Journal of KIIT. Vol. 21, No. 6, pp. 93-100, Jun. 30, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.6.93

분절모델 기반의 소리 이벤트 검출

김수종*, 정용주**

Sound Event Detection based on Segmental Models

Soo-Jong Kim*, Yong-Joo Chung**

요 약

분절 모델은 프레임 단위 대신 세그먼트 단위의 유사도 스코어를 계산함으로써 시퀀스 데이터를 인식하는 방법이다. 음성인식이나 자연어 처리 분야에서 일정 부분 효과를 보이고 있는 점에 착안하여, 본 논문에서는 소리 이벤트 검출을 위하여 분절 모델을 사용하였다. 딥뉴럴네트워크 기반의 인코딩 모듈을 사용함으로써 로그-멜-필터뱅크 값을 분절적 모델에 적합하도록 변형하였다. 분절적 모델에서는 다양한 길이의 세그먼트와 소리의 종류를 전부 고려해야하기 때문에 많은 계산량이 소요되는 단점이 있는데, 이를 보완하기 위한 계산량이 적은 세그먼트 스코어링 함수를 채택하였다. 비용함수로써 marginal log loss 함수를 사용함으로써, 강전사(Strong label) 학습데이터를 사용할 필요가 없도록 하였다. DCASE 챌린지 2018 오디오 데이터를 이용한 소리 이벤트 검출 실험 결과, 제안된 분절 모델을 이용함으로써 기존의 방식에 비해서 우수한 F-score를 얻을 수 있었다.

Abstract

Segmental models compute likelihood scores in segment units instead of frame units to recognize sequence data. Motivated by the fact that they have shown some promising effects in speech recognition and natural language processing, we used segmental models for sound event detection in this paper. We modified the log-mel filterbank to make it suitable for the segmental models by using an encoding module based on deep neural networks. Since various lengths and types of sounds of segments should be considered, the segmental models have the drawback of high cost in computation. To solve this problem, we use a segment scoring function with less computation amount. We used a marginal log loss as the loss function to train the segment model without having to rely on strong labels. From the experimental results using DCASE Challenge 2018 audio data, we could obtain better F-score compared to the conventional methods.

Keywords
sound event detection, segmental models, segment encoding, marginal log loss

* 계명대학교 전자공학과 석사과정
- ORCID: https://orcid.org/0000-0001-7492-1134
** 계명대학교 전자공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-0060-1178

· Received: Mar. 23, 2023, Revised: Apr. 12, 2023, Accepted: Apr. 15, 2023
· Corresponding Author: Yong-Joo Chung
Dept. of Electronics Engineering Keimyung University, 704-701
Shindang-dong, Dalseo-gu, Daegu-si, 1000, Republic of Korea
Tel.: +82-53-580-5925, Email: yjjung@kmu.ac.kr

Ⅰ. 서 론

소리(오디오) 이벤트 검출(Sound event detection)은 특정한 소리가 발생하는지의 여부와 함께 발생 시점을 탐지하는 기법이다. 전통적으로는 GMM (Gaussian Mixture Model)이나 SVM(Support Vector Machine)을 기반으로 한 방식들이 많이 사용되어 왔으나[1][2] 최근에는 기존의 머신러닝 기반의 방식보다 뛰어난 성능을 보여주는 딥뉴럴네트워크(Deep neural network) 기반 방식들이 많이 사용되고 있다. 소리 이벤트 검출 기법은 감시나 도심지 잡음 해석, 멀티미디어 컨텐츠로부터의 정보 탐색, 헬스케어 모니터링 및 새소리 탐지 등의 다양한 분야에서 활용 가능하므로 향후 발전 가능성이 매우 높다고 판단된다[3]-[6].

소리 이벤트 검출을 위하여 많은 방식들이 제안되었으며, 최근에 이르러서는 CNN(Convolutional Neural Network)과 RNN(Recurrent Neural Network)을 결합한 방식인 CRNN(Convolutional Recurrent Neural Network)이 가장 대표적인 오디오 인식 기법으로 자리 잡았다[7][8].

그러나 신뢰할 만한 학습 데이터의 확보가 어렵다는 현실적인 문제 때문에 소리 이벤트 검출 성능은 최근까지도 그리 만족스럽지 못한 실정이다. 특히, 대용량의 오디오 데이터의 확보가 가능한 web 기반 오디오 데이터에는 많은 양의 레이블 오류가 존재한다[9]. 특히, 전사자(Annotator)의 착오나 소리들 간의 의미론적인 애매함으로 인한 다수의 레이블 오류가 존재하는데 이를 부정확(Incorrect) 레이블 오류라고 한다. 그리고 하나의 오디오 파일에 다수의 소리가 존재하거나 긴 묵음 구간이 소리와 함께 존재하는 경우, 소리들 간의 정확한 경계가 알려져 있지 않은 경우를 불완전(Incomplete) 레이블 또는 약전사 레이블(Weak label) 오류라고 한다[10].

이러한 두 가지 레이블 오류 중에서, 부정확 레이블 오류에 대해서는 그 대처 방안에 대한 연구가 상당히 진행되었다. 특히 data cleansing 처리, noise tolerant 학습 그리고 레이블 잡음 모델링 등의 기법들이 대표적인 방법으로써, 영상 및 음성인식 등에서 사용되어 어느 정도 효과를 보았다[11][12]. 그러나 불완전 레이블 오류는 오디오 데이터에 고유한 레이블 오류로써 이에 대한 연구는 거의 진행되지 않았다.

본 논문에서는 이와 같이 오디오 데이터에서 발생하는 레이블 오류 중 불완전 레이블 오류에 대한 대처 방법으로써 분절 모델을 제안하고자 한다. 분절 모델은 학습과정을 통해서 소리들을 구분 짓는 최적의 세그먼트(분할)를 찾아가는 방식을 적용하기 때문에 불완전 레이블 오류를 가진 오디오 데이터를 이용한 학습에 매우 유리하다. 분절 모델에서는 오디오 파일에 포함된 여러 개의 소리 및 묵음에 대해서 프레임 단위의 구간 정보(Ground truth)를 제공할 필요가 없으며, 전체 오디오 파일 에 대한 레이블 정보만 제공하면 되기 때문에 불완전 레이블 오류 학습데이터를 이용한 소리 이벤트 검출에 있어, 기존의 프레임 기반 방식에 비해서 유리하다.

본 논문의 구성은 다음과 같다. 2장에서는 소리 이벤트 검출을 위한 특징 추출 방법과 본 논문에서 사용된 분절 모델의 구조에 대해서 설명하며 3장에서는 분절 모델을 이용한 오디오 이벤트 검출에 관한 실험결과를 제시하고 4장에서 결론을 맺는다.

Ⅱ. 특징 추출과 분절 모델

2.1 로그-멜-필터뱅크 추출

오디오 이벤트 검출을 위한 분절 모델의 학습 및 디코딩(Decoding)를 위해서는 오디오 웨이브(Wave)로부터 특징 추출이 필요하며, 전체적인 추출과정은 그림 1에 나타나 있다. 오디오 웨이브는 16kHz의 비율로 샘플링 되었으며 41.5ms의 프레임 간격마다 STFT(Short-Time Fourier Transform)이 계산된다. STFT 값으로부터 64 band의 멜-필터뱅크 값을 전체 0에서 8,000Hz의 주파수 구간에서 구한 후, 이를 로그변환 함으로써 64차원의 로그-멜-필터뱅크 값이 매 프레임마다 얻어지게 된다. 프레임 간격이나 로그-멜-필터뱅크 값의 차원 등은 본 논문에서 사용된 DCASE 2018 오디오 데이터의 배포시 추천된 값을 사용하였다[8]. 로그-멜-필터뱅크 값은 학습 데이터 전체의 평균값으로 빼주고 또한 표준편차 값으로 나누어주는 과정을 통해서 정규화한 후 사용하게 된다.

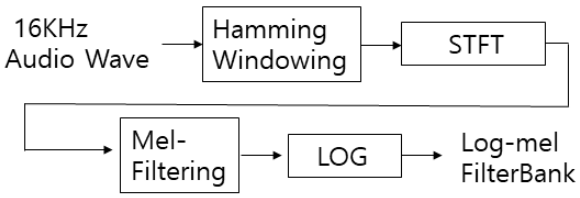


그림 1. 오디오 신호의 특징 추출 과정
Fig. 1. Feature extraction process of audio signal

2.2 분절 모델의 구조

본 논문에서 사용된 소리 이벤트 검출을 위한 분절 모델은 최근 음성인식 분야에서 사용하기 위하여 Bowen[13] 등에 의하여 제안된 모델에 기반하며 보다 상세한 구조는 그림 2 에 나타나 있다.

분절 모델은 2.1절에서 언급된 특징 추출(Feature extraction) 부분과 더불어 인코딩(Encoding) 부분, 세그먼트 임베딩(Embedding) 부분과 유사도 스코링(Scoring) 부분으로 크게 나누어진다. 또한, 학습을 위해서는 비용함수 계산 및 역전파 부분이 존재하며, 디코딩을 위해서는 동적프로그래밍을 활용한 최적의 세그먼트 결정 부분이 추가된다.

특징추출의 결과인 로그-멜 필터뱅크 출력 $x=\{x_1,x_2,\cdots,x_T\}$, $x_t\in R^B$ 는 인코딩 과정을 거쳐서 인식에 보다 적합한 새로운 특징 $h=Enc(x)$, $h=\{h_1,h_2,\cdots,h_{\hat{T}}\}$, $h_t\in R^F$으로 변하게 된다. 여기서 B와 F는 각각 로그-멜 필터뱅크 벡터와 인코딩 벡터의 차원을, T는 주어진 오디오 파일의 전체 프레임 길이를 $\hat{T}$는 인코딩 벡터의 전체 길이를 각각 의미한다. 인코딩 부분은 64개의 은닉 뉴런을 가진 bi-directional LSTM(Long Short-Term Memory)와 5x5의 커널 사이즈를 가진 CNN으로 구성된다. LSTM을 통해서는 오디오 신호들의 긴 시간상관관계를 모델링할 수 있으며, CNN의 pooling을 통하여 시간 구간 길이를 축소함으로써 불필요한 정보를 제한하고 계산량을 줄이는 효과를 얻고자 한다.

세그먼트 임베딩 부분에서는 인코딩 결과인 h로부터 각 세그먼트에 적합한 임베딩 벡터를 추출하게 된다. 예를 들어, 시간 구간 $[t,t+s]$에 해당하는 임의의 세그먼트 $h_{t:t+s}\in R^{s\times F}$에 해당하는 임베딩 벡터 $I(h_{t:t+s})\in R^D$는 아래와 같이 ReLU 활성함수를 가지는 FNN(Feed-forward Neural Network)를 이용하여 얻어진다.

$$I(h_{t:t+s})=W_1P(h_{t:t+s})+b_1 \tag{1}$$

여기서 $P(\bullet)$는 주어진 세그먼트에 대한 pooling을 의미하며 본 논문에서는 아래 식과 같이 각 세그먼트의 처음과 끝의 인코딩 벡터의 단순한 결합을 사용하였다. 여기서 W_1과 b_1은 각각 FNN의 가중치 벡터와 바이어스(Bias)를 나타낸다.

$$P(h_{t:t+s})=[h_t:h_{t+s}] \tag{2}$$

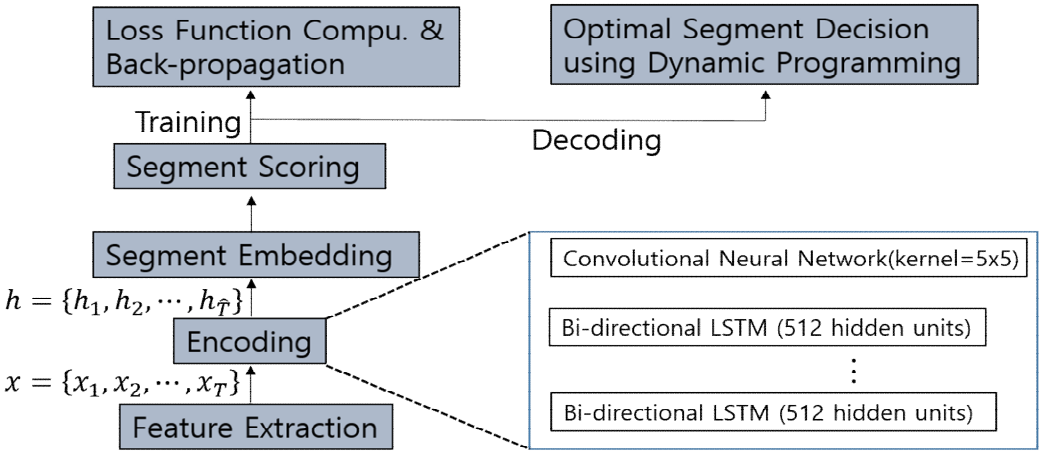


그림 2. 분절 모델의 구조
Fig. 2. Structure of segmental models

식 (2)와 같이 각 세그먼트에 대한 임베딩 벡터가 얻어지면 이를 이용하여 시작 시점이 t이고 그 길이가 s이며 소리의 종류가 y인 세그먼트에 대한 유사도 스코어(Score) $w_{t,s,y}$ 를 아래 식과 같이 내적(Dot product)을 이용하여 구하게 된다.

$$w_{t,s,y} = W_2 I(h_{t:t+s}) + b_2, \quad (3)$$
$$1 \le t < \hat{T},\ \ 1 \le s < S,\ 1 \le y \le Y$$

여기서, S는 하나의 세그먼트의 가능한 최대 길이이고 Y는 분류하고자 하는 소리 종류의 전체 개수를 나타내며, W_2와 b_2는 각각 가중치 벡터와 바이어스를 나타낸다.

2.3 학습

학습을 위한 비용함수로써, 강전사 레이블 정보가 필요한 log loss함수를 사용하는 대신에 본 연구에서는 marginal log loss를 사용하였다. 이를 통해서 오디오 분류기의 성능 저하를 발생시키는 불완전(Incomplete) 레이블 오류에 대해서 효과적으로 대처할 수 있을 것이라 판단된다[13]. 학습데이터로써 입력 특징벡터 $x = \{x_1, x_2, \cdots, x_T\}$ 와 레이블 시퀀스(Sequence) $Y = \{y_1, y_2, \cdots, y_L\}$가 주어진 경우에, marginal log loss 값은 아래식과 같이 정의 된다[13].

$$\mathcal{L}(Y,x) = -\log p(Y|x) = -\log \sum_z p(Y,z|x) \quad (4)$$

여기서, z는 입력 x가 가질 수 있는 모든 가능한 시간 분할(Segmentation)을 의미한다.

한편, 식 (4)로부터,

$$p(Y,z|x) = p(\pi|x) = \frac{1}{Z(x)} \exp(w(x,\pi)) \quad (5)$$

여기서, π는 $\{Y,z\}$로 구성되는 입력오디오 x에 대한 하나의 경로(Path)를 나타내며, $w(x,\pi)$는 해당 경로에 대한 전체 스코어 값을 의미하는데, 경로를 구성하는 각 세그먼트에 대해서 식 (3)을 이용하여 스코어 값을 각각 구한 후 더해줌으로써 얻을 수 있다.

식 (4)와 식 (5)로부터,

$$\begin{aligned}\mathcal{L}(Y,x) &= -\log \sum_z p(Y,z|x) \\ &= -\log \sum_z \exp(w(x,(Y,z))) + \log Z(x) \\ &= -\log \sum_{\bar{\pi}:\, L(\bar{\pi})=Y} \exp(w(x,\bar{\pi})) + \log Z(x)\end{aligned}$$

$$Z(x) = \sum_{\bar{\pi}} \exp(w(x,\bar{\pi})) \quad (6)$$

여기서 $L(\bar{\pi}) = Y$ 는 해당 경로 $\bar{\pi}$가 레이블 시퀀스 Y를 가짐을 의미한다.

식 (6)의 비용함수 및 그에 대한 gradient descent 값은 forward-backward 알고리듬[14]을 통하여 효과적으로 계산되며, 이를 통하여 분절 모델을 구성하는 뉴럴네트워크의 파라미터 값을 반복적 학습을 통하여 얻을 수 있다.

2.4 디코딩(Decoding)

오디오 이벤트 탐지를 위한 디코딩 과정은 아래 식과 같다.

$$p^* = \arg max_{\pi} w(x,\pi) \quad (7)$$

식 (7)을 통해서 스코어 값이 가장 큰 최적의 경로 $p^* = \{Y^*, z^*\}$가 얻어지며 이는 우리가 얻고자 하는 최적의 분할 정보와 그에 해당하는 소리의 종류들에 해당한다. 한편, 식 (7)은 Viterbi 알고리듬을 이용한 동적프로그래밍을 통하여 효과적으로 계산된다[13].

III. 실험 결과

3.1 데이터베이스

본 논문에서는 DCASE Challenge 2018 Task 4 의 오디오 데이터를 학습 및 테스트에 사용하였다[15]. 원래의 학습데이터는 약전사 레이블 데이터와 강전사 레이블 그리고 비전사 레이블로 구성되어 있으나, 본 논문에서는 앞 절에서 기술된 분할모델을 수

정 없이 바로 적용하기에 적합한 강전사 레이블 학습데이터 만을 사용하였다. 그러나 marginal log loss 비용 함수를 적용하기 위하여 강전사 레이블에 포함된 분할정보는 사용하지 않았다.

학습데이터의 각 파일의 길이는 10초로 동일하며 2548개의 파일로 구성되어 있다. 분류하고자 하는 소리의 종류는 10개이며 가정에서 흔히 발생하는 소리 이벤트로 구성되어 있다. 표 1에는 학습(Train), 테스트(Test) 그리고 검증(Validation) 데이터에 대한 상세한 정보가 나와 있다.

표 1. 오디오 데이터의 구성
Table 1. Contents of the audio data

	Train	Test	Validation
No. of clips	2548	692	1168
Clip length	10 sec.		
Classes(10)	Speech, Dog, Cat, Alarm bell ring, Dishes, Frying, Blender, Running water, Vacuum cleaner, Electric shaver toothbrush		

3.2 성능 평가 방법

소리 이벤트 검출기에 대한 성능은 F-score를 이용하여 평가하며, 이벤트기반(Event-based) 분석방법을 사용한다[16]. 이벤트기반 분석방법은 소리 이벤트 검출기에서 특정 이벤트가 발생한 경우에, 참 레이블 정보(Ground truth)와 비교하는 방법이다. 초기 판단은 TP(True Positive), FP(False Positive), FN(False Negative)의 3가지 형태로 하게 된다. TP는 검출된 소리 이벤트와 참 레이블 정보상의 시작 시간과 종료 시간이 겹치는 경우다. TP의 판단시, 시작 시간과 종료 시간에서 각각 200ms 오차가 허용되나, 겹치는 구간이 전체 소리 이벤트 길이의 20%를 넘어야 한다. FP는 TP와 상반되는 개념으로, 검출기에 의해서 소리 이벤트가 발생됨에도 불구하고 참 레이블 정보와 겹치는 구간이 발생하지 않는 경우이다. FN는 참 레이블 정보에는 소리 이벤트가 존재하나 검출기 출력이 존재하지 않는 경우이다.

F-score는 위에서 언급된 3가지 초기 판단을 근거로 계산되며 Precision과 Recall의 조화평균 값이다. Precision은 참인 문제에 대해 얼마나 잘 맞추었는지에 대한 값이며, Recall은 참인 문제에 대해 정확히 양성으로 식별한 비율을 말한다. F-score는 식(9)에서 계산된다.

$$P = \frac{TP}{TP+FP} \quad R = \frac{TP}{TP+FN} \tag{8}$$

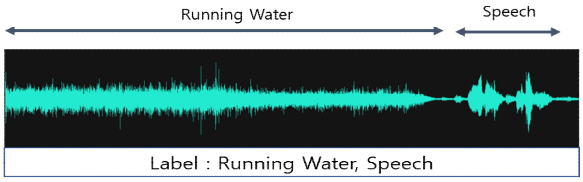
$$F = \frac{2 \cdot P \cdot R}{P+R} \tag{9}$$

3.3 소리 이벤트 검출 실험 결과

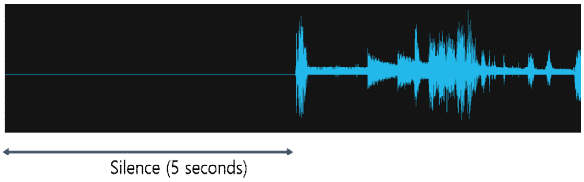
학습과 인식시에 적용되는 세그먼트의 최대 가능 길이는 파일 길이와 동일하게 10초로 설정하여 긴 지속시간을 가진 소리들에 대비하도록 하였다. Batch size는 16으로 설정하였으며, 모델 훈련은 marginal log loss를 손실함수로 삼아 Adam optimizer를 이용하였고 학습률(Learning rate)는 0.001로 하였다. 학습은 검증데이터에 대한 F-score 값을 기준으로 early stopping을 적용하였다.

그림 3에는 본 논문에서 사용된 오디오 데이터의 전형적인 예들을 보여주고 있다. 그림 3(a)에는 두 가지 종류의 소리가 이어지며 그 사이에 약간의 묵음 구간이 존재함을 알 수 있으며 그림 3(b)에는 파일의 앞부분에 상당한 길이의 묵음 구간이 존재하는 것을 알 수 있다. 이와 같이 오디오 파일에는 파일의 앞뒤와 서로 다른 소리가 연결되는 부분에 묵음 구간이나 의미 없는(Garbage) 소리가 존재하게 된다. 따라서 본 연구에서는 이와 같은 현상에 대처하기 위하여 파일의 시작과 끝 부분 그리고 소리의 연결부분에 “SIL” 이라고 하는 임의의 소리 레이블(Label)을 사용하였다. 이렇게 함으로써 실제 우리가 분류하고자 하는 타입의 소리가 아니거나 묵음 구간에 대해서 세그먼트 모델링을 할 수 있으며, 이를 통해서 성능이 심하게 저하되는 부분을 방지하고자 하였다. 이러한 “SIL” 레이블의 효과에 대해서 뒷부분에서 상세히 설명하고자 한다.

표 2에는 제안된 분절모델에서 RNN의 층수를 변화시키면서 관찰된 성능을 나타내었다. RNN의 층수가 증가하면서 F-score 값은 전반적으로 하락하였다. 이는 오디오 데이터의 양에 비해서 RNN의 파라미터의 수가 많다는 것을 의미한다.



(a) 하나의 오디오 클립에 다수의 레이블이 존재
(a) Multiple labels exist in a single audio clip



(b) 묵음구간(배경잡음)이 오디오 클립의 상당부분 차지
(b) Silence periods(background noise) cover significant parts of the audio clip

그림 3. 오디오 데이터의 전형적인 예들
Fig. 3. Typical examples of audio data

표 2. RNN 층 수에 따른 F-score(%) 성능 변화
Table 2. F-score variation as the number of RNN layers changes

# of RNNs	Test set	Validation set	Train set
1	14.87	9.23	57.07
2	10.49	9.26	42.44
3	13.91	8.83	45.55
4	11.48	7.92	33.65

이는 본 논문 실험에서 사용하는 오디오 데이터의 양이 다소 작은 것에 기인하는 것이라 판단된다. RNN의 층수가 1일 때 우리는 최고의 성능을 얻을 수 있었으며, 테스트 데이터에 대해서는 14.87% 그리고 검증데이터와 학습데이터에 대해서는 각각 9.23%과 57.07%의 F-score 값을 얻을 수 있었다.

표 3에는 본 논문에서 제안된 분절모델과 DCASE 2018에서 제안된 CRNN 기반의 오디오 분류기의 성능 비교를 나타내었다. 비교를 위해서 CRNN 오디오 분류기의 학습에서도 본 논문에서와 마찬가지로 DCASE 2018의 강전사 레이블 학습데이터 만을 사용하였으며, DCASE 2018 공식사이트에서 제시된 baseline 인식기의 학습절차를 그대로 따라하였다. 성능 비교 결과 제안된 분절 모델이 기존의 CRNN에 비해서 월등히 나은 성능을 보임을 알 수 있다. 물론 이 결과는 DCASE 2018에서 제시된 모든 학습데이터을 사용한 결과는 아니지만, 분절모델이 오디오 분류에 있어서 기존의 CRNN방식에 비해서 우월한 성능을 나타낼 수 있다는 가능성을 보여 준다는 점에서 큰 의미가 있다고 생각된다.

표 3. 본 연구에서 제안된 분절모델과 DCASE 2018의 CRNN 베이스라인 분류기의 성능비교
Table 3. Performance comparison between the proposed segmental model and the CRNN based baseline model in DCASE 2018

	Test set	
	DCASE 2018 CRNN baseline	Proposed segmental model
F-score (%)	2.74	14.87

표 4에는 분할모델에서 “SIL” 레이블을 사용하는 경우와 사용하지 않은 경우의 성능비교를 나타내었다. 우리의 예상대로, “SIL” 레이블을 사용하는 것이 성능 향상에 매우 유리함을 알 수 있었다. 이는 학습과 인식시에 사용되는 오디오 파일에는 우리가 분류하고자 하는 소리 외에도 많은 잡음과 묵음 구간이 존재하기 때문인 것으로 판단되며, 이 결과를 통해서 이들을 전체적으로 대표하는 레이블이 반드시 필요하다는 것을 알 수 있었다.

표 4. “SIL” 레이블을 사용할 경우와 그렇지 않은 경우의 F-score(%) 성능 비교
Table 4. Performance comparison between with and without using “SIL” label

# of RNNs		Test set	Validation set	Train set
With “SIL”	1	14.87	9.23	57.07
	2	10.49	9.26	42.44
Without “SIL”	1	6.41	4.28	7.48
	2	8.03	5.56	9.53

IV. 결 론

기존의 소리 이벤트 검출을 위하여 사용되던 프레임 단위의 인식 방식은 학습시 오디오 데이터에 대한 분할 정보가 정확히 주어져야 하는 단점이 있었다. 이를 극복하기 위하여 본 논문에서는 세그먼트 단위의 인식을 위한 분절모델을 소리 이벤트 검출에 사용하였다.

제안된 분절모델은 오디오 특징벡터에 대한 인코딩 부분과 각 세그먼트를 대표하는 벡터를 추출하기 위한 임베딩 부분 그리고 각 세그먼트에 대한 유사도 값을 추출하기 위한 스코어링 부분들이 연결되어 있으며, 이들 각 부분에 딥뉴럴네트워크들을 적절히 사용함으로써 전체적으로 end-to-end 기반의 분절 기반 오디오 분류기가 구축되도록 하였다.

제안된 오디오 분류기는 기존의 프레임기반의 오디오 분류기에 비해서 분할에 대한 정확한 정보가 필요 없이 학습이 가능할 뿐만 아니라, 오디오 정보를 프레임별로 인식할 경우에 발생하는 인식의 부자유스러움을 해소할 수 있어서 더욱 더 향상된 오디오 이벤트 검출 성능을 이끌어 낼 수 있을 것이라 판단된다.

DCASE 2018 오디오 데이터를 이용한 실험한 결과 제안된 분절 모델은 CRNN 기반의 기존의 오디오 이벤트 검출기에 비해서 뛰어난 성능을 보여 주었으며, 이를 통해서 오디오 분류에서 분절모델이 가지는 밝은 전망을 확인 할 수 있었다.

References

[1] J. J. Aucouturier, B. Defreville, and F. Pachet, "The Bag-of-Frames Approach to Audio Pattern Recognition: A Sufficient Model for Urban Soundscapes but Not for Polyphonic Music", The Journal Acoustical Society of America, Vol. 122, No. 2, pp. 881-891, Aug. 2007. https://doi.org/10.1121/1.2750160.

[2] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines", ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, pp. 1-27, Apr. 2011. https://doi.org/10.1145/1961189.1961199.

[3] M. Crocco, M. Christani, A. Trucco, and V. Murino, "Audio Surveillance: A Systematic Review", ACM Computing Surveys, Vol. 48. No. 4, pp. 1-46, May 2016. https://doi.org/10.1145/2871183.

[4] Y. Wang, L. Neves, and F. Metze, "Audio-based Multimedia Event Detection Using Deep Recurrent Neural Networks", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 2742-2746, Mar. 2016. https://doi.org/10.1109/ICASSP.2016.7472176.

[5] J. Salamon and J. P. Bello, "Feature Learning with Deep Scattering for Urban Sound Analysis", 23rd European Signal Processing Conference (EUSIPCO), Nice, France, pp. 724-728, Sep. 2015. https://doi.org/10.1109/EUSIPCO.2015.7362478.

[6] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic Classification of Multiple Simultaneous Bird Species: A Multi-instance Multi-Label Approach", The Journal of the Acoustical Society of America, Vol. 131, No. 6, pp. 4640-4650, Jun. 2012. https://doi.org/10.1121/1.4707424.

[7] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection", IEEE/ACM Trans. On Audio Speech and Language Processing, Vol. 25. No. 6, pp. 1291-1303, Jun. 2017. https://doi.org/10.1109/TASLP.2017.2690575.

[8] J. Y. Kwak and Y. J. Chung, "Sound Event Detection Based on CRNN Using Derivative Features", Journal of KIIT, Vol. 18, No. 6, pp. 89-96, Jun. 2020. https://doi.org/10.14801/jkiit.2020.18.6.89.

[9] E. Fonseca, J. Pons, X. Facory, F. Font, D. Bogganov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound Datasets: A Plaform for the Creation of Open Audio datasets", in Proc. of the 18th International Society for Music Information Retrieval Conference(ISMIR 2017), Suzhou, China, pp. 486-493, 2017.

[10] B. Frenay and M. Verleysen, "Classification in the Presence of Label Noise: a Survey", IEEE Transactions on Neural Networks and Learning

Systems, Vol. 25, No. 5, pp. 845-869, May 2014. https://doi.org/10.1109/TNNLS.2013.2292894.

[11] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning Sound Event Classifiers from Web Audio with Noisy Labels", in Proc. of International Conference on Acoustics, Speech, and Signal Processing, Brighton, UK. May 2019. https://doi.org/10.1109/ICASSP.2019.8683158.

[12] E. Beigman and B. B. Klebanov, "Learning with annotation noise", in Proceedings of Joint Conf. 47th Ann. Meeting ACL and 4th Int. Joint Conf. Natural Language Processing AFNLP, Suntec, Singapore, pp. 280-287, Aug. 2009. http://dx.doi.org/10.3115/1687878.1687919.

[13] H. Tang, L. Lu, L. Kong, and K. Gimple, "End-to-end Neural Segmental Models for Speech Recognition", IEEE J. Selected Topics in Signal Proc., Vol. 11, No. 8, pp. 1254-1264, Dec. 2017. https://doi.org/10.1109/JSTSP.2017.2752462.

[14] B. Shi, S. Settle, and K. Livescu, "Whole-word Segmental Speech Recognition with Acoustic Word Embeddings", in Proc. of IEEE Spoken Language Technology Workshop, Shenzen, China, Jan. 2021. https://doi.org/10.1109/SLT48900.2021.9383578.

[15] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis", Detection and Classification of Acoustic Scenes and Events 2019, New York, USA, Oct. 2019.

[16] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection", Applied Sciences, Vol. 6, No. 6, pp. 162-178, May 2016. https://doi.org/10.3390/app6060162.

저자소개

김 수 종 (Soo-Jong Kim)

2019년 : 계명대학교 전자공학과 (공학사)
2021년 ~ 현재 : 계명대학교 전자공학과 석사과정
관심분야 : 인공지능, 오디오 검출

정 용 주 (Yong-Joo Chung)

1995년 8월 : 한국과학기술원 (공학박사)
1999년 3월 ~ 현재 : 계명대학교 전자공학과 교수
관심분야 : 오디오 분류, 음성인식