스파이크그램을 이용한 심층 신경망 기반의 음악 장르 분류

*윤호원 장우진 신성현 조효진 장 원 박호종

광운대학교

*sleyard@kw.ac.kr

Music Genre Classification based on Deep Neural Network using Spikegram

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

Kwangwoon University

요약

본 논문에서는 인간의 청각 기관을 모델링 한 스파이크그램 (spikegram)을 이용한 심층 신경망 기반의 음악 장르 분류 기술을 제안한다. 분류 대상은 GTZAN 데이터 세트의 10개 장르로 정의한다. 본 논문에서는 청각 기관의 인식 방법을 모델링 한 방법을 이용하여 스파이크그램을 구하고, 스파이크그램에서 새로운 특성 벡터를 추출하는 방법을 제안한다. 제안하는 방법을 통해 심층 신경망에 적합한 특성 벡터를 구하고 이렇게 구한 특성 벡터로 신경망을 학습시켜 기존에 사용하던 다양한 방법들보다 높은 성능을 얻을 수 있다.

1. 서론

최근 심층 신경망에 대한 연구가 다양하게 진행되면서 학습의 핵심인 특성 벡터에 대한 연구도 활발하게 이루어지고 있다. 심층 신경망 (deep neural networks, DNN)은 인간의 신경망 구조를 모델링 한 기계학습 방법이기 때문에 기존에 오디오 분류의 핵심 특성으로 사용되던 MFCC나 스펙트로그램 기반의 특성 벡터뿐 아니라, 인간의 감각 기관의 인식 방법을 모델링 한 다양한 특성 벡터가 연구되고 있다. 과거에도 인간의 감각 기관에 대한 연구가 이루어졌으나, 많은 연산량의 문제로 실현에 제한이 있었다. 하지만 하드웨어의 발달과 그래픽 처리 장치(GPU)를 이용한 병렬 연산 처리가 가능해지면서 현실적으로 사용이 가능해졌다.

인간의 청각 시스템은 최대한의 정보를 최소한의 에너지와 신경 자원으로 전달해야하기 때문에 소리의 특성을 분석하는데 최적화 되어왔다[1]. 소리가 달팽이관에 도달하면 각 주파수를 인식하는 청세포에 자극이 가해지면서 전기신호가 발생하고, 뇌에서는 이 신호를 소리로 인식한다. 이때, 청세포에는 특정 주파수를 느끼는 세포에만 자극이 전달되고 그 이외의 주파수를 느끼는 부분은 어떤 자극도 없는 상태다. 이렇게 의미 있는 부분의 비율이 극단적으로 작은 상태를 스파스 (sparse)라고 하는데, 청각 인식의 과정에서 인간이 받는 자극은 상당히 스파스하다. 이것은 인간이 소리를 인식하는데 필요한 정보는 주어지는 정보에 비해 상당히 적은 것을 의미하며, 청각 시스템의 효율성을 보여준다.

본 논문에서는 이러한 청각 시스템을 기반으로 하는 스파이크그램을 이용한 오디오 특성을 구하는 방법을 제안한다. 제안하는 방법은 청각 필터를 모델링 한 감마톤 필터 (gammatone filter)와 MP (matching pursuit) 방법을 사용하여 5초 길이의 프레임 단위로 스파이크 (spike)를 구하고, 스파이크 기반의 특성 벡터로 학습한 심층 신경망이 다른 방법에 비해 성능이 향상된 것을 확인했다.

2. 제안하는 음악 장르 분류 방법

본 논문에서는 인간이 소리를 인식하는 시스템을 모델링 한 스파이크그램을 이용한 음악 장르 분류 방법을 제안한다. 식 (1)은 신호를 스파이크 코드로 표현하는 것을 나타낸다[2].

$$(t) = \sum_{=1 i=1}^{n} s_i^m \phi_m (t - \tau_i^n) + \epsilon(t) \quad (1)$$

는 감마톤 필터를 의미하며 각 필터를 커널 (kernel)이라고 한다. s는 커널이 가지는 크기 (gain)를, τ는 커널의 시간적 위치를 나타낸다.

은 알고리즘에 사용된 커널의 개수이며, n은 각 커널이 출현한 횟수이다. ϕ, s, τ 세 가지 값이 있으면 원본 신호의 복원도 가능하다. 본 논문에서는 청각 필터로 많이 쓰이는 64개 밴드의 감마톤 필터를 커널로 사용하며, 원본 신호와의 상관도 (correlation)를 구한다. MP 방법을 이용하여 상관도가 가장 높은 커널의 크기와 시간적 위치를 하나의 스파이크로 하여 시간-주파수 영역에 맵핑할 수 있다[2]. 정해진 개수까지 반복적으로 스파이크를 구하여 스파이크그램을 얻을 수 있다. 그림 1은 country 장르의 음원을 같은 시간에 대한 스펙트로그램과 스파이크그램으로 나타내고 있다. 그림 1의 (a)는 기존에 사용하는 스펙트로그램이며 (b)는 제안하는 방법의 스파이크그램이다.

MP 방법에 의해 구한 스파이크는 64개의 커널 중 몇 번째 커널이 어떤 크기를 가지고 사용되었는지에 대한 정보를 포함하고 있다. 본 논문에서는 정해진 개수에 대하여 각 커널의 출현 횟수와 커널의 크기의 합을 구한다. 정해진 개수의 스파이크를 모두 구했을 때의 원본 신호와 SNR을 구하고 최종으로 129차 특성 벡터를 구한다. 식 (2)와 식 (3)은 특성 벡터를 구하는 방법이다.

$$f_m = n_m (m = 1, \quad , 64) \quad (2)$$

$$_{+64} = \sum_{i=1}^{n} s_i^m, (m = 1, \quad ,64) \tag{3}$$

식 (2)의 커널의 출현 횟수는 입력 신호에 각 커널의 성분이 얼마나 많이 있는지를 나타내고, 식 (3)은 서로 중심 주파수가 다른 각 커널의 크기의 합은 각 커널의 영향이 얼마나 있는지를 나타낸다. 최종 SNR은 같은 수의 스파이크로 신호를 복원하는 정도를 나타내며 파형이 복잡한 장르일수록 SNR이 작고 파형이 단순한 장르일수록 SNR이 크다.

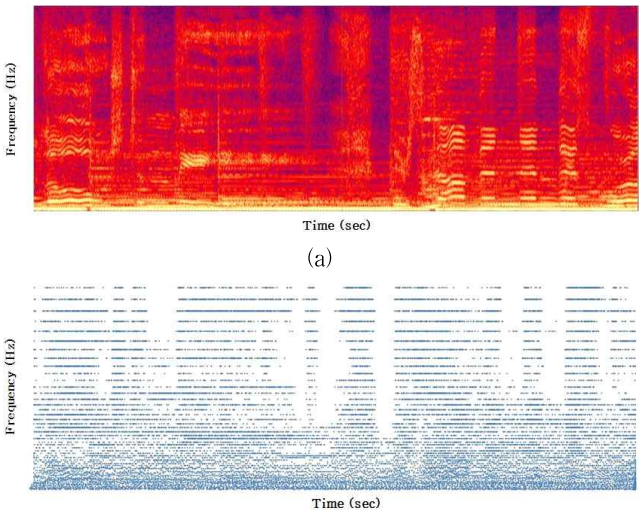

그림 1. Country 장르의 시간에 대한 주파수 특성 (a) 스펙트로그램 (b) 스파이크그램

Fig 1. Frequency response for time domain of country (a) Spectrogram (b) Spikegram (25,000 spikes)

3. 성능 평가

본 논문에서 성능 평가에 사용한 데이터 세트는 GTZAN 데이터 세트다. GTZAN 데이터 세트는 열 개의 장르 (classical, country, disco, hip hop, jazz, rock, blues, reggae, pop, metal)로 되어있으며, 각 장르는 30초 길이의 오디오 파일 100개다. 이 중 무작위로 10%를 실험 데이터 (test data)로 사용하고, 나머지 90%를 학습 데이터 (training data)로 사용한다. 또한, 데이터 세트를 10개의 같은 크기로 나누어 각 부분을 실험 데이터로 사용하여 열 번의 실험에 대한 평균적인 성능을 얻는 10-fold validation 방법을 사용했다.

실험에 사용한 심층 신경망의 구조 및 hyper-parameter는 반복적인 실험을 통해 최적화했다. 은닉층은 3개이며, 각 은닉층의 뉴런의 수는 입력 특성 벡터의 크기에 따라 설정하고, 출력층은 10개 장르를 분류하므로 10개로 설정했다. 학습 반복 횟수는 300회이며 mini-batch 크기는 1이다. 제안하는 방법은 5초 단위로 MP 방법을 사용하여 추출한 19,000개의 스파이크로 스파이크그램을 구하고 최종으로 129차 특성 벡터를 구한다.

스파이크그램을 이용한 장르 분류 성능은 표 1과 같다. 129차 스파이크그램 기반 특성 벡터를 이용한 평균 성능은 80.9%이다. 표 2는 다양한 분류기와 특성 벡터로 구한 성능을 나타내고 있다. 표의 괄호는 입력 특성 벡터의 차원을 의미한다. 옥타브 기반 스파스 코드의 성능이 제안하는 방법보다 높지만, 입력 특성 벡터의 차원이 약 4배 커지는 것을 확인할 수 있다.

표 1. 제안하는 스파이크그램을 이용한 GTZAN 장르 분류 인식률(%)

Table 1. The genre classification accuracy(%) using spikegram

	cl	co	di	hi	ja	ro	bl	re	po	me	Ave.
cl	**96**	1	0	0	1	1	0	0	0	1	
co	1	**78**	6	1	3	5	1	3	2	0	
di	2	2	**70**	7	0	6	0	6	5	2	
hi	0	1	6	**67**	0	1	0	15	6	4	
ja	1	2	2	0	**91**	1	2	0	0	1	**80.9**
ro	0	6	2	6	3	**69**	1	4	6	3	
bl	0	4	0	1	1	1	**93**	0	0	0	
re	0	3	3	5	3	4	2	**71**	8	1	
po	0	5	5	4	0	1	0	3	**82**	0	
me	0	1	0	2	1	4	0	0	0	**92**	

표 2. 다양한 방법의 GTZAN 장르 분류 인식률(%)

Table 2. The genre classification accuracy(%) of various methods

Classifier	Feature (# of feature)	Acc.
Linear SVM	Learned using PSD on octave[4] (512)	83.4
DNN	Spikegram (129)	**80.9**
CNN+BI+RNN	Spectrogram[3] (1024)	75
AdaBoost	Sparse code feature[4] (257)	63
GMM	MFCC+other[4] (30)	61

4. 결론

본 논문은 스파이크그램을 이용한 심층 신경망 기반의 음악 장르 분류 기술을 제안하였다. 제안 방법은 대표적으로 사용되는 MFCC나 스펙트로그램이 아닌 스파이크그램을 이용한 새로운 특성 벡터로 심층 신경망을 학습한다. 새로운 특성 벡터는 스파이크그램의 커널별 출현 횟수, 커널별 크기의 합, SNR을 구하여 얻을 수 있다. 제안 방법으로 신경망 학습에 효과적인 특성을 추출하며, 그 결과 음악 장르 분류에서 적은 수의 특성 벡터로 다른 방법보다 우수한 성능을 제공한다.

감사의 글

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

참고문헌

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

[2] 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.

[3] S. H. Kim, D. S. Kim and B. W. Suh, "Music Genre Classification Using Multimodal Deep Learning," Proc. of Human Computer Interaction(HCI) Korea 2016 Conf., pp. 389-395, Jan. 2016.

[4] M. Henaff, K. Jarrett, K. Kavukcuoglu and Y. LeCun, "Unsupervised Learning of Sparse Features for Scalable Audio Classification," Proc. of Int. Soc. Music Inf. Retrieval Conf. (ISMIR), pp. 681-686, Sep. 2011.