심층 신경망 기반의 사운드 분류를 위한 청각 특성 추출 기술

*장우진 신성현 윤호원 조효진 장원 박호종

광운대학교

*bbang0719@kw.ac.kr

Auditory Feature Extraction for Sound Classification based on Deep Neural Network

*Jang, Woo-Jin Shin, Seong-Hyeon Yun, Ho-Won Cho, Hyo-Jin Jang, Won Park, Ho-chong

Kwangwoon University

요약

본 논문에서는 심층 신경망 기반의 사운드 분류를 위한 청각 특성 추출 기술을 제안한다. 심층 신경망은 인간의 신경망을 모델링 하기 때문에 인간의 인식을 기반으로 하는 특성을 사용한다면 더 적합한 학습을 할 수 있다. 기존 방법인 MFCC와 스펙트로그램과는 달리 스파이크그램은 인간의 청각 시스템을 기반으로 파형을 해석하는 방법이기 때문에 심층 신경망에 더 효율적인 특성이라고 할 수 있다. 따라서 본 논문에서는 사운드 분류 기술의 특성으로 스파이크그램을 이용하는 방법을 제안한다. 제안한 방법을 사용하면 MFCC와 스펙트로그램을 사용하는 것보다 더 높은 분류 성능을 얻을 수 있다.

1. 서론

최근 심층 신경망을 이용한 딥 러닝 기술에 대한 관심이 높아지면서, 기존에 쓰이던 특성에 비해 심층 신경망에 더 적합한 특성을 찾으려는 다양한 연구가 진행되고 있다.

스파이크그램은 인간의 청각 시스템을 기반으로 파형을 해석하는 방법으로, 신호가 도달하면 특정 주파수를 인식하는 청세포가 반응하는 것을 이용한다. 즉 인간은 음향을 인지할 때 전체가 아닌 의미 있는 특정 주파수 대역들의 가중 합으로 받아들이게 되고, 스파이크그램은 이러한 스파스 (sparse)한 성질을 이용하여 인코딩한다[1, 2]. 또한, 심층 신경망은 인간의 신경망을 모델링 한 학습 방법으로써 그 특성으로 스파이크그램을 사용하는 것은 매우 자연스러운 접근이다.

본 논문에서는 스파이크그램을 이용하여 사운드의 특성을 추출하는 방법을 제안한다. 제안하는 방법은 스파이크 (spikes)를 추출하고, 이를 이용하여 사운드 분류에 적합한 새로운 특성을 얻는다. 제안하는 방법을 사용하면 기존의 MFCC와 스펙트로그램을 특성으로 사용하는 것보다 더 높은 인식률을 얻을 수 있다.

2. 제안하는 방법

기존의 사운드 분류에 널리 쓰이는 특성은 MFCC와 스펙트로그램이며, 이는 모두 주파수 영역으로부터 추출할 수 있다. 본 논문에서는 시간 영역으로부터 사운드 분류에 더 적합한 특성을 얻는 방법을 제안한다.

먼저 원본에서 1초 단위로 스파이크를 일정한 개수로 추출한다. 스파이크를 추출하는 것은 음원 1초 이내에 가장 큰 연관성 (correlation)을 갖는 커널 (kernel)의 정보를 저장하는 것이다. 각 스파이크당 커널의 종류와 위치, 그리고 게인 (gain), 이렇게 3가지 정보를 갖는다. 이 과정을 위해 MP (Matching Pursuit) 알고리즘을 적용하며, 추출한 스파이크를 이용하여 식 (1)과 같이 음원을 표현할 수 있다[3].

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} g_i^m \phi_m (t - \tau_i^m) + \epsilon(t) \quad (1)$$

위 식에서 M은 커널 (kernels)의 수, n은 커널별 스파이크의 출현 횟수, g는 각 스파이크의 게인, ϕ는 커널의 파형, τ는 각 스파이크의 위치를 의미한다. 이 과정에서 사용하는 커널로는 그림 1과 같은 주파수 특성을 갖는 감마톤 필터 (Gammatone filters)를 사용하였고, M은 64로 설정하였다. 그림 2는 추출한 스파이크로 스파이크그램을 만들고 이를 스펙트로그램과 비교한 것이다.

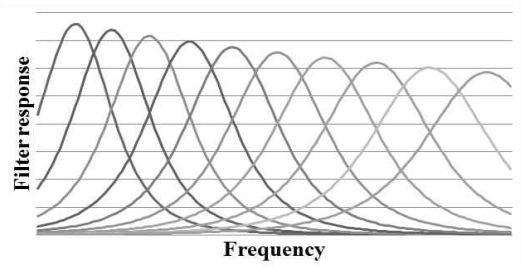

그림 1. 감마톤 필터

Fig. 1. Gammatone filters

추출한 스파이크를 이용하여 제안하는 특성을 구할 수 있다. 먼저 커널별 스파이크 출현 횟수 n과 식 (2)의 커널별 게인의 합 s를 이용하여 128차 벡터를 구한다. 여기에 스파이크를 이용하여 식 (3)과 같이 복원한 신호 $x'(t)$의 SNR (signal to noise ratio)을 추가하여 최종적으로 129차 특성 벡터를 얻는다.

$$s_m = \sum_{i=1}^{n_m} g_i^m, (m = 1, \cdots, 64) \quad (2)$$

$$x'(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} g_i^m \phi_m (t - \tau_i^m) \quad (3)$$

이 특성 벡터를 이용하여 추출하는 스파이크의 개수를 실험적으로 최적화한다. 본 논문에서는 3,000개로 설정하였다.

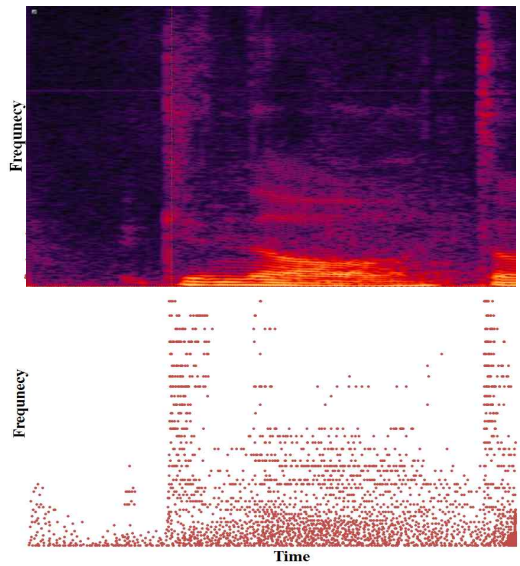

그림 2. 위: 스펙트로그램, 아래: 스파이크그램

Fig. 2. Top: spectrogram, Bottom: spikegram

3. 성능 평가

성능 평가에는 총 5개의 층을 가지는 인공 신경망 구조를 이용하여 딥 러닝을 수행한다. 각 층을 구성하는 뉴런 수는, 입력층은 129개, 출력층은 3개이며 나머지 3개의 은닉층은 각각 300, 60, 30개이다. mini-batch 크기는 1이고 학습 반복 횟수 (epoch)는 2,000이다. 각 층의 가중치는 입력 벡터의 크기에 따른 가우시안 랜덤 변수로 초기화한다.

사용한 사운드 데이터의 종류는 3가지이며 music, speech, effect로 구성된다. 이들은 뉴스, 다큐멘터리, 음악프로그램 등의 TV 방송에서 획득한 음원이다. 음원의 샘플링 주파수는 22.05kHz이고, 길이는 각각 32분이다. 그중 종류별로 10%를 무작위로 선택하여 실험데이터로 사용하고, 나머지 90%는 인공 신경망의 학습에 사용된다.

표 1은 기존 방법인 스펙트로그램을 이용한 심층 신경망 기반의 사운드 분류 성능이고[4], 표 2는 제안하는 방법인 스파이크그램을 이용한 심층 신경망 기반의 사운드 분류 성능이다.

표 1. 스펙트로그램을 이용한 사운드 분류 성능(%)

Table 1. Performance of sound classification using spectrogram(%)

True \ Estimated	Music	Speech	Effect	Average
Music	**97.40**	1.04	1.56	**95.66**
Speech	1.56	**94.27**	4.17	
Effect	0.52	4.17	**95.31**	

표 2. 스파이크그램을 이용한 사운드 분류 성능(%)

Table 2. Performance of sound classification using spikegram(%)

True \ Estimated	Music	Speech	Effect	Average
Music	**97.40**	2.60	0	**97.92**
Speech	0.52	**97.92**	1.56	
Effect	1.04	0.52	**98.44**	

표 3은 기존 방법과 제안하는 방법의 성능 차이를 나타낸다. 제안하는 방법의 분류 성능은 기존 방법보다 speech는 3.65%p, effect는 3.13%p 높아져 평균적으로 약 2.26%p 향상되었다.

표 3. 스펙트로그램 특성과 스파이크그램 특성의 성능 차이(%p)

Table 3. Performance difference between spectrogram and spikegram(%p)

	Music	Speech	Effect	Average
Difference	0	3.65	3.13	2.26

4. 결론

본 논문에서는 심층 신경망 기반의 사운드 분류를 위한 청각 특성 추출 기술을 제안하였다. 제안한 방법은 스파이크를 추출하고, 커널별 출현 횟수와 게인 합, 그리고 SNR을 이용하여 특성 벡터를 구한다. 이처럼 구한 특성 벡터를 이용하여 딥 러닝을 수행한다.

제안한 방법을 사용하면 심층 신경망을 이용한 딥 러닝에 더 적합한 특성을 얻을 수 있다. 그에 따라 기존의 스펙트로그램을 특성으로 사용하는 방법보다 높은 사운드 분류 성능을 얻을 수 있다.

감사의 글

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

참고문헌

[1] M. Lewicki, "Efficient Coding of Natural Sounds," Nature Neurosci. vol. 5, pp. 356-363, Mar. 2002.

[2] E. Smith and M. Lewicki, "Efficient Auditory Coding," Nature vol. 439, pp. 978-982, Feb. 2006.

[3] P. Manzagol, T. Bertin-Mahieux and D. Eck, "On The Use of Sparse Time-Relative Auditory Codes for Music," Proc. of Int. Soc. Music Inf. Retrieval Conf. (ISMIR), Sep. 2008.

[4] H. W. Yun, S. H. Shin, W. J. Jang and H. C. Park, "On-Line Audio Genre Classification using Spectrogram and Deep Neural Network," Journal of Broadcast Engineering vol. 21, No. 6, pp. 977-985, Nov. 2016.