심층 신경망 기반 감정 인식을 위한 스파이크 특성 추출 기술

*안순호 김재원 한석현 신성현 박호종
광운대학교
*knowledgeiin@naver.com

Spike Feature Extraction for Emotion Recognition based on Deep Neural Network

*An, Soonho Kim, Jaewon Han, Seokhyeon Shin, Seonghyeon Park, Hochong
Kwangwoon University

요약

본 논문에서는 심층 신경망을 기반으로 하는 감정 인식을 위해 스파이크 특성을 추출하는 기술을 제안한다. 기존의 심층 신경망을 이용한 감정 인식 기술은 대부분 MFCC를 특성 벡터로 사용한다. 그러나 프레임 단위의 연산인 MFCC는 높은 시간 해상도를 확보하기 어려워 시간적 특성의 영향을 받는 감정 인식에 한계가 있다. 이를 해결하기 위해 본 논문에서는 인간의 청각 필터를 모델링한 ERB에 따라 샘플 단위로 주파수의 특성을 나타내는 스파이크그램을 이용한 감정 인식 기술을 제안한다. 제안하는 방법이 감정 인식의 대표적 특성인 MFCC보다 높은 인식률을 제공하는 것을 확인하였다.

1. 서론

최근 사람의 음성을 이용한 감정 인식 기술이 활발하게 연구되고 있다. 이러한 연구는 심층 신경망을 중심으로 빠르게 발전하고 있으며 많은 연구에서 Mel-frequency cepstral coefficient (MFCC)를 핵심 입력 특성 벡터로 사용한다[1]. 그러나 MFCC 연산은 프레임 단위로 이루어지기 때문에 높은 시간 해상도를 확보하기 어렵다는 문제점이 있다.

기존의 감정 인식 기술에서 주로 사용하는 MFCC는 푸리에 변환을 기반으로 하고 있으며 푸리에 변환은 무한한 길이의 사인파와 코사인 파형으로 음성 신호를 분석하는 방법이다. 그러나 무한 길이의 파형을 사용한 분석은 정밀한 주파수 해상도를 취할 수 있으나 프레임 내의 시간 해상도를 포기해야 한다는 단점이 있다. 이러한 특성으로는 주파수 특성과 시간적 특성이 모두 중요한 인간의 감정을 분석하는 데에 한계가 있으며, 시간적 특성을 보완할 수 있는 새로운 특성 벡터가 필요하다.

본 논문은 인간의 청각 구조를 모델링한 스파이크그램으로부터 특성을 추출하여 이를 심층 신경망으로 학습하는 방식의 감정 인식 기술을 제안한다[2]. 스파이크그램은 샘플 단위로 주파수의 특성을 나타내기 때문에 더 정밀한 시간 해상도를 확보할 수 있으며, 주파수 특성 또한 효율적으로 표현할 수 있다[2, 3]. 성능 평가에는 Berlin Database of Emotional Speech (Emo-DB)[4]를 사용하였으며 스파이크그램을 이용하여 MFCC를 이용한 감정 인식보다 높은 인식률을 기록하였다.

2. 제안하는 방법

2.1 스파이크그램 생성

본 논문은 인간의 청각 알고리즘과 유사한 스파이크그램을 기반으로 감정 인식을 위한 특성을 추출하는 방법을 제안한다. 스파이크그램은 스펙트로그램과 유사하게 음성을 시간-주파수 그래프에 표현하는 기법이다. 이를 위해 본 논문에서는 음성을 감마톤 필터뱅크의 합으로 분해하였으며, 감마톤 파형은 식 (1)과 같다.

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft) \quad (1)$$

이때, f 는 중간 주파수, a 는 진폭, n 은 필터의 차수, b 는 필터의 대역폭이다. 감마톤 필터뱅크는 인간의 청각 필터를 모델링한 Equivalent Rectangular Bandwidth (ERB)에 맞게 32 밴드로 나눠서 생성하였다[5]. 그림 1은 32 밴드 감마톤 필터뱅크의 주파수 응답을 보여준다.

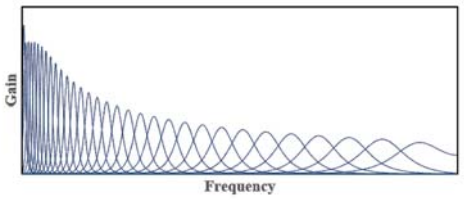

그림 1. 32 밴드 감마톤 필터뱅크의 주파수 응답
Fig. 1. Frequency response of 32 band gammatone filterbank

이와 같이 생성한 32 밴드 감마톤 필터뱅크로 음성 신호를 분해하여 각각의 감마톤 필터뱅크를 시간-주파수 그래프인 스파이크그램으로 표현하였다. 그림 2는 음성 신호의 스펙트로그램과 스파이크그램의 예시이다.

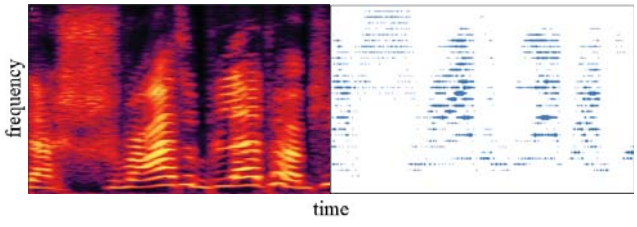

그림 2. 음성 신호의 스펙트로그램 (왼쪽)과, 스파이크그램 예시 (오른쪽)
Fig. 2. Example of spectrogram (left) and example of spikegram (right) of speech signal

2.2 특성 벡터 추출

생성한 스파이크그램 행렬 (***S***)로부터 크게 세 종류의 특성 벡터를 추출하였으며 그 과정은 그림 3과 같다. 먼저 각 밴드별로 약 100ms 길이의 프레임 마다 스파이크가 활성화되는 횟수 (***C***)와 스파이크의 에너지 (***G***)를 계산한다. 다음, 약 800ms 동안 8개의 frame에 대한 각 값의 평균과 표준편차를 계산하였다. 그리고 각 밴드별로 약 800ms 마다 스파이크가 활성화되는 간격의 평균과 표준편차 (***P***)를 계산하였으며 이를 (***C***), (***G***)와 연결하여 192-D 특성 벡터(***X***)를 추출하였다.

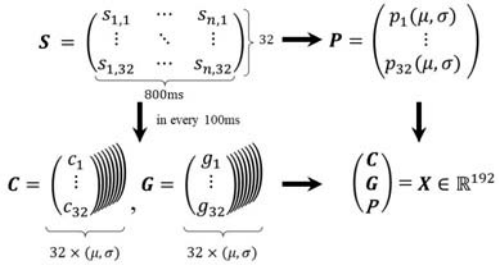

그림 3. 특성 벡터 ***X***의 추출 과정
Fig. 3. Extraction of feature vector ***X***

2.3 네트워크 구조

본 논문에서는 추출한 특성 벡터를 이용한 감정 인식을 위해 deep neural network (DNN)을 사용한다. 본 논문에서 사용한 DNN은 5개의 층으로 구성되어 있으며 각 층의 뉴런 개수는 [192, 600, 300, 100, 7]이다. DNN의 hyper-parameter는 실험을 통해 설정하였다. 은닉층의 활성화 함수는 rectified linear unit (ReLU)이며 출력층의 활성화 함수는 softmax function이다. Weight와 bias의 초기화에는 He 정규화를 사용하였고[6] optimizer는 Adam optimizer이며 dropout rate는 0.8로 설정하였다[7].

3. 성능 평가

성능 평가에는 Emo-DB를 사용하였다[4]. Emo-DB는 anger, boredom, disgust, fear, happiness, neutral version, sadness로 총 7개의 감정으로 구성되며 감정별로 각 [127, 81, 46, 69, 71, 79, 62] 개의 파일로 이루어져 있다. 각 파일의 길이는 1s에서 6s이며 각 파일은 약 800ms 단위로 추출된 1개에서 7개의 특성 벡터를 갖는다. 네트워크는 특성 벡터마다 해당 감정에 대한 확률을 출력하므로 이를 파일 단위로 더하여 가장 높은 확률을 갖는 감정으로 판정한다.

표 1은 밴드 수를 다르게 적용한 3가지 MFCC와 제안하는 방법의 감정 인식 기술 정확도이다. MFCC의 밴드 수는 32, 64, 96 밴드로 설정하였으며 약 30ms마다 계산하여 약 800ms 동안의 평균과 표준편차를 특성 벡터로 사용하였다. 정확도는 동일한 DNN을 기반으로 확인하였다. 제안하는 방법의 정확도가 가장 높았으며, MFCC 중 가장 높은 정확도를 보이는 MFCC_2 보다 2.9 %p 더 높다는 것을 알 수 있다.

표 1. MFCC와 제안하는 방법의 감정 인식 정확도
Table 1. Accuracy of emotion recognition for MFCC and proposed method

Feature	Band	Dimension	Accuracy (%)
MFCC_1	32	64	77.8
MFCC_2	64	128	78.2
MFCC_3	96	192	74.6
Proposed method	**32**	**192**	**81.1**

표 2는 제안하는 방법의 감정 인식 정확도에 대한 혼동 행렬이다. 3% 이하는 -로 표시하였으며 감정 인식 평균 정확도는 81.1%이다.

표 2. 제안하는 방법의 혼동 행렬
Table 2. Confusion matrix of proposed method

Predicted / True	An	Bo	Di	Fe	Ha	Ne	Sa	Recall(%)
Anger	90.6	-	-	-	6.3	-	-	90.6
Boredom	-	79.0	-	-	-	13.6	6.2	79.0
Disgust	-	-	84.8	6.5	-	-	-	84.8
Fear	4.3	-	-	73.9	10.1	-	4.3	73.9
Happiness	16.9	-	-	5.6	70.4	4.2	-	70.4
Neutral	-	13.9	-	-	-	78.5	5.1	78.5
Sadness	-	3.2	-	-	-	3.2	90.3	90.3
Precision(%)	81.0	79.2	88.1	82.8	78.0	75.0	83.6	81.1

4. 결론

본 논문은 스파이크그램을 이용한 심층 신경망 기반 감정 인식 기술을 제안하였다. 인간의 청각 필터를 모델링한 ERB 기반의 감마톤 필터뱅크를 사용해 정밀한 시간 해상도를 갖는 스파이크그램을 생성하였으며 이로부터 특성 벡터를 만들어 심층 신경망의 입력으로 사용하였다. 제안하는 방법을 사용한 감정 인식 기술이 MFCC를 사용한 감정 인식 기술보다 높은 성능을 보이는 것을 확인할 수 있다.

감사의 글

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016R1D1A1B03930923).

참고문헌

[1] M.S. Likitha, R. Gupta, K. Hasitha, A. Raju, "Speech Based Human Emotion Recognition Using MFCC," *IEEE International Conference (WiSPNET)*, 2017.

[2] EC. Smith, MS. Lewicki, "Efficient Auditory Coding," *Nature 439*, pp. 978-982, Feb. 2016.

[3] S.-H. Shin, H.-W. Yun, W.-J. Jang and H. Park, "Extraction of acoustic features based on auditory spike code and its application to music genre classification," *IET Signal Processing*, vol.13, no. 2, pp. 230-234, Apr. 2019.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A Database of German Emotional Speech," *INTERSPEECH* 2005.

[5] M. Slaney, "An Efficient Implementation of the Patterson - Holdsworth Auditory Filter Bank," *Apple Computer Technical Report* #35, 1993.

[6] K. He, X. Zhang, S. Ren, J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *ICCV*, pp. 1026-1034, 2015.

[7] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, The MIT Press, Cambridge and London, 2016.