Journal of KIIT. Vol. 18, No. 6, pp. 89-96, Jun. 30, 2020. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.6.89

차분특징을 이용한 CRNN 기반의 소리 이벤트 검출

곽진열*, 정용주**

Sound Event Detection Based on CRNN Using Derivative Features

Jin-Yeol Kwak*, Yong-Joo Chung**

요 약

본 논문에서는 딥뉴럴네트워크 기반의 소리 이벤트 검출을 위하여 차분특징을 사용하였다. 오디오 신호의 매 프레임 마다 주파수 분석을 통한 로그-멜-필터뱅크 값을 추출하고 프레임들 간의 상관관계를 이용한 1차 및 2차 차분 특징 값을 추출하여 이용하였다. 베이스라인 검출기로는 최근 소리 이벤트 검출에서 가장 많이 사용되는 CRNN(Convolutional Recurrent Neural Network)을 사용하였으며, 64차원의 로그-멜-필터뱅크 값과 그들의 1차 차분 및 2차 차분 값을 독립적인 입력 특징 맵으로 구성하였다. CRNN의 출력단에는 global average pooling을 추가하여 강전사(strong label) 오디오 데이터 뿐만 아니라 약전사(weak label) 및 비전사(un-label)데이터도 학습에 활용할 수 있도록 하였다. 다양한 학습 환경에서 차분 특징을 사용함으로서 일관된 성능 향상이 있음을 확인하였다. DCASE Challenge 2018/2019 오디오 데이터를 이용한 실험결과, 제안된 차분 특징을 이용하여 최대 16.9%의 상대적 F-score 향상을 얻을 수 있었다.

Abstract

In this paper, we used derivative features for sound event detection based on deep neural networks. We extracted log-mel-filterbank value for each frame of the audio signal by frequency analysis and its 1st and 2nd derivative features were extracted for the use by exploiting the correlation between the frames. CRNN which is recently most popular in audio event detection was used as the baseline detector and 64 dimensional log-mel-filterbank outputs and their 1st and 2nd derivatives were constructed as independent input feature maps. Global average pooling layer is added at the output of the CRNN to make use of weak and un-label audio data as well as strong label data in the training. In the various training environment, we could observe consistent performance improvement by using the derivative features. From the experimental results using DCASE Challenge 2018/2018 audio data, we could obtain maximally 16.9% relative improvement in F-score.

Keywords

sound event detection, CRNN, derivative features, deep neural networks

* 계명대학교 전자공학과 석사과정 졸업
- ORCID: https://orcid.org/0000-0001-5792-6267
** 계명대학교 전자공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-0060-1178

· Received: Apr. 06, 2020, Revised: Apr. 20, 2020, Accepted: Apr. 23, 2020

· 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(audio) 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 딥 뉴럴 네트워크(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].

딥뉴럴네트워크 기반의 방식 중에서 가장 먼저 활용된 FNN(Feed Forward Neural Network)은 그 구조가 간단하다는 장점이 있고 구현도 비교적 쉽게 이루어질 수 있으나, 영상이나 오디오 신호에서 발생하는 신호의 변이나 부가된 잡음신호에 대해서 강인하지 못한 것으로 알려져 있다.

영상인식에서 매우 훌륭한 성능을 보여준 CNN(Convolutional Neural Network)는 오디오 태깅(Audio tagging) 등의 소리 분류에도 적용되어 좋은 성과를 보여주었다[7][8][9]. 상이한 2차원의 필터들을 입력되는 소리 신호의 주파수-시간 스펙트럼에 적용하여 다양한 소리 특징 값을 추출함으로서, CNN은 FNN에서 단점으로 취급되었던 소리 신호의 변이나 잡음에 대한 취약점을 일정 부분 보완해 주는 것으로 알려져 있다. 그러나 CNN은 소리 신호의 샘플들 간의 시간 영역에서의 상관 관계를 모델링하는데 있어서는 다소 부족하다고 알려져 있다.

음성인식 분야에서 널리 사용되고 있는 RNN(Recurrent Neural Network)은 음성이나 오디오와 같은 시계열 신호의 시간 영역에서의 상관관계를 모델링하는데 있어서 다른 딥뉴럴네트워크 방식에 비해서 월등하다[10]. 최근에 들어서는 이러한 RNN의 장점과 앞에서 언급된 FNN 및 CNN의 장점을 결합함으로서 보다 나은 성능을 나타내기 위한 노력이 많이 진행되었으며, 특히 소리 이벤트 검출을 위한 CRNN(Convolutional Recurrent Neural Network)이 최근에 제안되어 매우 우수한 성능을 보여주었다[11].

CRNN의 학습과 테스트를 위해서는 소리 신호의 특징이 필요한데, 지금까지의 대부분의 연구들은 이를 위하여 소리 신호의 매 프레임마다 주파수 분석을 통하여 로그-멜-필터뱅크 값을 계산한 후 이를 단독으로 CRNN의 입력으로 사용하고 있다. 그러나 CNN이나 RNN을 이용한 음성인식 등의 연구에서는 특징벡터의 차분 값을 함께 사용함으로서 보다 나은 성능을 보여주고 있다[9]. 소리 신호와 음성신호의 유사성을 고려하였을 때, 소리 이벤트 검출에서도 특징벡터의 차분값을 원래의 로그-멜-필터뱅크 값과 함께 사용함으로서 보다 나은 성능을 보여 줄 것이라 기대된다.

이러한 점에 착안하여 본 연구에서는 일정한 길이의 오디오 신호가 주어지면, 이로부터 매 프레임마다 로그-멜-필터뱅크 값을 먼저 추출한 후 그들의 1차 및 2차 차분 특징 값을 프레임들 간의 상관 정보를 이용하여 구하였다. 이와 같이 구한 로그-멜-필터뱅크 값과 그들의 차분특징들을 각각 시간-주파수 영역의 2차원 특징맵(Feature map)으로 구성하여 CRNN의 입력으로 함께 사용함으로서 보다 나은 소리 이벤트 검출 성능을 유도하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 소리 이벤트 검출을 위한 특징 추출 방법과 본 논문에서 사용된 CRNN 인식기의 구조에 대해서 설명하며 3장에서는 본 연구에서 수행한 다양한 인식 실험결과를 제시하고 4장에서 결론을 맺는다.

Ⅱ. 특징 추출과 CRNN 구조

2.1 로그-멜-필터뱅크 추출

CRNN의 학습 및 테스트를 위해서는 오디오 웨이브(wave)로부터 특징 추출이 필요하며, 전체적인 추출과정은 그림 1에 나타나 있다. 오디오 웨이브는 16 kHz의 비율로 샘플링되었으며 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.

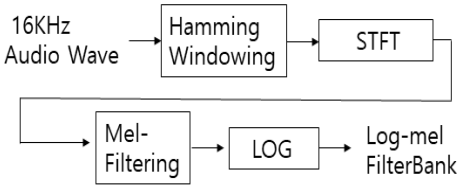

그림 1. 오디오 신호의 특징 추출 과정
Fig. 1. Feature extraction process of audio signal

프레임 간격이나 로그-멜-필터뱅크 값의 차원 등은 본 논문에서 사용된 DCASE 2018/2019 오디오 데이터의 배포시 추천된 값을 사용하였다[12]. CRNN의 학습시에 로그-멜-필터뱅크 값은 학습 데이터 전체의 평균값으로 빼주고 또한 표준편차 값으로 나누어주는 과정을 통해서 정규화한 후 사용하게 된다.

2.2 차분 특징

소리 이벤트 검출과 유사한 분류 과정을 가지고 있는 음성인식 분야에서는 로그-멜-필터뱅크 특징에 DCT(Discrete Cosine Transform)을 적용하여 얻은 MFCC(Mel-Frequency Cepstral Coefficient)를 사용한다. 그러나 음성인식에서는 단순한 MFCC 뿐만 아니라 이의 차분 특징을 구하여 함께 사용함으로서 성능의 향상을 이루고 있다.

차분특징에 대한 계산식은 식 (1)과 같다. d_t는 차분특징을 나타내며 c_t는 t번째 프레임에서의 정적 특징을, N은 차분특징을 계산할 때 사용되는 앞뒤 프레임의 개수이다.

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (1)$$

본 연구에서는 식 (1)을 이용하여 로그-멜-필터뱅크의 1차 차분 특징과 2차 차분 특징을 구한 후, 원래의 특징과 함께 각각 독립된 특징 맵으로 구성하여 그림 2에 나타난 바와 같이 CRNN의 입력으로 사용하였다.

2.3 CRNN의 구조

본 연구에서 사용된 CRNN의 구조가 그림 3에 나타나 있다. 기본적으로 [11]에서 사용된 CRNN의 구조와 유사하나 출력단에 global average pooling layer를 추가하였다. CNN으로 구성된 3개의 컨벌루션 블록(Conv. block), 1개의 bidirectional GRU (Gated Recurrent Unit) 그리고 FNN으로 구현된 하나의 분류층(Classification layer)이 연속적으로 연결되어 있다. 약전사(Weak label) 학습 데이터에 대비하기 위하여 분류층의 출력값을 클립(Clip) 전체 구간에 대해서 평균하기 위한 global average pooling 층이 마지막에 존재한다. 강전사(Strong label) 학습데이터인 경우에는 분류층의 출력값이 최종 결과 값이 된다.

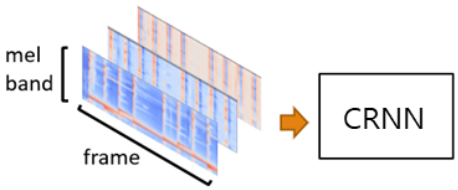

그림 2. 1, 2차 차분특징을 포함한 CRNN의 입력 특징맵 구성
Fig. 2. Construction of input feature map of CRNN including 1st and 2nd derivative features

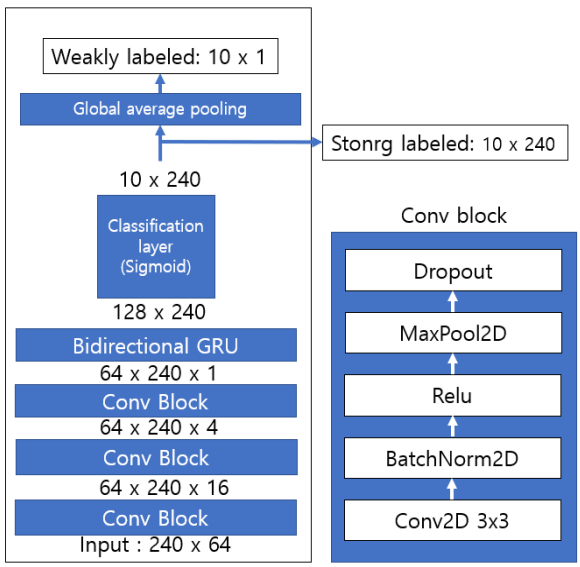

그림 3. CRNN의 구조
Fig. 3. Structure of CRNN

CRNN의 입력으로는 64차원의 로그-멜 필터뱅크 값과 그들의 1차 및 2차 차분 값들이 사용되며, 이들은 각각 독립된 240 프레임 길이의 특징 맵으로 구성된다. 이는 CRNN으로 입력되는 모든 오디오 클립의 길이가 10초이며 41.6ms의 프레임 간격 마다 로그-멜 필터뱅크 값이 생성되기 때문이다.

1, 2차 차분특징을 사용함으로서 CRNN의 파라미터 개수의 증가가 발생하지만 전체 파라미터 개수에 비하면 상대적으로 그리 크지 않으므로, 전체 시스템의 복잡도에는 큰 영향을 미치지 않는다. 참고로, 차분 특징을 사용할 경우에 CRNN의 전체 파라미터의 개수는 약 127,000개이며, 사용하지 않을 경우는 약 126,000개이다.

컨벌루션 블록은 64개의 특징맵으로 구성된다. 각 특징맵 마다, 3×3의 필터링 과정 이후 batch normalization을 적용하고 ReLU(Rectified Linear Unit) 활성함수를 사용하였으며 마지막으로 dropout이 적용되었다. 1x4의 max pooling을 통하여 주파수 영역에서의 정보의 양은 줄이되 시간 영역에서의 정보는 평균되지 않고 보전되도록 하였다. 이를 통하여 오디오 신호의 시간 관계 정보를 후단의 RNN에서 모델링 할 수 있도록 하였다. 마지막 3번째 컨벌루션 블럭의 출력은 bidirectional GRU의 입력으로 들어가게 되며 각각 64개의 unit을 가진 쌍방향 GUR의 출력들은 매 프레임 마다, 10개의 unit으로 구성된 분류층의 입력으로 사용된다.

분류층의 10개의 unit은 각각 분류하고자 하는 특정 오디오 클래스를 의미하며, sigmoid 활성함수를 가진다. 그 출력값은 10개 클래스별로의 매 프레임별 사후확률(Posterior probability) 값을 나타내게 된다. 학습데이터가 약전사 레이블을 가진 경우에는 분류층의 출력들은 global average pooling층의 입력으로 사용되어, 오디오 클립 전체 길이(10초) 동안의 클래스 별 사후 확률의 평균값을 출력하게 된다.

III. 실험 결과

3.1 데이터베이스

본 논문에서는 DCASE Challenge 2018 및 2019의 Task 4의 오디오 데이터를 학습 및 테스트에 사용하였다[13]. 학습데이터는 크게 약전사 레이블 데이터와 강전사 레이블 그리고 비전사 레이블로 나누어지며 표 1에 학습데이터에 대한 보다 자세한 내용이 나타나 있다.

표 1. 학습데이터의 구성
Table 1. Contents of the training data

Label type	Weak label	Strong label	Un-label
No. of clips	1578	2045	14412
Label properties	Clip-level	Frame-level	None
Clip length	10 sec.		
Classes(10)	Speech, Dog, Cat, Alarm bell ring, Dishes, Frying, Blender, Running water, Vaccum cleaner, Electric shaver toothbrush		

약전사 레이블은 시간정보가 없이 클립 단위의 소리 이벤트 정보를 제공하며 강전사 레이블은 시간 프레임 단위의 이벤트 발생 정보가 주어진다. 비전사 레이블은 원래 레이블 정보가 존재하지 않지만, 약전사 레이블 학습데이터를 이용하여 학습된 초기 CRNN을 이용하여 클립 단위의 레이블을 추후 얻을 수 있다. 각 클립의 길이는 10초로 동일하며, 약전사 레이블, 강전사 레이블 그리고 비전사 레이블 데이터는 각각 1578, 2045, 14412개의 클립으로 구성되어 있다. 분류하고자 하는 소리의 종류는 10개이며 가정에서 흔히 발생하는 소리 이벤트로 구성되어 있다.

표 2에는 테스트 데이터에 대한 정보가 나와 있으며, 2018 테스트 데이터는 208개의 클립이, 2019 테스트 데이터는 1168개의 클립이 존재한다. 인식 성능의 정확한 측정을 위하여 각 클립에는 프레임 단위의 레이블이 함께 제공된다.

표 2. 테스트 데이터의 구성
Table 2. Contents of the test data

	DCASE 2018 test	DCASE 2019 test
No. of clips	288	1168
Label properties	Frame-level	
Clip length	10 sec.	
Classes(10)	Same as training data	

3.2 성능 평가 방법

CRNN은 매 프레임마다 10개 클래스(Classes)의 소리 이벤트에 대해 사후 확률 값을 계산하고, 그 값이 문턱 값 0.5를 넘을 때 해당 프레임에 소리 이벤트가 존재한다고 간주한다. 보다 신뢰성 있는 판단을 위하여 중간값 필터(Median filter)를 거친 후 최종 결정을 하게 된다.

소리 이벤트 검출기에 대한 성능은 F-score와 ER(Error Rate)를 이용하여 평가하며, 이벤트기반(Event-based) 분석방법을 사용한다[14]. 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 (3)에서 계산된다.

$$P=\frac{TP}{TP+FP} \quad R=\frac{TP}{TP+FN} \tag{2}$$

$$F=\frac{2 \cdot P \cdot R}{P+R} \tag{3}$$

ER는 FN와 FP를 사용해 각각 대체(Substitution(S)), 삭제(Deletion(D)) 그리고 삽입(Insertion(I))을 식 (4)과 같이 계산하고 식 (5)와 같이 구한다.

$$\begin{aligned} S(k) &= \min(FN(k), FP(k)) \\ D(k) &= \max(0, FN(k)-FP(k)) \\ I(k) &= \max(0, FP(k)-FN(k)) \end{aligned} \tag{4}$$

$$ErrorRate=\frac{\sum S(k)+\sum D(k)+\sum I(k)}{\sum N(k)} \tag{5}$$

여기서 k는 발생한 이벤트의 수를 나타낸다[13].

3.3 소리 이벤트 검출 실험 결과

CRNN의 학습을 위하여 약전사, 강전사 그리고 비전사 레이블 데이터의 다양한 결합을 시도하였다. [약전사+비전사], [약전사+비전사+강전사], [강전사] 및 [약전사+강전사] 등의 총 4가지 조합을 사용하여 다양한 학습환경에서 차분 특징의 효과를 검증하였다.

모델 훈련은 이진 크로스엔트로피(Binary cross-entropy)를 손실함수로 삼아 Adam optimizer를 이용하였고 학습률(Learning rate)는 0.001로 적용하였다. Early stopping을 적용하였으며 최소 5회 epoch 수행 후 15회 patience를 적용하였다. 표에서 로그-멜-필터뱅크 값만을 적용한 경우를 1 channel 그리고 1차 차분과 2차 차분 특징값을 함께 적용한 경우를 3 channel이라 표시하였다.

표 3과 표 4에서 여러 가지 학습 환경 중에서 [Weak label+Un-label+Strong label]인 경우에 가장 나은 성능을 보임을 알 수 있다.

표 3. 차분특징을 이용한 CRNN 실험 결과(DCASE 2018 테스트 set 적용시)
Table 3. Experimental results of CRNN using derivative features (DCASE 2018 test set)

	DCASE 2018 test set			
	1 Channel		3 Channel	
	F-score(%)	ER	F-score(%)	ER
[Weak label + Un-label]	12.79	1.44	14.48	1.42
[Weak label+ Un-label+ Strong label]	17.57	2.42	18.85	2.41
Strong label	14.99	2.41	16.62	2.51
[Weak label+ Strong label]	15.25	2.42	17.83	2.37
Average	15.15	2.17	16.95	2.18
Relative improvement	-	-	11.6%	0.5%

표 4. 차분특징을 이용한 CRNN 실험 결과((DCASE 2019 테스트 set 적용시)
Table 4. Experimental results of CRNN using derivative features(DCASE 2019 test set)

	DCASE 2019 test set			
	1 Channel		3 Channel	
	F-score(%)	ER	F-score(%)	ER
[Weak label + Un-label]	11.28	1.55	11.93	1.54
[Weak label+ Un-label+ Strong label]	13.80	2.99	14.63	2.92
Strong label	12.85	3.07	13.11	3.09
[Weak label+ Strong label]	13.39	2.98	14.41	2.91
Average	12.83	2.65	13.52	2.62
Relative improvement	-	-	5.3%	1.1%

물론 가장 많은 학습데이터를 사용하기에 가장 좋은 성능을 보이는 것이 당연하다고 생각할 수도 있으나 un-label 학습 데이터의 양이 전체 학습데이터의 약 80%에 해당한다는 측면을 고려하면 성능 향상 정도가 오히려 기대에 못 미치는 것으로 보인다.

DCASE 2018 테스트 set을 사용한 표 3의 결과에서 우리는 모든 학습데이터의 조합에서 차분특징을 사용함으로서 일관된 F-score의 향상을 볼 수 있다.

특히, 가장 좋은 성능을 보이는 [Weak label+Un-label+Strong label]인 경우에 7.2%의 상대적인 F-score 향상을 볼 수 있으며, 전체 평균적으로는 11.6%의 향상을 얻을 수 있었다. ER의 경우에는 차분 특징을 사용함으로써 뚜렷한 성능 변화는 관찰되지 않았다.

DCASE 2019 테스트 set을 이용한 표 4의 경우에도 표 3과 유사한 결과를 얻을 수 있었다. 전체 평균으로는 5.3%의 상대적 F-score 향상을 보였고, 가장 좋은 성능을 보인 [Weak label+Un-label+Strong label]인 경우에는 6%의 성능 향상이 있었다.

표 5에는 학습데이터 중에서 [strong label]를 테스트 데이터로 활용한 경우에 대해서 그 결과를 나타내고 있다. 학습데이터를 테스트 데이터로 사용함으로서 인식 성능이 전반적으로 올라감을 알 수 있다([weak+un-label] 제외). 그러나 상대적으로 차분특징으로 인한 성능향상은 매우 미미한 것으로 보이는데, 전체 평균적으로 0.6%의 F-score 향상을 나타내었다. 이는 학습데이터와 테스트 데이터의 차이가 크지 않는 경우에 있어서 차분 특징이 그리 큰 효과가 없는 것을 나타낸다고 할 수 있다. 그러나 일반적인 소리 이벤트 검출에서는 학습데이터와 테스트 데이터가 동일한 경우는 거의 발생하지 않기 때문에 차분특징은 일반적인 환경에서 소리 이벤트 검출의 성능을 향상 시킨다고 말할 수 있다.

표 5. 차분특징을 이용한 CRNN 실험 결과(강전사 레이블 학습데이터 적용시)
Table 5. Experimental results of CRNN using derivative features(Strong label training data)

	Strong label training set			
	1 Channel		3 Channel	
	F-score(%)	ER	F-score(%)	ER
[Weak label + Un-label]	2.05	1.75	1.65	1.74
[Weak label+ Un-label+ Strong label]	63.47	0.74	64.09	0.73
Strong label	60.30	0.82	59.69	0.84
[Weak label+ Strong label]	61.15	0.80	62.68	0.76
Average	46.74	1.03	47.03	1.02
Relative improvement	-	-	0.6%	0.9%

그림 4에는 early stopping을 적용한 CRNN의 학습시의 F-score 러닝 커브를 보여주고 있다. Epoch =45에서 검증데이터(validation)에 대해서 최적의 성능을 보여 줌을 알 수 있다.

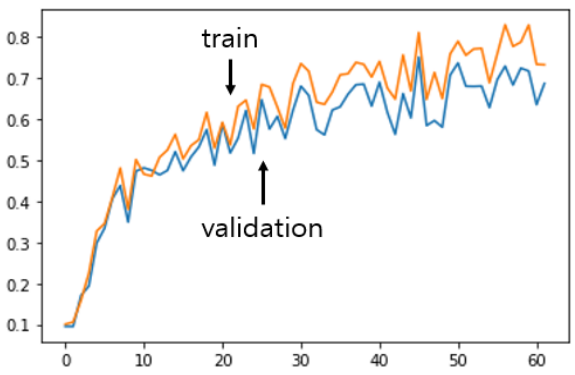

그림 4. CRNN의 F-score 러닝 커브
Fig. 4. F-score learning curve of CRNN

학습데이터에 대해서는 그 이후로 계속된 F-score 증가가 나타나지만 과학습(Over-training) 문제를 야기할 수 있기 때문에 Epoch=45에서 학습이 종료됨을 알 수 있다.

Ⅳ. 결 론

소리 이벤트 검출을 위한 다양한 기법 중에서 최근 주목을 받고 있는 방법은 딥러닝 기반의 CRNN 방식이다. 본 연구에서는 CRNN의 학습을 위한 입력 특징으로 오디오 신호의 차분 특징을 추가적으로 적용함으로서 성능향상을 이루었다.

일반적으로 소리 이벤트 검출을 위한 오디오 신호의 특징으로는 로그-멜-필터뱅크 값이 사용되는데, 본 연구에서는 이들의 1차 차분 과 2차 차분 특징을 구하여 이들을 각각 독립적인 3개의 특징 맵으로 구성한 후, CRNN의 입력으로 사용하였다.

차분 특징의 효과를 다양한 학습환경에서 검증하기 위하여 CRNN의 학습데이터를 약전사 레이블, 강전사 레이블, 비전사 레이블 등의 다양한 오디오 데이터를 결합하여 사용하였다. 차분 특징은 이러한 다양한 학습데이터의 결합에서 일관된 성능 향상을 보여 줌으로서, 제안된 방식이 효과적임을 알 수 있었다. 다만, 학습데이터와 테스트 데이터가 동일한 경우에는 성능 향상이 미미해 향후 이에 대한 추가적인 연구가 필요하리라 생각된다.

본 연구에서 차분특징은 기본적인 CRNN의 구조를 가정하고 인식실험을 하였으나 최근에는 CRNN의 개선된 구조를 통하여 보다 나은 소리 이벤트 검출 결과가 발표되고 있다. 따라서 이러한 개선된 CRNN에 대해서도 차분 특징이 효과가 있는지 향후 연구에서 검토할 예정이다.

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", J. Acoust. Soc. America., Vol. 122 No. 2, pp. 881-891, Sep. 2007.

[2] C. C. Chang, A. Ziaei, and J. H. L. Hansen, "LIBSVM: A Library for Support Vector Machines", Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, pp. 1-27, May 2011.

[3] M. Crocco, M. Christani, A. Trucco, and V. Murino, "Audio Surveillance: A Systematic Review", ACM Computing Surveys, Vol. 48. No. 4, pp. 52:1-52:46, May 2016.

[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.

[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.

[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. 4640-4640, Jun. 2012.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems, pp. 1097-1105, Dec. 2012.

[8] S. W. Jung and Y. J. Chung, "Performance Analysis of Sound Event Detection Based on CRNN", Journal of KIIT, Vol. 17, No. 5, pp. 83-90, May 2019.

[9] O. Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 10, pp. 1533-1545, Oct. 2014.

[10] A. Graves, A. Mohamed, and G. E. Hinton,

"Speech Recognition with Deep Recurrent Neural Networks", Proceedings of the IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, pp. 6645-6649, May 2013.

[11] 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.

[12] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments", Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Workshop, Surrey, UK, Nov. 2018.

[13] N. Turpault, R. Serizel, J. Salamon, and A. Shah, "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis", Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), New York, NY, USA, pp. 253-257, Oct. 2019.

[14] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection", Applied Sciences, Vol. 6, No. 6, pp. 162-178, May 2016.

저자소개

곽 진 열 (Jin-Yeol Kwak)

2020년 2월 : 계명대학교 전자공학과 대학원 석사과정 졸업

관심분야 : 머신러닝 및 오디오 분류

정 용 주 (Yong-Joo Chung)

1995년 8월 : 한국과학기술원 (공학박사)

1999년 3월 ~ 현재 : 계명대학교 전자공학과 교수

관심분야 : 오디오 분류, 음성인식