일반논문 (Regular Paper)
방송공학회논문지 제21권 제6호, 2016년 11월 (JBE Vol. 21, No. 6, November 2016)
http://dx.doi.org/10.5909/JBE.2016.21.6.977
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

스펙트로그램과 심층 신경망을 이용한 온라인 오디오 장르 분류

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

On-Line Audio Genre Classification using Spectrogram and Deep Neural Network

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

요 약

본 논문은 스펙트로그램과 심층 신경망을 이용한 온라인 오디오 장르 분류 방법을 제안한다. 제안한 방법은 온라인 동작을 위하여 1초 단위로 신호를 입력하여 speech, music, effect 중 하나의 장르로 분류하고, 동작의 범용성을 위하여 기존 오디오 분석에 널리 사용되는 MFCC 대신에 스펙트로그램 기반의 특성 벡터를 사용한다. 실제 TV 방송 신호를 사용하여 장르 분류 성능을 측정하였고, 제안 방법이 기존 방법보다 각 장르에 대하여 우수한 성능을 제공하는 것을 확인하였다. 특히 제안 방법은 기존 방법에서 나타나는 music과 effect 사이를 잘못 분류하는 문제점을 감소시킨다.

Abstract

In this paper, we propose a new method for on-line genre classification using spectrogram and deep neural network. For on-line processing, the proposed method inputs an audio signal for a time period of 1sec and classifies its genre among 3 genres of speech, music, and effect. In order to provide the generality of processing, it uses the spectrogram as a feature vector, instead of MFCC which has been widely used for audio analysis. We measure the performance of genre classification using real TV audio signals, and confirm that the proposed method has better performance than the conventional method for all genres. In particular, it decreases the rate of classification error between music and effect, which often occurs in the conventional method.

Keyword : audio genre, deep neural network, spectrogram, genre classification

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 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었습니다(IITP-2016-H8501-16-1014).
· Manuscript received September 6, 2016; Revised October 7, 2016; Accepted October 7, 2016.

Ⅰ. 서 론

오디오 콘텐츠의 특성을 분석하여 해당 특성에 특화된 서비스를 제공하려는 시도가 여러 플랫폼에서 진행 중이다[1]. 오디오 신호의 장르 (genre)는 오디오 콘텐츠의 특성을 나타내는 대표적인 분류 기준이며, 따라서 오디오 장르를 자동으로 분류하는 기술이 널리 연구되고 있다. 대부분의 기존 기술은 주어진 오디오 신호 전체를 분석하여 해당 오디오의 장르를 한 번 판정하는 방법을 사용한다[2,3]. 즉, 기존 방법은 단말기 또는 서버에 저장된 오디오 신호의 장르 분류에 적용하는 오프라인 (off-line) 방법에 해당한다. 최근에는 방송에서 콘텐츠 특성에 따른 특화된 서비스를 위하여 오디오 신호의 온라인 (on-line) 장르 분류 방법이 요구되고 있다[4]. 예로, TV 프로그램을 시청할 때 방송되는 음악에 따라 실시간으로 최적의 이퀄라이저를 적용하여 시청자에게 제공하려면 온라인 장르 분류가 필요하다. 따라서 짧은 시간 단위로 장르를 분류하는 온라인 장르 분류 기술 연구가 요구된다.

본 논문은 speech, music, 음향 효과 또는 자연음에 해당하는 effect 등의 3가지 장르를 1초 단위로 분류하는 온라인 장르 분류 기술을 제안한다. 기존의 대표적인 장르 분류 기술은 오디오 신호의 MFCC (mel-frequency cepstral coefficient), 스펙트럼 분포, 스펙트럼 변화량 등의 시간 통계 특성을 GMM (Gaussian mixture model) 또는 심층 신경망 (deep neural network, DNN)으로 모델링 하는 방법을 사용한다[2,3,5]. 실험 결과에 의하면 GMM 방법은 speech에 대하여 우수한 성능을 제공하지만 music과 effect 장르를 서로 잘못 분류하는 문제점을 가진다. 또한, 기존 DNN 방법은 30초 단위로 장르를 분류하는 오프라인 동작만 제공하고[3], MFCC와 같이 매우 특화된 특성을 사용하여 범용성에 한계를 가진다[3,5]. 본 논문은 이 문제를 해결하기 위하여 새로운 특성을 정의하고 DNN으로 모델링 하는 장르 분류 방법을 제안한다. 특히, 제안 방법은 GMM이 요구하는 특화된 성질을 가지는 MFCC를 사용하지 않고, 대신 오디오 신호의 가장 일반적인 표현 방식인 스펙트로그램 (spectrogram) 기반의 특성 벡터를 사용한다. 즉, 제안하는 방법은 모델링 방법과 오디오 정보에 특화된 특성 벡터가 아니라 일반적인 오디오 특성 벡터를 사용하고, 이로부터 다른 오디오 정보 인식과 특성 벡터를 공유하고 통합된 동작을 가능하게 한다.

본 논문에서 제안하는 방법은 프레임 단위로 스펙트럼과 mel-필터 출력을 구하여 1초 길이의 texture 프레임에 대한 스펙트로그램을 구하고, texture 프레임에서 각 대역별 평균과 분산을 구하여 최종 특성 벡터를 구한다. 이렇게 구한 특성 벡터를 3개의 은닉층 (hidden layer)을 가지는 DNN으로 모델링 하여 speech, music, effect 중 하나를 해당 texture 프레임의 장르로 최종 선택한다. 실제 TV 방송의 오디오 신호를 사용하여 장르 분류 성능을 평가하였으며, 제안한 방법이 기존의 GMM 방법에 비하여 향상된 장르 분류 성능을 제공하고, 특히 music과 effect 사이를 잘못 분류하는 문제를 해결한 것을 확인하였다.

Ⅱ. 기존 오디오 장르 분류 방법

1. 특성 모델링 방법

오디오 장르 분류의 일반적인 과정은 그림 1과 같다. 제공되는 훈련 데이터에서 특성 벡터를 구하고, 최상의 장르 분류 동작을 수행하도록 특성 벡터를 모델링 하여 모델 파라미터를 구한다. 다음, 입력 오디오 신호가 주어지면 특성 벡터를 추출하고 훈련된 모델 파라미터에 따라 특성을 모델링 하여 최종 장르를 결정한다.

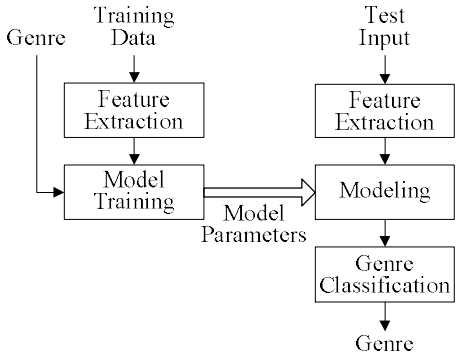

그림 1. 오디오 장르 분류 과정
Fig. 1. Overall structure of audio genre classification

오디오 장르 분류를 위한 대표적인 모델링 방법에 GMM과 DNN이 있다. GMM은 파라미터의 확률 분포를 여러 개 가우시안 확률 분포의 가중치 합으로 모델링 한다[6]. 각각의 가우시안 확률 분포를 가우시안 컴포넌트 (Gaussian component)라고 하며, 가우시안 컴포넌트의 수가 증가할수록 정교한 확률 모델링이 가능해진다.

GMM은 주어진 학습 데이터를 사용하여 각 장르별로 특성 벡터의 확률 분포를 M 개의 가우시안 컴포넌트로 모델링 한다. i번째 가우시안 컴포넌트는 식 (1)로 나타내며, 여기서 x는 각 D차 특성 벡터로 구성된 학습 데이터 집합, μ_i는 i번째 컴포넌트의 평균, $\sum_i$는 i번째 컴포넌트의 공분산 행렬을 나타낸다.

$$g(x|\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \quad (1)$$

$i=1,2,\cdots,M$

각 장르별 GMM 확률 분포는 식 (2)와 같이 각 장르별로 구한 $g(x|\mu_i,\Sigma_i)$의 가중치 합으로 표현된다. 여기서 λ는 각 장르를 의미하고, 각 장르별로 서로 다른 GMM 파라미터 $\{w_i, \mu_i, \Sigma_i\}$, $i=1,2,\cdots,M$를 가진다.

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i,\Sigma_i) \quad (2)$$

GMM 파라미터 $\{w_i, \mu_i, \Sigma_i\}$는 각 장르별로 주어진 훈련 데이터 x에 대하여 $p(x|\lambda)$를 최대로 하는 훈련 과정을 통하여 구한다. 입력 신호에 대한 장르 분류는 입력 특성 벡터 $\overline{x}$에 대하여 3가지 장르 λ_S, λ_M, λ_E에 대한 조건 확률 $p(\overline{x}|\lambda_S)$, $p(\overline{x}|\lambda_M)$, $p(\overline{x}|\lambda_E)$를 모두 구하여 최대 확률을 가지는 장르를 최종 장르로 출력한다.

신경망 (neural network)은 인간의 신경망 구조를 모방하는 기계학습 방법으로, 정보를 입력하는 입력층 (input layer), 입력 특성을 모델링 하는 은닉층 (hidden layer), 모델링 결과를 출력하는 출력층 (output layer)으로 구성된다. 하나의 층은 여러 개의 뉴런 (neuron)으로 구성되고, 인접한 두 층 사이의 뉴런은 가중치 (weight)와 바이어스 (bias)로 연결되고 최종 뉴런 값은 활성화 함수 (activation function)에 따라 결정된다. 계층 구조를 가지면서 보다 체계적이고 정확한 모델링을 위해 여러 개의 은닉층을 사용하는 것이 필요하며, 여러 은닉층을 가지는 신경망을 심층 신경망 (DNN)이라 하고, 본 논문에서는 심층 신경망을 사용하여 장르를 분류한다.

훈련 데이터에서 특성 벡터를 구하고, 이를 DNN에 입력하여 최상의 장르 분류 동작을 수행하도록 훈련하여 가중치와 바이어스로 구성된 DNN 파라미터를 구한다. 다음, 실제 오디오 신호가 입력되면 특성 벡터를 구하고 훈련된 DNN에 입력하여 출력층의 뉴런 값을 구한다. 본 논문에서는 3개 장르를 사용하므로 출력층은 3개의 뉴런으로 구성되고, 가장 큰 값을 가지는 뉴런에 해당하는 장르를 최종 장르로 결정한다.

본 논문에서는 DNN의 활성화 함수로 식 (3)의 시그모이드 (sigmoid) 함수를 사용하고, 비용함수로 식 (4)의 교차 엔트로피 (cross-entropy) 함수를 사용한다.

$$\sigma(z) = \frac{1}{1+e^{-z}},\ z = w\cdot x + b \quad (3)$$

$$C = -\frac{1}{N}\sum_x\sum_j \left[y_j \ln a_j^L + (1-y_j)\ln(1-a_j^L)\right] \quad (4)$$

여기서, N은 전체 학습 데이터의 수, x는 전체 학습 데이터, j는 출력층의 뉴런 인덱스, y_j는 뉴런 j에 대한 출력 목표값, a_j^L는 활성화 함수를 적용한 출력층의 최종 뉴런 값이다.

2. 특성 벡터

GMM 훈련의 연산량 감소를 위해 식 (1)에서 $\sum_i$ 이 대각 행렬이 되어야 하며, 이는 특성 벡터 성분끼리 직교하는 것을 의미한다. 그에 따라 대부분의 GMM 훈련은 특성 벡터가 직교성을 가진다고 가정하고 $\sum_i$ 를 대각 행렬로 제한하며, 만일 직교성이 약한 특성 벡터를 GMM으로 모델링

하면 성능이 저하된다. 따라서 오디오 특성을 GMM으로 모델링 할 때 직교성이 강한 MFCC가 널리 사용된다. MFCC는 mel-필터 출력 신호 f_i를 구한 후, 식 (5)과 같이 f_i에 DCT를 적용하여 구한다[7].

$$MFCC_k = \sum_{i=1}^{23} f_i \times \cos\left(\frac{\pi \times k}{23}(i-0.5)\right), 0 \le \ k \le \ 12 \quad (5)$$

장르 분류를 위한 기존 방법은 장르별 차별성을 가지는 스펙트럼 분포, 무게중심, 변화량 등을 MFCC와 결합하여 특성 파라미터로 사용한다[2,3,5]. 마지막으로, 여러 개의 연속된 프레임을 결합하여 긴 시간에 해당하는 texture 프레임을 정의하고, texture 프레임에서 MFCC와 각 특성 파라미터 값의 평균 (mean)과 분산 (variance)을 구하여 해당 texture 프레임의 특성 벡터를 정의한다. 이렇게 구한 특성 벡터를 각 장르별로 모델링 하여 장르를 분류한다. DNN은 GMM과 다르게 특성 벡터에 대한 특별한 요구 조건이 없으므로 MFCC를 포함하여 다양한 형태의 특성 벡터를 사용할 수 있다.

Ⅲ. 제안하는 오디오 장르 분류 방법

1. 스펙트로그램 기반의 특성 벡터

DNN은 특성 벡터에 대해 특별한 조건을 요구하지 않으므로 GMM에서 사용하였던 MFCC를 사용하지 않아도 된다. 따라서 본 논문에서는 MFCC보다 더 일반적인 특성인 스펙트로그램 기반의 특성 벡터를 제안한다. 또한, DNN은 시간 진행에 따른 특성을 모델링 하는데 한계가 있으므로 스펙트로그램의 시간 통계 특성으로부터 최종 46차 특성 벡터를 생성한다.

입력 신호의 특성 벡터를 추출하는 과정은 그림 2와 같다. 먼저, 그림 2(a)와 같이 샘플링 주파수 22.05kHz를 가지는 입력 신호 $x_k(n)$에 512-샘플 프레임 단위로 DFT를 적용하여 각 프레임의 스펙트럼을 구하고, mel-필터와 로그 연산을 적용하여 대역별 필터 출력 $f_{b,k}$ 를 구한다. 여기서, b는 mel 스케일 대역 인덱스이고 k는 프레임 인덱스이다. 다음, 그림 2(b)와 같이 43개 프레임을 연결하여 512 × 43 = 22,016 샘플 ≈ 1초 길이의 texture 프레임을 정의하고,

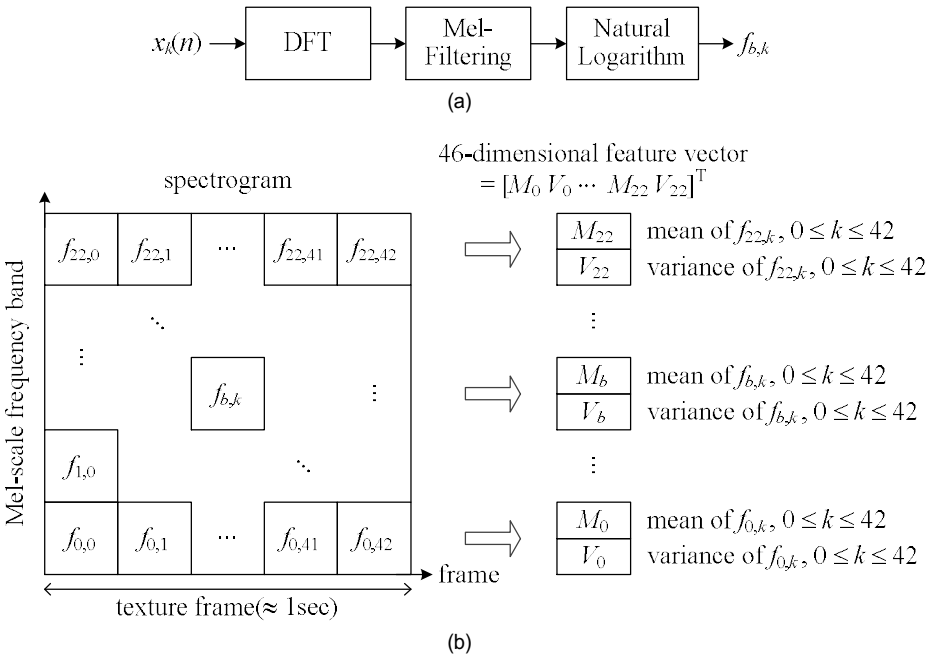

그림 2. 입력 신호의 특성 벡터를 구하는 과정. (a) 입력 신호의 스펙트로그램을 구하는 과정. (b) 스펙트로그램으로부터 특성 벡터를 구하는 과정
Fig. 2. Procedure of computing feature vector of input. (a) Computation of spectrogram. (b) Computation of feature vector from spectrogram

각 대역 b에 대하여 $f_{b,k}$, $0 \le k \le 42$ 의 평균 M_b과 분산 V_b를 구하여 46차 특성 벡터 $[M_0\ V_0\ \cdots\ M_{22}\ V_{22}]^T$를 최종 구한다.

2. 심층 신경망 훈련

DNN 훈련에서 가중치 초기값에 따라 최종 훈련 결과가 다르며, 특히 심층 구조에서는 가중치 초기값이 잘못 설정되면 학습이 매우 느리게 진행되거나 또는 가중치 값이 잘못된 값에 고정되는 문제가 발생할 수 있다. 따라서 DNN 훈련 성능을 향상시키고 그로부터 장르 분류 성능을 향상시키기 위해 가중치 초기값을 설정하는 체계적인 방법이 필요하며, 본 논문에서는 이를 위해 RBM (restricted Boltzmann machine)을 사용하다[8].

RBM은 가시층 (visible layer)과 은닉층의 두 층으로 구성된 네트워크를 사용하는 비지도 (unsupervised) 학습 방법이다. 그림 3이 RBM의 동작 구조를 나타내며, 가시층 입력 v, 가중치 w, 바이어스 b_1로부터 은닉층 h 값을 구한다. 이렇게 구한 은닉층 h 값에 w와 b_2를 적용하여 다시 가시층을 복원하는데, 초기 가중치 값은 임의의 값이므로 복원된 v'는 최초 입력 v과 다르게 된다. 다시, v'에 w와 b_1를 적용하여 h'를 구하고, 식 (6)의 갱신 식에 따라 w 를 계속 갱신한다.

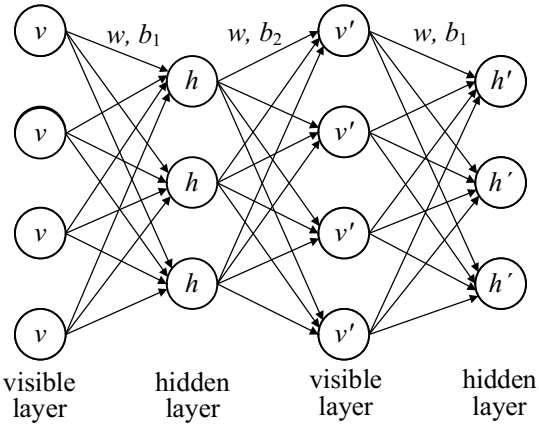

그림 3. RBM의 동작 진행 구조
Fig. 3. Structure of RBM processing

$$w_{t+1} = w_t + \epsilon\{vh^T - v'h'^T\} \tag{6}$$

그 결과 입력 v 를 모델링 하는 가중치가 구해지고, 해당 값을 DNN 첫 층의 가중치 초기값으로 결정한다. 심층 구조에서는 앞에서 구한 은닉층을 새로운 가시층으로 설정하여 다시 RBM을 통해 새로운 가중치를 구하고, 이 값은 DNN 두 번째 층의 가중치 초기값이 된다. 이과 같이 순차적으로 RBM을 수행하여 모든 층의 가중치 초기값을 정한다.

DNN 훈련은 역전파 (back propagation) 알고리즘으로 gradient를 구하고, 식 (7)에 따라 가중치를 갱신하는 과정으로 진행된다.

$$w_{t+1} = w_t - \eta\frac{\partial C}{\partial w_t} \tag{7}$$

여기서 η는 학습 속도를 결정하는 변수인 학습률 (learning rate)이다.

DNN 학습이 진행될수록 학습 데이터의 고유 특성에 과도하게 적응하여 잘못 훈련되는 과적응 (overfitting) 문제가 발생하며, 그 결과 학습 데이터에 포함되지 않는 데이터에 대해서는 장르 분류 성능이 저하된다. 과적응 문제를 해결하기 위한 가장 좋은 방법은 학습 데이터양을 증가시켜 학습 데이터의 범용성을 최대한 확보하는 것이다. 그러나 현실적으로 충분한 학습 데이터를 확보하지 못하므로 훈련 방법을 변경하여 과적응 문제를 해결해야 한다.

본 논문에서는 두 가지 방법을 적용하여 과적응 문제를 해결한다. 첫 번째 방법은 L2 정규화이며, 식 (8)과 같이 원 비용함수 C_0에 가중치 항이 추가된 새로운 비용함수를 정의하여 가중치 값이 비정상적으로 커지는 것을 방지하며, λ는 정규화 파라미터이다. L2 정규화를 적용하면 식 (9)와 같은 가중치 갱신 식이 얻어진다.

$$C = C_0 + \frac{\lambda}{2N}\sum_w w^2 \tag{8}$$

$$w_{t+1} = \left(1 - \frac{\eta\lambda}{N}\right)w_t - \eta\frac{\partial C_0}{\partial w_t} \tag{9}$$

두 번째 방법은 은닉층에 drop-out을 적용하는 것이다[9]. Drop-out은 비용함수를 변경하는 L2 정규화와는 다르게 신경망 자체를 변경한다. 각 학습 데이터에 대한 학습 과정에서 각 은닉층의 뉴런을 일정 비율 무작위로 삭제하여 학습

을 진행한다. Drop-out을 사용하면 학습 과정에서 신경망 구조가 계속해서 변하므로, 마치 여러 개의 학습된 신경망의 평균을 구하는 효과를 얻을 수 있고, 이로 인해 특정 값이 신경망에 과도하게 미치는 영향을 감소시켜 과적응을 방지한다.

3. DNN 매개 변수 결정

DNN 설계에 필요한 매개 변수 (hyper-parameter)는 신경망 구조를 결정하는 변수와 학습 과정을 결정하는 변수로 구성된다. 신경망 구조를 결정하는 변수는 은닉층의 수와 각 은닉층의 뉴런 수이다. 신경망이 커지면 필요한 가중치 수가 많아지고 초기화와 과적응 문제가 발생하여 학습이 올바르게 진행되지 않는다.

학습 과정을 결정하는 변수에는 학습률 η, epoch 수, drop-out 비율, mini-batch 크기, L2 정규화 변수 λ 등이 있다. Epoch는 반복 학습 횟수를 나타내며 모든 학습 데이터를 1회 학습시키는 것을 1 epoch라 하고, drop-out 비율은 각 은닉층에서 제외할 뉴런의 비율을 나타낸다. Mini-batch는 한 번의 갱신을 하기 위해 입력하는 훈련 데이터양을 의미하며, mini-batch 크기가 커지면 학습량이 감소해 연산량이 줄어드는 효과를 얻는다.

DNN 매개 변수를 구하는 이론적 방법은 없으며, 실험을 통하여 주어진 동작의 성능을 분석하여 결정해야 한다. 본 논문에서는 다양한 매개 변수 조합에 대한 반복적인 장르 분류 성능 분석을 통하여 최상의 매개 변수를 결정하였다.

3.1 신경망 구조

표 1은 심층 신경망 구조에 따른 장르 분류 성능을 나타내며, mini-batch 크기는 1, epoch는 1000번, 학습률은 0.07을 사용하였고, 신경망 구조에 따른 성능의 변화만을 보여주기 위해 drop-out과 L2 정규화는 제외하였다. 총 3개의 은닉층을 사용하며, 입력층과 출력층의 뉴런은 각각 46개와 3개이다. 은닉층의 뉴런이 많아질수록 가중치 수가 급증하기 때문에 초기화와 과적응에 의하여 오히려 성능이 저하될 수 있다. 본 논문에서는 표 1의 결과를 바탕으로 [120, 45, 30]을 최종 DNN 구조로 결정하였다.

3.2 학습률

그림 4가 학습률 η의 변화에 따른 장르 분류 성능의 차이를 보여준다. Mini-batch 크기는 1, Epoch는 1000번, L2 정규화 매개변수 λ는 0.01로 설정하고 오직 η의 변화에 따른 성능을 비교한다. η가 작으면 학습이 느리게 진행되고 너무 크면 과도한 갱신을 수행하여 잘못된 학습을 수행하게 된다. 본 논문에서는 그림 4의 결과에 따라 $\eta = 0.07$을 사용한다.

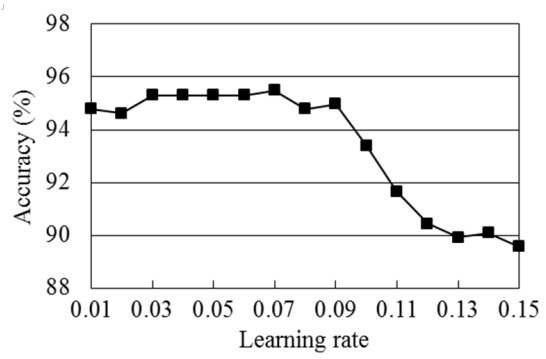

그림 4. 학습률 (learning rate)에 따른 장르 분류 성능
Fig. 4. The genre classification accuracy as a function of learning rate

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

# of Hidden Layers	# of Neurons in Each Hidden Layer	# of Parameters		Genre Classification Accuracy (%)
		Weights	Biases	
3	[60, 45, 30]	6,900	138	93.40
3	[90, 45, 30]	9,630	168	93.57
3	[120, 45, 30]	12,360	198	93.75
3	[150, 45, 45]	15,090	228	93.23
3	[180, 60, 30]	17,820	258	93.40

3.3 Epoch

표 2는 epoch 수에 따른 장르 분류 성능을 보여준다. DNN 훈련이 반복될수록 장르 분류 성능이 증가하다가 일정 epoch 수 이후에는 성능 증가가 미비한 포화상태(saturation)가 되는 것을 확인할 수 있다. 이 상태 이후에는 훈련을 계속해도 추가 학습이 거의 발생하지 않고, 오히려 과적응 문제가 발생할 수 있다. 따라서 실험을 통해 성능 향상이 없는 포화상태에 도달하는 epoch 수를 구하여 최종 epoch 수로 사용해야 한다.

표 2. Epoch 수에 따른 장르 분류 성능
Table 2. The genre classification accuracy as a function of the number of epochs

Number of Epochs	Genre Classification Accuracy (%)
300	91.66
400	93.23
500	94.60
1000	95.66
1500	95.48

3.4 Drop-out 비율

그림 5가 drop-out 비율에 따른 장르 분류 성능을 보여준다. 가중치 갱신은 역전파 방식에 따라 진행되기 때문에 영향력이 가장 큰 마지막 3번째 은닉층에 drop-out을 적용하는 것이 가장 효과적인 것을 확인하였다. 또한, 3번째 은닉층에서 drop-out을 강하게 적용하기보다는 10%를 적용했을 때 가장 우수한 성능을 가지고, drop-out으로 인해 과적응 현상이 감소하여 drop-out을 적용하지 않을 때 (drop-out 비율 0%)에 비하여 성능이 향상되는 것을 확인할 수 있다.

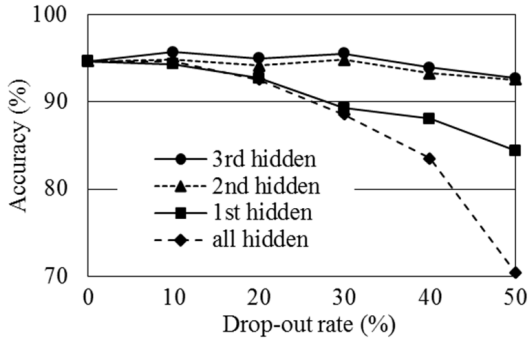

그림 5. Drop-out 비율에 따른 장르 분류 성능
Fig. 5. The genre classification accuracy as a function of drop-out rate

Ⅳ. 성능 분석

성능 평가에 사용된 오디오 데이터는 실제 TV 방송에서 획득한 음원이고, 데이터 길이는 각 장르별로 32분이다. 본 논문의 목표인 1초 단위의 온라인 장르 분류를 위해 수시로 장르가 변하도록 오디오 데이터를 구성하였고, 전체 오디오 데이터의 90%를 무작위로 선택하여 학습 데이터로 사용하고, 나머지 10%는 실험 데이터로 사용하여 성능을 평가한다. 1초마다 특성 벡터가 추출되므로 각 장르별로 1920개의 데이터를 사용하고, 그중 학습 데이터로 1728개, 실험 데이터로 192개를 사용한다.

MFCC와 GMM을 사용하는 기존의 장르 분류 방법의 성능을 측정하여 본 논문에서 제안하는 방법의 성능과 비교한다. 식 (5)의 $MFCC_k,\ 1 \le k \le 5$의 평균, 분산, 그리고 9개의 추가 스펙트럼 특성값으로 구성된 19차 특성 벡터를 사용한다[2]. GMM은 20개의 가우시안 컴포넌트를 사용하며, 학습 알고리즘인 EM 알고리즘의 반복 학습 횟수는 200번, 오차 허용 범위는 0.001로 설정했다.

제안 방법에서 다양한 매개 변수에 대한 성능을 분석하여 최종 심층 신경망 구조와 훈련 방법을 결정하였다. 은닉층의 뉴런 수는 [120, 45, 30], 학습률은 0.07, mini-batch 크기는 1, 정규화 매개 변수는 0.01, drop-out 비율은 세 번째 은닉층에 10%로 설정하였다. Epoch는 충분한 학습을 하면서 과적응을 방지하기 위해 1000번으로 설정하였다.

기존 방법과 제안 방법의 장르 분류 성능은 표 3과 같다. 기존 방법의 평균 장르 분류 정확도는 89.58%이며, music을 effect로 잘못 분류하는 비율이 7.81%이고 effect를 music으로 잘못 분류하는 비율이 13.02%이며, 이와 같이 상호 오분류 비율이 매우 높은 문제점이 나타난다. 제안 방법의 평균 장르 분류 정확도는 95.66%이고 기존 GMM 기반 방법보다 크게 향상된 것을 알 수 있다. 특히, music과 effect 사이를 서로 잘못 분류하는 비율이 뚜렷하게 감소한 결과를 얻는다.

이상의 성능 분석 결과에 의하면, 제안 방법을 사용하면

기존 방법보다 향상된 장르 분류 성능을 얻고, 기존 방법에서 발생하였던 music과 effect 사이를 서로 잘못 분류하는 문제점이 감소하고 각 장르에 대한 분류 정확도가 균등하게 되는 것을 확인할 수 있다. 이는 제안 방법이 사용하는 특성 벡터와 DNN이 기존 방법보다 music과 effect 사이의 우수한 차별화 모델링을 하기 때문으로 판단한다.

표 3. 기존 방법과 제안 방법의 장르 분류 정확도 (%)
Table 3. The genre classification accuracy (%) of the conventional method and the proposed method

Estimated / True	Conventional Method			Proposed Method		
	Speech	Music	Effect	Speech	Music	Effect
Speech	95.31	3.13	1.56	97.40	1.04	1.56
Music	3.13	89.06	7.81	1.56	94.27	4.17
Effect	2.6	13.02	84.38	0.52	4.17	95.31

V. 결 론

본 논문에서는 심층 신경망 기반의 온라인 오디오 장르 분류 방법을 제안하였다. 제안한 방법은 입력 오디오 신호를 1초 단위로 분석하여 speech, music, effect의 3 장르로 분류하며, 실시간으로 오디오 장르가 변하는 TV 방송에 적용 가능하고, 기존의 오프라인 분류 방법과 차별성을 가진다. 또한, 기존 GMM 기반의 분류 방법은 직교성 제약에 따라 MFCC를 사용하지만, 제안한 방법은 가장 일반적인 오디오 정보 표현 방식인 스펙트로그램 기반의 특성 벡터를 사용한다. 그에 따라 제안 방법은 특정 모델링 방법에 국한되지 않고 범용성을 확보할 수 있으며 다른 인식 모듈과의 통합을 가능하게 한다.

심층 신경망의 다양한 매개 변수에 대한 성능 분석을 통하여 최종 매개 변수를 결정하였고, 실제 TV 방송 신호에 대한 장르 분류 성능을 측정하였다. 제안 방법은 각 장르에 대하여 기존 방법보다 우수한 성능을 제공하며, 특히 기존 방법에서 문제가 되었던 music과 effect 사이의 오분류 비율을 뚜렷하게 감소시킨다.

참 고 문 헌 (References)

[1] Daeyoung Jang, Jeongil Seo, Yong Ju Lee, Jae-hyoun Yoo, Taejin Park and Taejin Lee, "A Study on Realistic Sound Reproduction for UHDTV," Journal of Broadcast Engineering, vol 20, no. 1, pp. 68-81, Jan. 2015.

[2] 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.

[3] Tao Feng, "Deep learning for music genre classification," private document.

[4] Jung-Sung Lee and Hyoung-Gook Kim, "Background Music Identification in TV Broadcasting Program Algorithm using Audio Peak Detection," Proc. of 2013 Korean Institute of Broadcast and Media Engineers Summer Conference, pp. 34-35, Jun. 2013.

[5] Z. Kons and O. Toledo-Ronen, "Audio event classification using deep neural networks," Proc. of Interspeech, pp. 1482-1486, 2013.

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

[7] 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.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," Science, vol. 313, pp. 504-507, Jul. 2006.

[9] N. Srivastava, G. Hinton, A. Krizhevsky and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Journal of Machine Learning Research, 15(1), pp. 1929-1958, Jun. 2014.

저 자 소 개

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

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

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