Yeo-Chan Yoon, Soo Kyun Kim, "IOptimizing CNN Structure to Improve Accuracy of Artwork Artist Classification", Journal of The Korea Society of Computer and Information, Vol. 28, No. 9, pp. 9-15, 2023.

[18] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", Published as a conference paper at ICLR 2015. arXiv:1409.1556, September 2014.

[19] DACON, https://dacon.io/competitions/official/236105/data

Authors

Seong-Gun Yun entered Multimedia Engineering from Andong National University in 2019. Seong-Gun Yun entered Andong University in 2019 and is majoring in multimedia engineering. He is currently studying machine learning, deep learning, and computer vision.



Hyeok-Chan Kwon entered Multimedia Engineering from Andong National University in 2019. Hyeok-Chan Kwon entered Multimedia Engineering at Andong University, in 2019. He is interested in Machine Learning, Deep Learning and cloud computing.

Eunju Park received the B.S. degree in Computational Statistics from Andong National University in 1993. M.S. degree in Computer Engineering from Andong National University in 2001. Ph. D. degree in Information Communication Engineering from Andong National University in 2016. She is currently teaching professor at Andong National University.

Young-Bok Cho received the M.S., and Ph.D. degrees in Computer Science from Chungbuk National University, Korea, in 2003 and 2012, respectively. also Dr. Cho received more Ph.D degrees in Medical and Law from Chungbuk National University and Chungnam National University , Korea, in 2019 and 2024, respectively. She has Professor of Information Security at Daejeon University, Daejeon, Korea , in 2018 to 2024, She is currently a Professor in the Computer Education at Andong National University, Andong, Korea, in 2024. Her research interests include AI medical image processing, information security and medical information protection, mobile security.