특집논문 (Special Paper)
방송공학회논문지 제22권 제6호, 2017년 11월 (JBE Vol. 22, No. 6, November 2017)
https://doi.org/10.5909/JBE.2017.22.6.693
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

스파이크그램과 심층 신경망을 이용한 음악 장르 분류

장 우 진[a], 윤 호 원[a], 신 성 현[a], 조 효 진[a], 장 원[a], 박 호 종[a]‡

Music Genre Classification using Spikegram and Deep Neural Network

Woo-Jin Jang[a], Ho-Won Yun[a], Seong-Hyeon Shin[a], Hyo-Jin Cho[a], Won Jang[a], and Hochong Park[a]‡

요 약

본 논문은 스파이크그램과 심층 신경망을 이용한 새로운 음악 장르 분류 방법을 제안한다. 인간의 청각 시스템은 최소 에너지와 신경 자원을 사용하여 최대 청각 정보를 뇌로 전달하기 위하여 입력 소리를 시간과 주파수 영역에서 부호화한다. 스파이크그램은 이러한 청각 시스템의 부호화 동작을 기반으로 파형을 분석하는 기법이다. 제안하는 방법은 스파이크그램을 이용하여 신호를 분석하고 그 결과로부터 장르 분류를 위한 핵심 정보로 구성된 특성 벡터를 추출하고, 이를 심층 신경망의 입력 벡터로 사용한다. 성능 측정에는 10개의 음악 장르로 구성된 GTZAN 데이터 세트를 사용하였고, 제안 방법이 기존 방법에 비해 낮은 차원의 특성 벡터를 사용하여 우수한 성능을 제공하는 것을 확인하였다.

Abstract

In this paper, we propose a new method for music genre classification using spikegram and deep neural network. The human auditory system encodes the input sound in the time and frequency domain in order to maximize the amount of sound information delivered to the brain using minimum energy and resource. Spikegram is a method of analyzing waveform based on the encoding function of auditory system. In the proposed method, we analyze the signal using spikegram and extract a feature vector composed of key information for the genre classification, which is to be used as the input to the neural network. We measure the performance of music genre classification using the GTZAN dataset consisting of 10 music genres, and confirm that the proposed method provides good performance using a low-dimensional feature vector, compared to the current state-of-the-art methods.

Keyword : music genre, genre classification, spikegram, deep neural network

a) 광운대학교 전자공학과(Dept. of Electronics Engineering, Kwangwoon University)
‡ Corresponding Author : 박호종(Hochong Park)
E-mail: hcpark@kw.ac.kr
Tel: +82-2-940-5104
ORCID: http://orcid.org/0000-0003-1600-6610
※ 본 연구는 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016R1D1A1B03930923).
※ 이 논문의 연구결과 중 일부는 “2017년 한국방송·미디어공학회 하계학술대회”에서 발표한 바 있음.
· Manuscript received August 23, 2017; Revised October 19, 2017; Accepted October 19, 2017.

Ⅰ. 서 론

디지털 미디어에 대한 접근이 쉬워지면서 사용자에게 제공되는 음악 콘텐츠의 양이 계속 증가하고 있고, 그에 따라 음악 콘텐츠 검색 서비스뿐만 아니라 콘텐츠 특성 기반의 특화된 서비스에 대한 요구가 발생하고 있다. 예로, 음악 미디어를 사용할 때 사용자의 음악 취향과 청취 환경에 따라 특정 장르의 음악을 재생 또는 추천해 주고, 재생되는 음악 장르에 따라 차별적으로 최적의 이퀄라이저(equalizer)를 적용해 주는 서비스가 요구된다. 이러한 서비스를 제공하기 위해 음악 파형으로부터 음악 장르를 추정하는 장르 분류 기술이 필요하다[1-6].

기존의 장르 분류 기술은 주로 파형으로부터 수학 기반의 파라미터를 추출하여 장르 분류를 위한 특성으로 사용한다. 예로, 스펙트로그램(spectrogram)[2], MFCC(Mel-frequency cepstral coefficients)와 크로마 주파수(chroma frequency)[4], MFCC와 음색(timbre)[6] 특성을 사용하여 장르를 분류한다. 그러나 이와 같은 방법은 인간의 청각 시스템 동작을 정확하게 모델링 하지 못하고 인간이 인지하는 음악 특성에 특화된 정보를 제공하지 못하는 문제점을 가지고[7], 그에 따라 장르 분류의 정확도에 한계를 가진다. 본 논문에서는 이와 같은 문제를 극복하기 위하여 인간의 청각 시스템 동작을 기반으로 신호를 분석하는 새로운 방법을 사용하고, 그 결과로부터 새로운 특성 벡터를 추출하여 장르 분류의 성능을 높이는 방법을 제안한다.

인간의 청각 시스템은 최소 에너지와 신경 자원으로 최대 청각 정보를 뇌로 전달하기 위하여 소리 정보를 시간과 주파수 영역에서 부호화 한다[7]. 인간의 달팽이관에 있는 청 세포들은 특정 시간에 특정 주파수에 반응하여 전기 신호를 발생시키는데, 특정 시간에 입력된 소리에 포함된 의미 있는 주파수 성분들에만 해당 청 세포들이 반응하고 그 외에는 반응하지 않는다. 인간의 뇌는 이러한 방식으로 소수의 중요한 정보만을 받아들여 소리를 인식하고, 이는 청각 시스템의 효율성을 보여준다. 이와 같은 인간 청각 기관의 부호화 동작 원리를 이용하여 파형을 분석하는 방법을 스파이크그램(spikegram)이라고 한다[5].

본 논문에서 제안하는 장르 분류 방법은 스파이크그램을 이용하여 특성 벡터를 추출한다. 먼저 청각 필터로 널리 사용되는 모델인 감마톤 필터뱅크(gammatone filterbank)를 사용하여 스파이크그램을 구한다. 이를 이용하여 주파수 기반 특성과 스파이크 개수에 따른 복원 SNR(signal-to-noise ratio)을 구하여 특성 벡터를 추출하고, 이를 심층 신경망에 입력하여 음악 장르 분류를 위한 학습을 진행한다. 제안한 방법으로 GTZAN 데이터 세트에 포함되어있는 classical, jazz, rock 등의 10가지 음악 장르를 분류하였고[6], 기존 방법에 비하여 낮은 차원의 특성 벡터를 사용하면서 우수한 성능을 가지는 것을 확인하였다.

Ⅱ. 스파이크그램 추출

인간의 달팽이관 내부의 기저막(basilar membrane)에는 소리 인지에서 중심적 역할을 하는 청 세포들이 붙어있는데, 장소 이론(place theory)에 의하면 소리의 주파수에 따라 반응하는 청 세포의 위치가 다르다[8]. 입력된 소리가 달팽이관에 도달하면 그 파형에 포함되어있는 의미 있는 주파수 성분들에만 청 세포들이 반응하여 해당 시각에 전기적 신호인 스파이크(spike)를 만들고, 인간의 뇌는 이를 입력하여 소리를 인지한다. 이러한 방법으로 인간은 필요한 에너지와 신경 자원을 최소화하면서 뇌에 전달되는 정보를 최대화할 수 있다[7]. 이와 같은 스파이크 정보를 시간과 주파수 영역에서 표현한 것을 스파이크그램이라 하고, 인간의 청각 시스템 동작 모델을 기반으로 소리 파형을 분석하는 새로운 방법을 제공한다.

청 세포가 반응하는 특정 주파수 대역의 파형을 커널(kernel)이라 하고, 각 커널이 발생한 시간 위치와 크기를 기반으로 신호 $x(t)$를 분해하면 식 (1)과 같이 표현된다[5].

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

위 식에서 $\phi_m(t)$는 주파수별 커널의 시간 축 함수, m은 커널 인덱스, M은 총 커널의 개수, n_m은 커널별 스파이크의 출현 횟수, g_i^m는 커널별 스파이크의 이득(gain), τ_i^m는 커널별 스파이크의 시간 위치, $\epsilon(t)$는 모델링 오차를 의미한다. 식 (1)을 통해 입력 신호 $x(t)$를 특정 주파수 커널의

특정 시간 위치에서의 가중 합으로 표현할 수 있고, 이는 인간 청각 시스템의 부호화 동작을 모델링 한 것이다.

신호 $x(t)$로부터 스파이크그램을 추출하기 위해서 MP (matching pursuit) 알고리즘을 사용하는데[9], 이를 위해 모든 커널 $\phi_m(t)$의 모든 시간 위치 τ_i^m에 대한 신호 집합 $\mathbf{\Phi}$를 미리 설정한다. 본 논문에서는 인간의 청 세포가 특정 주파수에 반응하는 과정을 모델링 하기 위해 청각 필터로 널리 쓰이는 감마톤 필터뱅크를 커널로 사용한다. 그림 1은 제안 방법에서 사용한 64개의 감마톤 필터 중 일부를 보여주고, 각 필터가 하나의 커널에 해당한다. 따라서 각 커널은 바크 (Bark) 단위로 하나의 주파수 성분을 나타낸다.

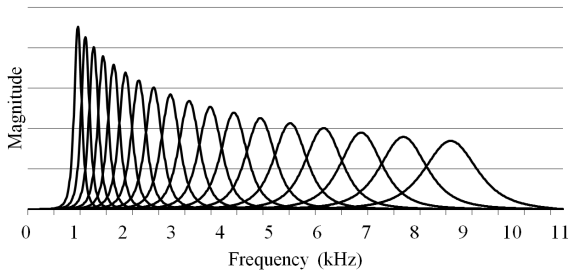

그림 1. 감마톤 필터
Fig. 1. Gammatone filters

MP 알고리즘을 사용하여 스파이크그램을 추출하는 과정은 그림 2와 같이 진행된다. 먼저 입력 $x(t)$와 미리 설정한 $\mathbf{\Phi}$의 각 요소 사이의 상관관계를 모두 구하여 가장 큰 상관관계를 가지는 $\mathbf{\Phi}$의 요소를 찾고, 해당 요소의 위상과 중심 주파수 정보를 저장한다. 이 때, 상관관계를 구하는 과정은 주파수 영역에서 수행하고 그래픽 처리 보드를 이용한 병렬 처리를 통하여 계산량을 줄일 수 있다. 다음, 해당 요소의 최적 이득을 구하고, $x(t)$에서 해당 요소의 성분을 제거하여 잔여 신호(residual signal) $r(t)$를 갱신한다. 여기까지의 과정이 스파이크 하나를 찾는 과정이고, 이 과정을 N개의 스파이크를 찾을 때까지 잔여 신호를 다시 입력해주면서 반복한다. 이렇게 찾은 N개의 스파이크에 해당하는 주파수, 위상, 이득 정보를 시간과 주파수 영역에 표현하여 스파이크그램을 구한다.

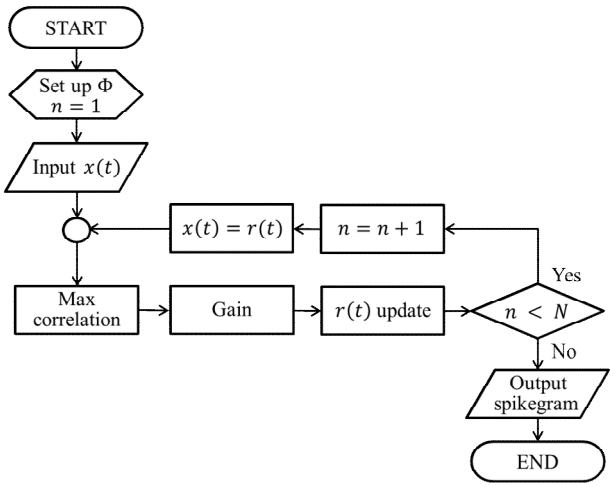

그림 2. MP 알고리즘을 이용하여 스파이크그램을 추출하는 과정
Fig. 2. Procedure of extracting spikegram using MP algorithm

그림 3은 특정 오디오 신호의 스펙트로그램과 그림 2 방법으로 추출한 스파이크그램을 비교한 것이다. 스파이크그램에 표시된 각 점은 하나의 스파이크를 나타내며, 스파이크의 시간 위치와 중심 주파수에 따라 점의 위치가 결정된다. 단, 그림 3에서 스파이크의 이득 정보는 표시하지 않았다. 스파이크그램은 해당 커널이 발생하는 시간 위치를 정확하게 표현할 수 있는 장점을 가지며, 이와 같은 높은 해상

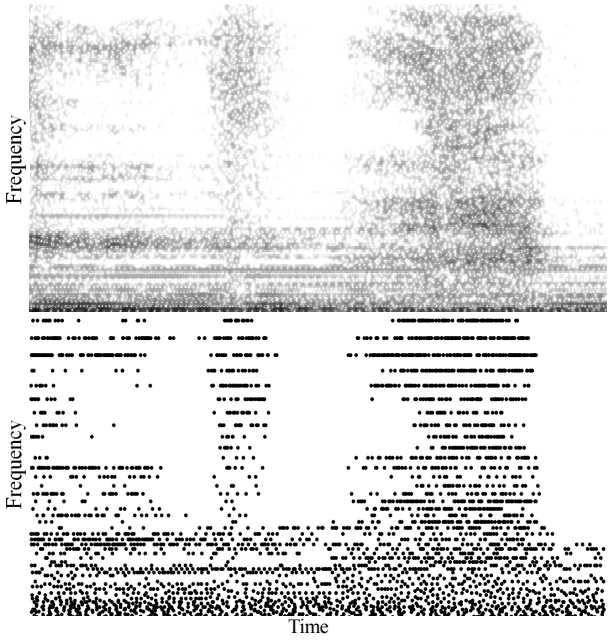

그림 3. Country 음원 1초에 대한 스펙트로그램 (위)과 스파이크그램 (아래)
Fig. 3. Spectrogram (top) and spikegram (bottom) of 1 second country music

도의 시간 정보는 프레임 단위로 스펙트럼을 구하여 나열한 스펙트로그램에서는 알 수 없는 정보이다.

Ⅲ. 제안하는 음악 장르 분류 방법

1. 개요

위에서 언급한 바와 같이 인간은 소리를 인식할 때 특정 주파수 커널의 가중 합으로 표현된 청각 정보를 사용하고, 스파이크그램은 이러한 인간의 청각 시스템 동작을 기반으로 파형을 표현한다. 본 논문에서는 스파이크그램 기반으로 신호의 특성을 추출하여 음악 장르를 분류하는 새로운 방법을 제안한다.

음악 장르 분류는 일정 시간에 대하여 신호의 특성을 추출하고 분석하는 과정으로 수행된다. 최적의 시간은 사용하는 특성과 분석 방법에 따라 달라지며 대개 3~5초를 사용한다[1]. 본 논문에서는 약 5초 단위로 특성 벡터를 추출하고 이를 심층 신경망의 입력 벡터로 사용하여 학습시킨다. 장르는 30초 단위로 분류하며, 30초 동안 6개의 특성 벡터에 대한 심층 신경망 출력값들의 평균이 가장 높은 장르를 최종 장르로 판단한다. 하나의 특성 벡터를 추출하기 위한 스파이크 개수는 실험을 통해 약 5초당 19,000개로 결정하였고, 그림 4는 사용한 스파이크 개수에 따른 음악 장르 분류 성능을 보여준다.

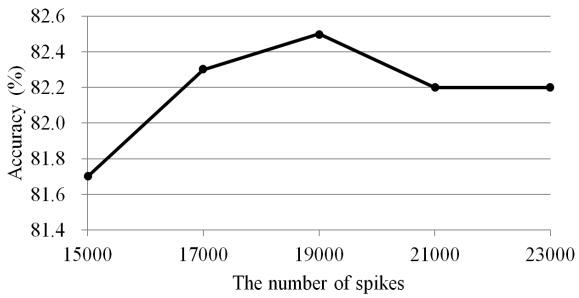

그림 4. 스파이크 개수에 따른 장르 분류 성능
Fig. 4. The genre classification accuracy as a function of the number of spikes

2. 스파이크그램 기반의 특성 벡터

신호에 포함된 주파수 정보는 인간이 음악 장르를 구분하는 데 있어서 중요한 정보가 된다. 특히 장르마다 주로 사용되는 주파수 대역이 다르므로 특정 주파수의 빈도와 크기에 대한 정보는 의미 있는 특성이라 할 수 있다. 본 논문에서는 이러한 정보를 포함하는 특성을 위해 그림 5와 같은 방법으로 스파이크그램으로부터 주파수 기반 특성을 추출한다. 그림 5에서 스파이크그램의 세로축은 각 커널의 중심 주파수 정보이고, 가로축은 약 5초에 해당하는 시간 위치(position) 정보이다. 시간 위치는 샘플 단위의 해상도로 표현된다. 데이터 세트의 샘플링 주파수는 22.05kHz이고, 2,048-샘플 프레임 단위로 커널을 설정하여 약 5초에 해당하는 2,048 × 53 = 108,544 샘플 길이로 시간 위치를 구성한다. $S_{k,p}$는 시간 위치 p에서 k번째 커널인 $\phi_k(t-p)$에 해당하는 스파이크의 이득을 나타낸다.

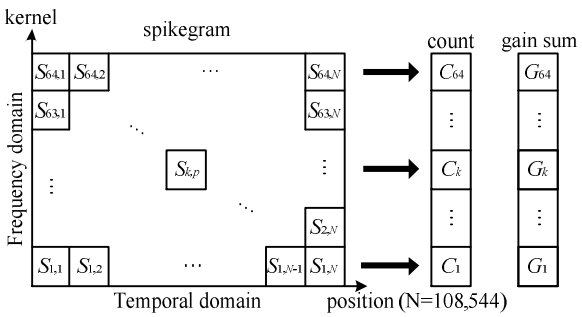

그림 5. 스파이크그램으로부터 주파수 기반 특성 벡터를 구하는 과정
Fig. 5. Procedure of computing frequency-based feature vector from spikegram

스파이크그램으로부터 주파수 기반 특성을 추출하기 위해 커널별 스파이크 중에서 이득이 0이 아닌 스파이크의 개수 C_k를 구하고, 커널별 스파이크 이득 합 G_k를 구한다[5]. 본 논문에서는 64개의 감마톤 필터를 커널로 사용하므로 C_k와 G_k는 각각 64차 벡터이고, 이 두 가지 특성을 순서대로 연결하여 128차의 주파수 기반 특성 벡터를 추출한다.

추출한 스파이크를 이용하여 식 (1)에 따라 신호를 복원할 수 있다. 복원 정확도를 알기 위해 원본 신호에서 복원 신호를 제거한 모델링 오차 신호 $\epsilon(t)$를 잡음으로 하여 원본 신호와의 SNR을 구한다. 복원이 정확히 될수록 $\epsilon(t)$ 성분이 줄어들고 SNR이 올라가게 된다. 같은 수의 스파이크를 사용해 복원하더라도 장르에 따라 SNR에 차이가 있고,

본 논문에서는 이 성질을 이용해 장르 분류를 위한 특성을 추출한다. 그림 6은 서로 다른 두 장르에 대해 3,000개의 스파이크를 이용하여 복원한 파형을 비교한 것이다. (a)의 classical 신호에서는 복원 신호와 원본 신호 사이에 거의 차이가 없는 반면, (b)의 metal 신호에서는 복원 신호와 원본 신호 사이에 큰 차이가 나타난다.

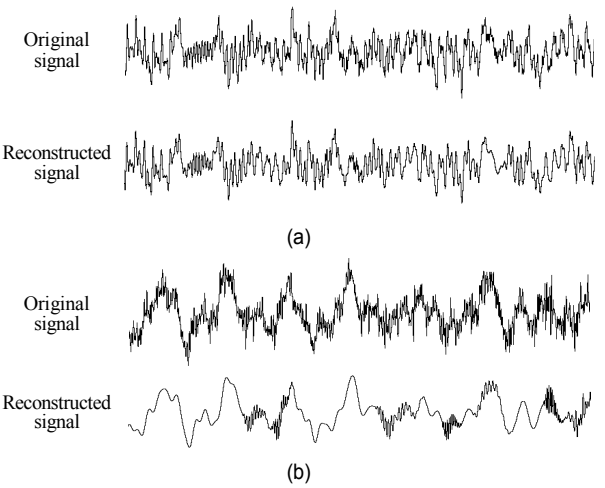

그림 6. 원본과 스파이크 3,000개를 이용하여 복원한 신호의 비교. (a) classical (b) metal
Fig. 6. Comparison between the original and the reconstructed signal using 3,000 spikes. (a) classical (b) metal

그림 7은 두 장르에 대하여 스파이크 개수가 1,000개씩 늘어날 때마다 SNR을 비교한 그래프이다. 두 장르를 같은 수의 스파이크를 이용하여 복원했지만 SNR에서 많은 차이가 나는 것을 확인할 수 있다. 따라서 본 논문에서는 스파이크 개수가 1,000개씩 늘어날 때마다 SNR을 구하고, 5초 구간에서 19,000개의 스파이크를 추출하므로 총 19개의 특성을 구한다. 따라서 앞에서 구한 128차 주파수 기반 특성 벡터와 스파이크 개수에 따른 SNR 기반의 19차 특성 벡터를 결합하여 최종적으로 147차 특성 벡터를 구한다.

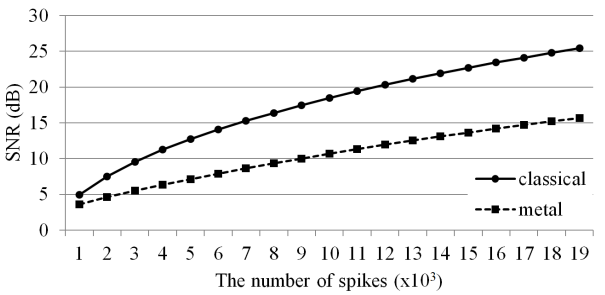

그림 7. 스파이크의 개수에 따른 SNR 변화
Fig. 7. Variation of SNR as a function of the number of spikes

3. 심층 신경망

심층 신경망의 학습은 여러 종류의 매개 변수(hyper-parameter)에 많은 영향을 받는다. 매개 변수를 제대로 설정해 주지 못하면 심층 신경망이 정확한 학습을 못 하거나 학습 데이터에 과적합(overfitting) 될 수 있고, 연산량이 비현실적으로 커지는 문제가 발생할 수 있다. 따라서 최적의 매개 변수 조합을 찾는 과정이 필수적이지만 이를 위한 이론적인 방법은 없으며, 실험을 통하여 다양한 매개 변수 조합에 대한 성능을 분석하여 최종값을 결정한다.

심층 신경망은 층(layer)의 수와 각 층의 뉴런(neuron)의 수에 따라 신경망 파라미터의 개수가 다르다. 반면에 학습 데이터의 양은 정해져 있으므로 신경망 파라미터의 개수가 과도하게 많은 경우 정확한 학습을 못 할 수 있고, 반대의 경우에는 과적합에 의해 오히려 성능이 저하될 가능성이 있다. 147차 특성 벡터를 사용하고 10개 장르로 분류하므로 입력층은 147개 뉴런, 출력층은 10개 뉴런으로 고정하고, 표 1과 같은 실험 결과를 통해 은닉층의 뉴런 수를 [250, 60, 30]로 최종 결정하였다.

표 1. 심층 신경망의 구조에 따른 장르 분류 성능
Table 1. The genre classification accuracy as a function of network structure

Number of hidden layers	Number of neurons in each hidden layer	Accuracy (%)
3	[200, 30, 30]	81.4
3	[250, 30, 30]	81.3
3	[250, 60, 30]	82.5
3	[300, 30, 30]	81.9
3	[300, 60, 30]	81.9

심층 신경망의 학습률이 잘못 설정되면 전역 최소값(global minimum)을 제대로 찾지 못하는 문제가 발생하기 때문에 최적의 학습률을 찾는 것이 중요하다. 본 논문에서는 그림 8의 실험 결과를 바탕으로 학습률을 0.007로 설정

하였다.

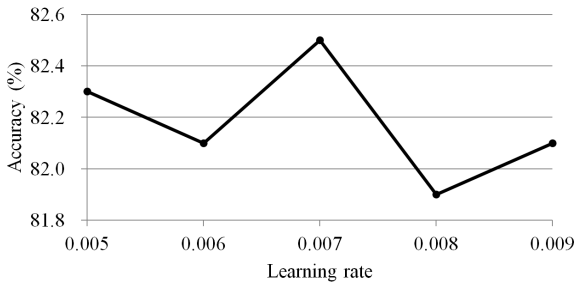

그림 8. 학습률에 따른 장르 분류 성능
Fig. 8. The genre classification accuracy as a function of learning rate

본 논문에서는 심층 신경망의 과적합 문제를 해결하기 위해 drop-out을 적용한다[10]. 그림 9는 drop-out 비율에 따른 장르 분류 성능을 나타낸다. 여기서 drop-out 비율은 학습 시 각 뉴런을 선택할 확률이며, 선택된 뉴런들로 신경망을 구성하여 학습한다. 본 논문에서는 입력층과 출력층을 제외한 모든 은닉층의 뉴런에 대해 같은 비율로 drop-out을 적용하였고, 그림 9과 같은 실험을 통해 drop-out 비율을 80%로 설정하였다.

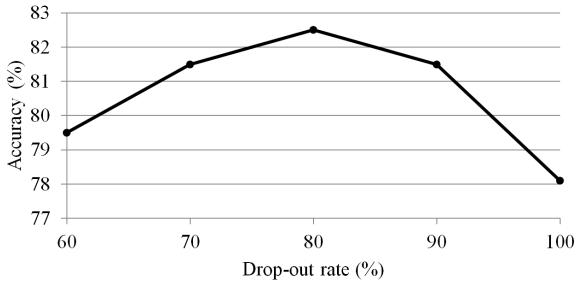

그림 9. Drop-out 비율에 따른 장르 분류 성능
Fig. 9. The genre classification accuracy as a function of drop-out rate

Ⅳ. 성능 분석

음악 장르 분류 기술의 성능 평가에 많이 사용되는 GTZAN 데이터 세트를 이용하여 제안 방법의 성능을 평가한다[6]. GTZAN 데이터 세트의 음악 장르는 총 10개 (classical, country, disco, hip hop, jazz, rock, blues, reggae, pop, metal)이며, 장르별로 30초 길이의 오디오 파일 100개로 구성되어 총 500분 길이의 음악 데이터를 포함한다. 제안 방법은 5초 단위로 특성 벡터를 추출하고, 30초 동안 6개 특성 벡터 각각에 대한 심층 신경망 출력값들의 평균이 가장 높은 장르를 최종 장르로 판단하고, 모든 장르에 대한 평균 정확도를 최종 성능으로 정의한다. 또한, 장르별로 데이터를 무작위로 10등분 하고, 각 부분을 실험 데이터 (test data)로 한 번씩 사용하여 평균 성능을 얻는 방법인 10-fold cross-validation을 사용한다.

표 2는 제안 방법을 사용한 장르 분류의 혼동행렬(confusion matrix)를 보여주고, 전체 10개 장르에 대한 평균 분류 정확도는 82.5%이다. 장르별로 약간의 성능 차이는 있으나 모든 장르에서 68% 이상의 높은 정확도를 가지는 것을 알 수 있다. 또한, hip hop과 reggae 사이에 오분류가 많이 발생하는데 이는 두 장르 사이에 근본적으로 높은 유사성이 있기 때문이다.

표 2. 음악 장르 분류에 대한 제안 방법의 혼동행렬
Table 2. Confusion matrix of the proposed method for music genre classification

True \ Est.	cl	co	di	hi	ja	ro	bl	re	po	me	Ave.
cl	**99**	0	0	0	0	1	0	0	0	0	**82.5**
co	2	**82**	1	0	4	4	2	3	1	1	
di	2	2	**70**	6	0	7	0	5	7	1	
hi	1	3	5	**68**	0	3	3	10	5	2	
ja	4	0	0	0	**92**	0	2	0	1	1	
ro	0	7	4	3	2	**75**	0	3	4	2	
bl	0	4	0	0	0	1	**91**	0	0	4	
re	0	1	1	7	4	6	3	**68**	9	1	
po	1	3	3	2	1	5	0	0	**85**	0	
me	0	0	1	0	2	2	0	0	0	**95**	

표 3은 제안 방법, 변형된 제안 방법, 기존 방법들의 성능을 비교한 것이다. [1]은 훈련 데이터로부터 학습된 인코더를 사용하여 옥타브(octave) 단위로 코드를 추출하고 SVM (support vector machine)을 사용하여 코드를 모델링 하는 방법을 사용한다. 이 방법의 성능은 제안 방법보다 약간 높지만, 특성 벡터가 512차로 제안 방법보다 약 3.5배 더 많은 특성 파라미터를 사용한다. 또한, [1] 방법은 코드 추출을

표 3. 제안 방법과 기존 방법들의 장르 분류 성능 비교
Table 3. Performance comparison between the proposed method and the conventional methods

Features	Dimension of feature vector	Classifier	Accuracy (%)
Learned using PSD on octave[1]	512	Linear SVM	83.4
Proposed spikegram-based feature (frequency-based + SNR)	**147**	**DNN**	**82.5**
Proposed spikegram-based feature (frequency-based only)	**128**	**DNN**	**80.3**
Spectrogram[2]	1024	CNN+BI-RNN	75
Sparse code feature[5]	257	AdaBoost	63
MFCC+other[6]	30	GMM	61

위한 인코더를 훈련을 통하여 설계하므로 매우 복잡한 훈련 과정이 필요하다. [2] 방법은 스펙트로그램이라는 간단한 특성 벡터를 사용하지만 차원이 매우 크고, 컨볼루션 신경망과 양방향 순환 신경망(bidirectional recurrent neural network)을 결합하여 제안 방법보다 매우 복잡한 신경망을 사용한다. [5] 방법은 본 논문과 같이 스파이크그램으로부터 특성을 추출하지만 추출된 특성의 성질과 표현하는 정보가 다르고, 그에 따라 성능에서 큰 차이를 가진다.

제안 방법은 128차 주파수 기반 특성 벡터와 스파이크 개수에 따른 19차 SNR 기반 특성 벡터를 사용한다. 만일 제안 방법에서 스파이크 개수에 따른 SNR 특성을 제외하고 128차 주파수 기반 특성 벡터만을 사용하면 평균 분류 정확도는 80.3%가 된다. 이로부터 본 논문에서 제안하는 128차 주파수 기반 특성 벡터만으로도 우수한 음악 장르 분류 성능을 가지는 것을 확인할 수 있다. 또한, 19차 SNR 기반 특성 벡터는 장르 분류 성능을 2.2% 포인트 높이는 역할을 하고, 이는 스파이크 개수에 따른 SNR 특성도 음악 장르를 구분하는 데 큰 역할을 하는 것을 보여준다.

V. 결 론

본 논문에서는 스파이크그램과 심층 신경망을 이용한 음악 장르 분류 기술을 제안하였다. 인간은 청각 기관에 입력된 소리의 주파수 성분 중 소수의 의미 있는 주파수의 가중합으로 신호를 모델링 하여 청각 정보를 인식한다. 스파이크그램은 이러한 인간의 청각 시스템 동작을 기반으로 파형을 효율적으로 표현하는 방법이다. 제안 방법은 스파이크그램을 이용하여 커널별 스파이크 출현 횟수와 이득의 합, 그리고 스파이크 개수에 따른 SNR을 구하여 147차 특성 벡터를 추출하고, 심층 신경망으로 특성 벡터를 모델링 하여 장르를 분류한다. GTZAN 데이터 세트를 이용하여 10개 음악 장르 분류 성능을 측정하였고, 제안 방법이 기존 방법과 비교하여 적은 특성 정보를 사용하여 우수한 성능을 제공하는 것을 확인하였다.

참 고 문 헌 (References)

[1] M. Henaff, K. Jarrett, K. Kavukcuoglu and Y. LeCun, "Unsupervised Learning of Sparse Features for Scalable Audio Classification," *Proceeding of International Society for Music Information Retrieval Conference (ISMIR)*, pp.681-686, Sep. 2011.

[2] S. H. Kim, D. S. Kim and B. W. Suh, "Music Genre Classification Using Multimodal Deep Learning," *Proceeding of Human Computer Interaction Korea*, pp.389-395, Jan. 2016.

[3] D. Bhalke, B. Rajesh and D. Bormane, "Automatic Genre Classification Using Fractional Fourier Transform Based Mel Frequency Cepstral Coefficient and Timbral Features," *Archives of Acoustics*, Vol.42, No.2, pp.213-222, 2017.

[4] M. Patil and U. Nemade, "Music Genre Classification Using MFCC, K-NN and SVM Classifier," *International Journal of Computer Engineering In Research Trends*, Vol.4, No.2, pp.43-47, Feb. 2017.

[5] P. Manzagol, T. Bertin-Mahieux and D. Eck, "On The Use of Sparse Time-Relative Auditory Codes for Music," *Proceeding of International Society for Music Information Retrieval Conference (ISMIR)*, pp.603-608, Sep. 2008.

[6] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, Vol.10, No.5, pp. 293-302, July 2002.

[7] E. Smith and M. Lewicki, "Efficient Auditory Coding," *Nature*, Vol.439, No.7079, pp.978-982, Feb. 2006.
[8] G. Mather, *Foundations of Perception*, Psychology Press, 2006.
[9] J. Tropp and A. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, Vol.53, No.12, Dec. 2007.
[10] N. Srivastava, G. Hinton, A. Krizhevsky and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, Vol.15, No.1, pp.1929-1958, June 2014.

저 자 소 개

장 우 진

- 2016년 2월 : 광운대학교 전자공학과 공학사
- 2016년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : http://orcid.org/0000-0003-0969-4582
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

윤 호 원

- 2016년 2월 : 광운대학교 전자공학과 공학사
- 2016년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : http://orcid.org/0000-0002-5998-2702
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

신 성 현

- 2016년 2월 : 광운대학교 전자공학과 공학사
- 2016년 3월 ~ 현재 : 광운대학교 전자공학과 석박사통합과정
- ORCID : http://orcid.org/0000-0002-2343-8983
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

조 효 진

- 2017년 2월 : 광운대학교 전자공학과 공학사
- 2017년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : http://orcid.org/0000-0003-2296-2270
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

저 자 소 개

장 원

- 2017년 2월 : 광운대학교 전자공학과 공학사
- 2017년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : http://orcid.org/0000-0002-4711-780X
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

박 호 종

- 1986년 2월 : 서울대학교 전자공학과 공학사
- 1987년 12월 : Univ. of Wisconsin-Madison 공학석사
- 1993년 5월 : Univ. of Wisconsin-Madison 공학박사
- 1993년 9월 ~ 1997년 8월 : 삼성전자 선임연구원
- 1997년 9월 ~ 현재 : 광운대학교 전자공학과 교수
- ORCID : http://orcid.org/0000-0003-1600-6610
- 주관심분야 : 오디오/음성 신호처리, 3D 오디오, 음악정보처리