CNN - LSTM 모델 기반 음성 감정인식

윤상혁*, 전다윤*, 박능수*
*건국대학교 컴퓨터공학과
sgryoon97@konkuk.ac.kr, jeondayoon27@gmail.com, neungsoo@konkuk.ac.kr

Speech emotion recognition based on CNN - LSTM Model

SangHyeuk Yoon*, Dayun Jeon*, Neungsoo Park*
*Dept. of Computer Science, Konkuk University

요 약

사람은 표정, 음성, 말 등을 통해 감정을 표출한다. 본 논문에서는 화자의 음성데이터만을 사용하여 감정을 분류하는 방법을 제안한다. 멜 스펙트로그램(Mel-Spectrogram)을 이용하여 음성데이터를 시간에 따른 주파수 영역으로 변화한다. 멜 스펙트로그램으로 변환된 데이터를 CNN을 이용하여 특징 벡터화한 후 Bi-Directional LSTM을 이용하여 화자의 발화 시간 동안 변화되는 감정을 분석한다. 마지막으로 완전 연결 네트워크를 통해 전체 감정을 분류한다. 감정은 Anger, Excitement, Fear, Happiness, Sadness, Neutral로, 총 6가지로 분류하였으며 데이터베이스로는 상명대 연구팀에서 구축한 한국어 음성 감정 데이터베이스를 사용하였다. 실험 결과 논문에서 제안한 CNN-LSTM 모델의 정확도는 88.89%로 측정되었다.

1. 서론

기계학습이 발전되면서 다양한 분야에 패턴 인식을 적용하는 사례가 늘어나고 있다. 사람의 감정 분석 또한 예외가 아니다. 컴퓨터 분야에서 인간의 감정을 이해하는 것은 가장 어려운 문제로 남아있지만, 감정을 분류하는 시도는 계속되고 있다. 사람의 감정은 표정, 목소리, 말 등에서 알아낼 수 있다. 본 논문에서는 화자의 음성에서 감정을 분류하는 방법에 관해 연구한다.

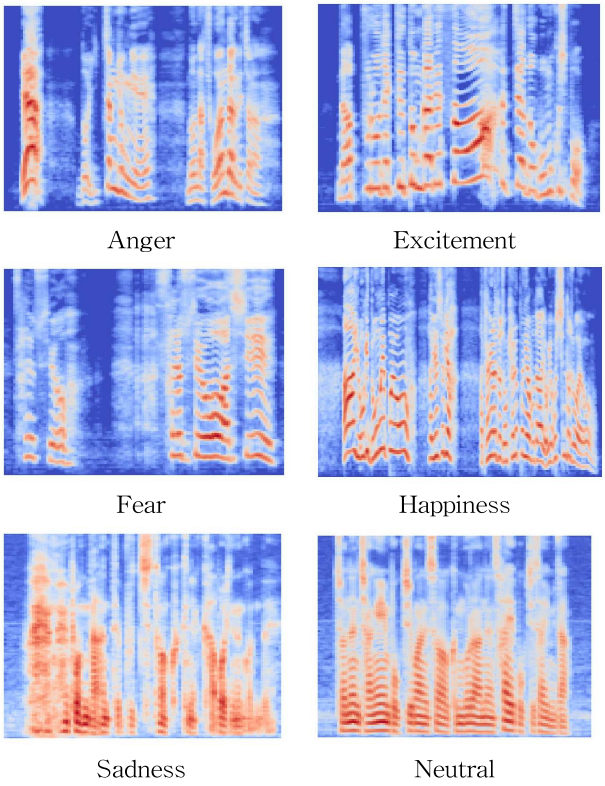
음성에서 감정을 인식하는 방법으로는 화자의 음성만 사용하는 방법과 발화 문장을 같이 사용하는 방법이 있다. 화자의 음성만 사용하여 감정을 분류한 기계학습의 모델 사례로는 CNN을 사용한 모델[1]과 LSTM을 사용하여 음성 감정인식을 연구한 방법[2]이 있다. 본 논문에서는 CNN과 LSTM을 함께 사용하여 화자의 음성데이터로부터 감정을 분류하는 모델을 제안한다.

2. 모델 학습을 위한 음성데이터 전처리

기계학습을 이용하여 오디오 데이터를 학습시킬 때 미가공된 원본 음성데이터를 사용하거나 음성데이터를 주파수 영역으로 변환하여 전처리할 수 있다. 음성데이터를 주파수 영역으로 변환시키는 방법에는 크게 멜 스펙트로그램(Mel-Spectrogram)과 MFCC(Mel-Frequency Cepstral Coefficient)가 있으며 논문에서 제안하는 모델은 멜 스펙트로그램을 전처리 방법으로 사용하였다.

멜 스펙트로그램은 음성데이터를 주파수 영역으로 변환시키기 위해 푸리에 변환(Fourier Transform)을 사용한다. 음성데이터를 작은 조각들로 나누어 각 조각에 푸리에 변환을 적용한 후 결과를 시간순으로 이어 저장한다. 이때 각 조각의 영역은 서로 겹칠 수 있다. 주파수 영역으로 변환이 끝나면 멜 필터(Mel-Filter)를 적용한다. 멜 필터는 주파수 영역을 여러 범위로 나누어 각 범위 안에 해당하는 값들을 합산한다. 이때 합산 방법은 단순 합산이 아닌 0부터 1 사이의 값을 갖는 비선형 함수를 이용해 합산한다. 또한 각 주파수 영역은 서로 겹칠 수 있다. 멜 필터를 적용하고 나면 결괏값을 스펙트로그램(Spectrogram)으로 변환한다. 스펙트로그램은 x축은 시간, y축은 주파수 영역으로 시간에 따른 주파수 영역 값의 변화를 기록한다. 각 시간 축은 0.128초에 해당하는 길이의 음성데이터를 주파수 영역으로 변화한 값이며 각 영역은 0.032초씩 겹치도록 하였다. 멜 필터 크기는 128로 정하여 각 시간 축의 특징값

개수는 128개이다. 그림 1은 학습에 사용한 음성데이터의 감정별 멜 스펙트로그램이다.

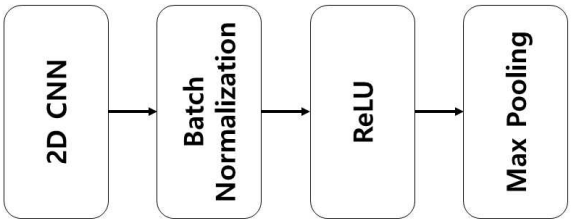


(그림 1) 감정별 멜 스펙트로그램

3. 음성 감정 인식 모델

멜 스펙트로그램은 시간에 따른 주파수 영역의 변화를 기록하는 2차원 데이터이다. 따라서 2차원 데이터의 특징값을 학습하기 위해 CNN(Convolutional Neural Network)을 사용하였다. 2차원 CNN을 사용하면 고정된 시간 축에서 인접한 주파수 영역의 특징을 학습할 수 있을 뿐만 아니라 고정된 주파수 축에서 시간에 따른 특징을 학습할 수 있다. 논문에서 제안하는 음성 감정 인식 모델은 3개의 CNN 모듈을 사용하며 각 모듈은 2D-CNN, Batch Normalization, ReLU(Rectified Linear Unit), Max Pooling 레이어 순으로 구성되어있다. 그림 2는 CNN 모듈에 관한 순서도이다.

감정은 한 발화에서 항상 일정하기보다는 시간에 따라 변화할 수 있다. 한 발화에서 특정 시간의 감정 정보를 지역 특징(Local Feature)이라 하며 발화 전체의 대표되는 감정 정보를 전역 특징(Global Feature)이라 한다. 하나의 음성데이터에서 하나의 감정을 유추하기 위해 지역 특징의 변화를 통한 전역 특징의 추론이 필요하다. 이를 위해 LSTM(Long Short-Term Memory) 네트워크를 사용하였다. 앞선 CNN 네트워크에서 시간에 따른 감정 정보를 압축된 특징 벡터로 전달하면 LSTM 네트워크에서 최종 감정 정보를 추론하게 된다. LSTM 네트워크는 양방향으로 구성된 Bi-Directional LSTM을 사용하였고 LSTM에서 출력한 특징 벡터를 감정 정보로 분류하기 위해 2개의 완전 연결 네트워크(Fully Connected Network)를 사용하였다. 표 1은 제안한 모델에서 사용한 모든 네트워크 정보를 정리한 표이다.



(그림 2) CNN 네트워크

<표 1> 모델 정보

타입	커널 크기	출력
2D CNN	(3, 3)	(256, 128, 64)
Batch Normalization	-	(256, 128, 64)
ReLU	-	(256, 128, 64)
Max Pooling	(2, 2)	(128, 64, 64)
2D CNN	(3, 3)	(128, 64, 64)
Batch Normalization	-	(128, 64, 64)
ReLU	-	(128, 64, 64)
Max Pooling	(4, 4)	(32, 16, 64)
2D CNN	(4, 4)	(32, 16, 256)
Batch Normalization	-	(32, 16, 256)
ReLU	-	(32, 16, 256)
Max Pooling	(4, 4)	(8, 4, 256)
Unfolding	-	(8, 1024)
Bi-Directional LSTM	(256)	(512)
Dense	(128)	(128)
Dropout	-	(128)
Dense	(6)	(6)

4. 한국어 음성 감정 데이터베이스

한국어 감정 음성 데이터베이스로는 상명대 연구팀에서 구축한 데이터베이스를 사용하였다[3][4]. Anger, Excitement, Fear, Happiness, Sadness, Neutral로 총 6가지 감정에 대해 구축된 데이터베이스로 총 20명의 전문 연기자가 감정별로 20개의 문장을 연기하였다. 이 중 감정당 150개씩, 총 900개의

데이터를 선별하였다.

5. 실험

학습, 검증, 테스트를 위해 900개의 데이터를 각각 720개, 90개, 90개로 나누어 학습을 진행하였다. 모든 데이터를 멜 스펙트로그램으로 변환하여 사용하였다. 모델에는 입력하는 멜 스펙트로그램의 길이는 256으로 정하였다. 모든 음성데이터의 길이가 다르므로 멜 스펙트로그램의 길이 또한 모두 다르다. 만약 길이가 256보다 짧다면 부족한 부분을 0으로 덧붙였으며 길다면 멜 스펙트로그램의 정중앙 부분만 사용하였다.

학습 데이터에 과적합을 막기 위해 학습 데이터와 검증 데이터의 손실 함숫값의 차이가 벌어지기 전까지 진행하였으며 그 중 검증 데이터의 정확도가 가장 높은 단계의 모델을 사용하였다. 실험 결과 테스트 데이터의 음성 감정 분류 정확도는 88.89%로 측정되었다. 표 2는 감정별 추론에 따른 분류 평가표이다. 실험 결과 Excitement, Happiness, Sadness, Neutral은 거의 완벽하게 분류하는 것을 볼 수 있지만, Anger 15개의 데이터 중 Excitement로 2개와 Happiness 2개로, Excitement는 15개의 데이터 중 2개는 Anger로 잘못 분류하는 때도 있음을 확인할 수 있다.

<표 2> 감정별 추론 평가표

	A	E	F	H	S	N
A	11	2	0	2	0	0
E	2	13	0	0	0	0
F	0	1	12	2	0	0
H	0	0	1	14	0	0
S	0	0	0	0	15	0
N	0	0	0	0	0	15

6. 결론

음성데이터로부터 감정을 추론하기 위하여 CNN-LSTM을 이용한 음성 감정 인식 모델을 제안하였다. 음성데이터를 멜 스펙트로그램으로 변환하여 시간별 주파수 영역의 변화를 기록하고 2D CNN을 이용하여 특징 벡터를 추출한다. 시간 흐름에 따라 특징 벡터를 추출하면 LSTM을 이용하여 감정의 변화를 학습하여 음성데이터 전체의 감정 특징 벡터를 추출한다. 마지막으로 완전 연결 네트워크를 통해 6가지 감정 중 하나로 감정 정보를 분류한다.

상명대 연구팀에서 구축한 한국어 음성 감정 데이터베이스를 이용하여 모델 학습 및 평가 결과 논문에서 제안한 모델의 정확도는 88.89%로 측정되었다.

감사의 글

본 논문은 2021년도 해양경찰 현장맞춤형 연구개발 사업(오션랩)의 재원(No.20016379)으로 수행된 결과임

참고문헌

[1] 박소은, 김대희, 권순일, 박능수."Spectrogram을 이용한 CNN 기반 음성 감정인식."정보 및 제어 논문집.(2018):240-241.

[2] 변성우, 이석필. "한국어 기반의 감정인식을 위한 연구 : 데이터 베이스 수집, 특징 벡터 선택, 인식 모델 설계." 대한전기학회 학술대회 논문집 . (2019): 1784-1785.

[3] 손희수, 변성우, 신보라, 이석필. "미디어 콘텐츠를 활용한 한국어 감정 음성 자료 구축." 정보 및 제어 논문집 . (2017): 138-139.

[4] 김주희, 오다경, 손명진, 이예진, 손희수, 변성우, 이석필. "한국어 감정 음성 DB 구축." 대한전기학회 학술대회 논문집 . (2019): 9-11.