한국빅데이터학회지
제8권 제2호, 2023, pp. 223-231 https://doi.org/10.36498/kbigdt.2023.8.2.223

음성 특징 필터를 이용한 딥러닝 기반 음성 감정 인식 기술

Deep Learning-Based Speech Emotion Recognition Technology Using Voice Feature Filters

신현삼[1] · 홍준기[2†]

한신대학교 정보통신학과[1], 국립공주대학교 스마트정보기술공학과[2]

요 약

본 연구에선 딥러닝 기반 음성 신호로부터 음성의 특징을 추출하고 분석하여 필터를 생성하고, 생성된 필터를 이용하여 음성 신호로부터 감정을 인식하는 모델을 제안하고 감정 인식 정확도 성능을 평가하였다. 제안한 모델을 사용한 시뮬레이션 결과에 따르면, DNN (Deep Neural Network)과 RNN (Recurrent Neural Network)의 평균 감정인식 정확도는 각각 84.59%와 84.52%으로 매우 비슷한 성능을 나타냈다. 하지만 DNN의 시뮬레이션 소요 시간은 RNN보다 약 44.5% 짧은 시뮬레이션 시간으로 감정을 예측할 수 있는 것을 확인하였다.

■ 중심어 : 딥러닝, DNN, RNN, 음성 감정인식

Abstract

In this study, we propose a model that extracts and analyzes features from deep learning-based speech signals, generates filters, and utilizes these filters to recognize emotions in speech signals. We evaluate the performance of emotion recognition accuracy using the proposed model. According to the simulation results using the proposed model, the average emotion recognition accuracy of DNN and RNN was very similar, at 84.59% and 84.52%, respectively. However, we observed that the simulation time for DNN was approximately 44.5% shorter than that of RNN, enabling quicker emotion prediction.

■ Keyword : Deep Learning, DNN, RNN, Speech Emotion Recognition

2023년 11월 27일 접수; 2023년 12월 12일 수정본 접수; 2023년 12월 22일 게재 확정.

† 교신저자 (jkhong@kongju.ac.kr)

Ⅰ. 서론

1.1 연구 배경

감정은 사람의 마음을 표출할 수 있는 가장 중요한 요소이며 인간의 감정은 오랜 시간을 거쳐 심리학자와 신경 과학자들에 의해 인간의 감정과 표현 방식을 이해하고 분류하는 연구를 진행해 왔다[1]. 이러한 연구들은 음성에서의 감정 표현의 복잡성을 이해하기 위한 기반이 되었다. 또한, 음성은 우리가 어떤 감정을 느끼고 있을 때 그것을 대외적으로 나타내는 주요한 수단 중 하나이다. 단순히 언어의 내용만으로는 전달하기 어려운 미묘한 뉘앙스나 감정의 깊이를 음성의 높낮이나 강도를 통해 표현하는 것이 가능하다. 따라서 음성이라는 매체를 통해 감정을 인식하는 연구는 매우 중요하다.

음성인식 기술의 핵심적인 목적은 화자의 의도를 정확하게 해석하고 음성에 내재되어 있는 감정 상태를 정확하게 인식하는 것이다. 이러한 감정 인식을 통해 화자의 음성에 내재되어 있는 감정 상태를 정확하게 인식하여 화자의 명령이나 요청을 더욱 명확하게 이해하여 심도 높은 서비스를 제공할 수 있다. 이처럼 음성 감정 인식 기술은 우리의 일상생활에서 더욱 개인화되고 효율적인 서비스 제공을 가능하게 하는 중요한 기술 중 하나이다[2]. 또한, 최근 헬스케어, 엔터테인먼트, 스마트 홈, 고객 서비스 등 다양한 영역에서 음성 감정 인식 기술은 인간과 기계 간의 음성 기반 상호작용을 통해 더욱 중요한 역할과 기능을 수행하고 있다.

1.2 기존 연구 및 한계점

최근 딥러닝 기술의 발전과 대규모의 음성 데이터 데이터의 확보가 가능해짐에 따라 다양한 딥러닝 기술을 통해 감성을 인식하는 연구가 활발히 진행되고 있다. DNN(Deep Neural Network), RNN(Recurrent Neural Network) 모델과 어텐션 기법을 이용하여 감정 특징을 추출하여 로컬 어텐션과 가중 폴링 방법을 제시하였다[3]. RNN을 이용하여 오토 인코더를 작용하는 모델과 감정 예측에 사용되는 모델을 제시하고 성능을 평가한 연구가 진행되었으며[4], 음성 스펙트로그램 프레임을 DNN을 사용하여 감정을 인식하는 방법을 제안하고 End-to-end 기법을 추가로 활용하는 기술이 제안되었다[5].

LSTM(Long Short Term Memory)의 확장 모델인 Dual-Sequence LSTM 아키텍처 모델을 제시하고 원시 오디오 신호에서 파생된 MFCC(Mel-frequency cepstral coefficients) 특징과 멜-스펙트로그램을 모두 사용하여 음성 감정 인식에 대한 새로운 접근 방식을 제안하였다[6]. 또한, 음성 감정 데이터 세트에서 일관되지 않은 샘플 지속 시간과 불균형한 샘플 범주의 문제를 해결하기 위해 Bi-GRU(Bidirectional-Gated Recurrent Units) 모델을 사용하여 초점 손실을 활용하는 모델을 제안하였다[7].

하지만 기존 연구들에서는 하나의 특징을 추출하고 이를 학습시켜 감정을 예측하는 방식이지만, 높은 정확도를 도출하기 위한 특징의 적합성 및 호환성을 보장하기 어려운 문제가 있다. 따라서 본 논문에서는 복수의 음성 특징을 추출하여 감정 분류의 정확도를 높이기 위한 특수 필터를 제작하고 이를 딥러닝의 학습 과정에 반영하여 정확도를 높이는 기술을 제안하고 성능을 평가하였다.

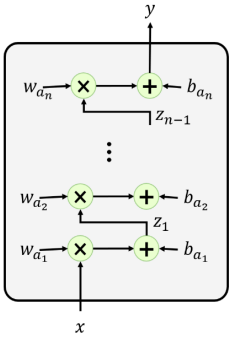
Ⅱ. 관련 연구

2.1 딥러닝 기술

2.1.1 DNN

DNN의 각 은닉층은 여러 개의 뉴런 또는 노드로 구성되어 있으며, 각 뉴런은 입력값을 가

중치와 활성화 함수를 통해 처리하고 출력을 생성한다. DNN은 주로 비선형 문제를 해결하기 위해 사용되며, 다층 구조를 통해 데이터의 추상적인 표현을 학습할 수 있다. 각 은닉층은 이전 은닉층의 결과를 입력으로 받아 다음 층으로 전달하면서 데이터의 특징을 더욱 추상화한다[8].

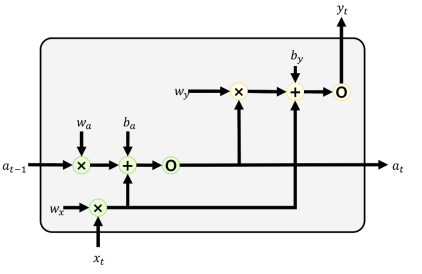


〈그림 1〉 DNN의 구조

여기서 그림 1의 w_{an}과 b_{an}은 각각 은닉층의 가중치와 편향을 나타내며 n은 은닉층의 개수를 의미한다. 각 은닉층의 가중치와 편향은 모두 독립적인 변수로 학습을 통해 업데이트된다. DNN에서 첫 번째 은닉층은 입력 x로 해당 은닉층의 출력인 z를 계산하여 두 번째 은닉층으로 전달한다. 각 은닉층은 이전 은닉층의 출력을 입력으로 사용하며 다음 은닉층으로 출력을 전달한다. 다음 은닉층으로 출력을 전달하기 전에 활성화 함수를 사용하도록 정의할 수 있다. 모든 은닉층을 통과한 마지막 은닉층의 출력은 DNN의 최종 예측 결과를 출력한다.

2.1.2 RNN

RNN은 입력과 출력을 처리하고 방향을 가진 엣지로 히든 노드가 연결되는 순환구조를 이루는 신경망이다[9]. RNN 모델의 구조는 각 시간 스텝에서, 현재의 입력값과 이전 시간 스텝의 은닉 상태를 받아서 새로운 은닉 상태를 계산하고, 이 새로운 은닉 상태는 다음 시간 스텝에 전달되며, 동시에 현재 시간 스텝의 출력을 생성한다.



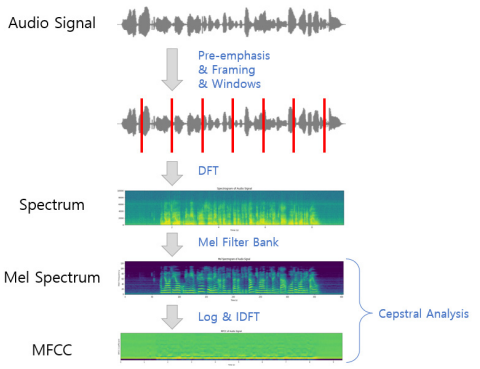
〈그림 2〉 RNN의 구조

그림 2의 x_t와 y_t는 각각 시간 t에서 입력과 출력을 의미하며 w와 b는 각각 가중치와 편향을 나타낸다. a_{t-1}, a_t는 메모리 쉘이며, RNN의 은닉층에서 활성화 함수를 통해 결과를 출력하는 역할을 한다. 메모리 쉘은 일종의 메모리 역할로 이전 시간대의 값을 기억하려고 하는 기능을 수행한다. 시점 t에서의 메모리 쉘 a_t는 이전 시점의 메모리 쉘 a_{t-1}을 활용해 계산하며, g_2은 활성화 함수를 의미한다. 하지만 RNN 구조에는 '장기 의존성 '문제를 갖고 있다. 장기 의존성 문제란 시퀀스가 길어질수록, 네트워크의 가중치를 업데이트하는 그래디언트가 진행함에 따라 시퀀스의 앞부분에 있는 정보를 기억하지 못하는 '그래디언트 소실(vanishing gradient)' 문제가 발생한다. 이러한 그라디언트 소실 문제를 해결하기 위해, LSTM과 GRU와 같은 RNN 기술들이 제안되었다.

2.2 MFCC

MFCC는 음성 신호에서 특징을 추출하기 위해 사용되는 기술로, 음성 신호의 주파수 특징을 캡처하고 음성의 특징을 캡처하고 주파수 음성의 특성을 나타낸다[10]. MFCC는 음성인식,

화자 인식, 음성 합성 등 다양한 음성 처리 작업에 활용한다.



〈그림 3〉 MFCC 특징 추출 과정

그림 3은 MFCC 특징 추출 과정을 나타낸다. 첫 번째 단계로 시간적인 특성을 캡처하기 위해 음성 신호를 작은 프레임으로 나누어 처리한다. 그 다음 각 프레임에 DFT(discrete Fourier transform)를 적용하여 주파수 도메인으로 신호를 변환하여 주파수 성분을 추출한다. 이후 음성은 멜 스케일에서의 주파수에 민감하게 반응하므로, 주파수를 멜 스케일로 변환한 후 일정 간격으로 나눈 멜 필터를 적용하여 주파수 대역에 대한 강도를 계산하기 위해 사용된다. 로그 변환은 주파수 성분을 로그 스케일(log scale)로 변환하여 인간의 청각 시스템의 특성을 모방하여 높은 주파수에서의 변동을 감소시키고, 낮은 주파수에서의 세부 정보를 강조한다. 로그 변환된 스펙트럼에 대해 케프스트럼 계수(Cepstral coefficient)를 추출하고 시간에 따른 변화를 고려하기 위해, 케프스트럼 계수에 대한 델타 및 이중 델타 특징을 계산한다.

MFCC 값은 멜 필터 뱅크 분석을 통해 압축된 파워 스펙트럼 정보에 Cepstral 분석을 한 단계 더 적용해 스펙트럼 포락 성분(계수)만을 추출하는 것이 목적이다. MFCC는 로그 멜 필터 뱅크 특징 값에 이산 코사인 변환을 적용하여 계산한다. 이산 코사인 변환은 다음과 같이 정의된다[10].

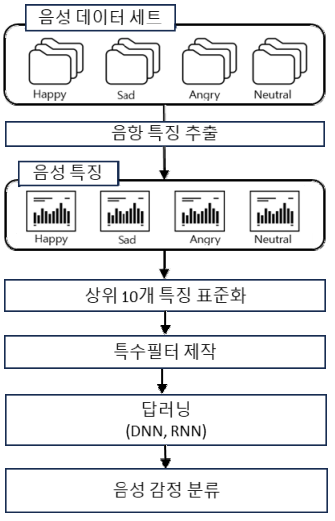
$$\phi_d(l) = \begin{cases} \frac{1}{\sqrt{L}} & (d=0) \\ \sqrt{\frac{2}{L}} \cos\frac{(2l+1)d\pi}{2L} & (d=1,2,\ldots,L-1) \end{cases} \quad (1)$$

$$MFCC(d) = \sum_{l=0}^{L-1} \phi_d(l) \log(b^{mel}(l)) \quad (2)$$

여기서 d는 MFCC의 차원을 의미하며, $\phi_d(l)$는 이산 코사인 변환의 기저 함수이다. $b^{mel}(l)$은 L차원 Mel 필터 뱅크 특징 값을 나타낸다.

Ⅲ. 제안한 특수 필터 기반 음성 감성인식 모델

제안한 특수 필터 방식은 감정별로 파일의 특징을 추출하고 공통된 음향 특징들을 수집하여 특수 필터를 생성하고, 생성된 특수 필터를 사용하여 음성으로부터 감정을 인식하는 방법이다. 아래 그림 4는 제안한 특수 필터 방식 기반의 음성 감성 인식 모델의 순서도를 나타낸다.



〈그림 4〉 음성 특징 인식 모델의 순서도

제안한 특수 필터 방식 기반의 음성 감정 인식 모델은 가장 먼저 주요 감정들의 데이터셋의 음향 특징을 추출하고 해당 음향 특징 중 상위 10개를 감정별로 선정한다. 이후 각 감정별로 선정된 상위 10개의 음성 특징을 평균화하여 10개의 공통 상위 음성 특징을 선정한다. 선정한 10개 음성 특징을 이용하여 감정을 효율적으로 추출하기 위한 특수 필터를 생성한다. 특수 필터를 생성한 이후 딥러닝 기술들을 통해 감정을 분류하고 성능을 비교 분석한다.

3.1 한국 음성 데이터 세트

본 연구에서는 한국지능정보 사회진흥원의 AI-HUB에서 제공하는 '감성 및 발화 스타일별 음성합성 데이터'를 사용하였다. 이 음성 감정 데이터 세트는 다양한 말투, 억양, 감정 표현 등을 포함한 음성 신호를 담고 있으며 화자의 표정, 감정 상태, 감정 종류 등을 레이블로 제공한다.

다양한 감성과 발화 스타일을 포함한 음성 감정 기술 개발을 위한 학습용 음성합성 데이터로 7가지 대표 감정(기쁨, 슬픔, 분노, 불안, 상처, 당황, 중립)과 5가지 발화 스타일, 3가지 발성 캐릭터, 12가지 감정 및 발화 스타일 조합으로 분류된 음성 데이터 세트이다. 본 실험에서는 실험의 효율을 위해 대표적인 7가지 감정을 선정하였다.

〈표 1〉 한국어 데이터 세트의 구성

라벨	의미	데이터 수
Angry	분노	32,000
Happy	기쁨	33,000
Sad	슬픔	32,000
Hurt	상처	32,000
Embarrassed	당황	32,000
Neutral	중립	32,000
Anxious	불안	32,000

표 1은 제안한 감정인식 모델의 정확도를 평가하기 위해 사용된 한국어 감정의 종류와 데이터 수를 나타낸 표이다. 본 연구에서는 제안한 특수 필터 기반 음성 감정 인식 모델을 사용하여 감정 인식 정확도를 계산하기 위해 각 감정 라벨별 32,000개 데이터들을 학습 데이터와 테스트 데이터의 비율을 각각 70%와 30%로 구성하여 시뮬레이션을 진행하였다.

3.2 특징 추출 및 표준화

오디오 음성 파일은 다양한 음성 특징들을 내포하고 있으며 이러한 특징들로부터 감정을 판단하고 분류할 수 있다. 음성 신호는 일반적으로 다음 세 가지 속성이 있으며 제안한 모델의 특징 추출 단계에선 주파수, 진폭, 음속 특징을 추출하였다.

〈표 2〉 오디오 신호의 주요 속성

주요 속성	특징
주파수	주파수는 소리의 높낮이를 나타내는 특성이다. 더 높은 주파수는 높은음을 나타내며, 낮은 주파수는 낮은음을 나타낸다. 주파수는 주기적인 진동의 빈도를 나타낸다. 즉, 초당 진동하는 횟수로 표현된다.
진폭	진폭은 소리의 크기나 강도를 나타내는 특성이다. 진폭이 높으면 소리는 크고 강하게 들린다. 진폭은 소리 파형의 진폭 변화로 표현되며, 음성 신호에서는 음성의 세기나 크기를 나타낸다. 음성 감정 분석에서는 진폭 변화가 음성의 감정 상태를 나타내는 데 활용될 수 있다.
음속	음속은 소리의 전파 속도를 의미한다. 이는 소리가 단위 시간당 이동하는 거리로 정의되며, 일반적으로 음파가 공기를 통해 전파될 때의 속도를 나타낸다.

표 2는 제안한 특수 필터 기반 음성 감성 인식 모델에서 추출한 오디오의 주요 속성과 특징을

나타낸다. 본 연구에선 음성의 감정적 영향력을 측정하기 위해 음성 신호로부터 총 37개의 음성 특징을 추출하였다. 음성 특징 추출은 파이썬의 음성 분석 라이브러리인 librosa를 사용하여 이용하여 음성 파일로부터 음성의 특징을 추출하였다.

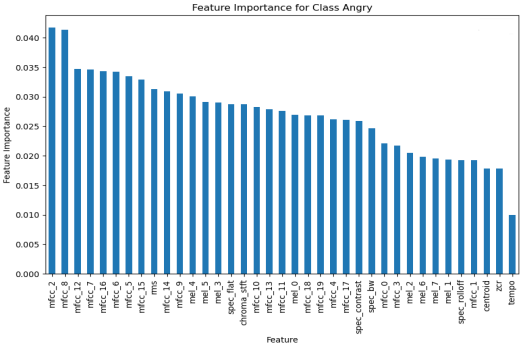
〈표 3〉 librosa로 추출 가능한 음성 특징

37가지 음성 특징	
MFCC (20 features)	Tempo
Mel spectrogram (8 features)	Chroma
Spectral bandwidth	Energy
Spectral contrast	Zero crossing rate
Spectral flat	Centroid
Spectral roll off	

표 3은 특수 필터 제작을 위해 librosa로 추출 가능한 음성 특징을 나타낸다.

3.3 특수 필터 및 딥러닝

제안한 특수 필터는 음성 데이터로부터 추출한 다양한 특징 중 가장 의미 있는 부분들을 감지하여 감정을 분류하는 기능을 수행한다. 또한 특수 필터는 추출된 특징으로부터 최상위 10개의 가중치를 추출하여 제작된다. 특징의 가중치를 통해 감정을 분류하고 예측하기 위한 관련성



〈그림 5〉 분노 음성의 음성 특징 중요도 분석 결과

이 가장 높은 팩터를 알 수 있다. 그림 5는 '분노' 감정의 음성 특징 중요도 분석 결과를 나타낸다.

그림 5에 따르면, 주요 음성 특징으로는 상위 MFCC_2, MFCC_8, MFCC_12, MFCC_7, MFCC_16, MFCC_5, MFCC_14, MFCC_9, Mel_4, MFCC_9 특성이 '분노'의 감정에 가장 많은 영향을 미치는 것을 확인하였다. 이 다섯 가지 특징은 음성 데이터의 복잡한 음향적 특성 중에서 '분노'의 감정을 예측하기 위해 사용된다. 본 연구에서는 음성 신호로부터 감정을 예측하기 위해 각 감정별 상위 10개의 감정에 대한 주요 음성 특징들을 추출하여 특수 필터를 제작하였다. 이러한 특징의 가중치를 기반으로 상위 10개의 특징의 값에 각 가중치를 곱하고 곱한 결과를 모두 추가해서 각 감정별로 감정 지수를 생성한다. 이렇게 생성된 감정 지수를 음성의 특징과 함께 모델의 입력으로 사용한다. 이러한 감정 지수의 계산식은 아래의 수식과 같이 나타낼 수 있다.

$$e_i = \sum_{k=0}^{n} (V_n^f \times W_n^f) \tag{3}$$

여기서 e_i는 감정 지수에 대한 최종 계산 값을 나타낸다. 총 n개 합의 상한으로 양의 정수를 표현한다. V_n^f는 n번째 피처와 관련된 값을 나타낸다. W_n^f는 n번째 피처와 관련된 중요 가중치를 나타낸다. 중요도는 각 특징의 값과 그에 따른 가중치를 합산하여 계산되며, 이를 통해 각 음성 특징의 상대적인 중요성을 확인할 수 있다.

수식 (3)은 감정 특징들의 중요도를 계산하는 과정을 나타내며, 음성 특징의 상대적인 영향을 평가할 수 있다. 감정 특징들의 가중치는 음성 신호의 어떤 특징들이 음성 감정 인식에 영향을 끼치는지를 나타내므로 모델의 의사 결정 과정을 이해하고 해석하는 데 사용할 수 있다. 또한 감정 지수는 감정 카테고리에 맞게 변형되고 음

성에 포함된 각 감정 사이의 미묘한 차이를 표현할 수 있기 때문에 다양한 감정을 구별할 수 있다. 따라서 특수 필터를 통해 음성 데이터로부터 추출한 다양한 특징 중 가장 의미 있는 부분을 감지하고, 이를 통해 감정을 분류하기 위해 사용된다. 따라서 음성으로부터 감정을 더욱 정확하게 구분하고 해석할 수 있으며, 감정 간의 차이를 명료화하기 위해 사용된다. 특징 추출 전처리 과정을 거친 음성 데이터는 딥러닝 모델을 통해 감정이 분류된다.

Ⅳ. 시뮬레이션 결과

본 장에서는 한글과 영어 데이터 세트를 활용하여 총 2가지 딥러닝 모델을 적용하고 각각 모델들의 성능을 비교 분석하였다. 제안한 모델의 감정 분류 정확도 평가를 위한 정밀도는 아래 수식 (4)와 같이 표현할 수 있다.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \quad (4)$$

여기서 TP(True Positive)는 실제 True인 결과를 True로 예측한 결과를 나타내며, FP(False Positive)는 실제 False인 결과를 True로 예측한 결과를 나타낸다. 반면, 재현율은 실제 True인 것 중에서 제안한 모델이 True라고 예측한 것의 비율이며 FN(False Negative)은 실제로 False인 답을 False라고 예측한 값이다.

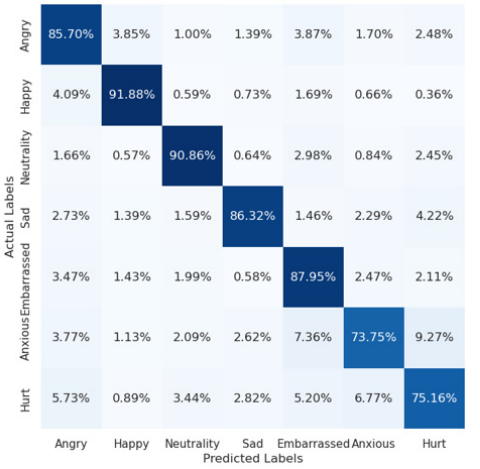
$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \quad (5)$$

그림 6과 그림 7은 제안한 모델에서 DNN과 RNN을 이용하여 감정을 분류했을 때 시뮬레이션 결과를 혼동 행렬로 나타낸 그림이다. 여기서 x축은 감정의 예측 결과를 나타내는 재현율(Recall)을 나타내며, y축은 정밀도(precision)를 나타낸다. 아래 표 4는 감정 카테고리별 DNN과

〈표 4〉 카테고리별 예측 정확도

Model	Angry	Happy	Neutrality	Sad	Embarrassed	Anxious	Hurt	Average
DNN	83.30	93.27	91.90	82.68	86.85	75.98	78.15	84.59
RNN	85.70	91.88	90.86	86.32	87.95	73.75	75.16	84.52

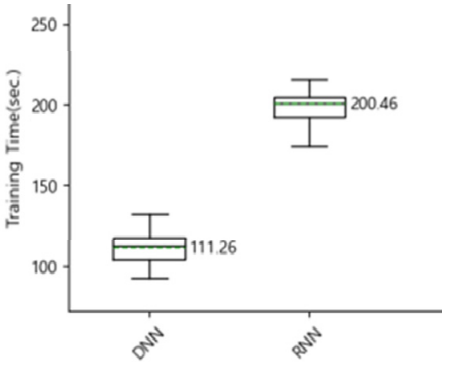
〈그림 6〉 DNN의 혼돈행렬



〈그림 7〉 RNN의 혼돈행렬

RNN의 예측 정확도를 나타낸 표이다.

표 4에 따르면 DNN과 RNN 기술 모두 행복한 음성의 감정 인식 정확도가 가장 높고 불안감성의 인식 정확도가 가장 낮은 것을 확인하였다. 또한 DNN과 RNN 기술의 음성 감정 인식 정확도는 각각 84.59%와 84.52%로 두 기술 모두 매우 비슷한 예측 성능을 나타냈다. 다음 그림 7은 제안한 모델에서 DNN과 RNN 기술들을 사용하여 감정을 인식할 때 소요되는 시뮬레이션 시간을 나타낸 그림이다.



〈그림 8〉 시뮬레이션 소요시간 비교

그림 8에 따르면 DNN과 RNN이 감정 인식을 위해 소요되는 시간은 각각 111.26초와 200.46초로 DNN이 RNN보다 약 44.5% 짧은 시뮬레이션 시간으로 감정을 예측할 수 있는 것을 확인하였다. 따라서 제안한 특수 필터 기반 감정 모델을 사용하여 감정을 예측할 때 DNN이 RNN보다 효율적인 것을 확인하였다. 이와 같은 결과는 DNN이 RNN보다 단순한 구조로 이루어져 있기 때문이다.

Ⅴ. 결론

본 연구에선 음성 데이터로부터 감정 인식을 위해 특수 필터 기반 감정 인식 추출 모델을 제안하고 DNN과 RNN 기술들을 적용하여 감정 인식 정확도와 시뮬레이션 소요 시간을 비교 분석하였다. 제안한 특수 필터는 MFCC를 사용하여 음성 신호로부터 감정을 판단하는 중요 가중치들을 추출하여 생성하였으며, 특징 추출 전처리 과정을 거친 음성 데이터는 DNN과 RNN 모델들을 통해 감정이 분류된다. 제안한 모델의 시뮬레이션 결과에 따르면 DNN과 RNN의 예측 정확도는 각각 84.59%와 84.52%로 매우 유사하지만, DNN의 시뮬레이션 시간은 약 44.5% 적게 소요되는 것을 확인하였다.

또한 추후 제안한 특수 필터 기반 감정 인식 모델은 음성 연속성이 보장되어야 하는 대화형 음성에서 효과적으로 예측할 수 있는 모델로 확장될 예정이다. 제안한 모델의 확장을 위해 특정 구간의 묵음이나 비정상인 특수 음을 배제하여 감정 인식 정확도를 향상시키는 연구를 진행할 예정이다.

참 고 문 헌

[1] Russell, J. A. “A Circumplex Model of Affect Journal of Personality and Social Psychology 39.”, 161-178, 1980.D. Bird, “Direct Marketing Is as Relevant Now as It Was in 1900,” Marketing, pp. 28, 2000.

[2] 김남수(2009), “감정인식 기술의 현황과 전망,” Telecommunications Review 19권 5호, 에스케이텔레콤.

[3] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention,” presented at the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2227–2231.

[4] J. Han, Z. Zhang, F. Ringeval, and B. Schuller,

"Reconstruction-error-based Learning for Continuous Emotion Recognition in Speech," presented at the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 2367–2371.

[5] H. M. Fayek, M. Lech, and L. Cavedon, "Towards Real-time Speech Emotion Recognition using Deep Neural Networks," presented at the 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), IEEE, pp. 1–5, 2015.

[6] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech Emotion Recognition with Dual-Sequence LSTM Architecture," presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6474–6478, 2020.

[7] Z. Zhu, W. Dai, Y. Hu, and J. Li, "Speech Emotion Recognition Model based on Bi-GRU and Focal Loss," Pattern Recognit. Lett., vol. 140, pp. 358–365, 2020.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019.

[9] "Recurrent Neural Networks," Stanford University. Accessed 4 Oct. 23. https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks.

[10] M. Sahidullah and G. Saha, "Design, Analysis and Experimental Evaluation of Block based Transformation in MFCC Computation for Speaker Recognition," Speech Communication, vol. 54, no. 4, pp. 543–565, May 2012.

저 자 소 개

신 현 삼(Shin Hyun Sam)

- 1999년: 동아대학교 대학원 컴퓨터공학과 (공학석사)
- 2007년 2월~현재: (주)퓨렌스 대표이사
- 2021년~현재: 한신대학교 정보통신학과 박사과정

<관심분야> 음성처리, 음성 감정인식 등

홍 준 기(Jun-Ki Hong)

- 2010년 11월: Carleton University 컴퓨터시스템공학과 (공학사)
- 2017년 2월: 연세대학교 전기전자공학과 (공학박사)
- 2016년 8월~2017년 7월: 한국정보통신기술협회(TTA) 선임연구원
- 2017년 8월~2020년 2월: 영산대학교 전기전자공학과 조교수
- 2020년 3월~2023년 2월: 배재대학교 컴퓨터공학과 조교수
- 2023년 3월~현재: 국립공주대학교 스마트정보기술공학과 조교수

<관심분야> 인공지능, 항공체, 차세대통신 등