음악 장르 분류를 위한 스파이크그램 기반의 시간 및 주파수 특성 추출 기술

*장 원 조효진 신성현 박호종
광운대학교
*h575h@kw.ac.kr

Extraction of Temporal and Spectral Features based on Spikegram for Music Genre Classification

*Jang, Won Cho, Hyo-Jin Shin, Seong-Hyeon Park, Hochong
Kwangwoon University

요약

본 논문에서는 음악 장르 분류를 위한 시간 및 주파수 기반 스파이크그램 특성 추출 기술을 제안한다. 기존의 음악 장르 분류 시스템에서는 푸리에 변환 기반의 입력 특성을 주로 사용해 왔다. 푸리에 변환은 시간 축에서 프레임 단위로 평균적인 주파수 정보를 취하므로 낮은 시간 해상도를 갖지만, 스파이크그램은 샘플 단위의 주파수 정보를 갖고 있어 고해상도의 특성을 추출할 수 있다. 제안하는 기술은 이러한 시간 기반 특성을 추출하여 주파수 기반 특성 및 SNR 특성과 함께 심층 신경망의 입력으로 사용한다. 제안하는 특성을 사용하여 시간 기반 특성을 사용하지 않은 기존 스파이크그램 특성 기반 분류기의 성능을 개선하였으며, 다른 특성 및 분류기에 비해 적은 수의 특성 입력으로도 우수한 성능을 얻는 것을 확인하였다.

1. 서론

최근 디지털 음악 시장은 단순한 온라인 음원 스트리밍 제공을 벗어나 소셜 네트워크 서비스 (SNS)나 동영상 콘텐츠 공유 플랫폼과 연계하여 음악 콘텐츠에 특화된 다양한 서비스를 제공하는 방향으로 성장하고 있다. 청취한 음원의 장르나 악기, 가수의 성별 및 연령대 등의 메타데이터를 자동으로 추출하는 음악 정보 처리 (music information retrieval, MIR) 기술로 개인의 취향에 따라 맞춤형 음악 콘텐츠를 제공하는 등의 서비스를 기대할 수 있다[1]. 인간의 신경망 구조를 모델링한 심층 신경망 (deep neural network)이 최근 다양한 음악 정보 처리 분야에서 성능을 향상시키고 있으며, 심층 신경망으로 더욱 좋은 성능을 얻기 위해서는 적합한 입력 특성을 찾는 노력이 필수적이다.

기존의 음악 장르 분류 시스템에서는 입력 특성으로 스펙트로그램 (spectrogram)을 그대로 사용하거나 MFCC (Mel-frequency cepstral coefficients)로 변환하는 등 푸리에 변환을 기반으로 계산된 입력 특성을 사용한다. 하지만 푸리에 변환은 시간 축에서 프레임 단위로 평균적인 주파수 정보를 취하므로 낮은 시간 해상도를 갖는다. 인간의 청각 신호인 스파이크 (spike)를 수학적으로 모델링하고 이를 추출하여 시간 및 주파수 축에 나열한 구조를 스파이크그램 (spikegram)이라고 하는데, 스파이크그램은 샘플 단위의 주파수 정보를 갖고 있어 시간 축에서 보다 고해상도의 특성을 추출할 수 있다[2, 3]. 기존의 스파이크그램 특성은 특정 길이의 측정 시간에 걸쳐 각 주파수 대역을 담당하는 스파이크의 발생 빈도와 이득의 합을 계산해 사용했는데, 이러한 방법은 스파이크그램의 시간 고해상도 특징을 살리지 못하는 문제가 있다[3].

본 논문에서는 음악 장르 분류를 위한 새로운 스파이크그램 특성 추출 기술을 제안한다. 시간 기반 특성을 추출해 주파수 기반 특성 및 SNR (signal-to-noise ratio) 특성과 함께 심층 신경망의 입력으로 사용하여 푸리에 변환과 대조되는 스파이크그램의 장점을 살린다. 제안하는 특성을 사용하여 시간 기반 특성을 사용하지 않은 기존 스파이크그램 특성 기반 분류기의 성능을 개선하였으며, 다른 특성 및 분류기에 비해 적은 수의 특성 입력으로도 우수한 성능을 얻는 것을 확인하였다.

2. 제안하는 특성 구조

인간의 청각 시스템을 기반으로 한 특성을 구성하기 위해 먼저 감마톤 필터뱅크 (gammatone filterbank) 방식으로 스파이크를 모델링하며, 이를 커널 (kernel)이라고 한다. 커널은 바크 (Bark) 단위로 특정 주파수 성분을 나타내는 기본 단위이다[2, 3]. Matching pursuit (MP) 알고리즘으로 측정 시간 내에서 가장 큰 상관도 (correlation)를 갖는 커널별 스파이크의 종류와 위치, 그리고 이득을 추출해 저장한다. 이후 그 다음으로 큰 상관도를 갖는 스파이크를 원하는 개수만큼 반복해서 추출하고 이를 시간-주파수 축에 나열하여 스파이크그램을 완성한다.

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

식 (1)은 음원 $x(t)$를 각 스파이크의 가중 합으로 표현한 것이다. $\phi_m(t)$는 특정 주파수 대역의 커널 함수, M은 커널의 종류의 수, τ_i^m는 각 스파이크 별 시간 위치, n_m은 커널별 스파이크의 발생 횟수, g_i^m는 각 스파이크의 이득, $\epsilon(t)$는 모델링 오차를 의미한다.

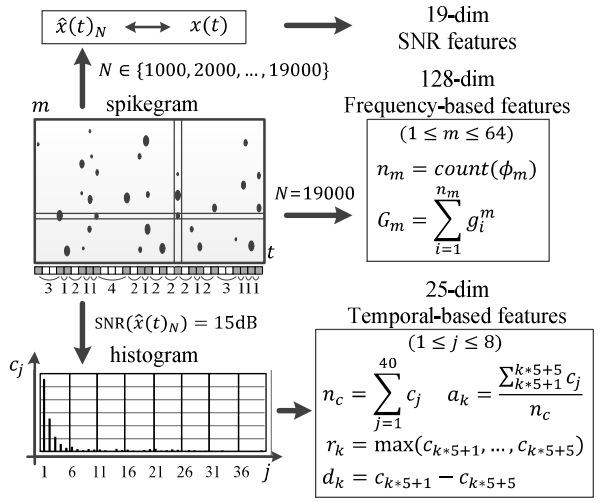

그림 1. 스파이크그램 기반 172차 특성 추출 구조. (위) 19차 SNR 특성 (중간) 128차 주파수 기반 특성 (아래) 25차 시간 기반 특성
Fig. 1. Spikegram based 172-th order features extraction structure. (top) 19-th order SNR features (mid) 128-th order frequency based features (bottom) 25-th order temporal based features

그림 1은 스파이크그램을 기반으로 제안하는 특성을 추출하는 구조이다. 커널의 종류 M은 64이고, 전체 특성 크기는 172이다. 먼저 추출한 스파이크로 복원 신호 $\hat{x}(t)_N$을 만들고, 음원 $x(t)$와의 오차 $\epsilon(t)$를 잡음으로 하여 SNR을 계산한다. 전체 스파이크 개수 N은 1000씩 19번 증가시켜 19개의 SNR을 계산해 특성으로 사용한다. 다음, N이 19000일 때의 스파이크그램으로부터 커널별 스파이크의 발생 횟수 n_m과 커널별 스파이크의 이득 총합 G_m을 계산한다. M=64이므로 총 128차 주파수 기반 특성이 추출된다[3].

제안하는 시간 기반 특성 추출 방법은 다음과 같다. SNR이 15dB가 될 때 까지 스파이크를 추출하고, 샘플 단위로 스파이크가 한 번이라도 존재하는지 아닌지를 체크한다. 그 후 스파이크가 존재하는 샘플 간 거리를 모두 구해 히스토그램 c_j를 계산한다. 이 때 40을 초과하는 값은 버리고 5개 간격으로 8개의 묶음을 만든다. 그리고 모든 c_j의 합 N, 각 묶음의 합을 N으로 나눠 정규화한 값 a_k, 각 묶음의 최대값 r_k, 각 묶음의 양 끝 값의 차 d_k를 추출해 25차 시간 기반 특성을 완성한다.

3. 성능 평가

음악 장르 분류에 사용한 신경망의 은닉층 뉴런 수는 각각 300, 60, 30이다. 은닉층은 ReLU (rectified linear unit), 출력층은 softmax 활성화 함수를 사용하였고, 학습률 0.007인 SGD (stochastic gradient descent) 방법으로 1000번 반복해 신경망을 학습시켰다.

심층 신경망 학습을 위해 10개의 장르로 구성된 GTZAN 데이터 세트를 사용하였다. 5초마다 특성을 추출해 신경망에 입력하고, 전체 음원의 길이 30초 동안 6번의 출력 평균이 가장 높은 장르를 최종 장르로 선택하였다. 성능 평가에는 데이터를 무작위로 10등분해 한 번씩 테스트 데이터로 사용하는 10-fold cross validation을 사용하였다.

표 1은 제안하는 특성을 사용한 분류 결과의 혼동행렬 (confusion matrix)이다. 특히 classical과 blues, metal을 잘 분류해 내고 있으며, 전체 음악 장르 분류의 평균 정확도는 84.7%이다.

표 1. 제안하는 특성을 사용한 분류 결과의 혼동행렬
Table 1. Confusion matrix of classification result using proposed features

Est. / True	cl	co	di	hi	ja	ro	bl	re	po	me	recall(%)
classic	**99**	1	0	0	0	0	0	0	0	0	99.0
country	1	**82**	4	0	4	2	2	3	1	1	82.0
disco	2	2	**78**	3	3	3	0	3	4	1	78.0
hiphop	0	2	3	**77**	0	2	1	7	4	4	77.0
jazz	2	1	0	0	**89**	2	4	0	1	1	89.0
rock	1	4	2	3	2	**74**	4	4	2	4	74.0
blues	0	4	0	0	2	1	**90**	0	0	3	90.0
reggae	0	3	2	5	1	2	1	**77**	7	2	77.0
pop	1	3	2	2	1	3	0	0	**88**	0	88.0
metal	0	0	2	0	2	3	0	0	0	**93**	93.0
precision(%)	93.4	80.4	83.9	84.6	85.6	80.4	88.2	81.9	82.2	85.3	**84.7**

표 2는 제안하는 특성과 기존 방법들과의 장르 분류 정확도 비교를 나타낸다. 제안하는 특성이 다른 특성 및 분류기를 사용한 방법보다 적은 수의 특성으로 우수한 분류 성능을 제공함을 확인할 수 있다.

표 2. 제안하는 특성과 기존 방법들과의 장르 분류 정확도 비교
Table 2. Genre classification accuracy comparison among the proposed features and other conventional methods

Features	Classifier	Dim.	Acc.(%)
Proposed features	**DNN**	**172**	**84.7**
Learned using PSD on octave[4]	SVM	512	83.4
Spikegram based features[3]	DNN	147	82.5
Spectrogram[1]	CNN+Bi-RNN	1024	75.0

4. 결론

본 논문에서는 음악 장르 분류를 위한 스파이크그램 기반 시간 및 주파수 특성 추출 기술을 제안하였다. 스파이크그램 특성으로 주로 사용되던 주파수 기반 특성과 SNR 특성에 시간 기반 특성이 추가된 172차 특성을 추출해 장르를 분류한다. 제안하는 특성이 기존의 분류 기술에 비해 적은 수의 특성으로도 높은 성능을 제공하는 것을 확인하였다.

감사의 글

본 연구는 2017년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016R1D1A1B03930923).

참고문헌

[1] S. H. Kim, D. S. Kim and B. W. Suh, "Music Genre Classification Using Multimodal Deep Learning," *Proc. of HCI Korea 2016 Conf.*, pp.389-395, Jan. 2016.

[2] E. Smith and M. Lewicki, "Efficient Auditory Coding," *Nature*, Vol.439, No.7079, pp.978-982, Feb. 2006.

[3] Woo-Jin Jang, Ho-Won Yun, Seong-Hyeon Shin, Hyo-Jin Cho, Won Jang, and Hochong Park, "Music Genre Classification using Spikegram and Deep Neural Network," *JBE*, Vol. 22, No. 6, Nov. 2017.

[4] M. Henaff, K. Jarrett, K. Kavukcuoglu and Y. LeCun, "Unsupervised Learning of Sparse Features for Scalable Audio Classification," *ISMIR*, pp.681-686, 2011.