pISSN 2005-8063
eISSN 2586-5854
2019. 12. 31.
Vol.11 No.4
pp. 89-96

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2019.11.4.089

Speech detection from broadcast contents using multi-scale time-dilated convolutional neural networks*

Byeong-Yong Jang · Oh-Wook Kwon**

School of Electronics Engineering, Chungbuk National University, Cheongju, Korea

Abstract

In this paper, we propose a deep learning architecture that can effectively detect speech segmentation in broadcast contents. We also propose a multi-scale time-dilated layer for learning the temporal changes of feature vectors. We implement several comparison models to verify the performance of proposed model and calculated the frame-by-frame F-score, precision, and recall. Both the proposed model and the comparison model are trained with the same training data, and we train the model using 32 hours of Korean broadcast data which is composed of various genres (drama, news, documentary, and so on). Our proposed model shows the best performance with F-score 91.7% in Korean broadcast data. The British and Spanish broadcast data also show the highest performance with F-score 87.9% and 92.6%. As a result, our proposed model can contribute to the improvement of performance of speech detection by learning the temporal changes of the feature vectors.

Keywords: speech detection, multi-scale time-dilated convolution, deep learning, broadcast data

1. 서론

방송 데이터는 음성, 음악, 효과음, 배경 잡음 등과 같은 다양한 오디오 신호를 포함하고 있다. 특히 방송 데이터에서의 음성 신호는 다양한 환경과 화자, 그리고 발화 스타일을 포함하고 있기 때문에 음성 관련 기술을 향상시키는데 많은 도움을 줄 수 있다. 이러한 이유로 방송 데이터에서 음성 분할(speech segmentation) 또는 음성 구간 검출(speech detection)은 이전부터 흥미로운 연구 주제로 다뤄지고 있다. 또한 심층 학습(deep learning) 기술이 발달함에 따라 많은 데이터가 필요하게 되었고, 방송 데이터는 방대한 데이터 양이란 장점 때문에 더욱 주목을 받게 되었다. 하지만 방송 데이터에서의 음성은 다양한 오디오 신호와 혼합되어 있기 때문에 음성 구간을 검출하는데 어려움이 있고, 이러한 부분이 음성 구간 검출 연구에서 해결해야 하는 중요 과제이다.

IberSPEECH에서는 오디오 분할 경진대회(audio segmentation challenge)를 2010년과 2014년에 개최하였으며, 여기에는 방송 데이터에서의 음성 검출 과제가 포함되어 있다(Butko et

* This research project was supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2019 [2018-micro-9500, Intelligent Micro-Identification Technology for Music and Video Monitoring].

** owkwon@cbnu.ac.kr, Corresponding author

Received 10 September 2019; Revised 4 November 2019; Accepted 8 November 2019

al., 2011; Castan et al., 2015). 이 경진대회는 카탈루니아어(Catalan)의 방송 뉴스 데이터를 사용하였기 때문에 장르의 다양성은 다소 부족하지만, 많은 접근법과 알고리즘이 소개되었다. 2014년 경진대회에서 가장 좋은 성능을 나타낸 접근법은 2개의 시스템을 결합하여 구현되었다. 첫 번째 시스템은 은닉 마르코프 모델(hidden Markov model, HMM)을 이용하여 오버랩(overlap)이 없는 클래스에 대하여 분류를 하고, 두 번째 시스템은 첫 번째 시스템의 결과와 가우시안 혼합 모델(Gaussian mixture model, GMM)과 다층 퍼셉트론(multilayer perceptron, MLP)를 이용하여 세부적인 클래스를 분류하였다. 이 시스템의 입력은 멜스케일 켑스트럼 계수(Mel-frequency cepstral coefficient; MFCC)와 i-vector(Dehak et al., 2010)를 사용하였다.

2015년과 2018년에는 MIREX에서 방송 데이터의 음성/음악 구간 검출 과제에 대한 경진대회가 열렸다(MIREX, 2015; MIREX, 2018). 이 경진대회에서는 영국, 스페인, 독일, 프랑스에서 수집한 다양한 TV 프로그램 방송 데이터를 사용하였다. Doukhan et al.(2018)은 MIREX 2018의 경진대회에서 음성 구간 검출 과제의 가장 높은 성능을 보여주었고, 이들은 4개의 합성곱 층(convolutional layer)과 4개의 완전 연결 층(fully-connected layer)을 사용한 합성곱 신경망(convolutional neural network, CNN) 구조를 사용하고, 입력 특징으로 멜스케일 스펙트로그램(Mel-scaled spectrogram)을 사용하였다.

Tsipas et al.(2017)은 거리기반 방법과 모델기반 방법을 이용하여 멀티미디어 데이터에서 음성과 음악을 검출하는 알고리즘을 소개하였다. 이들은 자기 유사성 행렬(self-similarity matrix)을 이용하여 음성과 음악의 경계 부분을 찾고, 서포트 벡터 머신(support vector machine, SVM)을 이용하여 음성과 음악을 분류하였다. 이들은 경계 검출과 분류를 위하여 zero crossing rate, flux, spectral roll-off, root mean square energy, MFCC, spectral flatness를 특징으로 추출하여 사용하였다. 다만 이 알고리즘은 경계구간을 검출한 후 음성과 음악으로 분류하는 순서로 진행되기 때문에 오버랩 구간을 처리할 수 없는 한계점이 있다.

분할(segmentation) 문제는 영상 분야에서도 많이 연구되고 있는 과제이다. 영상 분야에서 특정 영역을 분할하기 위한 접근법으로 확장 합성곱 층(dilated convolutional layer)이 많이 사용되고 있는데 이는 합성곱의 시야가 넓어지면서 보다 넓은 범위에서의 특징 간의 변화 정보를 같이 학습할 수 있기 때문이다. Yu et al.(2015)은 확장 합성곱을 제안하여 이미지의 특정 영역 분할 성능을 향상시켰으며, Zhang et al.(2017)은 피라미드 확장 합성곱 단(pyramid dilated convolution unit)을 제안하여 이미지 분할 성능을 향상시켰다.

본 논문에서는 방송데이터의 음성 구간을 검출하기 위하여 합성곱 신경망 기반의 심층 학습 모델을 제안한다. 제안하는 심층 학습 모델은 시간 축에 대하여 더 넓은 시야를 가지도록 학습하기 위하여 시간 확장 합성곱 층(time-dilated convolutional layer)을 정의하고 사용한다. 또한, 특징 벡터의 시간적 변화 정보를 잘 추출하기 위한 다중 스케일 시간 확장 합성곱 층(multi-scale time-dilated convolutional layer)을 제안하여 사용한다. 이렇게 구성된 심층 학습 모델은 특징 벡터의 정보 손실을 줄이고, 시간적 변화 정보를 학습함으로써 음성 구간 검출을 위한 모델 학습에 유리할 것으로 예상된다. 따라서 본 논문은 다음 2장에서 제안 알고리즘을 설명하고, 3장에서 실험 및 결과를 제시하여 제안 알고리즘의 성능을 검증할 것이며, 4장에서 결론을 도출할 것이다.

2. 제안 알고리즘

방송 데이터의 오디오에 포함되어 있는 신호는 크게 음악, 음성, 잡음으로 분류할 수 있다. 여기서 음성은 다른 신호에 비하여 시간적 변화가 다양하다는 특성이 있다. 음성은 발음하는 음소 내에서도 다양한 시간적 변화가 존재하며, 하나의 음소는 매우 짧은 시간 안에 발화되기 때문에 음성은 시간에 대하여 동적 변화가 많을 수밖에 없고, 방송 데이터의 음성 검출 시스템은 이를 고려하여 구현할 필요가 있다.

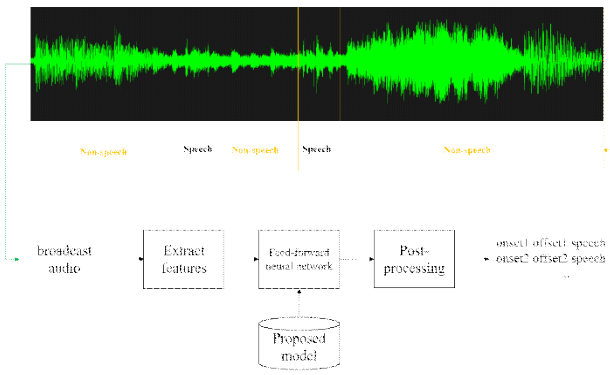

그림 1. 음성 구간 검출 시스템의 전체 구조도
Figure 1. Structure of speech detection system

본 논문에서 제안하는 음성 구간 검출의 전체 구조도와 입출력은 그림 1과 같다. 제안 시스템은 방송 데이터의 오디오 신호를 입력 받아 특징을 추출하고, 심층 학습 모델을 이용하여 프레임 단위로 음성/비음성으로 구분한다. 이후 사후 처리를 통하여 최종적인 음성 구간을 검출하며, 최종 출력은 음성 구간의 시작 시간과 끝 시간으로 나타난다.

2.1. 특징 추출

방송 데이터 또는 음성 데이터를 다루는 심층 학습 모델에서 주로 스펙트로그램과 멜스케일 스펙트로그램, 그리고 멜스케일 켑스트럼 계수(MFCC)가 입력으로 사용된다(Butko, 2011; Castan, 2015; Doukhan, 2018; Tsipas, 2017). 여기서 멜스케일은 인간의 청각 특성을 반영한 필터로 주파수 영역에서의 정보 압축 역할을 한다. 음성의 경우 청각 특성과 많은 연관성이 있기 때문에 멜스케일 필터는 음성의 정보

손실을 최소화하는 특징을 추출할 수 있도록 도와준다. 이렇게 정보의 손실을 최소화하며 입력 차원을 감소시키는 멜스케일 스펙트로그램은 음성 데이터를 다루는 심층 학습 모델에서 좋은 성능을 나타내고 있다(Doukhan, 2018).

본 논문에서는 이러한 점을 고려하여 멜스케일 스펙트로그램을 심층 학습 모델의 입력 특징으로 사용하였으며, 이는 스펙트로그램에서 주파수 영역에 필터를 적용함으로써 추출된다. 하지만 이 특징은 시간 영역에서의 특징 변화 정보는 고려되어 있지 않기 때문에 본 논문에서는 시간 영역에서의 특징들 간의 관계를 학습할 수 있도록 고안된 시간 확장 합성곱을 정의하여 사용한다.

2.2. 시간 확장 합성곱

확장 합성곱은 Yu & Koltun(2015)에 의하여 제안되었으며, 이는 기존 합성곱 층에 확장 비율(dilation rate) 파라미터를 도입한 방법이다. 확장 비율은 커널(kernel) 사이의 간격을 정의하며, 이는 커널이 더 넓은 시야를 갖도록 해주는 역할을 한다. 이러한 성질 때문에 확장 합성곱은 비전 분야에서 영역 분할을 위한 접근법으로 주로 사용되고 있다.

시간 확장 합성곱은 입력으로 스펙트로그램을 사용할 때 정의할 수 있다. 스펙트로그램을 2차원 이미지라고 가정할 때, 보통 가로축은 시간을 의미하고, 세로축은 주파수를 의미한다. 여기서 확장 비율을 가로축에만 적용을 하면 시간에 대한 확장 합성곱, 즉 시간 확장 합성곱이 된다.

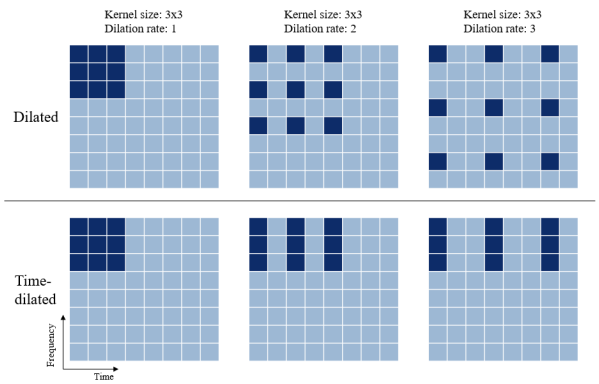

그림 2. 확장 합성곱과 시간 확장 합성곱
Figure 2. Dilated convolution and time-dilated convolution

그림 2의 위 부분은 확장 합성곱이고, 아래 부분은 시간 확장 합성곱이다. 합성곱 층의 커널 사이즈를 3×3으로 가정할 때, 확장 비율의 값이 1, 2, 3으로 변화함에 따라 커널의 변화는 그림 2와 같으며, 이는 확장 합성곱과 시간 확장 합성곱의 차이를 쉽게 볼 수 있다. 확장 합성곱은 확장 비율이 변화함에 따라 커널이 주파수 축(세로축)과 시간 축(가로축)으로 확장되는데 반해, 시간 확장 합성곱은 커널이 시간 축으로만 확장됨을 볼 수 있다. 즉 주파수 영역에서는 기존 커널과 똑 같은 범위로 필터를 적용하지만, 시간 영역에서는 기존 커널보다 더 넓은 시야로 필터를 적용하는 효과를 보여준다. 그러므로 스펙트로그램에 적용한 시간 확장 합성곱은 커널이 시간 영역에 넓은 시야를 가짐으로써 심층 학습 모델이 시간에 따른 특징 벡터의 변화를 학습할 수 있도록 도와준다. 즉, 시간에 대하여 동적인 변화가 많은 데이터 또는 특징을 학습함에 있어 시간 확장 합성곱은 기존 합성곱보다 유리한 특성을 갖는다.

2.3. 제안하는 심층 학습 모델

본 논문에서 제안하는 심층 학습 모델은 그림 3과 같다. 2.1절에서 언급한 것과 같이 제안하는 심층 학습 모델은 멜스케일 스펙트로그램을 추출하여 특징으로 사용한다. 이 때, 멜스케일의 빈(bin) 개수는 64개로 설정하였으며, 시간적 정보를 관찰하기 위하여 분류하고자 하는 프레임의 앞뒤 50개의 프레임을 함께 심층 학습 모델의 입력으로 사용한다. 즉 심층 학습 모델에 입력되는 멜스케일 스펙트로그램의 차원은 64(bin)×101(frame)이다. 첫 번째 합성곱 층은 하나의 확장 비율로 구성되는 기존의 확장 합성곱 층과 다르게 각각 다른 확장 비율을 갖는 시간 확장 합성곱 단(time-dilated convolution unit) 3개를 병렬로 구성한 후 결합한 합성곱 층이며, 그림 3의 A에 해당한다. 즉, 멜스케일 스펙트로그램을 입력으로 하는 시간 확장 합성곱 층 3개를 각각 구성한 후 합성곱 결과를 이어 붙여 다음 합성곱 층의 입력으로 사용한다. 그림 3의 A의 첫 번째(빨간색)는 확장 비율이 1이고, 필터 개수가 2인 시간 확장 합성곱 층이고, 두 번째(노란색)은 확장 비율이 2이고, 필터 개수가 2인 시간 확장 합성공 층이다. 마찬가지로 세 번째(파란색)는 확장 비율이 3이고, 필터 개수가 2인 시간 확장 합성곱 층이며, 이 3개의 합성곱 출력을 이어 붙여 다중 스케일 시간 확장 합성곱 층을 구성하였다. 이러한 구조는 정보가 압축되어 있지 않은 순수한 입력인 멜스케일 스펙트로그램을 다양한 스케일의 시야를 갖는 커널로 관측함으로써 특징 벡터 사이의 시간적 변화 정보를 학습할 수 있도록 도와준다. 그림 3의 A의 시간 확장 합성곱 층은 공통적으로 5×5의 커널 크기로 1의 stirde 크기로 필터를 적용하였으며, 이때 필터 개수는 2개이고, 하이퍼볼릭 탄젠트(hyperbolic tangent) 활성화 함수로 구성하였다. 그림 3의 A부분 이후 3개의 시간 확장 합성곱 층과 2개의 완전 결합 층을 연결하였으며, 이는 각각 그림 3의 B와 C에 해당한다. 3개의 시간 확장 합성곱 층은 각각 1, 2, 4의 확장 비율을 가지고, 16, 32, 64개의 필터 개수를 가지며, 공통적으로 3×3의 커널 크기와 활성화 함수 Rectified linear unit(ReLU)으로 구성된다. 합성곱 층 사이에는 stride size 2의 average pooling을 적용하였고, 마지막 층 이후 softmax를 적용하여 음성/비음성에 대한 확률을 출력하도록 구성하였다. 제안 모델에서 설정한 확장 비율 파라미터는 실험을 통하여 가장 적절한 파라미터로 선정된 것이다.

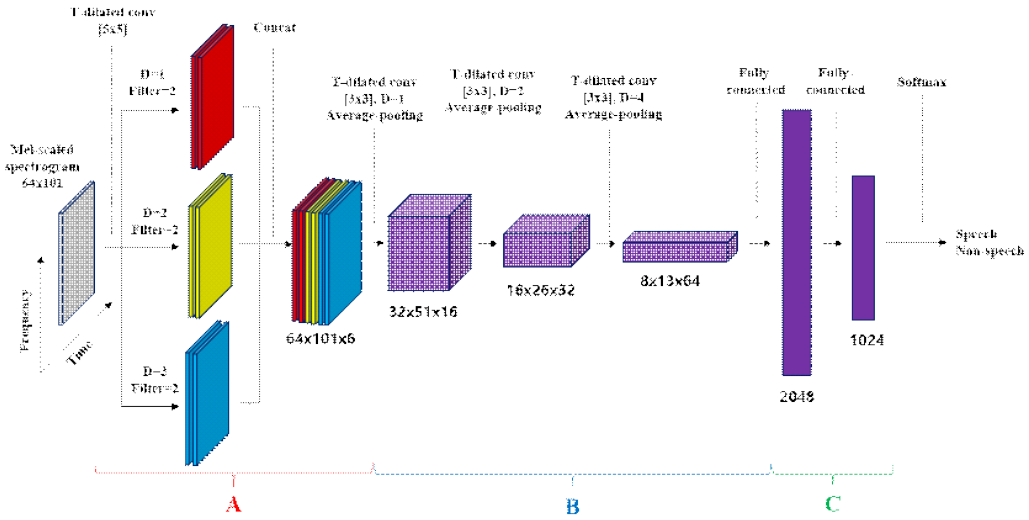

그림 3. 제안하는 심층 학습 모델 구조
Figure 3. Proposed deep learning model

2.4. 사후 처리

심층 학습 모델의 출력인 음성/비음성(speech/non-speech) 확률은 0.5를 기준으로 음성/비음성으로 분류된다. 하지만 이는 프레임 단위로 계산되기 때문에 평활화(smoothing)가 필요하다. 본 시스템에서는 프레임 단위의 결과를 평활화하기 위하여 중간 값 필터(median filter)를 사용하였다. 중간 값 필터는 설정한 윈도우내의 값들을 중간 값으로 치환하여 고주파 특성을 갖는 noise 성분들을 제거해 주는 역할을 한다. 중간 값 필터의 윈도우 크기는 101 프레임(1.01초)으로 설정하여 최소 음성 구간이 1초가 넘도록 평활화를 하였다.

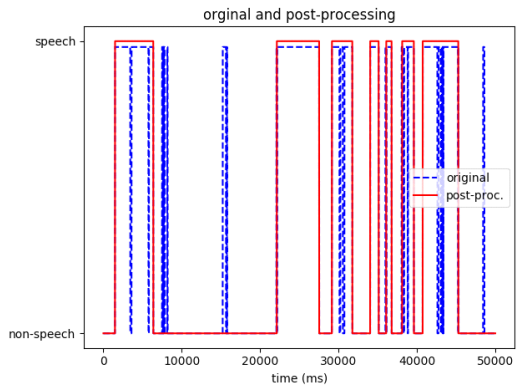

그림 4. 사후 처리 결과
Figure 4. Result of post-processing

그림 4는 심층 학습 모델 출력의 원본 결과와 중간값 필터를 이용하여 사후처리가 수행된 후의 값을 보여주고 있다. 파란 점선은 원본 결과이며, 빨간 실선은 사후처리 후 결과이다. 원본 결과와 달리 사후 처리 후 결과는 매우 짧은 음성 구간은 제거하고, 음성 구간 사이의 짧은 비음성 구간은 병합하여 결과를 평활화하고 있음을 알 수 있다.

3. 실험 및 결과

3.1. 데이터베이스

본 연구에서는 음성 구간 검출을 위한 심층 학습 모델을 학습하기 위하여 방송 데이터를 수집하였다. 수집한 방송 데이터는 다큐멘터리, 쇼, 드라마, 어린이 장르를 포함하는 44시간 분량이며, 수동으로 음성 구간을 모두 표시하였다. 본 연구에서는 학습을 위하여 32시간(train set), 검증을 위하여 12시간(test set1)으로 데이터를 분할하였고, 학습 데이터와 검증 데이터의 프로그램명은 겹치지 않으며, 검증 데이터는 다양한 장르를 포함하도록 구성하였다. 이 데이터와 별개로 모델 선정과 파라미터 결정을 위하여 드라마 데이터 3시간(development set)을 수집하여 사용하였다.

본 연구에서는 제안 시스템의 객관적인 검증을 심화하기 위하여 다른 나라의 방송데이터를 수집하여 검증에 사용하였다. 외국어 방송 데이터는 다양한 장르를 포함하는 영국 영어 방송 데이터 8시간(test set2)과 스페인어 방송 데이터 12시간(test set3)으로 구성되었으며, 한국어 데이터와 동일하게 수동으로 음성 구간을 표시하여 사용하였고, 모든 오디오 데이터는 16,000 Hz 샘플링 주파수와 16 bit, 모노(mono)로 구성되었다.

표 1은 본 연구에서 사용한 데이터의 분량 및 음성 구간 비율을 보여준다. 방송 데이터는 그 종류에 따라 음성 구간의 비율이 최소 30.3%, 최대 64.8%로 나타났다.

표 1. 데이터베이스
Table 1. Database

	이름	총 시간	음성 구간	비율(%)
Train	Korean_broadcast_32h	32	21	64.8
Develop	Korean_drama_3h	3	0.9	30.3
Test	Korean_broadcast_12h	12	7.6	59.4
	British_broadcast_8h	8	4.3	53.9
	Spanish_broadcast_12h	12	7.7	64.2

3.2. 비교 모델

본 연구에서는 제안 모델의 성능 비교를 위하여 비교 모델을 선정 및 구현하여 성능 실험을 수행하였다. 성능 비교를 위한 첫 번째 모델은 범용적으로 사용하는 LeNet-5(LeCun et al., 1998)에 기반하여 구성한 합성곱 신경망 구조의 모델(CNN)로 LeNet-5와의 차이점은 합성곱 층과 pooling 층이 3개이고, 완전결합 층은 2개로 구성되었으며, pooling 층은 average pooling 기법을 사용하였다. 두 번째 모델은 특징 벡터의 시간적 변화를 고려할 수 있는 순환 신경망(recurrent neural network, RNN)(Sak et al., 2014) 구조로 구성하였으며, 이는 2개의 양방향 순환 게이트 유닛(bidirectional gated recurrent unit, bi-GRU)(Lu & Duan, 2017) 층으로 구성하였다. 세 번째 모델은 합성곱 신경망과 순환 신경망을 결합한 합성곱 순환 신경망(convolutional recurrent neural network, CRNN) 구조로 3개의 합성곱 층과 pooling 층 이후에 시간 축으로 순환하는 양방향 순환 신경망 층 2개를 연결하여 구성하였다(Zuo et al., 2015). 마지막으로 합성곱 신경망 구조에서 강인하다고 알려져 있는 ResNet(He et al., 2016) 모델도 학습하여 성능 검증을 수행하였다. 이 모델은 영상 및 음성 관련 연구에서 널리 사용되고 있으며, 본 연구에서는 TensorFlow library[1]를 이용하여 실험을 수행하였고, 이 모델은 총 50개의 레이어 개수를 가진다. 마지막으로 본 논문에서 제안한 다중 스케일 시간 확장 합성곱 층과 시간 확장 합성곱 층이 성능에 미치는 영향을 확인하기 위하여 각각을 따로 적용하여 성능 실험을 수행하였다. 실험에서 비교 대상으로 사용되는 CNN 모델은 그림 3의 구조에서 시간 확장 비율이 없는 기존의 합성곱 층을 적용한 B+C로 구성된다. 여기에 다중 스케일 시간 확장 합성곱 층(그림 3의 A부분)을 추가한 모델을 CNN+A라 하고, CNN 모델의 합성곱 층에 시간 확장 비율(그림 3의 B부분)을 추가한 CNN+B 모델이라 정의한다. 마지막으로 그림 3의 A와 B부분을 모두 추가한 모델은 CNN+A+B이다.

3.3. 모델 학습 및 선정

본 연구는 제안 모델과 비교 모델의 객관적 비교를 위하여 같은 파라미터로 특징 추출 및 학습을 진행하였다. 스펙트로그램을 추출하기 위하여 10 ms마다 25 ms의 윈도우 크기로 STFT(short-time Fourier transform)을 512 크기로 수행 후 log power coefficient를 계산하였고, 멜스케일 스펙트로그램은 스펙트로그램에 64개의 빈으로 구성된 멜스케일 필터를 적용하여 추출하였다. 심층 학습 모델 학습은 leaning rate 0.001, epoch 42, minibatch 300, dropout probability 0.4로 수행되었다. 제안 모델과 비교 모델들은 모델마다 수렴 또는 과적합(overfitting)되는 구간이 다를 수 있으므로 epoch 횟수를 크게 설정하였다.

본 연구에서는 심층 학습 모델을 학습하면서 과적합에 의한 성능 저하를 배제하기 위해 일정 반복 횟수(1,000 iteration)마다 모델을 저장하고, develop 데이터 셋(Korean_ drama_3h)에 대한 성능을 확인하였다. 그리고 이 중 가장 높은 성능의 모델을 선정하여 사용하였다. 비교 모델과 제안 모델은 10,000 iteration(약 15 epoch) 근방에서 가장 높은 성능을 보여주었다.

3.4. 성능 평가

본 연구에서는 제안한 모델과 비교 모델의 성능을 평가하기 위하여 sed_eval 도구를 사용하였다(Mesaros et al., 2016). 이 도구는 프레임 단위의 F-score, precision, recall을 성능 지표로 보여주며, 각 지표는 True Positive(TP), False Positive(FP), False Negative(FN)으로 계산되고, 그 식은 다음과 같다.

$$\mathrm{Precision}\ (P) = \frac{TP}{TP+FP} \tag{1}$$

$$Recall\ (R) = \frac{TP}{TP+FN} \tag{2}$$

$$F-score = 2\frac{P\times R}{P+R} \tag{3}$$

3.4. 실험 결과

본 연구에서는 앞서 언급하였던 것과 같이 한국 드라마 3시간 데이터를 이용하여 모델 선정을 하였다. 표 2는 가장 좋은 성능으로 선정된 모델들의 평가 결과이다. 제안 모델을 제외한 비교 모델 중 CNN 모델의 성능이 88.8%의 F-score로 가장 높게 나타났다. 제안 모델인 CNN+A+B는 89.3%의 F-score로 가장 높은 성능을 보여주었으며, 다중 스케일 시간 확장 합성곱 층만 추가한 CNN+A의 성능은 CNN 모델 성능보다 높은 89.0%의 F-score를 보여주었다. 다만 시간 확장 합성곱 층만 추가한 CNN+B 모델의 성능은 CNN 모델 성능보다 낮은 87.10%의 F-score로 나타났다. 본 연구에서는 이렇게 선별된 모델들을 이용하여 검증 데이터의 성능 평가를 수행하였다.

1 https://github.com/tensorflow/tensorflow/blob/master/tensorflow/contrib/slim/python/slim/nets/resnet_v2.py

표 2. Korean drama 3h 데이터(dev)의 평가 결과
Table 2. Evaluation result of Korean drama 3h

Structure	F-score	Precision	Recall
CNN (LeCun et al.)	88.8	90.2	87.6
bi-GRU (Lu & Duan)	88.2	88.2	88.3
CRNN (Zuo et al.)	87.1	87.2	87.0
ResNet (He et al.)	88.1	89.4	86.9
CNN+A	89.0	89.0	89.1
CNN+B	87.1	90.8	83.7
CNN+A+B	89.3	89.6	89.1

표 3은 한국 방송 데이터 12시간에 대한 평가 결과를 보여준다. 한국 드라마 데이터 3시간 결과와 다르게 제안 모델을 제외한 모델 중 ResNet 모델의 성능이 91.5%의 F-score로 가장 높게 나타났다. 하지만 이 데이터에서도 CNN+A+B 모델은 91.7의 F-score로 가장 높게 나타났으며, CNN+A 모델의 성능 또한 ResNet보다 높은 91.6%의 F-score를 보여주고 있다. 드라마 데이터의 양상과 마찬가지로 CNN+B의 성능은 CNN의 성능보다 낮은 87.1 %의 F-score로 나타났다.

표 3. Korean broadcast 12h 데이터의 평가 결과
Table 3. Evaluation result of Korean broadcast 12h

Structure	F-score	Precision	Recall
CNN (LeCun et al.)	90.1	94.8	85.9
bi-GRU (Lu & Duan)	90.1	94.0	86.5
CRNN (Zuo et al.)	91.0	92.6	89.5
ResNet (He et al.)	91.5	93.2	89.9
CNN+A	91.6	94.2	89.4
CNN+B	89.6	94.9	84.9
CNN+A+B	91.7	94.2	89.2

표 4와 5는 영국과 스페인 방송 데이터에 대한 평가 결과이다. 본 연구에서 사용한 학습 데이터에는 영국 영어와 스페인어가 없음에도 전반적으로 85% 이상의 좋은 성능을 보여주고 있다. 그리고 영국 영어와 스페인어 데이터에서 모두 CNN+A+B 모델이 87.9%와 92.6%의 F-score로 비교 모델보다 좋은 성능을 보여주었다. 다만 한국어 데이터의 성능과 다르게 영국 영어 데이터와 스페인어 데이터에서는 CNN+A+B 모델보다 CNN+A 모델의 성능이 89.3%와 93.1%의 F-score로 높게 나타났으며, 이 모델의 성능이 가장 높은 성능을 보여주었다. 또한 영국 영어 데이터에서 제안 모델을 제외한 모델 중 CRNN 모델이 87.7%의 F-score로 좋은 성능을 보여주었으며, 스페인어 데이터에서는 ResNet 모델이 92.5%의 F-score로 좋은 성능을 보여주었다.

표 4. British broadcast 8h 데이터의 평가 결과
Table 4. Evaluation result of British broadcast 8h

Structure	F-score	Precision	Recall
CNN (LeCun et al.)	86.5	91.7	81.8
bi-GRU (Lu & Duan)	85.1	89.4	81.3
CRNN (Zuo et al.)	87.7	88.7	86.7
ResNet (He et al.)	87.3	88.8	85.8
CNN+A	89.3	92.4	86.4
CNN+B	85.1	91.6	79.5
CNN+A+B	87.9	91.4	84.7

표 5. Spanish broadcast 12h 데이터의 평가 결과
Table 5. Evaluation result of Spanish broadcast 12h

Structure	F-score	Precision	Recall
CNN (LeCun et al.)	92.4	93.9	90.9
bi-GRU (Lu & Duan)	90.9	92.8	89.2
CRNN (Zuo et al.)	92.1	91.8	92.4
ResNet (He et al.)	92.5	92.5	92.5
CNN+A	93.1	93.5	92.8
CNN+B	90.9	92.8	89.2
CNN+A+B	92.6	93.2	92.1

마지막으로 표 6은 CNN+A+B 모델에 사후 처리를 적용한 성능과 적용하지 않은 성능을 보여준다. 프레임 단위의 결과를 평활화하기 위하여 적용한 중간 값 필터는 확실히 모델의 성능을 높여주고 있음을 볼 수 있다. 다만 복잡한 알고리즘이 아닌 간단한 필터 적용이기 때문에 성능 상승폭이 작다는 한계점이 존재한다.

표 6. CNN+A+B 모델의 사후 처리 성능 비교
Table 6. Comparison of post-processing for CNN+A+B model

Data	Post-proc.	F-score	Precision	Recall
Korean drama 3h	적용	89.3	89.6	89.1
	미적용	88.5	88.3	88.8
Korean broadcast 12h	적용	91.7	94.2	89.2
	미적용	90.8	93.6	88.2
British broadcast 8h	적용	87.9	91.4	84.7
	미적용	87.1	90.3	84.1
Spanish broadcast 12h	적용	92.6	93.2	92.1
	미적용	92.1	92.7	91.6

4. 결론

본 논문에서는 방송데이터에서 다른 오디오 신호와 혼합되어 있는 음성 구간을 검출하기 위하여 새로운 심층 학습 모델을 제안하였다. 제안한 모델은 음성의 시간 축에서의 동적 변화가 많은 특성을 고려하기 위하여 다중 스케일 시간 확장 합성곱 층과 시간 확장 합성곱 층을 제안하여 사용하였다. 본 논문에서 제안 모델의 성능은 다른 심층 학습 모델보다 높은 성능을 보여주었다. 제안 모델은 한국어 드라마 데이터 3시간과 한국 방송 데이터 12시간을 사용한 실험에서 89.3과 91.7의 가장 높은 F-score를 나타냈다. 또한 영국 영어와 스페인어 방송 데이터에 대해서도 87.9와 92.6

의 F-score로 가장 높은 성능을 보여주었다. 그리고 영국 영어와 스페인어 데이터에서는 다중 스케일 시간 확장 합성곱 층만 사용한 모델(CNN+A)이 89.3%와 93.1%의 F-score로 시간 확장 합성곱 층을 사용한 모델(CNN+A+B)보다 높은 성능을 보여주었다. 이렇듯 제안 모델은 방송 데이터에 종류와 상관없이 가장 높은 성능을 보여 주였으며, 이는 본 논문에서 제안한 시간 확장 합성곱 층이 특징 벡터 간의 시간적 변화 정보를 잘 관찰하고, 음성 구간 검출 성능 향상에 기여하고 있다고 판단할 수 있으며, 특히 영국 영어 데이터와 스페인어 데이터에서는 다중 스케일 시간 확장 합성곱 층이 성능 향상에 더욱 많은 기여를 하는 것으로 보여졌다.

결론적으로 본 논문에서는 음성의 특성을 고려한 심층 학습 모델 구조를 제안하였고, 실험 결과를 통하여 제안 모델이 음성 구간 검출에 적합한 모델임을 검증하였다. 추후에는 최신의 심층 신경망 모델에 다중 스케일 시간 합성곱 층을 적용하는 연구를 할 계획이다.

References

Butko, T., & Nadeu, C. (2011). Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, *2011*(1), 1-10.

Castan, D., Tavarez, D., Lopez-Otero, P., Franco-Pedroso, J., Delgado, H., Navas, E., Docio-Fernandez, L., ... Lleida, E. (2015). Albayzin-2014 evaluation: audio segmentation and classification in broadcast news domains. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(33), 1-9.

Doukhan, D., Lechapt, E., Evrard, M., & Carrive, J. (2018). Ina's MIREX 2018 music and speech detection system. *Music Information Retrieval Evaluation eXchange (MIREX)*.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788-798.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Lu, R., & Duan, Z. (2017). Bidirectional GRU for sound event detection. *Detection and Classification of Acoustic Scenes and Events*.

Mesaros, A., Heittola, T., & Virtanen, T. (2016). Metrics for polyphonic sound event detection. *Applied Sciences*, *6*(6), 162.

Mirex (2015). Music/speech classification and detection. Retrieved from http://www.music-ir.org/mirex/wiki/2015:Music/Speech_Classifi-cation_and_Detection

Mirex (2018). Music and/or speech detection. Retrieved from http://www.music-ir.org/mirex/wiki/2018:Music_and/or_Speech_Detection

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *15th Annual Conference of the International Speech Communication Association (Interspeech-2014)* (pp. 338-342). Singapore.

Tsipas, N., Vrysis, L., Dimoulas, C., & Papanikolaou, G. (2017). Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination. *Multimedia Tools and Applications*, *76*(24), 25603-25621.

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. Retrieved from https://arxiv.org/abs/1511.07122.

Zhang, Q., Cui, Z., Niu, X., Geng, S., & Qiao, Y. (2017). Image segmentation with pyramid dilated convolution based on ResNet and U-Net. In *International Conference on Neural Information Processing* (pp. 364-372).

Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., & Chen, Y. (2015, June). Convolutional recurrent neural networks: Learning spatial dependencies for image representation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 18-26).

• **장병용 (Byeong-Yong Jang)**
충북대학교 전자공학부 박사과정
충북 청주시 서원구 충대로 1
Tel: 043-261-3374
E-mail: byjang@cbnu.ac.kr
관심분야: 음성인식, 심층 학습, 음성 및 오디오 신호 처리

• **권오욱 (Oh-Wook Kwon)** 교신저자
충북대학교 전자공학부 교수
충북 청주시 서원구 충대로 1
Tel: 043-261-3374
E-mail: owkwon@cbnu.ac.kr
관심분야: 음성인식, 심층 학습, 음성 및 오디오 신호 처리

다중 스케일 시간 확장 합성곱 신경망을 이용한 방송 콘텐츠에서의 음성 검출*

장병용 · 권오욱

충북대학교 전자공학부

국문초록

본 논문에서는 방송 콘텐츠에서 음성 구간 검출을 효과적으로 할 수 있는 심층 학습 모델 구조를 제안한다. 또한 특징 벡터의 시간적 변화를 학습하기 위한 다중 스케일 시간 확장 합성곱 층을 제안한다. 본 논문에서 제안한 모델의 성능을 검증하기 위하여 여러 개의 비교 모델을 구현하고, 프레임 단위의 F-score, precision, recall을 계산하여 보여준다. 제안 모델과 비교 모델은 모두 같은 학습 데이터로 학습되었으며, 모든 모델은 다양한 장르(드라마, 뉴스, 다큐멘터리 등)로 구성되어 있는 한국 방송데이터 32시간을 이용하여 모델을 학습되었다. 제안 모델은 한국 방송데이터에서 F-score 91.7%로 가장 좋은 성능을 보여주었다. 또한 영국과 스페인 방송 데이터에서도 F-score 87.9%와 92.6%로 가장 높은 성능을 보여주었다. 결과적으로 본 논문의 제안 모델은 특징 벡터의 시간적 변화를 학습하여 음성 구간 검출 성능 향상에 기여할 수 있었다.

핵심어: 음성 구간 검출, 다중 스케일 시간 확장 합성곱, 심층 학습 모델, 방송 데이터

* 본 연구는 문화체육관광부 및 한국저작권위원회의 2019 년도 저작권기술개발사업의 연구결과로 수행되었다[2018-micro-9500, 음악 및 동영상 모니터링을 위한 지능형 마이크로 식별 기술개발].