



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사학위논문

잔향 환경에서의 사운드 이벤트 분류
성능 개선 기법

Performance Enhancement Techniques for Sound Event
Classification in Reverberant Environment

2019년 8월

서울대학교 융합과학기술대학원
융합과학부 디지털정보융합전공
이재준

공학석사학위논문

잔향 환경에서의 사운드 이벤트 분류
성능 개선 기법

Performance Enhancement Techniques for Sound Event
Classification in Reverberant Environment

2019년 8월

서울대학교 융합과학기술대학원
융합과학부 디지털정보융합전공
이재준

잔향 환경에서의 사운드 이벤트 분류 성능 개선 기법

Performance Enhancement Techniques for Sound Event
Classification in Reverberant Environment

지도교수 이 교 구

이 논문을 공학석사 학위논문으로 제출함

2019년 8월

서울대학교 융합과학기술대학원

융합과학부 디지털정보융합전공

이 재 준

이재준의 공학석사 학위 논문을 인준함

2019년 8월

위 원 장: _____

부위원장: _____

위 원: _____

요약

본 연구에서는 잔향 환경에서의 사운드 이벤트 분류시 성능을 개선하는 기법을 제안한다. 사운드 이벤트 분류는 교통 상황, 방범 상황 감지 시스템 등 다양한 응용분야에 활발하게 적용되고 있고 응용분야의 특성상 실제 환경의 잡음과 잔향에 강인한 성능을 갖는 것이 중요한 문제이다. 하지만 이런 잡음과 잔향 환경에서의 사운드 이벤트 분류 성능 저하에 대한 연구는 저조하며 특히 잔향 환경에서의 사운드 이벤트 분류 연구는 전무한 실정이다.

따라서 본 연구에서는 잔향 환경에서 사운드 이벤트 분류 성능이 저하되는 것을 관찰하고 이를 해결하기 위한 개선 기법을 제안한다. 먼저, 잔향 환경을 모델링 하기 위해 원본 데이터셋을 잔향이 존재하는 실제 환경에서 재녹음한 녹음 테스트셋과 공간 임펄스 응답 데이터셋을 이용하여 합성한 합성 테스트셋을 제작하였고, 이를 이용하여 잔향 환경에서 사운드 이벤트 분류 성능이 저하됨을 관찰하였다.

성능 저하에 대한 개선 기법으로 인위적으로 제작한 가상 공간 임펄스 응답을 이용한 데이터 증가 방법과 공간 임펄스 응답을 네트워크에 컨디셔닝하는 기법을 제안하였다. 실험을 통해 제안한 데이터 증가 방법이 잔향 환경에서의 성능을 개선함을 검증하며, 특히 데이터 증가 방법과 컨디셔닝 기법을 함께 사용했을 때 추가적으로 성능이 향상됨을 보인다. 또한 제안한 컨디셔닝 기법이 정확한 공간 임펄스 응답 오디오를 모를 때라도 대략적 잔향 시간 정보를 통해 성능을 향상시킬 수 있음을 보인다.

주요어: 사운드 이벤트 분류, 잔향, 데이터 증가 방법, 컨디셔닝 네트워크

학번: 2017-28919

차 례

| | |
|---------------------------------|-----------|
| 요약 | i |
| 제 1 장 서론 | 6 |
| 1.1 연구 배경 | 6 |
| 1.2 연구 목표 | 9 |
| 제 2 장 배경 이론 및 관련 연구 | 10 |
| 2.1 배경 이론 | 10 |
| 2.1.1 사운드 이벤트 분류 | 10 |
| 2.1.2 딥러닝 연구 | 12 |
| 2.1.3 잔향 및 공간 임펄스 응답 | 16 |
| 2.2 관련 연구 | 19 |
| 2.2.1 사운드 이벤트 분류 연구 | 19 |
| 2.2.2 제안 기법 관련 연구 | 25 |
| 제 3 장 제안 기법 | 28 |
| 3.1 가상 공간 임펄스 응답을 이용한 데이터 증가 방법 | 28 |
| 3.2 공간 임펄스 응답 컨디셔닝 네트워크 | 31 |
| 제 4 장 실험 | 34 |
| 4.1 실험 준비 | 34 |
| 4.1.1 데이터셋 | 34 |
| 4.1.2 테스트셋 제작 방법 | 35 |
| 4.1.3 실험 상세 설정 | 38 |

| | | |
|--------------|-------------------------------------|-----------|
| 4.2 | 실험 결과 및 토론 | 42 |
| 4.2.1 | 간향 환경에서의 사운드 이벤트 분류 성능 저하 | 42 |
| 4.2.2 | 딕컨볼루션 적용 시 성능 및 한계점 | 47 |
| 4.2.3 | 데이터 증가 방법을 이용한 성능 향상 | 49 |
| 4.2.4 | 컨디셔닝 네트워크를 이용한 성능 향상 | 50 |
| 제 5 장 | 결 론 | 58 |
| 5.1 | 연구 의의 | 58 |
| 5.2 | 한계점 | 60 |
| 5.3 | 향후 연구 | 61 |
| | ABSTRACT | 68 |
| | 감사의 글 | 70 |

표 차례

| | | |
|-------|--|----|
| 표 3.1 | 가상 공간 임펄스 응답 생성을 위한 9 개 가상 공간 설정 값 . . . | 29 |
| 표 4.1 | <i>RWCP</i> 사운드 이벤트 클래스 (50 종류) | 35 |
| 표 4.2 | 합성 테스트셋용 공간 임펄스 데이터셋 정보 | 36 |
| 표 4.3 | 녹음 테스트셋 녹음 환경 및 RT60 | 37 |
| 표 4.4 | 베이스라인 모델의 분류 성능 (%) | 42 |
| 표 4.5 | 딕컨볼루션 적용 시 베이스라인 모델의 분류 성능 (%) | 47 |
| 표 4.6 | 데이터 증가 방법을 적용한 모델의 분류 성능 (%) | 49 |
| 표 4.7 | 컨디셔닝 네트워크를 적용한 모델의 분류 성능 (%) | 51 |
| 표 4.8 | Fake 컨디셔닝 실험 결과 | 55 |
| 표 4.9 | 딕컨볼루션 적용 베이스라인 모델과 제안한 컨디셔닝 네트워크 모델의 분류 성능 (%) | 56 |

그림 차례

| | | |
|--------|--|----|
| 그림 1.1 | Citygram 온라인 인터페이스 | 7 |
| 그림 2.1 | 사운드 이벤트 분류 개요도 | 11 |
| 그림 2.2 | 인공 신경망의 구조 예시 | 13 |
| 그림 2.3 | 합성곱 신경망의 구조 예시 | 14 |
| 그림 2.4 | 전이학습 개요도 | 15 |
| 그림 2.5 | 실내 공간에서의 잔향 | 16 |
| 그림 2.6 | 잔향에 의한 사운드 신호의 왜곡 | 17 |
| 그림 2.7 | 이미지 분야의 데이터 증가 방법 예시 | 21 |
| 그림 2.8 | 요소간 변환을 이용한 컨디셔닝 예시 | 26 |
| 그림 3.1 | 가상 공간 임펄스 응답 샘플 오디오 분석 | 30 |
| 그림 3.2 | 컨디셔닝 네트워크 개요도 | 31 |
| 그림 3.3 | 전이학습을 적용한 컨디셔닝 네트워크 개요도 | 32 |
| 그림 4.1 | <i>RWCP</i> 데이터셋 무향실 데이터 수집 예시 (클래스 : <i>coffcan</i>) | 34 |
| 그림 4.2 | Reverberation Time (RT60) 측정 예시 | 37 |
| 그림 4.3 | 트레이닝, 테스트셋 개요도 | 39 |
| 그림 4.4 | 베이스라인 합성곱 신경망 모델 구조 | 40 |
| 그림 4.5 | 제안한 컨디셔닝 네트워크 구조 | 41 |
| 그림 4.6 | 전이학습을 적용한 컨디셔닝 네트워크 구조 | 42 |
| 그림 4.7 | 베이스라인 모델의 RT60에 따른 분류 성능 (%) | 43 |
| 그림 4.8 | 4 개 테스트 공간에서의 샘플 오디오 분석 (클래스 : <i>cherry</i>) | 44 |
| 그림 4.9 | 4 개 테스트 공간에서의 샘플 오디오 분석 (클래스 : <i>ring</i>) | 45 |

| | |
|---|----|
| 그림 4.10 녹음 테스트셋과 합성 테스트셋의 샘플 오디오 비교 (클래스 : <i>cherry</i>) | 46 |
| 그림 4.11 디컨볼루션 샘플 오디오 분석 (클래스 : <i>sandpp1</i>) | 48 |
| 그림 4.12 데이터 증가 방법 적용 모델의 RT60에 따른 분류 성능 (%) . . . | 50 |
| 그림 4.13 컨디셔닝 네트워크 모델의 RT60에 따른 분류 성능 (%) | 51 |
| 그림 4.14 테스트셋에 사용한 공간 임펄스 응답 임베딩 벡터의 t-SNE 표현 | 53 |
| 그림 4.15 Fake 컨디셔닝 실험 개요도 | 54 |

제 1 장 서 론

1.1 연구 배경

사운드 이벤트 분류(sound event classification) 혹은 어쿠스틱 이벤트 분류(acoustic event classification)는 어쿠스틱한 신호를 처리하여 신호 속에 있는 다양한 사운드 이벤트를 청자가 인지할 수 있는 심볼릭한 형태로 표현 및 분류하는 분야이다. [1] 이벤트 단위로 분할 되지 않은 긴 오디오를 입력으로 받아 사운드 이벤트의 종류 분류 및 발생한 시간적 위치까지 찾는 이벤트 감지(event detection)와는 달리 사운드 이벤트 분류는 이미 시간적으로 분할되어 입력한 오디오 클립에 대해 이벤트의 종류만을 분류한다.

이렇게 오디오 클립에 대해 자동으로 심볼릭한 레이블 태깅이 가능한 사운드 이벤트 분류 기술을 이용한 다양한 응용 연구가 증가하고 있다. 자동차 경적소리 분류를 통해 교통량을 측정하거나 [2], 도로 상황에서의 이상 소음 자동 감지시스템을 위해 자동차 충돌음, 타이어 제동 마찰음 등을 분류하는 등 [3,4] 교통 상황에서의 자동 감지시스템에 활용할 수 있다. 또 비명소리, 총소리 등을 분류하여 자동으로 방범 환경을 감지하거나 [5] 포유류의 울음소리를 감지 및 분류하는 연구도 진행되었다. [6]

뉴욕대학교의 CUSP¹에서는 실시간 오디오를 입력받고 특징 추출(feature extraction)을 수행하는 센서 장치를 뉴욕 시내에 배치하여 이 데이터를 이용한 사운드 이벤트 분류를 통해 사이렌소리, 음악소리 등의 도시 내에서 발생하는 소음들의 지역별 분포 등을 분석하여 시민들에게 도시내 소음정보를 제공하는 Citygram [그림

¹Center for Urban Science And Progress, <https://cusp.nyu.edu/>

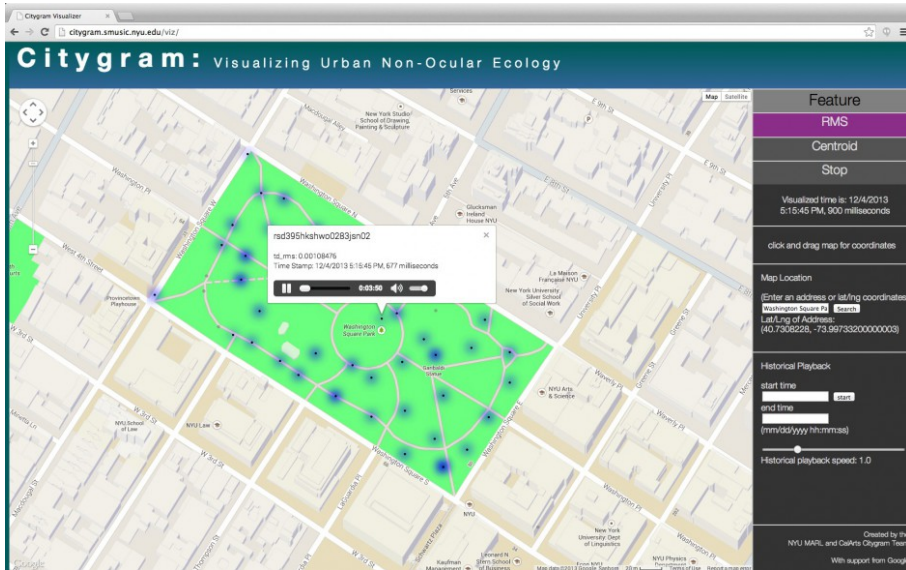


그림 1.1: Citygram 온라인 인터페이스

1.1] 프로젝트를 진행하였다. [7] 이외에도, DCASE² 챌린지에서는 사운드 이벤트 감지 및 분류를 포함한 다양한 종류의 과제에 대한 챌린지가 매년 활발하게 진행되고 있다. [8]

고전적인 사운드 이벤트 분류 연구에서는 도메인 지식을 활용하여 특징을 추출하고 여러 확률모델로 분석을 하는 것이 일반적이었다. 하지만 딥러닝 기술이 컴퓨터 비전 분야에서 두각을 나타낸 이후 이를 넘어 음성인식 [9-11] 등 다양한 분야에 적용되기 시작하면서 최근 사운드 이벤트 분류 분야에서도 합성곱 신경망 등 딥러닝을 이용한 분류 모델이 더 높은 성능을 보이고 있다. [12-14]

사운드 이벤트 분류 기술이 앞선 사전 연구들과 [2-7] 같이 다양한 실제 환경에 응용되기 시작함에 따라, 실제 환경에서 강인한 분류 성능을 갖는 것이 중요한

²Detection and Classification of Acoustic Scenes and Events, <http://www.cs.tut.fi/sgn/arg/dcse2017/>

문제로 자리잡고 있다. 기존의 사운드 이벤트 분류 연구들, 특히 딥러닝을 이용한 연구에서 제안한 방법을 검증하기 위해 통상적으로 타겟으로 삼은 데이터셋을 트레이닝셋과 이와 중복되지 않은 테스트셋으로 나누며 이는 트레이닝셋에 과적합 되지 않은 일반화 가능성을 지닌 테스트 성능을 확인하기 위함이다. [15] 하지만 연구의 목표가 실제 환경에서의 강인한 성능을 갖는 것이라면 테스트셋 또한 이러한 환경에서의 사운드 이벤트 분류 성능을 검증할 수 있도록 구성되어야 할 것이다. 이는 모델이 적용될 실제 환경을 잘 모방하지 못하는 테스트셋으로 검증한 모델은 실제 환경에서의 성능을 보장할 수 없음을 의미한다.

분류해야 할 오디오 자체의 다양성을 제외하고, 테스트셋과 모델이 적용될 실제 환경의 가장 큰 차이는 잡음(noise)과 잔향(reverberation)에 의한 오디오 신호의 왜곡이라고 할 수 있다. 타겟으로 삼은 오디오를 직접 녹음하고 다양한 환경 잡음들을 여러 signal to noise ratio (SNR)로 선형적으로 더하는 합성방법을 이용하여 다양한 잡음이 섞인 타겟 오디오 데이터셋을 제작한 연구도 있었으나 [3, 16], 이 합성방법에는 잔향에 의한 신호 왜곡이 배제되어 있다.

한편 유저들의 협업 방식으로 다양한 오디오에 대한 온라인 데이터베이스를 구축하고 있는 *Freesound*³에서 키워드 ‘field recordings’ 태깅이 되어있는 오디오를 수집하여 실제 환경과 유사한 데이터셋을 구성한 연구도 진행 되었으나 [17, 18] 이러한 데이터셋의 숫자가 제한적이고, 수집된 데이터셋의 특성상 잡음과 잔향을 제어하며 성능을 검증하기 어렵기에 잡음 및 잔향 정도에 따른 성능 저하 경향 등을 파악하기 힘들다는 단점이 있다. 다시 말하여, 잡음과 잔향이 실제환경에서 사운드 이벤트 분류 성능을 저하시키는 두드러지는 요인임에도 이를 실험을 통해 확인하거나 개선하는 기법에 관한 연구는 저조하며 특히 잔향환경에서의 사운드 이벤트 분류 성능에 대한 연구는 전무한 실정이다.

³Freesound, <http://freesound.org>

1.2 연구 목표

본 연구에서는 잔향에 의한 오디오 신호 왜곡과 이에 의한 사운드 이벤트 분류 성능 저하 현상을 확인하고, 이를 개선하기 위한 기법을 제안한다.

구체적으로는, 먼저 잔향에 의해 왜곡된 테스트셋을 제작하여 잔향에 의한 사운드 이벤트 분류 성능 저하를 확인한다. 테스트셋은 잔향이 존재하는 실제 환경에서 테스트셋을 재녹음한 녹음 테스트셋과, 다양한 잔향 환경에서 녹음된 공간 임펄스 응답 데이터셋과 컨볼루션을 통해 합성한 합성 테스트셋을 이용한다.

제작한 테스트셋을 이용하여 잔향 환경에서 사운드 이벤트 분류 성능이 저하됨을 확인하고, 이에 대한 개선 기법을 제시한다. 개선 기법으로는 첫째, 인위적으로 제작한 가상 공간 임펄스 응답을 이용한 데이터 증가 방법과 둘째, 데이터 증가 방법과 함께 공간 임펄스 응답을 네트워크에 컨디셔닝 하는 방법을 제시하며 특히 사전 훈련된 가상 공간 임펄스 응답 간 분류 정보를 네트워크에 전이하는 기법을 제시한다. 실험을 통하여 제안한 기법을 적용했을 때 분류 성능이 개선됨을 보이며 특히, 공간의 정확한 공간 임펄스 응답을 모를 때라도 대략적 잔향 시간을 통해 성능 향상 가능성이 있음을 검증한다.

제 2 장 배경 이론 및 관련 연구

본 장에서는 연구의 배경이 되는 이론과 연구와 직접적인 관련이 있는 연구들을 설명한다. 배경 이론에서는 사운드 이벤트 분류에 대한 설명과 본 연구에서 사용하는 모델의 기반이 되는 딥러닝 연구, 그리고 잔향 및 공간 임펄스 응답과 이로 인한 사운드 이벤트 신호 왜곡에 대해 간략히 설명한다.

관련 연구에서는 사운드 이벤트 분류에 관련한 연구, 특히 잡음에 강인한 성능을 갖기 위한 기술을 제안한 연구와 다양한 기존 데이터 증가 방법을 사운드 이벤트 분류에 적용한 사전 연구에 대해 설명한다. 그 다음 본 연구에서 성능 향상 기법으로 제시한 컨볼루션을 이용한 데이터 증가 방법에 사용하기 위한 가상 공간 임펄스 응답을 이미지 방법(image methods)을 이용해 인위적으로 제작한 연구와, 마지막으로 본 연구에서 마찬가지로 성능 향상 기법으로 제시한 컨디셔닝 기법의 근간이 되는 요소간 변환(feature-wise transformation)을 적용한 사전 연구를 살펴본다.

2.1 배경 이론

2.1.1 사운드 이벤트 분류

사운드 이벤트 분류(sound event classification)는 어쿠스틱 이벤트 분류(acoustic event classification) 혹은 환경 사운드 분류(environmental sound classification)라고도 불리며, [그림 2.1]과 같이 오디오 클립에 대하여 오디오의 의미론적(semantic) 정보에 부합하는 레이블(label)을 자동으로 분류하는 과제이다. 이는 오디오 자체가 가진 콘텐츠(contents)가 오디오에 해당하는 레이블에 부합하는 충분한 정보를 담고 있다는 가정에 근거하며 사람의 감독 없이 자동으로 오디오 레이블링이 가능하다는

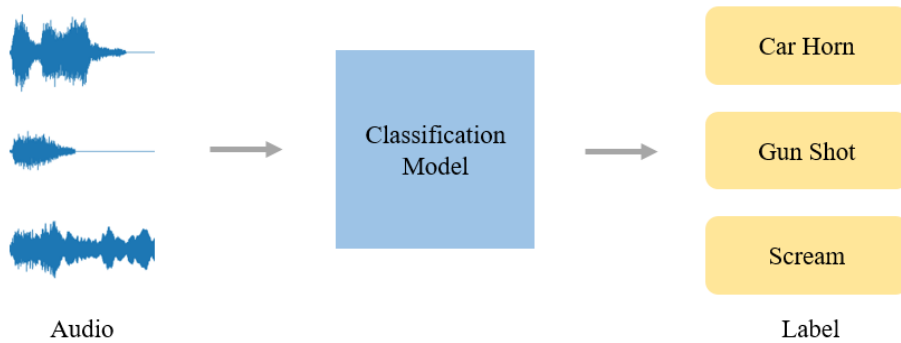


그림 2.1: 사운드 이벤트 분류 개요도

장점으로 다양한 분야에의 적용이 가능하다. 교통 상황에서 발생하는 이상 소음들의 분류를 통해 교통 상황 감지 시스템에 적용할 수 있고 [2-4], 총 소리 및 비명 소리의 분류를 통해 자동 방범 시스템에 적용가능하다. [5] 이런 감지 시스템의 적용 외에도, 일상에서의 소리 자동 감지등을 통해 홈 매니지먼트에 적용하거나 도시 내에서의 특정 소음 관리 등에도 사용할 수 있다. [7]

분할되지 않은 긴 오디오 스트림(stream)에서 이벤트가 존재하는 시간적 위치 까지 까지 찾는 사운드 이벤트 감지(sound event detection) 과제와는 다르게 사운드 이벤트 분류에서는 이미 잘려진 특정 오디오 클립을 입력으로 받아 그 레이블을 분류한다. 일반적으로 오디오 클립 그 자체를 모델의 입력으로 이용하기 보다 오디오로부터 사운드 이벤트 분류에 적합한 특징을 추출하여 입력으로 사용한다. 대표적으로 mel-frequency cepstral coefficients (MFCC) 혹은 멜-스펙트로그램(mel-spectrogram) 등을 이용한다.

기존 사운드 이벤트 분류 연구에서는 고전적인 오디오 분석 기술을 이용하여 사운드 신호에서 스펙트로그램(spectrogram), MFCC 등의 특징을 추출한 뒤 support vector machine (SVM), gaussian mixture model (GMM), hidden markov model

(HMM) 등의 확률모델을 통하여 분석하는 것이 일반적이었다. [19,20] 하지만 최근 딥러닝 기술이 여러 분야에서 두각을 나타냄에 따라 사운드 이벤트 분류 분야에서도 합성곱 신경망 등 딥러닝을 이용한 분류 모델이 더 높은 성능을 보이고 있다. [12-14]

2.1.2 딥러닝 연구

딥러닝(deep learning)은 인공지능(artificial intelligence)에 속하는 개념으로 [15] 인공신경망(artificial neural network)을 기반으로 발전한 기계 학습(machine learning) 알고리즘이다. 뇌의 뉴런과 유사한 정보 전달 구조를 가지며 입력 데이터와 출력 사이에 이어진 층(layer)에서 데이터 분포를 통해 여러 유의미한 정보를 학습하게 되며 이 층이 깊어질수록 딥(deep)하다고 말한다. 최근 딥러닝기술이, 비약적인 도약을 이루었던 이미지 분야 외에도 음성 인식(speech recognition), 자연어 처리(natural language processing), 음악 정보 검색(music information retrieval) 등 다양한 분야에서 높은 성과를 내고있다. [21]

인공 신경망은 [그림 2.2]과 같이 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 나누어지며 일반적으로 은닉층의 층 개수가 2 개 이상일 경우 심층 신경망(deep neural network)이라고 부른다. 각 층마다 뉴런 들이 존재하며 각 층의 뉴런들은 이전 및 다음 층의 뉴런들과 가중치를 통해서 연결되어 있다. 이전 층의 뉴런의 출력 값과 다음층의 뉴런의 연결된 가중치를 선형 결합한 후 비선형 변환(nonlinear transformation)을 통해 그 뉴런에서의 출력 값이 결정되며 이 형태를 뉴런 및 층마다 쌓아 복잡한 데이터를 파악할 수 있도록 구성된다. 비선형 변환은 선형 결합 이후 활성화 함수(activation function)를 이용하며 대표적으로 *Sigmoid*, *Tanh*, *ReLU*, 그리고 *Softmax* 함수가 있다. 특히 *Softmax* 함수는 입력된 데이터에 대하여 정해진 클래스 개수에 대한 각각의 확률값을 출력해주는 함수로, 사운드 이벤트 분류 등 분류 과제에서 많이 사용된다. 출력층의 최종 출력 값과 정답과의 차이를 계산하는

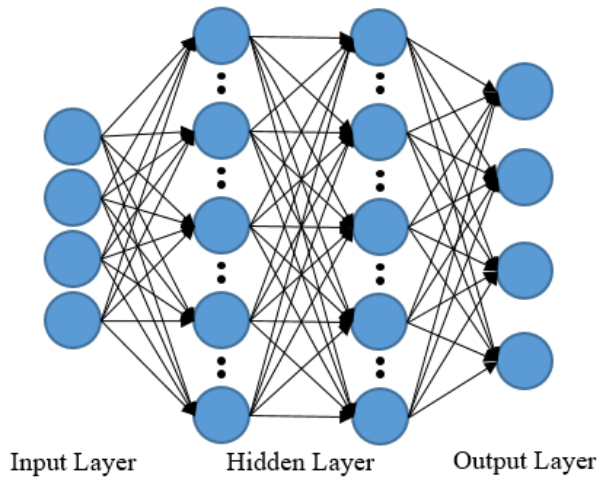


그림 2.2: 인공 신경망의 구조 예시

함수를 손실함수(loss function)라고 하며, 그 종류로는 평균 제곱 오차(mean squared error), 크로스 엔트로피(cross-entropy) 등이 있다.

일반적으로 분류 과제에서는 모델의 출력층에서 나오는 각 클래스에 대한 확률 값과 원 핫 벡터(one-hot vector)로 만든 정답지 사이의 크로스 엔트로피를 계산한다. 이 손실함수를 이용하여 출력층에서 계산된 활성화함수 값부터 역전파(back propagation) 기법을 이용하여 손실함수 값이 줄어드는 방향으로 네트워크 전체의 각 가중치 값이 업데이트된다.

합성곱 신경망(convolutional neural network)의 경우 특히 이미지 분석 분야에서 사람의 인식을 뛰어넘는 높은 성능을 보이는 네트워크로, 이미지의 각 픽셀 값을 연속된 수치로 인식하는 심층 신경망과 달리 필터(filter)를 이용하여 이미지의 각 부분을 인지한다. 이는 사람이 이미지 속의 물체를 인식할 때 전체 이미지의 픽셀 값을 통해서가 아닌 이미지의 부분에 존재하는 픽셀들의 집합으로 물체를 인식한다는 점에서 착안하였다.

합성곱 신경망의 구조는 [그림 2.3]과 같이 일반적으로 합성곱 층(convolution layer), 풀링 층(pooling layer), FC 층(fully connected layer)으로 나누어진다. 합성곱 층은 특정 크기를 갖는 필터(filter) 혹은 커널(kernel)을 일정 간격(stride)에 따라 이동하며 필터 안에 들어오는 이미지의 픽셀 값과 필터 가중치의 원소 곱 후 총합을 계산한다. 설계에 따라 연산 이후 편향(bias) 값이 더해지며 동일한 필터가 이미지 전체에 적용되고 합성곱의 효과로 필터는 학습된 특징이 데이터에 있는지 없는지를 검출할 수 있다. 필터의 크기, 개수 및 이동 간격은 하이퍼 파라미터(hyper parameter)로 설계에 따라 변경할 수 있고, 학습에 의해 필터의 가중치와 편향 값이 업데이트 된다.

풀링 층에서는 세로, 가로 방향의 공간을 줄이는 풀링 연산을 수행한다. 풀링은 영역에서 최댓값 혹은 평균값을 취하는 연산으로 학습해야 할 매개변수가 없으며, 풀링 층은 입력 데이터의 변화에 강인하다는 특징을 가진다. 이렇게 합성곱 층, 풀링 층 등을 반복적으로 거친 후 일반적으로 최종 FC 층으로 연결되며, 이후 인공 신경망과 같이 정답지와 손실함수 계산 및 역전파 기법을 통해 가중치 및 편향 값이

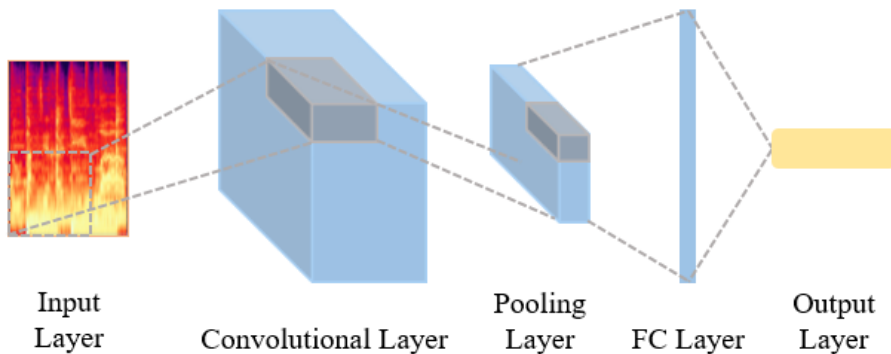


그림 2.3: 합성곱 신경망의 구조 예시

업데이트 된다.

전이학습(transfer learning) 혹은 지식이전(knowledge transfer)은 원본 태스크에서 학습시 얻은 정보를 대상 태스크에 전이(transfer)하는 기술을 말한다. 최근 괄목할 만한 성과를 보이고 있는 딥러닝 기술은 주로 데이터에 의존적이며 이는 훈련 데이터와 테스트 데이터가 같은 분포를 가졌을 때 효율적이다. 따라서 데이터의 분포가 바뀌거나 훈련 데이터의 분포가 테스트 데이터의 분포를 충분히 포함하지 못할 경우 성능을 장담할 수 없다. 전이학습은 이렇게 데이터의 분포가 바뀌거나 부족할 경우 유용하게 적용 가능한 기술이다.

[그림 2.4]과 같이 대용량의 훈련 데이터셋으로 학습한 원본 태스크의 정보를 다양한 대상 태스크에 적용함으로써, 대상 태스크의 데이터가 부족한 경우에도 전이된 정보로 인하여 성능 향상을 기대할 수 있다. 이는 원본 태스크와 대상 태스크의 데이터셋에서 추출해야 하는 정보가 유사하다는 전제가 있을 경우 효율적이다. 또한 데이터셋의 크기로 인한 성능 저하를 극복하기 위한 용도 외에도, 대상 태스크에서 더 좋은 특징 표현(feature representations)을 추출하기 위해 원본 태스크의 특징 표

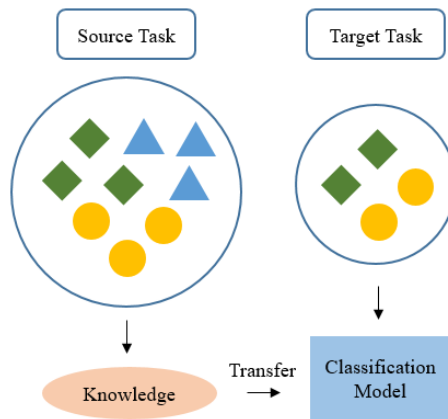


그림 2.4: 전이학습 개요도

현을 추출하는 정보를 전이하기도 한다. [22]

2.1.3 잔향 및 공간 임펄스 응답

사운드 이벤트가 발생하고, 청자 혹은 마이크 등의 수신단에 전달되기까지 사운드 신호는 물체와 벽 등에 반사되며 왜곡된다. 사운드 신호는 이벤트 발생원에서 수신단까지 직진경로 만으로 이동하지않고 [그림 2.5]과 같이 물체와 벽 등에 반사되어 이동하며 반사된 신호는 원본 신호와 유사한 형태로 그 크기가 점차 감소하며 반복되며 나타나는 신호 시퀀스 형태를 가진다. 이 반사된 신호를 잔향(reverberation)이라고 하고 이 잔향에 의해 원본 사운드 이벤트 신호의 어쿠스틱한 특성이 변질되게 된다. 잔향은 실내 공간에만 국한되는 용어가 아닌 원본 사운드 신호가 반사될 수 있는 물체를 지닌 모든 공간에 적용되나, 벽 등이 있는 실내 공간에서 더 두드러지게 나타난다.

이렇게 잔향이 섞여 수신단으로 들어온 사운드 신호는 원본 사운드 신호가 왜

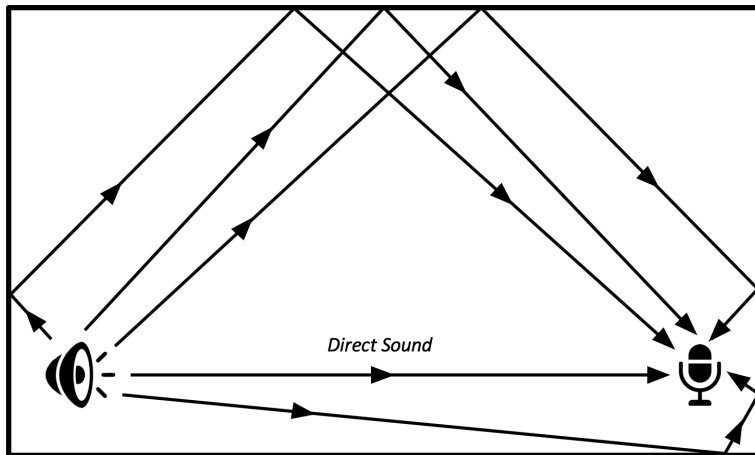


그림 2.5: 실내 공간에서의 잔향

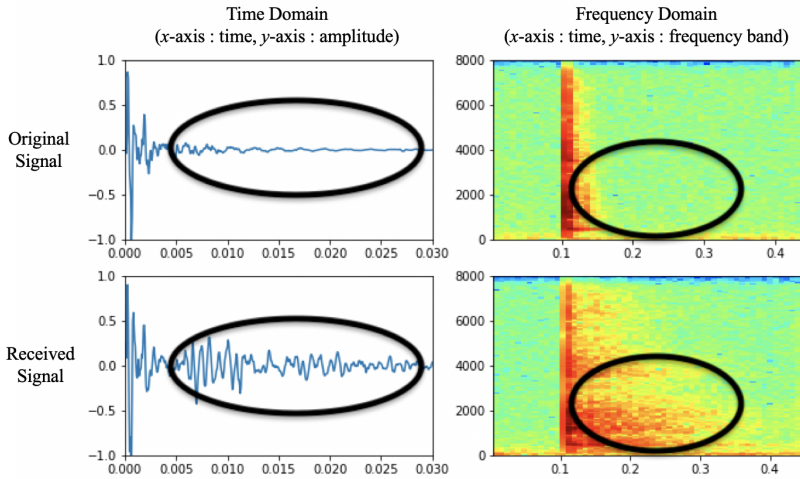


그림 2.6: 잔향에 의한 사운드 신호의 왜곡

곡된 형태로 나타난다. [그림 2.6]의 시간 영역 파형을 보면, 원본 사운드 신호에는 없었던 파형이 수신된 사운드 신호에는 잔향에 의해 발생했음을 볼 수 있다. 마찬가지로 시간축으로 구간에 따라 short time fourier transform (STFT)를 한 주파수 영역 스펙트로그램을 보면, 원본 사운드 신호의 스펙트로그램에는 없었던 시간에 따른 주파수 성분들이 수신된 신호에 존재함을 볼 수 있다.

잔향에 의해 왜곡된 신호는 원본 사운드 신호와 잔향이 존재하는 공간의 공간 임펄스 응답(room impulse response) 간의 컨볼루션 연산(convolution)으로 모델링할 수 있다. [23] 공간 임펄스 응답은 임펄스 사운드에 대한 공간에서의 반응이며, 사운드 이벤트 신호가 원본 발생지로부터 수신단까지 잔향에 의해 왜곡되는 변화를 보여준다고 할 수 있다. 따라서 원본 사운드 신호와 공간에서 수신된 사운드 신호, 공간 임펄스 응답을 각각 $x(t)$, $y(t)$, $h(t)$ 라고 했을 때,

$$y(t) = \sum_{\tau=1}^{T_h} h(\tau)x(t - \tau) = h(t) * x(t) \quad (2.1)$$

로 표기할 수 있으며, 여기서 * 는 컨볼루션 연산을, T_h 는 공간 임펄스 응답의 길이를 의미한다. 또한 시간 영역 신호를 주파수 영역 신호로 변환하는 푸리에 변환 (Fourier transform) 성질에 의해 시간 영역 컨볼루션 연산은 주파수 영역의 곱셈으로 표현 가능하며, 따라서 (식 2.1)은 다음과 같이 표현 가능하다.

$$\mathcal{F}\{h(t) * x(t)\} = k \cdot \mathcal{F}\{h(t)\} \cdot \mathcal{F}\{x(t)\} \quad (2.2)$$

여기서 \cdot 는 곱셈을, \mathcal{F} 는 푸리에 변환을 나타내고 k 는 푸리에 변환의 정규화 (normalization) 상수이다. 따라서 공간에서 수신된 사운드 신호 $y(t)$ 는 다음과 같이 표현 가능하고,

$$y(t) = l \cdot \mathcal{F}^{-1}\{\mathcal{F}\{h(t)\} \cdot \mathcal{F}\{x(t)\}\} \quad (2.3)$$

여기서 \mathcal{F}^{-1} 는 주파수 영역의 신호를 시간 영역 신호로 변환하는 역 푸리에 변환 (inverse Fourier transform) 을 나타내며, l 은 역 푸리에 변환의 정규화 상수이다. (식 2.3)에 따라 공간 임펄스 응답 $h(t)$ 에 의해 왜곡되어 수신된 사운드 신호 $y(t)$ 로부터 원본 사운드 신호 $x(t)$ 를 다음 식과 같이 디컨볼루션 (deconvolution) 연산을 통해 얻을 수 있다.

$$x(t) = m \cdot \mathcal{F}^{-1}\{\mathcal{F}\{y(t)\} / \mathcal{F}\{h(t)\}\} \quad (2.4)$$

여기서 m 은 디컨볼루션 정규화 상수를 의미한다. (식 2.4)을 통해 수신된 사운드 이벤트 신호를 왜곡시킨 정확한 공간 임펄스 응답을 알 수 있을 경우, 디컨볼루션 연산을 통해 원본 사운드 이벤트 신호를 이론적으로 복원 가능함을 알 수 있다. 하지만 실제 잔향 환경에서는 사운드 이벤트 발생 위치에 따라 공간 임펄스 응답 신호가 변하게 되고, 이로 인해 정확한 공간 임펄스 응답 신호가 아닌 신호로 디컨볼루션 시 불완전하게 복원되며 이에 대한 결과는 4.2.2에서 보인다.

2.2 관련 연구

본 절에서는 사운드 이벤트 분류 분야의 사전 연구들을 비롯한 본 연구에서 제안한 기법에 근간이 된 연구에 대해서 설명한다. 먼저 사운드 이벤트 분류 분야에서 주어진 데이터셋 내에서 제안기법을 검증한 여러 연구들, 특히 데이터 증가 방법을 이용하여 분류 성능을 향상시킨 연구를 살펴본 뒤 이어서 환경 잡음에 강인한 성능을 검증하기 위해 환경 잡음을 포함한 테스트셋으로 제안기법을 검증한 연구를 살펴보고 그 한계점을 고찰한다. 다음으로 본 연구에서 트레이닝을 위해 사용한 가상 공간 임펄스 응답을 이미지 방법(image method)을 이용하여 인위적으로 제작하는 알고리즘을 제안한 연구와, 요소간 변환 기법을 사용하여 유의미한 네트워크 컨디셔닝 결과를 보인 연구에 대해 설명한다.

2.2.1 사운드 이벤트 분류 연구

고전적인 사운드 이벤트 분류 연구에서는 zero-crossing rate, 스펙트럼 에너지(spectral energy) 등의 저수준 특징이나 MFCC 등을 특징으로 이용하여 여러 확률 모델을 통해 분석하는 것이 일반적이었다. 대표적으로는 MFCC를 입력 특징으로 이용하여 HMM 확률 모델을 이용하여 사운드 이벤트 분류 및 감지를 수행한 연구

가 있었다. [19]

또 비지도(unsupervised) 학습을 통해 특징을 추출한 뒤 랜덤 포레스트 분류기(random forest classifier)를 이용하여 사운드 이벤트 분류를 수행하여 MFCC 특징을 이용한 기준(baseline) 모델보다 높은 성능을 가짐을 보인 연구도 있었다. [24]

딥러닝 기술이 컴퓨터 비전분야를 넘어서 다양한 분야에서 두드러지는 성능 향상을 보인 이후, 사운드 이벤트 분류 분야에도 딥러닝 기술이 적용되기 시작하였다. *Piczak* [12]은 2 개의 합성곱 층, 풀링 층, FC 층으로 구성된 합성곱 신경망 네트워크를 이용하여 기존대비 높은 성능 향상을 보였다. 데이터셋으로는 *ESC-10*, *ESC-50* 데이터셋 [18]과 *Urbansound8k* 데이터셋 [17]을 이용했고 시간 지연(time delaying), 시간 스트레칭(time stretching), 피치 이동(pitch shifting) 등의 데이터 증가 방법을 적용했다. 데이터셋의 사운드 신호를 로그 스케일의 멜-스펙트로그램으로 전처리한 특징을 네트워크의 입력으로 사용하여 합성곱 신경망이 사운드 이벤트 분류 과제에도 높은 성능을 가짐을 보였다.

이후 딥러닝 기술의 적용이 더욱 활발해 지면서 convolutional restricted boltzmann machine (ConvRBM)을 이용하여 원 오디오 신호로부터 특징 추출을 위한 필터를 비지도 학습을 통해 학습하고 합성곱 신경망을 통해 사운드 이벤트 분류를 수행하는 연구도 진행되었다. [13] 이 연구에서는 특히 기존 연구에서 많이 사용되던 mel filterbank energy (FBE) 특징과 ConvRBM을 이용하여 추출한 특징을 함께 입력으로 사용하여 *ESC* [18] 데이터셋에서 당시 최고 성능을 가짐을 보였다.

딥러닝을 이용한 모델은 훈련셋에서 보지 않은 데이터에 대한 일반화된 성능을 얻기 위하여 많은 양의 훈련 데이터셋을 확보하는 것이 성능에 상당한 영향을 끼치며 특히 심층 신경망 등 모델 용량(capacity)이 큰 여러 층의 깊은 모델에서 이러한

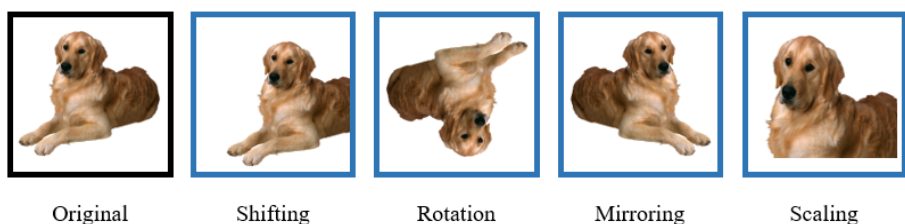


그림 2.7: 이미지 분야의 데이터 증가 방법 예시

경향이 두드러지게 나타난다. 훈련 데이터셋의 크기가 작을 경우, 데이터 증가 방법을 적용하여 인위적으로 데이터를 증가시켜 성능을 향상시킬 수 있다. 데이터 증가의 주요 목적은 레이블링 된 훈련 데이터에 대하여 각 데이터의 레이블과 의미론적으로 부합하는 정보의 손실 없이 데이터를 변형하여 증가시키는 것이다. 이는 같은 원본에서 나온 데이터지만 모델이 보기에 다양성을 갖추어, 증가된 데이터 사이에서 레이블에 해당하는 일반화된 특징만을 학습시키려는 의도가 반영된 것이다. 예를들어 이미지 분야의 경우, [그림 2.7]과 같이 원본이미지의 이동(shifting), 회전(rotation), 반전(mirroring), 스케일링(scaling) 등은 원본 이미지가 담고있는 의미론적 정보는 손실시키지 않으면서 데이터에 다양성을 부여할 수 있다.

오디오 분야에도 마찬가지로 원본 오디오가 가진 의미론적 정보를 손실시키지 않으면서 데이터에 다양성을 부여하는 데이터 증가 방법이 시도되었다. *Piczak* [12], *Salamon* [14]은 사운드 이벤트 분류 과제의 훈련 데이터가 부족할 시 적용 가능한 데이터 증가 방법을 검증하였다. *Piczak*은 클래스 마다 입력 오디오에 무작위로 시간 지연, 시간 스트레칭, 피치 이동 등의 데이터 증가 방법을 이용하여 사운드 이벤트 분류를 수행하였으나, 일부 데이터셋에서는 데이터 증가 효과가 미비하고 훈련 시간이 급증하는 문제가 있다고 기술하였다. [12]

반면 *Salamon* [14]은 사운드 이벤트의 데이터 증가 방법으로 시간 스트레칭, 피

치 이동, 동적 범위 압축(dynamic range compression), 배경 잡음(background noise)을 섞는 방법 그리고 여러 증가 방법을 혼합한 방법을 사용하여 데이터 증가 방법에 의한 성능 향상을 검증하였다. 데이터셋은 *Urbansound8k* [17]를 이용했고, 네트워크 구조로 작은 수용 필드(receptive field)를 가진 합성곱 필터를 이용한 깊은 합성곱 신경망을 이용하여 데이터 증가 방법 뿐만 아닌 깊은 합성곱 신경망의 성능 또한 검증했다.

실험 결과, 데이터 증가 방법과 깊은 합성곱 신경망을 이용한 모델로 벤치마크 데이터셋에서 당시 최고 성능을 기록했고, 데이터 증가 방법을 사전 연구의 모델에 적용했을 때도 분류 성능이 큰 폭으로 향상하여 데이터 증가 방법의 성능 향상 효과를 검증했다. 다만 모든 데이터 증가 방법에서 전체 분류 성능이 향상되었지만, 일부 데이터 증가 방법을 적용했을 때 특정 클래스의 분류 성능이 저하되었다. 예를 들어, 동적 범위 압축 방법과 배경 잡음을 섞는 방법을 사용할 경우 ‘*air_conditioner*’ 클래스를 ‘*engine_idling*’으로 혼동하는 비율이 높아져 이 클래스에 한해서 분류 성능이 매우 저하된다. 이에 대하여 저자는 ‘*air_conditioner*’ 클래스의 어쿠스틱한 특성상 배경 잡음과 유사한 일정한(steady) 사운드로 구성되어 있어 배경 잡음을 섞는 등의 증가 방법은 오히려 해당 클래스의 의미론적인 정보를 가려 성능 저하가 나타난 것으로 추측했다. 이와 더불어 각 클래스마다 성능 향상 비율이 높은 데이터 증가 방법이 달라, 분류하려는 타겟의 사운드 특성에 따라 적절한 데이터 증가 방법이 적용되어야 한다고 기술했다.

두 연구 [12,14]는 데이터 증가 방법이 사운드 이벤트 분류에 사용하는 특징 혹은 모델에 상관없이 원 오디오 단에서 적용 가능하기에 그 성능 향상을 검증한 의의가 있다고 할 수 있다. 하지만 이를 포함한 앞선 연구 모두 [12-14, 19, 24] 데이터셋 내에서 성능을 검증하여 잡음 및 잔향에 강인한지에 대한 성능 고찰이 이루어지지 않았다. 따라서 데이터셋에서의 검증된 성능이 다양한 실제 환경에서도 적용되는지

예측할 수 없다는 한계가 있다.

최근에는 데이터셋 내에서의 성능 검증 이외에 다양한 잡음 환경에서 강인한 성능을 검증하는 연구가 진행되었다. 예를 들어, 실제 도로 환경의 잡음이 반영된 테스트 환경을 모델링을 하고자 잡음을 선형적으로 더한 테스트셋을 사용하여 제안한 모델의 성능을 검증한 연구가 있었다. [3] 이 연구에서는 사운드 신호로부터 신호 크기, 에너지, zero-crossing rate, 스펙트럼 중심(spectral centroid) 등의 저수준 특징과 MFCC, Bark 밴드 에너지 등을 추출한 뒤 bag-of-words 방식을 오디오 도메인에 적용한 audio words 방식으로 보다 고수준의 특징으로 매핑하였다. 이렇게 추출한 고수준 특징을 SVM, k nearest neighbor (kNN) 모델을 이용하여 사운드 이벤트 분류를 수행하여 각 특징별, 모델별 성능을 검증하였다. 분류한 클래스는 도로에서 발생할 수 있는 타이어 제동 마찰음(tire skidding)과 자동차 충돌음(car crash) 이고, 실제 도로 환경을 모델링 하기위해 감지를 수행하기 위한 기기가 설치되는 도로 옆의 장소와 실제 이벤트 음원과 도로 배경 잡음이 발생하는 도로위에서의 장소 사이의 거리를 계산하여 다양한 SNR의 음원을 사용하여 검증하였다. 음원과 일정 SNR로 섞기 위한 잡음은 도로 환경에서 발생할 수 있는 도로 소음을 녹음하여 사용하였다.

이와 유사하게 잡음에 강인한 사운드 이벤트 분류 성능을 갖기 위하여, 사운드 이벤트 음원에 군중소리, 공장소리 등의 환경 잡음을 선형적으로 더하여 테스트셋을 제작한 연구도 진행되었다. [25] 이 연구에서는 stabilized auditory image (SAI) 특징과 spectrogram image features (SIF) 특징을 입력으로 사용하였다. SAI 특징은 사람의 청각 자극 처리에서 영감을 얻은 것으로 실제 청자가 안정적이라고 인지하는 반복적인 사운드의 특징을 부각시킨다. 기본적으로 2 차원 데이터로, 각 필터 구간에 대한 주파수 값이 있는 차원과 피크(peak)값이 발생하는 구간에서 스트로브된(strobed) 시간 축 차원으로 나뉜다. 일반적으로 검출하려는 이벤트는 피크값이

상대적으로 규칙적으로 나타나며, 잡음 등은 주기성이 없다는 점에 착안하여 규칙적인 성문 펄스(glottal pulse)가 발생하는 음성분야에 주로 사용되던 특징이다.

SIF 특징은 사운드 신호에서 생성된 스펙트로그램에 일정 크기의 블락들(blocks)로 구간을 나누어 중심 모멘텀(central momentum)을 구하고 각 블락마다 이 값을 연결하여 특징 벡터를 만든다. 이 특징은 스펙트로그램의 값들을 양자화(quantization)하는 효과가 있어 일반적으로 작은 동적 영역(dynamic region)에 속하게 되는 잡음의 영향을 줄일 수 있다는 장점이 있다. 이 연구에서는 SAI, SIF 두가지 특징에 SVM과 심층신경망을 적용하여 잡음에 강인한 사운드 이벤트 분류 성능을 가짐을 보였다.

SIF 특징의 잡음 강인성 특징을 이용하여, SIF 특징과 함께 합성곱 신경망을 이용하여 더 높은 잡음 강인성을 가진 사운드 이벤트 분류 모델을 제안한 연구 또한 진행되었다. [26] 이 연구에서도 앞선 연구와 마찬가지로 사운드 이벤트 음원에 환경 잡음을 선형적으로 더한 테스트셋을 사용하여 성능을 검증하였다.

이와같이 데이터셋 뿐만 아니라 잡음 환경에서의 강인한 성능을 가진 사운드 이벤트 분류를 위한 기법들을 제안하는 연구가 진행되었으나 [3, 25, 26], 세 연구 모두 잡음에 강인한지 검증하기 위한 테스트셋 제작 시 원본 트레이닝셋에 잡음을 SNR에 따른 일정 비율로 선형적으로 더하는 합성방법을 사용하여, 실제 환경에 대한 충분한 고찰이 이루어졌다고 보기 힘들다. 실제 환경에서는 잡음 뿐 아니라 사운드가 발생하는 공간의 벽, 물체 등에 의한 잔향에 의해 선형적 덧셈과는 다른 복잡한 어쿠스틱한 간섭이 생겨 수신되는 사운드는 잔향에 의한 추가적인 신호 왜곡이 일어났을 가능성이 높다.

2.2.2 제안 기법 관련 연구

Allen [27]은 작은 방의 음향을 효율적으로 나타낼 수 있는 이미지 방법을 제안하였다. 이는 곧 인위적으로 크기를 가정한 가상의 공간에 대하여 공간의 특성을 나타내는 공간 임펄스 응답을 추정하는 방법을 말한다. 구체적으로는 먼저 임펄스 발생 지점과 수신 지점을 가정한 뒤, 임펄스 발생 지점과 수신 지점까지의 거리와 각 주파수 구간 별 속도 차이를 고려하여 수신 지점에서의 시간에 따른 각 주파수 구간 별 파워를 구한다. 여기에 벽에서 반사된 이미지(image) 신호들의 파워를 임펄스 발생 지점 및 수신 지점과 벽 사이의 거리를 고려하여 더해준다.

이미지 방법은 닫힌 공간에서의 음향 속성을 분석하는데에 널리 사용되지만, 실제 환경에서는 벽이 완전히 단단하지 않고 유한한 임피던스를 가지기 때문에 이미지 방법 만으로는 이를 완전히 모델링 할 수 없다. *Allen*은 벽의 임피던스가 음향의 투사각에 비례한다고 가정하였으며, 벽의 반사 계수를 조정 가능한 하이퍼 파라미터로 두어 완전히 단단하지 않은 유한한 임피던스를 가진 공간에서의 공간 임펄스 응답을 추정하였고, 이는 곧 유한한 임피던스를 가진 실제 환경에서의 잔향을 모델링하기 위함이다.

*Allen*이 제안한 가상 공간 임펄스 응답 추정 방법은 직사각형 모양의 공간, 100 ~ 4 kHz 사이의 주파수 영역, 0.7 이상의 반사계수와 임펄스 발생 지점과 수신 지점이 벽에서 너무 가깝지 않은 조건에서만 높은 신뢰도로 공간 임펄스 응답을 모델링 할 수 있다. 이런 한계점이 있음에도 제안한 가상 공간 임펄스 응답 추정 방법은 그 구현에의 용이함으로 다양한 연구에 빈번하게 이용되고 있으며, 본 연구에서도 데이터 증가 방법을 위한 가상 공간 임펄스 응답을 제작하는데에 사용하였다.

Perez [28]는 요소간 변환을 이용하여 네트워크에 특정 정보를 컨디셔닝하는 기법을 제안하였다. 이미지에 특정 사물이 몇 개 있는지 자동으로 세거나 오디오와

비디오가 존재하는 동영상을 분석하는 등의 문제는 이미지 정보만으로는 완전히 이해할 수 없다. 실제 문제들은 이미지와 텍스트, 혹은 이미지와 오디오 등 다양한 소스의 정보들이 집적되어 나타나기 때문에 이를 처리하기 위해 한 소스의 정보를 다른 소스의 문맥으로 이해할 필요가 있다. 이렇게 특정 소스를 다른 소스의 문맥으로 변환하여 정보를 부여하는 기법을 컨디셔닝(conditioning)이라고 하며 효율적이고 효과적인 컨디셔닝 기법에 관한 연구가 활발히 진행되고 있다.

요소간 변환(feature-wise transformations)은 [그림 2.8]과 같이 소스 입력에서 추출한 특징의 각 요소가 보조 입력에서 추출한 특징의 각 요소에 의해 요소간 변환되어 보조 입력에서의 정보를 컨디셔닝하는 기법이다. 대표적인 요소간 변환 방법에는 스케일링(scaling)과 바이어싱(biasing)이 있으며 두 요소가 곱해지는 스케일링 방법은 두 입력간의 관계에 대한 정보를 얻기 유용하고, 두 요소가 더해지는 바이어싱 방법은 서로 의존적이지 않은 독립적인 정보를 얻기 유용하다. [29] 스케일링 방법은 로지스틱(logistic) 함수를 적용하여 출력력을 0 과 1 사이로 제한하여 특정 요소를 0 으로 만드는 방법을 취하기도 한다. 이는 컨디셔닝 정보를 부여하는 보조 입력의 특징이 소스 입력의 특징 중 특정 부분만 사용하게끔 선택하는 효과를 부여한다.

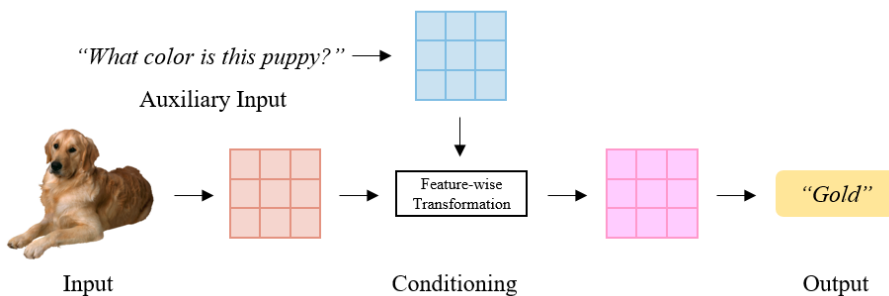


그림 2.8: 요소간 변환을 이용한 컨디셔닝 예시

*Perez*는 이 요소간 변환 컨디셔닝 기법을 사용하여 다양한 도형에 대한 이미지와 도형의 개수, 색깔, 모양 등, 이미지에 페어링된 텍스트가 존재하는 *CLEVR* 데이터셋 [30]에서의 성능을 검증하였다. 구체적으로는 텍스트를 보조 입력으로 받아 인공 신경망의 일종인 순환 신경망(recurrent neural network)을 이용하여 텍스트에서 특징을 추출하고, 합성곱 신경망을 이용하여 이미지로부터 특징을 추출한 뒤 요소간 변환을 수행하는 요소간 선형 변조 (feature-wise linear modulation) 층을 두어 텍스트 정보를 이미지에 컨디셔닝하였다. 특히 요소간 선형 변조 층 이후에 비선형 변환을 수행하는 *ReLU* 활성화 함수를 통해 추출한 특징중 특정 요소만을 사용하였다. 이 연구를 통하여 *Perez*는 당시 *CLEVR* 데이터셋에서 최고 성능을 기록하였으며, 요소간 변환 기법이 서로 다른 도메인의 정보를 컨디셔닝하는데 효과가 있음을 검증하였다. 본 연구에서는 요소간 변환이 보조 입력의 정보를 네트워크에 컨디셔닝하는데 효과가 있다는 점에서 착안하여, 공간 임펄스 응답을 사운드 이벤트 분류 네트워크에 컨디셔닝하는데에 사용하였다.

제 3 장 제안 기법

본 장에서는 잔향 환경에서의 사운드 이벤트 분류 성능을 개선하기 위해 본 연구에서 제안한 기법들에 대해 서술한다. 먼저 가상 공간 임펄스 응답을 이용한 데이터 증가 방법과, 데이터 증가 방법과 함께 사용했을 때 추가적인 성능 향상을 보이는 컨디셔닝 네트워크, 그리고 컨디셔닝 네트워크에 가상 공간 임펄스 응답간 분류 정보를 전이한 네트워크와 그 방법에 대해 설명한다.

3.1 가상 공간 임펄스 응답을 이용한 데이터 증가 방법

사운드 이벤트 분류 분야에서 사용하는 데이터 증가 방법은 *Salamon*의 [14]에서와 같이 시간 스트레칭, 피치 이동, 동적 범위 압축, 배경 잡음을 섞는 방법 등이 있다. 이외에도 오디오를 2 차원의 이미지로 변환한 스펙트로그램에서 특정 구간을 마스킹(masking) 하는 방법등이 있으며 이는 2.2.1에서 서술했듯, 데이터의 의미론적 정보를 왜곡시키지 않으면서 데이터에 다양성을 부여하여 데이터의 수에 의존적인 딥러닝 모델의 성능을 향상시키기 위함이다.

이런 여러 데이터 증가 방법은, 적용되는 방법에 의한 데이터의 변화가 마치 노이즈처럼 작용하여 딥러닝 모델로 하여금 이 노이즈와 관련 없는 의미론적 정보에만 집중하게하며 이는 곧 노이즈에 강인한 성능을 가지게 함을 의미한다. 따라서 이러한 데이터 증가 방법은 주로 실제 환경에서의 강인한 성능을 갖기 위한 고찰에서 이루어졌다고 보다 데이터의 의미론적인 정보에 더 집중하기 위한 측면이 강하다고 할 수 있다.

| Room No. | Room Width (m) | Room Height (m) | Room Depth (m) | Reverberation Time (s) |
|----------|----------------|-----------------|----------------|------------------------|
| 1 | 3 | 3 | 2 | 0.15 |
| 2 | 3 | 3 | 2 | 0.3 |
| 3 | 7 | 7 | 3 | 0.2 |
| 4 | 7 | 7 | 3 | 0.35 |
| 5 | 15 | 15 | 5 | 0.25 |
| 6 | 15 | 15 | 5 | 0.35 |
| 7 | 15 | 15 | 5 | 0.55 |
| 8 | 30 | 30 | 10 | 0.5 |
| 9 | 30 | 30 | 10 | 0.7 |

표 3.1: 가상 공간 임펄스 응답 생성을 위한 9 개 가상 공간 설정 값

본 연구에서는 이와 유사하게 데이터의 의미론적인 정보에만 집중하는 한편, 실제 환경에서 사운드 이벤트 분류 성능 저하의 요인이 되는 잔향에 강인한 성능을 갖게하기 위해 공간 임펄스 응답을 이용한 데이터 증가 방법을 제시한다. 공간에서의 잔향은 공간에서의 임펄스 응답과 컨볼루션 연산을 통해 모델링 할 수 있으며(식 2.1), 이를 이용한 데이터 증가 방법을 통해 모델이 잔향에 강인하게 하고, 잔향을 제외한 데이터의 의미론적 정보에만 집중하게 한다.

데이터 증가 방법을 위해 사용한 가상 공간 임펄스 응답은 2.2.2에서 설명한 *Allen* [27]의 이미지 방법을 이용하여 제작하였다. 이를 위한 구현 코드는 공개된 MATLAB 코드¹를 이용하였다. 공개된 코드는 가상 공간 임펄스 응답을 구하기 위한 공간의 가로(width), 높이(height), 세로(depth) 길이와 잔향 시간(reverberation time) 및 임펄스 발생 위치와 수신하는 마이크 위치를 설정할 수 있다. 다양한 실제 공간의 크기와 잔향 시간을 반영하는 총 9 개 종류의 가상 공간을 설정하였으며 각 방의 설

¹MATLAB code, <https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator>

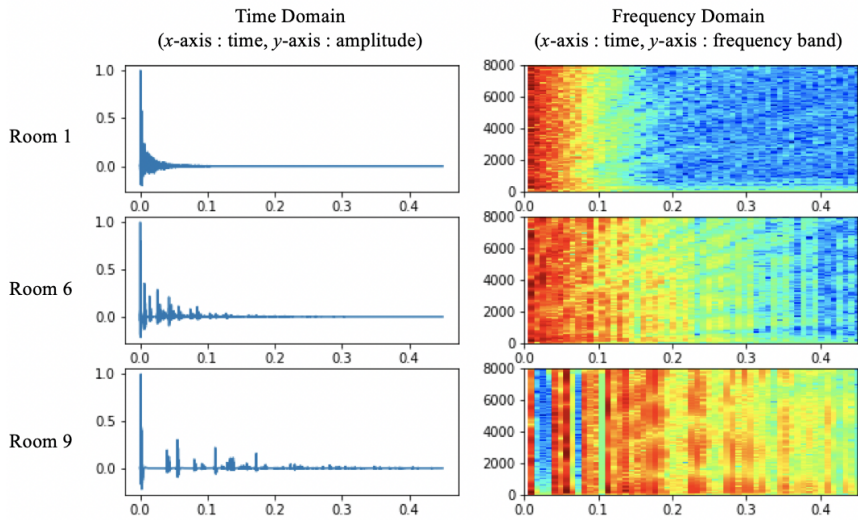


그림 3.1: 가상 공간 임펄스 응답 샘플 오디오 분석

정 값은 [표 3.1]와 같다. 임펄스 발생 위치는 각 가상 공간의 중앙으로 설정하였으며 수신하는 마이크 위치는 각 가상 공간 내에 무작위 지점을 골라 공간마다 가상 공간 임펄스 응답 100 개씩을 생성하였다. 이는 원본 데이터셋의 이벤트에 최대 900 개의 잔향 특성이 반영될 수 있음을 의미하며 학습시 생성한 가상 공간 임펄스 응답 중 무작위로 선택하여 컨볼루션 연산을 통해 데이터를 증가시켰다.

생성한 가상 공간 임펄스 응답의 예시는 [그림 3.1]과 같으며 [표 3.1] 중 Room 1, 6, 9 를 골라 각 가상 공간 내에서 무작위 샘플을 선택하여 시간 축 신호와 주파수 축 스펙트로그램을 분석하였다. 가상 공간의 설정 크기 및 잔향 시간이 증가함에 따라 (Room 1, 6, 9) 시간 축 신호에 존재하는 잔향에 의한 추가 신호가 더 넓은 시간축에 걸쳐 나타남을 볼 수 있고, 이는 주파수 축 스펙트로그램을 통해서도 확인할 수 있다.

3.2 공간 임펄스 응답 컨디셔닝 네트워크

제안한 컨디셔닝 네트워크는 3.1에서 제안한 데이터 증가 방법과 함께 적용 가능한 구조로서, 잔향이 존재하는 실제 공간에서 사운드 이벤트 분류를 수행하는 하드웨어가 해당 공간의 임펄스 응답을 손쉽게 수집할 수 있다는 점에서 착안하였다. 실제 공간에서의 사운드 이벤트들은 공간의 잔향에 의해 왜곡되며 잔향은 공간 및 수신하는 하드웨어의 위치에 따라 달라진다. 공간 임펄스 응답은 이러한 공간의 잔향을 나타내는 해당 공간의 특징이 되며, 이러한 공간 임펄스 응답에서 유의한 정보를 추출함으로써 사운드 이벤트 분류를 수행함에 있어 특정 공간 잔향에 의한 성능 저하를 개선할 수 있을 것이라 가정하였다. 따라서 공간 임펄스 응답에서 추출한 정보를 사운드 이벤트 분류 네트워크에 컨디셔닝하는 네트워크를 설계하였고 그 개요도는 [그림 3.2]과 같다.

소스 입력에는 잔향에 의해 왜곡된 사운드 이벤트 오디오가 들어가며, 보조 입력으로 소스 입력의 사운드 이벤트 오디오를 왜곡시킨 공간 잔향을 나타내는 공간 임펄스 응답이 들어간다. 컨디셔닝 방법은 2.2.2에서 설명한 요소간 변환을 이용하였다. 자세한 네트워크 구조는 4.1.3에서 서술한다. 2.2.2에서 설명한 Perez [28]에서는 이미지와 텍스트라는 서로 다른 도메인의 정보를 컨디셔닝하기 위해 각각 다른 임베딩 블록을 사용하였지만 본 연구에서 제안하는 컨디셔닝 방법은 잔향에 의해

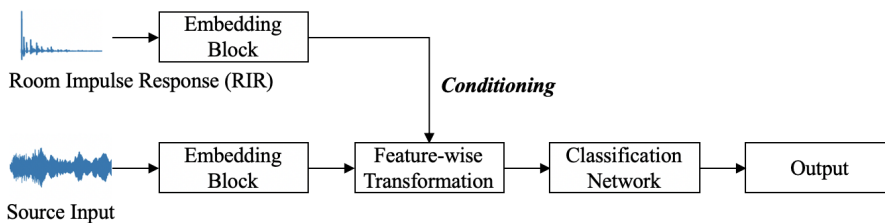


그림 3.2: 컨디셔닝 네트워크 개요도

왜곡된 사운드 이벤트와 공간 임펄스 응답이라는 서로 같은 오디오 도메인의 정보를 컨디셔닝한다. 따라서 소스 입력인 사운드 이벤트 오디오에서 정보를 추출하는 임베딩 블록과 보조 입력인 공간 임펄스 응답 오디오에서 정보를 추출하는 임베딩 블록의 구조를 동일하게 설정하였다.

제안한 컨디셔닝 네트워크는 소스 입력이 이미 잔향에 의해 왜곡되어 있다고 가정하고 이 잔향 특성을 나타내는 공간 임펄스 응답을 보조로 입력한다. 따라서 다양한 가상 공간 임펄스 응답을 이용한 데이터 증가 방법과 함께 적용하였다. 단, 학습시 데이터 증가 방법이 적용된 데이터들과 더불어 왜곡되지 않은 원본 소스 데이터들을 함께 학습하였는데 이 데이터들은 컨볼루션 연산 후에도 동일한 값이 나오는 임펄스 신호를 보조로 입력하였다. 이는 원본 소스 데이터들이 잔향에 의한 왜곡이 없는 이상적인 공간에서 수집된 데이터들이라고 가정한 것으로 볼 수 있다.

본 연구에서는 제안한 컨디셔닝 네트워크의 성능을 향상시키기 위하여 2.1.2에서 설명한 전이학습을 추가로 사용하였다. 먼저 3.1에서 생성한 9 개 공간 설정의 가상 공간 임펄스 응답과 잔향에 의한 왜곡이 없는 이상적 공간의 공간 임펄스 응답

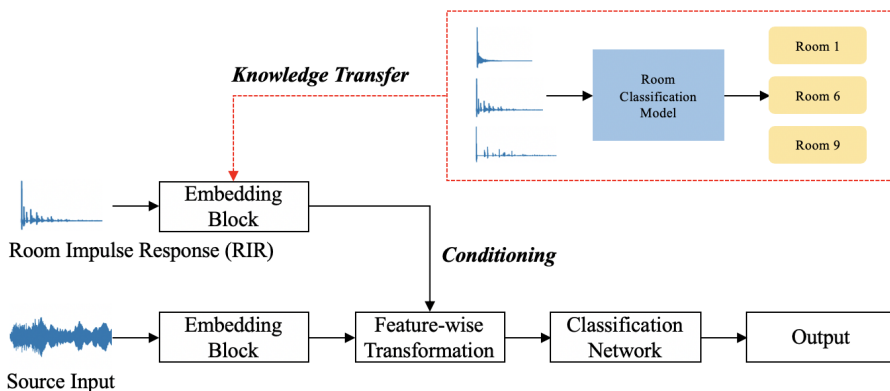


그림 3.3: 전이학습을 적용한 컨디셔닝 네트워크 개요도

인 임펄스 신호를 포함한 10 개 방을 레이블로 두고 이를 분류하는 네트워크를 학습하였다. 이후 학습된 공간 임펄스 응답 분류 네트워크의 임베딩 블록의 가중치들을 전이하여 제안한 컨디셔닝 네트워크의 임베딩 블록의 가중치 학습 초기값으로 설정한다. 이런 전이학습을 적용함은, 컨디셔닝 네트워크의 임베딩 블록을 통해 공간 임펄스 응답에서 정보를 추출함에 있어 잔향의 공간별 특징을 구분짓는 두드러지는 차이에 더 집중하게하기 위한 의도가 반영된 것이다. 전이학습이 적용된 컨디셔닝 네트워크의 개요도는 [그림 3.3]과 같다.

제 4 장 실 험

본 장에서는 본 연구에서 제안한 기법을 검증하는 실험에 대해 설명한다. 먼저 실험 준비에서는 사용한 데이터셋, 테스트셋 제작 방법, 실험 상세 설정에 대해 설명한다. 실험 결과 및 토론에서는 잔향에 의해 사운드 이벤트 분류 성능이 저하되는 현상을 실험을 통해 확인하고 2.1.3에서 설명한 디컨볼루션을 적용했을 때의 성능과 그 한계점에 대해 고찰한다. 이후 제안한 기법이 잔향 환경에서의 성능 저하 현상을 개선하고, 디컨볼루션 적용 시의 한계점을 극복함을 보인다.

4.1 실험 준비

4.1.1 데이터셋

사운드 이벤트 분류 과제를 위한 데이터셋은 Real World Computing Partnership (RWCP) [31] 데이터셋을 사용했다. RWCP 데이터셋은 [그림 4.1]과 같이 무향실에서 녹음한 데이터셋으로 잔향 등 공간 특성에 의한 원본 사운드 신호의 왜곡에서 자유롭다.

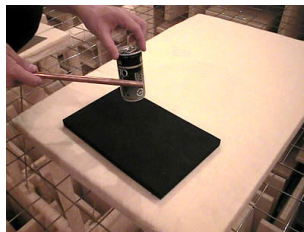


그림 4.1: RWCP 데이터셋 무향실 데이터 수집 예시 (클래스 : *coffcan*)

| | | | | |
|----------|----------|----------|----------|----------|
| aircap | bank | bells5 | book1 | bottle1 |
| bowl | buzzer | candybwl | cap1 | case1 |
| cherry1 | clap1 | clock1 | coffcan | coffmill |
| coin1 | crumple | cup1 | cymbals | dice1 |
| doorlock | drum | dryer | file | horn |
| kara | magno1 | maracas | mechbell | metal05 |
| pan | particl1 | phone1 | pipong | pump |
| punch | ring | sandpp1 | saw1 | shaver |
| snap | spray | stapler | sticks | string |
| teak1 | tear | trashbox | whistle1 | wood1 |

표 4.1: RWCP 사운드 이벤트 클래스 (50 종류)

클래스는 나무 부딪히는 소리, 쇠 부딪히는 소리, 벨 소리 등 90 가지 이상의 다양한 소리이며 클래스 당 약 100 개의 샘플을 녹음하였고 40 ~ 50 db의 높은 SNR 을 가진다. 이 중 사전 연구 [25,26]에서 사용했던 50 개 클래스를 동일하게 사용했으며, 클래스마다 무작위로 선별한 60 개의 샘플을 트레이닝셋으로, 20 개의 샘플을 테스트셋으로 사용하였다. 클래스는 [표 4.1]와 같다.

4.1.2 테스트셋 제작 방법

본 연구에서는 잔향에 의한 왜곡이 반영된 테스트셋을 만들기 위해 합성 테스트셋과 녹음 테스트셋을 사용했다. 합성 테스트셋은 잔향이 존재하는 공간에서 녹음된 4 종류의 공간 임펄스 응답 데이터셋 내 7 개 공간의 공간 임펄스 응답을 이용하여 생성하였고, 녹음 테스트셋은 잔향이 존재하는 2 개 장소를 선정 후 4.1.1의 데이터셋을 해당 공간에서 재녹음하여 생성하였다.

| Dataset Name | Room Name | Number of Data | Width × Height × Depth (m) | RT60 (s) |
|--------------|-----------|----------------|----------------------------|----------|
| AIR | Booth | 12 | 3.00 × 1.80 × 2.20 | 0.267 |
| AIR | Lecture | 24 | 10.8 × 10.9 × 3.15 | 0.682 |
| AIR | Office | 12 | 5.00 × 6.40 × 2.90 | 0.385 |
| AIR | Stairway | 78 | 5.20 × 7.00 × - | 0.771 |
| MARDY | - | 73 | - | 0.550 |
| QMUL | Classroom | 130 | 7.50 × 9.00 × 3.50 | 1.341 |
| WDR | CR7 | 360 | 7.26 × 7.33 × 2.90 | 0.294 |

표 4.2: 합성 테스트셋용 공간 임펄스 데이터셋 정보

먼저 합성 테스트셋에 사용한 공간 임펄스 응답 데이터셋은 Aachen impulse response (*AIR*) [32], multichannel acoustic reverberation database at York (*MARDY*) [33], Queen Mary University of London (*QMUL*) [34], Westdeutscher Rundfunk (*WDR*) [35] 이며 데이터셋 내에서 사용한 공간별 상세는 [표 4.2]와 같다. 공간별로 테스트셋을 합성하여 총 7 개의 합성 테스트셋을 생성하였으며 합성시 이벤트 음원에 해당 공간 내 공간 임펄스 응답을 무작위로 선택하여 (식 2.1)의 컨볼루션 연산을 통해 합성하였다.

[표 4.2]의 RT60은 사운드 이벤트 피크값에서 60 db 만큼 떨어지기까지 걸리는 시간으로, 일반적으로 공간의 잔향 특성을 나타내는 대표값으로 사용된다. RT60의 측정은 [그림 4.2]과 같이 공간 임펄스 응답 오디오의 피크값에서 5 db 감소된 시점에서 60 db가 더 떨어지기까지의 시간을 구하지만, 일반적으로 각각 20 db, 30 db가 떨어지는데 걸리는 시간인 RT20, RT30을 구한 뒤에 외삽(extrapolation)하여 각각 3, 2 를 곱하여 RT60을 구하는 방법을 사용한다. 본 연구에서는 RT20을 구한뒤 외삽하여 RT60을 구하는 방법을 사용하였으며, 사운드 이벤트 분류 모델의 입력 주파수

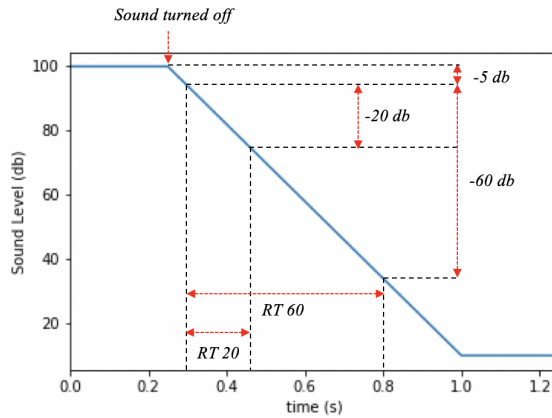


그림 4.2: Reverberation Time (RT60) 측정 예시

범위인 125 ~ 7500 Hz 사이의 주파수 별 RT60을 구하여 평균을 내어 공간 임펄스 응답 오디오 1 개에 대한 RT60으로 취하였다. 공간 내의 모든 공간 임펄스 응답에 대하여 동일한 방식으로 RT60을 구한뒤 평균값을 취해 해당 공간의 RT60을 구하였다.

녹음 테스트셋은 [표 4.3]와 같이 잔향이 존재하는 실제 환경인 복도, 회의실 두 가지 장소를 선정하여 데이터셋 전체를 재녹음하였다. 복도는 두 면이 벽으로 막혀



| Room Name | Recording Environment | RT60 (s) |
|------------------|---|----------|
| Recorded Hallway |  | 0.206 |
| Recorded Room |  | 0.216 |

표 4.3: 녹음 테스트셋 녹음 환경 및 RT60

있고 다른 두 면은 넓게 뚫려있는 실내 공간이고, 회의실은 네 면이 벽으로 막혀있는 실내 공간이다. 녹음용 마이크는 ‘RODE MiCon-5’, 음원 재생용 스피커는 ‘Genelic 8040 APM’를 사용하였다. RT60은 각 녹음 장소의 데이터셋을 녹음한 지점에서 임펄스 응답을 재생한 뒤 녹음하여 합성 테스트셋을 위한 공간 임펄스 응답의 RT60과 동일한 방법으로 측정하였다.

4.1.3 실험 상세 설정

RWCP 데이터셋의 대부분의 소스 음원들은 1 초 미만의 길이를 가진다. 딥러닝 모델의 입력 사이즈를 맞추기 위해 소스 음원의 1 초 만을 사용하였으며, 1 초 보다 짧은 음원의 경우 제로 패딩(zero padding)하였다. 이후 1 초로 길이가 제한된 오디오를 오디오의 최대 값으로 나누는 정규화(normalization)를 적용하였고 이에대한 수식은 다음과 같다.

$$normalized_x(i) = x(i) / \max(\forall |x(i)|) \quad (4.1)$$

(식 4.1)에 따라 정규화 한 뒤 멜-스펙트로그램으로 변환하였다. 멜-스펙트로그램은 시간축으로 윈도우 길이(window length)에 따라 STFT를 수행한 뒤 주파수 빈(bin)의 크기를 선형이 아닌 사람의 청각 지각과 유사하게 로그 스케일로 변형한 오디오 특징이다. 샘플링 율(sampling rate)은 16 kHz, 프레임(frame) 수는 100, 주파수 밴드(band) 수는 64, 윈도우 길이는 32 ms, 홉(hop) 길이는 10 ms, 주파수 축의 최저 주파수는 125 Hz, 최고 주파수는 7500 Hz로 제한하였다. 합성 테스트셋과 녹음 테스트셋도 이와 같은 전처리 과정을 적용하였다.

4.1.1에서 서술한 대로, 50 개 클래스는 각각 80 개의 데이터를 갖는다. 이중 60

개를 트레이닝, 20 개를 테스트셋으로 사용하였으며 트레이닝셋의 20%를 검증셋(validation set)으로 사용하였다. 따라서 데이터 증가 방법이 적용되지 않은 경우 클래스당 48 개 데이터로 트레이닝, 12 개 데이터로 검증, 이후 20 개 데이터로 테스트가 이루어지며 이는 전체 클래스로 보면 에폭(epoch)마다 2,400, 600, 1,000 개 데이터가 각각 트레이닝, 검증, 테스트로 사용된다.

제안한 데이터 증가 방법을 적용한 실험의 경우 기존 2,400 개 트레이닝셋, 600 개 검증셋에 데이터 증가 방법이 적용된 2,400 개 트레이닝셋, 600 개 검증셋을 추가로 학습한다. 이 추가된 데이터들은 기존 훈련시키는 트레이닝셋, 검증셋과 같은 의미론적(semantic) 정보를 가지는 오디오에 3.1의 데이터 증가 방법을 적용한 것이다. 증가된 데이터들은 에폭마다 트레이닝 및 검증셋의 각각의 오디오에 9 개의 가상 공간 설정 중 무작위로 선택하여 컨볼루션 하였었고, 따라서 데이터 증가 방법이 적용된 실험의 경우 4,800 개 트레이닝셋, 1,200 개 검증셋을 사용한다.

또한 테스트셋은 4.1.2에서 설명한 방법대로 합성 테스트셋 7 개와 녹음 테스트셋 2 개를 사용하였고, 잔향에 의한 왜곡을 반영하지 않은 Clean 테스트셋까지 총 10 개 테스트셋을 사용하였다. 트레이닝셋과 테스트셋에 대한 개요는 [그림 4.3]과 같다. 또한 특정 테스트셋에 과적합(overfit)된 결과를 피하기 위해 서로 독립된 10



그림 4.3: 트레이닝, 테스트셋 개요도

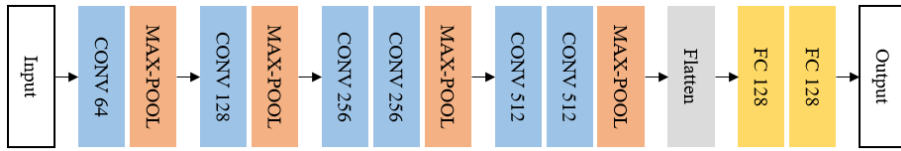


그림 4.4: 베이스라인 합성곱 신경망 모델 구조

쌍의 트레이닝셋, 테스트셋 쌍을 만들어 모든 실험에서 10 개 쌍에 대한 테스트 결과를 구해 평균값을 취하였다. 따라서 10 개의 트레이닝셋, 테스트셋 쌍에 훈련된 모델이 각각 존재하며 이 모델은 서로 독립적인 학습과정 및 성능을 가진다.

분류 모델의 네트워크는, 먼저 공간 임펄스 응답을 컨디셔닝하지 않은 베이스라인 모델은 [그림 4.4]과 같은 합성곱 신경망 구조를 사용하였다. ‘CONV’는 합성곱 층을 의미하며 합성곱 층의 숫자는 필터 수를 의미한다. ‘MAX-POOL’은 최대값-풀링 층을, ‘Flatten’은 다차원의 구조를 갖는 이전 층의 출력을 일차원의 형태로 변환시키는 층이며 ‘FC 128’은 128 개의 차원을 갖는 FC 층을 의미한다. 이후 출력 (output)층에서 각 클래스에 대한 확률값을 출력한다. 매 합성곱 층 이후에 배치 정규화(batch normalization)를 적용하고 *ReLU* 활성화 함수를 통과시켰다.

3.2에서 서술한 제안한 컨디셔닝 네트워크의 구조는 [그림 4.5]과 같다. 구체적으로는, 베이스라인 합성곱 신경망 모델의 FC 층 전까지의 구조와 동일한 구조의 room impulse response (RIR) 임베딩 블록(embedding block)을 추가하였다. 이 임베딩 블록을 통해 보조 입력으로 준 공간 임펄스 응답에서 정보를 추출한 뒤 ‘Scaling’, ‘Biasing’ 층에서 베이스라인 합성곱 신경망 모델에 컨디셔닝한다. ‘Scaling’층에서 선형 곱셈, ‘Biasing’층에서 선형 덧셈을 수행하며 각 층에서 선형 변환 직후 비선형 변환인 *ReLU* 활성화 함수를 통과시킨다.

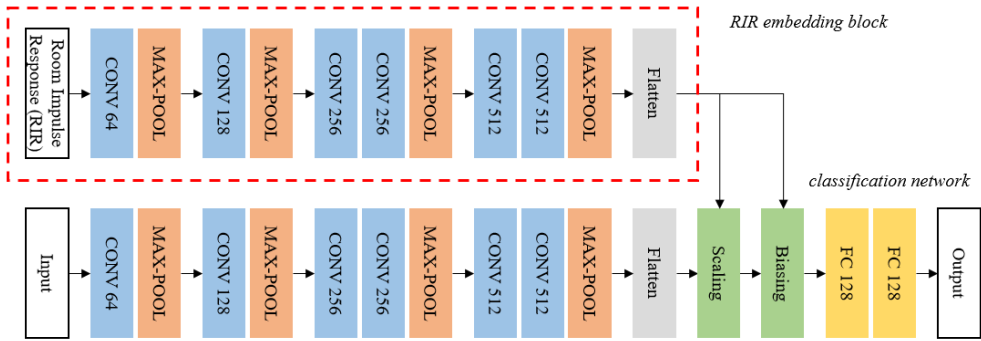


그림 4.5: 제안한 컨디셔닝 네트워크 구조

트레이닝시 3.1에서 생성한 가상 공간 임펄스 응답과 데이터셋 사운드 이벤트를 컨볼루션한 오디오를 소스 입력으로 주고, 컨볼루션에 사용한 가상 공간 임펄스 응답을 RIR 임베딩 블록의 보조 입력으로 준다. 테스트시에는 트레이닝셋에 사용한 가상 공간 임펄스 응답이 아닌 4.1.2에서 서술한, 데이터셋으로 제작한 합성 테스트셋의 공간 임펄스 응답 혹은 직접 녹음한 녹음 테스트셋의 공간 임펄스 응답을 보조 입력으로 주었다.

또한 전이학습을 적용한 제안한 컨디셔닝 네트워크의 구조는 [그림 4.6]과 같으며, 이는 생성한 10 개 종류의 가상 공간 임펄스 응답을 분류하는 네트워크를 훈련시킨 뒤, Flatten 층 이전까지의 가중치들을 제안한 컨디셔닝 네트워크의 RIR 임베딩 블록으로 전이한다. 전이한 방법은 가져온 가중치들로 임베딩 블록의 초기값을 설정하고, 이후 재 훈련하였다.

구현은 *Tensorflow* [36] 백엔드(backend)의 *Keras* [37]를 이용했다. 최대 에폭값을 100 으로 제한한 뒤 검증셋의 검증 손실값(validation loss)이 최저일 때의 모델을 선택하였다. 배치사이즈는 64, 손실 함수는 크로스엔트로피를 사용했고, 최적화 함수(optimize function)는 0.0001 학습률(learning rate)의 *Adam* [38]을 사용하였다.

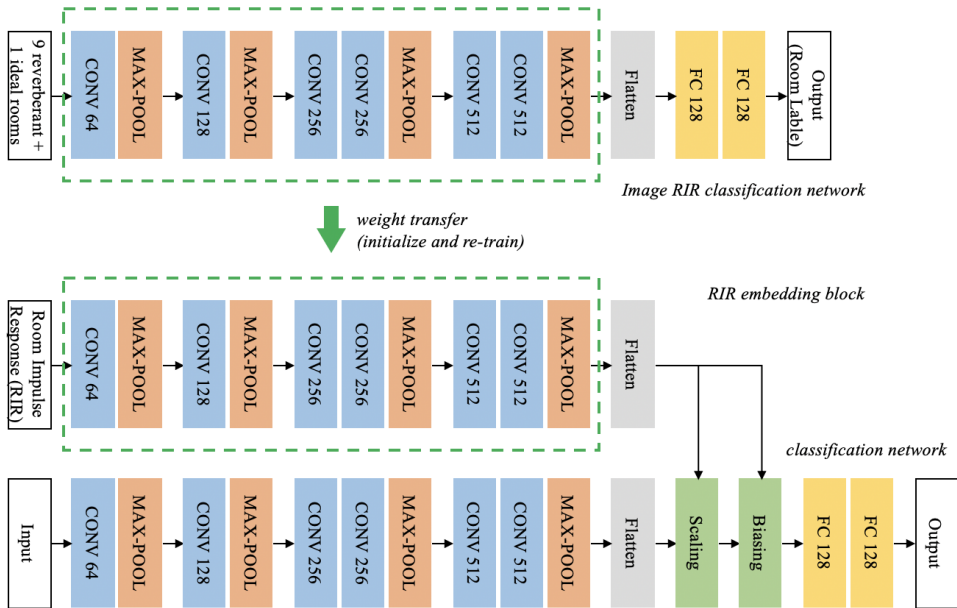


그림 4.6: 전이학습을 적용한 컨디셔닝 네트워크 구조

4.2 실험 결과 및 토론

4.2.1 잔향 환경에서의 사운드 이벤트 분류 성능 저하

[표 4.4]는 데이터셋만으로 훈련시킨 베이스라인 모델의 합성 테스트셋과 녹음 테스트셋에 대한 사운드 이벤트 분류 성능이다. 분류 모델은 [그림 4.4]의 베이스라

| Model | Clean | Recorded Hallway | Recorded Room | AIR Booth | WDR CR7 | AIR Office | MARDY | AIR Lecture | AIR Stairway | QMUL Classroom |
|----------|-------|------------------|---------------|-----------|---------|------------|-------|-------------|--------------|----------------|
| Baseline | 99.32 | 56.56 | 57.74 | 88.11 | 75.66 | 61.28 | 53.27 | 47.10 | 46.89 | 32.25 |

표 4.4: 베이스라인 모델의 분류 성능 (%)

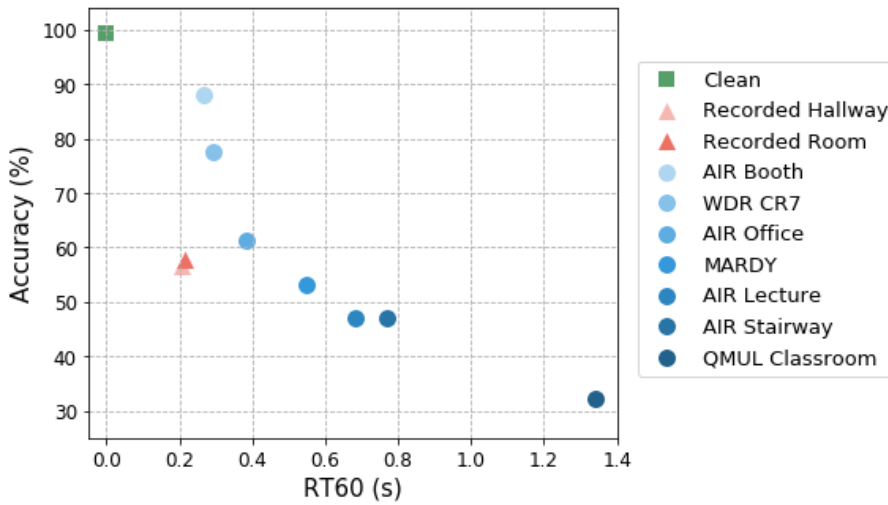


그림 4.7: 베이스라인 모델의 RT60에 따른 분류 성능 (%)

인 모델을 사용하였고 분류 성능은 10 개의 트레이닝셋, 테스트셋 쌍에 대한 10 번의 실험의 평균 분류 정확도(%)를 계산하였다. ‘Clean’은 원본 데이터셋의 테스트셋이며 이는 잔향이 없는 무향실에서의 테스트 성능이다. ‘Recorded Hallway’, ‘Recorded Room’은 녹음 테스트셋이며 나머지는 합성 테스트셋이다. 테스트셋은 RT60의 크기에 따라 오름차순으로 정렬하였다.

잔향이 없는 환경의 테스트셋인 ‘Clean’의 성능은 99.32%로 높은 성능을 보이거나 잔향 환경의 테스트셋으로 검증할 시 성능이 급격하게 저하되는 것을 확인할 수 있다. 모든 잔향 환경 테스트셋의 각 10 번의 실험에 대하여 ‘Clean’ 테스트셋과 통계적으로 유의미한 성능 하락이 있었다(Paired t -test, $p < 0.001$). 특히, 합성 테스트셋에서 [그림 4.7]과 같이 RT60이 증가함에 따라서 성능이 저하되는 경향성을 확인하였다.

[그림 4.8]은 데이터셋 내 *cherry* 클래스의 샘플 오디오에 대하여, 4 개 테스트

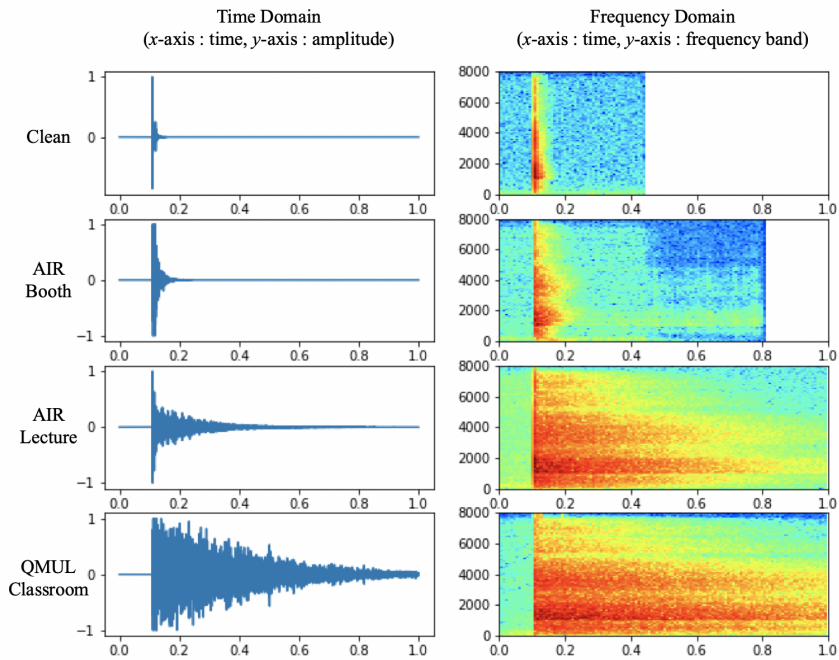


그림 4.8: 4 개 테스트 공간에서의 샘플 오디오 분석 (클래스 : *cherry*)

공간에서의 시간축 신호 및 주파수축 스펙트로그램 분석이다. 4 개 테스트 공간은 각각 잔향에 의한 왜곡이 없는 원본 테스트셋인 ‘Clean’, 그리고 각각 0.267, 0.682, 1.341의 RT60 길이를 갖는 ‘AIR Booth’, ‘AIR Lecture’, ‘QMUL Classroom’을 선정하였다.

RT60이 짧은 AIR Booth 테스트 공간에서의 이벤트 파형은 시간축 신호와 주파수축 스펙트로그램 모두 원본 테스트 공간인 ‘Clean’과 유사함을 볼 수 있다. 반면 ‘AIR Lecture’, ‘QMUL Classroom’ 테스트 공간에서는 RT60이 길어짐에 따라 파형이 원본 ‘Clean’에 비해 크게 왜곡되었다. [그림 4.9]은 *ring* 클래스의 샘플 오디오에 대한 동일한 분석이며, 마찬가지로 RT60이 길어짐에 따라 원본 ‘Clean’에 비해 왜곡되는 정도가 증가한다.

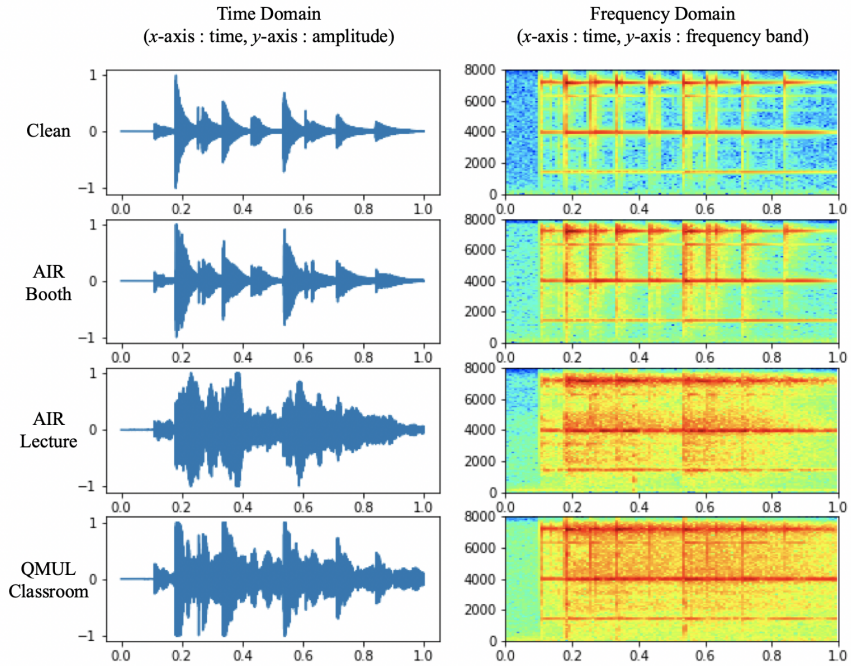


그림 4.9: 4 개 테스트 공간에서의 샘플 오디오 분석 (클래스 : ring)

[표 4.4]와 [그림 4.7]을 보면 ‘Recorded Hallway’, ‘Recorded Room’의 녹음 테스트셋의 성능이 합성 테스트셋의 RT60에 따른 경향에 비해 현저하게 낮는데 이는 제작한 녹음 테스트셋에 잔향외에 잡음에 의한 추가적인 왜곡이 반영되었기 때문으로 추측한다. 합성 테스트셋의 경우 무향실에서 녹음된 원본 데이터셋과 공간 임펄스 데이터셋과의 컨볼루션으로 생성하여, 두 데이터셋이 잡음에 의한 왜곡이 없다는 가정하에 합성 테스트셋 또한 잡음에 의한 왜곡이 없다.

반면 녹음 테스트셋의 경우, 원본 데이터셋을 복도와 회의실 두 공간에서 재생하고 마이크로 재녹음하는 과정에서 마이크의 잡음과, 복도와 회의실에 존재하는 잡음에 의한 추가적인 왜곡이 반영되었다. 실제로 [그림 4.10]과 같이 합성 테스트셋과

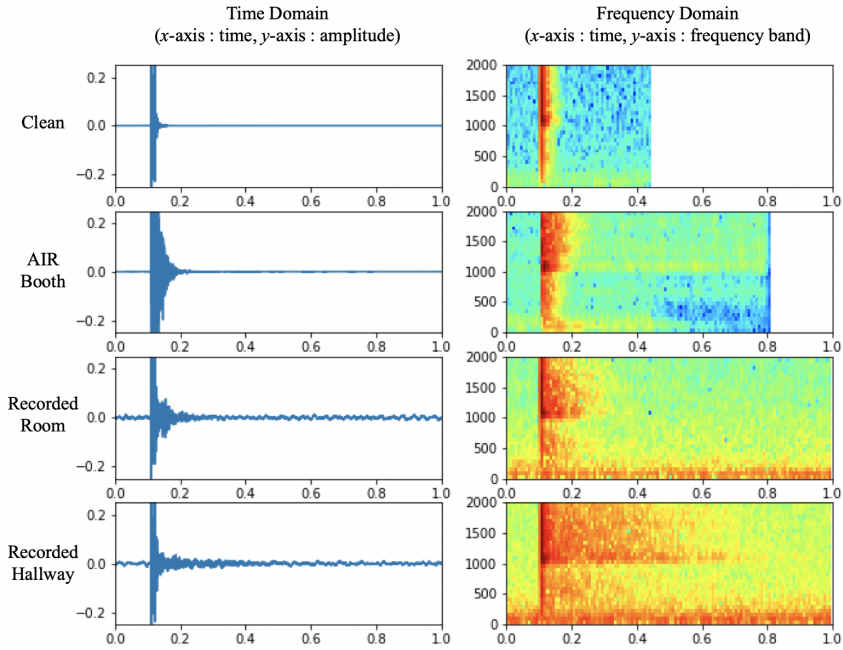


그림 4.10: 녹음 테스트셋과 합성 테스트셋의 샘플 오디오 비교 (클래스 : *cherry*)

녹음 테스트셋의 샘플 오디오를 비교했을 때, 합성 테스트셋의 오디오는 이벤트신호 발생 전 및 잔향 신호가 종료된 시점에서 신호가 없는 반면, 녹음 테스트셋의 오디오는 오디오 전역에 걸쳐 잡음이 존재함을 확인하였다. 주파수 축 스펙트로그램을 통해서도 녹음 테스트셋에서는 특히 저주파수 대역에 잡음으로 인한 추가 왜곡이 발생한 것을 볼 수 있다. 이 잡음이 합성 테스트셋의 RT60에 따른 성능 경향에 비해 녹음 테스트셋에서 더 낮은 성능을 보이는 이유라고 추측한다.

4.2.2 디컨블루션 적용 시 성능 및 한계점

[표 4.5]는 2.1.3에서 설명한 디컨블루션을 적용한 베이스라인 모델의 성능이다. ‘Baseline’은 앞서 설명한 베이스라인 모델이며, ‘Deconv’는 이 베이스라인 모델에 입력전에 테스트셋에 해당하는 공간 임펄스 응답으로 디컨블루션을 적용한 후 입력한 결과이고 디컨블루션 연산 전 입력 길이에 제한을 두지 않았다. ‘Deconv short’는 테스트셋의 입력 길이를 기존 실험 설정과 같은 1 초로 제한한 후 디컨블루션을 적용한 모델이며, ‘Deconv shuffle’은 입력 오디오와 같은 공간 내에 있되, 다른 위치의 공간 임펄스 응답으로 디컨블루션 한 모델로 입력 길이에 제한을 두지 않았다.

먼저 ‘Deconv’ 모델의 경우 입력 오디오 길이에 제한을 두지 않는 이상 합성 테스트셋에서는 이론적으로 온전한 원본 사운드 오디오 복원이 가능하기에, ‘Clean’ 성능과 같은 99.32%의 성능을 보였다. 다만, 녹음 테스트셋의 경우 베이스라인 성능보다 하락한 것을 볼 수 있는데, 이는 사운드 이벤트를 왜곡시킨 잔향과 디컨블루션에 사용한 공간 임펄스 응답이 완전히 매핑되지 않기 때문으로 추측하였다.

‘Deconv short’의 경우 테스트셋의 RT60이 증가함에 따라 성능이 저하되며 이는 디컨블루션 연산을 이용한 온전한 복원을 위해서는 원본 사운드 이벤트 오디오 길이와 공간 임펄스 응답 오디오 길이가 포함된 충분히 긴 오디오 길이가 필요하기 때문이다. 따라서 공간의 RT60이 길어질 수록 온전한 복원을 위한 오디오 길이가 길어지게 되고, 이는 곧 입력의 길이를 제한했을 시의 성능 저하로 나타난다.

| Model | Clean | Recorded Hallway | Recorded Room | AIR Booth | WDR CR7 | AIR Office | MARDY | AIR Lecture | AIR Stairway | QMUL Classroom |
|----------------|-------|------------------|---------------|-----------|---------|------------|-------|-------------|--------------|----------------|
| Baseline | 99.32 | 56.56 | 57.74 | 88.11 | 75.66 | 61.28 | 53.27 | 47.10 | 46.89 | 32.25 |
| Deconv | 99.32 | 49.55 | 50.05 | 99.32 | 99.32 | 99.32 | 99.32 | 99.32 | 99.32 | 99.32 |
| Deconv short | 99.32 | 48.33 | 47.02 | 99.30 | 98.99 | 99.02 | 98.44 | 97.95 | 95.65 | 61.12 |
| Deconv shuffle | 99.32 | 20.97 | 32.81 | 56.81 | 31.33 | 32.11 | 25.80 | 22.59 | 24.33 | 18.76 |

표 4.5: 디컨블루션 적용 시 베이스라인 모델의 분류 성능 (%)

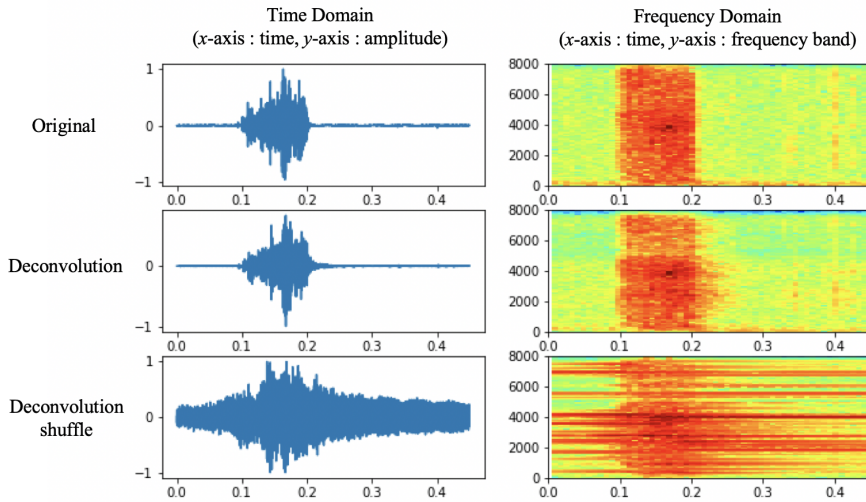


그림 4.11: 디컨볼루션 샘플 오디오 분석 (클래스 : *sandpp1*)

‘Deconv shuffle’은 베이스라인 성능에 비해 현저히 떨어지는 결과를 보이며, 이는 디컨볼루션 연산이 원본 사운드 이벤트 오디오를 왜곡시킨 잔향과 완벽히 매핑되는 공간 임펄스 응답이 아니라면, 같은 공간 내 공간 임펄스 응답이라고 하더라도 디컨볼루션 연산 시 원본 사운드 이벤트 오디오를 불완전하게 복원함을 의미한다.

[그림 4.11]는 ‘AIR Stairway’ 테스트셋 샘플 오디오의 디컨볼루션 이후 복원된 원본 오디오 분석이다. ‘AIR Stairway’는 합성 테스트셋으로 [그림 4.11]의 ‘Deconvolution’을 보면, 원본 ‘Original’과 매우 유사하게 신호 복원이 이루어졌으며 주파수 영역 스펙트로그램 또한 유사한 형태를 보인다. 반면, 동일 공간 내 다른 위치의 공간 임펄스 응답으로 디컨볼루션 한 ‘Deconvolution shuffle’을 보면 원본 ‘Original’에 비해 매우 왜곡되어 있음을 볼 수 있고, 주파수 영역 스펙트로그램을 통해서도 원본 신호에는 없던 주파수 성분들이 넓은 영역에 걸쳐 분포한 것을 볼 수 있다.

따라서 디컨블루션 연산을 적용한 잔향 환경에서의 사운드 이벤트 분류의 한계 점으로는 첫째, 특히 긴 잔향 시간을 갖는 공간에서 온전한 복원을 위해 긴 오디오 길이가 필요하며 이는 실제 환경의 사운드 이벤트 분류의 실시간성을 저하시킬 수 있다. 둘째, 같은 공간이라도 사운드 이벤트 발생 위치 등에 따라 수집한 공간 임펄스 응답과 정확히 매핑되지 않는 잔향의 영향을 받을 시, 불완전하게 복원되어 성능을 하락시킬 수 있다. 특히 실제 환경에서는 사운드 이벤트 오디오가 공간 내 매우 다양한 위치에서 발생하고, 공간 내 사람, 물체 등에 의해 잔향 특성이 쉽게 변화하기 때문에, 수집한 공간 임펄스 응답이 수신한 사운드 이벤트 오디오의 잔향과 완전히 매핑된다고 보장할 수 없다.

4.2.3 데이터 증가 방법을 이용한 성능 향상

[표 4.6]는 3.1에서 설명한, 제안한 데이터 증가 방법을 적용한 모델의 각 테스트셋에서의 분류 성능이다. ‘Baseline’은 제안한 기법을 적용하지 않은 베이스라인 모델이고 ‘Aug’는 제안한 데이터 증가 방법을 적용한 모델로 네트워크 구조는 베이스라인 모델과 같다. ‘Clean’ 테스트셋을 제외한 모든 데이터셋에서 ‘Baseline’에 비해 ‘Aug’에서 통계적으로 유의미한 성능 향상이 있었다(Paired *t*-test, $p < 0.001$).

제안한 데이터 증가 방법을 적용한 ‘Aug’ 모델 또한 RT60의 증가에 따라 분류 성능이 저하되지만([그림 4.12]), 잔향이 없는 ‘Clean’ 테스트셋을 제외한 모든 테스트셋에서 유의미한 성능 향상이 있었다는 점에서, 제안한 데이터 증가 방법이 잔향 환경에서의 사운드 이벤트 분류 성능을 개선시킴을 실험적으로 확인하였다. 특히

| Model | Clean | Recorded Hallway | Recorded Room | AIR Booth | WDR CR7 | AIR Office | MARDY | AIR Lecture | AIR Stairway | QMUL Classroom |
|----------|-------|------------------|---------------|-----------|---------|------------|-------|-------------|--------------|----------------|
| Baseline | 99.32 | 56.56 | 57.74 | 88.11 | 75.66 | 61.28 | 53.27 | 47.10 | 46.89 | 32.25 |
| Aug | 99.38 | 81.18 | 75.39 | 98.38 | 98.39 | 96.95 | 94.90 | 91.14 | 87.94 | 67.56 |

표 4.6: 데이터 증가 방법을 적용한 모델의 분류 성능 (%)

이미지 방법을 이용하여 인위적으로 제작한 가상 공간 임펄스 응답을 이용하여 실제 녹음된 공간 임펄스 응답으로 제작한 테스트셋에서 성능 향상이 있었다는 점에서 그 의의가 있다고 할 수 있다.

4.2.4 컨디셔닝 네트워크를 이용한 성능 향상

[표 4.7]는 3.2에서 설명한, 제안한 컨디셔닝 네트워크 모델의 각 테스트셋에서의 분류 성능이다. ‘Aug+Cndt’는 데이터 증가 방법과 제안한 공간 임펄스 응답 컨디셔닝 네트워크를 함께 사용한 모델이고, ‘Aug+Cndt+transfer’는 이 모델에 전이학습을 이용하여 추가 정보를 전이한 모델이다.

먼저 ‘Aug+Cndt’ 모델의 경우 ‘Aug’ 모델과 비교했을 때 ‘MARDY’, ‘AIR Lecture’, ‘AIR Stairway’, 그리고 ‘QMUL Classroom’ 테스트셋에서 통계적으로 유의미한 성능 향상이 있었다(Paired t -test, $p < 0.05$). 통계적으로 유의미한 성능 향상을

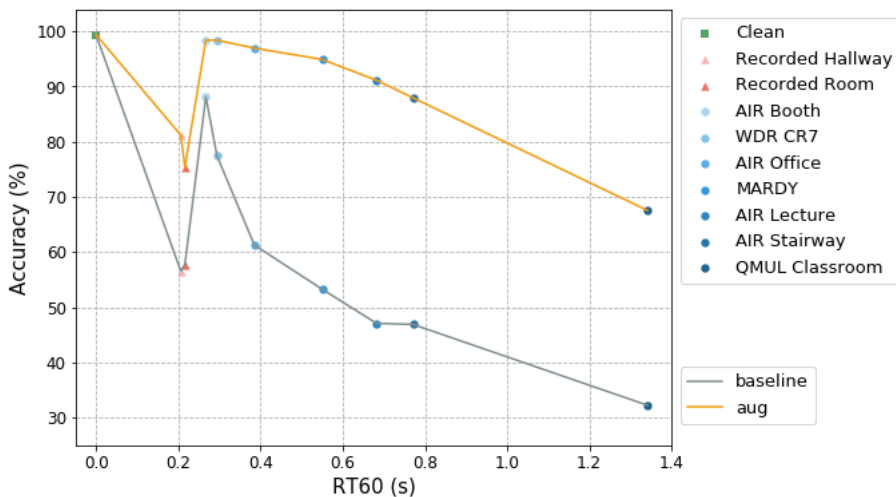


그림 4.12: 데이터 증가 방법 적용 모델의 RT60에 따른 분류 성능 (%)

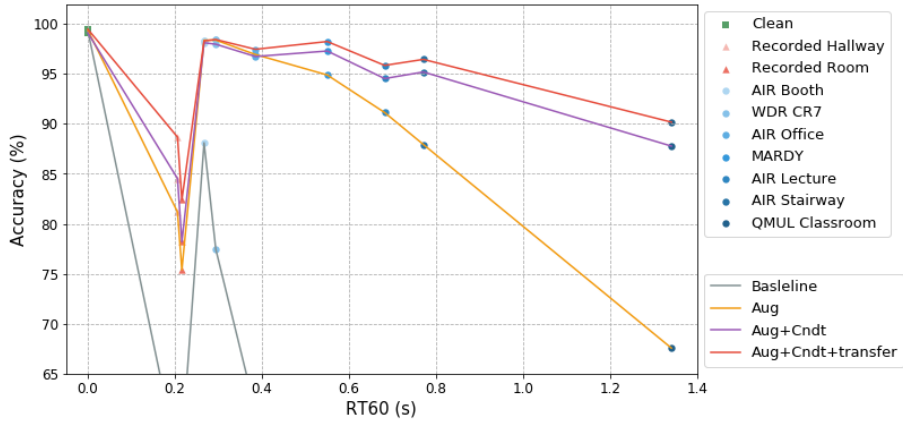


그림 4.13: 컨디셔닝 네트워크 모델의 RT60에 따른 분류 성능 (%)

보인 4 개 테스트셋은 RT60 길이가 긴 상위 4 개 테스트셋이며, 제안한 컨디셔닝 네트워크가 데이터 증가 방법만으로 성능 향상의 한계가 있는 잔향 공간에서 추가로 성능을 향상시킴을 확인하였다.

나머지 테스트셋에서는 ‘Aug’ 모델과 비교했을 때 통계적으로 차이가 없었으며, 특히 ‘Clean’, ‘AIR Booth’, ‘WDR CR7’, 그리고 ‘AIR Office’ 테스트셋은 ‘Aug’ 모델 만으로도 성능이 포화점에 근접하게 도달했기 때문으로 추측한다.

다음으로 제안한 데이터 증가 방법과 컨디셔닝 네트워크, 그리고 전이학습까지 적용한 ‘Aug+Cndt+transfer’ 모델의 경우, 모든 테스트셋에서 4 가지 모델중 가

| Model | Clean | Recorded Hallway | Recorded Room | AIR Booth | WDR CR7 | AIR Office | MARDY | AIR Lecture | AIR Stairway | QMUL Classroom |
|-------------------|-------|------------------|---------------|-----------|---------|------------|-------|-------------|--------------|----------------|
| Baseline | 99.32 | 56.56 | 57.74 | 88.11 | 75.66 | 61.28 | 53.27 | 47.10 | 46.89 | 32.25 |
| Aug | 99.38 | 81.18 | 75.39 | 98.38 | 98.39 | 96.95 | 94.90 | 91.14 | 87.94 | 67.56 |
| Aug+Cndt | 99.11 | 84.53 | 78.20 | 98.13 | 97.98 | 96.75 | 97.30 | 94.54 | 95.20 | 87.78 |
| Aug+Cndt+transfer | 99.48 | 88.68 | 82.42 | 98.31 | 98.45 | 97.47 | 98.25 | 95.86 | 96.46 | 90.18 |

표 4.7: 컨디셔닝 네트워크를 적용한 모델의 분류 성능 (%)

장 높은 분류 성능을 보인다. ‘Aug+Cndt’ 모델과 비교했을 때 ‘Recorded Hallway’, ‘Recorded Room’, ‘MARDY’ 테스트셋에서 통계적으로 유의한 성능 향상을 보였다 (Paired t -test, $p < 0.05$).

‘Aug’ 모델과 비교했을 때에는 ‘Recorded Hallway’, ‘Recorded Room’, ‘MARDY’, ‘AIR Lecture’, ‘AIR Stairway’, 그리고 ‘QMUL Classroom’ 테스트셋에서 ‘Aug+Cndt’ 모델에 비해 통계적으로 더욱 유의한 성능 향상을 보였다(Paired t -test, $p < 0.01$). [그림 4.13]은 4 가지 모델의 RT60에 따른 각 테스트셋에서의 성능이며 ‘Aug+ Cndt+transfer’ 모델이 가장 높은 성능을 가짐을 보여준다.

t -stochastic neighbor embedding (t -SNE) [39]는 고차원의 벡터들간의 구조적 관계를 유지하는 2 차원의 벡터를 학습하여 고차원의 데이터를 2 차원 지도로 표현하는 알고리즘으로, 벡터 시각화에 많이 사용된다. [그림 4.14]은 테스트셋에 사용한 공간 임펄스 응답들의 임베딩 벡터를 t -SNE로 표현한 것이다. 임베딩 네트워크는 제안한 ‘Aug+Cndt+transfer’ 모델의 임베딩 블록을 사용하였고([그림 4.6]의 ‘RIR embedding block’), 각 테스트셋 별 포인트 수는 테스트셋내의 공간 임펄스 응답 개수이다([표 4.2] 참조). 테스트셋과 해당하는 RT60(s)을 함께 표기하였다.

[그림 4.14]을 보면 동일 테스트셋 내의 공간 임펄스 응답끼리 군집해 있는 것을 볼 수 있으며, 또한 비슷한 RT60 길이를 갖는 테스트셋끼리 군집하여 3 개의 큰 그룹을 이루는 것을 볼 수 있다. 0.3 초 미만의 RT60을 갖는 ‘Clean’, ‘AIR Booth’, ‘WDR CR7’, ‘Recorded Hallway’, ‘Recorded Room’ 그룹(‘짧은 RT 그룹’)과 0.3 초 이상 1 초 미만의 RT60을 갖는 ‘AIR Office’, ‘MARDY’, ‘AIR Lecture’, ‘AIR Stairway’ 그룹(‘중간 RT 그룹’), 그리고 1 초 이상의 RT60을 갖는 ‘QMUL Classroom’ 그룹(‘긴 RT 그룹’)으로 나뉘어진다.

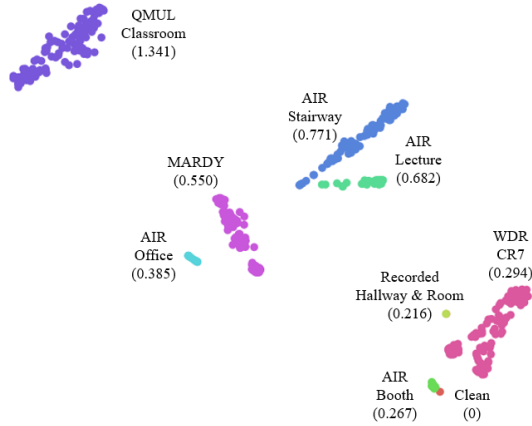


그림 4.14: 테스트셋에 사용한 공간 임펄스 응답 임베딩 벡터의 t-SNE 표현

t-SNE의 random state (RS)를 변경함에 따라 결과 표현이 조금씩 달라지는데, 대부분의 t-SNE 결과에서 3 개의 군집 그룹 및 그 구성 테스트셋의 종류는 유지되었다. 단, ‘Recorded Hallway’, ‘Recorded Room’ 테스트셋의 경우 ‘중간 RT 그룹’에 근접하게 표현되는 결과가 발견되었는데, 이는 RT60이 0.3 초 미만임에도 4.2.1에서 서술했듯 녹음 테스트셋의 공간 임펄스 응답 오디오 전역에 합성 테스트셋의 공간 임펄스 응답보다 잡음이 많이 존재하기 때문에 잡음의 영향이 마치 잔향의 영향과 유사하게 동작했을 것으로 추측한다. 전이학습을 적용하지 않은 ‘Aug+Cndt’ 모델을 이용하여 t-SNE 결과를 분석했을 때에도 이와 유사한 경향성이 발견되었다.

제안한 컨디셔닝 네트워크가 성능 향상 효과가 있음을 보이는 [표 4.7]와, RT60의 길이로 군집이 이루어지는 [그림 4.14]에 따라 제안한 컨디셔닝 네트워크가 보조 입력으로 주어지는 공간 임펄스 응답에서, RT60과 관계된 정보를 추출한 뒤 이를 분류 네트워크에 활용하는 것이라 추측하였고, 이를 검증하기 위해 소스 입력과 상관 없는 다른 공간의 임펄스 응답을 보조 입력으로 주는 Fake 컨디셔닝 실험을 수행하였고 그 개요도는 [그림 4.15]과 같다.

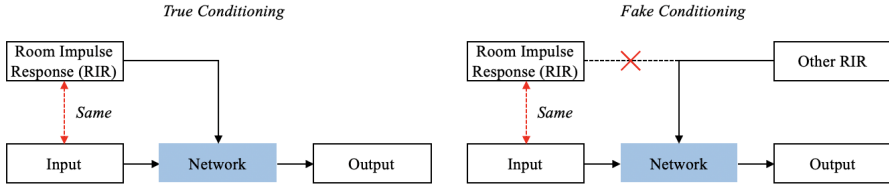


그림 4.15: Fake 컨디셔닝 실험 개요도

[그림 4.15]에서 True 컨디셔닝 방법은 Input에서 오디오를 왜곡시킨 공간 임펄스 응답을 네트워크의 보조 입력으로 주어 컨디셔닝하지만, Fake 컨디셔닝 방법에는 Input의 공간 임펄스 응답과 관련없는 다른 공간 임펄스 응답으로 Fake 컨디셔닝한다. [그림 4.14]의 세 그룹(‘짧은 RT 그룹’, ‘중간 RT 그룹’, ‘긴 RT 그룹’) 별 1 개 테스트셋을 골라 Fake 컨디셔닝을 실험한 결과는 [표 4.8]와 같다. 각 표의 상단은 Input에 들어가는 테스트셋의 이름과 RT60을 나타내며, 각 표 내에서는 Input과 상관 없는 다른 공간의 테스트셋의 공간 임펄스 응답으로 컨디셔닝한 결과이며 분류 성능의 내림차순으로 정렬하였다. 단, Room Name 중 True는 True 컨디셔닝을 적용한 본래의 결과이고, True가 아님에도 표 상단의 테스트셋과 동일한 이름을 갖는 Room Name (예: 첫번째 표의 4번째 행 AIR Booth)은 동일한 테스트셋 내에서 서로 다른 공간 임펄스 응답을 무작위로 컨디셔닝한 결과이다. p 값 유의성 표기는 True 결과와 나머지 결과 사이의 통계 검정 결과이다(Paired *t*-test).

먼저 [표 4.8]의 AIR Booth를 보면 AIR Booth가 속해있는 ‘짧은 RT 그룹’의 다른 테스트셋 공간 임펄스 응답으로 Fake 컨디셔닝 할시 성능이 크게 감소하지 않으며 테스트셋의 RT60이 증가함에 따라 성능이 감소한다. 반면 ‘긴 RT 그룹’인 QMUL Classroom의 경우 RT60이 감소함에 따라 성능이 감소하고, ‘중간 RT 그룹’인 AIR Lecture의 경우 같은 ‘중간 RT 그룹’에 속하는 테스트셋으로 Fake 컨디셔닝 했을때

| Room Name | Accuracy (%) | RT60 (s) |
|------------------|--------------|----------|
| Clean | 98.38 | 0 |
| True | 98.31 | 0.267 |
| AIR Booth | 98.29 | 0.267 |
| WDR CR7 | 98.02* | 0.294 |
| Recorded Room | 97.97* | 0.216 |
| Recorded Hallway | 97.87** | 0.206 |
| AIR Office | 97.41** | 0.385 |
| MARDY | 97.18** | 0.550 |
| AIR Lecture | 96.11** | 0.682 |
| AIR Stairway | 95.60** | 0.771 |
| QMUL Classroom | 89.76*** | 1.341 |

| Room Name | Accuracy (%) | RT60 (s) |
|------------------|--------------|----------|
| MARDY | 96.22 | 0.550 |
| AIR Office | 96.01 | 0.385 |
| AIR Lecture | 95.93 | 0.682 |
| True | 95.86 | 0.682 |
| AIR Stairway | 95.39* | 0.771 |
| Recorded Hallway | 95.06 | 0.206 |
| Recorded Room | 94.75 | 0.216 |
| WDR CR7 | 94.47* | 0.294 |
| AIR Booth | 92.89** | 0.267 |
| QMUL Classroom | 90.99*** | 1.341 |
| Clean | 89.91*** | 0 |

| Room Name | Accuracy (%) | RT60 (s) |
|------------------|--------------|----------|
| QMUL Classroom | 90.19 | 1.341 |
| True | 90.18 | 1.341 |
| AIR Stairway | 89.96 | 0.771 |
| AIR Lecture | 89.33 | 0.682 |
| MARDY | 87.38* | 0.550 |
| AIR Office | 84.65** | 0.385 |
| Recorded Hallway | 80.76*** | 0.206 |
| Recorded Room | 80.20*** | 0.216 |
| WDR CR7 | 79.26*** | 0.294 |
| AIR Booth | 74.29*** | 0.267 |
| Clean | 69.93*** | 0 |

*** : p < 0.001, ** : p < 0.01, * : p < 0.05

표 4.8: Fake 컨디셔닝 실험 결과

보다 ‘짧은 RT 그룹’ 혹은 ‘긴 RT 그룹’에 속하는 테스트셋으로 컨디셔닝했을 시 성능이 더 감소한다. 이를 통해 제안한 컨디셔닝 네트워크가 공간의 RT60과 관련된 정보를 네트워크에 부여함으로써 공간 간향에 의한 성능 저하를 개선함을 확인하였고, 이는 임의의 공간의 정확한 공간 임펄스 응답 오디오를 알지 못할 때에도 RT60 정보만으로도 성능 향상의 가능성이 있음을 시사한다.

[표 4.9]는 [표 4.5]의 디컨볼루션을 적용한 모델과 제안한 컨디셔닝 네트워크 모델의 성능을 비교한 결과이다. ‘Baseline’은 베이스라인 모델, ‘Deconv’는 베이스라인 모델의 입력 전에 디컨볼루션을 적용한 모델이며, ‘Deconv shuffle’은 디컨볼루션 적용 시 입력과 정확히 매핑되지 않은, 같은 공간 내 다른 위치의 공간 임펄스 응답으로 디컨볼루션을 적용한 모델이다. ‘Aug+Cndt+transfer’는 최종 제안한 전이

| Model | Clean | Recorded Hallway | Recorded Room | AIR Booth | WDR CR7 | AIR Office | MARDY | AIR Lecture | AIR Stairway | QMUL Classroom |
|------------------------------|-------|---------------------|------------------|--------------|------------|---------------|-------|----------------|-----------------|-------------------|
| Baseline | 99.32 | 56.56 | 57.74 | 88.11 | 75.66 | 61.28 | 53.27 | 47.10 | 46.89 | 32.25 |
| Deconv | 99.32 | 49.55 | 50.05 | 99.32 | 99.32 | 99.32 | 99.32 | 99.32 | 99.32 | 99.32 |
| Deconv shuffle | 99.32 | 20.97 | 32.81 | 56.81 | 31.33 | 32.11 | 25.80 | 22.59 | 24.33 | 18.76 |
| Aug+Cndt+transfer | 99.48 | 88.68 | 82.42 | 98.31 | 98.45 | 97.47 | 98.25 | 95.86 | 96.46 | 90.18 |
| Aug+Cndt+transfer shuffle | 99.48 | 88.41 | 80.93 | 98.29 | 98.48 | 97.51 | 98.12 | 95.93 | 96.21 | 90.19 |

표 4.9: 디컨볼루션 적용 베이스라인 모델과 제안한 컨디셔닝 네트워크 모델의 분류 성능 (%)

학습이 적용된 컨디셔닝 네트워크 모델이며, ‘Aug+Cndt+transfer shuffle’은 제안한 모델에서 공간 임펄스 응답을 컨디셔닝 할 때, 입력과 정확히 매핑되지 않은, 같은 공간 내 다른 위치의 공간 임펄스 응답으로 컨디셔닝 한 모델이다.

‘Deconv’ 와 ‘Deconv shuffle’의 결과를 보면 4.2.2에서 서술한 대로, 합성 테스트 셋의 경우 이상적인 디컨볼루션 연산이 가능하기에 ‘Clean’ 테스트셋과 동일한 성능을 보이지만(‘Deconv’), 녹음 테스트셋의 결과와 ‘Deconv shuffle’의 결과를 통해 원본 오디오를 왜곡시킨 잔향과 완벽히 매핑되는 공간 임펄스 응답을 알 수 없다면, 같은 공간 내 공간 임펄스 응답이라 하더라도 디컨볼루션 연산 후 불완전하게 복원 되어 베이스라인 보다 하락한 성능을 보인다.

하지만 제안한 컨디셔닝 네트워크 모델의 경우, 원본 오디오를 왜곡시킨 잔향과 완벽히 매핑되는 공간 임펄스 응답이 아닌 다른 위치의 공간 임펄스 응답으로 컨디셔닝 했을 때에도 성능이 하락하지 않는 것을 볼 수 있다(‘Aug+Cndt+transfer shuffle’). 실제로 통계 검정 시, ‘Recorded Room’ 테스트셋을 제외한 모든 테스트셋에서 통계적으로 차이가 없었다(Paired *t*-test).

실제 환경에서 사운드 이벤트 분류 수행 시 공간 내 사람, 물체 등 신호가 반사될 수 있는 모든 대상의 움직임에 따라 공간 임펄스 응답이 변화할 수 있고, 또 사운드

이벤트 분류를 수행하는 하드웨어가 특정 위치에 고정되어 있다 하더라도 발생하는 사운드 이벤트의 위치가 변화하기 때문에 이벤트 발생 시 마다 사운드 이벤트를 왜곡시킨 정확한 잔향의 공간 임펄스 응답을 얻기 힘들다. 따라서, 제작한 테스트셋 중 녹음 테스트셋 결과와 모델 중 ‘Deconv shuffle’, ‘Aug+Cndt+transfer shuffle’ 모델의 결과가 정확한 공간 임펄스 응답이 아닌 같은 공간 내 유사 공간 임펄스 응답을 사용하므로 실제 환경에 더 부합하는 결과라고 할 수 있다.

결과적으로 제안한 기법 중 데이터 증가 방법은 공간의 정보를 모를 때라도 적용 가능하며, 컨디셔닝 네트워크는 공간의 정확한 공간 임펄스 응답 오디오를 알 수 없을 때에도 유사한 공간 임펄스 응답 혹은 RT60 정보만으로도 추가적인 성능 향상을 이룰 수 있다는 점에서, 디컨볼루션을 적용한 방법의 한계점을 극복할 수 있다.

제 5 장 결 론

5.1 연구 의의

본 연구에서는 잔향 환경에서 사운드 이벤트 분류 성능이 저하됨을 실험적으로 확인하고 이에 대한 개선 기법을 제안하였다. 먼저 잔향 환경에서의 사운드 이벤트 분류 성능 저하를 확인하기 위해 공간 임펄스 응답 데이터셋과 컨볼루션하여 생성한 합성 테스트셋 7 개와 데이터셋을 잔향이 존재하는 공간에서 재녹음한 녹음 테스트셋 2 개를 제작하였고, 통계적으로 유의미한 성능 저하를 확인하였다. 특히, 테스트셋의 RT60이 증가함에 따라 성능이 하락하는 경향성을 확인하였고, 디컨볼루션 방법을 적용했을 때의 잔향 테스트셋에서의 성능과 그 한계점에 대해 고찰하였다.

잔향 환경에서의 성능 저하에 대한 개선 방법으로, 먼저 이미지 방법을 통해 인위적으로 제작한 가상 공간 임펄스 응답과의 컨볼루션 연산을 통해 데이터를 증가시키는 데이터 증가 방법을 제시하였고, 제안한 데이터 증가 방법이 설정한 모든 잔향 환경 테스트셋에서 성능 향상의 효과가 있음을 실험적으로 검증하였다. 또한 실제 환경에서 사운드 이벤트 분류를 수행함에 있어 해당 공간의 임펄스 응답을 쉽게 수집 가능하다는 점에서 착안하여 제안한 데이터 증가 방법과 함께 사용 가능한, 공간 임펄스 응답을 네트워크에 컨디셔닝하는 기법을 제안하였다. 구체적으로는 스케일링 및 바이어싱 요소간 변환 기법을 이용하였고 제안한 컨디셔닝 네트워크가 긴 RT60 길이를 갖는 테스트셋에서 성능 향상의 효과가 있음을 검증했다.

또한 제안한 컨디셔닝 네트워크의 추가 성능 향상을 위하여, 가상 공간 임펄스 응답을 분류하는 네트워크를 따로 훈련시켜 얻은 정보를 컨디셔닝 네트워크에 전이하는 기법을 사용하였고, 이를 통해 성능이 포화된 것으로 판단되는 일부 테스트셋

을 제외한 테스트셋에서 추가적으로 성능이 향상됨을 확인하였다.

최종 제안한 컨디셔닝 네트워크의 공간 임펄스 응답 임베딩 블록의 출력을 분석하여 네트워크가 공간 임펄스 응답 컨디셔닝 시 RT60과 관련된 정보를 사용하는 것이라 추측하였고, 실험을 통해 소스 입력과 완전히 동일하지 않더라도 유사한 RT60 길이를 갖는 공간 임펄스 응답을 컨디셔닝함으로써 성능을 향상시킬 수 있음을 보였다. 이는 실제 잔향 환경에서의 정확한 공간 임펄스 응답을 모를지라도, RT60의 정보만으로도 사운드 이벤트 분류 성능을 향상시킬 수 있음을 의미한다.

이에 따라 제안한 기법이, 정확한 공간 임펄스 응답을 사용할 경우에만 높은 성능을 보이는 디컨볼루션 적용 방법의 한계점을 극복할 수 있음을 확인하였다. 또, 공간의 잔향 시간이 길어짐에 따라 온전한 복원을 위해 더 긴 오디오 입력을 필요로 하는 디컨볼루션 방법과 달리 제안한 기법은 공간의 잔향 시간과 관계 없이 짧은 오디오 길이만으로도 향상된 성능을 보여, 실제 환경의 사운드 이벤트 분류 시 실시간성에도 유리할 것으로 기대된다. 궁극적으로 제안한 기법은 디컨볼루션 등의 중간 연산의 필요 없이 잔향에 의해 왜곡된 오디오 입력 시 이에 해당하는 클래스를 출력해주는 엔드 투 엔드(end-to-end) 네트워크로 그 구조적으로 효율적이라 할 수 있다.

저자의아는 바에 의하면, 잔향 환경에서의 사운드 이벤트 분류 성능에 관한 연구는 전무하다. 따라서 본 연구의 의의는 첫째, 잔향 환경을 모델링한 합성 테스트셋과 녹음 테스트셋을 제작하여 잔향 환경에서 사운드 이벤트 분류 성능이 저하됨을 실험적으로 확인하였고 둘째, 인위적으로 제작한 가상 공간 임펄스 응답과의 컨볼루션 연산을 통한 데이터 증가 방법으로 분류 성능을 향상시켰다. 셋째, 요소간 변환 및 전이학습을 이용하여 공간 임펄스 응답을 네트워크에 컨디셔닝하는 기법을 제안하여, 데이터 증가 방법과 함께 사용하였을 때 추가적인 성능 향상이 있음을 검증했다. 특

히 제안한 기법들이 가상의 공간 임펄스 응답을 이용하여 실제 공간 임펄스 응답을 이용한 합성 테스트셋과 실제 잔향이 존재하는 공간에서 녹음한 녹음 테스트셋에서 성능을 검증했으며, 정확한 공간 임펄스 응답 오디오를 모를 때라도 대략적인 잔향 시간 정보를 이용하여 적용 가능하다는데서 그 추가적 의의가 있다고 할 수 있다.

5.2 한계점

본 연구의 한계점으로는, 먼저 제안한 기법의 성능을 검증하는 벤치마크 이벤트 데이터셋이 1 개로, 이 데이터셋에 국한된 검증 결과라는데에 있다. 사용한 데이터셋 클래스가 50 개로 다양한 종류의 오디오를 담고있긴 하지만, 제안한 기법이 효과를 가지는 오디오 특성이 있는지에 대한 추가적인 검증이 필요하다.

둘째, 데이터셋을 잔향이 존재하는 환경에서 재녹음한 녹음 테스트셋에 대한 다양하고 면밀한 검증이 추가로 필요하다. 녹음 테스트셋 제작 시 실제 환경에 존재하는 잡음 등의 문제로 합성 테스트셋에 사용한 데이터셋에 비해 많은 양의 잡음이 녹음 테스트셋에 포함되었다. 잔향 뿐만 아니라 잡음 또한 사운드 이벤트 분류 성능을 저하시키는 주 요인이기에 잔향과 잡음이 확실하게 제어된 녹음 테스트셋을 제작하여 제안한 기법을 검증할 필요가 있다. 또한 이 과정에서 사용한 마이크와 스피커에 따른 성능 경향성 등을 파악할 필요가 있다.

셋째, 본 연구에서 제안한 기법이 베이스라인 모델에 비해 성능이 향상됨을 검증했지만 잔향을 제거하는 등 기존 잔향 관련 연구의 기법들과 성능을 비교하지 않았다는데에 그 한계가 있다. 사운드 이벤트 분류 분야 외에도, 음성 인식 등의 분야에서 잔향에 의한 왜곡을 해결하기 위해 잔향 제거(dereverberation) 등의 기법을 사용한다. 본 연구에서는 적용되는 실제 환경의 대략적 공간 임펄스 응답을 쉽게 얻

을 수 있다고 가정하여 디컨블루션 연산과의 성능을 비교했지만, 다양한 잔향 제거 기법을 적용했을 때의 성능과의 비교가 추가로 필요하다.

5.3 향후 연구

향후 연구 계획으로는, 먼저 5.2에서 제시한 한계점들을 해결하기 위한 검증을 시도할 계획이다. 다양한 오디오 특성을 가진 데이터셋, 특히 이미 잔향에 의해 왜곡된 데이터셋에도 적용되는지 검증할 계획이며, 또 면밀하게 설정한 다양한 녹음 환경에서 녹음 테스트셋을 제작하고 제안 기법을 검증할 계획이다. 특히, 실제 사운드 이벤트 분류를 수행하는 하드웨어와 마이크 등을 제어하여 실제 잔향 환경에서의 사운드 이벤트 분류 성능을 면밀히 검증할 계획이다.

두 번째로, 본 연구에서 제안한 컨디셔닝 네트워크가 입력에 사용한 공간 임펄스 응답 뿐만 아니라 이와 유사한 RT60 길이를 갖는 다양한 공간 임펄스 응답을 컨디셔닝해도 성능이 향상된다는 점에서 착안하여, 공간 임펄스 응답 오디오가 아닌 RT60 수치를 네트워크에 컨디셔닝하는 구조를 설계할 계획이다. 이는 제안한 컨디셔닝과 유사한 성능을 가진다는 가정하에 네트워크 용량을 줄일 수 있다는 장점을 갖는다.

세 번째로, 본 연구를 진행하면서 잔향에 의한 성능 저하 및 제안한 기법을 적용했을 때 성능 향상이 두드러지는 클래스가 존재함을 확인하였다. 이는 잔향에 의해 성능이 저하되는 이벤트의 특성 및 그에 따른 적절한 성능 향상 기법이 다를 수 있음을 시사하며, 클래스 특성에 따른 잔향 환경에서의 성능 저하 예측 및 적절한 성능 향상 기법을 제시하는 연구를 계획하고 있다.

또한 마지막으로, 본 연구에서 제안한 컨디셔닝 네트워크가 잔향 뿐 아니라 잡음

환경에서의 성능 또한 개선할 수 있는지를 연구할 계획이다. 실제 환경에서 성능을 저하시키는 주된 요인은 잡음과 잔향으로, 잡음을 이용한 컨디셔닝 네트워크를 통해 잡음 환경에서의 성능 향상 경향성을 확인하고 궁극적으로는 잡음과 잔향이 모두 존재하는 환경에서의 사운드 이벤트 분류 성능을 개선하기 위한 연구를 수행할 계획이다.

참고 문헌

- [1] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, “Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems,” *Cough.*, vol. 65, no.48, pp. 5, 2006.
- [2] R. Banerjee, A. Sinha and A. Saha, “Participatory sensing based traffic condition monitoring using horn detection,” *In Proceedings of the 28th annual ACM symposium on applied computing*, 2013. pp. 567-569.
- [3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio and M. Vento, “Audio surveillance of roads: A system for detecting anomalous sounds,” *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279-288, 2016.
- [4] 이재준, 김완수, 이교구, “합성곱 신경망 기반 환경잡음에 강인한교통 소음 분류 모델,” *한국음향학술지*, 제37권, 제6호, 469-474쪽, 2018년.
- [5] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, Sep 2007. pp. 21-26.
- [6] S. P. Van IJsselmuide and S. P. Beerens, “Detection and classification of marine mammals using an LFAS system,” *Canadian Acoustics*, vol. 32, no. 2, pp. 93-106, 2004.
- [7] T. H. Park, J. Turner, M. Musick, J. H. Lee, C. Jacoby, C. Mydlarz, and J. Salamon, “Sensing Urban Soundscapes,” *In EDBT/ICDT Workshops*, 2014. pp. 375-382.

- [8] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, ... and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, Nov 2017.
- [9] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, ... and T. Sainath, "Deep neural networks for acoustic modeling in speech recognition," in *IEEE Signal processing magazine*, vol. 29, 2012.
- [10] G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," in *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30-42, 2012.
- [11] A. Graves, A. R. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, May 2013. pp. 6645-6649.
- [12] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep 2015. pp. 1-6.
- [13] H. B. Sailor, D. M. Agrawal and H. A. Patil, "Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification," in *INTERSPEECH*, Aug 2017. pp. 3107-3111.
- [14] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," in *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279-283, 2017.
- [15] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*, MIT press, 2016.

- [16] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio and M. Vento, "Reliable detection of audio events in highly noisy environments," in *Pattern Recognition Letters*, vol. 65, pp. 22-28, 2015.
- [17] J. Salamon, C. Jacoby and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, Nov 2014. pp. 1041-1044.
- [18] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, Oct 2015. pp. 1015-1018.
- [19] A. Mesaros, T. Heittola, A. Eronen and T. Virtanen, "Acoustic event detection in real life recordings," In *2010 18th European Signal Processing Conference*, Aug 2010. pp. 1267-1271.
- [20] X. Zhuang, X. Zhou, M. A. Hasegawa-Johnson and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543-1551, 2010.
- [21] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," in *nature*, vol. 521, no. 7553, pp. 436, 2015.
- [22] S. J. Pan and Q. Yang, "A survey on transfer learning," in *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2009.
- [23] H. Kuttruff, *Room acoustics*, Crc Press, 2016.
- [24] J. Salamon and J. P. Bello, (2015, April). "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2015. pp. 171-175.

- [25] I. McLoughlin, H. Zhang, Z. Xie, Y. Song and W. Xiao, “Robust sound event classification using deep neural networks,” in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 540-552, 2015.
- [26] I. Ozer, Z. Ozer and O. Findik, “Noise robust sound event classification with convolutional neural network,” in *Neurocomputing*, vol. 272, pp. 505-512, 2018.
- [27] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” in *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943-950, 1979.
- [28] E. Perez, F. Strub, H. De Vries, V. Dumoulin and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, Apr 2018.
- [29] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. D. Vries, A. Courville and Y. Bengio, “Feature-wise transformations,” in *Distill*, vol. 3, no. 7, pp. e11, 2018.
- [30] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2901-2910.
- [31] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” 2000.
- [32] M. Jeub, M. Schafer and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *2009 16th International Conference on Digital Signal Processing*, Jul 2009, pp. 1-5.

- [33] MARDY dataset, <https://www.commsp.ee.ic.ac.uk/~sap/resources/mardy-multichannel-acoustic-reverberation-database-at-york-database>
- [34] R. Stewart and M. Sandler, "Database of omnidirectional and B-format room impulse responses," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar 2010, pp. 165-168.
- [35] P. Stade, B. Bernschütz and M. Rühl, "A spatial audio impulse response compilation captured at the WDR broadcast studios," in *Proceedings of the VDT International Convention*, Nov 2012.
- [36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, ... and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016. pp. 265-283.
- [37] Keras, <http://owl.mcmater.ca/solarmesh>
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.
- [39] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," in *Journal of machine learning research*, vol. 9, pp. 2579-2605, Nov 2008.

ABSTRACT

In this paper, we propose techniques to enhance performance of sound event classification in reverberant environment. Sound event classification is actively applied to various application fields such as anomaly detection system, and it is important to maintain robust performance in real-world environments. In real-world environments, noise and reverberation are the main factors that degrade the performance of sound event classification. However, the research on sound event classification in noisy and especially reverberant environments is poor.

Therefore, in this paper, we observe the degradation phenomenon of sound event classification in reverberant environments and propose performance enhancement techniques for this phenomenon. To do this, we build a test set that models the reverberant environments and observe that sound event classification performance of the test set is degraded.

In order to improve the performance, we propose a data augmentation method using an artificially synthesized room impulse response and a method of conditioning the room impulse response to the network. Experimental results show that the proposed data augmentation method improves performance in reverberant environments. It also demonstrates additional performance improvements when using with the proposed conditioning method together. Finally, we show that the proposed method improves the performance by using approximate reverberation time information even when accurate room impulse response audio is not known.

주요어: Sound event classification, reverberation, data augmentation, conditioning network

학 번: 2017-28919

감사의 글

나의 가는 길을 오직 그가 아시나니

그가 나를 단련하신 후에는 내가 정금 같이 나오리라 (욥 23:10)