한국음향학회지 제39권 제6호 pp. 600~605 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.6 (2020)
https://doi.org/10.7776/ASK.2020.39.6.600

pISSN : 1225-4428
eISSN : 2287-3775

청각 장애인용 홈 모니터링 시스템을 위한 다채널 다중 스케일 신경망 기반의 사운드 이벤트 검출

Sound event detection based on multi-channel multi-scale neural networks for home monitoring system used by the hard-of-hearing

이기용,[1] 김형국[1†]

(Gi Yong Lee[1] and Hyoung-Gook Kim[1†])

[1]광운대학교 전자융합공학과

(Received August 31, 2020; revised October 12, 2020; accepted October 27, 2020)

초 록: 본 논문에서는 청각 장애인을 위한 소리 감지 홈 모니터링을 위해 다채널 다중 스케일 신경망을 사용한 사운드 이벤트 검출 방식을 제안한다. 제안하는 시스템에서는 홈 내의 여러 무선 마이크 센서들로부터 높은 신호 품질을 갖는 두 개의 채널을 선택하고, 그 신호들로부터 도착신호 지연시간, 피치 범위, 그리고 다중 스케일 합성 곱 신경망을 로그 멜 스펙트로그램에 적용하여 추출한 특징들을 양방향 게이트 순환 신경망 기반의 분류기에 적용함으로써 사운드 이벤트 검출의 성능을 더욱 향상시킨다. 검출된 사운드 이벤트 결과는 선택된 채널의 센서 위치와 함께 텍스트로 변환되어 청각 장애인에게 제공된다. 실험결과는 제안한 시스템의 사운드 이벤트 검출 방식이 기존 방식보다 우수하며 청각 장애인에게 효과적으로 사운드 정보를 전달할 수 있음을 보인다.
핵심용어: 사운드 이벤트 검출, 다채널 오디오 특징 값, 다중 스케일 신경망, 양방향 게이트 순환 신경망

ABSTRACT: In this paper, we propose a sound event detection method using a multi-channel multi-scale neural networks for sound sensing home monitoring for the hearing impaired. In the proposed system, two channels with high signal quality are selected from several wireless microphone sensors in home. The three features (time difference of arrival, pitch range, and outputs obtained by applying multi-scale convolutional neural network to log mel spectrogram) extracted from the sensor signals are applied to a classifier based on a bidirectional gated recurrent neural network to further improve the performance of sound event detection. The detected sound event result is converted into text along with the sensor position of the selected channel and provided to the hearing impaired. The experimental results show that the sound event detection method of the proposed system is superior to the existing method and can effectively deliver sound information to the hearing impaired.
Keywords: Sound event detection, Multichannel audio features, Multi-scale neural networks, Bidirectional gated recurrent neural networks

PACS numbers: 43.60.Bf, 43.60.Lq

I. 서 론

무선 센서와 네트워크 기술의 발달로 인해 사람과 환경을 모니터링 하는 기술은 보안, 의료, 에너지 등의 산업 분야뿐 만 아니라 일상생활의 다양한 분야에도 적용되고 있다. 주변 환경의 데이터를 수집하고 분석 및 표현해주는 모니터링 시스템은 정보 전달을 위한 보조 시스템으로서 주변 환경의 정보 취득이 힘든 장애인, 노인, 어린이를 위해 비상시 중요

†**Corresponding author:** Hyoung-Gook Kim (hkim@kw.ac.kr)
Department of Electonics Convergence Engineering, KwangWoon University, 20 Gwangun-ro, Nowon-gu, Seoul 01897, Republic of Korea
(Tel: 82-2-940-5574, Fax: 82-2-913-5006)

한 기능을 수행한다. 특히 청각장애인은 사운드를 통한 정보취득이 어렵기 때문에 일상생활의 많은 위험에 노출되어 있다. 이에 따라 청각장애인들이 위험한 상황에 듣지 못하는 소리를 감지해 경고해 주는 사운드 이벤트 검출(Sound Event Detection, SED) 방법에 대한 연구가 활발히 진행되고 있다.

Kim *et al.*[1] 은 다채널 SED기반의 청각 장애인을 위한 홈 모니터링 시스템을 제안하였다. 이 방법에서는 단 채널 특징 값이 아닌 다채널 오디오 기반의 특징 추출 방법을 적용하여 SED성능을 향상시켰다. 하지만 검출된 사운드 이벤트 정보에는 발생 위치가 반영되어 있지 않다. 상황의 맥락 이해가 필요한 청각장애인에게 보다 정확한 정보 전달을 위해서는 SED의 정확도와 함께 발생 위치를 제공해야 한다.

최근에는 Convolutional Neural Network(CNN), Long Short Term Memory(LSTM), Convolutional and Recurrent Neural Network(CRNN)과 같은 다양한 구조의 심층 신경망이 SED에 적용되고 있다. 그 중, Zhang *et al.*[2] 은 단 채널 사운드 신호의 시간 및 주파수 영역에서 고수준의 시프트 불변 특징을 학습하기 위해 다중 스케일의 합성 곱 필터를 CRNN에 적용한 SED 방법을 제안하였다. 이 방식에서는 다양한 스케일의 합성 곱 필터를 통해 서로 다른 발생 길이의 사운드 이벤트로부터 효과적으로 주파수 정보와 시간 정보를 학습하여 SED의 성능을 향상시켰다.

이에 본 논문에서는 다채널 사운드 신호에 Multi-scale Convolutional Neural network(MCN)과 Bidirectional Gated Recurrent Neural Network (BGRNN)을 적용하여 SED의 성능의 향상시키고, 검출된 사운드 이벤트의 명칭과 발생위치를 함께 전달하는 청각장애인을 위한 홈 모니터링 시스템을 제안한다.

본 논문의 구성은 다음과 같다. II장에서는 제안하는 홈 모니터링 시스템과 다채널 SED방법에 대해 설명하고, III장에서는 실험결과를 제시한다. 마지막으로 IV장에서는 결론을 맺는다.

II. 제안하는 방법

2.1 홈 모니터링 시스템

Fig. 1은 본 논문에서 제안하는 홈 모니터링 시스템의 전체 구조도이다. 제안하는 시스템은 무선 마이크 센서들과 싱크를 통해 검출된 사운드 이벤트 정보를 텍스트로 변환하여 사용자에게 전달한다.

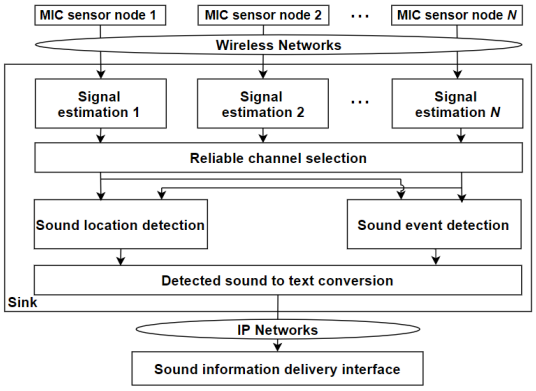

Fig. 1. Architecture of the proposed home monitoring system.

먼저 각 무선 마이크 센서는 사운드를 수신하며 녹음된 사운드를 패킷으로 인코딩하여 무선 음향센서 네트워크를 통해 싱크로 전달한다. 이후 싱크에서는 수신된 사운드 패킷을 신호 프레임으로 디코딩하고 무선 멀티 홉 통신 과정 중에 손실된 패킷을 복원하기 위해 Recursive Linear Prediction and Synthesis(RLPS)[3] 기반의 패킷 손실 은닉 방식을 통해 손실된 패킷을 복원한다. 다음으로 내외장벽 마감재로 인해 발생될 수 있는 위치 추정 오류의 감소 및 SED 성능의 향상을 위해 가장 상관관계가 높고 신호의 발생원으로부터 근접한 두 개의 마이크 채널을 선택한다. 채널 선택 방법은 패킷 손실이 적고 에너지 임계 값 보다 RMS값이 큰 채널들을 시간 순으로 두 개씩 쌍을 이룬 후 가장 큰 Multi-Channel Cross-Correlation Coefficient(MCCC)[4] 값을 갖는 하나의 채널 쌍을 선택한다. 선택된 두 채널의 위치가 사운드 이벤트의 발생 위치 정보로 메모리에 저장되며, 수신된 두 채널의 사운드 신호들은 다채널 다중스케일 신경망 기반의 SED에 적용된다. 검출된 사운드 이벤트의 명칭과 위치 정보는 텍스트로 변환되며 IP네트워크를 통해 청각 장애인의 기기로 전달된다.

2.2 다채널 사운드 이벤트 검출

본 논문에서는 다채널 오디오 신호에 MCN과 BGRNN을 사용한 SED 방법을 제안한다. 제안하는 SED 방법

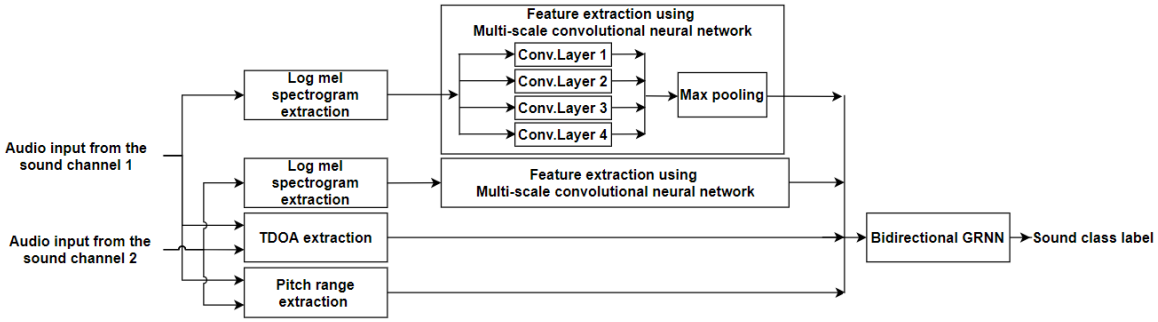

Fig. 2. Architecture of the proposed sound event detection using multi-channel multi-scale neural networks.

은 Fig. 2와 같다. 먼저 사운드 이벤트의 발생 위치 인식과 중첩된 모든 사운드 이벤트에 대한 효과적인 검출을 위해 다채널 오디오 신호로부터 인간의 청각 특성을 반영한 세 가지의 특징을 추출한다. 그다음 시간-주파수 영역에서 다양한 분포의 사운드 이벤트를 구분할 수 있도록 MCN기반의 특징 추출 방법을 적용한다. 추출된 특징들은 모두 BGRNN에 입력되어 사운드 이벤트로 예측된다.

2.2.1 다채널 기반의 오디오 특징 추출

제안하는 시스템은 일상 환경에서 발생하는 다양한 사운드 이벤트를 검출하며 검출된 이벤트의 발생 위치 인식을 목표로 한다. 하지만 일상 환경의 사운드로부터 녹음된 오디오 신호에는 다양한 사운드 이벤트가 중첩되어 존재하며 어떠한 위치 정보도 포함되어 있지 않다. 따라서 효과적인 사운드 이벤트 검출을 위해 중첩된 각 사운드 이벤트의 특성을 구분할 수 있는 특징 추출 방법과 사운드를 공간적으로 분리하여 위치를 인식할 수 있는 방법이 필요하다.

본 논문에서는 다채널 오디오 신호로부터 사운드 이벤트를 검출하며 채널에 해당하는 무선 마이크 센서의 위치 정보를 통해 검출된 사운드 이벤트의 발생 위치를 예측한다. 또한 사람의 청각 특성을 반영한 Log Mel Spectrogram(LMST), Time Difference Of Arrival(TDOA),[5] Pitch Range(PR)[6]의 세 가지 특징 값을 사용하여 효과적으로 중첩된 사운드 이벤트를 구분한다. 두 채널의 오디오 신호로부터 추출된 TDOA와 PR 은 각각 사운드가 두 귀에 도달하는 시간의 지연차를 통해 사운드를 국지화시키고 음의 높낮이를 활용하여 사운드를 분리하는 인간의 청각 특성을 반영한다. 따라서 TDOA특징 값은 사운드 이벤트의 주파수 대역에 따른 시간차 특성을 포함하며 PR특징 값은 피크사이의 주기 정보를 통해 추출된 피치 값을 포함한다. 또한 각 채널로부터 추출된 LMST는 저주파에 민감하고 고주파에 둔감한 인간의 청각 특성을 반영하며 시간-주파수 영역에서 다양한 분포로 발생하는 사운드 이벤트들의 윤곽특징을 검출하기 위해 MCN기반의 특징 추출 방법에 입력된다.

2.2.2 MCN기반의 특징 추출

Fig. 3은 일상생활에서 녹음된 오디오 신호의 시간-주파수 영역을 보여준다. 이러한 오디오 신호에 포함된 각 사운드 이벤트는 서로 다른 주파수 대역 및 대역폭으로 존재할 뿐만 아니라 다양한 지속 길이로 여러 시점에서 발생할 수 있기 때문에 SED를 위해서는 각 사운드 이벤트에 대한 시간-주파수 영역의 윤곽 특징을 효과적으로 분석할 수 있는 방법이 필요하다. 따라서 좁은 주파수 대역폭의 사운드 이벤트에 대해서는 작은 크기의 합성 곱 필터를 통해 주파수 영역의 세밀한 특징을 추출할 수 있어야 하며 넓은 주파수 대역폭의 사운드 이벤트에 대해서는 큰 크기의 합성 곱 필터를 통해 주파수 영역의 전체적인 특징을 추출할 수 있어야 한다. 이와 마찬가지로 각 사운드 이벤트의 서로 다른 발생 길이를 고려한 합성 곱 필터들을 통해 시간영역의 특징 또한 추출할 수 있어야 한다. 이에 따라 본 논문에서는 주파수 축 방향과 시간 축 방향 각각에 대한 다중 스케일 합성 곱 필터를 포함하는 MCN을 통해 각 채널의 LMST로부터 다양한 해상도의 윤곽 특징을 추출한다.

제안하는 MCN의 구조는 Fig. 4와 같다. MCN은 4

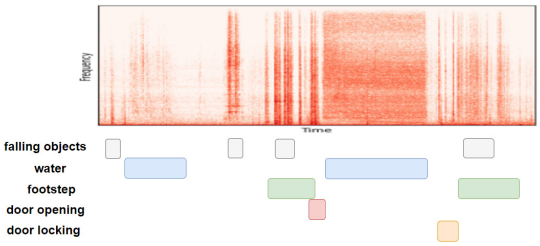

Fig. 3. (Color available online) Different lengths and frequency bands depending on the sound events in the spectrogram.

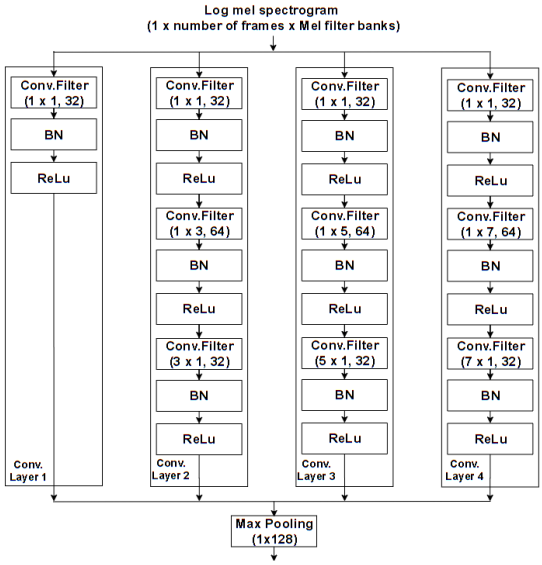

Fig. 4. Structure of MCN.

개의 합성 곱 층이 병렬로 연결되어 있으며 각 합성 곱 층에는 다양한 크기의 합성 곱 필터가 직렬로 연결되어 있다. 첫 번째 합성 곱 층은 1×1크기의 합성 곱 필터와 Batch Normalization(BN), Rectified Linear Unit(ReLu) 활성화 함수로 구성되며 위치 정보를 유지한 채 특징 채널을 조정하기 위해 사용된다. 나머지 세 개의 합성 곱 층은 1×1크기의 합성 곱 필터, 시간 축 방향의 정보를 유지하며 주파수 영역의 특징을 얻기 위한 $1 \times n(n=3, 5, 7)$크기의 합성 곱 필터, 그리고 주파수 축 방향의 정보를 유지하며 시간 영역의 특징을 얻기 위한 $n \times 1(n=3, 5, 7)$크기의 합성 곱 필터로 구성된다. 또한 BN과 ReLU 함수가 각 합성 곱 필터 이후에 적용된다. 마지막으로 1×128 크기의 최대 풀링을 적용하여 시간 축에 대한 정보를 보존하며 주파수 축의 차원만을 압축한다.

따라서 제안하는 SED 방법은 여러 크기의 합성 곱 필터로부터 고수준의 시프트 불변 특징을 추출하여 시간-주파수 영역에서 다양한 위치 및 크기로 발생하는 사운드 이벤트들의 검출 성능을 향상시킨다.

2.2.3 양방향 게이트 순환 신경망

본 논문에서 제안하는 다채널 SED 방법에서는 MCN으로부터 추출된 특징과 TDOA 그리고 PR의 3가지 특징 값을 BGRNN에 적용하여 시퀀스 정보를 추출한다. BGRNN은 장기의존성 문제를 해결하는 LSTM을 더욱 간단하게 변형한 구조로 적은 매개 변수와 빠른 학습 속도의 이점이 있다.

BGRNN은 리셋게이트와 업데이트게이트로 구성된 Gated Recurrent Unit(GRU)을 사용하며 t 번째에 대한 입력 x에 대해 다음과 같은 계산과정을 갖는다.

$$r_t = \sigma(W_r h_{t-1} + U_r x_t). \tag{1}$$

$$u_t = \sigma(W_u h_{t-1} + U_u x_t). \tag{2}$$

$$\overline{h_t} = \tau(W h_{t-1} \cdot r_t + U x_t). \tag{3}$$

$$h_t = (1-u_t) \cdot h_{t-1} + u_t \cdot \overline{h}, \tag{4}$$

여기서 r, u는 리셋게이트, 업데이트게이트를 나타낸다. 또한 h, W, U, σ, τ는 각각 은닉층의 출력, 은닉층간의 가중치, 입력층과 은닉층 사이의 가중치, 시그모이드 함수, 하이퍼볼릭 탄젠트 함수를 의미한다. Eq. (1)에서 리셋게이트는 시그모이드 함수를 통해 0과1 사이의 출력 값을 가지며 Eq. (3)에서 현재 상태의 입력과 이전 메모리의 정보를 제어하기 위해 사용된다. Eq. (2)에서 계산된 업데이트 게이트의 출력은 Eq. (4)의 은닉 층의 계산과정에 적용되어 현재 입력과 이전 메모리의 비율을 결정하게 된다. 또한 BGRNN의 은닉층은 Fig. 5와 같이 정방향 과 역방향 시퀀스로 나뉘어 이전 시간과 다음 시간의 특징 정보를 모두 고려하여 현재 상태의 학습을 진행 한다.

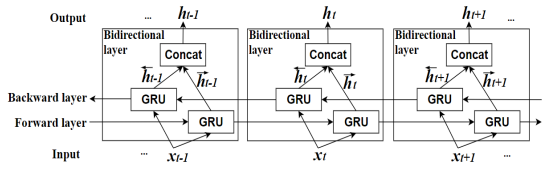

Fig. 5. Architecture of bidirectional GRNN with three consecutive steps.

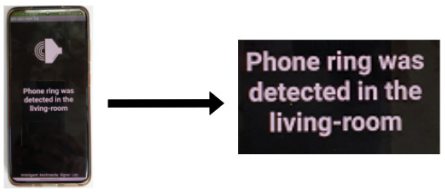

Fig. 6. (Color available online) Sound information delivery interface.

2.3 사운드 정보 전달 인터페이스

본 논문에서는 청각 장애인에게 효과적으로 사운드 정보를 제공하기 위해 검출된 사운드 이벤트의 명칭과 발생 위치를 텍스트로 변환한다.

Fig. 6는 실험 환경에서 발생한 사운드를 싱크에서 인식한 후에 인식된 사운드 이벤트의 정보가 스마트 기기에 자동으로 전송된 결과를 보여준다.

III. 실험 결과

3.1 실험 데이터

본 논문에서 제안한 방법의 성능을 평가하기 위해 59 m^2크기의 아파트에서 데이터베이스를 구성하였다. 아파트는 주방, 복도, 거실, 침실, 서재, 화장실로 구성되어 있으며 각 방에는 6개의 마이크를 설치하였다. 마이크를 통해 수집된 데이터 세트는 총 2300개이며 모두 16 kHz 샘플링레이트와 24비트 해상도로 녹음되었다. 데이터는 5 min ~ 15 min 길이의 다양한 사운드로 총 재생시간은 25000 min이다. 데이터에는 유리 깨지는 소리, 물체 또는 접시 떨어지는 소리, 비명 소리, 전화 또는 현관문 벨 소리, 대화 소리, 음악 소리, 망치소리, 물소리, 걷는 소리, 도로 소음, 강아지 소리, 문 잠그는 소리, 아기 울음소리, 청소기 소리, 박수 소리, 웃음소리 등이 포함되며 길고 짧은 소리가 1 ~ 3개씩 섞여 존재하나 일부 길이가 짧은 소리는 개별적으로 존재하는 경우도 있다. 그리고 물소리는 화장실과 주방에서만 발생하고, 현관문 벨소리, 문 잠그는 소리는 현관에서만 발생한다. 전체 데이터베이스 중 50 %는 학습 데이터로 사용하였고 20 %는 검증데이터, 나머지는 테스트 데이터로 사용하였다. 또한 무선 음향센서 네트워크를 통한 SED의 테스트 베드는 다음과 같이 설정하였다. 직선으로 균일하게 배치된 36개의 노드들 중 한쪽 끝의 노드는 싱크, 다른 한쪽 끝의 노드는 패킷을 생성하는 노드로 설정하였다. 전송 전력을 0 dBm으로 설정하여 약 4 미터의 전송 범위를 갖으며 무선 주파수는 890 MHz로 설정하였다. 패킷 크기는 36 바이트로 고정되어 링크 당 최대 133 패킷/초(pkt /s)의 용량을 갖는다. 무작위의 트래픽 로드에 대한 무선 음향센서 네트워크 연결을 시뮬레이션하기 위해 0.5 pkt/s의 패킷 생성 속도에 지연 (25 ms ~ 80 ms), 지터 (40 ms ~ 300 ms) 및 패킷 손실(2 % ~ 10 %)을 적용하였다. 또한 본 논문의 실험을 위해 8 % 패킷 손실률을 사용하였다.

3.2 측정 방식

제안하는 방법과 다른 SED 방법의 성능 비교를 위해 다채널 및 단 채널 오디오 신호로부터 TDOA, PR, LMST, MCN, SCN(단일 스케일 합성 곱 신경망) 의 특징 추출 방법을 사용하였다. 또한 분류기로는 GRNN과 BGRNN을 사용하여 성능을 비교하였으며 모두 200개의 GRU로 구성된 3개의 은닉층을 사용하였다. 신경망 학습의 손실 함수로는 binary cross-entropy를 사용하였다. 실험의 측정 지표는 1 s 단위의 세그먼트 기준으로 Error Rate (ER)와 F-score를 계산하였다.

3.3 실험 결과

Table 1은 실험 결과를 나타낸다. 제안하는 방식인 다채널 기반의 오디오 특징과 MCN을 사용한 경우 ER 0.46, F-score 92.5로 가장 우수한 성능을 보였으며 기존의 다채널 기반 오디오 특징만을 사용한 경우보다 성능이 증가하였다. 반면 SCN을 사용한 경우 성능이 크게 저하되었다. 따라서 시간-주파수 영역에서 다양한 분포를 갖는 사운드 이벤트들을 효과적으로 검출하기 위해서는 여러 크기의 합성 곱 필터를 사용하는 특징 추출 방법이 효과적임을 알 수 있다. 또한 MCN만을 사용한 경우 TDOA 혹은 PR과 함께

Table 1. Comparison of the error rate and F-score for different combinations of classifiers and features.

Methods			
Classifier	Feature	ER	F-score
BGRNN	TDOA, PR, MCN	0.46	92.5
BGRNN	TDOA, PR, SCN	0.60	87.3
BGRNN	TDOA, MCN	0.50	91.2
BGRNN	PR, MCN	0.51	90.1
BGRNN	MCN	0.58	89.4
BGRNN	TDOA, LMST [1]	0.48	90.4
GRNN	TDOA, PR, MCN	0.49	90.3

사용하는 경우보다 성능이 감소하였다. 이는 다채널 기반의 특징 값을 사용하는 방법이 중첩된 사운드 이벤트를 효과적으로 구분 짓는다는 것을 확인할 수 있다. 분류기 성능 비교로는 GRNN을 적용한 경우가 BGRNN에 적용한 경우 보다 낮은 성능을 보였다. 이는 이전 시간의 특징 정보뿐만 아니라 다음 시간의 특징 정보 또한 고려한 학습 방법이 SED에 효과적임을 알 수 있다. 또한 6개의 방들 중 사운드 이벤트가 발생한 하나의 방을 감지하는 위치 추정은 채널 선택 과정을 적용한 경우 센서 네트워크만을 사용한 경우보다 5 % 향상된 98 %의 높은 정확도를 보였다.

IV. 결 론

본 논문에서는 청각 장애인을 위한 소리 감지 홈 모니터링을 위해 다채널 다중 스케일 신경망을 사용한 SED 방식을 제안하였다. 제안한 방식은 다채널 오디오 신호로부터 추출한 MCN기반의 특징을 BGRNN에 적용하였다. 실험 결과를 통해 제안한 방식의 성능이 기존 방법보다 뛰어남을 확인할 수 있었다. 향후 본 연구를 바탕으로 다양한 신경망 기반의 다중 스케일 특징 추출 방법을 SED에 적용하고자 한다.

감사의 글

본 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018R1D1A1B07041783).

References

1. G. Y. Kim, S.-S. Shin, and H.-G. Kim, “Home monitoring system based on sound event detection for the hard-of-hearing” (in Korean), J. Acoust. Soc. Kr. **38**, 427-432 (2019).
2. K. Zhang, Y. Cai, Y. Ren, R. Ye, and L. He, “MTF-CRNN: Multiscale time-frequency convolutional recurrent neural network for sound event detection,” IEEE Access, **8**, 147337-147348 (2020).
3. B. H. Kim, H.-G. Kim, J. Jeong, and J. Y. Kim, “VoIP receiver-based adaptive playout scheduling and packet loss concealment technique,” IEEE Trans. Consum. Electron. **59**, 250-258 (2013).
4. K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj, “Channel selection based on multichannel cross-correlation coefficients for distant speech recognition,” Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays, 1-6 (2011).
5. D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, “Real-time multiple sound source localization and counting using a circular microphone array,” IEEE Trans. Audio, Speech, Lang. Process. **21**, 2193-2206 (2013).
6. B. Uzkent, B. D. Barkana, and H. Cevikalp, “Non-speech environmental sound classification using svms with a new set of features.” Int. J. ICIC. **8**, 3511-3524 (2012).

저자 약력

▸ 이 기 용 (Gi Yong Lee)

2020년 2월 : 광운대학교 전자융합공학과 학사
2020년 3월 ~ 현재 : 광운대학교 전자융합공학과 석사과정

▸ 김 형 국 (Hyoung-Gook Kim)

1999년 ~ 2002년 : 독일 SIEMENS/Cortologic AG 책임연구원
2002년 ~ 2005년 : 독일 베를린 공과대학교 Assistant Professor
2005년 ~ 2007년 : 삼성종합기술원 수석 연구원
2007년 3월 ~ 현재 : 광운대학교 전자융합공학과 교수