특집논문 (Special Paper)
방송공학회논문지 제25권 제6호, 2020년 11월 (JBE Vol. 25, No. 6, November 2020)
https://doi.org/10.5909/JBE.2020.25.6.845
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

산업현장에서의 선택적 소음 제거를 위한 환경 사운드 분류 기술

최 현 국[a)], 김 상 민[a)], 박 호 종[a)‡]

Environmental Sound Classification for Selective Noise Cancellation in Industrial Sites

Hyunkook Choi[a)], Sangmin Kim[a)], and Hochong Park[a)‡]

요 약

본 논문에서는 산업현장에서의 선택적 소음 제거를 위한 환경 사운드 분류 기술을 제안한다. 산업현장에서의 소음은 작업자의 청력 손실의 주요 원인이 되며, 소음 문제를 해결하기 위한 소음 제거 기술이 널리 연구되고 있다. 그러나 기존 소음 제거 기술은 모든 소리를 구분 없이 차단하는 문제를 가지며, 모든 소음에 공통된 제거 방법을 적용하여 각 소음에 최적화된 소음 제거 성능을 보장할 수 없다. 이러한 문제를 해결하기 위해 사운드 종류에 따라 선택적 동작을 하는 소음 제거가 필요하고, 본 논문에서는 이를 위해 딥 러닝 기반의 환경 사운드 분류 기술을 제안한다. 제안 방법은 기존 오디오 특성인 멜-스펙트로그램의 한계를 극복하기 위해 새로운 특성으로서 멜-스펙트로그램 기반의 시간 변화 특성과 통계적 주파수 특성을 사용하며, 합성곱 신경망을 이용하여 특성을 모델링 한다. 제안하는 분류기를 사용하여 3가지 소음과 2가지 비소음으로 구성된 총 5가지 클래스로 사운드를 분류하였고, 제안하는 오디오 특성을 사용하여 기존 멜-스펙트로그램 특성을 사용할 때에 비하여 분류 정확도가 6.6% 포인트 향상되는 것을 확인하였다.

Abstract

In this paper, we propose a method for classifying environmental sound for selective noise cancellation in industrial sites. Noise in industrial sites causes hearing loss in workers, and researches on noise cancellation have been widely conducted. However, the conventional methods have a problem of blocking all sounds and cannot provide the optimal operation per noise type because of common cancellation method for all types of noise. In order to perform selective noise cancellation, therefore, we propose a method for environmental sound classification based on deep learning. The proposed method uses new sets of acoustic features consisting of temporal and statistical properties of Mel-spectrogram, which can overcome the limitation of Mel-spectrogram features, and uses convolutional neural network as a classifier. We apply the proposed method to five-class sound classification with three noise classes and two non-noise classes. We confirm that the proposed method provides improved classification accuracy by 6.6% point, compared with that using conventional Mel-spectrogram features.

Keyword : industrial noise, sound classification, deep neural network, industrial site

Ⅰ. 서 론

산업현장에서의 소음은 작업자의 청각 기관에 심각한 손상을 주는 환경 오염원으로서 청력 손실로 인한 소음성 난청을 일으킨다[1]. 이러한 문제를 예방하기 위해서 청력보호를 위한 능동형 소음 제거 (active noise cancellation, ANC)에 관한 많은 연구가 진행되고 있고, 산업현장에서 널리 사용되고 있다[2]. ANC는 외부 사운드를 측정하고 측정된 신호를 적응 필터에 통과시켜 소음 방지 신호 (anti-noise)를 생성하여 외부 사운드가 귀에 입력되는 것을 차단한다. 그러나 ANC에 의하여 모든 외부 사운드는 구분 없이 차단되고, 그에 따라 사용자는 주변 사람의 음성이나 특정 경고 상황을 알리는 알람, 사이렌 등의 중요한 소리를 선택적으로 듣지 못하는 문제가 발생한다. 즉, ANC 사용에 따라 작업자의 의사 전달 및 정보 전달을 힘들게 하여 업무의 효율을 하락시키고 위험 상황에 대한 빠른 인지를 막아 안전에 취약한 환경을 만든다. 또한, 다양한 신호 특성을 고려하지 않은 공통된 ANC 동작은 각 소음에 최적화된 성능을 보장할 수 없다.

이와 같은 ANC 사용의 문제점을 해결하기 위하여 주변 사운드 특성에 따른 선택적 소음 제거 기능을 제공하는 ANC 동작이 필요하다. 본 논문에서는 선택적 소음 제거에 필요한 핵심모듈로서 딥 러닝 기반으로 산업현장 사운드를 분류하는 기술을 제안한다. 본 논문에서는 공사장, 공장, 비행장 등과 같은 산업현장에서 발생하는 사운드를 다루며, 타격음 (attack), 회전음 (rotation), 마찰음 (friction), 음성 (speech), 경고음 (alarm) 등의 5개 클래스를 정의하고 입력 사운드를 5개 클래스로 분류한다. 타격음은 망치질과 같은 순간적인 충격 소음을, 회전음은 모터가 회전하는 소리와 같은 지속적인 소음을, 마찰음은 톱질과 같은 불규칙한 소음을, 경고음은 특정 경고 상황을 알리는 알람과 사이렌을 의미한다. 또한, ANC에서 차단할 소음과 차단하지 않고 청취해야 할 비소음을 각각 정의하고, 입력 사운드를 2개 클래스로 분류하는 과정도 진행한다. 여기서 타격음, 회전음, 마찰음이 소음 클래스이고, 음성과 경고음이 비소음 클래스이다.

기존의 환경 사운드 분류 방법은 오디오 특성으로 멜-스펙트로그램 (Mel-spectrogram)을 가장 널리 사용한다[3,4]. 그러나 본 논문에서 다루는 산업현장의 사운드 분류에서 멜-스펙트로그램을 특성으로 사용할 때 두 가지 문제점이 발생한다. 첫 번째로, 다양한 종류의 사운드 특성을 분석한 결과에 의하면 시간에 따른 스펙트럼의 변화가 사운드 종류를 구분하는 주요한 특성인데, 멜-스펙트로그램은 시간에 따른 변화 특성을 직접적으로 반영하지 못하여 분류를 어렵게 한다. 두 번째로, 멜 필터를 통과한 멜-스펙트로그램에서 저대역의 해상도는 높지만 주파수 대역이 올라갈수록 해상도가 낮아지며, 그에 따라 비소음의 고유 특성인 하모닉 구조를 왜곡시켜 소음과 비소음의 분류를 어렵게 한다.

이와 같은 멜-스펙트로그램 특성의 문제점들을 해결하기 위해 본 논문에서는 멜-스펙트로그램 기반의 시간 변화 특성과 스펙트럼 기반의 통계적 주파수 특성을 새롭게 제안한다[5]. 최종적으로 멜-스펙트로그램을 포함하여 세 종류의 오디오 특성을 사용하고 합성곱 신경망(convolutional neural network, CNN)으로 특성을 모델링 하여 환경 사운드를 분류한다[6]. 제안하는 특성 벡터를 사용하여 5-클래스 분류와 2-클래스 분류를 각각 수행하고 분류 성능을 측정하였고, 멜-스펙트로그램만을 단독으로 사용하는 기존 방법에 비하여 모든 클래스에 대한 분류 정확도가 향상되는 것을 확인하였다. 특히, 선택적 ANC 동작의 핵심이 되는 비소음 분류에서 매우 우수한 성능을 제공하는 것을 확인하였다.

a) 광운대학교 전자공학과 (Dept. of Electronics Engineering, Kwangwoon University)
‡ Corresponding Author : 박호종(Hochong Park)
E-mail: hcpark@kw.ac.kr
Tel: +82-2-940-5104
ORCID: https://orcid.org/0000-0003-1600-6610
※ 이 논문의 연구 결과 중 일부는 “2020년 한국방송·미디어공학회 하계학술대회”에서 발표한 바 있음.
※ 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2018-0-01407). (The present Research has been conducted by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.2018-0-01407)).
· Manuscript received September 7, 2020; Revised October 21, 2020; Accepted October 21, 2020.

Ⅱ. 제안하는 사운드 분류 방법

1. 특성 벡터 추출

타격음, 회전음, 마찰음, 음성, 경고음의 분류를 위해 각

클래스의 특징을 스펙트로그램으로 분석하였다. 그림 1은 각 클래스 신호의 스펙트로그램의 예이다. 타격음은 순간적인 충격이 가해진 이후 에너지가 급격히 감소하는 것을 알 수 있고, 회전음은 시간에 따라 에너지의 큰 변화 없이 일정한 모습을 보여준다. 마찰음은 에너지가 급격히 감소하는 타격음의 특성과 에너지가 일정한 회전음의 특성을 모두 가진다. 음성은 하모닉 구조를 가지며 시간에 따라 그 구조가 변화하는 것을 알 수 있다. 경고음은 음성과 같이 강한 하모닉 구조를 가지지만 시간에 따른 변화는 크게 존재하지 않는다. 이러한 각 클래스 신호의 고유 성질 차이를 특성 벡터에 반영하기 위해 총 3가지의 특성 벡터를 사용한다. 첫 번째 특성 벡터로 기존의 멜-스펙트로그램 (X_{mel})을 사용하고, 멜-스펙트로그램의 한계를 극복하기 위하여 본 논문에서 새로운 두 가지 특성 벡터를 제안한다. 즉, 시간에 따른 에너지의 변화를 활용하기 위해 두 번째 특성 벡터로 멜-스펙트로그램 기반의 시간 특성 (X_{temp})을 제안하고, 하모닉 구조의 특성을 활용하기 위해 세 번째 특성 벡터로 주파수의 통계적 특성 (X_{stat})을 제안한다.

샘플링 주파수가 16 kHz인 입력 신호에 프레임 길이 40 ms와 50% 오버랩으로 640-포인트 Fourier 변환을 적용하여 500 ms 신호에 대한 스펙트로그램을 생성한다. 이 스펙트로그램으로부터 3가지 특성 벡터를 추출하며, 각 특성 벡터의 구조와 추출 방법은 다음과 같다. 첫 번째 특성 벡터는 멜-스펙트로그램 X_{mel}이다. 인간의 인지 기준에 따라 헤르츠 단위의 주파수를 멜 스케일 (Mel-scale)로 변환하고, 일정한 멜 간격으로 멜 필터를 정의한다. 다음, 각 프레임의 스펙트럼에 멜 필터를 적용하고 각 멜 밴드의 로그 에너지 값 $m_{i,j}$ 을 구하여 X_{mel}을 얻는다. 여기서 i는 멜 밴드 인덱스로 $0 \le i < p$의 범위를 가지고 p는 멜 밴드 개수이다. j는 시간 축 프레임 인덱스로 $0 \le j < 25$ 의 범위를 가진다. 본 논문에서는 j의 범위를 500 ms의 텍스처 프레임 (texture frame)으로 정의한다. 따라서 X_{mel}는 $p \times 25$ 크기의 2차원 구조를 가진다. 그림 2는 각 클래스 신호의 30 밴드 멜-스펙트로그램의 예를 보여준다.

두 번째 특성 벡터인 X_{temp}은 멜-스펙트로그램의 시간 기반 특성으로, 멜-스펙트로그램의 시간에 따른 에너지 변화를 특성으로 사용한다. X_{temp}는 $n_{i,j}$로 구성되고, 식 (1)과 같이 텍스처 프레임 내에서 각 멜 밴드별로 $m_{i,j}$의 최대값

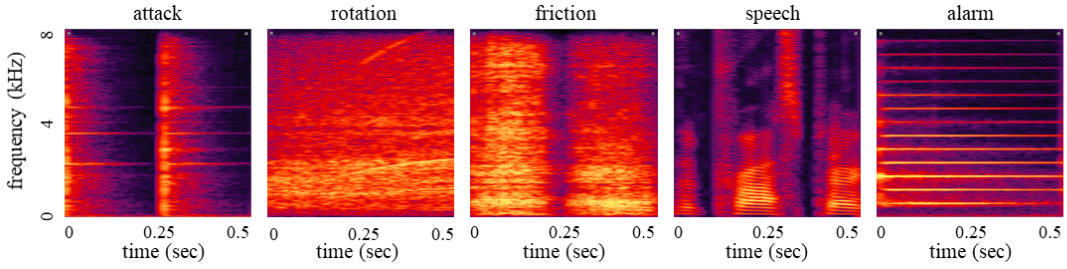

그림 1. 각 클래스 신호의 스펙트로그램
Fig. 1. Spectrogram of signal in each class

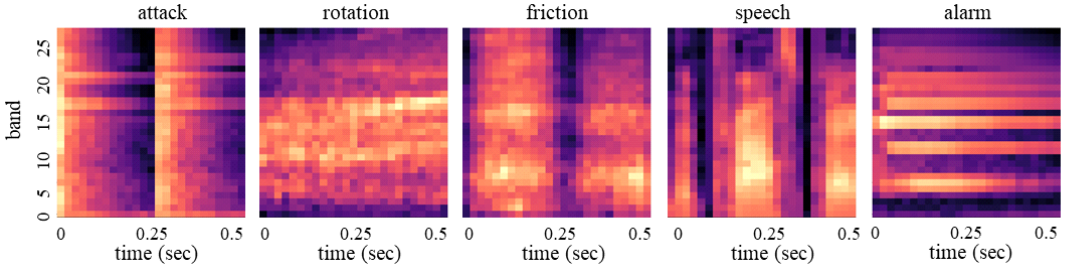

그림 2. 각 클래스 신호의 30 밴드 멜-스펙트로그램 X_{mel}
Fig. 2. 30-band Mel-spectrogram X_{mel} of signal in each class

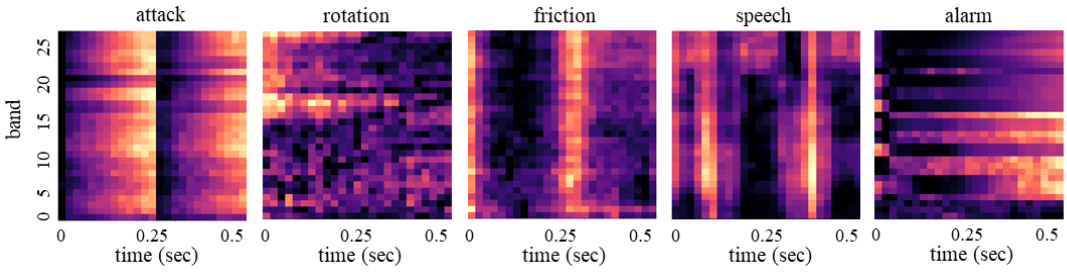

그림 3. 각 클래스 신호의 시간 기반 특성 X_{temp}

Fig. 3. Temporal features X_{temp} of signal in each class

인 $m_{i,max}$를 구하고, 식 (2)와 같이 $m_{i,max}$와 $m_{i,j}$의 차이에 해당하는 $n_{i,j}$을 구한다. X_{temp}은 X_{mel}과 동일하게 $p \times 25$ 크기의 2차원 구조를 가진다. 그림 3은 각 클래스 신호의 30 밴드 X_{temp}의 예를 보여주며, X_{temp}이 멜-스펙트로그램에서 시간에 따라 에너지가 변화하는 구간을 강조해 주는 것을 알 수 있다.

$$m_{i,max} = max(m_{i,0}, \ldots, m_{i,24}) \tag{1}$$

$$n_{i,j} = m_{i,max} - m_{i,j} \tag{2}$$

세 번째 특성 벡터 X_{stat}은 스펙트럼 기반의 통계적 주파수 특성이다. X_{stat}은 g_b로 구성되고, b는 스펙트럼의 주파수 bin 인덱스로 $0 \leq b < 321$의 범위를 가지고, 멜 밴드에 비하여 매우 높은 주파수 해상도를 제공한다. j번째 프레임의 로그 스펙트럼을 $f_{b,j}$라 할 때, g_b는 각 b에 대하여 $f_{b,j}$의 텍스처 프레임 평균이다. 그림 4는 각 클래스 신호의 X_{stat} 예를 보여주며, 강한 하모닉 구조를 가지고 그 구조가 시간에 따라 변하지 않고 일정하게 유지되는 경고음의 고유 특성을 뚜렷하게 반영하는 것을 알 수 있다.

그림 5는 제안하는 분류 방법에서 사용하는 전체 특성 벡터의 구성을 보여준다. X_{mel}와 X_{temp}의 크기는 $p \times 25$이고 주파수 축과 시간 축의 2차원 구조를 가지고, X_{stat}의 크기는 321이고 주파수 축의 1차원 구조를 가진다. 이러한 구조적 차이 때문에 CNN 기반의 분류기에서 X_{mel}와 X_{temp}는 2차원 CNN에 입력하고 X_{stat}는 1차원 CNN에 입력하는 이중 CNN 구조를 가져야 한다.

$$X_{mel} = \begin{bmatrix} m_{p,0} & \cdots & m_{p,24} \\ \vdots & \ddots & \vdots \\ m_{0,0} & \cdots & m_{0,24} \end{bmatrix} \quad X_{temp} = \begin{bmatrix} n_{p,0} & \cdots & n_{p,24} \\ \vdots & \ddots & \vdots \\ n_{0,0} & \cdots & n_{0,24} \end{bmatrix} \quad X_{stat} = [g_0 \ldots g_b \ldots g_{320}]$$

$$m_{i,max} = max(m_{i,0}, \ldots, m_{i,24}) \qquad g_b = \frac{1}{25}\sum_{j=0}^{24} f_{b,j}$$

$$n_{i,j} = m_{i,max} - m_{i,j}$$

그림 5. 제안하는 특성 벡터의 구성

Fig. 5. Composition of proposed feature vectors

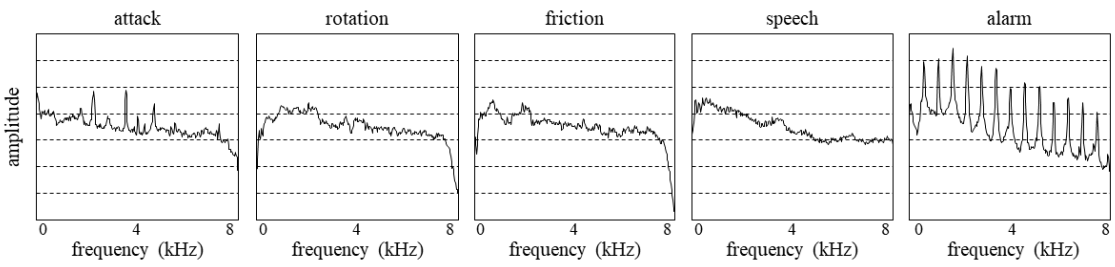

그림 4. 각 클래스 신호의 통계적 주파수 특성 X_{stat}

Fig. 4. Statistical frequency features X_{stat} of signal in each class

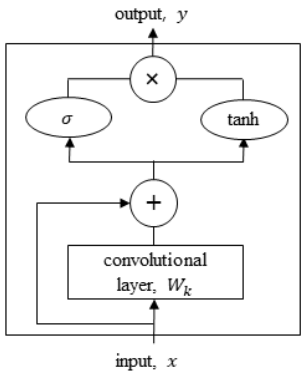

그림 6. 첫 번째 합성곱 단의 구조
Fig. 6. Structure of the first convolution stage

2. 네트워크 구조

본 논문에서는 입력 신호에서 추출한 특성 벡터를 CNN을 사용하여 모델링한다[6-8]. 크기가 동일한 2차원 구조의 X_{mel}과 X_{temp}을 2 채널 구조로 입력하는 2차원 CNN과, 1차원 구조의 X_{stat}을 입력하는 1차원 CNN를 각각 독립적으로 동작시키고, 평탄화 (flattening) 단계에서 각 CNN 출력을 합친 후 fully-connected layer를 통과시켜 최종 분류를 진행한다. 첫 번째 합성곱 (convolution) 단에서는 그림 6의 gated 활성화 함수를 사용하고[9], 그 이후의 단에서는 rectified linear unit (ReLU) 활성화 함수를 사용하며, 각 단마다 max pooling을 통과시킨다. 마지막 출력층은 softmax 활성화 함수를 사용한다. 그림 7은 제안한 분류 방법에서 사용하는 네트워크의 전체 구조를 보여주며, 2차원 CNN과 1차원 CNN의 이중 구조를 가진다. 2차원 CNN 과정에서의 괄호는 (밴드 수, 프레임 수, 채널 수)를 의미하고, 1차원 CNN 과정에서는 (bin 수, 채널 수)를 의미한다.

Ⅲ. 성능 평가

성능 평가에는 TIMIT와 Sound-Ideas의 Industry & Office 데이터 세트를 사용하였다[10,11]. Industry & Office 데이터 세트로부터 타격음, 회전음, 마찰음, 경고음을 분류하여 각 클래스를 구성하였고, TIMIT를 사용하여 음성 클래스를 구성하였다. 각 인스턴스 (instance)는 500 ms 단위로 추출하였고, 표 1은 각 클래스의 인스턴스 개수이다. 모든 특성 벡터는 zero mean과 unit variance로 정규화 하여 사용하였다. 성능 평가에서의 최종 판정은 500 ms 입력마다 클래스를 판정하고 이웃한 1초에 대한 2번의 판정을

표 1. 각 클래스의 인스턴스 개수
Table 1. The number of instances for each class

class	noise			non-noise	
	attack	rotation	friction	speech	alarm
no.	1215	4943	1044	927	1357

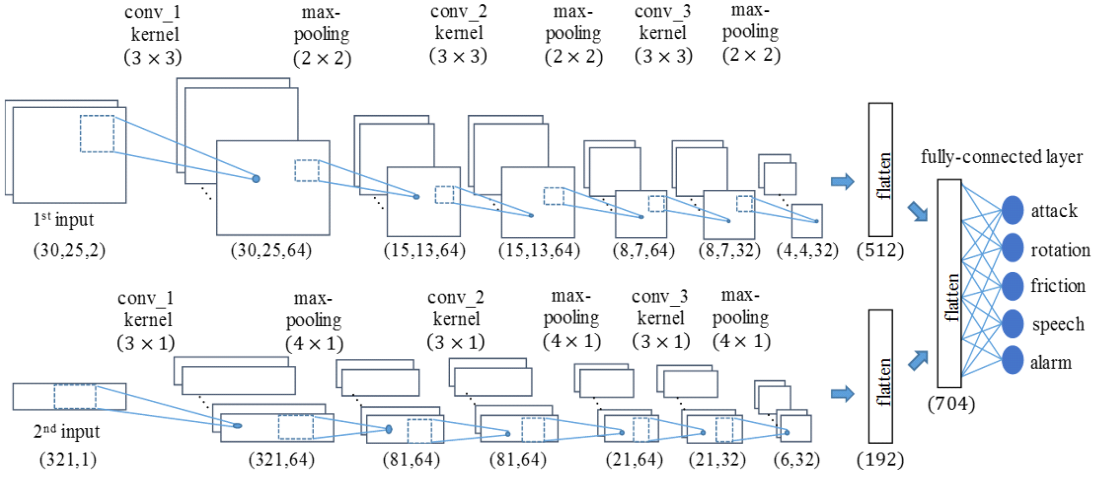

그림 7. 제안하는 분류기의 네트워크 구조
Fig. 7. Network structure of the proposed classifier

soft voting 하여 1초 단위로 진행하였다. 음원의 길이가 짧아 2번의 판정이 불가능할 때는 1번의 판정을 최종 판정으로 사용하였다.

네트워크 학습 과정에서 가중치와 바이어스의 초기화는 He 초기화로 하였고[12], Adam 최적화를 사용하였으며[13], mini-batch size는 256, learning rate는 0.001로 설정하였다. 학습에서의 과적합을 막기 위해 early stopping과 dropout을 사용하였고[14], early stopping의 patience는 30으로 설정하고 dropout의 keep probability는 0.8로 설정하였다.

작은 학습 데이터를 사용하는 학습에서 성능 평가의 신뢰성을 높이기 위해 10-fold 교차 검증 (cross validation)을 사용하였다[15]. 전제 데이터 세트를 10개 fold로 나누고 이 중 1개 fold는 실험 (testing) 데이터 세트, 다른 1개 fold는 검증 (validation) 데이터 세트, 나머지 8개 fold는 학습 데이터 세트로 사용하였다. 이것을 순환시켜 총 10번의 학습으로 모든 fold가 한 번씩 평가되도록 하였다. 음원의 길이가 서로 다르므로 fold 분할을 음원 단위로 하면 fold마다 각 클래스별 인스턴스 개수의 차이가 심해진다. 따라서 음원 단위 대신에 인스턴스 개수를 기준으로 fold마다 클래스별 인스턴스의 개수가 거의 균일하도록 fold 분할을 하였다. 또한, 하나의 음원으로부터 추출된 인스턴스는 여러 fold에 분리되지 않고 하나의 fold에만 들어가도록 하였다.

성능 비교를 위한 베이스라인 (baseline)을 설정하기 위해 대표적인 오디오 특성인 멜-스펙트로그램 X_{mel}만을 사용하는 분류기를 설계하였고, 베이스라인의 네트워크는 그림 7의 2차원 CNN과 동일하고 입력 채널 수를 1로 변경하여 사용하였다. 멜 밴드 수에 따른 성능을 평가하였고, 표 2에서 보듯이 성능이 가장 좋은 30개 멜 밴드를 베이스라인의 규격으로 설정하였다. 표 3은 30 밴드를 사용하는 베이스라인의 혼동 행렬을 보여주고, 0.3% 이하의 값은 -로 표시하였다. 마찰음과 산업현장에서의 안전에 큰 영향을 미치는 경고음의 분류 성능이 낮은 문제를 가진다. 그림 1에서 보듯이 마찰음의 고유 특징은 시간에 따른 에너지의 급격한 감소이고 경고음의 고유 특징은 강한 하모닉 구조인데, X_{mel}이 이와 같은 특징을 잘 표현하지 못하기 때문에 해당 클래스의 분류 성능이 낮다.

표 2. 멜 밴드 수에 따른 분류 성능
Table 2. Classification accuracy for each number of mel bands

number of mel bands, p	accuracy (%)
10	82.7
15	82.8
20	84.7
25	87.5
30	88.3
35	85.8
40	85.5
45	86.6
50	86.6

표 3. 멜-스펙트로그램을 사용하는 베이스라인의 혼동 행렬
Table 3. Confusion matrix of baseline using mel-spectrogram

true \ pred.	attack	rotation	friction	speech	alarm	recall (%)
attack	93.6	4.2	1.8	-	-	93.6
rotation	0.7	94.1	1.7	-	3.5	94.1
friction	4.9	15.0	71.1	3.4	5.6	71.1
speech	-	0.4	-	99.6	-	99.6
alarm	-	13.1	2.4	1.3	83.0	83.0
precision(%)	93.3	92.2	83.6	94.5	83.1	88.3

표 4는 X_{mel}과 X_{temp}을 입력으로 사용할 때의 혼동 행렬이고, 이 때 사용한 네트워크는 그림 7의 2차원 CNN과 동일하다. X_{temp}은 멜-스펙트로그램의 시간에 따른 에너지 변화를 반영한 특성 벡터이고, 뚜렷한 에너지 변화 특성을 가지는 타격음과 마찰음에서의 성능 향상이 가능하며, 각각 베이스라인과 비교하여 2.7%p와 3.1%p의 recall 향상을 얻고 전체 recall이 1.9%p 향상된 것을 확인하였다.

표 4. X_{mel}과 X_{temp}을 사용할 때의 혼동 행렬
Table 4. Confusion matrix when using X_{mel} and X_{temp}

true \ pred.	attack	rotation	friction	speech	alarm	recall (%)
attack	96.3	1.6	1.6	-	-	96.3
rotation	1.0	95.6	1.5	-	1.9	95.6
friction	5.1	18.8	74.2	0.9	0.9	74.2
speech	-	-	-	99.8	-	99.8
alarm	4.6	8.4	1.8	-	85.1	85.1
precision(%)	88.5	93.5	86.2	98.5	91.5	90.2

표 5는 X_{mel}과 X_{stat}을 사용할 때의 혼동 행렬이고, 그림 7에서 2차원 CNN의 입력 채널 수를 1개로 변경하여 사용하였다. 하모닉 구조의 특성을 반영하는 X_{stat}을 통해 경고음의 성능 향상이 가능하고, 베이스라인 대비 7.4%p의 recall 향상을 확인하였다. 또한, 베이스라인에서 마찰음이 음성과 경고음으로 오분류 되던 문제점이 개선되어 마찰음 recall이 12.7%p 향상되었고, 전체 recall이 4.2%p 향상된 것을 확인하였다.

표 5. X_{mel}과 X_{stat}을 사용할 때의 혼동 행렬
Table 5. Confusion matrix when using X_{mel} and X_{stat}

true \ pred.	attack	rotation	friction	speech	alarm	recall (%)
attack	92.1	4.7	1.6	-	1.3	92.1
rotation	1.7	96.1	0.5	-	1.6	96.1
friction	4.7	11.5	83.8	-	-	83.8
speech	-	-	-	100.0	-	100.0
alarm	-	9.4	-	-	90.4	90.4
precision(%)	90.0	93.9	94.7	99.4	92.8	92.5

표 6은 X_{mel}, X_{temp}, X_{stat}을 모두 사용하는 최종 제안 분류기의 혼동 행렬이고, 그림 7의 네트워크를 사용하였다. 베이스라인에 비하여 마찰음과 경고음의 recall이 각각 18.6%p와 8.6%p 향상되었고 전체 recall은 6.6%p 향상된 것을 확인하였다. 특히, 베이스라인에 비하여 클래스 사이의 recall 편차가 감소한 결과를 얻었다. 표 7은 소음과 비소음의 2-클래스 분류에 대한 성능을 보여준다. Class merge는 기존 5-클래스 분류기를 그대로 사용하고 타격음, 회전음, 마찰음을 소음으로 결합하고 음성과 경고음을 비소음 결합하여 2개 클래스로 분류할 때의 성능이다. New label은 모든 신호를 소음과 비소음으로 새로 라벨링하고 2-클래스 분류기를 학습하여 분류할 때의 성능이다. 두 방법 모두에서 제안 방법이 베이스라인보다 1.2 ~ 3.7%p의 정확도 향상을 제공한다.

표 6. X_{mel}, X_{temp}, X_{stat}을 사용할 때의 혼동 행렬
Table 6. Confusion matrix when using X_{mel}, X_{temp} and X_{stat}

true \ pred.	attack	rotation	friction	speech	alarm	recall (%)
attack	96.8	2.8	-	-	-	96.8
rotation	1.6	96.9	-	-	1.2	96.9
friction	-	10.2	89.7	-	-	89.7
speech	-	-	-	99.8	-	99.8
alarm	0.6	7.7	-	-	91.6	91.6
precision(%)	93.3	95.1	98.8	99.4	95.2	94.9

표 7. 소음과 비소음의 분류 성능
Table 7. Classification accuracy for noise and non-noise

class	accuracy (%)			
	baseline		proposed method	
	class merge	new label	class merge	new label
noise	96.3	97.5	99.1	98.5
non-noise	90.7	93.5	95.2	95.0
total	93.4	95.5	97.1	96.7

이상의 성능 평가를 통하여 제안하는 분류기가 베이스라인에 비하여 모든 클래스에서 분류 성능을 향상시키는 것을 확인하였다. 또한, 제안 방법이 5-클래스 분류와 2-클래스 분류에서 모두 성능 향상을 제공하고, 특히 각 클래스의 성능 편차를 줄이는 효과도 제공한다. 따라서 제안하는 특성이 기존 멜-스펙트로그램의 한계점을 보완하여 산업현장에서의 환경 사운드 분류에 필요한 핵심 특성을 추가로 제공하는 것을 확인할 수 있다.

Ⅳ. 결 론

본 논문에서는 산업현장에서의 선택적 소음 제거를 위한 딥 러닝 기반의 환경 사운드 분류 기술을 제안하였다. 산업현장에서의 사운드를 차단할 소음과 차단하지 않고 청취해야 할 비소음으로 분류하고, 또한 소음 제거 기술의 성능 향상을 위해 사운드 특징을 세분화 하여 총 5개 클래스로 사운드를 분류하였다. 오디오의 대표적 특성인 멜-스펙트로그램의 한계를 극복하기 위하여 시간 축 변화와 주파수 통계 특성을 반영하는 새로운 오디오 특성을 제안하였고, CNN 기반의 분류기를 사용하여 5개 클래스에 대한 분류

성능과 2개 클래스에 대한 분류 성능을 각각 측정하였다. 그 결과, 제안한 특성 벡터를 사용한 5-클래스 분류의 평균 정확도는 94.9%이고, 기존 멜-스펙트로그램 방법에 비하여 6.6%p 향상된 성능을 제공한다. 또한, 소음과 비소음의 2-클래스 분류에서도 기존 대비 1.2 ~ 3.7%p 향상된 성능을 제공한다.

참 고 문 헌 (References)

[1] A. Dzhambov and D. Dimitrova, "Occupational noise exposure and the risk for work-related injury: a systematic review and meta-analysis," *Annals of Work Exposures and Health,* Vol. 61, No. 9, pp. 1037-1053, Nov. 2017.

[2] S. Kuo and D. Morgan, "Active noise control : a tutorial review," *Proceedings of the IEEE*, Vol. 87, No. 6, pp. 943 - 973, June 1999.

[3] S. Suh, W. Lim, Y. Jeong, T. Lee and H. Kim, "Dual CNN structured sound event detection algorithm based on real life acoustic dataset," *J. of Broadcast Engineering*, Vol. 23, No. 6, pp. 855-865, 2018.

[4] K. J. Piczak, "Environmental sound classification with convolutional neural networks," *in Proc. of IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, pp. 1-6, Sep. 2015.

[5] H. W. Yun, S. H. Shin, W. J. Jang and H. Park, "On-line audio genre classification using spectrogram and deep neural network," *J. of Broadcast Engineering*, Vol. 21, No. 6, pp. 977-985, Nov. 2016.

[6] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, 521.7553, pp. 436-444, May 2015.

[7] X. Glorot, A. Bordes and Y. Bengio, "Deep sparse rectifier neural networks," *in Proc. of Int. Conf. on Artificial Intelligence and Statistics,* pp. 315 - 323. 2011.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.

[9] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv:1605.07146*, 2016.

[10] V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, Vol. 9, No. 4, pp. 351-356, Aug. 1990.

[11] https://www.sound-ideas.com/Collection/54/2/0/Industry-Machinery-Tools-and-Office-SFX (accessed May 2019)

[12] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," *in Proc. IEEE Int. Conf. on Computer Vision*, Chile, pp. 1026 - 1034, 2015.

[13] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv:1412.6980*, 2014.

[14] N. Srivastava, G. Hinton, A. Krizhevesky and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. of Machine Learning Research*, Vol. 15, No. 1, pp. 1929-1958, June 2014.

[15] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Advances in Neural Information Processing Systems*, pp. 231 - 238, 1995.

저 자 소 개

최 현 국

- 2020년 2월 : 광운대학교 전자공학과 학사
- 2020년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0001-6536-4997
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

저 자 소 개

김 상 민

- 2020년 2월 : 신한대학교 전자공학과 학사
- 2020년 3월 ~ 현재 : 광운대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0000-0001-9090-5513
- 주관심분야 : 오디오/음성 신호처리, 딥 러닝

박 호 종

- 1986년 2월 : 서울대학교 전자공학과 공학사
- 1987년 12월 : Univ. of Wisconsin-Madison 공학석사
- 1993년 5월 : Univ. of Wisconsin-Madison 공학박사
- 1993년 9월 ~ 1997년 8월 : 삼성전자 선임연구원
- 1997년 9월 ~ 현재 : 광운대학교 전자공학과 교수
- ORCID : https://orcid.org/0000-0003-1600-6610
- 주관심분야 : 오디오/음성 신호처리, 3D 오디오, 음악정보처리