음성 특징을 이용한 딥러닝 기반 감정 인식

최은빈, 김수형*
전남대학교, *전남대학교
iidmsqlss@gmail.com, *shkim@jnu.ac.kr

Deep Learning-Based Emotion Recognition using Speech Features

Choi Eun Bin, Kim Soo-Hyung*
Chonnam National Univ., Chonnam National Univ.

요 약

본 논문은 음성으로부터 추출되는 특징을 이용하여 딥러닝 기반 감정을 예측하는 모델을 소개한다. 사람의 감정은 얼굴 표정, 음성, 생체신호, 행동 등 여러 가지 형태로 나타나게 된다. 감정을 직관적으로 인식할 수 있는 음성으로부터 여러 가지 음성 특징을 추출하여 감정을 인식하고자 한다. 데이터셋은 RAVDESS Speech와 RAVDESS Song Dataset의 음성 데이터를 결합하여 사용하며 음성 특징(ZCR, Chroma STFT, MFCC, RMS, Tonnetz, Spectral Contrast, Mel Spectrogram)을 추출한다. 이에 따라 본 논문에서 음성 특징을 이용한 딥러닝 기반 감정 인식 알고리즘을 구현하고 각 음성 특징에 따른 결과를 비교한다.

Ⅰ. 서론

감정은 사람의 얼굴 표정, 음성, 생체신호, 뇌파 등 여러 방법으로 표현될 수 있다[1]. 음성은 발화자의 신체적 움직임이나 시각적 장애요소 등에 큰 영향을 받는 얼굴 표정에 비해서 거의 영향을 받지 않으며, 가장 감정을 쉽게 획득할 수 있는 방법이다[2]. 음성의 높낮이와 억양에 감정 정보가 포함되어 있기 때문에 수신자가 발화자의 감정 정보를 획득하고 이해할 수 있다[3]. 따라서, 음성 신호를 분석하여 사용자의 감정을 분류하는 음성 감정 인식(Speech Emotion Recognition; SER)[4]에 대한 활발한 연구가 진행 중이며, 인간과 기계 간의 상호작용에 대해 연구하는 HCI(Human-Computer Interaction)에서 중요한 역할을 한다[5].

음성 기반 감정 인식 시스템은 5가지 단계인 음성 신호 입력, 특징 추출, 특징 선택, 분류, 분류된 감정 추출로 이루어져 있으며, 특징 추출에서는 감정을 잘 나타내는 적절한 음성 특징을 추출해야 한다[1]. 음성신호로부터 추출되는 특징들은 Energy, 템포, 크로마그램, RMS, Tonnetz 등 여러 특징이 있다.

본 논문에서는 RAVDESS Speech[6]와 Song[7]의 음성 데이터셋을 이용하여 여러 가지 음성 특징을 이용한 CNN 기반 감정 인식 알고리즘을 구현하고 각 음성 특징에 따른 성능을 비교한다.

Ⅱ. 음성 특징

1. Zero Crossing Rate(ZCR)

ZCR은 특정 프레임의 지속시간 동안 신호의 부호(Sign) 변화율이며[3], 이는 0을 많이 지나는 비율을 뜻한다. ZCR이 높을수록 음성 안에 잡음이 많다는 것을 의미한다. 따라서, 음성과 잡음의 지배적인 구간을 추정할 수 있다[8].

2. Chroma STFT와 MFCC(Mel-Frequency Cepstral Coefficient)

크로마는 서양 음악의 12개의 음계이며[8], Chroma STFT는 waveform이나 power spectrogram으로부터 생성한 크로마그램이다[10].

MFCC는 음성에서 일정 구간을 나누어 구간에 대한 Mel-Spectrum에서 Cepstral 분석을 통해 고유 특성을 추출한 것이다[3].

3. Root Mean Squere(RMS)와 Tonnetz

RMS는 오디오의 평균 음량을 의미하며, 특정 시간 동안의 평균 음량(신호 제곱의 평균 제곱근)이다[3]. Tonnetz는 tone의 중심을 계산한 것이다[8].

4. Spectral Contrast와 Mel Spectrogram

Spectral Contrast는 각 프레임의 서브 밴드에서 스펙트럼 대비를 계산한 것이며[8], Mel Spectrogram은 주파수의 비선형 변환으로 mel-scale의 에너지 스펙트럼을 계산한 것이다[9].

Ⅲ. 본론

1. 데이터셋

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) 데이터셋은 24명의 전문 성우(남:12명, 여:12명)가 북미 억양으로 두 개의 문장("Kids are talking by the door", "Dogs are sitting by the door")을 각각 말하고 노래할 때의 비디오 및 오디오의 데이터를 제공하는 멀티 모달 데이터셋이다. RAVDESS Speech에는 중립, 차분함, 행복, 슬픔, 분노, 두려움, 역겨움, 놀람의 총 8가지의 감정을 포함하며, RAVDESS Song은 중립, 차분함, 행복, 슬픔, 분노, 두려움의 총 6가지의 감정을 포함한다. 또한, 두 개의 데이터셋에서 감정의 강도는 보통과 강함으로 나누어져 녹음된다[6]. 본 논문에서는 RAVDESS Speech[7]와 RAVDESS Song[8]의 음성 데이터만을 결합하여 사용한다.

2. 데이터 전처리

노이즈, 타임 스트레칭, 피치 변화의 세 가지 방법을 적용하여 데이터를 증강하여 본래의 데이터와 함께 사용한다. 노이즈는 Over-fitting을 피하

기 위하여 white noise를 추가하여 원본 오디오에 잡음을 생성하는 방법이며, 타임 스트레칭은 음성의 피치(Pitch)의 변화 없이 속도만을 바꾸는 방법이다. 증강한 음성 데이터에서 오디오 신호를 분석하는 Python 모듈인 librosa[11]를 이용하여 ZCR, Chroma STFT, MFCC, RMS, Tonnetz, Spectral Contrast, Mel Spectrogram을 추출한다.

3. 시스템 구성 및 결과

본 논문에서는 딥러닝의 한 종류인 CNN(Convolution Neural Network)을 이용하여 모델을 학습시켜 음성의 감정을 인식한다.

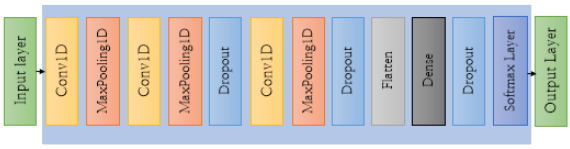

그림 1. 음성 특징 학습을 위한 CNN 구조

모든 Conv1D과 Dense 층에서 Activation Function은 ReLU(Rectified Linear Unit)을 사용하며, 첫 번째와 두 번째의 Dropout은 20%, 마지막 Dropout은 30%를 적용한다. 마지막의 Softmax Layer는 다중 클래스 분류에 사용되는 Softmax Activation Function을 적용한 Dense 층이다. Batch size는 32, Epoch는 100으로 설정하여 학습을 진행한 결과, 단일 특징 및 모든 음성 특징을 사용한 결과는 표 1과 같다.

음성 특징	Accuracy(%)
ZCR	19.41
Chroma_stft	39.31
MFCC	**67.10**
RMS	27.57
Tonnetz	37.14
Spectral-contrast	38.83
Mel-Spectrogram	59.16
All Features	**73.08**

표 1. 음성 특징에 따른 결과

단일 음성 특징을 사용하여 정확도를 계산한 결과, 가장 음성의 특징을 반영한 것은 MFCC가 67.10%로 높았다. 그 뒤는 Mel-Spectrogram이 59.16%로 높았으며 모든 음성 특징을 사용하였을 경우, 73.08%의 정확도를 보여주었다.

데이터셋	Accuracy(%)
RAVDESS Speech(AO)	62
RAVDESS Song(AO)	57

표 2. 기존 연구[10]의 성능

표 2는 기존 연구[10]인 RAVDESS Speech(AO, Audio Only)와 RAVDESS Song(AO)의 정확도를 나타낸 것이다. 모든 음성 특징을 사용한 본 시스템과 비교한 결과, Speech와 Song 데이터셋에서 기존 연구 결과에 비해 각각 약 11%, 16% 향상했다.

Ⅳ. 결론

본 논문에서는 여러 음성 특징을 사용하여 딥러닝 기반 감정 인식 모델을 소개하였다. Over-fitting을 피하기 위해 노이즈, 타임 스트레칭, 피치 변화를 주어 데이터를 증강하였고 ZCR, Chroma_stft, MFCC, RMS, Tonnetz, Spectral-contrast, Mel-Spectrogram의 7가지 특징을 추출하였다. 설계한 CNN 모델을 이용하여 음성 단일 특징에 따른 결과를 비교한 결과, MFCC가 67.10%로 가장 높은 정확도를 보여주었으며 모든 음성 특징을 사용하였을 때는 73.08%의 정확도를 보여주었다. 또한, Speech와 Song 데이터셋에서 기존 연구 결과에 비해 각각 약 11%, 16%의 향상을 보여주었다.

앞으로의 연구 과제는 더 많은 음성 특징을 추출하고 단일이 아닌 여러 가지의 특징 결합을 하여 비교를 할 것이다. 또한, 음성뿐만 아니라 얼굴과 생체신호를 결합하여 감정을 인식하는 연구를 진행하고자 한다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (NRF-2020R1A4A1019191) and also supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1I1A3A04036408).

참 고 문 헌

[1] 김정인, 김진경, and 임완수. "휴먼 인터페이스를 위한 음성 신호 기반 감정인식." 한국통신학회 학술대회논문집 (2022): 1727-1729.

[2] Fahad, Md Shah, et al. "A survey of speech emotion recognition in natural environment." Digital Signal Processing 110 (2021): 102951.

[3] 민동진, 원종호, 강동현, & 김덕환. (2022). 음성언어 감정 인식을 위한 시계열 특징추출 및 다양한 순환신경망 모델의 성능비교. 한국차세대컴퓨팅학회 학술대회, 173-176.

[4] Jae Hun Bang and Sungyoung Lee, 2014, “Call Speech Emotion Recognition for Emotion based Services”, Journal of KISS : Software and Applications, Vol. 41, No.3, pp. 208~213

[5] Khalil, Ruhul Amin, et al. "Speech emotion recognition using deep learning techniques: A review." IEEE Access 7 (2019): 117327-117345.

[6] RAVDESS Emotional Speech Dataset, https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[7] RAVDESS Emotional Song Dataset, https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-song-audio

[8] Lhoest, Lancelot, et al. "MosAIc: a classical machine learning multi-classifier based approach against deep learning classifiers for embedded sound classification." Applied Sciences 11.18 (2021): 8394.

[9] Aslam, Muhammad Ahsan, et al. "Acoustic classification using deep learning." International Journal of Advanced Computer Science and Applications 9.8 (2018).

[10] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13.5 (2018): e0196391.

[11] librosa, https://librosa.org/doc/latest/index.html