온라인 오디오 장르 분류의 성능 분석

*윤호원 장우진 신성현 박호종

광운대학교

*sleyard@kw.ac.kr

The Performance Analysis of On-line Audio Genre Classification

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

Kwangwoon University

요약

본 논문에서는 온라인 오디오 장르 분류의 성능을 비교 분석한다. 온라인 동작을 위해 1초 단위의 오디오 신호를 입력하여 music, speech, effect 중 하나의 장르로 판단한다. 학습 방법은 GMM과 심층 신경망을 사용하며, 특성은 MFCC와 스펙트로그램을 포함하는 네 가지 종류의 벡터를 사용한다. 각 성능을 비교 분석하여 장르 분류에 적합한 학습 방법과 특성 벡터를 확인한다.

1. 서론

본 논문은 오디오 장르 분류의 여러 가지 방법에 대한 성능을 비교 분석한다. 기계학습 방법으로 GMM (Gaussian mixture model)과 심층 신경망 (deep neural network, DNN)을 이용하고, 특성으로 MFCC (mel-frequency cepstral coefficient)와 스펙트로그램 (spectrogram)을 이용한다. 분류 장르는 music, speech, 음향효과를 나타내는 effect 3가지이며, 장르의 특성을 나타내기 위한 최소한의 길이를 1초로 설정하여 1초 단위 온라인 장르 분류를 진행한다. 각 방법에 대한 온라인 장르 분류의 성능을 비교 분석하여 더 적합한 학습 방법과 오디오 특성을 제시하고자 한다.

2. 기계학습 방법과 오디오 특성

1. GMM과 심층 신경망

GMM은 장르 분류에 사용되는 대표적인 모델링 방법으로, 특성 파라미터의 확률 분포를 여러 개의 가우시안 확률 분포의 가중치 합으로 모델링 한다[1]. 각각의 가우시안 확률 분포를 가우시안 컴포넌트라 하며, 컴포넌트의 수가 많아지면 더욱 정교한 모델링이 가능하며, EM 알고리즘을 통해 학습을 진행한다.

신경망은 인간의 신경망을 모방한 학습 방법으로, 입력층, 은닉층, 출력층으로 구성된다. 하나의 층은 여러 개의 뉴런으로 구성되고, 인접한 두 층 사이의 뉴런은 가중치와 바이어스로 연결된다. 여러 개의 은닉층을 사용하면 정확한 파라미터 모델링이 가능하며, 이러한 구조를 심층 신경망이라고 한다[2].

2. 오디오 특성

MFCC는 기존 음성 인식에서 널리 쓰이는 오디오 특성으로, 인간의 귀가 고주파로 갈수록 민감하지 않은 특성에 따라 mel 스케일을 이용한다[3]. 스펙트로그램은 오디오 신호의 가장 일반적인 표현 방식으로 시간 축과 주파수 축의 진폭의 변화를 동시에 알 수 있다. 그 밖에 스펙트럼 분포, 무게중심, 변화량 등을 MFCC와 함께 사용할 수 있다[4].

3. 성능 분석

성능 평가에 사용된 오디오 데이터는 실제 TV 방송에서 얻은 음원이며, 장르별로 32분이다. 1초 단위의 장르 분류를 위해 수시로 장르가 변하도록 데이터를 구성하였고, 전체 데이터의 90%를 무작위로 뽑아 학습 데이터로 사용하고, 나머지 10%를 실험 데이터로 사용한다.

본 논문에서는 GMM을 이용한 오디오 장르 분류와 심층 신경망을 이용한 오디오 장르 분류에 대한 성능을 분석하며, 각 방법에 대해서 오디오 특성을 바꿔가며 실험을 진행한다. 오디오 신호는 22.05kHz 샘플링 주파수를 가지고, 512-샘플 프레임 단위로 특성을 구하며, 여러 개 프레임을 연결하여 1초 길이의 texture 프레임을 정의하고 texture 프레임에서 각 특성 파라미터의 평균과 분산을 구하여 최종 특성 벡터를 구한다. 실험에 사용한 특성 벡터는 DC를 제외한 저대역 5개의 MFCC를 사용한 10차 특성 벡터, 13개의 MFCC 전부를 사용한 26차 특성 벡터, DC를 제외한 저대역 5개의 MFCC를 기반한 19차 특성 벡터, 23개의 스펙트로그램 밴드를 모두 사용한 46차 특성 벡터로 총 4가지 특성 벡터로 성능을 비교한다.

GMM은 20개의 가우시안 컴포넌트를 사용하며, 학습 알고리즘인 EM 알고리즘의 반복 학습 횟수는 200번, 오차 허용범위는 0.001로 설정했다. 심층 신경망은 3개의 은닉층을 사용하며, 각 은닉층은 120, 45, 30개의 뉴런으로 구성되었다. 학습률은 0.07이며, 학습 반복 횟수는 500번이다.

표 1은 MFCC 10차 특성 벡터를 이용한 장르 분류 성능을 보여준다. DC를 제외한 5개의 저대역 MFCC만을 사용하여도 비교적 높은 성능을 얻어 MFCC의 저대역이 핵심적인 역할을 하는 것을 확인할 수

있다.

표 1. MFCC 10차 특성 벡터를 이용한 장르 분류 성능 (%)
Table 1. The performance of genre classification using 10-D MFCC feature vector(%)

Estimated / True	GMM			DNN		
	Speech	Music	Effect	Speech	Music	Effect
Speech	**92.19**	3.13	4.68	**94.27**	4.17	1.56
Music	3.65	**75.00**	21.35	2.60	**89.06**	8.33
Effect	6.25	17.19	**76.56**	7.29	9.90	**82.81**
Precision	90.30	78.68	74.63	90.50	86.36	89.33

표 2는 MFCC기반 19차 특성 벡터를 이용한 장르 분류 성능이다. GMM은 4가지 특성 벡터 조건 중 가장 높은 성능을 보인다. 하지만 GMM은 speech와 effect의 성능 차이가 10%p 이상으로 장르 간의 큰 성능 차이를 보인다. 반대로 심층 신경망은 장르에 관계없이 비슷한 성능을 얻으며, 전체 성능 또한 GMM에 비해 높게 나와 오디오 장르 분류에 있어 심층 신경망이 GMM보다 더 효율적인 모델링을 하는 것을 확인할 수 있다.

표 2. MFCC기반 19차 특성 벡터를 이용한 장르 분류 성능(%)
Table 2. The performance of genre classification using 19-D feature vector based on MFCC(%)

Estimated / True	GMM			DNN		
	Speech	Music	Effect	Speech	Music	Effect
Speech	**95.31**	3.13	1.56	**96.35**	1.04	2.60
Music	3.13	**89.06**	7.81	1.04	**96.88**	2.08
Effect	2.60	13.02	**84.38**	1.56	5.21	**93.23**
Precision	94.33	84.65	90.01	97.37	93.94	95.21

표 3은 MFCC 26차 특성 벡터를 이용한 장르 분류 성능을 보여준다. 특성 벡터의 차원 증가로 MFCC 10차 특성 벡터보다 성능이 증가하였으나, MFCC기반의 19차 특성 벡터보다 성능이 떨어져 MFCC의 중, 고대역의 성분이 효율적인 오디오 장르 특성 역할을 하지 못하는 것을 확인할 수 있다. 또한, 19차 특성 벡터와 비슷한 성능을 얻은 speech, effect와 다르게 music의 성능이 떨어진 것을 확인할 수 있다.

표 3. MFCC 26차 특성 벡터를 이용한 장르 분류 성능(%)
Table 3. The performance of genre classification using 26-D MFCC feature vector(%)

Estimated / True	GMM			DNN		
	Speech	Music	Effect	Speech	Music	Effect
Speech	**95.83**	3.13	1.04	**96.35**	0.52	3.13
Music	3.65	**85.94**	10.41	1.56	**93.75**	4.69
Effect	3.65	11.97	**84.38**	2.60	4.17	**93.23**
Precision	92.92	85.06	88.05	95.85	95.24	92.27

표 4는 스펙트로그램 46차 특성 벡터를 이용한 장르 분류 성능을 보여 준다. 특성 벡터의 차원이 증가하였으나 GMM은 성능이 떨어지는 것을 확인할 수 있다. 이러한 이유로 GMM은 스펙트로그램보다 MFCC가 더 효율적인 특성 역할을 하는 것을 확인할 수 있다. 심층 신경망은 4가지 특성 벡터 조건 중 스펙트로그램을 특성 벡터로 사용할 때 가장 높은 성능을 보인다.

표 4. 스펙트로그램 46차 특성 벡터를 이용한 장르 분류 성능(%)
Table 4. The performance of genre classification using 46-D spectrogram feature vector(%)

Estimated / True	GMM			DNN		
	Speech	Music	Effect	Speech	Music	Effect
Speech	**89.58**	6.77	3.65	**97.40**	0.52	2.08
Music	5.21	**75.00**	19.79	2.08	**94.79**	3.13
Effect	11.46	22.92	**65.63**	1.56	3.13	**95.31**
Precision	84.31	71.64	73.68	96.39	96.30	94.82

GMM은 music과 effect가 서로 잘못 분류하는 문제를 보이며, MFCC가 포함된 특성을 사용한 성능에 비해 스펙트로그램을 특성으로 사용 성능이 많이 떨어진다. 반면, 심층 신경망은 GMM에 비해 전체적인 성능이 향상되었으며, music과 effect의 오분류 비율이 감소한 것을 확인할 수 있다.

4. 결론

본 논문에서는 1초 단위의 온라인 장르 분류에 대한 GMM과 심층 신경망의 성능을 분석하였다. 각 오디오 특성에 대하여 심층 신경망의 성능이 GMM보다 전반적으로 높은 것을 확인할 수 있다. GMM은 MFCC에 기반한 19차 특성 벡터를 사용할 때 가장 높은 성능을 보이며, 심층 신경망은 스펙트로그램을 특성 벡터로 사용할 때 가장 높은 성능을 보였다. 이 결과로 실시간 오디오 장르 분류에서 GMM보다 심층 신경망의 성능이 뛰어나며, 심층 신경망에는 MFCC보다 스펙트로그램이 특성 벡터로 적합한 것을 알 수 있다.

감사의 글

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었습니다(IITP-2016-H8501-16-1014).

참고문헌

[1] D. Reynolds, "Gaussian Mixture Models," Encyclopedia of Biometrics, pp. 827-832, Jul. 2015.

[2] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning," Nature, 521.7553, pp. 436-444, May. 2015.

[3] ETSI ES 202 211, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Extended Front-End Feature Extraction Algorithm; Compression Algorithm; Back-End Speech Reconstruction Algorithm," Nov. 2003.

[4] 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, Jul. 2002.