선택적 노이즈 캔슬링을 위한 딥 러닝 기반의 환경 인지 기술

*최현국 김상민 한석현 신성현 박호종
광운대학교
*chk0222@naver.com

Deep learning based environmental sound classification for selective noise canceling

*Choi, Hyunkook Kim, Sangmin Han, Seokhyeon Shin, Seong-Hyeon Park, Hochong
Kwangwoon University

요약

본 논문에서는 선택적 노이즈 캔슬링을 위한 환경 인지 기술을 제안한다. 기존의 노이즈 캔슬링은 모든 소리를 구분 없이 차단하여 여러 가지 문제를 유발할 수 있으며 공통된 노이즈 캔슬링 동작으로 각 소음에 최적화된 성능을 보장할 수 없다. 이러한 문제를 해결하기 위해 제안하는 방법은 대표적 오디오 특성인 멜-스펙트로그램과 스펙트로그램 기반의 시간적 특성 벡터를 사용하여 환경 인지를 진행한다. 본 논문에서는 attack, rotation, sawing으로 구성된 3가지 소음과 speech, tonal로 구성된 2가지 비 소음으로 총 5가지 클래스를 분류한다. 제안하는 방법에서 특성 벡터로 멜-스펙트로그램만을 사용했을 때 87.5%의 분류 성능을 보였으며, 스펙트로그램 기반의 시간적 특성을 추가했을 때 분류 성능이 91.2%로 향상되었다.

1. 서론

최근 특정 직업군의 작업환경에서 발생하는 소음으로 인해 청력 손실을 겪는 사람들이 많아지고 있다[1]. 특히 소음이 심한 공사장과 같은 작업장에서 일을 하는 군인, 공사장 근로자 등의 직업군에서 소음성 난청이 문제 되고 있다. 이러한 문제를 예방하기 위해서 노이즈 캔슬링 기술에 대한 연구가 진행되고 있으며, 실제 현장에서 사용되고 있다[2]. 하지만 기존의 노이즈 캔슬링 기술은 모든 소리를 구분 없이 차단하는 것을 목적으로 하여 사람의 음성이나 특정 상황을 알리는 알림, 경적소리 등의 중요한 소리를 선택적으로 듣지 못한다는 한계가 있다. 또한 환경을 고려하지 않은 공통된 노이즈 캔슬링 동작은 각 소음에 최적화된 성능을 보장할 수 없다. 따라서 본 논문에서는 선택적 노이즈 캔슬링 기술을 위한 딥 러닝 기반의 소음과 비 소음을 분류하는 환경 인지 방법을 제안한다.

본 논문은 환경 인지의 대상을 크게 공사 현장에서 주로 발생하는 소음과 차단하지 않고 들어야 하는 비 소음으로 나누었으며, 노이즈 캔슬링 기술의 성능 향상을 위해 소음과 비 소음의 세부 클래스를 추가적으로 구분하였다. 본 논문에서는 소음을 각 특성에 따라 attack, rotation, sawing 3가지의 특성으로 나누었으며, 비 소음을 speech와 tonal로 나누었다.

본 논문에서는 멜-스펙트로그램과 스펙트로그램 기반의 시간적 특성을 사용하여 이를 convolutional neural network (CNN)으로 학습하는 방식의 환경 인지 방법을 제안한다. 대표적 오디오 특성인 멜-스펙트로그램만을 사용했을 때 평균 87.5%의 분류 성능을 보였으며, 스펙트로그램 기반의 2가지 시간적 특성을 추가했을 때 평균 91.2%로 분류 성능이 3.7%p 향상되었다. 특히 소음 클래스인 attack, rotation, sawing 사이의 에러가 크게 감소하였다.

2. 제안하는 방법

2.1 특성 벡터 추출

샘플링 주파수가 16 kHz인 입력신호로부터 프레임 길이 40 ms, 50%의 오버랩으로 short time fourier transform (STFT)를 적용하여 약 500 ms 음원에 대한 스펙트로그램을 생성한다. 이 스펙트로그램으로부터 3가지의 특성 벡터를 추출하고 하나의 벡터로 합쳐 최종 특성 벡터를 구성한다. 3가지 특성 벡터는 멜-스펙트로그램 (X_{mel}), 시간적 특성 (X_{delta}, $X_{temporal}$)이다. 그림 1은 특성 벡터의 구성과 각 특성의 추출 과정을 보여준다.

첫 번째 특성 벡터 X_{mel}은 스펙트로그램에 멜-필터를 적용한 멜-스펙트로그램이다. 멜-스펙트로그램을 단독으로 사용한 실험에서는 24 밴드를 사용하고 스펙트로그램 기반의 시간적 특성과 함께 사용한 실험에서는 8 밴드를 사용하였다. k는 시간 축 인덱스로 $0 \leq k \leq 24$ 범위의 정수이다. 이 범위를 texture frame으로 정의한다.

두 번째 특성 벡터인 X_{delta}와 세 번째 특성 벡터 $X_{temporal}$은 스펙트로그램으로부터 주파수 축을 축소한 S_k를 사용하여 생성하였다. S_k는 스펙트럼에서 320개의 주파수 정보를 40개마다 하나의 밴드로 합하여 총 8 밴드의 주파수 정보로 구성하였다. X_{delta}는 $D_k = S_{k+1} - S_k$로 구

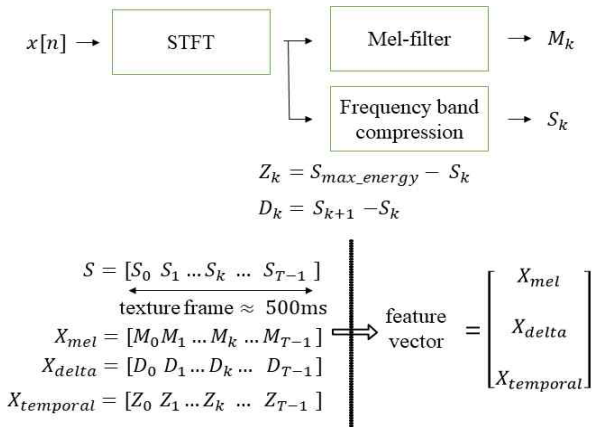

그림 1. 특성 벡터의 구성과 추출 과정
Fig. 1. Construction of the feature vector.

성하였으며, $X_{temporal}$은 $Z_k = S_{max_energy} - S_k$로 구성하였다. $X_{temporal}$에서 사용한 S_{max_energy}는 S_k 중 에너지가 가장 큰 벡터이다. 따라서 특성 벡터 $X_{temporal}$의 요소들은 가장 큰 에너지를 갖는 벡터와의 차이이다. X_{delta}와 $X_{temporal}$은 X_{mel}과 동일한 texture frame 길이를 갖기 위해 마지막 프레임과 동일한 1개의 프레임을 패딩하여 구성하였다. 따라서 각 특성 벡터는 모두 주파수 축 8 밴드, 시간 축 25개의 프레임으로 8 × 25 크기로 구성하였으며, 이를 하나의 특성 벡터로 합쳐 24 × 25 크기의 최종 특성 벡터를 구성하였다.

2.2 네트워크 구조

본 논문에서는 추출한 특성 벡터를 이용한 환경 인지를 위해 CNN을 사용하였다[3]. 본 논문에서 사용한 CNN의 층은 3개 convolutional layer와 2개의 fully connected layer로 구성하였다. 은닉층의 활성화 함수는 rectified linear unit (ReLU)를, 출력층의 활성화 함수는 softmax 를 사용하였다. 가중치와 바이어스의 초기화에는 Xavier 초기화를, 최적화기는 Adam 최적화기를 사용하였으며, mini-batch size는 256, dropout rate는 0.4로 설정하였다[4].

3. 성능 평가

성능 평가에는 TIMIT-DB와 sound-ideas의 Industry & Office DB를 사용하였다. Industry & Office DB로부터 attack, rotation, sawing, tonal을 구성하였고 TIMIT-DB의 test core set를 speech로 사용하였다. 데이터는 1 s 단위로 분리하여 사용하였고 특성 벡터는 약 500 ms 단위로 추출하여 soft voting을 통해 판정하였다. 표 1은 각 클래스의 데이터 수를 보여준다.

표 1. 클래스 별 데이터 개수
Table 1. Dataset specification.

	noise			non-noise	
class	attack	Rota.	sawing	speech	tonal
Num.	685	2,572	536	516	732

표 2는 멜-스펙트로그램만을 사용했을 때의 혼동 행렬이며 표 3은 스펙트로그램의 두 가지 시간적 특성 벡터를 추가했을 때의 혼동 행렬이다. 두 방법 모두 동일한 구조의 CNN을 사용하였다. 표 2의 분류 성능은 평균 87.5%, 표 3의 분류 성능은 평균 91.2%를 기록했다. 표 2, 3에서 1% 이하의 칸은 -로 표시하였다.

표 3을 보면 소음 클래스인 attack, rotation, sawing 사이의 에러가 표 2에 비해 약 1.3-6%p 줄어든 것을 알 수 있다.

표 2. 멜-스펙트로그램 (24 밴드)의 혼동 행렬
Table 2. Confusion matrix of Mel-spectrogram (24 band).

true \ predicted	attack	Rota.	sawing	speech	tonal	recall(%)
attack	89.1	4.2	5.8	-	-	89.1
rotation	-	94.9	2.5	-	1.7	94.9
sawing	4.1	18.5	74.8	2.2	-	74.8
speech	1.2	-	-	98.6	-	98.6
tonal	4.1	13.5	2.0	-	80.1	80.1
precision(%)	88.4	91.5	77.0	96.4	92.4	87.5

표 3. 스펙트로그램 기반의 시간적 특성을 반영한 혼동 행렬
Table 3. Confusion matrix of the spectrogram based feature.

true \ predicted	attack	Rota.	sawing	speech	tonal	recall(%)
attack	94.0	2.3	3.2	-	-	94.0
rotation	-	97.3	1.5	-	-	97.3
sawing	2.8	12.5	84.5	-	-	84.5
speech	-	-	-	99.4	-	99.4
tonal	1.8	13.5	1.6	2.0	81.0	81.0
precision(%)	93.7	93.2	86.0	97.2	96.6	91.2

4. 결론

본 논문은 선택적 노이즈 캔슬링을 위한 멜-스펙트로그램과 스펙트로그램 기반의 시간적 특성 벡터를 이용한 딥 러닝 기반의 환경 인지 기술을 제안하였다. 스펙트로그램의 에너지 변화를 특성 벡터에 반영하기 위해 현재와 다음 프레임 간의 에너지 차이, 가장 큰 에너지를 갖는 프레임과의 차이를 추가했다. 오디오의 대표적인 특성인 멜-스펙트로그램만을 사용한 경우 분류 성능은 87.5%를 보였고 시간적 특성을 반영한 분류 기술에서는 91.2%로 3.7%p 향상된 결과를 보였다.

감사의 글

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.2018-0-01407).

참고문헌

[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, Issue 9, pp. 1037-1053, Nov. 2017.

[2] S.M. Kuo, and D.R. Morgan, "Active noise control : A tutorial review," *Proceedings of the IEEE*, Vol. 87, pp. 943-973, June 1999.

[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature,* pp. 436-444, May 2015.

[4] N. Srivastava, G. Hinton, A. Krizhevesky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, pp. 1929-1958, June 2014.