JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 1, pp. 183-190, Feb. 2017

깊은 신경망을 이용한 오디오 이벤트 검출

임 민 규 · 이 동 현 · 박 호 성 · 김 지 환*
서강대학교 컴퓨터공학과

Audio Event Detection Using Deep Neural Networks

Minkyu Lim · Donghyun Lee · Hosung Park · Ji-Hwan Kim*
Department of Computer Science and Engineering, Sogang University, Seoul 04107, Korea

[요 약]

본 논문에서는 깊은 신경망을 이용한 오디오 이벤트 검출 방법을 제안한다. 오디오 입력의 매 프레임에 대한 오디오 이벤트 확률을 feed-forward 신경망을 적용하여 생성한다. 매 프레임에 대하여 멜 스케일 필터 뱅크 특징을 추출한 후, 해당 프레임의 전후 프레임으로부터의 특징벡터들을 하나의 특징벡터로 결합하고 이를 feed-forward 신경망의 입력으로 사용한다. 깊은 신경망의 출력층은 입력 프레임 특징값에 대한 오디오 이벤트 확률값을 나타낸다. 연속된 5개 이상의 프레임에서의 이벤트 확률값이 임계값을 넘을 경우 해당 구간이 오디오 이벤트로 검출된다. 검출된 오디오 이벤트는 1초 이내에 동일 이벤트로 검출되는 동안 하나의 오디오 이벤트로 유지된다. 제안된 방법으로 구현된 오디오 이벤트 검출기는 UrbanSound8K와 BBC Sound FX자료에서의 20개 오디오 이벤트에 대하여 71.8%의 검출 정확도를 보였다.

[Abstract]

This paper proposes an audio event detection method using Deep Neural Networks (DNN). The proposed method applies Feed Forward Neural Network (FFNN) to generate output probabilities of twenty audio events for each frame. Mel scale filter bank (FBANK) features are extracted from each frame, and its five consecutive frames are combined as one vector which is the input feature of the FFNN. The output layer of FFNN produces audio event probabilities for each input feature vector. More than five consecutive frames of which event probability exceeds threshold are detected as an audio event. An audio event continues until the event is detected within one second. The proposed method achieves as 71.8% accuracy for 20 classes of the UrbanSound8K and the BBC Sound FX dataset.

Key word : Audio event detection, Deep neural network, Feed forward neural network

색인어 : 오디오 이벤트 검출, 깊은 신경망, Feed-forward 신경망

http://dx.doi.org/10.9728/dcs.2017.18.1.183

Received 19 January 2017; **Revised** 16 February 2017
Accepted 25 February 2017

***Corresponding Author; Ji-Hwan Kim**

Tel: +82-2-705-8924
E-mail: kimjihwan@sogang.ac.kr

Ⅰ. 서 론

사용자의 미디어 사용 분포가 점차 TV/영화 등 전문적인 대중미디어에서 UCC등 개인 미디어로 옮겨가는 추세에 있다[1]. 그러나 아직까지는 미디어 분류의 대부분을 메타데이터에 의존하고 있다. 이에 따라 유수의 글로벌 SW 기업들은 영상을 이용한 컨텐츠 분석을 통해 사용자의 맞춤지식을 생성하는 연구를 활발히 진행하고 있다. 동영상 컨텐츠의 의미를 분석하기 위해서는 해당 영상에 포함된 소리 이벤트를 검출하는 기술은 필수적이다. 기존의 소리 이벤트 인식의 경우 오디오 시그널로부터 spectal flux, ZCR (zero crossing rate), band periodicity 등 다양한 특징을 추출하여 해당 특징의 성능을 검증하는 연구와, GMM (gaussian mixture model), 규칙기반 (rule-based) 등 분류기에 대한 연구가 주를 이루었다[2]-[4]. 하지만 대부분의 연구에서는 소리 이벤트의 클래스 수가 음악/음성/기타소리 등 제한적이었다.

DNN (deep neural network)은 다양한 머신러닝 분야에서 기존의 방법보다 높은 성능 보이는 기술로서 주목 받고 있다. DNN은 2개 이상의 은닉층으로 구성된 인공 신경망으로서 단일 은닉층의 인공 신경망보다 더 복잡하고 비선형적인 학습 경계를 구분 지을 수 있어 분류 문제에 있어 더 좋은 성능을 얻을 수 있다. 다만 DNN의 수많은 파라미터를 추정하는 데에 있어서 많은 자료량이 필요하고, 높은 연산량이 요구되어 왔다. 최근 빅데이터의 발전, 하드웨어 기술의 발전으로 다양한 응용 분야에서 DNN을 적용하고 있다.

DNN은 이미지 분류 및 음성인식에 적용되어 괄목할만한 성능향상을 보였으나, 소리 이벤트 검출에 대한 적용은 그 사례가 많지 않다. 본 논문에서는 DNN을 이용한 오디오 이벤트 분류기로부터 오디오 이벤트 검출 방법을 제안한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 오디오 이벤트 인식을 위한 기존 연구들에 대하여 서술하고, 3장에서는 DNN을 이용한 오디오 이벤트 검출 방법에 대해 서술한다. 4장에서는 DNN을 이용한 오디오 이벤트 검출기에 대한 실험을 진행한다. 5장에서는 결론을 맺는다.

Ⅱ. 관련 연구

DBN (deep belief network)을 음악가 및 음악 장르 구분 분류 문제에 적용한 연구가 있었다[5]. DBN은 정답 스크립트가 존재하지 않는 많은 학습자료를 이용하여 은닉층을 순차적으로 학습시킨 후에, 소량의 정답 스크립트가 존재하는 자료를 이용하여 최종적으로 출력층을 학습시키는 방법이다. DBN의 은닉층 학습은 RBM (restricted boltzmann machine)을 이용하여 무감독 학습시킨 은닉층을 층층이 쌓는다. 최종적으로 출력층에 대해서에만 소량의 정답 스크립트가 존재하는 학습 자료를 이용하여 감독 학습을 수행한다. 실험 결과에 따르면 5개의 음악 장르 구분 실험에 대하여 약 73%의 분류 성능을 나타내었고, 음악가 분류에 적용 하였을 경우, 4개의 음악가 분류 실험에서는 약 80%의 분류성능을 나타내었다. 이 연구는 정답 스크립트가 존재하지 않는 다량의 학습자료를 이용할 수 있다는 장점을 보였다. 그러나 분류 실험의 클래스가 적다는 한계가 있다.

기존의 SVM (support vector machine)과 DNN과의 차이를 분석 기술한 연구가 있었다[6]. 이 연구에서 SVM은 shallow architecture로, DNN은 deep architecture로 구분 짓는다. SVM의 경우 커널을 통하여 차원을 줄이고, 클래스 분포를 구분짓도록 하는 선을 긋는 방식이며 이는 하나의 은닉층으로 구성되는 인공신경망의 형태로 볼 수 있다. 반면 다중 은닉층을 가지는 DNN의 경우 비선형 경계를 계층적으로 쌓기 때문에 SVM이나 단일 은닉층으로 구성된 인공신경망보다 복잡한 decision boundary를 생성할 수 있다. 한 예로, 단일 은닉층의 인공신경망은 XOR 구분 문제를 모델링 할 수 없지만, 다중 은닉층의 인공신경망으로는 구분이 가능하다. 다만 DNN 파라미터를 학습하기 위해서는 많은 자료가 필요하다.

HMM-SVM 기반 오디오 이벤트 분류기에 대한 연구가 있었다[7]. 15개의 오디오 이벤트 분류기를 MFCC (mel frequency cepstral coefficient), PLP (perceptual linear prediction), ZCR (zero crossing rate) 등 다양한 특징벡터 조합으로 학습시켰으며, 특징벡터의 조합에 따라 인식 성능 차이가 발생 할 수 있음을 보였다. 토크쇼, 뉴스, 영화, 다큐멘터리에서 등장하는 소리 이벤트들로 부터 직접 정답 스크립트를 생성하여하여 이벤트를 검출하였다. 실험 결과 PLP 특징벡터를 사용한 경우의 성능이 가장 높았다.

스포츠 중계 영상에서 소리 정보를 이용하여 다섯 종류의 이벤트 분류를 DBN을 사용하여 실험한 연구가 있다[8]. 다섯 개의 이벤트는 관중소리, 해설, 관중소리+해설, 흥분된 해설이며, DBN 모델을 학습시켜 SVM 모델을 사용한 분류기와 성능을 비교 평가하였다. 실험 결과 DNN보다 SVM의 성능이 조금 더 높았으며, 많지 않은 양의 학습 자료로부터 학습된 결과로 분석되었다.

한 샘플에 다수의 이벤트가 존재하는 소리 이벤트 시퀀스 분류에 대한 연구가 있었다[9]. 기존의 소리 이벤트 분류의 경우 하나의 샘플에는 하나의 이벤트만 있는 것으로 가정되었지만 실제로는 다양한 이벤트가 존재하는 경우가 있기 때문에 이벤트 시퀀스를 분류하였다. 각각의 소리 이벤트들을 GMM을 기반으로 모델링을 한 후, 소리 이벤트 시퀀스를 분류하기 위해 3-state 기반 HMM (hidden markov model)을 사용하였다. 이는 소리 이벤트가 일련의 순서대로 나타나는 경우 높은 성능을 보이는 반면 정답 스크립트가 존재하는 학습 자료를 수집하기가 어려운 단점이 있다.

RBM을 이용하여 traffic, music, crowd, applause, 네 종류의 이벤트에 대해서 구분하는 연구가 있었다[10]. 이 연구에서 은닉층의 경우 입력 특징벡터에 대한 출력 특징벡터를 생성하도록 RBM을 학습시킨 후, 최상위 출력층에 대해서만 FFNN을 구성하여 소리 이벤트가 출력되도록 하였다. RBM을 이용한 깊은

신경망을 SVM 및 GMM과 성능 평가를 하였고, RBM의 분류 성능이 GMM, SVM 보다 높게 나왔다. 하지만 이 연구 역시 소리 이벤트의 수가 적다는 한계가 있다.

DNN기반의 오디오 이벤트를 프레임 단위로 분류 연구가 있었다[11]. 매 프레임 마다 소리 이벤트를 분류하는 방법을 제안하였고, 하나의 샘플 입력을 분류 할 시에는 모든 프레임에 대한 각 이벤트별 확률값을 더해 가장 높은 확률값을 가지는 이벤트를 분류 결과로 출력 하였다. 오디오 이벤트의 개수는 10개이었으며, 샘플 단위의 오디오 이벤트 분류 결과는 70% 후반의 성능을 보였다. 그러나 프레임 단위 오디오 분류기로는 긴 영상에서 일부를 차지하고 있는 오디오 이벤트들을 검출하지는 못하는 문제가 있다. 본 연구에서는 프레임 단위 오디오 이벤트 분류 결과로부터 오디오 이벤트 구간을 검출하는 방법을 제안한다.

Ⅲ. DNN 기반 오디오 이벤트 검출기

제안한 DNN 기반 오디오 이벤트 검출기는 세 가지 모듈로 구성되어 있다. 각 모듈은 오디오 특징 추출기, 프레임 단위 오디오 이벤트 분류기, 그리고 오디오 이벤트 검출기로 구성되어 있다. 그림 1.은 전체적인 오디오 이벤트 검출 과정을 나타낸다. 오디오 특징 추출기 및 프레임 단위 오디오 이벤트 분류기는 [11]에서 제안된 방법을 취한다. 3.1절에서는 오디오 특징 추출기에 대해서 기술하며 3.2절에서는 DNN을 이용한 프레임 단위 오디오 이벤트 분류기에 대해서 기술한다. 3.3절에서는 프레임 단위 오디오 이벤트 분류 결과로부터 오디오 이벤트 구간을 검출하는 방법을 기술한다.

3-1 오디오 특징 추출기

오디오 특징 추출기는 주어진 입력신호에 대한 특징 벡터열을 생성하고, 이 벡터열은 DNN의 입력층에 사용된다. 입력 오디오 신호는 16kHz로 샘플링 되며, 샘플당 2byte이고, mono channel로 구성된다. 20ms 크기의 해밍 윈도우가 입력 신호에 대하여 10ms 단위로 이동하면서 STFT (short time fourier transform) 이 수행된다. 모든 윈도우에 대하여 mel-scale로 증가하는 삼각 bin을 씌워 각각의 주파수 에너지마다 가중치를 곱하여 삼각 bin 별로 하나의 특징 값을 추출한다. 이 특징 값들을 하나의 벡터로 구성하여 FBANK (mel-scale filter bank) 특징벡터를 생성한다.

하나의 프레임으로부터 추출된 특징 벡터는 20 ms 길이에 해당되는 특징값으로서, 소리에 대하여 매우 짧은 구간의 특징만을 표현하며, 10ms 단위로 이동하면서 특징 벡터가 생성되기 때문에 동일 소리 내에서도 윈도우 이동에 따라 특징벡터의 변화가 매우 크다. 이러한 현상을 보정하기 위해 좌우 컨텍스트 정보를 반영한다. 현재 프레임 기준 좌/우 N개의 윈도우로부터 생성된 특징벡터를 하나의 벡터로 결합하고 이를 DNN 분류기 입력층에 사용한다. 컨텍스트에 반영되는 윈도우 크기 N에 따라 DNN 기반 분류기 입력층에 사용되는 벡터 크기가 달라진다.

3-2 프레임 단위 오디오 이벤트 분류기

본 절에서는 프레임 단위의 오디오 이벤트 분류 방법을 서술한다. 오디오 이벤트 분류기는 FFNN (feed forward neural network)으로 이루어져 있다. FFNN은 하나의 입력층과 하나

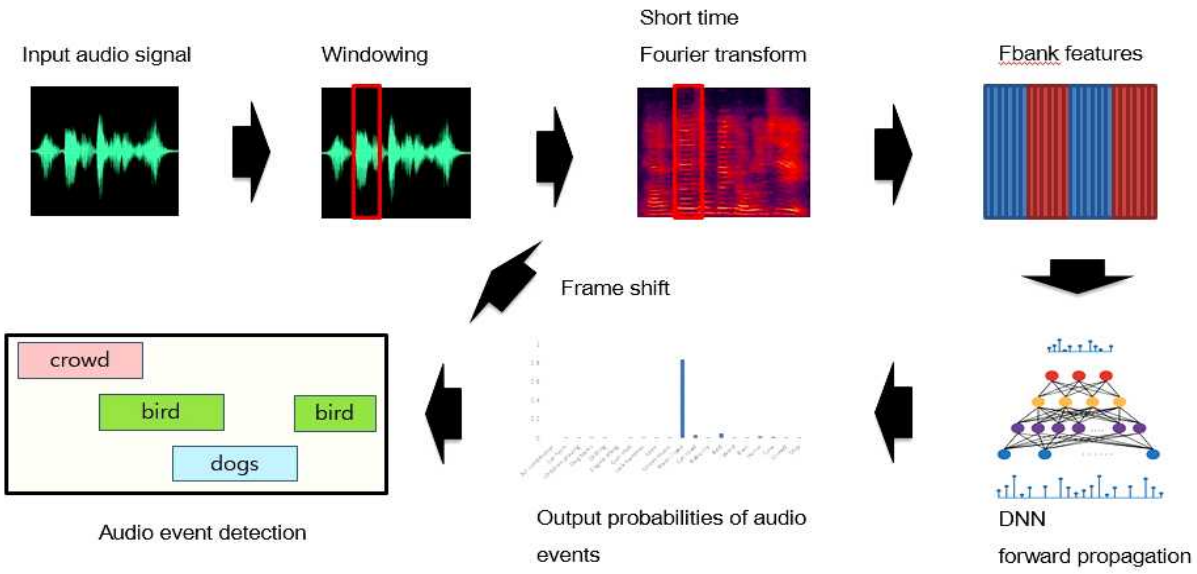

그림 1. 오디오 이벤트 검출 과정
Fig. 1. Audio event detection process

의 출력층, 그리고 둘 이상의 은닉층으로 구성되어 있다. 각 층은 weight와 bias값을 파라미터로 가지는 뉴론들로 구성되어 있다. 각 뉴론의 입력은 해당 뉴론의 아래층의 모든 뉴론들의 출력이 된다. 3.1절의 오디오 특징 추출기에서 생성된 특징벡터는 프레임 단위 오디오 이벤트 분류기를 구성하는 FFNN의 입력 $\overline{x}$가 된다. 입력 $\overline{x}$ 로부터 수식 (1)과 (2)를 통하여 FFNN의 출력층까지 계산 된다. 입력벡터 $\overline{x}$에 대한 출력층에서의 출력값은 해당 클래스에 대한 발생확률을 의미한다. 수식에서 사용된 기호는 <표 1>에서 정의한다.

$$a_i^k(\overline{x}) = b_i^k + \sum_{j=1}^{N^{k-1}} w_j^k h_j^{k-1}(\overline{x}) \; (1 \le k \le L+1) \tag{1}$$

$$h_i^k(\overline{x}) = g(a_i^k(\overline{x})) (1 \le k \le L+1) \tag{2}$$

$$\begin{aligned} P(y=i|\overline{x}, W, b) &= h_i^{L+1}(\overline{x}) \\ &= softmax(a_i^{L+1}(\overline{x})) \\ &= \frac{e^{a_i^{L+1}(\overline{x})}}{\sum_{j=1}^{N^{L+1}} e^{a_j^{L+1}(\overline{x})}} \end{aligned} \tag{3}$$

수식 (1)에서 $a_i^k(\overline{x})$는 입력 $\overline{x}$가 들어왔을 때 k번째 층의 번째 i번째 뉴론의 선활성함수의 결과이다. 해당 뉴론에서의 최종 출력 값은 선활성함수의 계산 결과에 활성 함수를 취한 결과가 된다. 은닉층의 활성함수는 주로 sigmoid, tanh 등의 비선형 함수가 사용된다. 비선형 활성 함수는 선형 활성 함수보다 복잡한 분류 경계를 표현할 수 있기 때문에 비선형 활성 함수를 사용한 모델의 성능이 더 높은 것으로 알려져 있다[12]. 최근에는 ReLU (rectified linear unit)가 비선형 활성함수로 많이 사용되고 있다[13]. ReLU 함수의 수식은 $g(x) = \max(0,x)$ 와 같다. 이는 선활성함수의 계산결과가 양수인지 음수인지에 따라 다음 층의 입력에 값을 전달 하는지에 대한 필터 역할을 하게 된다. 또한 sigmoid, tanh 등과 비교하여 연산이 간단하며, sigmoid, tanh를 사용함에 따라 발생하는 vanishing gradient 문제가 감소되는 장점이 있다. 출력층에서의 활성 함수는 softmax가 사용된다. 출력층의 최종 출력 값은 수식 (3)에 따라 계산된다. 모델 파라미터 W, b 와 입력 $\overline{x}$가 주어졌을 때 프레임단위 분류기는 각 클래스에 대한 발생확률을 출력한다.

파라미터 학습은 수식 (5)와 같은 NLL (negative log likelihood)을 loss function으로 정의하여 모든 자료에 대하여 수식 (4)의 loss가 최소가 되도록 한다. 학습은 stochastic gradient descent 방법을 따른다[14].

표 1. FFNN 수식 기호 정의
Table. 1. Notations in FFNN equation

Notation	Definition
$\overline{x}$	Input vector
L	Number of hidden layer
$\overline{w}^k$	Weight vector of k-th layer
w_i^k	Weight value of k-th layer, I-th neuron
W	Total weight matrix of FFNN
b	Bias
N^k	Number of neurons of k-th layer
$a()$	Preactivation function
$g()$	Activation function
$h_i^k(\overline{x})$	Output value of k-th layer, I-th neuron for input vector $\overline{x}$
y	Output class of classifier

$$E(\Theta, D) = NLL(\Theta, D) + \lambda \| \Theta \|_p^p \tag{4}$$

$$NLL(\Theta, D) = -\sum_{s=1}^{|D|} \log P(y^{(s)}|\overline{x}^{(s)}, \Theta) \tag{5}$$

$$\| \Theta \|_p = \left(\sum_{j=1}^{|\Theta|} |\Theta_j|^p \right)^{\frac{1}{p}} \tag{6}$$

수식 (4), (5)에서 D는 학습 자료로부터 추출된 특징 벡터열 집합을 나타내고, Θ는 앞에서 학습된 모든 weight 및 bias 파라미터 집합을 의미한다. j는 모든 파라미터의 뉴론 index를 표현한다. 수식 (6)에서 $\| \Theta \|_p$는 학습자료에 과적합 되는 것을 방지하기 위한 regularizer를 나타낸다.

테스트 샘플에 대해서 10ms 단위로 이동하는 모든 프레임에 대한 클래스별 출력 확률값이 계산되고, 모든 프레임에 대한 각 이벤트들의 발생 확률은 3절에서 기술하는 오디오 이벤트 검출기의 입력이 된다.

3-3 오디오 이벤트 검출기

프레임 단위의 인식 결과로부터 구간 검출은 그림 2.와 같이 이루어진다. 매 프레임마다 모든 오디오 객체에 대한 발생확률을 DNN forward propagation을 통하여 계산한다. 발생확률이 연속된 5개의 프레임 이상에서 기 학습된 threshold를 넘는 객체에 대해서 1차 구간 검출을 수행한다. 1차로 검출된 구간들이 1초 이내에 인접하여 있는 경우 해당하는 두 구간을 하나의 구간으로 합친다. 최종적으로 오디오 객체명과 시작시간 및 종료시간이 오디오 객체 검출 결과로 도출된다.

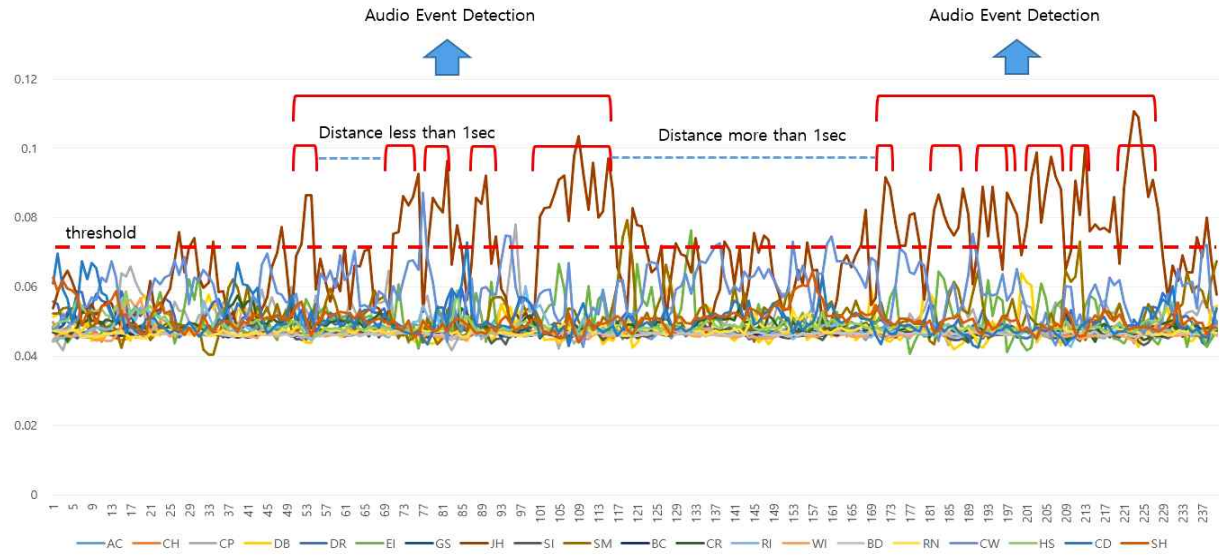

그림 2. 잭해머 소리에 대한 오디오 이벤트 구간 검출 예
Fig. 2. Example of event detection for jack hammer sound

Ⅳ. 오디오 이벤트 검출 성능 평가

4-1 학습 및 실험 자료

오디오 이벤트 분류 연구의 경우 대부분 별도로 자료를 수집하여 검증하기 때문에 공통된 평가 지표를 구하기 어렵다. 본 연구에서는 접근 가능한 자료 중 퍼스널 미디어에서 주로 나오는 오디오 이벤트를 포함하고 있는 학습 자료들을 선정하였다. 선정된 학습자료는 UrbanSound8K와 BBC Sound FX 이다. UrbanSound8K 와 BBC Sound FX에서 각 10개의 오디오 이벤트를 선정하였다.

1) UrbanSound8K

UrbanSound8K는 NYU에서 수집 배포한 오디오 이벤트 자료로서, www.freesound.org 로부터 수집한 오디오 자료를 이벤트 구간에 맞게 잘라 태깅한 자료이다[15]. 하나의 오디오 샘플은 4초 이하로 구성되어 있으며 총 8,732개의 파일로 제공된다. 10개의 오디오 이벤트 클래스를 가지며, 총 9시간 분량의 길이를 가진다. 표2는 UrbanSound8K dataset의 10개의 오디오 이벤트 종류를 나타낸다.

2) BBC Sound FX

BBC Sound FX 자료는 영화/TV방송 등에서 사용된 음향 이펙트 모음 자료로써 CD 40장 분량으로 구성되어 있다[16]. 모든 트랙에 대해서 직접 이벤트에 제한을 두지 않고 자유롭게 수동 태깅을 진행하였고, 태깅 결과 총 160가지의 오디오 이벤트가 존재 하였다. 이 중 빈도수 기준 상위 10개의 오디오 이벤트를 선정하여 학습 자료에 포함 하였다. 표2는 BBC Sound FX자료에서 선정된 10개의 오디오 이벤트를 보여준다.

4-2 실험 환경

모든 자료는 16bit-mono의 16kHz로 일괄 변환하였고 자료의 9/10은 학습에 사용되었고, 1/10은 평가에 사용되었다. 특징 벡터 추출은 음성인식 툴킷인 HTK[17] 를 사용하였고, DNN기반 프레임 단위 오디오 이벤트 분류기는 Tensorflow[18]를 사용하여 구현하였다. 오디오 이벤트 검출기는 python으로 구현하였다. DNN학습 시 optimizer는 SGD (stochastic gradient descent)를 사용하였고, minibatch 크기는 100 이다. 초기 learning rate은 0.01로 시작하였고, validation error가 30 회 동안 감소하지 않으면 learning rate을 10%씩 감소시켰다. 최소 learning rate은 0.001 이다. 수식 (4), (6)에서 $\lambda = 0.001$, $p = 2$로 설정 하였다.

표 2. UrbanSound8K dataset 과 BBC Sound FX에서 선정된 이벤트 종류
Table. 1. Event list of UrbanSound8K and BBC Sound FX dataset

Air conditioner	Engine idling	Baby cry	Rain drop
Car horn	Gun shot	Car road	Crowd
Children playing	Jack hammer	River	Cow
Dog barks	Siren	Wind	Horse
Drilling	Street music	Bird	Ship horn

표 3. 은닉층당 뉴론 수 변화에 따른 오디오 이벤트 검출 성능
Table. 3. Performance of audio event detection according to the number

Number of hidden layer	Number of neurons per hidden laer	Frame-level accuracy (%)	Audio event detection accuracy (%)
2	500	42.35	50.9
2	1,000	44.57	52.3
2	2,000	47.53	65.5
2	3,000	49.76	71.8

4-3 실험 결과

DNN 모델의 은닉층 당 뉴론 수를 다르게 적용하여 실험을 진행 하였다. 그에 대한 실험 결과는 <표 3>과 같다. 프레임 단위 인식률은 테스트 자료의 모든 프레임 별로 프레임 단위 오디오 이벤트 분류 결과의 정답률을 의미 한다. 오디오 이벤트 검출율은 하나의 샘플에 대해서 오디오 이벤트가 정상적으로 검출 되었을 경우 정답으로 판정하고, 모든 테스트 샘플에 대한 정답 판정된 샘플의 비율을 나타낸다.

2-은닉층 / 3,000-은닉층 당 뉴론 수를 사용하여 학습한 모델이 가장 높은 성능을 보였다. 10개의 클래스에 대한 오디오 이벤트 분류기를 구현한 기존 연구 대비 동일한 수준을 유지 하면서 오디오 이벤트 구간을 검출 함과 동시에 검출 오디오 이벤트를 10개 추가하여 20개를 달성 하였다.

Ⅴ. 결 론

본 연구에서는 프레임 단위 오디오 이벤트 분류 결과로부터 오디오 이벤트를 검출하는 방법을 제안하였고, 실험을 통하여 성능을 검증하였다. 은닉층은 2개를 사용하고 은닉층당 뉴론 수를 달리 하여 검출 성능을 측정하여 비교 분석 하였다. 실험 결과 뉴론 수가 3,000개의 오디오 이벤트 분류기가 71.8%의 최대 검출 성능을 보였다. 이는 프레임 단위 오디오 이벤트 분류 성능을 유지하면서 긴 구간의 동영상 내에서 특정 소리 구간을 검출 할 수 있는 장점을 가진다.

참고문헌

[1] K. Kim and H. Kim, “Scaling learning algorithms towards AI,” *Journal of Digital Content Society*, Vol. 14, No.4, pp.481-491, December, 2013.

[2] L. Lu, H. Jiang, and H. Zhang, “A robust audio classification and segmentation method,” in *Proceeding of ACM International Conference on Multimedia*, Ottawa, pp.203-211, 2001.

[3] M. Xu, N. Maddage, C. Xu, M. Kankanhalli, and Q. Tian, “Creating audio keywords for event detection in soccer video,” in *Proceeding of IEEE International Conference on Multimedia and Expo*, Baltimore: MD, pp.281-284, 2003.

[4] W. Cheng, W. Chu, and J. Wu, “Semantic context v detection based on hierarchical audio models,” in *Proceeding of ACM SIGMM International Workshop on Multimedia Information Retrieval*, Berkeley: CA, pp.109-115, 2003.

[5] H. Lee, P. Pham, Y. Largman, and Y. Ng, “Unsupervised feature learning for audio classification using convolutional deep belief networks,” in *Proceeding of Advances in Neural Information Processing Systems*, Vancouver, pp.1096-1104, 2009.

[6] Y. Bengio and Y. LeCun, “Scaling learning algorithms towards AI,” *Large-scale Kernel Machines*, Vol. 34, No.5, pp.321-360, August, 2007.

[7] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, “Non-speech audio event detection,” in *Proceeding of Internationa Conference on Acoustics, Speech and Signal Processing*, Taipei, pp.1973-1976, 2009.

[8] L. Ballan, A. Bazzica, M. Bertini, A. Bimbo, and G. Serra, “Deep networks for audio event classification in soccer videos,” in *Proceeding of International Conference on Multimedia and Expo*, Cancun, pp.474-477, 2009.

[9] T. Heittola, A. Mesaros, A. Eronen, T. Virtanen, “Scaling learning algorithms towards AI,” *EURASIP Journal on Audio, Speech, and Music Processing*, Vol.1, pp.1-13, January, 2013.

[10] K. Zvi, and T. Orith, “Audio event classification using deep neural networks,” in *Proceeding of INTERSPEECH*, Lyon, pp.1482-1486, 2013.

[11] M. Lim and J. Kim, “Audio Event Classification Using Deep Neural Networks,” *Phonetics and Speech Sciences*, Vol. 7, No. 4, pp.27-33, January, 2015.

[12] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, “An empirical evaluation of deep architectures on problems with many factors of variation,” in *Proceeding of International Conference on Machine learning*, Corvaliis: OR, pp.473-480, 2007.

[13] E. Dahl, N. Sainath, and E. Hinton, “Improving deep neural networks for LVCSR using rectified linear units and dropout,” in *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, Vancouver, pp.8609-8613, 2013.

[14] L. Bottou, *Advanced Lectures on Machine Learning,* Springer, pp. 146-168, 2004.

[15] J. Salamon, C. Jacoby, and J. Bello, “A dataset and taxonomy for urban sound research,” in *Proceeding of ACM International Conference on Multimedia*, Orlando: FL, pp.1041-1044, 2014.

[16] M. Slaney, “Semantic-audio retrieval,” in *Proceeding of International Conference on Acoustics, Speech and Signal Processing*, Orlando: FL, pp.1408-1411, 2002.

[17] S. Young, G. Evermann, M. Gales, and P. Woodland, *The HTK book (for HTK version 3.4),* Cambridge, U.K.: Entropic, 2006.

[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, Tensorflow: Large-scale machine learning on heterogeneous distributed systems, Available: https://www.tensorflow.org/

임 민 규 (Minkyu Lim)

2008년 : 서강대학교 (학사)
2010년 : 서강대학교 대학원 (석사)
2002년~2008년: 서강대학교 기계/컴퓨터공학 학사
2008년~2010년: 서강대학교 컴퓨터공학 석사
2012년~2015년: 휴맥스/아이큐브
2010년~현 재: 서강대학교 컴퓨터공학과 박사과정
※관심분야 : 음성인식, 오디오 컨텐츠 분석

이 동 현 (Donghyun Lee)

2013년 : 서강대학교 (학사)
2009년~2013년: 서강대학교 컴퓨터공학과 학사
2013년~현 재: 서강대학교 컴퓨터공학과 석박사 통합과정
※관심분야 : Speech Recognition using Deep Learning, Artificial Intelligence, Multimedia Content Search

박 호 성 (Hosung Park)

2016년 : 한동대학교 (학사)

2009년~2016년: 한동대학교 컴퓨터공학과 학사
2016년~현 재: 서강대학교 컴퓨터공학과 석사과정
※관심분야 : 음성인식

김 지 환 (Ji-Hwan Kim)

1996년 : KAIST (학사)
1998년 : KAIST (석사)
2001년 : University of Cambridge (박사)

1992년~1996년: KAIST 전산학 학사
1996년~1998년: KAIST 전산학 석사
1998년~2001년: University of Cambridge 컴퓨터공학 박사
2001년~2007년: LG 전자 책임연구원
20017년~현 재: 서강대학교 컴퓨터공학과 교수
※관심분야 : Spoken Multimedia Content Search, Speech Recognition using Cloud Computing and Dialogue Understanding