스펙트로그램을 이용한 딥 러닝 기반의 오디오 장르 분류 기술

*장우진 윤호원 신성현 박호종

광운대학교

*bbang0719@kw.ac.kr

Audio Genre Classification based on Deep Learning using Spectrogram

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

Kwangwoon University

요약

본 논문에서는 스펙트로그램을 이용한 딥 러닝 기반의 오디오 장르 분류 기술을 제안한다. 기존의 오디오 장르 분류는 대부분 GMM 알고리즘을 이용하고, GMM의 특성에 따라 입력 성분들이 서로 직교한 성질을 갖는 MFCC를 오디오의 특성으로 사용한다. 그러나 딥 러닝은 입력의 성질에 제한이 없으므로 MFCC보다 가공되지 않은 특성을 사용할 수 있고, 이는 오디오의 특성을 더 명확히 표현하기 때문에 효과적인 학습을 할 수 있다. 본 논문에서는 딥 러닝에 효과적인 특성을 구하기 위하여 스펙트로그램(spectrogram)을 사용하여 오디오 특성을 추출하는 방법을 제안한다. 제안한 방법을 사용하면 MFCC를 특성으로 하는 딥 러닝보다 더 높은 인식률을 얻을 수 있다.

1. 서론

인터넷과 정보통신기술의 발달로 멀티미디어 데이터의 양이 기하급수적으로 증가하면서, 많은 양의 데이터를 효과적으로 관리하는 것이 중요해졌다[1]. 특히 오디오는 멀티미디어 데이터들이 공통으로 포함하는 중요한 정보이므로 더 정교한 분류 기술이 필요하다. 따라서 오디오 데이터의 효과적인 관리를 위한 자동 분류 기술들이 개발되었으며, 대표적으로 Gaussian Mixture Models (GMM)을 이용한 오디오 장르 분류 기술이 개발되었다.

GMM은 입력의 성분이 서로 직교해야 한다는 조건을 가지기 때문에 주로 Mel Frequency Cepstral Coefficients (MFCC)를 이용하여 장르 분류를 수행한다[2]. 그러나 MFCC는 신호에 인위적인 변형이 많이 적용된 상태이기 때문에 성능에 악영향을 미칠 수 있다. 이와 다르게 딥 러닝은 입력의 성질에 제한이 없어 어떠한 입력이 들어와도 학습할 수 있다. 따라서 딥 러닝 기반 장르 분류의 오디오 특성으로 MFCC를 그대로 사용하는 것은 딥 러닝의 장점을 활용하지 못한 것이고, 좀 더 가공되지 않은, 딥 러닝에 더 적합한 특성을 추출하는 방법이 필요하다.

본 논문에서는 스펙트로그램을 이용하여 오디오 특성을 추출하는 방법을 제안한다. 제안한 방법은 딥 러닝이 입력의 성질에 관계없이 학습이 가능하다는 점을 이용하여, MFCC를 구하기 전 과정에서 얻을 수 있는 스펙트로그램을 오디오의 특성으로 사용한다. 제안하는 방법을 사용하면 MFCC에 비해 가공되지 않은 특성을 얻을 수 있으므로 인식률이 높아진다.

2. 제안하는 방법

기존에 음성 인식에서 널리 쓰이는 오디오 특성은 MFCC이며, 이는 기존의 GMM 방식에 적합한 특성이다. 본 논문에서는 MFCC와 비슷한 의미가 있지만 딥 러닝에 더 적합한 특성을 얻는 방법을 제안한다.

그림 1은 제안하는 특성인 스펙트로그램을 추출하는 과정을 나타낸다[3]. 입력 신호 $x(n)$에 윈도를 씌워 프레임 단위로 나눈 후에 FFT(Fast Fourier Transform)를 적용하여 크기 스펙트럼(magnitude spectrum)을 구한다. 그 후 멜-필터링(mel-filtering)과정을 적용하여 f_i를 구한다. 기존의 MFCC는 f_i를 DCT하여 구할 수 있다.

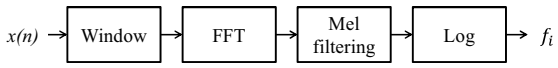

그림 1. 스펙트로그램 추출 과정

Fig. 1. Block diagram of spectrogram extraction

멜-필터링과정은 그림 2의 멜-필터 뱅크를 사용하여 식(1)과 (2)를 적용한다[4].

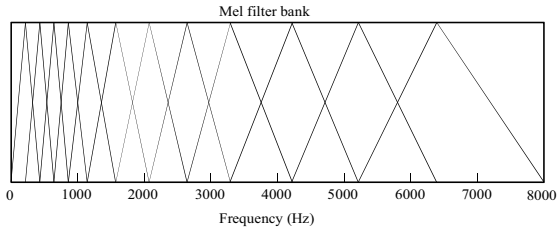

그림 2. 멜-필터 뱅크

Fig. 2. Mel-filter bank

$$fb_k = \sum_{i=cb_{k-1}}^{cb_k} \frac{i - cb_{k-1} + 1}{cb_k - cb_{k-1} + 1} b_i + \sum_{i=cb_k+1}^{cb_{k+1}} (1 - \frac{i - cb_k}{cb_{k+1} - cb_k + 1}) b_i \quad (1)$$

$$f_i = \ln(fb_i), i = 1, ..., 23 \quad (2)$$

여기서, b는 크기 스펙트럼, cb는 각 멜-필터 뱅크의 중심주파수를 나타낸다.

제안하는 방법은 먼저 식(2)의 f_i 23개를 그림 3의 (a)와 같이 프레임마다 구한다. 다음으로 1초에 해당하는 조직(texture) 프레임 단위로 평균과 분산을 구하고, 그 결과물을 특성으로 사용한다. 즉, 하나의 조직 프레임마다 그림 3의 (b)와 같이 f_i 의 평균과 분산으로 구성된 46차 벡터를 특성으로 사용한다. 여기서 m은 평균(mean), v는 분산(variance)을 나타낸다.

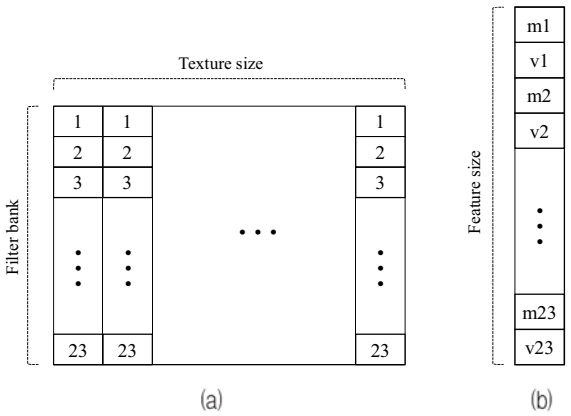

그림 3. 오디오 특성 벡터를 구하는 과정 (a) 프레임 단위 필터 뱅크 (b) 조직 프레임 단위 필터 뱅크

Fig. 3. Process of obtaining audio feature vector (a) filter bank of each frame (b) filter bank of each texture frame

3. 성능 평가

제안하는 방법의 성능 평가를 위하여 총 5개의 층을 가지는 인공신경망 구조로 딥 러닝을 수행한다. 3개의 은닉층은 각각 120, 45, 30개, 입력층은 46개, 출력층은 3개의 뉴런으로 구성된다. mini-batch 크기는 1이고 학습 반복 횟수(epoch)는 100이다. 각 층의 가중치들을 초기화하기 위해 RBM(Restricted Boltzmann Machines) 과정을 적용한다.

성능 평가로 사용된 데이터는 music, speech, effect의 3가지 장르로 구분되며 이들은 뉴스, 인터뷰, 영화, 음악프로그램 등의 TV 방송으로부터 얻을 수 있다. 장르별 음원의 길이는 32분이다. 이 중 장르별로 90%를 무작위로 선택하여 학습을 시키고, 나머지 10%로 실험하여 성능을 평가한다.

기존 방법과의 성능 비교를 위하여 그림 3과 같은 방법으로 MFCC 13개를 프레임마다 구한 다음 조직 프레임 단위로 평균과 분산을 구하여 26차 벡터를 기존 방법의 특성으로 사용한다. 학습 및 실험에 사용되는 데이터와 딥 러닝을 수행하기 위한 인공신경망의 구조는 모두 같은 것으로 사용한다.

다음의 표 1은 기존 방법인 MFCC를 특성으로 이용한 오디오 장르 분류의 인식률을 나타낸 것이고 평균 인식률은 89.24%이다. 표 2는 제안하는 방법인 스펙트로그램을 특성으로 이용한 오디오 장르 분류의 인식률을 나타낸 것이고 평균 인식률은 91.84%이다. 표 3에서 확인할 수 있듯이, 제안하는 방법은 기존 방법보다 장르별 인식률이 music은 1.6%p, speech는 1.0%p, effect는 4.7%p 향상되었고, 평균 약 2.6%p 향상되었다.

표 1. MFCC를 이용한 오디오 장르 분류 인식률(%)

Table 1. The accuracy of audio genre classification using MFCC(%)

True \ Estimated	Music	Speech	Effect
Music	88.0	4.2	7.8
Speech	2.1	93.8	4.2
Effect	2.6	11.5	85.9

표 2. 스펙트로그램을 이용한 오디오 장르 분류 인식률(%)

Table 2. The accuracy of audio genre classification using spectrogram(%)

True \ Estimated	Music	Speech	Effect
Music	89.6	2.6	7.8
Speech	2.6	94.8	2.1
Effect	4.2	5.2	90.6

표 3. 스펙트로그램 특성과 MFCC 특성의 인식률 차이(%p)

Table 3. The accuracy difference between spectrogram and MFCC(%p)

	Music	Speech	Effect	Average
Difference	1.6	1.0	4.7	2.6

4. 결론

본 논문에서는 스펙트로그램을 이용한 딥 러닝 기반의 오디오 장르 분류 기술을 제안하였다. 제안한 방법은 기존에 특성으로 사용하는 MFCC를 필터 뱅크로 바꾸고 평균과 분산으로 특성 벡터를 구한다. 이처럼 구한 특성 벡터를 이용하여 딥 러닝을 수행한다.

제안한 방법을 사용하면 딥 러닝에 효과적인 특성을 추출할 수 있고, 이에 따라 기존 MFCC를 특성으로 사용하는 방법보다 높은 오디오 장르 분류 인식률을 얻을 수 있다.

참고문헌

[1] W. J. Yoon, K. K. Lee and K. S. Park, "A study on the signal processing for content-based audio genre classification", IEEK, Nov. 2004.

[2] L. Liu and J. He, "On the use of orthogonal GMM in speaker recognition", IEEE, vol. 2, pp. 845-848, Mar. 1999.

[3] J. E. Kim and I. S. Lee, "Speech/Mixed content signal classification based on GMM using MFCC", journal of IEIE, vol. 50, no. 2, pp. 185-192, Feb. 2013.

[4] ETSI ES 202 211, v1.1.1, pp. 10-14, Nov. 2003.