

일반논문 (Regular Paper)

방송공학회논문지 제24권 제3호, 2019년 5월 (JBE Vol. 24, No. 3, May 2019)

<https://doi.org/10.5909/JBE.2019.24.3.472>

ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

심층신경망을 이용한 시간 영역 음향 이벤트 검출 알고리즘

김 범 준^{a)}, 문 현 기^{a)}, 박 성 옥^{b)}, 정 영 호^{c)}, 박 영 철^{a)†}

Time-domain Sound Event Detection Algorithm Using Deep Neural Network

Bum-Jun Kim^{a)}, Hyeonggi Moon^{a)}, Sung-Wook Park^{b)}, Youngho Jeong^{c)}, and Young-Cheol Park^{a)†}

요 약

본 논문에서는 심층신경망을 이용한 시간 영역 음향 이벤트 검출 알고리즘을 제시한다. 본 시스템에서는 주파수 영역으로 변환되지 않은 시간 영역의 음향 데이터를 심층신경망의 입력으로 사용한다. 전반적인 구조는 CRNN 구조를 사용하였으며, GLU, ResNet, Squeeze-and-excitation 블록을 적용하였다. 그리고 여러 계층에서 추출된 특징을 함께 고려하는 구조를 제안하였다. 또한 본 연구에서는 강한 라벨이 있는 훈련 데이터를 확보하는 것이 현실적으로 어렵다는 전제 아래에서 약한 라벨이 있는 훈련 데이터 약간 그리고 다수의 라벨이 없는 훈련 데이터를 활용하여 훈련을 수행하였다. 적은 수의 훈련 데이터를 효과적으로 사용하기 위해 타임 스트레칭, 피치 변화, 동적 영역 압축, 블록 혼합 등의 데이터 증강 방법을 적용하였다. 라벨이 없는 데이터에는 의사 라벨을 붙여 부족한 훈련 데이터를 보완하였다. 본 논문에서 제안한 신경망과 데이터 증강 방법을 사용하는 경우, 종래의 방식으로 CRNN 구조의 신경망을 훈련하여 사용하는 경우보다, 음향 이벤트 검출 성능이 약 6 % (f-score 기준)가 개선되었다.

Abstract

This paper proposes a time-domain sound event detection algorithm using DNN (Deep Neural Network). In this system, time domain sound waveform data which is not converted into the frequency domain is used as input to the DNN. The overall structure uses CRNN structure, and GLU, ResNet, and Squeeze-and-excitation blocks are applied. And proposed structure uses structure that considers features extracted from several layers together. In addition, under the assumption that it is practically difficult to obtain training data with strong labels, this study conducted training using a small number of weakly labeled training data and a large number of unlabeled training data. To efficiently use a small number of training data, the training data applied data augmentation methods such as time stretching, pitch change, DRC (dynamic range compression), and block mixing. Unlabeled data was supplemented with insufficient training data by attaching a pseudo-label. In the case of using the neural network and the data augmentation method proposed in this paper, the sound event detection performance is improved by about 6 % (based on the f-score), compared with the case where the neural network of the CRNN structure is used by training in the conventional method.

Keyword : Sound Event Detection (SED), Time-domain based DNN structure, ResGLU-SE, Data augmentation, pseudo-labeling

a) 연세대학교(Yonsei University)

b) 강릉원주대학교(Gangneung-Wonju National University)

c) 한국전자통신연구원(ETRI)

† Corresponding Author : 박영철(Young-Cheol Park)

E-mail: young00@yonsei.ac.kr

Tel: +82-033-760-2756

ORCID: <https://orcid.org/0000-0003-3274-076X>

※ 본 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구이다. (No.2017-0-00050, 신체기능의 이상이나 저하를 극복하기 위한 휴먼 청각 및 근력 증강 원천 기술 개발)

※ This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support)

· Manuscript received December 4, 2018; Revised April 22, 2019; Accepted April 22, 2019.

Copyright © 2016 Korean Institute of Broadcast and Media Engineers. All rights reserved.

“This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered.”

I. 서론

가정에서 음성으로 대화를 나눌 수 있는 인공지능 스피커가 다량 보급되고 있으며, 이러한 인공지능 스피커는 음성과 함께 주변 음향을 분석하여 상황을 이해하고, 이를 바탕으로 향상된 서비스를 제공하고자 시도하고 있다. 상황을 이해하기 위해서는 입력되는 주변 음향 신호를 분석하여 특정 시간에 어떤 사건(즉, 이벤트)이 있는지를 인식하는 기능이 필요하며, 이를 음향 이벤트 검출(SED: sound event detection)라고 한다^[1]. 음향 이벤트 검출 기능을 이용한다면 구조 요청, 긴급 출동, 정보 검색^[2]과 같은 서비스를 제공할 수 있다.

음향 이벤트 검출 알고리즘은 입력된 음향 데이터에서 언제 어떤 이벤트가 발생했는지를 알아낸다. 종래의 이벤트 검출 알고리즘은 SVM (Support Vector Machines), HMM (Hidden Markov Models)을 기반으로 설계되었으나, 최근에는 음성 신호 처리에서 높은 성능을 보이는 심층신경망을 기반으로 보다 높은 성능을 보이고 있다^[3]. 심층신경망은 기계학습의 한 종류로 비선형 연산들을 중첩을 통하여 입력 신호의 추상적인 특징들을 스스로 추출할 수 있는 능력이 있다.

본 논문에서는 시계열 데이터인 음향 신호를 처리하기 위하여 CRNN (Convolutional Recurrent Neural Networks)^[4]을 기본 구조로 사용하였다. CRNN은 합성곱 신경망(CNN: Convolutional Neural Network)과 순환 신경망(RNN: Recurrent Neural Network), 완전 접속망(FNN: Fully-connected Neural Network)이 결합된 구조이다. 또한 기본 구조에서 GLU (Gated Linear Units)^[5], ResNet (Residual Neural Network)^[6], SE 블록 (Squeeze-and-excitation block)^[7]을 결합하여 추가적인 성능 향상을 달성하였다.

본 논문에서 음향 이벤트 검출용 심층신경망 훈련에 사용하는 훈련용 데이터 세트는 구글의 오디오 세트의 일부를 사용하였다. 논문에서 사용한 데이터 세트는 적은 수의 약한 라벨(weakly labeled)이 부여된 데이터와 다량의 라벨이 없는(unlabeled) 데이터로 구성되어 있다. 이는 모든 훈련용 데이터에 정교한 강한 라벨(strong label)을 부여하기 어려운 현실을 반영한 것이다. 여기서 약한 라벨은 정답인

라벨이 데이터에 부여 되었으나 그 라벨이 데이터의 어떤 시점에 해당하는지를 설명하는 시간 정보가 없는 라벨을 말하며, 강한 라벨은 시간 정보를 포함하는 라벨을 말한다. 그러므로 음향 이벤트 검출은 입력 시계열 신호에 강한 라벨을 부여하는 작업이라고 할 수 있다.

훈련 데이터의 수가 많아지고, 다양해질수록 성능이 좋아지는 것이 일반적인 심층신경망의 특성이다. 성능을 높이기 위해 주어진 적은 훈련 데이터로 보다 많은 데이터를 확보하는 데이터 증강 방법을 추가로 사용하였다. 이렇게 하여 총 세 종류의 훈련용 데이터들, 즉 약한 라벨 데이터, 데이터 증강방법으로 증강된 데이터, 라벨이 없는 데이터를 확보하였다. 사용된 훈련용 데이터는 제안된 구조가 최적의 성능을 내는 것을 확인하기 위해 사용되었다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 본 연구에서 제안한 심층신경망의 구조를 소개하였고, 3장에서 데이터 증강 방법에 대해 소개하였다. 4장에서 데이터 세트, 훈련 절차 및 평가 방법을 설명하였으며, 5장에서 제안된 알고리즘의 성능을 분석하였다. 그리고 6장에서 논문을 마무리하였다.

II. 심층신경망 구조

본 논문에서는 Fig. 1과 같은 변형된 CRNN 구조^[8]를 가지는 심층신경망을 사용하였다. (a)는 의사 라벨(pseudo label) 적용을 위해 사용한 ‘의사 라벨 적용 신경망’이며, (b)는 음향 이벤트 검출에 사용된 ‘SED 신경망’의 구조를 보여준다. 제안된 두 신경망은 모두 입력된 시간축 음향 신호 한 프레임에 대하여 1차 합성곱 계층^[4]을 적용하여 시간축 특징을 추출한다. 제안된 신경망은 추출된 시간축 특징을 ResNet^[6]과 GLU^[5], SE^[7]가 결합된 ResGLU-SE 계층에 연속적으로 통과시켜 입력 신호가 가지고 있는 추상적인 특징들을 계층별로 추출한다, 계층별로 추출된 특징들은 추상화 정도가 각각 다르다. 여러 계층에서 추출된 추상적인 특징들은 각각 시간축에서의 상관관계를 고려하는 양방향 순환 신경망 계층을 통과한다. 제안된 신경망은 3 개의 추상화 레벨에서 추출한 특징을 각각의 순환 신경망 계층으로 처리하고, 그 출력들을 종합적으로 활용하기 위하여

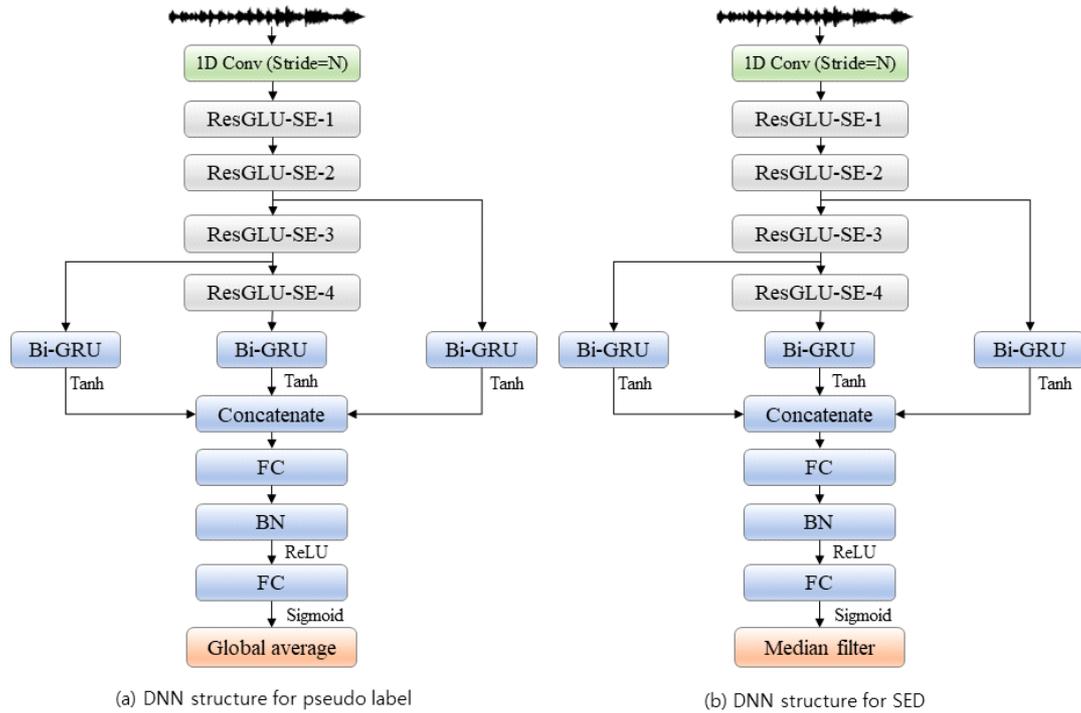


그림 1. 제안된 DNN 구조의 블럭 다이어그램, (a) 의사 라벨 DNN 구조, (b) SED DNN 구조
 Fig. 1. Block diagram of the proposed DNN structure, (a) DNN structure for pseudo label, (b) DNN structure for SED

concatenate 계층으로 이들을 결합한 후, FC 계층의 입력으로 사용하였다.

1. 1차 합성곱 계층

본 논문에서는 시간축 신호를 신경망의 입력으로 사용하였다. 잘 훈련된 합성곱 계층의 계수는 저주파수 대역에

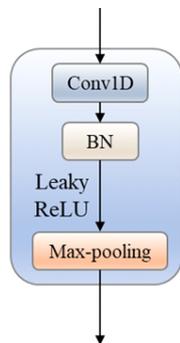


그림 2 . N 스트라이드 1차원 합성곱 계층의 블럭 다이어그램
 Fig. 2. Block diagram of the 1D convolution layer with stride N

서 높은 분해능을 갖는 필터 बैं크와 유사한 주파수 영역 특성을 갖는다^[9]. 그러므로 Fig. 2와 같이 구성된 1차 합성곱 계층은 입력된 시간축 신호의 저주파 영역에 해당하는 특징들을 섬세하게 추출하고^[4], 신경망의 상위 계층들은 추출된 저주파 영역 특징들을 위주로 하여 분류에 필요한 추상적인 특징들을 파악한다.

1차원 합성곱 연산을 고속 병렬 처리하기 위하여 1차원인 시간 영역 음향 데이터를 2차원인 행렬의 형태로 바꾸어 입력으로 사용한다. 이를 위하여 일정 샘플 수만큼 이동하는 윈도우를 통하여 입력된 신호를 행벡터로 하는 행렬을 구성한다. 제안된 구조에서 사용된 1차 합성곱 계층은 최하위 계층에 위치해 있다. 합성곱 계층 이후에는 배치 정규화, leaky ReLU 활성화 계층, 최대 풀링을 적용 하였다.

2. GLU

GLU는 심층신경망 구조의 문제점 중 하나인 기울기 소실 문제를 줄이는데 도움이 된다^[10]. 이전 계층의 출력 두

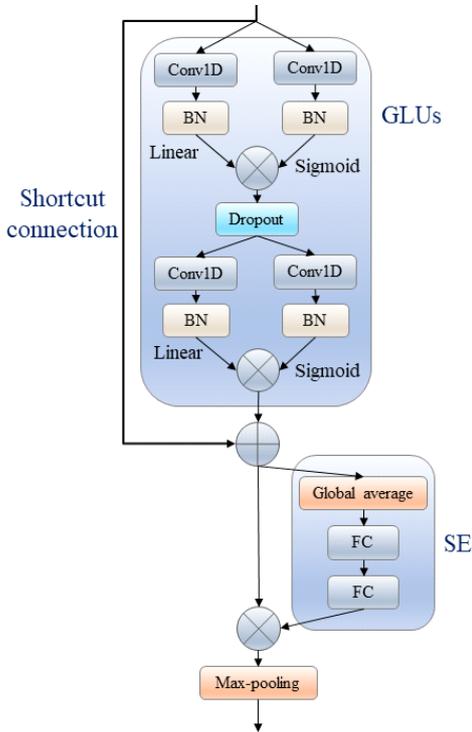


그림 3. ResGLU-SE 블록의 블록 다이어그램
 Fig. 3. Block diagram of the ResGLU-SE block

개의 합성곱, 배치 정규화 계층을 병렬로 통과하게 되는 데, 각각 선형, 시그모이드 함수를 적용한다. 그리고 그 결과들은 서로 곱해지는데, 이때 시그모이드 함수의 결과는 선형 함수의 결과가 얼마만큼의 크기로 출력되어야 하는 지를 결정한다. 그러므로 GLU의 두 합성곱들은 통상의 비선형 활성화 함수 역할을 담당하는 것으로 해석할 수 있다. 본 논문에서는 위의 과정이 두 번 반복 하되, 중간에 드롭아웃 (20%)을 적용하고 하나의 블록으로 묶었으며, Fig. 3에서는 GLUs로 표기되어 있다.

3. ResNet

이전 계층의 출력을 shortcut 연결을 통해 GLUs 블록을 건너뛰어서 기울기 소실 문제를 줄이는 역할을 한다. 바로 가기 연결로 인해 초기 훈련 속도가 빨라진다⁶⁾. 시간 영역 입력을 사용하는 경우 주파수 영역 입력과 동일한 성능을 내기 위해 더 많은 데이터로 훈련을 수행시켜야 하는

문제가 있다는 점을 고려한다면, ResNet의 훈련 속도의 증가는 제안된 구조에서 큰 이점으로 작용한다. 본 논문에서는 GLUs와 바로 가기 연결을 함께 ResGLU 블록이라고 이름 지었다.

4. SE block

SE 블록은 매 입력에 대하여 채널 단위의 특징을 재조정한다⁷⁾. 제안된 구조에서 SE 블록은 바로 가기 연결 이후에 배치하여 ResGLU 블록의 출력으로 얻은 특징을 더욱 도드라지게 하였다. 전역 평균은 ResGLU 블록의 3차원 텐서 출력에 대하여 채널 단위로 평균을 취하는 역할을 한다. 전역 평균 풀링을 통해 얻은 평균 정보는 채널 간의 상호 의존성을 모델링하는 두 개의 FC 계층을 거친다. 첫 번째 FC 계층의 노드의 개수는 전역 평균 풀링의 출력 데이터의 채널 개수와 동일하다. 두 번째 FC 계층의 노드의 개수는 SE 블록의 첫 번째 FC 계층의 노드의 개수의 12배가 되도록 정하였다. 이 배율은 그리드 탐색을 통해 결정하였다. 본 논문에서는 ResGLU 블록과 SE 블록을 함께 ResGLU-SE 블록이라고 이름 지었다.

5. Multi-level feature의 결합

Fig. 1과 같이 제안된 구조의 ResGLU-SE 2, 3, 4 번째 블록의 출력은 각각 양방향 순환 신경망 계층을 통과한 뒤 하나의 벡터로 결합된다. 이러한 구조는, 입력 신호가 여러 ResGLU-SE 계층을 통과되면서 각각 서로 다른 추상화 수준의, 점차 추상적인, 특징으로 추출되며, 한 계층의 특징만을 사용하는 것 보다 여러 수준의 추상화 특징을 사용하는 것이 보다 좋은 결과를 보이는 경향이 있다는 자체 실험 결과를 반영한 구조이다.

결합된 벡터는 두 개의 FC 계층을 거친다. 첫 번째 FC 계층의 크기는 입력 벡터의 크기와 동일하며, 배치 정규화를 거친 후 ReLU 활성화 함수를 통과한다. ReLU 활성화 함수를 거쳐 나온 결과는 두 번째 FC 계층의 입력으로 사용되는데, 두 번째 계층은 이벤트의 수만큼 노드를 갖는다.

III. 데이터 증강

심층신경망 훈련에 있어 사용되는 훈련 데이터(약한 라벨이 붙은)의 수가 1,578 개로 매우 적기 때문에 신경망이 충분한 훈련이 되지 못하는 문제가 있다. 이를 개선하기 위해 데이터 증강 방법을 사용한다. 전통적인 방법인 타임 스트레칭, 피치 변화, 동적 영역 압축, 그리고 블럭 혼합 등 네 가지 방법을 이용하여 부족한 데이터를 보완한다. 이 중에서 타임 스트레칭, 피치 변화, 동적 영역 압축은 약한 라벨을 검출하는 오디오 표식(audio tagging)을 위한 데이터 증강 방법으로 널리 사용되는 방법들이며^[11], 본 연구에서는 타임 스트레칭 비율을 [0.81, 0.93, 1.07, 1.23]로, 피치

변화를 [-2, -1, 1, 2] 음 만큼 주었다. 또한 DRC를 위해서는 Fig. 4의 두 번째 그림과 같은 변화를 주었다.

1. 블럭 혼합

블럭 혼합은 오디오 표식에서는 사용되지 않으며, 강한 라벨을 검출하는 음향 이벤트 검출을 위한 훈련에서 사용된다. 블럭 혼합은 특정 조건에 부합하는 서로 다른 두 음향 데이터를 음향 데이터 전 구간에 대하여 혼합한다. 본 논문에서는 훈련된 신경망을 최대한 일반화시키기 위하여 두 음향 데이터를 혼합할 때 여타의 방법으로 증강이 된 데이터를 포함하여 블럭 혼합하였다. 본 논문에서 사용한 혼합 방법은 Eq.(2)이며, 혼합 대상이 되는 두 음향 데이터를 선정할 때 두 가지 선택 조건을 적용하였다. 첫 번째 조건은 서로 다른 두 음향 데이터는 한 단일 이벤트를 갖는 경우를 선택하였다. 두 번째 조건은 서로 다른 두 음향 데이터가 각각 두 개의 이벤트를 가지며, 한 개 또는 두 개의 이벤트가 동일한 경우이다.

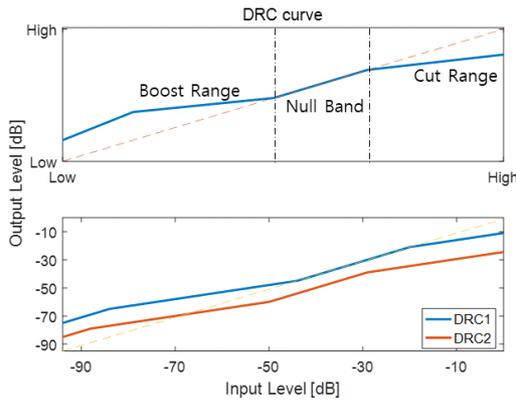


그림 4. DRC 커브 예제(위), 사용된 DRC 커브(아래)
Fig. 4. A DRC curve example (above) and the DRC curves used (below)

$$M = F + S \tag{1}$$

$$M_{norm} = \frac{M}{\max(|M^v|)} \tag{2}$$

F 는 첫 번째로 선택된 음향 데이터를 나타내며, S 는 두 번째로 선택된 음향 데이터를 나타낸다. M 은 혼합된 음향

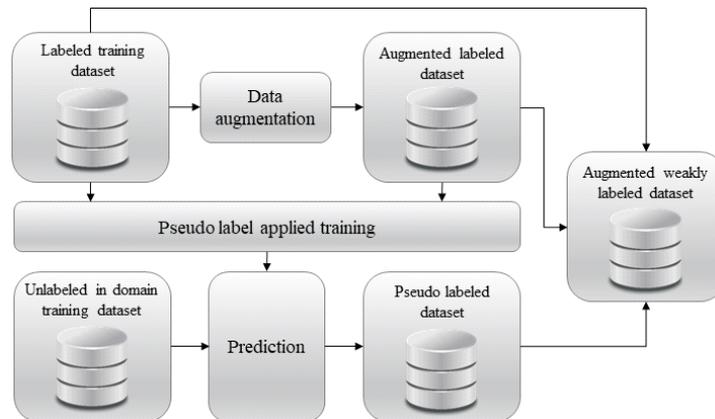


그림 5. 의사 라벨이 적용된 훈련 블럭 다이어그램
Fig. 5. Block diagram of pseudo label applied training

데이터를 나타낸다. 두 신호의 덧셈으로 얻은 M 은 지나치게 큰 값을 가질 수 있기에 정규화를 통해 $-1 \sim 1$ 사이의 값을 갖도록 하며 M_{norm} 으로 나타낸다.

2. 의사 라벨 (pseudo-label) 적용

데이터를 지도 학습에 사용하기 위해서는 라벨이 부여되어 있어야 한다. Fig. 5는 의사 라벨을 부여하기 위한 블록 다이어그램을 나타낸다. 본 논문에서는 라벨을 부여하기 위하여 주어진 약한 라벨이 있는 데이터와 타임 스트레칭, 피치 변화, 동적 영역 압축 방식으로 증강된 데이터를 이용하여 의사 라벨을 부여하는 신경망 (의사 라벨 적용 신경망)을 훈련시켰다. 그리고 이렇게 훈련된 의사 라벨 적용 신경망에 라벨이 없는 데이터를 입력하여 그 데이터의 라벨을 예측하였고, 이를 의사 라벨이라고 하였다. 의사 라벨은 시간 정보가 없는 약한 라벨이다.

IV. 알고리즘 구현

1. 데이터 세트

심층신경망의 훈련을 위하여 구글에서 제공하는 오디오 세트^[12] 중 일부를 훈련 데이터 세트로 사용한다. 구글 오디오 세트는 음향 이벤트 10 종류를 가지며, 각 음향 데이터 (즉, 클립)는 하나 이상의 이벤트를 갖는 멀티 라벨 데이터이다. 약한 라벨이 부여된 데이터는 1,578 개, 라벨이 없는 데이터 14,412 개, 강한 라벨이 부여된 테스트 데이터 288 개로 구성되어있다. 약한 라벨이 부여된 훈련 데이터는 Table 1과 같이 이벤트에 대하여 불균형한 분포를 가지고 있다. 심층신경망 훈련에 사용할 수 있는 라벨이 있는 데이터 수가 적어 발생하는 문제를 해결하고자 3장에서 제시한 데이터 증강 방법을 적용하여 데이터를 추가로 확보하였다. 이 때 적용된 데이터 증강 비율은 Table 2와 같다. 데이터 증강을 통하여 최종적으로 약한 라벨을 가진 데이터 17,358 개, 의사 라벨을 가진 데이터 14,412개의 데이터를 확보하여 음향 이벤트 검출을 수행하는 SED 신경망의 첫 번째 훈련에 사용하였다. 그리고 조건에 따라 다른 결과를 만든

는 블록 혼합을 적용하여 6,312개의 데이터를 추가로 증강하였고 이를 포함하여 두 번째 훈련에 사용하였다. 이렇게 훈련한 SED 신경망을 강한 라벨이 부여된 288개의 테스트 데이터를 이용하여 성능을 검증하였다.

표 1. 이벤트 별 수 (약한 라벨)

Table 1. Number of clips per event (weak label)

Event	# of clips
Speech	550
Dog	214
Cat	173
Alarm / bell / ringing	205
Dishes	184
Frying	171
Blender	134
Running water	343
Vacuum cleaner	167
Electric shaver / toothbrush	103
Total	2244

표 2. 데이터 증강 비율과 증강된 클립의 수

Table 2. Ratio of data augmentation and resultant number of augmented clips

Event	Ratio	# of clips
Original	1	1,578
Time stretching	4	6,312
Pitch shifting	4	6,312
Dynamic Range Compression	2	3,156
Block mixing	4	6,312
Total	11	23,670

2. 훈련 절차

음향 이벤트 검출을 위한 SED 신경망 훈련을 2회에 걸쳐 진행하였다. 사용한 데이터 중 80 %는 훈련 데이터로, 20 %는 첫 번째와 두 번째 훈련 결과에 대한 검증 데이터로 사용하였다. 손실 함수로 크로스 엔트로피(cross-entropy)가 사용되었으며, Adam^[4] 최적화 방법을 사용하였다.

첫 번째 훈련은 주어진 데이터를 블록 믹싱을 제외한 증강 방법으로 확보한 약한 라벨 데이터 세트를 이용하여 수행하였다. 이 때 입력되는 음향 데이터의 전 구간에 대해 이벤트가 존재한다고 가정하고 훈련하였다. 두 번째 훈련

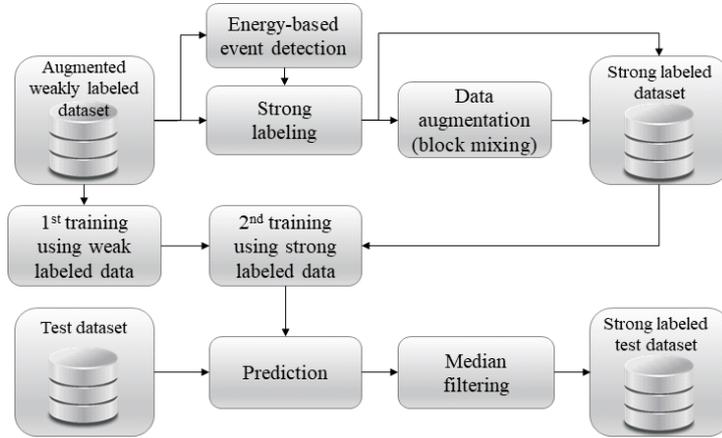


그림 6. 전반적인 구조의 블럭 다이어그램
Fig. 6. Block diagram for overall structure

의 방향성을 잡아주는 역할을 하며, 이를 통하여 얻은 SED 신경망은 본격적인 음향 이벤트 검출(프레임 단위의 음향 이벤트 검출)을 수행하는 SED 신경망을 얻기 위한 두 번째 훈련의 초기 값으로 사용하였다.

두 번째 훈련은 첫 번째 훈련을 통해 얻은 SED 신경망을 음향 이벤트 검출에 보다 적합하도록 훈련하는 역할을 한다. 프레임 별로 이벤트가 있는지 없는지를 구분하기 위하여 강한 라벨이 붙은 데이터로 훈련하였다. 강한 라벨이 부여된 훈련 데이터를 얻기 위하여, 먼저 약한 라벨이 부여된 클립을 프레임 단위로 읽고, 읽은 프레임의 RMS(Root Mean Square) 값이 역치를 넘으면 그 프레임에 클립의 약한 라벨에 해당하는 이벤트가 발생하였다고 간주하는 방식을 사용하였다.

3. 성능 평가 방법

성능을 평가할 지표로서 f-score와 오류율이 있다. 정답이라고 분류했는데 실제로 정답인 TP(True Positive), 오답이라고 분류했는데 실제로 정답인 FN(False Negative), 정답이라고 분류했는데 실제로 오답인 FP(False Positive)을 이용하여 측정한다.

F-score는 검사의 정확도에 대한 척도를 표기하며 보통 f1-score를 측정한다. 정답이라고 검출한 결과를 얼마나 신뢰할 수 있는지를 나타내는 정밀도(precision), 정답을 얼마나 섬세하게 감지하는 지를 나타내는 검출율(recall) 등 두

요소가 사용된다.

$$precision = \frac{TP}{TP+FP} \quad (3)$$

$$recall = \frac{TP}{TP+FN} \quad (4)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (5)$$

구해진 F_1 은 확률의 값을 가지며 0 ~ 100 %의 값을 갖는다. 100 %에 가까울수록 예측이 정확하다고 판단될 수 있으며, 시험 데이터에 대한 f-score를 통해 비교한다.

본 논문에서 다루는 음향 이벤트 검출에 대한 f-score 측정을 하기 위해 이벤트 기반 측정을 수행한다. 구체적으로는 onset 과 offset 각각에 대하여, 정답과 예측 값의 차이 크기가 기준 값 대비 작은지 여부를 따진다. 이때 기준 값으로 결정 간격(DI: decision interval), 길이 비율(PL: percentage of length), 정답에 대한 길이(AL: annotated length) 등 세 요소를 사용한다^[13].

$$|onset_{ref} - onset_{pred}| \leq DI \quad (6)$$

$$|offset_{ref} - offset_{pred}| \leq \max(DI, AL \cdot PL) \quad (7)$$

Eq.(6)은 정답 이벤트의 onset과 추정된 이벤트의 on-

set의 차이가 결정 간격 내에 있는지를 검증한다. Eq.(7)은 클래스의 정답 offset과 추정된 offset의 차이가 시간 영역의 일정 비율 이내 여부를 검증한다. 두 수식을 모두 만족하는 경우 TP로 간주한다. 두 수식을 모두 만족하지 못하는 경우 누락이 된다. FP는 추정된 이벤트의 전체 구간에 대해 TP와 누락의 차, FN은 정답 이벤트의 전체 구간에 대해 TP와 누락의 차로 구한다. 본 논문에서 사용된 결정 간격은 200ms, 길이 비율은 20 %를 적용한다¹⁴⁾.

오류율은 검사에 대한 오류의 빈도를 나타낸다. 기존 정답 대신 다른 값으로 대체된 대체(S: substitutions), 기존 정답이 누락된 경우인 제거(D: deletions), 새롭게 추가된 경우인 삽입(I, insertions), 항목의 수를 나타내는 N , 전체 구간의 개수를 나타내는 K 를 통해 측정한다.

$$S = \sum_{k=1}^K S(k), S(k) = \min(FN(k), FP(k)) \quad (8)$$

$$D = \sum_{k=1}^K D(k), D(k) = \max(0, FN(k) - FP(k)) \quad (9)$$

$$I = \sum_{k=1}^K I(k), I(k) = \max(0, FP(k) - FN(k)) \quad (10)$$

$$ER = \frac{S+D+I}{N} \quad (11)$$

위의 제시된 Eq.(11)와 같이 전체 구간에 대한 대체,

제거, 삽입을 계산하고 항목의 수로 나눠 줌으로서 오류율을 계산한다¹³⁾. 0 ~ 100 %의 값을 가지며, 0 %에 가까울수록 예측이 잘 되었다고 판단할 수 있다.

V. 실험 결과

훈련을 위해 앞서 언급된 구글 오디오 세트를 사용하였다. 구글 오디오 세트는 44.1 kHz의 샘플링 율을 갖는다. 하지만 10 kHz 이상의 대역에 대해 음향이 매우 미비하게 존재하거나 없기에 22.05 kHz로 다운 샘플링을 하여 사용하였다. 오디오 세트 중 10 초 미만 혹은 초과하는 데이터(클립)에 대해서는 영 삽입을 하여 모두 10 초로 맞추었다. 각 데이터는 0.5 초 단위(hop size = 0.1 초)로 나뉘어져 제안된 SED 신경망의 입력으로 공급되었다.

SED 신경망을 이용하여 첫 번째 훈련은 최대 200 epoch, 두 번째 훈련은 최대 50 epoch 까지 훈련이 진행되도록 하였다. 사용한 중앙값 필터는 1 초의 크기를 갖도록 설계하였다.

1. ResGLU-SE 블록을 사용한 심층신경망의 1차 합성곱 계층 분석

Fig. 7은 제안한 SED 신경망에서 음향 이벤트 검출 훈련을 마친 후 얻은 1차 합성곱 계층 계수의 크기 스펙트럼

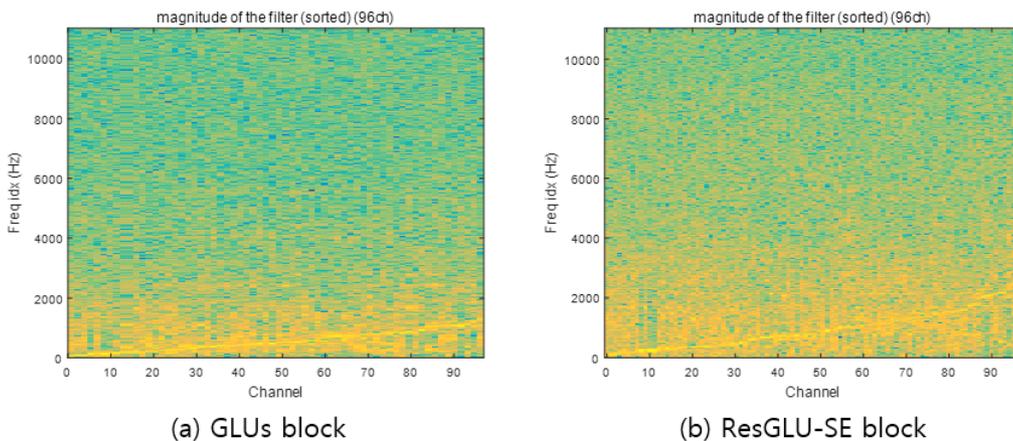


그림 7. 스트라이드된 1차 합성곱 계층의 크기 스펙트럼, (a) GLUs 블록, (b) ResGLU-SE 블록
 Fig. 7. Magnitude spectrum of the strided 1D convolutional layers, (a) Using GLUs block, (b) Using ResGLU-SE block

이다. 노란색에 가까울수록 해당 주파수에 해당하는 특징을 추출하는 것을 나타낸다. 시간 영역의 입력을 처음 처리하는 1차 합성곱 계층은 96개의 채널이 있다. Fig. 7의 (a)는 SED 신경망에서 ResGLU-SE 블록 대신에 GLUs 블록을 사용했을 때의 크기 스펙트럼을 나타내며, Fig. 7의 (b)는 ResGLU-SE 블록을 사용했을 때의 크기 스펙트럼을 나타낸다. 두 가지 모두 저주파 대역에 대해 강하게 반응하는 것을 확인할 수 있다. 하지만 (a)의 경우 1500 Hz까지의 특징을 추출하고, (b)의 경우 채널들이 2 kHz까지의 특징을 추출한다. V 장 2절의 성능 평가에 따르면 입력 신호를 더 넓은 주파수 대역에서 분석할 수 있는 ResGLU-SE 블록을 사용하는 경우, GLUs 블록을 사용하는 경우 보다, f-score 기준으로 더 나은 성능을 보인다.

2. 제안된 블록이 사용된 심층신경망의 음향 이벤트 검출 성능

음향 이벤트 검출 성능을 평가하기 위해 288개의 강한 라벨이 된 시험 데이터를 사용하였다. Table 3은 주파수 영역의 입력 신호를 처리하는 기준 신경망^[4]과 시간 영역의 입력 신호를 처리하는 SED 신경망을 비교하였다. 주어진 데이터를 입력으로 사용한 기준 신경망보다 높은 성능을 내는 것을 확인하고자, 기준 신경망의 경우 데이터 증강 방법을 적용하지 않았다. SED 신경망의 경우 내부 구성에

표 3. 음향 이벤트 검출 성능

Table 3. Performance of sound event detection

Structure		F-score [%]	Error rate [%]
Baseline (CRNN)		13.99	1.52
SED neural network	GLUs	14.39	1.58
	ResGLU	15.76	1.73
	ResGLU-SE	19.96	1.60

서 2 ~ 5 계층을 GLUs 블록, ResGLU 블록, ResGLU-SE 블록으로 사용하는 경우 각각 성능 향상에 얼마나 기여하는지를 보여준다. 기본 구조에 비하여, 제안된 ResGLU-SE 블록을 사용한 신경망은 데이터 증강을 사용하여 f-score가 약 6 % point가 되었다. ResGLU 블록을 사용한 경우 GLUs 블록을 사용했을 때 보다 1.37 % point가 있었으며, ResGLU-SE 블록을 사용한 경우 ResGLU 블록을 사용했을 때 보다 4.2 % point를 보여주었다.

4장에서 f-score는 시험 데이터에 대한 예측의 정확도를 나타내며 100 %에 가까울수록 예측이 잘 되는 것을 나타내고, 오류율은 오류의 빈도를 나타내며 0 %에 가까울수록 오류가 감소함을 나타낸다고 하였다. Table 3을 보면 f-score가 14.39 %일 경우의 오류율은 1.58 %, f-score가 19.64 %로 가장 높을 때의 오류율은 1.60 %이다. 즉 f-score가 크다고 하여 반드시 오류율이 낮은 것이 아님을 확인할 수 있다.

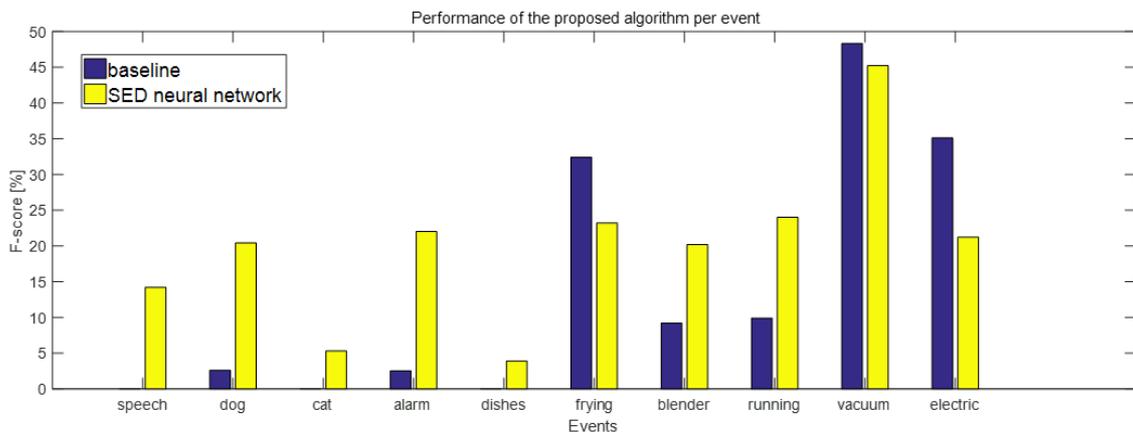


그림 8. 제안된 알고리즘의 이벤트 별 성능

Fig. 8. Performance of the proposed algorithm per event

3. ResGLU-SE 블럭을 사용한 심층신경망의 이벤트별 검출 성능 분석

Fig. 8은 테스트 데이터에 대한 두 신경망의 성능 결과를 이벤트 별로 나타내었다. 데이터 증강 방법으로 증가된 데이터를 이용하여 훈련을 수행하였으며, 기준 신경망을 사용했을 때의 결과와 SED 신경망에서 ResGLU-SE 블럭을 사용했을 때의 결과를 나타낸다. Fig. 8을 통하여 신경망에 따라서 이벤트 별 f-score가 다른 양상을 보인다는 것을 알 수 있다. 제한된 신경망은 speech, dog, cat, alarm, dishes, blender, running water 의 경우 상대적으로 우월한 성능을 보이지만, frying event, vacuum cleaner, electric shaver 에서는 상대적으로 열등한 성능을 보인다.

비교된 신경망들이 성능의 차이를 보이는 이벤트들의 속성을 파악하기 위하여 대표적인 이벤트들에 대해 스펙트럼을 검토하였다. Fig. 9는 훈련 데이터에서 음향 데이터 두 개의 스펙트럼을 보여준다. Fig. 9 (a)는 speech와 vacuum cleaner 이벤트가 있는 음향에 대한 스펙트럼이다. 2.1 ~ 6.9 초까지는 vacuum cleaner에 해당하며, 6.9 초 이후에는 speech에 해당한다. Vacuum cleaner의 경우 전체 주파수 대역에 대해 에너지를 갖는 것이 확인된다. Speech는 저주파 대역에 에너지가 모여 있으며, 고주파 대역에 대해서는 상대적으로 적은 에너지를 갖는 것을 확인할 수 있다. Fig. 9 (b)는 blender 이벤트가 있는 음향에 대한 스펙트럼이다. 전체 구간에 대하여 고르게 이벤트가 분포해 있으며, Fig. 9 (a)의 vacuum cleaner와 비교했을 때 2 kHz 이하에서

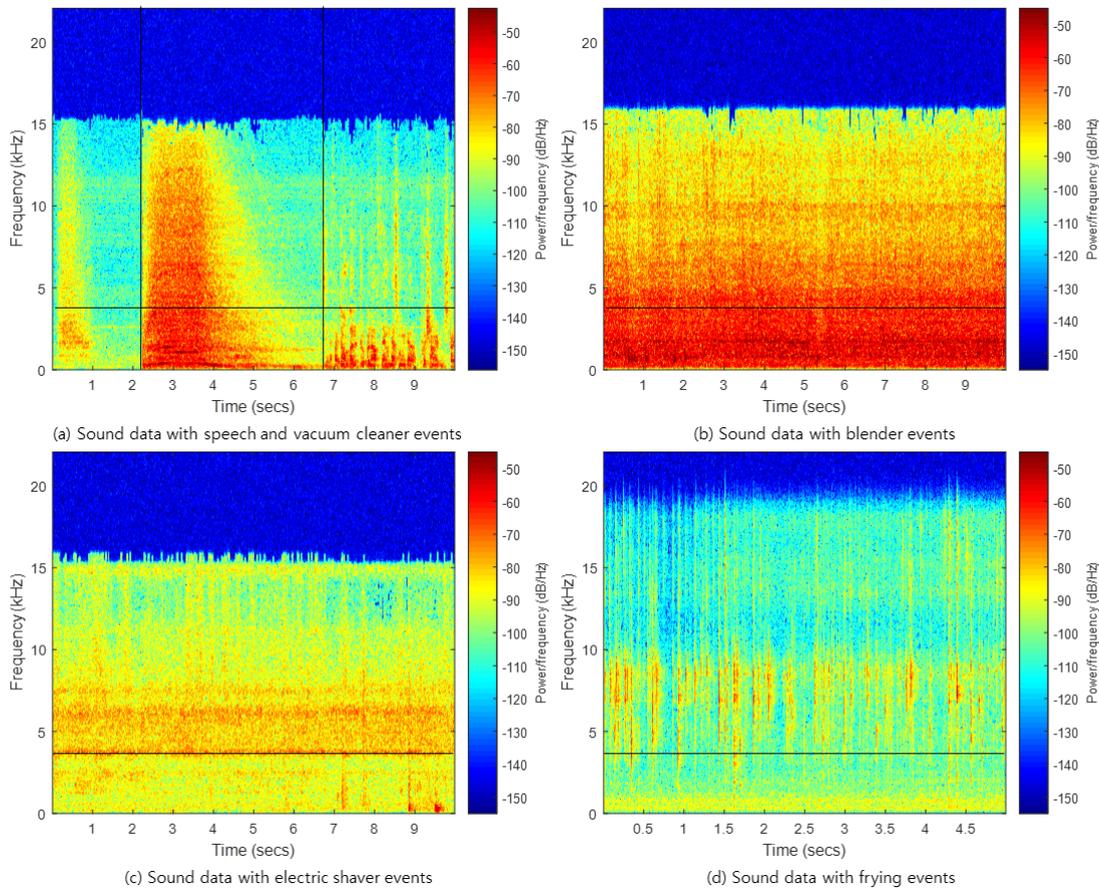


그림 9. 훈련 데이터의 스펙트럼
 Fig. 9. Spectrum of training data

두 장치의 스펙트럼이 상당히 유사함이 확인할 수 있다. Fig. 9 (c)와 (d)는 각각 electric shaver와 frying 이벤트의 스펙트럼을 나타낸다. 두 이벤트 모두 4 kHz 이상의 고주파 대역에 에너지가 모여 있으며, 저주파 대역에 대해서는 상대적으로 적은 에너지를 갖는다.

Vacuum cleaner와 electric saver, frying의 경우, 제안된 SED 신경망은 기준 신경망 보다 각각 3.1 %, 13.9 %, 9.2% 낮은 성능을 보였다. Fig. 7에서 확인하였듯이 제안된 구조는 저주파 대역에 대해 민감하게 반응한다. 이 사실을 고려한다면, 넓은 대역에 걸쳐 특징이 있는 이벤트에 대해서는 제안된 신경망의 검출 성능은 제한적이라고 생각할 수 있다. Vacuum cleaner와 blender의 경우 저주파 대역의 에너지가 유사한 문제가 있으며, electric shaver와 frying은 저주파 대역에 에너지가 너무 작기에 환경 잡음과 뚜렷한 구분이 되지 않는 문제가 있다. 다만 blender 의 경우 제안된 신경망의 성능이 기준 신경망 보다 우수하다. 이는 제안된 신경망이 blender 와 vacuum cleaner 두 이벤트에 해당하는 입력 신호에 대하여 주로 blender 로 편향되게 판정하는 경향이 있기 때문으로 보인다. 그러므로 이 편향 역시 넓은 대역에 특징이 걸쳐 있는 두 이벤트를 구분하는 성능이 제한된 결과로 볼 수 있다.

이벤트의 종류 따른 결과의 차이를 확인하고자 테스트 데이터 중 두 개의 음향 데이터에 대해 분석하였다. Fig.

10은 테스트 데이터 중 정답에 vacuum cleaner가 포함된 음향 데이터에 대한 정답과 예측된 결과의 구간을 나타낸 것이다. 이벤트가 없을 경우 0의 값을 가지며, 이벤트가 존재하는 경우 0보다 큰 값을 갖는다. 각 이벤트 별로 서로 다른 값을 갖도록 하여 이벤트 존재여부를 시각적으로 확인할 수 있도록 하였다.

Fig. 10 (a)의 경우, 정답과 다른 라벨이 붙은 결과를 나타낸다. 정답 vacuum cleaner는 전 구간에 이벤트가 있다. 예측된 결과는 0 ~ 3.1 초 구간(3.1 초)에 대해서는 vacuum cleaner를 정확하게 맞추지만, 3.1 ~ 10 초 구간(6.9 초)에 대해서는 blender로 잘못 라벨이 붙은 결과가 나왔다. Fig. 7 (b)의 결과에서 알 수 있듯이, 제안된 구조는 2 kHz까지의 특징을 추출한다. Fig. 9의 (a)의 vacuum cleaner가 있는 구간과 Fig. 9 (b)의 blender의 2 kHz까지의 대역은 큰 차이가 없기에 두 이벤트를 혼동하는 것으로 추정된다.

Fig. 10 (b)는 시간 정보 예측이 잘못된 경우를 보여준다. 정답 vacuum cleaner는 전 구간에 이벤트가 있다고 되어 있지만, 예측된 결과는 0.5 ~ 1.2 초 구간(0.7 초)에 대해서는 이벤트가 없다고 판단한다. 반면 dishes는 0.2 ~ 0.5 초의 짧은 구간(0.3 초)에 대해 이벤트가 있지만 예측된 결과는 음향 데이터 전 구간에 대해 dishes가 없다고 판단했다. 그 결과 vacuum 이벤트의 TP는 0보다 큰 값을 가지지만, vacuum 이벤트 보다 짧은 구간을 놓친 dishes 이벤트의 TP는

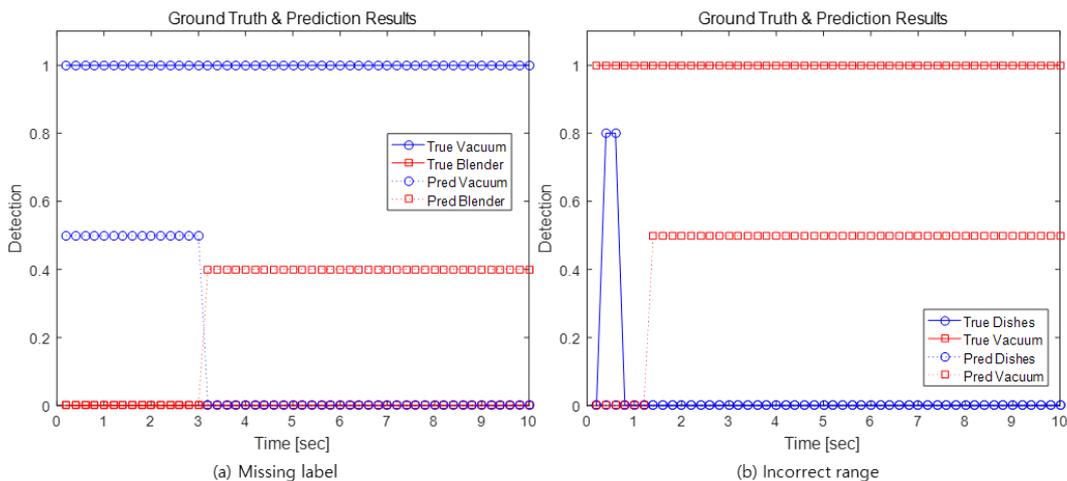


그림 10. 정답과 추정 결과
Fig. 10. Ground truth and prediction results

0의 값을 갖게 된다. 그에 따라 dishes 이벤트의 f-score는 0이 되어 dishes 이벤트의 전체 f-score가 일부 구간을 맞춘 vacuum 이벤트의 전체 f-score 보다 더 많이 낮아지는 결과를 초래한다. 그러므로 성능 측정 척도의 특성상 제안된 신경망은, 상대적으로 짧은 이벤트를 놓치지 않는 방향으로 훈련된 것으로 보인다. 그 결과 저주파 영역에 특징이 집중되어 있는 짧은 이벤트들에 보다 민감하도록 최하단의 1차 신경곱 계층이 훈련된 것으로 추정된다.

VI. 결 론

본 논문에서는 심층신경망을 이용한 시간 영역 음향 이벤트 검출 알고리즘을 제안하였다. 이를 위해 ResGLU-SE라는 새로운 블록을 사용하고 여러 수준의 계층에서 추출된 특징을 함께 고려하는 신경망 구조를 제안하였다. 또한 제안한 구조와 함께 사용할 수 있는 데이터 증강 방법을 제시하였다. 제안된 SED 신경망은 CRNN 구조를 가지는 기존 신경망과 비교하여 약 6 % point (f-score 기준)를 보였다.

본 논문에서 제안한 SED 신경망은 저주파 영역에서 분류에 활용 가능한 특징이 많은 이벤트들은 비교적 분류가 잘되지만, 그렇지 않은 이벤트들은 분류 성능이 떨어지는 경향을 보였다. 입력 신호를 최초로 처리하는 1차 합성곱 계층이 입력신호의 약 2 kHz 대역까지만 민감하게 반응하도록 훈련된 것이 SED 신경망의 최종 성능을 제한하고 있는 것으로 분석된다. 음향 이벤트 검출 성능을 보다 향상시키기 위해서는 고주파 영역에 대한 특징을 추출할 수 있는 신경망 구조, 그리고 훈련 방법에 대한 추가적인 연구가 필요하다.

참 고 문 헌 (References)

- [1] Mesaros, A., Heittola, T, and Virtanen, T, "TUT database for acoustic scene classification and sound event detection," 2016 24th EUSIPCO, Hungary, Budapest, pp.1128-1132, August 2016.
- [2] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," IEEE Multimedia, Vol.3, No.3, pp.27 - 36, 1996.
- [3] DENG, Ltsc, et al. "Recent advances in deep learning for speech research at Microsoft," In ICASSP, Vol. 26, pp. 64, May 2013.
- [4] Mun, Seongkyu, et al. "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," Proceeding of DCASE, pp.93-97, 2017.
- [5] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," arXiv preprint arXiv preprint arXiv:1612.08083, 2016.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," In European Conference on Computer Vision (ECCV). Springer, pp.630 - 645, 2016.
- [7] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," arXiv preprint arXiv:1709.01507, 2017.
- [8] Hyeonggi Moon, Joon Byun, Bum-Jun Kim, Shin-hyuk Jeon, Youngho Jeong, Young-cheol Park and Sung-wook Park, "End-to-end CRNN Architectures for Weakly Supervised Sound Event Detection," DCASE 2018 Challenge, Sep. 2018.
- [9] Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, Oriol Vinyals, "Learning the speech front-end with raw waveform CLDNNs," Proceeding of INTERSPEECH, Germany, Dresden, September 2015.
- [10] Yong Xu, Qiuqiang Kong, Wenwu Wang and Mark D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," Proceeding of ICASSP, Canada, Calgary, pp.121-125, April 2018.
- [11] Justin Salamon and Juhan Pablo Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," IEEE Signal Processing Letters, pp.279-283, 2017
- [12] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," Proceeding of ICASSP, USA, New Orleans, pp.776-780, March 2017.
- [13] Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection," Applied Sciences, 6.6: 162, 2016.
- [14] Romain Serizel, Nicolas Turpault, Hamid Eghbal-Zadeh, Ankit Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," arXiv preprint arXiv:1807.10501, 2018.

저 자 소 개



김 범 준

- 2017년 2월 : 연세대학교 컴퓨터공학 학사
- 2017년 3월 ~ 현재 : 연세대학교 전산학과 석사과정
- ORCID : <https://orcid.org/0000-0002-8977-2917>
- 주관심분야 : 디지털 신호처리, 음질 개선, 음성 신호처리, 적응 신호처리



문 현 기

- 2013년 2월 : 연세대학교 전기전자공학부 학사
- 2013년 3월 ~ 현재 : 연세대학교 전기전자공학부 석박 통합과정
- ORCID : <https://orcid.org/0000-0001-8806-6335>
- 주관심분야 : 오디오 신호처리, 3D 오디오, 오디오 부호화



박 성 욱

- 1993년 2월 : 연세대학교 전자공학과 학사
- 1995년 2월 : 연세대학교 신호처리 석사
- 1998년 8월 : 연세대학교 신호처리 박사
- 2009년 3월 ~ 현재 : 국립강릉원주대학교 전자공학과 부교수
- 주관심분야 : VLSI 신호처리, 멀티미디어 시스템



정 영 호

- 1992년 : 전북대학교 전자공학과 공학사
- 1994년 : 전북대학교 전자공학과 공학석사
- 2006년 : 충남대학교 전자공학과 공학박사
- 2011년 ~ 2017년 : 과학기술연합대학원대학교(UST) 이동통신 및 디지털 방송공학과 겸임 교수
- 1994년 ~ 현재 : ETRI 실감 AV연구그룹 책임연구원
- ORCID : <https://orcid.org/0000-0001-9552-8593>
- 주관심분야 : 음향인식, 머신러닝, 오디오 신호처리



박 영 철

- 1986년 2월 : 연세대학교 전자공학과 학사
- 1988년 2월 : 연세대학교 전자공학과 석사
- 1988년 2월 : 연세대학교 전자공학과 박사
- 2002년 3월 ~ 현재 : 연세대학교 컴퓨터정보통신공학부 교수
- ORCID : <https://orcid.org/0000-0003-3274-076X>
- 주관심분야 : 디지털 신호처리, 오디오 신호처리, 음성 신호처리, 적응 신호처리