

석사학위논문  
Master's Thesis

음향 인тен시티를 사용한 심층신경망 기반  
엔드투엔드 다채널 음성 잔향 제거 기법

End-to-End Multichannel Speech Dereverberation Using  
Acoustic Intensity Based on Deep Neural Networks

2020

정경석 (鄭京錫 Jeong, Kyeong-Seok)

한국과학기술원

Korea Advanced Institute of Science and Technology

석사학위논문

음향 인텐시티를 사용한 심층신경망 기반  
엔드투엔드 다채널 음성 잔향 제거 기법

2020

정경석

한국과학기술원

전기및전자공학부

# 음향 인텐시티를 사용한 심층신경망 기반 엔드투엔드 다채널 음성 잔향 제거 기법

정 경 석

위 논문은 한국과학기술원 석사학위논문으로  
학위논문 심사위원회의 심사를 통과하였음

2020년 6월 19일

심사위원장 최정우 (인)

심사위원 김회린 (인)

심사위원 한민수 (인)

# End-to-End Multichannel Speech Dereverberation Using Acoustic Intensity Based on Deep Neural Networks

Kyeong-Seok Jeong

Advisor: Jung-Woo Choi

A dissertation submitted to the faculty of  
Korea Advanced Institute of Science and Technology in  
partial fulfillment of the requirements for the degree of  
Master of Science in Electrical Engineering

Daejeon, Korea  
June 19, 2020

Approved by

---

Jung-Woo Choi  
Professor of Electrical Engineering

The study was conducted in accordance with Code of Research Ethics<sup>1</sup>.

---

<sup>1</sup> Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MEE

정경석. 음향 인텐시티를 사용한 심층신경망 기반 엔드투엔드 다채널 음성  
잔향 제거 기법. 전기및전자공학부 . 2020년. 60+iv 쪽. 지도교수: 최정우.  
(한글 논문)

Kyeong-Seok Jeong. End-to-End Multichannel Speech Dereverberation Using Acoustic Intensity Based on Deep Neural Networks. School of Electrical Engineering . 2020. 60+iv pages. Advisor: Jung-Woo Choi. (Text in Korean)

### 초 록

객체 기반 오디오 시스템은 마이크로폰 어레이로 녹음된 신호에서 개별 음원에 대한 정보와 공간 정보를 분리하여 저장한 후, 렌더링 시에 현장과 유사하게 음장을 재현하는 시스템이다. 이를 구현하기 위해선 녹음된 신호로부터 잔향을 제거할 필요가 있다. 최근 심층신경망을 이용해 다채널 신호를 입력으로 받아 잔향을 제거하는 기법이 좋은 성능을 보이나, 크기 스펙트로그램만을 다루기에 왜곡된 위상으로 인해 복원한 신호의 품질이 떨어지는 단점이 있다. 본 연구에서는 그러한 문제를 완화하고자 시간 영역의 방향 특징을 입력으로 하는 엔드투엔드 모델을 사용할 것을 제안한다. 방향 특징은 삼차원 음향 인텐시티의 근사값으로 잔향의 양이나 직접파와 반사파의 방향을 판단하는 데에 유용한 정보이다. 훈련에 사용한 모델은 변형을 통해 보다 효율적인 학습을 가능케 하였으며 기존 모델 및 기법들과 비교하여 우수한 성능을 보임을 확인하였다.

핵심 날말 잔향 제거, 심층신경망, 음향 인텐시티, 시간 영역 특징

### Abstract

In the object-based audio system, information of sound sources and reverberant room are saved separately so that a renderer reproduces realistic and interactive sound field. Reverberation of recorded sources should be removed to implement the system. Recent studies show that deep neural networks (DNNs) can learn dereverberation using multi-channel reverberant signals. However, they've mainly dealt with the magnitude spectrogram except the phase, which led to degraded speech quality. To relieve the issue, an end-to-end DNN model using time-domain directional features is proposed in the research. The features are related to 3-dimensional acoustic intensity, useful to infer the degree of reverberation and the direction of direct and reflected waves. Compared with the original model and prior techniques, the modified DNN model for efficient learning shows de-reverberation performance at the significantly lower computational complexity.

**Keywords** Dereverberation, Deep Neural Networks, Acoustic Intensity, Time-domain Feature

# 차 례

차 례 . . . . .	i
표 차례 . . . . .	iii
그림 차례 . . . . .	iv
제 1 장      머릿말	1
1.1      연구 배경 및 목적 . . . . .	1
1.2      논문의 구성 . . . . .	5
제 2 장      이론적 배경	7
2.1      음향 인텐시티 . . . . .	7
2.2      실수 구면 조화 영역 신호 . . . . .	10
2.3      Griffin-Lim 알고리즘 . . . . .	11
2.4      심충 학습 . . . . .	12
2.4.1      인공신경망 및 심충신경망 . . . . .	14
2.4.2      합성곱 신경망 . . . . .	16
제 3 장      기준의 잔향 제거 기법	19
3.1      문제 정의 . . . . .	19
3.2      신호 처리 기법 . . . . .	20
3.2.1      빔 형성 기법 . . . . .	20
3.2.2      선형 예측 기법 . . . . .	22
3.3      심충신경망 기반 기법 . . . . .	25
3.3.1      신호 처리 기법의 보조용으로 심충신경망을 사용하는 기법 . . . . .	25
3.3.2      심충신경망이 직접 잔향을 제거하는 기법 . . . . .	26
제 4 장      제안 기법	29

4.1	방향 특징	29
4.2	심충신경망 설계	31
4.2.1	입출력 전처리	32
4.2.2	심충신경망 구조	33
<b>제 5 장</b>	<b>실험 및 결과</b>	<b>37</b>
5.1	실험 설정	37
5.1.1	성능 평가 지표 및 비교 기법	37
5.1.2	데이터 세트 구성	38
5.1.3	하이퍼파라미터 설정	39
5.2	실험 결과	39
5.2.1	테스트 데이터 세트에 대한 잔향 제거 성능	41
5.2.2	추정한 시간 영역 신호에 포함된 위상의 유의 미성	44
<b>제 6 장</b>	<b>맺음말</b>	<b>47</b>
<b>사 사</b>		<b>58</b>
<b>약 력</b>		<b>60</b>

## 표 차례

4.1 Wave-U-Net의 구조 변형 및 파라미터 수의 변화 . . . . .	34
5.1 심층신경망 훈련을 위한 하이퍼파라미터 . . . . .	39
5.2 모델 별 파라미터 수 . . . . .	41

# 그림 차례

1.1	객체 기반 오디오 시스템 . . . . .	1
2.1	전결합 레이어와 MLP의 도식화 . . . . .	15
2.2	합성곱 신경망의 도식화 . . . . .	18
3.1	DAS 빔 형성 기법 . . . . .	21
3.2	단일 채널 신호를 이용한 잔향 제거 기법의 구조도 . . . . .	26
3.3	방향 특징을 사용한 다채널 잔향 제거 기법의 구조도 . . . . .	27
4.1	제안하는 잔향 제거 기법의 구조도 . . . . .	29
4.2	프리엠퍼시스 및 디엠퍼시스 필터의 주파수에 따른 이득 . . . . .	33
4.3	Wave-U-Net 구조 . . . . .	35
4.4	본 연구에서 제안한 Wave-U-Net의 변형 구조 . . . . .	36
5.1	합성에 사용한 Eigenmike 사진 . . . . .	38
5.2	훈련 이포크에 따른 검증 데이터의 손실 함수값 변화 . . . . .	40
5.3	훈련된 심층신경망 모델들과 테스트 데이터 세트에서의 잔향 제거 성능 .	41
5.4	잔향이 포함된 입력 대비 훈련된 심층신경망 모델들의 잔향 제거 성능 향상도	42
5.5	테스트 샘플의 입력 및 정답 스펙트로그램과 각 모델의 출력 스펙트로그램	44
5.6	추정 위상에 따른 앤드투엔드 모델의 성능 지표 값 . . . . .	46

# 제 1 장 머릿말

## 1.1 연구 배경 및 목적

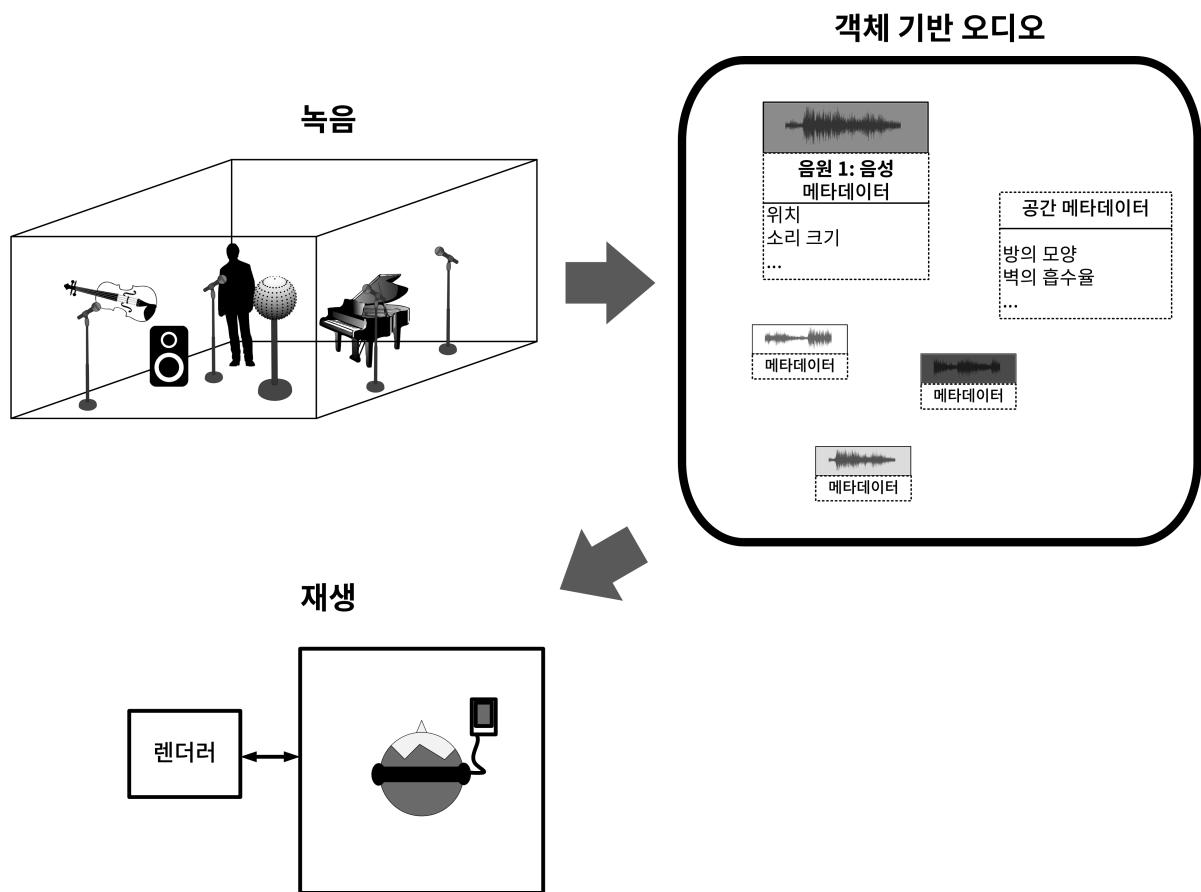


그림 1.1: 객체 기반 오디오 시스템 개요도

최근 가상현실 시스템이 대중화됨에 따라 그림 1.1과 같은 객체 기반 오디오 시스템 [1]에 대한 연구의 필요성이 높아지고 있다. 객체 기반 오디오 시스템에서는 녹음 시에, 공간상에 분포된 다수의 마이크로폰이나 마이크로폰 어레이를 사용하여 음원에 대한 정보와 공간에 대한 정보를 분리하여 개별로 저장한다. 이때 저장된 오디오 데이터는 각 음원의 위치와 소리의 크기 등의 메타데이터를 포함하며, 공간의 구조나 물체의 반사 계수 (reflection coefficient) 등의 녹음 현장에 대한 메타데이터도 포함한다. 재생 시에는 렌더러가 메타데이터를 종합하여 가상의 음장을 청취자의 환경에서 재현할 수 있다. 청취자의

환경이란, 가상현실 시스템에서는 헤드폰을 사용하는 바이노럴 (binaural) 환경인 경우가 일반적이고, 영화관 등의 장소에서는 스피커 어레이를 사용하는 환경일 수 있다. 이때 사용자는 메타데이터를 선택적으로 변형함으로써 실시간 상호작용이 가능한 감상 방식을 구현할 수 있다.

해당 시스템을 구현하기 위해서는 개별 음원의 정보와 공간에 대한 정보의 정확한 분리가 요구된다. 일반적으로 이를 위해 개별 음원은 별도의 무향 (anechoic) 공간에서 따로 녹음하고, 공간에 대한 정보는 사전에 측정한 정보를 사용한다 [1]. 하지만 이와 같은 방식은 현장감을 담아내는 데에 한계가 있기 때문에, 생생한 현장감을 위해서는 렌더링할 공간에서 음원을 직접 녹음하는 방식을 고려할 수 있다. 이처럼 공간의 특성이 포함된 음원을 녹음하는 경우, 마이크로폰 어레이로 녹음된 신호에서 공간에 의한 잔향을 제거해야 음원에 대한 정보만 따로 저장할 수 있다. 따라서 마이크로폰 어레이로 녹음된 다채널 신호로부터 무향 음원을 추출하는 연구가 필요하다.

다양한 음원이 객체 기반 오디오 시스템을 위해 녹음될 수 있기 때문에 음원의 종류와 무관하게 잔향을 제거할 수 있는 연구가 필요하지만, 대다수의 기법들이 음성 데이터 세트로 훈련 또는 검증되었으므로 본 연구 또한 음성 데이터에 집중하였다.

기존의 잔향 제거 기법들 중에는 신호 처리 이론을 기반으로 한 기법이 오랜 시간 연구되었다. 신호 처리 기법은 수식으로 신호를 모델링하고 신호의 통계적 특성을 이용해 수식을 만족하는 해를 찾는 방법으로, 대표적으로는 범 형성 기반 기법 [2]–[5]이나 선형 예측 기반 기법 [6]이 있다. 이러한 방법들은 잔향 환경에 따라 수동으로 조절해야 하는 파라미터가 많다는 문제가 있다. 그러한 파라미터들에는 신호를 모델링하기 위해 정의하는 파라미터들이 있고, 신호가 모델의 가정과 맞지 않을 경우 발생하는 문제를 해결하기 위해 도입하는 기법의 파라미터들이 있다.

이러한 문제를 해결하기 위해 심층신경망 (deep neural network)을 보조로서 사용하는 기법이 연구되었다 [7]–[14]. 심층신경망은 패턴을 학습하는 데에 강점을 보이기 때문에, 충분히 다양한 환경에서 훈련된다면 어떤 환경이든 적응하여 문제를 해결할 수 있다 [15]. [7]–[14]는 심층신경망의 패턴 학습 능력을 이용하여 환경에 따라 조절되어야 하는 파라

미터 수를 줄이는 연구들이다. 기존의 신호 처리 기법보다 좋은 성능을 보였지만 위의 연구들에서 잔향 제거 연산 자체는 신호 처리 기법과 동일하므로, 신호가 모델과 맞지 않을 경우 발생하는 성능 제약은 여전히 존재한다 [16].

최근에는 도메인 지식의 활용 없이 심층신경망의 훈련만으로 잔향 제거를 수행하는 연구도 진행 중에 있다 [16]–[23]. 심층신경망을 사용할 경우 단일 채널 데이터만으로도 신호 처리 기법에 비해 우수한 성능을 얻을 수 있기 때문에 단일 채널 데이터를 사용한 연구가 많다. 학습 방식은 크게 두 가지로 나뉜다. 하나는 무향 음원의 크기 (magnitude) 스펙트로그램을 직접 추정하는 스펙트럼 대응 방식의 연구이고 [17]–[19], [21]–[23], 다른 하나는 잔향이나 잡음이 포함된 스펙트로그램과 무향 음원의 스펙트로그램의 파워의 비율을 의미하는 마스크를 추정하는 스펙트럼 마스킹 방식의 연구이다 [20].

심층신경망을 사용할 경우 단일 채널 신호만 사용해도 훌륭한 잔향 제거 성능을 얻을 수 있지만, 객체 기반 오디오 시스템을 구현하기 위해서는 보다 완벽한 잔향 제거 기법을 연구할 필요가 있다. 공간상에서 잔향의 에너지가 커서 직접음 대 잔향비 (direct-to-reverberant ratio)가 떨어지면 음장이 확산 음장 (diffuse field)에 가까워진다 [24]. 확산 음장에 가까울수록 서로 떨어진 두 지점에서 녹음되는 신호의 상관성 (correlation)이 떨어지기 때문에 [25], 공간상에 분포된 다수의 마이크로폰 데이터는 음장의 확산도 (diffuseness)를 판단하는 데에 용이하다. 그러므로 다채널 데이터로 잔향의 에너지를 간접적으로 알 수 있고, 심층신경망이 이러한 정보를 이용하면 잔향 제거 성능을 높일 수 있다.

따라서 심층신경망 기반 기법 중 다채널 신호를 이용하는 기법도 연구되었다 [16]. [16]에서는 다채널 신호의 스펙트로그램 데이터를 심층신경망의 입력으로 하여 잔향과 잡음이 없는 신호를 위한 마스크를 추정하고 이를 이용해 무향의 크기 스펙트로그램을 추출한다. 해당 논문에서는 이렇게 심층신경망이 직접 잔향과 잡음을 제거하는 방식이 빔 형성 기법을 심층신경망이 보조하는 방식보다 성능이 우수함을 보였다.

그러나 [16]와 같이 모든 마이크로폰 채널의 데이터를 심층신경망에 입력하면 마이크로폰 채널 수가 많아짐에 따라 입력의 차원이 커지는 것이 문제가 될 수 있다. 기계학습 분야에서 널리 알려진 대로, 입력의 차원이 클수록 과적합의 위험성이 높아진다 [26]. 데

이터 세트가 거대할 경우 차원이 높은 데이터를 추가적인 가공 없이 거대한 심층신경망에 입력하여 좋은 성능을 얻어낼 수 있지만, 잔향을 제거하는 문제에서는 그러한 잔향 데이터 세트가 부재하다. REVERB 챌린지 [27]와 CHiME 챌린지 [28]에 각각 8개, 32개의 room impulse response (이하 RIR) 데이터가 있고, Bar-Ilan 대학교의 데이터 세트 [29]에 78개의 RIR이 있으며, SMIR generator (spherical microphone array impulse response generator) [30]를 이용해 직육면체 방에 대한 RIR을 시뮬레이션하여 합성할 수 있을 뿐이다. 따라서 다양한 환경에 대한 데이터 세트를 구성하기 어려운 상황이며, 이 경우 적은 데이터로 좋은 성능을 얻기 위해 과적합의 발생 확률을 줄일 필요가 있다. 이를 위해서는 다채널 신호를 압축하여 입력의 차원을 줄이는 방법이 있으며, [31]에서 해당 방법을 채택하였다.

[31]는 다채널 신호가 담고 있는 공간 정보를 포함하며 차원을 낮춘 음향 인텐시티와 연관된 두 가지의 입력 특징을 제안하였다. 각각은 방향 벡터 (directional vector, 이하 DV)와 공간 평균 인텐시티 (spatially-averaged intensity vector, 이하 SIV)로 저차와 고차 앰비소닉 신호를 사용해 계산되며, 한 지점 또는 보다 넓은 영역에서의 평균 활성 인텐시티 (mean active intensity)를 나타낸다. 이는 음파가 일정 시간 동안 단위 면적에 전달하는 평균 파워를 의미하며, 음장의 확산도가 높을 수록 인텐시티의 벡터 길이가 짧아지는 특성이 있다 [24]. 따라서 인텐시티와 관련된 특징을 심층신경망의 입력으로 사용하면 음장의 확산도와 유사한 정보를 신경망이 내부적으로 추정하여 잔향 제거에 활용할 가능성이 있기에 인텐시티와 관련된 특징을 사용하였다. 해당 연구에서는 압축을 거치지 않은 특징을 입력으로 훈련한 심층신경망 모델의 결과보다 제안 기법이 뛰어난 성능을 보임을 확인하였다.

그럼에도 불구하고 위의 심층신경망을 이용한 잔향 제거 기법들은 크기 스펙트로그램만을 추정하기 때문에, 왜곡된 위상 정보를 사용하여 복원한 음성 신호의 품질이 떨어지는 문제가 있다. 따라서 크기 스펙트로그램뿐만 아니라 위상도 심층신경망으로 다루는 기법들이 [20], [32]–[39] 연구되고 있으나, 이러한 기법들은 주로 잡음 제거를 스펙트럼 마스킹 방식으로 수행하거나 그러한 잡음 제거 기법들에 추가하는 것으로 성능을 검증한 기법들이다.

위상을 함께 추정하는 기법들을 잔향 제거에 적용할 수 있겠지만 역 국소 푸리에 변환 (inverse short time Fourier transform, 이하 ISTFT)과 같이 시간-주파수 영역에서 시간 영역 신호로 복원하는 과정이 필요하다. 본 연구에서는 영역 간의 전환이 필요 없는 앤드 투엔드 (end-to-end) 형태의 심층신경망 모델을 훈련시켜 잔향 제거를 수행하고자 한다. 입력 특징은 공간 정보를 유지하며 차원을 축소하되 위상 정보를 포함한 시간 영역의 방향 특징인 순간 인텐시티 벡터 (instantaneous intensity vector, 이하 IIV)를 사용한다.

훈련에 사용하는 모델은 Wave-U-Net [40]이다. 모델의 성능 대비 훈련 되어야하는 파라미터 수가 많았기에 성능을 유지하면서 파라미터를 줄이는 방향으로 구조를 변형하였다. 음성을 다루는 앤드투엔드 모델 역시 많은 연구가 되어왔으나 [41]–[45], 모델의 훈련에 오랜 시간이 걸리거나 단일 채널 데이터를 다루기 때문에 본 연구에는 적합하지 않았다. 또한 신호를 일정 시간 길이만큼 잘라 입력으로 사용할 경우 해당 신호 전후의 잔향에 대한 정보를 잃게 되는 문제가 있으므로, 그러한 방식 또한 부적합하다고 판단하였다. [40]의 경우 다채널 신호를 다루고, 음원 분리를 수행하기 위해 제안되었지만 [46]에서 음성의 잡음 제거 역시 성공적으로 수행한 바가 있다. 덧붙여, 모든 레이어가 합성곱 레이어로 구성되었기 때문에 입력의 길이에 제약이 없다는 장점이 있어 선택하였다.

## 1.2 논문의 구성

본 논문은 다음과 같이 구성되었다. 제 2 장에서는 본 연구의 이론적 배경을 다룬다. 음향 인텐시티의 정의, 실수 구면 조화 함수를 기저로 한 구면 조화 영역에서의 음장, Griffin-Lim 알고리즘, 그리고 심층 학습 분야의 지식을 설명한다. 제 3 장에서는 잔향 제거 문제를 정의하고, 기존의 잔향 제거 기법을 정리한다. 기존 기법은 신호 처리 기법과 심층 신경망 기반 기법으로 나뉘며, 두 기법을 모두 서술한다. 제 4 장은 본 연구에서 제안하는 음향 인텐시티와 관련된 방향 특징을 입력으로 하는 심층신경망 기반의 잔향 제거 기법을 설명한다. 방향 특징의 정의, 심층신경망에 사용할 입출력의 전처리, 심층신경망의 구조를 다룬다. 제 5 장에서는 제안 기법의 성능을 확인하기 위한 실험 설정과 결과를 설명한다. 실험은 가상의 공간에서 훈련된 심층신경망을 테스트 데이터 세트로 테스트하였다. 실험 설정으로는 제안 기법의 성능과 비교하는 기법과 데이터 세트 구성, 하이퍼파라미터 설정을

서술한다. 실험 결과에서는 음성 신호의 품질 평가 지표 값을 각 기법들 간에 대해 계산하여 비교하고, 엔드투엔드 모델이 추정한 신호의 위상이 얼마나 유의미한지를 이야기한다. 마지막 제 6 장에서는 제안 기법의 장점과 한계, 그리고 향후 연구 방향을 정리한다.

## 제 2 장 이론적 배경

### 2.1 음향 인텐시티

순간 음향 인텐시티 (instantaneous acoustic intensity)  $\mathbf{I}$ 는 단위 시간 및 단위 면적 당 음파가 전달하는 일의 양이다 [24]. 음파가 입자에 하는 일  $W$ 는 다음과 같이 힘  $\mathbf{F}$ 와 입자의 변위  $\mathbf{s}$ 의 내적의 적분으로 나타난다.

$$W = \int \mathbf{F} \cdot d\mathbf{s} \quad (2.1)$$

일률  $dW/dt$ 은 다음과 같이 압력  $p$ 와 입자 속도  $\mathbf{v}$ 로 나타낼 수 있다.

$$\frac{dW}{dt} = \mathbf{F} \cdot \left( \frac{d\mathbf{s}}{dt} \right) = p \Delta \mathbf{S} \cdot \mathbf{v} \quad (2.2)$$

$\Delta \mathbf{S}$ 는 단위 면적 벡터를 의미하며, 단위 수직 벡터  $\mathbf{n}$ 을 사용하여  $\Delta S \mathbf{n}$ 로 나타내면 단위 면적에 음파가 수직 방향으로 전달하는 일률은 다음과 같다.

$$(dW/dt) / \Delta S = p \mathbf{v} \cdot \mathbf{n} \quad (2.3)$$

따라서 위치 벡터  $\mathbf{r}$ 과 시간  $t$ 에서의 순간 음향 인텐시티  $\mathbf{I}(\mathbf{r}, t)$ 는 아래와 같이 정의된다.

$$\mathbf{I}(\mathbf{r}, t) = p(\mathbf{r}, t) \mathbf{v}(\mathbf{r}, t) \quad (2.4)$$

음압  $p$ 와 입자 속도  $\mathbf{v}$ 가 각각 다음과 같이 복소 진폭 (complex amplitude)  $P$ 와  $\mathbf{V}$ 를

갖는 단일 주파수  $\omega$ 의 음파로 이루어져 있다고 가정한다.

$$P(\mathbf{r}) = \hat{P}(\mathbf{r}) \exp(i\phi_p(\mathbf{r})) \quad (2.5)$$

$$\mathbf{V}(\mathbf{r}) = \hat{\mathbf{V}}(\mathbf{r}) \odot \begin{bmatrix} \exp(i\phi_{v,1}(\mathbf{r})) \\ \exp(i\phi_{v,2}(\mathbf{r})) \\ \exp(i\phi_{v,3}(\mathbf{r})) \end{bmatrix} \quad (2.6)$$

$\odot$ 은 벡터의 원소 곱 (element-wise product)을 의미한다. 이로부터  $p$ 와  $\mathbf{v}$ 는 다음과 같이 나타낼 수 있다.

$$p(\mathbf{r}, t) = \operatorname{Re}\{P(\mathbf{r}) \exp(-i\omega t)\} \quad (2.7)$$

$$\mathbf{v}(\mathbf{r}, t) = \operatorname{Re}\{\mathbf{V}(\mathbf{r}) \exp(-i\omega t)\} \quad (2.8)$$

$\operatorname{Re}\{\cdot\}$ 은 복소수의 실수부를 취하는 연산이다. 이때, 순간 인텐시티는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \mathbf{I}(\mathbf{r}, t) &= \hat{P}(\mathbf{r}) \cos(-\omega t + \phi(\mathbf{r})) \hat{\mathbf{V}}(\mathbf{r}) \odot \begin{bmatrix} \cos(-\omega t + \phi_{v,1}(\mathbf{r})) \\ \cos(-\omega t + \phi_{v,2}(\mathbf{r})) \\ \cos(-\omega t + \phi_{v,3}(\mathbf{r})) \end{bmatrix} \\ &= \operatorname{Re}\left\{\frac{1}{2}P(\mathbf{r})\mathbf{V}(\mathbf{r})^* [1 + \exp[2i(-\omega t + \phi_p(\mathbf{r}))]]\right\} \end{aligned} \quad (2.9)$$

이때,  $\frac{1}{2}P(\mathbf{r})\mathbf{V}(\mathbf{r})^*$ 의 실수부를 평균 액티브 인텐시티, 허수부를 리액티브 인텐시티의 진폭이라고 부른다. 즉, 평균 액티브 인텐시티  $\mathbf{I}_{\text{act}}$ 와 리액티브 인텐시티의 진폭  $\mathbf{I}_{\text{re}}$ 는 다음과 같이 정의된다.

$$\mathbf{I}_{\text{act}}(\mathbf{r}) = \frac{1}{2} \operatorname{Re}\{P(\mathbf{r})\mathbf{V}(\mathbf{r})^*\} \quad (2.10)$$

$$\mathbf{I}_{\text{re}}(\mathbf{r}) = \frac{1}{2} \operatorname{Im}\{P(\mathbf{r})\mathbf{V}(\mathbf{r})^*\} \quad (2.11)$$

$(\cdot)^*$ 는 복소 결례 (complex conjugate)를 의미하고,  $\operatorname{Im}\{\cdot\}$ 은 복소수의 허수부를 취하는 연

산이다. 매질의 밀도  $\rho$ 에 대해 선형 오일러 방정식은  $i\omega\rho\mathbf{v} = \nabla p$ 로 나타난다 [47]. 오일러 방정식을 이용하여  $\mathbf{V}$ 를  $P$ 에 대한 식으로 나타낸 후, 이를 식 2.9에 대입하면 순간 인텐시티를 다음과 같이  $\mathbf{I}_{\text{act}}$ 와  $\mathbf{I}_{\text{re}}$ 에 대해 나타낼 수 있다.

$$\mathbf{I}(\mathbf{r}, t) = \mathbf{I}_{\text{act}}(\mathbf{r}) [1 + \cos[2(-\omega t + \phi_p(\mathbf{r}))]] + \mathbf{I}_{\text{re}}(\mathbf{r}) \sin[2(-\omega t + \phi_p(\mathbf{r}))] \quad (2.12)$$

위의 식으로부터 순간 인텐시티의 시간 평균이  $\langle \mathbf{I}(\mathbf{r}) \rangle = \mathbf{I}_{\text{act}}(\mathbf{r})$ 임을 알 수 있다. 따라서 일정 시간 동안 음파가 단위 면적 및 단위 시간 당 전달하는 평균 에너지는 평균 액티브 인텐시티와 같고, 리액티브 인텐시티는 전달되는 평균 에너지에 기여하지 않는다.

다음으로 음장이 확산 음장 (diffuse field)일 때를 가정해보자. 확산 음장은 모든 방향에서 평면파가 관측될 확률이 동일한 음장을 의미한다 [48]. 확산 음장에서는 모든 위치  $\mathbf{r}$ 마다  $\phi_p(\mathbf{r}) - \phi_{v,1}(\mathbf{r})$ ,  $\phi_p(\mathbf{r}) - \phi_{v,2}(\mathbf{r})$ ,  $\phi_p(\mathbf{r}) - \phi_{v,3}(\mathbf{r})$ 의 값의 확률이 균일한 분포 (uniform distribution)를 갖는다 [24]. 따라서  $P\mathbf{V}^*$ 의 공간 평균은 확산 음장에서 다음과 같이 나타난다.

$$\begin{aligned} \mathbb{E}_{\mathbf{r}} \{ P(\mathbf{r}) \mathbf{V}^*(\mathbf{r}) \} &= \mathbb{E}_{\mathbf{r}} \left\{ \hat{P}(\mathbf{r}) \hat{\mathbf{V}}(\mathbf{r}) \right\} \odot \begin{bmatrix} \exp(i\mathbb{E}_{\mathbf{r}} \{\phi_p(\mathbf{r}) - \phi_{v,1}(\mathbf{r})\}) \\ \exp(i\mathbb{E}_{\mathbf{r}} \{\phi_p(\mathbf{r}) - \phi_{v,2}(\mathbf{r})\}) \\ \exp(i\mathbb{E}_{\mathbf{r}} \{\phi_p(\mathbf{r}) - \phi_{v,3}(\mathbf{r})\}) \end{bmatrix} \\ &= \mathbf{0} \end{aligned} \quad (2.13)$$

즉, 음장이 확산 음장에 가까울수록 평균 활성 인텐시티 벡터의 길이는 0에 가까워진다.

## 2.2 실수 구면 조화 영역 신호

구면 좌표계에서 나타낸 위치벡터  $\mathbf{r} = [r \ \theta \ \phi]^T$ 에 대해, 실수 구면 조화 함수 (real spherical harmonics)는 다음과 같이 정의된다 [49].

$$Y_{nm}(\theta, \phi) = \mathcal{N}_{n|m|} \mathcal{P}_{n|m|}(\cos \theta) \times \begin{cases} \sqrt{2} \sin(|m|\phi) & m < 0 \\ 1 & m = 0 \\ \sqrt{2} \cos(m\phi) & m > 0 \end{cases} \quad (2.14)$$

$$\mathcal{N}_{n|m|} = \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} \quad (2.15)$$

$\mathcal{P}_{nm}$ 은 베금 르장드르 함수 (associated Legendre polynomial)  $\mathcal{P}_n^m$ 에 Condon-Shortley 위상인  $(-1)^m$ 을 곱한 함수이고,  $\mathcal{N}_{nm}$ 은 정규화 상수이다 [50].  $n$ 은 구면 조화 함수의 차수,  $m$ 은 각도이며, 각각  $n \geq 0$ 과  $-n \leq m \leq n$ 를 만족한다. 음장  $p(\mathbf{r}, t)$ 를  $Y_{nm}$ 의 계수인 구면 조화 영역 (spherical harmonic domain, 이하 SHD) 신호  $p_{nm}$ 으로 변환하는 과정을 구형 푸리에 변환(spherical Fourier transform, 이하 SFT)이라 하며, 역 과정인 구형 푸리에 변환 (inverse spherical Fourier transform, 이하 ISFT)와 함께 아래와 같이 정의된다.

$$p_{nm}(t) = \int_0^{2\pi} \int_0^\pi p(\mathbf{r}, t) Y_{nm}(\theta, \phi) \sin \theta d\theta d\phi \quad (2.16)$$

$$p(\mathbf{r}, t) = \sum_{n=0}^{\infty} \sum_{m=-n}^n p_{nm}(t) Y_{nm}(\theta, \phi) \quad (2.17)$$

$p_{nm}(t)$ 는 다음과 같이 마이크로폰 어레이의 모드 강도 (modal strength)  $b_n(kr)$ 의 역 푸리에 변환 (inverse Fourier transform, 이하 IFT)과 거리  $r$ 에 독립적인 MC-SHD(mode-compensated spherical harmonic domain) 신호  $a_{nm}(t)$ 의 합성곱으로 나타낼 수 있다.

$$p_{nm}(t) = \text{IFT}\{b_n(kr)\} * a_{nm}(t) \quad (2.18)$$

$k$ 는 파수 (wavenumber)이며, 음속  $c$ 에 대해  $k = \omega/c$ 로 정의된다. 음향학적으로 투명하여 음장에 영향을 미치지 않는 구형 마이크로폰 어레이를 사용할 때의 모드 강도는 제1종 구면

Bessel 함수  $j_n(kr)$ 에 대해 다음과 같이 표현된다 [51].

$$b_n(kr) = 4\pi i^n j_n(kr) \quad (2.19)$$

반지름  $r_a$ 인 강체 구의 표면에 마이크로폰이 위치한 어레이를 사용할 때의 모드 강도는 제2종 구면 Hankel 함수  $h_n^{(2)}(kr)$ 과 그 도함수  $h_n^{(2)'}(kr)$ , 그리고  $j_n(kr)$ 의 도함수  $j'_n(kr)$ 에 대해 다음과 같이 표현된다 [51].

$$b_n(kr) = 4\pi i^n \left[ j_n(kr) - j'_n(kr_a) \frac{h_n^{(2)}(kr)}{h_n^{(2)'}(kr_a)} \right] \quad (2.20)$$

### 2.3 Griffin-Lim 알고리즘

국소 푸리에 변환 (short time Fourier transform, 이하 STFT)은 오디오 신호의 시간-주파수의 구조를 잘 나타내는 장점으로 인해 널리 사용된다. 또한 선형성과 가역성을 지닌다 [52]. 그러나 STFT는 신호의 일정 구간이 겹치게 하여 시간 프레임을 나눈 후, 프레임 별로 푸리에 변환을 수행하기 때문에 불필요하게 중복되는 정보를 포함한다는 문제가 있다. 따라서 STFT 스펙트로그램을 변형할 경우 해당 스펙트로그램과 대응하는 시간 영역의 신호가 존재하지 않을 수 있다 [52]. 이 문제를 일관성 문제 (consistency problem)라 하며, 이를 만족하는 스펙트로그램을 일관성 있는 스펙트로그램 (consistent spectrogram)이라 한다. 스펙트로그램은 복소수이며 크기와 위상으로 분리할 수 있다. 크기 스펙트로그램은 해석이 가능한 형태를 띠지만, 위상은 구간  $[-\pi, \pi]$ 의 값을 갖는 균일한 분포의 잡음으로 나타난다 [53], [54]. 따라서 오디오 신호 처리 분야의 여러 알고리즘은 원하는 크기 스펙트로그램은 획득하지만 위상 스펙트로그램을 다루지 못하기 때문에 앞서 말한 일관성 문제가 발생한다.

Griffin-Lim 알고리즘은 복원한 시간 영역의 신호가 고정된 크기 스펙트로그램을 갖도록 STFT와 ISTFT를 반복하여 일관성 문제를 해결하고자 한다 [55]. 스펙트로그램을 일관성 있는 스펙트로그램의 집합으로 사영 (projection)하고, 다시 주어진 크기 스펙트로그램의 집합에 사영하는 과정을 번갈아가며 여러 번 반복한다. 아래는 Griffin-Lim 알

고리즘을 나타낸 것이다.  $t$ 는 시간 샘플 인덱스,  $\tau$ 는 시간 프레임 인덱스,  $f$ 는 주파수 빈 인덱스를 의미하며, 고정된 크기 스펙트로그램은  $M(\tau, f)$ , 초기 위상 스펙트로그램은  $\phi_0(\tau, f)$ 로 나타내었다.

---

**Algorithm 1** Griffin-Lim 알고리즘

---

```

1: Initialization:  $n \leftarrow 0$ ,  $\phi_n(\tau, f) \leftarrow \phi_0(\tau, f)$ 

2: while  $n < N_{\text{GL}}$  do

3:    $P(\tau, f) \leftarrow M(\tau, f) \exp(i\phi_n(\tau, f))$ 

4:    $p(t) \leftarrow \text{ISTFT}\{P(\tau, f)\}$ 

5:    $\hat{P}(\tau, f) \leftarrow \text{STFT}\{p(t)\}$ 

6:    $\phi_{n+1}(\tau, f) \leftarrow \angle \hat{P}(\tau, f)$ 

7:    $n \leftarrow n + 1$ 

8: end while

9: Output:  $p(t)$ 

```

---

위의 알고리즘은 스펙트로그램이 적당한 신호로 수렴할 때까지 두 집합 사이에서 사영을 반복하는 방식이기 때문에 위상의 초기값  $\phi_0$ 에 따라 결과가 달라질 수 있다. 그러므로 원하는 시간 영역 신호와 거리가 먼 위상을 초기값으로 사용할수록 해당 신호에 수렴하지 않을 확률이 높아지며, 복원한 신호의 품질 역시 나빠질 수 있다.

## 2.4 심층 학습

심층 학습 (deep learning)은 기계 학습 (machine learning)의 부분집합으로, 심층신경망을 많은 양의 데이터로 학습시켜 주어진 일을 수행하도록 하는 알고리즘을 일컫는다 [56]. 기계 학습은 시스템이 입출력 데이터로부터 특정한 패턴을 추출하는 방법을 스스로 터득하여, 둘 사이의 관계를 설명하는 모델을 만들어 내는 방식이다. 그러므로 시스템의 학습을 위한 데이터가 필요하다. 모든 경우에 대한 입출력 데이터를 훈련에 사용할 수 있다면 좋으나 이는 자원의 한계로 불가능하며, 또한 훈련에 사용되지 않은 입력 데이터에

대해 잘못된 출력을 내놓을 가능성이 있다. 따라서 충분한 양과 편향되지 않은 분포를 가진 훈련 데이터 세트를 학습에 사용하고, 훈련에 사용하지 않은 테스트 데이터 세트로 모델의 일반화 성능을 확인해야 한다.

알고리즘이 올바른 모델을 학습하기 위해서는 어느 정도의 가이드라인이 필요하다 [56], [57]. 이는 사람이 직접 정하는 것으로, 대표적으로는 모델의 용량 (capacity) 또는 복잡도 (complexity)가 있다. 모델이 학습하고자 하는 입출력 데이터 간의 관계에 비해 모델의 용량이 작다면, 관계를 지나치게 단순히 설명하는 모델이 만들어진다. 이 경우를 과소 적합 (underfitting)이라 하며 훈련 데이터 세트에서조차도 작지 않은 오류를 발생시킨다. 반면에 모델의 용량을 높이면 훈련 데이터 세트에 대해 완벽한 성능을 보이는 모델을 학습하게 된다. 하지만 해당 경우에는 훈련 데이터 세트가 충분히 크지 않거나 편향된 분포를 가졌을 시, 이전에 학습하지 않은 새로운 데이터에 대해 큰 오류가 발생하게 되며 이를 과적합 (overfitting)이라고 한다. 그러므로 입출력 데이터의 관계의 복잡도를 적절히 설명할 수 있을 만큼의 적당한 용량을 설정해야 한다.

기계 학습 알고리즘은 크게 세 가지 갈래로 나뉜다 [58]. 첫 번째는 지도 학습 (supervised learning)으로, 입출력 데이터가 쌍으로 주어져 있을 때 둘 사이의 관계를 알고리즘이 학습하는 방식이다. 입력이 무엇을 나타내는지를 판단하는 분류 문제와, 주어진 데이터가 어떤 함수로 이뤄졌는가를 예측하는 회귀 문제가 대표적이다. 두 번째는 비지도 학습 (unsupervised learning)으로, 입력 데이터로부터 정답 출력 없이 특정한 패턴을 학습하는 방식이다. 유사한 데이터를 그룹으로 묶는 클러스터링 (clustering)이나 입력 데이터의 분포를 추정하는 밀도 추정 (density estimation) 등이 대표적이다. 마지막은 강화 학습 (reinforcement learning)이다. 주어진 환경에서 보상을 최대화하기 위해 어떤 행동이 적절한가를 학습하는 방식이다. 게임과 같이 환경과 알고리즘이 상호작용하며 전략을 학습해야 하는 문제에 많이 사용된다. 본 연구의 주제인 잔향 제거는 잔향이 포함된 데이터와 잔향이 없는 데이터 사이의 대응 관계를 알고리즘이 학습하므로 지도 학습 방식에 해당한다. 따라서 본 절에서는 지도 학습을 중심으로 설명한다.

### 2.4.1 인공신경망 및 심층신경망

인공신경망 (artificial neural network)은 복잡한 문제를 풀기 위해 생물학적 신경망의 기능성을 활용한 모델링 툴이다 [59]. 신경세포는 주변의 신경세포로부터 전기 신호를 받아 들여 역치를 넘어섰을 때 또 다른 세포로 전기 신호를 전송한다. 이를 수학적으로 모방한 것이 퍼셉트론 (perceptron)이다 [60]. 입력 신호  $\mathbf{x}$ , 신호의 세기를 나타내는 가중치  $\mathbf{w}$ , 역치를 나타내는 바이어스 (bias)  $b$ , 그리고 활성함수 (activation function)를 나타내는  $a$ 를 사용하여 다음과 같이 퍼셉트론의 작동 방식을 표현할 수 있다.

$$z = a(\mathbf{w}^T \mathbf{x} + b) \quad (2.21)$$

$a$ 는 어떤 비선형함수든 가능하며, 대표적으로 sigmoid ( $\sigma(x) = 1 / (1 + e^{-x})$ )와 ReLU (rectified linear unit,  $\text{ReLU}(x) = \max(x, 0)$ )가 있다. 여러 개의 퍼셉트론이 하나의 입력  $\mathbf{x}$ 를 받아 한 층을 이룰 때, 이를 퍼셉트론 레이어 (perceptron layer) 또는 전결합 레이어 (fully-connected layer)라고 부른다. 다음과 같이 입력  $\mathbf{x}$ 의 모든 원소가 가중치  $\mathbf{W}$ 를 통해 출력  $\mathbf{z}$ 의 모든 원소와 연결되는 형태이기 때문이다.

$$\mathbf{z} = a(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad (2.22)$$

$\mathbf{W}$ 와  $\mathbf{b}$ 는 각 퍼셉트론의  $\mathbf{w}^T$ 와  $b$ 를 연결 (concatenate)한 것이다. 이러한 전결합 레이어를 쌓으면 여러 개의 층을 가진 인공신경망이 되며, 이를 MLP (multi-layer perceptron)라고 부른다. 그림 2.1에 간단한 예시가 나타나 있다.

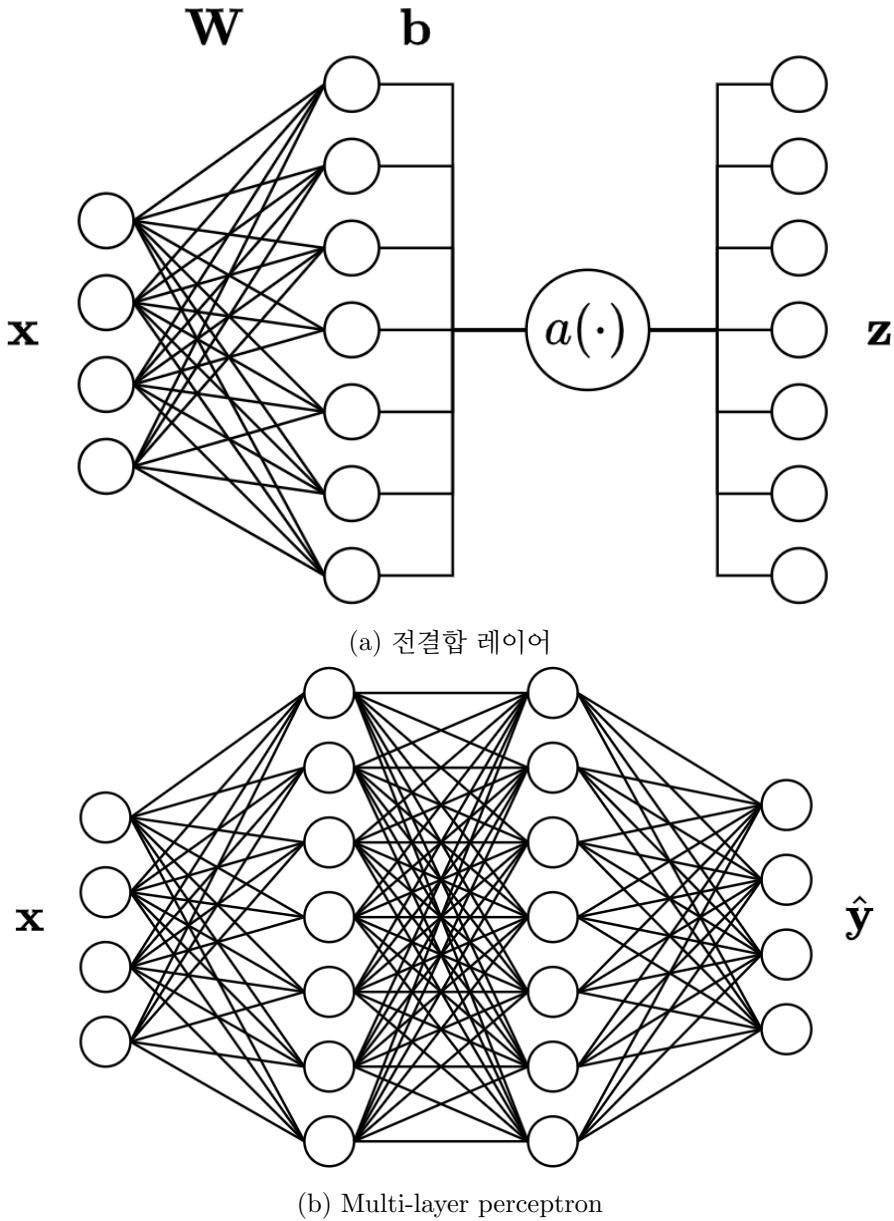


그림 2.1: 전결합 레이어와 MLP의 도식화

인공신경망을 학습시키는 방법은 경사 하강법 (gradient descent)으로, 모델의 파라미터에 대한 손실함수  $\mathcal{L}$ 의 미분 값인 경사의 반대 방향으로 파라미터를 업데이트하는 방식이다 [61], [62]. 식 2.22를 예로 들면 다음과 같다.

$$\mathbf{W}^{(n+1)} \leftarrow \mathbf{W}^{(n)} - \eta \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial \mathbf{W}} \quad (2.23)$$

$$\mathbf{b}^{(n+1)} \leftarrow \mathbf{b}^{(n)} - \eta \frac{\partial \mathcal{L}(\mathbf{W}, \mathbf{b})}{\partial \mathbf{b}} \quad (2.24)$$

손실 함수  $\mathcal{L}$ 은 MLP의 출력  $\hat{y}$ 와 정답 출력  $y$  간의 차이를 의미하는 함수로, 평균 제곱 오차 (mean squared error), 크로스 엔트로피 (cross entropy) 등이 대표적이다.  $\eta$ 는 학습률 (learning rate)로 모델의 파라미터를 얼마나 업데이트할 것인가를 결정한다. 경사 하강법은 손실 함수 값이 작아지는 방향으로 학습을 진행하기 때문에 전역 최솟값에 가까워질 가능성을 높일 수 있으나, 심층신경망의 손실 함수는 일반적으로 볼록하지 않으며 (non-convex), 따라서 경사가 0이 되는 값에 도달하여도 그 값이 전역 최솟값이라는 보장은 없다. 그러나 이러한 방법으로 적절한 모델을 얻을 수 있다는 경험적 근거가 다양한 분야에 축적되어 널리 사용되고 있다. 또한 모델의 효율적인 학습을 위하여 경사 하강법 기반의 변형 기법들이 연구되어 왔으며 [63], 최근에는 Adam [64]과 AdamW [65]가 우수한 최적화 성능을 보여 많이 사용되고 있다. 최적화 방법과 더불어, 학습률 역시 모델의 성능에 영향을 끼치기 때문에 이를 다루는 다양한 기법들이 연구되었다 [66], [67].

두 개의 전결합 레이어를 포함하는 MLP는 각 레이어의 퍼셉트론 수를 충분히 크게 설정하면 입력과 출력 사이의 어떠한 관계든 학습할 수 있음이 증명되었다 [68]. 그러나 어떠한 관계든 학습할 수 있는 모델은 과적합에 취약하기 때문에 전결합 레이어가 아닌 다른 레이어를 사용하거나, 적은 수의 퍼셉트론을 포함하는 레이어를 사용하는 대신 더 많은 수의 레이어를 사용하는 인공신경망이 연구되었다. 그러한 인공신경망을 심층신경망이라고 한다.

#### 2.4.2 합성곱 신경망

합성곱 신경망(Convolutional Neural Network, 이하 CNN)은 널리 사용되는 심층신경망 중 한 종류이다 [69]–[72]. 합성곱 신경망은 합성곱 레이어 (convolution layer)로 구성된 신경망을 의미하며, 합성곱 레이어는 이미지로부터 신경망에 넣을 특징을 추출하기 위한 특징 추출기 (feature extractor)로 사용하기 위해 제안되었다. 그림 2.2에 간단한 합성곱 신경망의 도식이 나타나 있다. MLP를 특징 추출기로 사용하게 되면 몇 가지의 문제가 발생한다. 이미지의 많은 수의 픽셀을 입력으로 다루기 위해 많은 수의 뉴런이 요구되어 과적합으로 이어질 가능성이 크고, 이미지의 변형에 효율적으로 대처하기 힘든 구조를 가지고 있으며, 가까운 픽셀들 간의 상관관계가 강하다는 구조상의 특징을 반영하지 못한다는

점이다. 합성곱 신경망은 위의 문제를 해결하는 데에 탁월함을 보이며, 세 가지의 메커니즘을 전제로 한다. 국소 수용 영역 (local receptive field), 가중치 공유 (weight sharing), 폴링 (pooling)이 그것이다. 레이어에 사용할 뉴런의 수를 줄이고 뉴런을 이차원으로 배치 함으로써 이미지의 국소 영역에 대해 특징을 추출하도록 한다. 이 뉴런들을 커널 (kernel)이라 하며 커널을 상하좌우로 이동하며 국소 영역의 특징을 추출한다. 커널의 가중치는 이동 시마다 달라지는 게 아니라 고정된 값을 가지며 이로 인해 훈련이 필요한 파라미터의 수가 줄어드는 결과를 가져온다. 고정된 가중치는 모든 국소 영역에서 동일한 특징을 추출한다는 의미를 지닌다. 여러 개의 커널을 사용한다면 커널 별로 서로 다른 특징을 추출할 수 있으며, 추출된 결과를 특징맵 (feature map)이라 하고, 생성된 특징맵 개수만큼의 채널 (channel)을 갖는다.

합성곱 레이어가  $S \times S$  사이즈의 커널  $\mathbf{W}$ 을 갖는다면 다음의 연산을 수행한다 [69], [72].

$$\mathbf{z}(s_1, s_2) = a \left( \sum_{s'_1=0}^{S-1} \sum_{s'_2=0}^{S-1} [\mathbf{W}(s'_1, s'_2) \mathbf{x}(s_1 - s'_1, s_2 - s'_2)] + \mathbf{b} \right) \quad (2.25)$$

$s_1, s_2$ 는 특징맵의 가로축과 세로축에 대한 인덱스이고,  $\mathbf{x}(s_1, s_2), \mathbf{z}(s_1, s_2)$ 의 원소는 채널을 의미한다. 위의 식은 전결합 레이어의 연산을 나타내는 식 2.22와 유사함을 확인할 수 있다. 이로부터 합성곱 레이어는 특징맵의 가로축과 세로축에 대해서는 합성곱을, 채널 축에 대해서는 전결합 레이어와 동일한 연산을 수행함을 알 수 있다.

풀링 레이어는 합성곱 레이어 다음에 주로 위치하며, 이미지 또는 특징맵을  $2 \times 2$ 와 같은 작은 구역으로 나누고 구역별로 정해진 규칙에 따라 하나의 값을 출력하여 전체 사이즈를 줄이는 역할을 한다. 대표적인 규칙은 평균값을 취하거나 최댓값을 취하는 것이며, 각각을 평균 풀링과 최대 풀링이라고 한다. 인접한 영역의 데이터는 비슷한 정보를 담고 있기 때문에, 대푯값이나 두드러진 정보만을 추출하여 신경망이 효율적인 학습을 할 수 있도록 한다.

전치 합성곱 레이어(transposed convolution layer)는 풀링 레이어와 반대로 이미지의 사이즈를 키워야할 때 사용하기 위해 제안되었다 [73]. 전치 합성곱 레이어는 입력 이미지를 업샘플링 (upsampling)하고 비선형 함수를 적용하는 레이어이다. 합성곱 레이어와

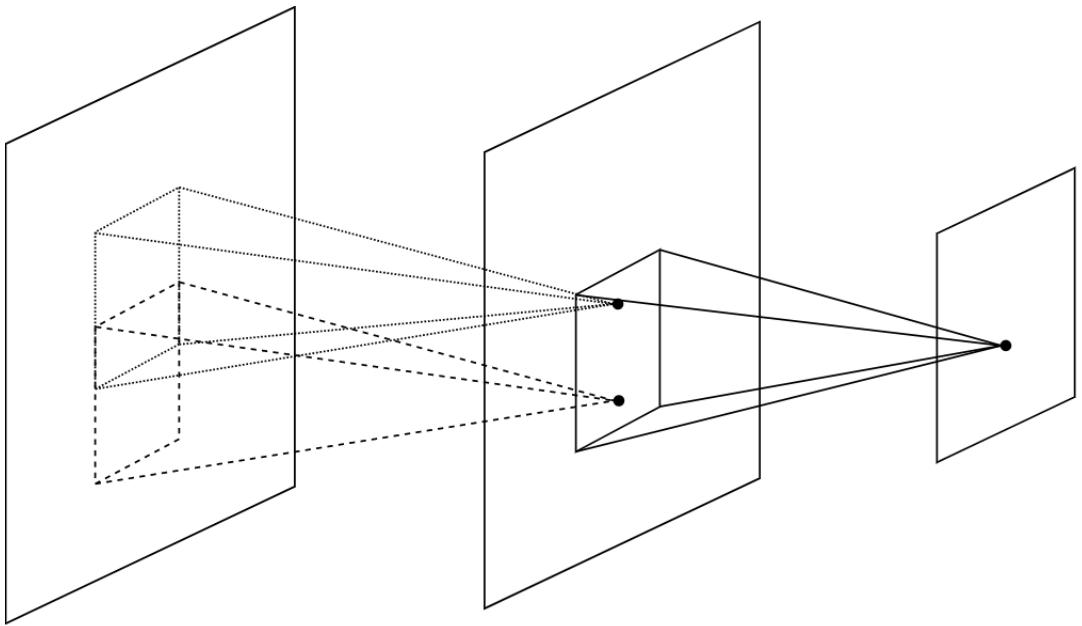


그림 2.2: 합성곱 신경망의 도식화. 왼쪽부터 입력 이미지, 합성곱 레이어, 풀링 레이어를 나타낸다 [70].

마찬가지로 전치 합성곱 레이어에 사용할 커널을 학습시키거나, 중요한 특징을 추출이 목적인 풀링 레이어와는 달리 높은 해상도의 이미지를 얻기 위한 목적으로 사용한다.

위의 설명은 이차원 입력을 중심으로 서술하였으나, 음성 신호와 같은 일차원 입력에 대해서도 같은 작용을 한다. 본 연구에서는 일차원 입력을 사용하기 때문에 사용하는 심층신경망 역시 일차원 합성곱 레이어로 구성되어있다.

## 제 3 장 기존의 잔향 제거 기법

### 3.1 문제 정의

잔향 제거 문제란, 잔향이 존재하는 환경에서 녹음된 신호가 잔향이 없는 환경에서는 어떻게 녹음될지를 추정하는 문제다. 음원  $s(t)$ 로부터 거리가  $R$ 만큼 떨어진 곳에 위치한 무지향성의 마이크로폰으로 녹음한 압력 신호  $p_a(t)$ 는 다음과 같이 나타난다.

$$p_a(t) = \frac{1}{R} s(t - t_0) \quad (3.1)$$

이때  $t_0$ 는 신호가 거리  $R$ 만큼 전파되는 데에 걸린 시간이다. 즉, 음원이 전달 거리만큼 감쇠하고 전달 시간만큼 지연된 형태이다. 한 편, 잔향이 있는 환경에서 무지향성 마이크로폰으로 녹음한 압력 신호는  $p_r(t)$ 는 다음과 같이 RIR  $h(t)$ 와 음원의 합성곱으로 나타난다.

$$p_r(t) = h(t) * s(t) \quad (3.2)$$

$N$ 개 채널의 마이크로폰 어레이로 녹음된 신호 벡터  $\mathbf{p}_r(t)$ 는 식 3.5와 같이 각 마이크로폰과 음원 간의 RIR로 이뤄진 벡터  $\mathbf{h}(t)$ 와 음원의 합성곱으로 표현할 수 있다.

$$\mathbf{p}_r(t) = [p_{r,1}(t) \ p_{r,2}(t) \ \cdots \ p_{r,N}(t)]^T \quad (3.3)$$

$$= [h_1(t) \ h_2(t) \ \cdots \ h_N(t)]^T * s(t) \quad (3.4)$$

$$= \mathbf{h}(t) * s(t) \quad (3.5)$$

본 논문과 같이 시간 영역의 잔향 제거 기법은, 단일 채널의 경우  $p_r(t)$ 로부터  $p_a(t)$ 의 추정 값인  $\hat{p}_a(t)$ 를 찾고, 다채널의 경우  $\mathbf{p}_r(t)$ 로부터  $\hat{p}_a(t)$ 를 찾는다.

다른 방식으로는, 시간과 주파수 영역을 동시에 표현한 스펙트로그램을 사용하여 잔향

제거 문제에 접근할 수 있다. 해당 경우는 아래의 식들을 이용하여 문제를 정의한다.

$$P_a(\tau, f) = \text{STFT} \{p_a(t)\} = M_a(\tau, f) \exp(i\phi_a(\tau, f)) \quad (3.6)$$

$$P_r(\tau, f) = \text{STFT} \{p_r(t)\} = M_r(\tau, f) \exp(i\phi_r(\tau, f)) \quad (3.7)$$

$$\mathbf{P}_r(\tau, f) = [P_{r,1}(\tau, f) \ P_{r,2}(\tau, f) \ \cdots \ P_{r,N}(\tau, f)]^T \quad (3.8)$$

위의 식들은 각 신호를 STFT한 것이며  $M_a, M_r$ 은 크기 스펙트로그램을,  $\phi_a, \phi_r$ 은 위상을 의미한다. 단일 채널 신호를 이용한 잔향 제거 기법은  $P_r(\tau, f)$ 로부터  $P_a(\tau, f)$ 의 추정 값인  $\hat{P}_a(\tau, f)$ 를 찾고, 다채널의 경우  $\mathbf{P}_r(\tau, f)$ 로부터  $\hat{P}_a(\tau, f)$ 를 찾는다. 이후  $\hat{P}_a(\tau, f)$ 를 ISTFT하여  $\hat{p}_a(t)$ 를 구한다.

## 3.2 신호 처리 기법

신호 처리 기반의 잔향 제거 기법은 크게 빔 형성 기법 [2]–[5]와 선형 예측 기법 [6]으로 나뉜다.

### 3.2.1 빔 형성 기법

빔 형성 기법은 마이크로폰 어레이로 녹음한 신호에 필터를 적용한 후, 이를 합하여 특정 방향에서 온 신호의 이득을 최대화하고 이외의 방향에서 온 신호의 이득을 줄이는 방식이다 [74]. 기초적인 방식 중 하나는 Delay-and-Sum (이하 DAS)이다. 특정 방향에서 입사하는 평면파가 각 마이크로폰에 도달하는 시간 차이를 평면파 전파 모델로 계산하고, 각 신호의 시간차를 보상하여 더하는 기법이다. 그림 3.1에 해당 기법이 나타나 있다.  $\mathbf{w}(\theta, \omega)$ 는 방향  $\theta$ 와 주파수  $\omega$ 에 대한 필터 또는 스캔 벡터 (scan vector)이며, 출력 신호는  $b(\theta, \omega) = \mathbf{w}(\theta, \omega)^H \mathbf{p}(\omega)$ 로 나타난다.

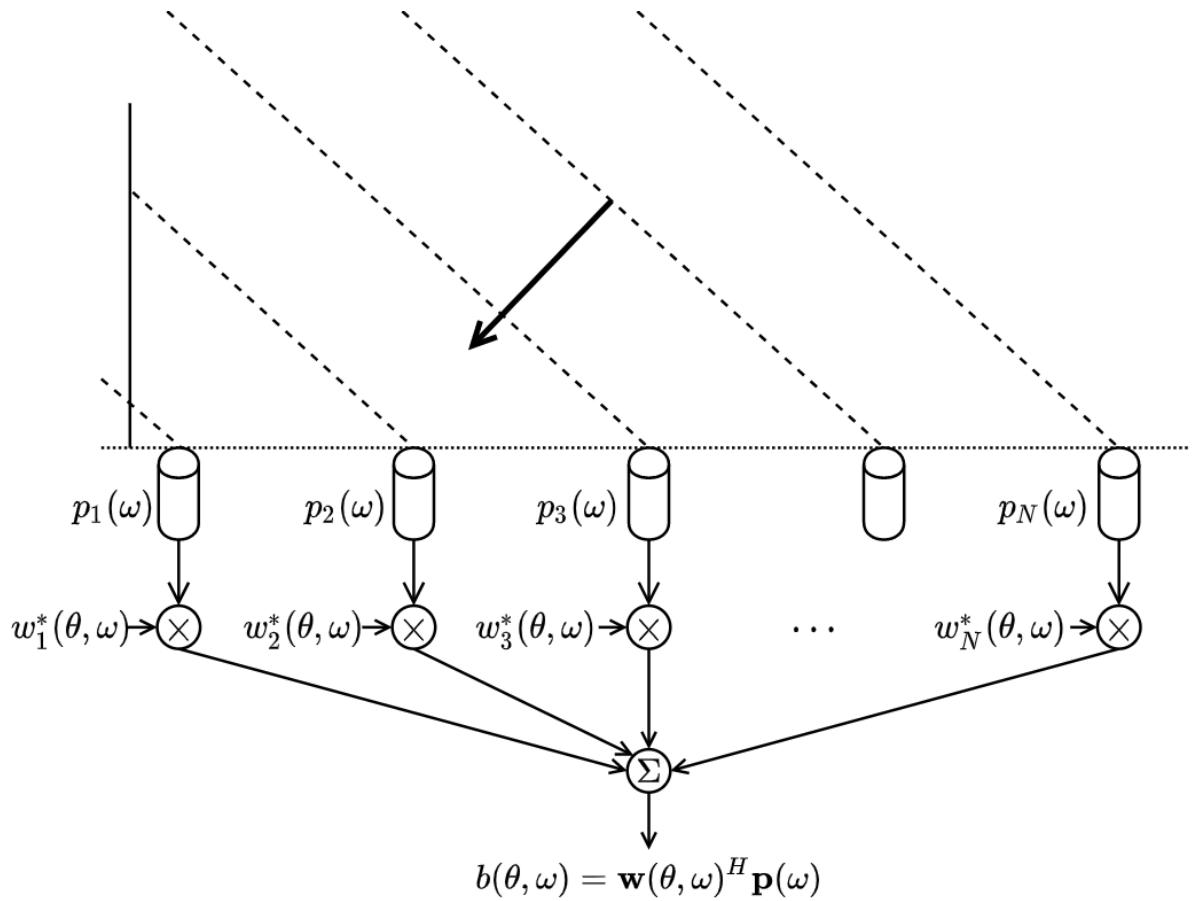


그림 3.1: DAS 빔 형성 기법 [74]

많은 연구에서 사용되는 빔 형성 기법 중 하나는 MVDR (minimum variance distortionless response)로 전파 모델에 따라 필터를 구하는 DAS와 달리, 신호를 기반으로 하기 때문에 빔 패턴이 측정된 신호에 따라 적응하여 변하고 지향성이 크다는 특징이 있다 [75]. MVDR의 스캔 벡터를 구하는 과정은 다음과 같다 [75]. 빔 형성의 출력의 제곱에 대한 기댓값을 빔 파워  $\beta$ 라고 하며 식 3.10과 같다.

$$\beta(\theta, \omega) = \mathbb{E} [|b(\theta, \omega)|^2] \quad (3.9)$$

$$= \mathbf{w}(\theta, \omega)^H \mathbf{R}(\omega) \mathbf{w}(\theta, \omega), \text{ where } \mathbf{R}(\omega) = \mathbb{E} [\mathbf{p}(\omega) \mathbf{p}(\omega)^H] \quad (3.10)$$

MVDR의 목적은 특정 방향  $\theta$ 에 대한 신호의 응답을 유지하되 다른 방향으로부터의 신호의 응답을 최소화하는 것이다. 즉, 빔 파워  $\beta$ 를 최소화해야 하므로 다음과 같이 문제를 정의할

수 있다 [75].

$$\min \beta(\theta) = \mathbf{w}(\theta)^H \mathbf{R} \mathbf{w}(\theta), \text{ subject to } \mathbf{w}(\theta)^H \mathbf{h}(\theta) = 1 \quad (3.11)$$

이때,  $\mathbf{h}(\theta)$ 는 방향  $\theta$ 에 대한 평면파 모델의 전달 함수이다. 위의 문제를 라그랑주 승수법 (Lagrange multiplier method)을 이용해 다음과 같은 최적화 문제로 변환할 수 있다.

$$\min \mathcal{L}(\mathbf{w}, \mu) = \mathbf{w}^H \mathbf{R} \mathbf{w} + \mu(\mathbf{w}^H \mathbf{h} - 1) \quad (3.12)$$

이때  $\mu$ 는 라그랑주 승수이다. 함수  $\mathcal{L}$ 을 최소화하는  $\mathbf{w}$ 와  $\mu$ 를 구하기 위해 각각에 대해 편미분하여 0이 되는 지점을 찾는다.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{R} \mathbf{w} + \mu \mathbf{h} = 0 \Rightarrow \mathbf{w} = -\mu \mathbf{R}^{-1} \mathbf{h} \quad (3.13)$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \mathbf{w}^H \mathbf{h} - 1 = 0 \Rightarrow -\mu^* (\mathbf{h}^H \mathbf{R}^{-1} \mathbf{h}) = 1 \quad (3.14)$$

식 3.13과 3.14로부터 다음과 같이 MVDR의 스캔 벡터를 구할 수 있다.

$$\mathbf{w}(\theta) = \frac{\mathbf{R}^{-1} \mathbf{h}(\theta)}{\mathbf{h}(\theta)^H \mathbf{R}^{-1} \mathbf{h}(\theta)} \quad (3.15)$$

그러나 대부분의 빔 형성 기법은 얻고자 하는 신호와 그 외의 신호 사이에 상관 (correlation)이 없음을 가정하기 때문에 [75], 음원과 상관이 높은 잔향을 제거하는 데에는 성능에 한계가 있다. 이를 보완할 수 있는 diagonal loading 기법 [76]과 공간 스무딩 기법 [77]이 있으나, 환경에 따라 조절해야 하는 파라미터의 수가 많아지는 단점이 존재한다. 덧붙여, 빔 형성 기법은 직접파와 동일한 방향에서 오는 반사파의 영향을 줄일 수 없다는 문제도 있다.

### 3.2.2 선형 예측 기법

선형 예측 기법 중 가장 널리 사용되는 것은 WPE (weighted prediction error)이다 [6]. 다르게는 NDLP (variance-normalized delayed linear prediction)이라고도 불린다. WPE에서는 현재 신호가 이전에 한 번 직접파로 녹음된 적이 있는 신호가 감쇠되어 잔향으로

녹음된 사실로부터 다음과 같이 신호를 나타낸다.

$$p_{r,m}(t) = \sum_{m'=1}^{N_m} \sum_{l=0}^{N_l-1} c_{m,m'}(l) p_{r,m'}(t - t_D - l) + p_{e,m}(t) \quad (3.16)$$

이때  $p_{e,m}$ 은 시간 샘플  $t_D$ 개 보다 이전에 녹음된  $N_l$ 개의 시간 샘플의 가중합 (weighted sum)으로 나타낼 수 없는 신호로, 직접파와 초기 반사를 포함한 것이다. 예측 지연 시간  $t_D$ 와 선형 예측 필터  $c_{m,m'}$ 의 길이  $N_l$ 은 환경에 따라 조절해야 하는 파라미터이다. 이상적으로,  $t_D$ 는 직접파가 도달한 후 후기 잔향 (late reverberation)이 시작되기 전까지의 시간 샘플 수로 설정할 수 있고,  $N_l$ 은 후기 잔향의 길이로 설정할 수 있다. 이를 벡터 형태로 나타내기 위해 다음과 같이 크기가  $N_m N_l \times 1$ 인 벡터  $\bar{\mathbf{c}}_m$ 과  $\bar{\mathbf{p}}_r$ 을 정의한다.

$$\mathbf{c}_m(l) = [c_{m,1}(l) \quad \cdots \quad c_{m,N_m}(l)]^T \quad (3.17)$$

$$\bar{\mathbf{c}}_m = [\mathbf{c}_m^T(0) \quad \mathbf{c}_m^T(1) \quad \cdots \quad \mathbf{c}_m^T(N_l - 1)]^T \quad (3.18)$$

$$\bar{\mathbf{p}}_r(t) = [\mathbf{p}_r^T(t) \quad \mathbf{p}_r^T(t-1) \quad \cdots \quad \mathbf{p}_r^T(t-N_l+1)]^T \quad (3.19)$$

이로부터 식 3.16은 다음과 같이 나타낼 수 있다.

$$p_{r,m}(t) = \bar{\mathbf{c}}_m^T \bar{\mathbf{p}}_r(t - t_D) + p_{e,m}(t) \quad (3.20)$$

가능도가 최대화되는 해를 구하기 위해  $L_f$  길이의 짧은 시간 구간  $t - \frac{L_f}{2} < t' \leq t + \frac{L_f}{2}$  동안  $p_{e,m}$ 이 가우시안 분포를 갖는다고 가정한다 ( $p_{e,m} \sim \mathcal{N}(0, \sigma^2(t))$ ). 녹음된 신호  $p_{r,m}$ 의 전체 시간 샘플 길이를  $T$ 라고 할 때, 추정해야 하는 변수  $\Theta_m = \{\bar{\mathbf{c}}_m, \sigma^2(1), \dots, \sigma^2(T)\}$ 에 대한 가능도  $\mathcal{L}$ 은 다음과 같다.

$$\begin{aligned} \mathcal{L}(\Theta_m) &= \sum_{t=1}^T \log \Pr [p_{e,m}(t) = p_{r,m}(t) - \bar{\mathbf{c}}_m^T \bar{\mathbf{p}}_r(t - t_D)] \\ &= -\frac{1}{2} \sum_{t=1}^T \frac{|p_{r,m}(t) - \bar{\mathbf{c}}_m^T \bar{\mathbf{p}}_r(t - t_D)|^2}{\sigma^2(t)} \\ &\quad - \frac{1}{2} \log \sigma^2(t) + \text{const.} \end{aligned} \quad (3.21)$$

$\sigma^2$ 을 특정 값으로 고정하면 가능도  $\mathcal{L}$ 을 최대화하기 위해서는  $|p_{r,m}(t) - \bar{\mathbf{c}}_m^T \bar{\mathbf{p}}_r(t - t_D)|^2$ 를 최소화함으로써 달성할 수 있다. 따라서  $\sigma^2$ 을 알면 최소 제곱법 (least square method) 으로부터  $\bar{\mathbf{c}}_m$ 을 추정하고 식 3.20을 이용하여  $p_{e,m}$ 을 추정할 수 있다. WPE에서는  $\sigma^2$ 과  $\bar{\mathbf{c}}_m$ 을 변갈아가며 반복적으로 추정하는 방식을 사용한다. 추정을 위한 반복 횟수를  $N_{\text{WPE}}$ 라고 했을 때, WPE는 다음과 같은 알고리즘으로 나타낼 수 있다.

---

**Algorithm 2** WPE 알고리즘

---

```

1: Initialization:  $n \leftarrow 0$ ,  $\hat{\sigma}^2(t) \leftarrow \max \left\{ \frac{1}{L_f} \sum_{t'=t-\frac{L_f}{2}+1}^{t+\frac{L_f}{2}} |p_{r,1}(t')|^2, \epsilon \right\}$  for  $1 \leq t \leq T$ 
2: while  $n < N_{\text{WPE}}$  do
3:    $\bar{\mathbf{R}} \leftarrow \sum_{t=1}^T \frac{\bar{\mathbf{p}}_r(t - t_D) \bar{\mathbf{p}}_r^T(t - t_D)}{\hat{\sigma}^2(t)}$ 
4:    $m \leftarrow 1$ 
5:   while  $m \leq N_m$  do
6:      $\bar{\mathbf{b}}_m \leftarrow \sum_{t=1}^T \frac{\bar{\mathbf{p}}_r(t - t_D) p_{r,m}(t)}{\hat{\sigma}^2(t)}$ 
7:      $\hat{\mathbf{c}}_m \leftarrow \bar{\mathbf{R}}^\dagger \bar{\mathbf{b}}_m$ ,  $([\cdot]^\dagger : \text{Moore-Penrose pseudo-inverse})$ 
8:      $\hat{p}_{e,m}(t) \leftarrow p_{r,m}(t) - \hat{\mathbf{c}}_m^T \bar{\mathbf{p}}_r(t - t_D)$  for  $1 \leq t \leq T$ 
9:      $m \leftarrow m + 1$ 
10:  end while
11:   $\hat{\sigma}^2(t) \leftarrow \max \left\{ \frac{1}{L_f} \sum_{t'=t-\frac{L_f}{2}+1}^{t+\frac{L_f}{2}} |\hat{p}_{e,1}(t')|^2, \epsilon \right\}$  for  $1 \leq t \leq T$ 
12:   $n \leftarrow n + 1$ 
13: end while
14: Output:  $\hat{p}_{e,m}(t)$  for  $1 \leq m \leq N_m$  and  $1 \leq t \leq T$ 

```

---

WPE 기법은 파라미터  $t_D$ ,  $N_l$ ,  $L_f$  등을 음원이나 잔향 환경에 따라 적절한 값을 설정 해야 좋은 성능을 얻을 수 있다는 한계가 있다. 또한 선형 예측을 기반으로 하기 때문에 RIR의 pole만을 추정하며, 따라서 zero를 만들 가능성성이 높은 초기 반사는 추정하지 않도록  $t_D$ 를 선택하는 데에 주의해야 한다. 그러므로 공간 정보를 완벽히 제거해야 하는 객체

기반 오디오 시스템에 적용하기에는 부적합하다.

### 3.3 심층신경망 기반 기법

심층신경망을 사용하여 잔향 제거 문제를 해결하는 방식은 크게 두 가지로 나뉜다. 하나는 기존의 신호 처리 기법의 보조용으로 심층신경망을 사용하고 [7], [10]–[14], 다른 하나는 심층신경망이 직접 음원의 잔향을 제거하는 방식이다 [16]–[19], [21]–[23]. 후자의 경우 스펙트로그램에 직접 매핑하는 방식과 [17]–[19], [21]–[23], 시간-주파수 빈마다 0부터 1 사이의 값을 갖는 이득을 곱하여 잔향이 제거된 스펙트로그램을 얻는 마스킹 방식으로 나뉜다 [16].

#### 3.3.1 신호 처리 기법의 보조용으로 심층신경망을 사용하는 기법

신호 처리 기법은 신호의 통계적 특성을 이용하여 잔향을 제거한다. 해당 특성을 정확하게 알면 보다 좋은 성능의 잔향 제거가 가능하지만, 실제로 이를 추정하는 것은 쉽지 않다. 따라서 심층신경망의 패턴 인식 능력을 이용해 통계적 특성을 추정함으로써 신호 처리 기법의 성능 향상을 도모할 수 있다. 신호 처리 기법에 필요한 파라미터는 잔향 환경에 따라 조절 되어야 정확한 잔향 제거를 수행할 수 있다. 심층신경망이 여러 환경에 대해 신호의 통계적 특성을 잘 추정할 수 있다면 조절해야하는 파라미터의 수가 줄어들 것이다.

위에서 언급한 기법을 살펴보면 다음과 같다. [7], [11], [12], [14]에서는 심층신경망으로 IRM (Ideal Ratio Mask)을 추정하여 잔향이 포함된 신호와 함께 PSD를 계산하고 이를 범 형성에 사용한다. 해당 기법들은 PSD를 정확히 추정하기 위한 파라미터 조절이 필요 없는 장점이 있지만 범 형성 기법으로부터 잔향을 제거하기 때문에 범 형성 기법이 가진 단점을 여전히 갖고 있다. WPE는 이전 단계에서 예측한 신호로부터 PSD를 추정하여 반복적으로 잔향을 제거해 나가는 신호처리 기법이다. [10]에서는 PSD를, [13]에서는 PSD 계산을 위한 예측 신호를 추정하는 데에 심층신경망을 사용한다. 추정한 PSD로 선형 예측 필터를 구하여 WPE 기법으로 잔향을 제거하기 때문에, 예측 지연 시간과 필터 길이를 조절해야 하는 단점을 여전히 갖고 있다.

### 3.3.2 심층신경망이 직접 잔향을 제거하는 기법

[17], [21], [22]에서는 심층신경망이 잔향이 있는 단일 채널 스펙트로그램을 잔향이 제거된 스펙트로그램으로 매핑하는 방식을 사용한다. 해당 방식의 구조도를 그림 3.2에 나타내었다. [17]은 잔향을 제거하고자 하는 한 개의 시간 프레임과 좌우의 추가적인 시간 프레임을 신경망의 입력으로 하여 잔향이 제거된 시간 프레임을 추정한다. 잔향이 제거된 스펙트로그램을 얻은 후에는 Griffin-Lim 알고리즘 [55]을 사용하여 위상 정보를 추정해 시간 영역의 신호로 복구한다. 한편, [22]은 생성적 적대 신경망 (generative adversarial network, GAN) 형태의 모델로 잔향 제거를 수행한다. 생성기 (generator)는 U-Net [78] 구조를 차용하였으며 잔향이 포함된 크기 스펙트로그램을 이미지로 간주해 전체 신호에 대한 잔향 제거를 수행한다. 판별기 (discriminator)는 생성된 크기 스펙트로그램을 입력으로 하여 원본 신호의 것인지 또는 생성기로부터 출력된 것인지를 판별한다. [21]은 단일 채널 스펙트로그램의 전체 시간 프레임을 신경망의 입력으로 하여, 합성곱 레이어으로 이뤄진 인코딩 경로와 게이트 순환 유닛 (Gated Recurrent Unit, GRU)으로 구성된 디코딩 경로를 거쳐 잔향이 제거된 스펙트로그램을 출력한다. 하지만 시간 영역 신호로 복구 시에 잔향이 포함된 위상 정보를 사용하기 때문에 왜곡이 발생하는 단점이 있다.

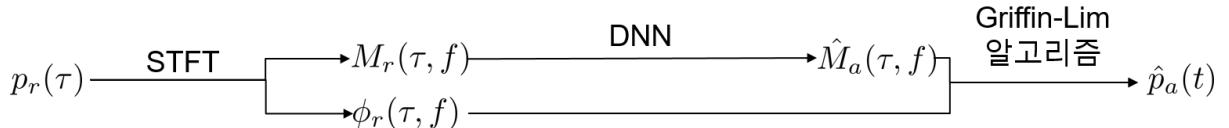


그림 3.2: 단일 채널 신호를 이용한 잔향 제거 기법의 구조도

[31]에서는 다채널 신호가 내포한 공간 정보를 유지하며 입력의 차원을 축소하기 위해 음향 인텐시티 기반의 방향 특징을 제안하였다. 제안한 방향 특징은 두 가지로 하나는 3D 오디오의 녹음과 압축에 주로 쓰이는 directional audio coding (이하 DirAC) 시스템 [79]에서 정의된 DV와, DirAC 시스템에서 사용하는 것보다 많은 마이크로폰이 있을 때 사용할 수 있는 SIV이다. 두 방향 특징 모두 소위 x축, y축, z축 방향에 대한 공간 정보를 포함하고 있으며, DV는 측정 지점에서의 인텐시티를, SIV는 그 보다 넓은 범위에서의 인텐시티를 나타낸다. 그림 3.3에 [31]의 구조도를 나타내었다. 잔향이 포함된 방향 특징을 입력으로

하여 심층신경망이 잔향이 제거된 크기 스펙트로그램으로 매핑하는 방식이다. 시간 영역 신호로 복원할 때에는 Griffin-Lim 알고리즘 [55]을 사용하며, 초기 위상 값으로는 DAS로 얻은 신호의 위상을 사용한다. 하지만 DAS로 얻은 신호 역시 왜곡이 존재하기 때문에 정답 신호에 비해 품질이 떨어지는 문제가 존재한다. 해당 연구에서는 신호 처리 기법이나 가공하지 않은 다채널 신호를 입력으로 사용한 경우보다 압축된 입력을 사용한 모델의 성능이 뛰어남을 보였다.

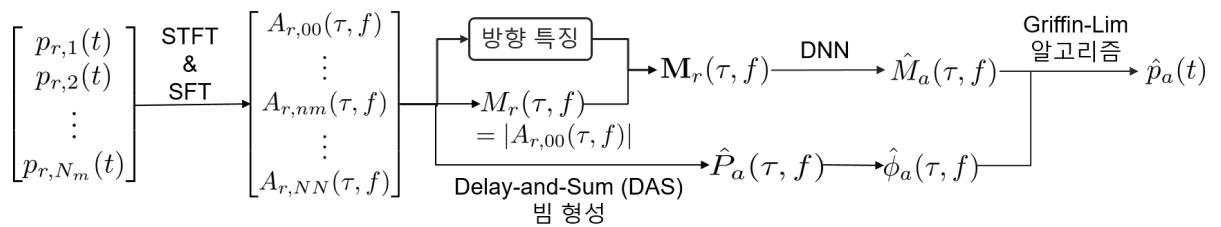


그림 3.3: 방향 특징을 사용한 다채널 잔향 제거 기법의 구조도 [31]

[16]에서는 잔향이 있는 다채널 스펙트로그램으로부터 잔향과 잡음을 제거하기 위한 마스크를 추정한다. IRM은  $|P_a(\tau, f)| / |P_{r,1}(\tau, f)|$ 로 정의되며, 이때  $P_a$ 는 잔향이 없는 환경에서 1번 마이크로폰으로 녹음된 신호의 스펙트로그램이다. 논문에서는 제안한 기법이 신호 처리 기법이나 3.3.1항의 유형과 같은 기법보다 우수한 성능을 나타냄을 보였다. 하지만 앞서 설명한 기법들과 마찬가지로, 잔향과 잡음이 포함된 위상 정보를 이용하여 시간 영역의 신호로 복구하기 때문에 왜곡이 발생하는 단점이 존재한다.

[80], [81]는 엔드투엔드 심층신경망 모델을 사용하였으며, 잔향이 포함된 단일 채널의 시간 영역의 신호를 입력으로 하여 잔향이 제거된 시간 영역의 신호를 출력한다. [80]에서는 [82]에서 제안한 TasNet을 변형하여 잔향 제거에 적용하였다. 신호는 특정 길이만큼 잘라 사용하고 출력 신호를 일정 구간을 겹치고 더하여 (overlap-and-add, 이하 OLA) 본래 신호를 복원하였다. 이 경우 구간 밖의 잔향의 정보를 놓칠 수 있다는 문제와, 단일 채널을 다루기 때문에 공간 정보를 활용하지 못한다는 문제가 있다. [81]는 denoising WaveNet [44]과 생성적 적대 신경망을 결합한 구조이다. 생성기는 denoising WaveNet 형태로, 잔향이 포함된 시간 영역 신호를 입력으로 받아 잔향이 제거된 시간 영역 신호를 출력한다.

판별기는 생성된 신호의 멜-스펙트로그램 (mel-spectrogram)을 입력으로 하여 신호가 원본인지 또는 생성기로부터의 출력 신호인지를 판별한다. 마찬가지로 단일 채널을 다루기 때문에 공간 정보의 활용이 불가하다는 단점이 있다.

## 제 4 장 제안 기법

본 연구에서는 심층신경망의 잔향 제거 성능과 음성 신호의 품질을 향상시키기 위해, 입력 특징으로 음향 인텐시티와 관련된 시간 영역의 방향 특징을 사용할 것을 제안한다. 전체 과정은 다음과 같다. 구형 마이크로폰 어레이로부터 취득한 잔향이 포함된 다채널 신호에 구형 푸리에 변환과 모드 강도 보정을 수행하여 MC-SHD 신호를 얻는다. MC-SHD 신호로부터 방향 특징을 계산한 후 0차 MC-SHD 신호와 연결하여 심층신경망의 입력으로 사용한다. 심층신경망은 잔향이 포함된 방향 특징을 입력으로 받아 잔향이 없는 음성 신호를 추정한다. 아래의 그림 4.1에 제안하는 기법의 구조도를 나타내었다.

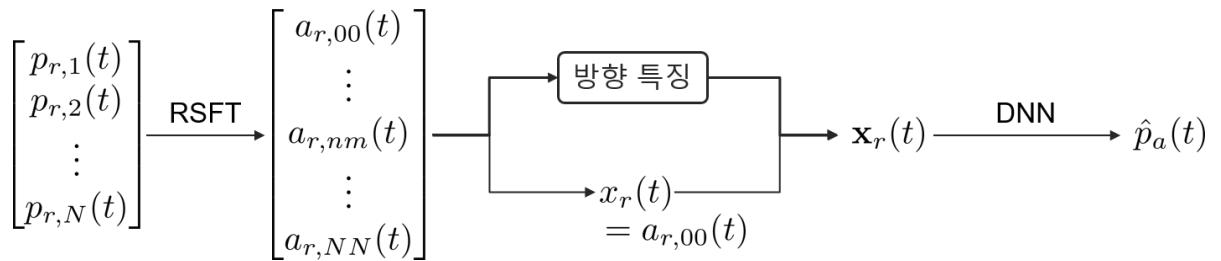


그림 4.1: 제안하는 잔향 제거 기법의 구조도

### 4.1 방향 특징

주파수  $\omega$ 와 파수 벡터  $\mathbf{k}$ 를 갖는 평면파에 대해 읍압  $p$ 와  $\mathbf{v}$ 는 다음과 같이 나타난다.

$$p(\mathbf{r}, t) = P \cos(\mathbf{k} \cdot \mathbf{r} - \omega t) \quad (4.1)$$

$$\mathbf{v}(\mathbf{r}, t) = [V_1 \ V_2 \ V_3]^T \odot \cos(\mathbf{k} \cdot \mathbf{r} - \omega t) \quad (4.2)$$

매질의 밀도  $\rho$ 에 대한 선형 오일러 방정식 [47]을 이용하여  $\mathbf{v}$ 를  $p$ 에 대해 정리하고 순간 인텐시티  $\mathbf{I}$ 를 구하면 다음과 같다.

$$\mathbf{v}(\mathbf{r}, t) = \frac{p(\mathbf{r}, t)}{\rho\omega} \mathbf{k} \quad (4.3)$$

$$\mathbf{I}(\mathbf{r}, t) = p(\mathbf{r}, t) \mathbf{v}(\mathbf{r}, t) = \frac{p^2(\mathbf{r}, t)}{\rho\omega} \mathbf{k} \quad (4.4)$$

심층신경망의 입력으로 사용되는 순간 인텐시티 벡터 (instantaneous intensity vector, 이하 IIV)는 B포맷 앰비소닉 신호와 식 4.4를 이용해 계산한다. B포맷 앰비소닉 신호는 실수 구면 조화 함수를 기저로 하여 구형 푸리에 변환을 수행하고, 모드 강도를 보정한 MC-SHD 신호 중 0차와 1차 신호만을 포함하는 것을 말한다. 0차 신호는 W, 1차 신호는 각각 X, Y, Z 채널이라고 부른다. W, X, Y, Z 채널을 각각  $a_W, a_X, a_Y, a_Z$ 라 할 때 IIV  $\mathbf{i}$ 는 다음과 같이 정의된다.

$$\hat{\mathbf{v}}(t) = [a_X(t) \ a_Y(t) \ a_Z(t)]^T \quad (4.5)$$

$$\mathbf{i}(t) \triangleq a_W(t) \odot \hat{\mathbf{v}}(t) \quad (4.6)$$

W 채널은 측정 원점에서 무지향성 마이크로폰으로부터 얻은 압력 신호와 동등하고, X, Y, Z 채널로 구성된 벡터  $\hat{\mathbf{v}}$ 는 측정 원점에서의 입자 속도를 직교좌표계에서 나타낸 것을 근사한 값이다. 따라서 IIV는 측정 원점에서의 순간 인텐시티의 근사값으로 볼 수 있다.

IIV는 잔향이 포함된 압력 신호인  $x_r$ 과 연결하여 심층신경망의 입력  $\mathbf{x}_r$ 로 사용된다. W 채널은 측정 원점에서의 압력 신호와 동등하므로  $x_r$ 의  $a_{r,W}$ 와 같으며, 방향 특징  $\mathbf{x}_r$ 은 다음으로 정의된다.

$$\begin{aligned} \mathbf{x}_r(t) &= [x_r(t) \ \mathbf{i}_r^T(t)]^T \\ &= [x_r(t) \ i_{r,1}(t) \ i_{r,2}(t) \ i_{r,3}(t)]^T \end{aligned} \quad (4.7)$$

## 4.2 심층신경망 설계

이전 절에서의 방향 특징을 다룰 수 있는 심층신경망 구조가 필요하여 다양한 구조의 모델을 검토하였다. 가장 기본적인 MLP 형태의 심층신경망은 구조상 고정된 크기의 입력이 필요하다. 잔향의 길이에 따라 잔향 제거를 위해 신경망이 보아야 할 시간 길이가 달라지므로 [17], 긴 잔향을 대비하기 위해서는 많은 시간 데이터를 입력으로 받을 수 있도록 설계해야 한다. 하지만 이 경우, 모델의 크기가 지나치게 커지기 때문에 충분히 많은 양의 데이터가 없다면 과적합이 발생할 가능성이 높후하다 [69].

시간 데이터 혹은 순차 데이터 (sequential data)를 다루는 데에는 LSTM 또는 BLSTM이 널리 사용된다. 둘은 한 번에 하나의 시간 샘플을 입력으로 받고 전후의 시간 샘플의 특징을 셀 (cell)에 저장함으로써 데이터의 시간적 맥락을 고려할 수 있다 [83]. 그러나 시간 영역 음성 신호는 길이가 매우 길기 때문에 역전파 (back-propagation)를 이용한 모델의 훈련에 어려움이 있을 수 있으며, MLP와 마찬가지로 과적합이 발생할 수 있다.

기존에 제안되었던 음성 신호를 다루는 엔드투엔드 모델들은 합성곱 레이어를 주로 사용하였으며, 이를 여러 층 쌓아 신경망이 충분한 길이의 신호를 볼 수 있도록 하였다 [40], [41], [44], [45]. WaveNet [41]은 음성 신호를 다루는 엔드투엔드 모델 중에서 가장 고품질의 출력을 보장하지만, 모델의 훈련과 출력에 지나치게 오랜 시간이 필요하다는 문제가 있다. 현재 연구 및 실험 환경에서는 이를 소화할 만큼의 장비가 없기 때문에 제외하였다. Denoising WaveNet [44]은 WaveNet을 변형한 구조이며 이름 그대로 음성의 잡음 제거를 수행하는 모델이다. [81]에서 음성의 잔향 제거에 사용된 바가 있으나, 단일 채널 신호만을 다루기 때문에 공간 정보를 활용하지 못한다는 단점이 있다. Conv-TasNet [45] 또한 같은 이유로 제외하였다. Wave-U-Net [40]은 음원의 분리를 수행하는 모델로 일차원의 U-Net 구조이며 모든 레이어가 합성곱 레이어로 이루어져 있기 때문에 입력의 길이에 제약이 없다는 장점이 있다. 다채널 신호를 입력으로 다룰 수 있으며, 단일 채널보다 모델의 성능 향상이 있음을 연구에서 보였다. 또한, 다채널 음성 잡음 제거를 수행하는 데에 사용된 바가 있다 [46].

### 4.2.1 입출력 전처리

음성 신호는 저주파 성분에 비하여 고주파 성분이 약하다. 전처리 과정을 거치지 않고 심층신경망의 입력으로 넣을 경우 고주파 특성을 살리지 못한 결과가 나오게 된다. 따라서 프리эм페시스 필터 (pre-emphasis filter)를 사용해 저주파 성분을 줄이고 고주파 성분을 증폭시켜주는 과정을 거친다. 이는 SEGAN [43]에서도 전처리 과정으로 사용되었다. 프리эм페시스 필터는 Z-domain에서 다음과 같이 정의되며, 그림 4.2에서 주파수에 따른 이득을 확인할 수 있다.

$$H_{\text{emph}}(z) = 1 - 0.95z^{-1} \quad (4.8)$$

시간 영역에서의  $H_{\text{emph}}$ 를  $h_{\text{emph}}(t)$ 라고 할 때, 프리эм페시스 필터를 통과한 IIV의 원소  $i_{r,j}$ 와 신호  $x_r, p_a$ 는 각각 다음과 같이 합성곱의 형태로 나타난다.

$$\tilde{i}_{r,j}(t) = i_{r,j}(t) * h_{\text{emph}}(t), \quad (j \in \{1, 2, 3\}) \quad (4.9)$$

$$\tilde{x}_r(t) = x_r(t) * h_{\text{emph}}(t) \quad (4.10)$$

$$\tilde{p}_a(t) = p_a(t) * h_{\text{emph}}(t) \quad (4.11)$$

필터를 통과한 입력은  $\tilde{\mathbf{x}}_r = [\tilde{x}_r(t) \ \tilde{i}_{r,1}(t) \ \tilde{i}_{r,2}(t) \ \tilde{i}_{r,3}(t)]^T$ 로 표현된다.

또한 효율적인 심층신경망 훈련을 위해 입출력의 정규화 과정이 필요하다 [84]. 모든 훈련 데이터를 모아 각 채널 (방향 특징 세 채널, 잔향 신호 한 채널) 별로 평균이 0, 분산이 1이 되도록 정규화 한다. 단, 방향 특징은 벡터 값이므로 각 채널의 스케일 (scale)을 유지하기 위해 벡터 크기에 대한 표준편차  $\sigma_i = \sqrt{\sigma_{i,X}^2 + \sigma_{i,Y}^2 + \sigma_{i,Z}^2}$ 으로 정규화 한다.  $\tilde{\mathbf{x}}_r$ 과  $\tilde{p}_a$ 를 정규화 한 것을 각각  $\mathbf{x}'_r$ 과  $p'_a$ 로 표기한다.

심층신경망은 전처리 된 입력  $\mathbf{x}'_r$ 에서 전처리 된 신호  $p'_a$ 로의 대응을 학습한다. 따라서 신경망의 출력  $\hat{p}'_a$ 로부터 음성 신호를 얻기 위해서는 전처리 과정을 역으로 수행하여  $\hat{p}_a$ 를 계산하여야 한다. 역과정에 사용되는 디эм페시스 필터 (de-emphasis filter)는 Z-domain에서 식 4.12와 같이 나타나며, 그림 4.2에서 주파수에 따른 이득을 확인할 수 있다.

$$H_{\text{emph}}^{-1}(z) = \frac{1}{1 - 0.95z^{-1}} \quad (4.12)$$

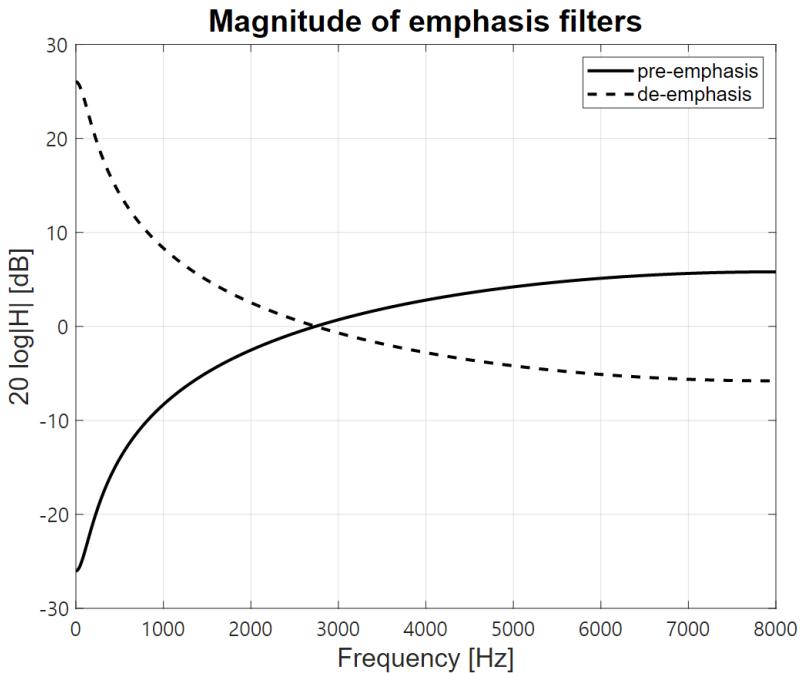


그림 4.2: 프리эм퍼시스 및 디эм퍼시스 필터의 주파수에 따른 이득. 샘플링 레이트 16 kHz를 기준으로 주파수 범위를 설정하였다.

#### 4.2.2 심충신경망 구조

심충신경망 구조는 시간 영역의 음성 신호를 다루는 여러 구조 중 원저자에 의해 변형된 Wave-U-Net [85] (그림 4.3)을 기본 모델로 설정하였다. 보다 효율적인 학습이 가능하도록 실험을 통해 변형한 구조는 그림 4.4에 나타나있다. 표 4.1에는 Wave-U-Net을 변형 과정에 따른 파라미터 수의 변화를 나타내었다. 기존 모델의 성능을 유지하되 파라미터 수를 감소하는 방향으로 구조를 바꾸었다. 전체 구조는 입력의 크기를 줄이는 대신 채널의 수를 늘리는 인코더 (encoder) 경로와 반대의 역할을 수행하는 디코더 (decoder) 경로, 그리고 둘 사이의 병목으로 이뤄져있다.

본래의 U-Net [78]과 같이 Wave-U-Net [40], [85]은 인코더 경로의 중간 출력을 동일한 층의 디코더 경로로 채널축에 대해 연결 (concatenate)하는 스kip 커넥션 (skip connection)을 사용한다. 변형한 구조에서는 이를 합으로 연산하는 레지듀얼 커넥션 (residual connection)으로 대체하여 파라미터수를 감소시켰다. 병목에서는 기존의 합성곱 레이어 대신 전결합 레이어와 레지듀얼 연산을 사용한다. 구조의 특성 상 채널의 수가 층이 깊어짐에 따라

표 4.1: Wave-U-Net의 구조 변형 및 파라미터 수의 변화

Wave-U-Net	192,106,562
레지듀얼 커넥션	170,074,562
전결합 레이어 병목	162,794,282
레지듀얼 블록	114,790,442
채널 수 감소	41,355,002

라 증가하면 각 채널 별로 다양한 특징맵을 추출할 수 있다. 그러나 모델의 성능에 반드시 도움이 된다는 보장은 없기 때문에, 채널축에 대한 전결합 레이어 연산을 적용하여 채널의 유용성을 높이고자 하였다. 인코딩과 디코딩 연산을 이행하는 블록에는 ResNet [86]에서 제안한 레지듀얼 블록 (residual block)을 삽입하였다. 레지듀얼 블록은 모델이 효율적인 학습을 하는 데에 도움이 된다는 연구 결과가 있어 이를 차용하였다.

신경망의 학습을 위한 손실 함수는 널리 사용되는 함수 중 하나인 평균 절대 오차 (mean absolute error)를 사용하였다.  $T_a$ 를  $p_a$ 의 시간 길이라고 할 때, 손실 함수는 다음과 같이 나타낼 수 있다.

$$L(p'_a, \hat{p}'_a) = \frac{1}{T_a} \sum_{t=0}^{T_a-1} |p'_a(t) - \hat{p}'_a(t)| \quad (4.13)$$

많이 사용되는 평균 제곱 오차 (mean squared error)보다 실험적으로 좋은 성능과 학습 진행을 보여주었기 때문에 평균 절대 오차를 선택하였다.

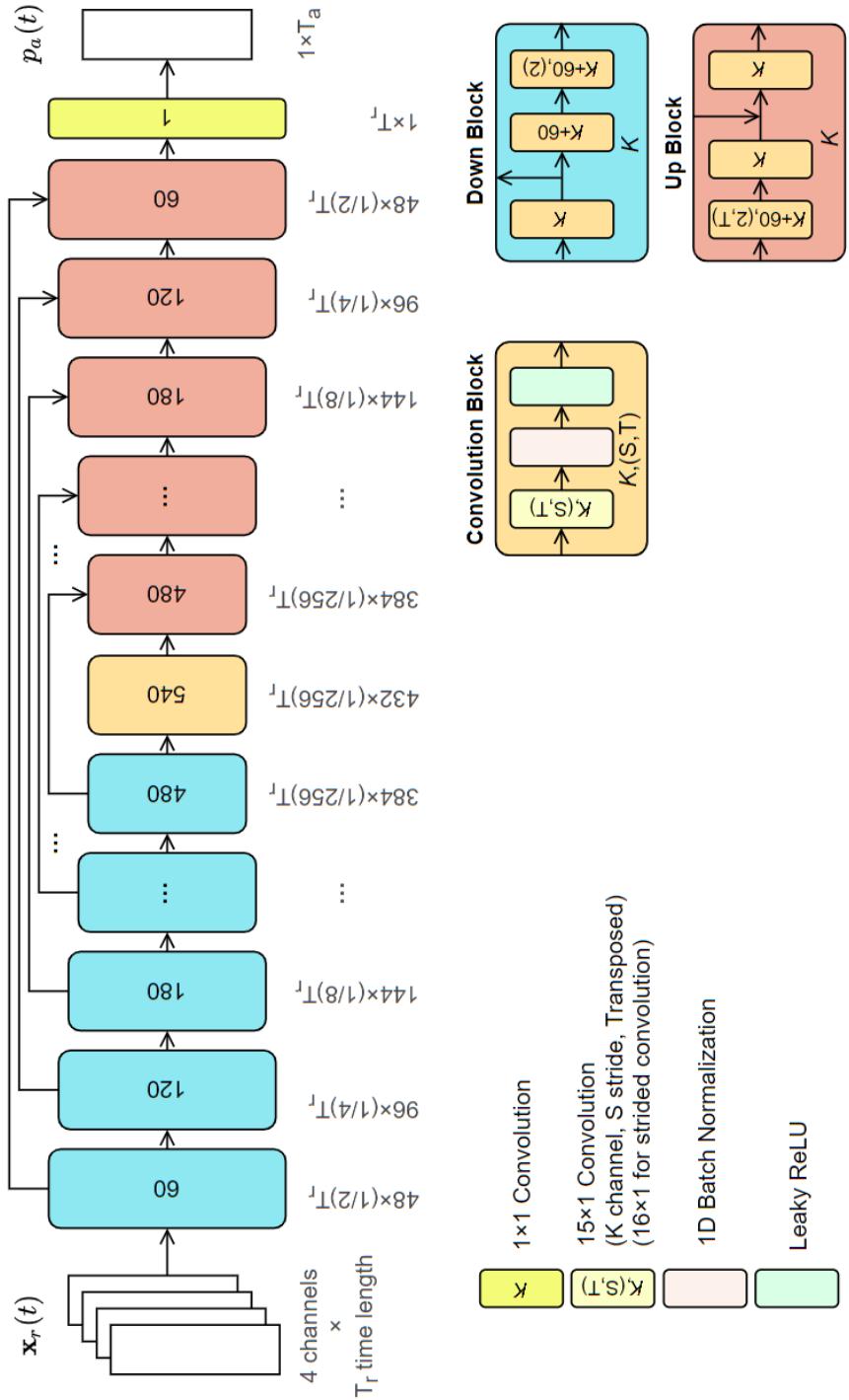


그림 4.3: Wave-U-Net [85] 구조. 원저자가 변형한 형태로 본래 구조는 [40]이다.

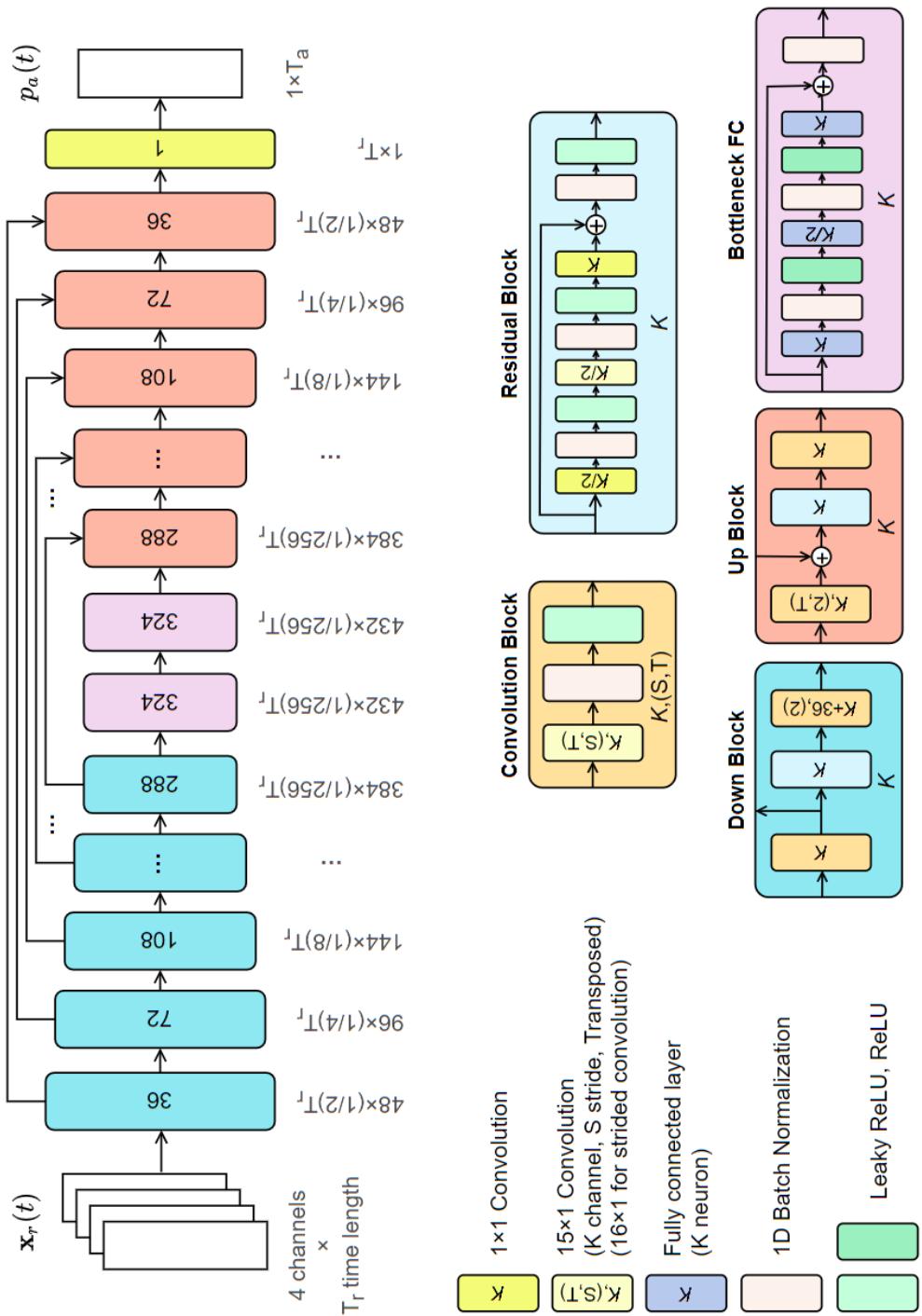


그림 4.4: 본 연구에서 제안한 Wave-U-Net [85]의 변형 구조. 전결합 레이어로 구성된 병목과 레지듀얼 블록을 추가하였다.

## 제 5 장 실험 및 결과

시간 영역의 방향 특징과 변형한 심층신경망 모델이 잔향 제거 성능에 얼마나 영향을 주는지 확인하기 위해 4개의 심층신경망 모델을 훈련하여 비교하였다. 시뮬레이션을 통해 하나의 방에서 마이크로폰 어레이와 음원의 위치를 달리하며 RIR을 생성하고, 음성과 합성곱하여 데이터 세트를 만들어 사용하였다.

### 5.1 실험 설정

#### 5.1.1 성능 평가 지표 및 비교 기법

성능 평가 지표로 사람이 느끼는 음성의 품질을 나타내는 PESQ (Perceptual Evaluation of Speech Quality) [87]와 명료도를 나타내는 STOI (Short-Time Objective Intelligibility) [88], 그리고 신호 대 잡음비를 나타내는 fwSegSNR (frequency-weighted Segmental Singla-to-Noise Ratio) [89]을 사용하였다. PESQ는 구간  $[-0.5, 4.5]$ 의 값을, STOI는 구간  $[0, 1]$ 의 값을 가진다. 세 개의 지표 모두 값이 클수록 음성의 품질과 명료도가 우수함을 의미한다.

위의 성능 지표들을 4개의 심층신경망 기법 (IIV-MW 모델, IIV-W 모델, DV 모델, Single-MW 모델) 결과에 대해 계산하였다. 각각의 기법은 다음과 같이 수행하였다.

1. IIV-MW 모델: IIV를 포함하는 방향 특징을 입력으로 하여 훈련된 변형된 Wave-U-Net (modified Wave-U-Net) 구조의 심층신경망 모델
2. IIV-W 모델: IIV를 포함하는 방향 특징을 입력으로 하여 훈련된 Wave-U-Net 구조의 심층신경망 모델
3. DV 모델 [31]: 방향 벡터 (directional vector, 이하 DV)를 포함하는 방향성 스펙트로그램을 입력으로 하여 훈련된 심층신경망 모델. 위상 추정 초기값은 잔향이 포함된 입력의 위상을 그대로 사용했다.
4. Single-MV 모델: 단일 채널 음성 신호를 입력으로 하여 훈련된 심층신경망 모델

### 5.1.2 레이터 세트 구성

구형 마이크로폰 어레이의 응답을 시뮬레이션할 수 있는 프로그램인 SMIR generator [30]를 사용하여 RIR을 합성하였다. 사용한 구형 마이크로폰 어레이는 그림 5.1의 Eigenmike [90]와 동일한 사양으로, 반지름이 4.2 cm인 강체 구 표면에 32개의 마이크로폰이 배치된 형태이다. IIV를 계산할 때에는 32개 채널 데이터를 실수 구면 조화 함수를 기저로 SFT하여 1차까지의 SHD 신호를 사용하였다. DV를 계산할 때에는 [31]의 방식을 따라 실수 구면 조화 함수를 기저로 SFT하여 1차까지의 SHD 신호를 사용하였다. SHD 신호에서 강체 구의 모드 강도를 보정하기 위해 다음의  $b_n^{-1}(kr)$ 을 사용하여 MC-SHD신호를 계산하였다.

$$b_n^{-1}(kr) = \frac{1 + \sqrt{\lambda}}{|b_n(kr)| + \sqrt{\lambda}} \exp(-i\angle b_n(kr)) \quad (5.1)$$

$$\lambda = \left(1 - \sqrt{1 - 1/G^2}\right) / \left(1 + \sqrt{1 - 1/G^2}\right) \quad (5.2)$$

이때  $\lambda$ 는 규제 계수 (regularization factor)이며 상수  $G$ 는  $10/32$ 로 설정하였다.



그림 5.1: RIR 합성에 사용한 Eigenmike [90] 사진

RIR을 합성하는 데에 사용한 가상방의 환경은  $[너비, 깊이, 높이] = [10.5 \text{ m}, 7.1 \text{ m}, 3.0 \text{ m}]$ ,  $T60 = 0.31$  초이다. 마이크로폰 어레이와 음원의 위치 선정은 둘 모두 벽에서 50 cm 이상 떨어지도록 하였고, 둘 사이의 거리 또한 50 cm 이상이 되도록 하였다. 총 8개의

위치를 선정하여 데이터 세트를 구성하는데 사용하였다.

음성 데이터 세트는 TIMIT [91]을 사용하였다. TIMIT의 훈련용 음성 샘플 4,620개를 균일하게 무작위로 9,600번, 150번 선택하여 합성한 RIR과 합성곱하여 훈련 데이터 세트와 검증 데이터 세트를 만들었다. 그리고 TIMIT의 테스트용 음성 샘플 1,680개를 균일하게 무작위로 805번 선택하여 합성한 RIR과 합성곱하여 테스트 데이터 세트를 만들었다. IIV 및 단일 채널을 심층신경망의 입력으로 사용할 경우 규제 효과를 주기 위해 마이크로폰 어레이와 음원 위치를 동일하게 하여 직접파만 존재하는 RIR과 음성 샘플을 합성하여 960 개의 추가 훈련 데이터세트를 구성하였다. 정리하면, 훈련 데이터 세트는 총 10,560개의 샘플, 검증 데이터 세트는 총 150개의 샘플, 그리고 테스트 데이터 세트는 총 805개의 데이터 세트로 구성되었다.

### 5.1.3 하이퍼파라미터 설정

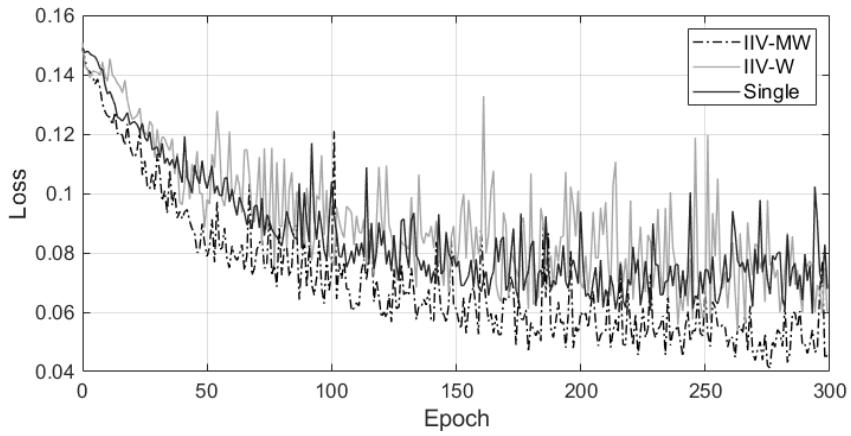
표 5.1: 심층신경망 훈련을 위한 하이퍼파라미터

샘플 레이트	16 kHz
옵티마이저	AdamW [65]
가중치 감쇠	0
모멘텀	$\beta_1 = 0.9, \beta_2 = 0.999$
학습률	$5 \times 10^{-4}$
배치 크기	16, IIV-W의 경우 8

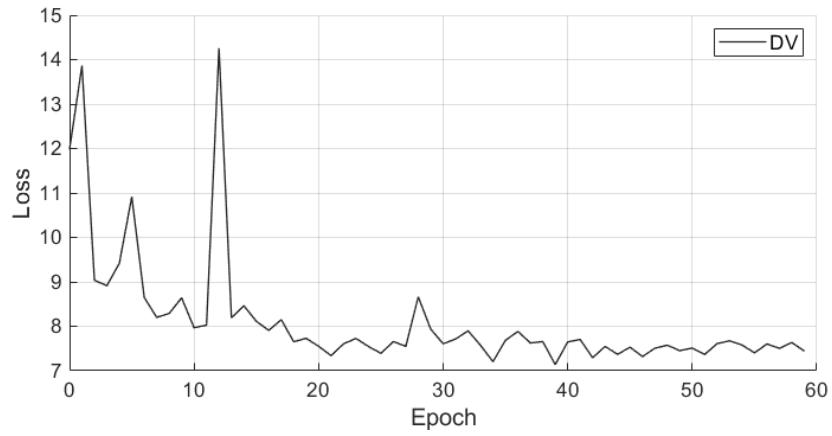
표 5.1에 심층신경망 훈련을 위한 하이퍼파라미터를 나타내었다. 스펙트로그램 매핑 방식을 이용하는 DV 모델의 경우 [31]의 하이퍼파라미터를 그대로 사용하였다. 배치 내의 샘플들은 시간 또는 시간 프레임의 길이가 다르기 때문에 가장 긴 샘플의 길이에 맞춰 0을 패딩 하였고, 손실 함수 및 성능 평가 지표를 계산할 시에는 패딩을 제거한 후 계산하였다.

## 5.2 실험 결과

그림 5.2는 훈련 이포크에 따른 모델 별 손실 함수값의 변화를 나타낸다. DV는 로그 (log)를 취한 크기 스펙트로그램으로 구성되며 손실 함수로 평균 제곱 오차를 사용했기



(a) 시간 영역 특징 사용 모델



(b) DV를 사용한 모델

그림 5.2: 훈련 이포크에 따른 검증 데이터의 손실 함수값 변화. (a) 시간 영역 특징을 사용한 모델, (b) DV를 사용한 모델의 손실 함수값을 나타낸다.

때문에, 시간 영역 특징을 사용한 모델과 데이터 스케일의 차이가 있어 둘을 분리하여 나 타냈다. 시간 영역의 특징을 사용한 모델은 300 번의 훈련 이포크를 거친 모델 중 검증용 데이터 세트에서 성가 지표 결과가 가장 뛰어난 모델을 선택하였다. DV 모델은 60 번의 이포크를 거친 모델을 사용하였다.

### 5.2.1 테스트 데이터 세트에 대한 잔향 제거 성능

표 5.2: 모델 별 파라미터 수

IIV-MW	41,355,002
IIV-W	192,106,562
DV	122,025,474
Single	41,351,762

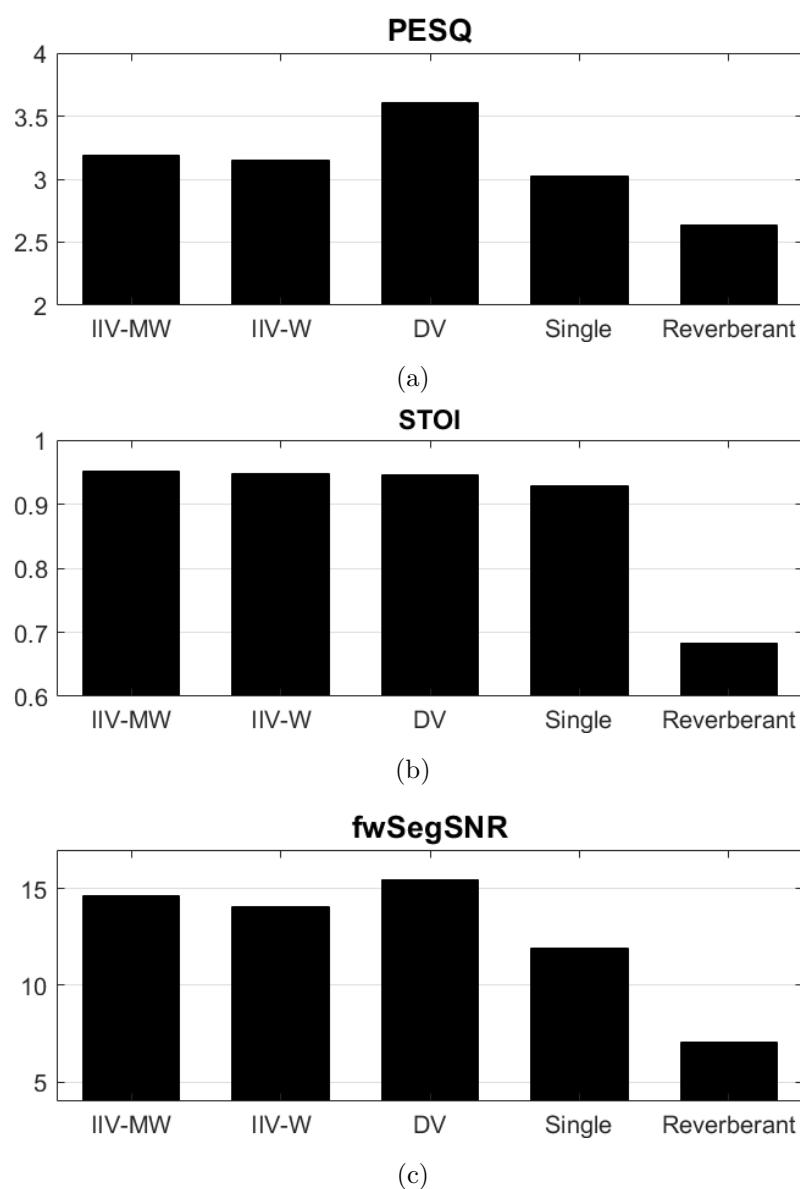


그림 5.3: 훈련된 심층신경망 모델들과 테스트 데이터 세트에서의 잔향 제거 성능. 각각 (a) PESQ, (b) STOI, (c) fwSegSNR의 값을 나타낸다.

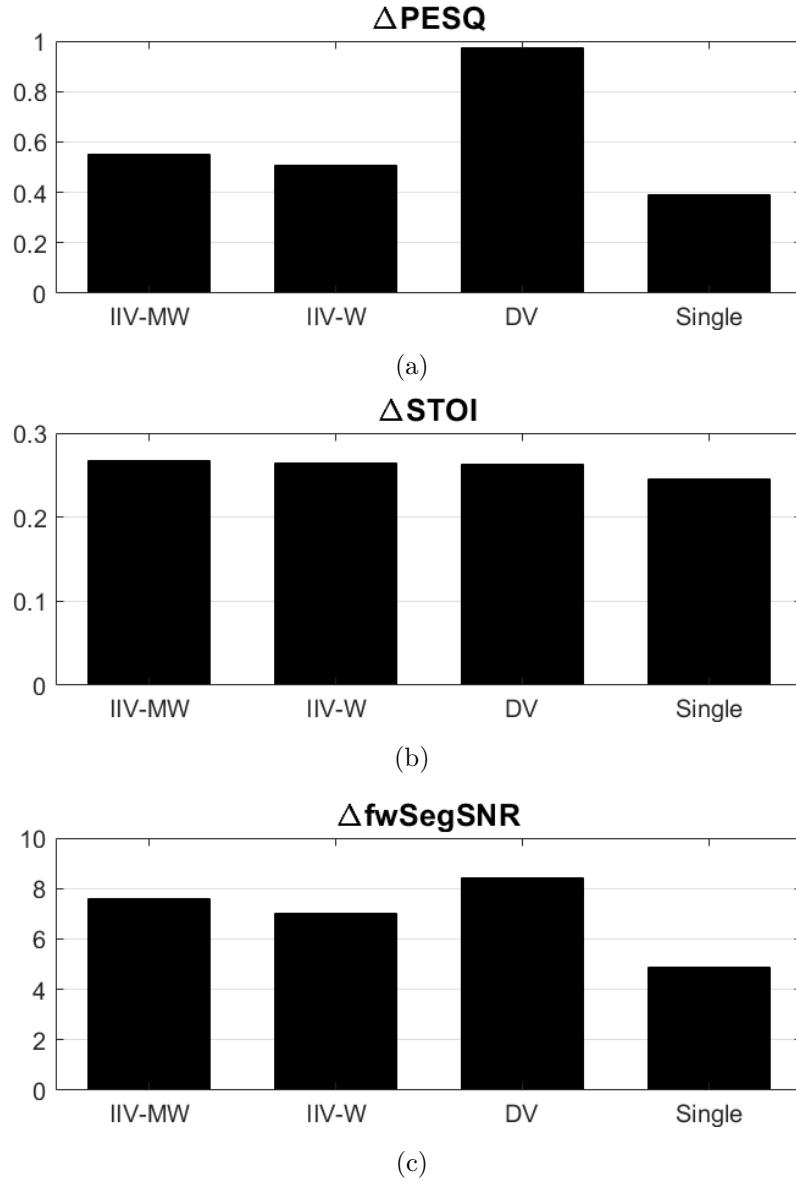


그림 5.4: 잔향이 포함된 입력 대비 훈련된 심층신경망 모델들의 잔향 제거 성능 향상도. 각각 (a) PESQ, (b) STOI, (c) fwSegSNR의 값을 나타낸다.

훈련이 완료된 심층신경망 모델들을 테스트 데이터 세트로 테스트 했을 때의 잔향 제거 성능을 그림 5.3에 나타내었다. 해당 그림은 잔향이 포함된 입력에 비해 각 모델의 출력의 성능 지표 값을 나타낸 것이다. 또한, 좀 더 보기 쉽게 잔향이 포함된 입력에 대비하여 각 모델의 성능 지표가 향상된 정도를 그림 5.4에 나타내었다.

STOI의 향상도가 Single을 제외하고 비등함을 고려했을 때, DV 모델이 가장 우수한 성능을 보임을 알 수 있다. 특히 PESQ에서는 여타 모델보다 두드러지는 증가폭을 보이

는데 이는 PESQ의 연산 방식으로부터 추측해 볼 수 있다. 사람의 청각 모델로 변환하는 단계에, 참조 (또는 정답) 신호와 왜곡된 신호 사이의 평균 바크 스펙트럼 (mean Bark spectrum)의 비로부터 전달 함수를 추정하여, 추정 값으로 참조 신호를 왜곡된 신호에 대해 주파수 이퀄라이즈 (frequency equalize)하는 과정이 포함된다. 또한 특정 역치 값 이하의 왜곡에 대한 마스킹 과정이 포함된다 [87]. 앞의 두 과정에서 DV 모델의 출력이 지난 왜곡이 PESQ의 연산 방식에 의해 완화 혹은 상쇄되어 다른 모델에 비해 보다 높은 성능 지표 값을 얻었을 가능성이 존재한다.

시간 영역의 특징을 사용하는 모델 중에서는 IIV-MW 모델이 가장 우수한 성능을 보인다. 표 5.2에서 확인할 수 있듯이, IIV-W의 파라미터 수에 비해 약 5배 적은 수의 파라미터를 사용하지만 모든 지표에 대해 작은 차이일지라도 더 큰 값을 기록하였다. 이로부터 모델의 변형된 구조가 더 효율적임을 말할 수 있다. Single 모델은 비교 모델 중 가장 나쁜 성능을 보인다. 따라서 공간에 대한 정보를 포함한 입력이 심층신경망이 잔향을 파악하는데에 도움이 되는 것으로 간주할 수 있다.

테스트 데이터 샘플 중 하나를 선택하여 그림 5.5에 입력, 정답, 그리고 각 모델의 출력 스펙트로그램을 나타내었다. 샘플을 실제로 들었을 때 왜곡이 발생하는 유성음 부분을 빨간색 사각형으로 표시하였다. IIV-MW 모델의 출력 (그림 5.5(b))은 잔향이 적당히 제거된 모습을 보이지만, IIV-W 모델의 출력 (그림 5.5(e))은 잔향이 과도하게 제거되어 포먼트 (formant)가 끊기는 형태를 보인다. DV의 출력 (그림 5.5(c))은 정답과 가장 비슷한 형태를 보일만큼 잔향이 잘 제거된 모습을 보인다. Single의 출력은 (그림 5.5(f))은 포먼트가 뭉개진 것을 보이며, 실제로 들었을 때 왜곡이 발생함을 알 수 있다.

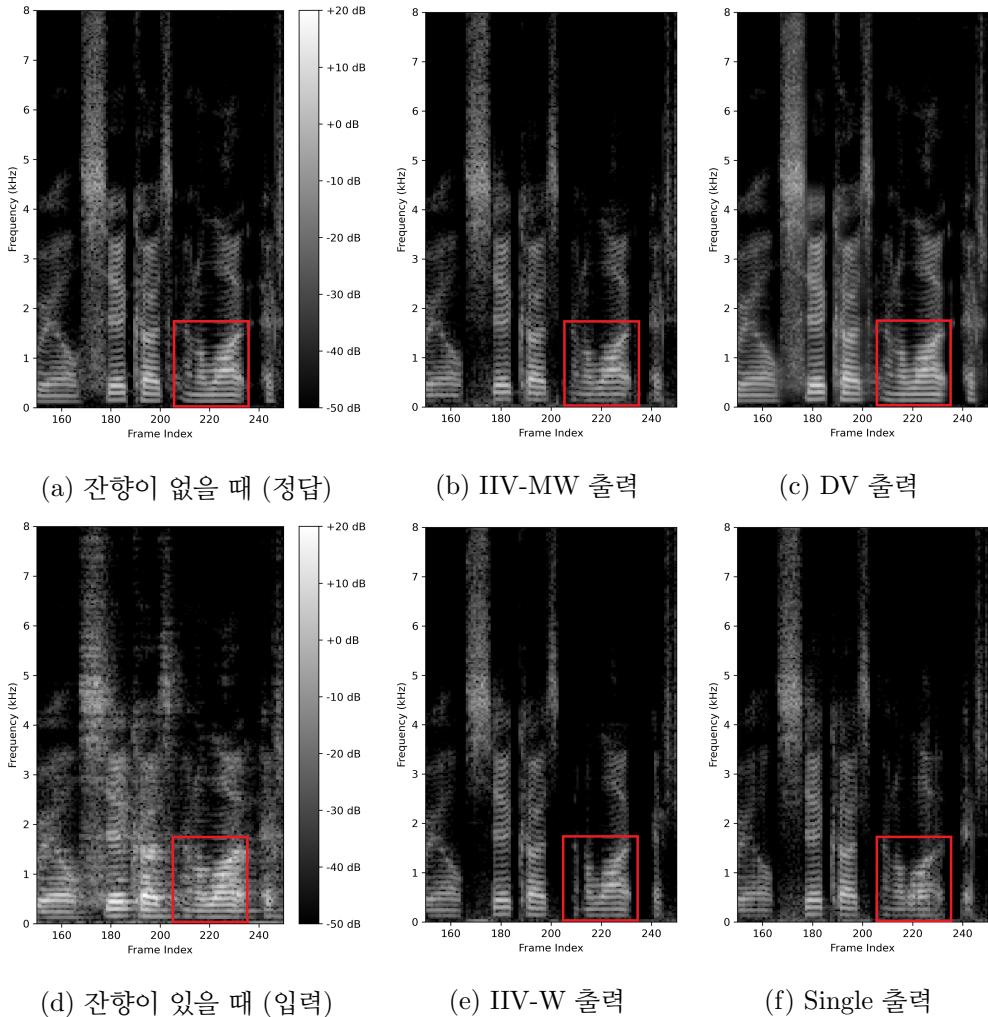


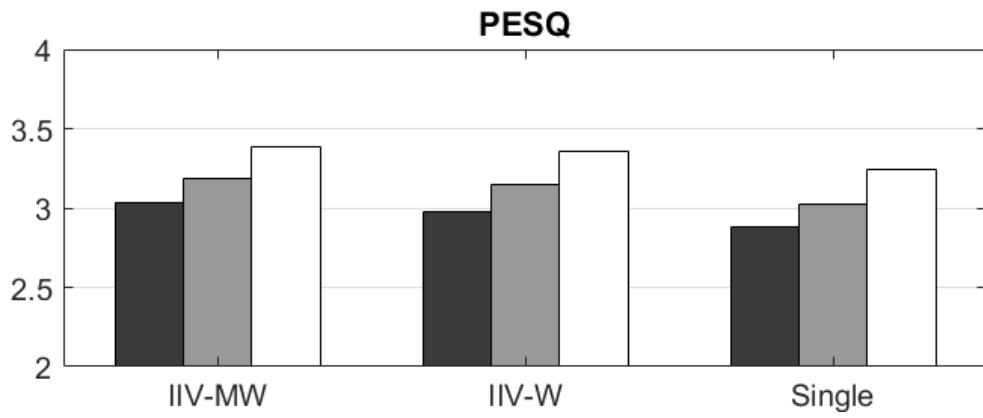
그림 5.5: 테스트 샘플의 입력 및 정답 스펙트로그램과 각 모델의 출력 스펙트로그램. 차이가 있는 부분을 빨간색 사각형으로 표시하였다.

### 5.2.2 추정한 시간 영역 신호에 포함된 위상의 유의미성

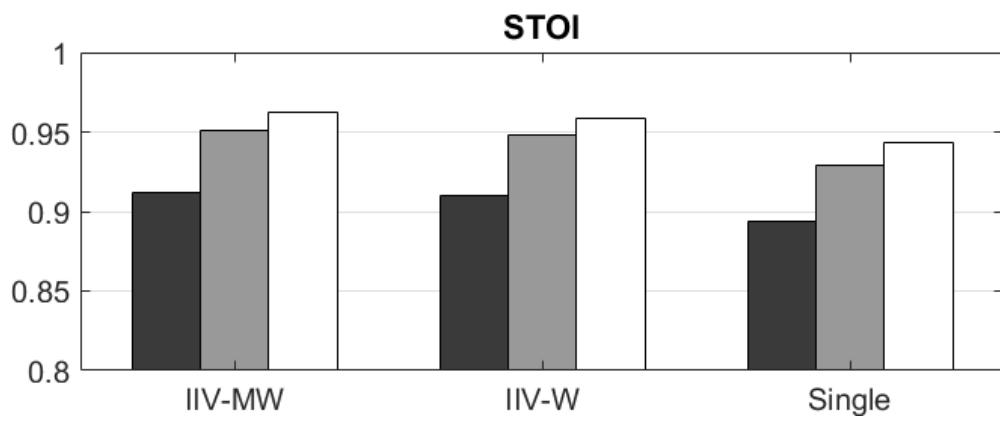
비교 기법 중 엔드투엔드 모델이 추정한 시간 영역 신호가 갖는 위상이 얼마나 유의미한지 알아보기 위해 다음과 같은 실험을 진행하였다. 추정된 신호를 STFT하여 크기를 고정한 후, 잔향이 포함된 입력의 위상을 초기값으로 하여 Griffin-Lim 알고리즘을 사용하거나, 무향 신호의 위상과 함께 ISTFT하여 시간 영역의 신호를 복원하였다. 테스트 데이터 세트에 대한 성능 지표 값을 그림 5.6에 나타내었다. ‘Reverberant’은 잔향이 포함된 위상을, ‘Anechoic’은 무향 신호의 위상을 사용한 경우이고 ‘Time’은 모델의 출력에 대한 평가 결과이다.

잔향이 포함된 위상 ('Reverberant')으로 시간 영역의 신호를 복원한 경우는 모델이

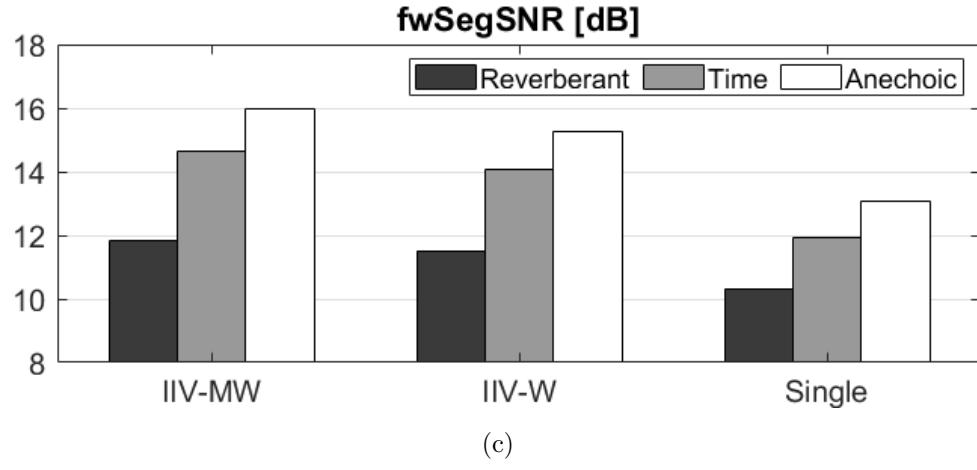
추정한 신호 ('Time')에 비해 음성의 품질이 떨어지는 것을 보인다. 모델이 추정한 신호와 무향 신호의 위상 ('Anechoic')으로 시간 영역의 신호를 복원한 경우도 품질의 차이가 있으나, 명료도 면에서는 비슷한 것으로 보인다. 이로부터 엔드투엔드 모델이 추정한 음성 신호의 위상은 유의미하다고 말할 수 있다. 향후 연구에서는 위상과 함께 신호의 크기 역시 효과적으로 추정할 수 있는 기법에 대한 연구가 필요할 것이다.



(a)



(b)



(c)

그림 5.6: 엔드투엔드 모델로 추정한 신호의 서로 다른 위상에 대한 성능 지표 값. 각각 (a) PESQ, (b) STOI, (c) fwSegSNR의 값을 나타낸다.

## 제 6 장 맷음말

본 연구에서는 구형 마이크로폰 어레이로 녹음된 신호에서 심층신경망이 잔향을 제거할 수 있도록 훈련시키기 위해 시간 영역의 방향 특징을 입력으로 사용할 것과, 기존에 제안된 심층신경망 모델을 변형함으로써 보다 효율적인 학습을 도모하고자 하였다. 방향 특징은 한 지점에서의 순간 음향 인텐시티를 근사한 IIV로 0차와 1차 구면 조화 신호로 계산된다. 다채널 신호를 다루고 음성 잡음 제거에 사용된 바가 있는 앤드투엔드 모델인 Wave-U-Net을 심층신경망 모델로 사용하였다.

시간 영역의 방향 특징과 변형한 모델의 유용성을 보이기 위하여 IIV를 입력으로 하되 원형의 구조를 사용한 심층신경망 모델 (IIV-W), 시간-주파수 영역의 방향 특징을 포함한 DV가 입력인 심층신경망 모델 (DV), 단일 채널 데이터를 입력으로 하는 심층신경망 모델 (Single)과 비교하였다. 테스트 데이터 세트로 테스트하였을 때, DV 모델이 가장 높은 성능을 보였으며 이를 제외하고는 IIV-MW 모델이 앤드투엔드 모델 중에 가장 좋은 잔향 제거 성능을 보였다. IIV-W 모델은 IIV-MW 모델에 버금가는 성능을 보였으나 훨씬 많은 수의 파라미터로 구성되어있기에 변형한 모델이 더 효율적임을 보였다. 단일 채널 데이터로 훈련한 모델은 가장 나쁜 잔향 제거 성능을 나타냈다. 이로부터 공간 정보가 잔향 제거에 도움이 되는 것을 확인할 수 있었다.

제안한 기법에서 시간 영역의 입출력은 위상 정보를 포함하기 때문에, 이를 심층신경망이 학습하면 기존의 위상 추정 방식보다 좋은 성능을 보일 것이라 가정하였다. 시간 영역의 입출력이 내포한 위상 정보가 유의미함을 확인하기 위해, 출력의 크기 스펙트로그램은 고정하고 잔향이 포함된 위상을 Griffin-Lim 알고리즘의 초기값으로 하여 얻은 신호와, 원본 신호가 갖는 위상을 사용하여 ISTFT한 신호에 대해 성능 평가 지표를 제안 기법과 비교하였다. 앤드투엔드 모델의 출력은 무향 신호의 위상과 근접한 위상을 포함하는 것을 보였다.

향후 연구에서는 잔향 제거의 성능을 향상시키고 다양한 환경에 대한 일반화가 진행

되어야 한다. 현재의 데이터 세트는 여덟 개의 RIR만으로 구성되어있기 때문에, 모델의 성능 향상 및 일반화를 위해서는 보다 다양한 환경에 대한 데이터 세트 구성이 필요하다. 덧붙여, 엔드투엔드 모델의 특성상 수행하는 과제의 복잡도가 높기 때문에, 훈련을 위해선 데이터 세트의 크기를 키울 필요가 있다. 또한 학습에 보다 유리한 구조와 손실 함수의 존재 가능성이 있으며 그에 따른 연구와 최적화가 진행되어야 한다.

## 참 고 문 헌

- [1] P. Coleman *et al.*, “Object-based reverberation for spatial audio,” *J. Audio Eng. Soc.*, vol. 65, no. 1, pp. 66–77, Jan. 2017.
- [2] J. Benesty *et al.*, “On microphone-array beamforming from a mimo acoustic signal processing perspective,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1053–1065, 2007.
- [3] S. Braun *et al.*, “An informed spatial filter for dereverberation in the spherical harmonic domain,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Vancouver, Canada, 2013, pp. 669–673.
- [4] O. Schwartz, S. Gannot, and E. A. P. Habets, “Multi-microphone speech dereverberation and noise reduction using relative early transfer functions,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 2, pp. 240–251, Feb. 2015.
- [5] B. Cauchi *et al.*, “Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech,” *EURASIP J. Advances in Signal Process.*, vol. 2015, no. 1, p. 61, Jul. 2015.
- [6] T. Nakatani *et al.*, “Speech dereverberation based on variance-normalized delayed linear prediction,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Aug. 2010.
- [7] J. Heymann, L. Drude, and R. Haeb-Umbach, “Neural network based spectral mask estimation for acoustic beamforming,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 196–200.
- [8] H. Erdogan *et al.*, “Improved MVDR beamforming using single-channel mask prediction networks,” in *Proc. INTERSPEECH*, San Francisco, 2016, pp. 1981–1985.
- [9] Z. Q. Wang and D. L. Wang, “All-neural multi-channel speech enhancement,” in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 3234–3238.

- [10] S. R. Chetupalli and T. V. Sreenivas, “LSTM based AE-DNN constraint for better late reverb suppression in multi-channel LP formulation,” arXiv: [1812.01346 \[cs\]](https://arxiv.org/abs/1812.01346), Dec. 2018.
- [11] T. May, “Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 2, pp. 406–414, 2018.
- [12] Y. Liu *et al.*, “Neural network based time-frequency masking and steering vector estimation for two-channel mvdr beamforming,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6717–6721.
- [13] A. S. Subramanian *et al.*, “An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions,” arXiv: [1904.09049 \[eess.AS\]](https://arxiv.org/abs/1904.09049), Apr. 2019.
- [14] L. Pfeifenberger, M. Zohrer, and F. Pernkopf, “Eigenvector-based speech mask estimation for multi-channel speech enhancement,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2162–2172, Sep. 2019.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016, p. 163.
- [16] S. Chakrabarty and E. A. P. Habets, “Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks,” *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 787–799, Apr. 2019.
- [17] K. Han *et al.*, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 6, pp. 982–992, Mar. 2015.
- [18] Y. Zhao, Z. Q. Wang, and D. Wang, “A two-stage algorithm for noisy and reverberant speech enhancement,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, New Orleans, Louisiana, 2017, pp. 5580–5584.
- [19] B. Wu *et al.*, “A reverberation-time-aware approach to speech dereverberation based on deep neural networks,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 1, pp. 98–107, Oct. 2017.

- [20] D. S. Williamson and D. Wang, “Time-frequency masking in the complex domain for speech dereverberation and denoising,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Apr. 2017.
- [21] J. F. Santos and T. H. Falk, “Speech dereverberation with context-aware recurrent neural networks,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 7, pp. 1232–1242, Apr. 2018.
- [22] O. Ernst *et al.*, “Speech dereverberation using fully convolutional networks,” in *Proc. European Signal Process. Conf.*, Rome, Italy, 2018, pp. 390–394.
- [23] W. J. Lee *et al.*, “Speech dereverberation based on integrated deep and ensemble learning algorithm,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, 2018, pp. 5454–5458.
- [24] F. Fahy, *Sound Intensity*, 2nd ed. CRC Press, 2002, pp. 38–60, 72–77.
- [25] A. D. Pierce, *Acoustics: An introduction to its physical principles and applications (McGraw-Hill series in mechanical engineering)*. New York: McGraw-Hill Book Company, 1981, pp. 305–310.
- [26] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 33–38.
- [27] K. Kinoshita *et al.*, “The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *Proc. IEEE Workshop Appl. Signal Process. to Audio and Acoust.*, New Paltz, New York, 2013, pp. 1–4.
- [28] H. Christensen *et al.*, “The CHiME Corpus: A resource and a challenge for computational hearing in multisource environments,” in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1918–1921.
- [29] E. Hadad *et al.*, “Multichannel audio database in various acoustic environments,” in *Proc. 14th Int. Workshop Acoust. Signal Enhancement*, French Riviera, 2014, pp. 313–317.
- [30] D. P. Jarrett *et al.*, “Rigid sphere room impulse response simulation: Algorithm and applications,” *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

- [31] J. Liu, J.-W. Choi, and B. Jo, “Dereverberation based on deep neural networks with directional feature from spherical microphone array recordings,” in *International Commission for Acoustics (ICA)*, International Commission for Acoustics (ICA), 2019.
- [32] S. W. Fu *et al.*, “Complex spectrogram enhancement by convolutional neural network with multi-metrics learning,” in *Proc. IEEE Int. Workshop Machine Learning for Signal Process. (MLSP)*, Tokyo, Japan, 2017, pp. 1–6.
- [33] Y. Wakabayashi *et al.*, “Single-channel speech enhancement with phase reconstruction based on phase distortion averaging,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 9, pp. 1159–1169, Apr. 2018.
- [34] H.-S. Choi *et al.*, “Phase-aware speech enhancement with Deep Complex U-Net,” in *Int. Conf. Learning Representations*, 2019, arXiv: [1903.03107 \[cs\]](https://arxiv.org/abs/1903.03107).
- [35] Z. Ouyang *et al.*, “A fully convolutional neural network for complex spectrogram processing in speech enhancement,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brighton, United Kingdom, 2019, pp. 5756–5760.
- [36] K. Tan and D. Wang, “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brighton, United Kingdom, 2019, pp. 6865–6869.
- [37] N. Zheng and X. L. Zhang, “Phase-aware speech enhancement based on deep neural networks,” arXiv: [1608.01953](https://arxiv.org/abs/1608.01953), *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 1, pp. 63–76, Sep. 2019.
- [38] Y. Masuyama *et al.*, “Deep Griffin–Lim iteration,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Brighton, United Kingdom, 2019, pp. 61–65. [Online]. Available: <https://ieeexplore.ieee.org/document/8682744>.
- [39] J. Lee and H. G. Kang, “A joint learning algorithm for complex-valued T-F masks in deep learning-based single-channel speech enhancement systems,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 1098–1109, Apr. 2019.
- [40] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation,” arXiv: [1806.03185 \[cs\]](https://arxiv.org/abs/1806.03185), Jun. 2018.

- [41] A. v. d. Oord *et al.*, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [42] A. v. d. Oord *et al.*, “Parallel wavenet: Fast high-fidelity speech synthesis,” *arXiv preprint arXiv:1711.10433*, 2017.
- [43] S. Pascual, A. Bonafonte, and J. Serra, “Segan: Speech enhancement generative adversarial network,” *arXiv: 1703.09452 [cs]*, Jun. 2017.
- [44] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5069–5073.
- [45] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [46] H. Lee *et al.*, “End-to-end multi-channel speech enhancement using inter-channel time-restricted attention on raw waveform,” *Proc. Interspeech 2019*, pp. 4285–4289, 2019.
- [47] A. D. Pierce, *Acoustics: An introduction to its physical principles and applications (McGraw-Hill series in mechanical engineering)*. New York: McGraw-Hill Book Company, 1981, pp. 8–11, 26–27.
- [48] ———, *Acoustics: An introduction to its physical principles and applications (McGraw-Hill series in mechanical engineering)*. New York: McGraw-Hill Book Company, 1981, pp. 255–258.
- [49] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Springer International Publishing, 2017, vol. 9, pp. 27–29.
- [50] B. Rafaely, *Fundamentals of Spherical Array Processing*. Springer-Verlag Berlin Heidelberg, 2015, pp. 1–20.
- [51] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*. Springer International Publishing, 2017, vol. 9, pp. 33–36.

- [52] J. Le Roux *et al.*, “Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency,” in *Proc. Int. Conf. Digital Audio Effects*, vol. 10, 2010.
- [53] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, “Phase processing for single-channel speech enhancement: History and recent advances,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.
- [54] P. Mowlaei, R. Saeidi, and Y. Stylianou, “Advances in phase-aware signal processing in speech communication,” *Speech communication*, vol. 81, pp. 1–29, 2016.
- [55] D. Griffin and Jae Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.
- [56] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016, pp. 1–26.
- [57] ——, *Deep Learning*. Cambridge: MIT Press, 2016, pp. 108–118.
- [58] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 1–57.
- [59] I. A. Basheer and M. Hajmeer, “Artificial neural networks: Fundamentals, computing, design, and application,” *Journal of microbiological methods*, vol. 43, no. 1, pp. 3–31, 2000.
- [60] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.,” *Psychological review*, vol. 65, no. 6, p. 386, 1958.
- [61] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015, pp. 1–38.
- [62] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016, pp. 78–95.
- [63] S. Ruder, “An overview of gradient descent optimization algorithms,” arXiv: [1609.04747 \[cs\]](https://arxiv.org/abs/1609.04747), *arXiv preprint arXiv:1609.04747*, 2016.

- [64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv: [1412.6980 \[cs\]](https://arxiv.org/abs/1412.6980), Dec. 2014.
- [65] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Int. Conf. Learning Representations*, 2019, arXiv: [1711.05101 \[cs\]](https://arxiv.org/abs/1711.05101).
- [66] L. N. Smith, “Cyclical learning rates for training neural networks,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2017, pp. 464–472.
- [67] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in *Int. Conf. Learning Representations*, 2017, arXiv: [1608.03983 \[cs\]](https://arxiv.org/abs/1608.03983).
- [68] K. Hornik, “Approximation capabilities of multilayer feedforward networks,” *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.
- [69] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time-series,” in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., MIT Press, 1995.
- [70] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 266–267.
- [71] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015, pp. 169–185.
- [72] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge: MIT Press, 2016, pp. 326–366.
- [73] V. Dumoulin and F. Visin, *A guide to convolution arithmetic for deep learning*, 2016, arXiv: [1603.07285 \[stat\]](https://arxiv.org/abs/1603.07285).
- [74] 김양한, 최정우, *Sound Visualization and Manipulation*. John Wiley & Sons, Ltd, 2013, pp. 137–140.
- [75] ——, *Sound Visualization and Manipulation*. John Wiley & Sons, Ltd, 2013, pp. 157–163.

- [76] B. D. Carlson, “Covariance matrix estimation errors and diagonal loading in adaptive arrays,” *IEEE Transactions on Aerospace and Electronic systems*, vol. 24, no. 4, pp. 397–401, 1988.
- [77] M. Elmaraazey, “On spatial smoothing techniques for beamforming in the presence of correlated arrivals,” *Signal Processing*, vol. 37, no. 2, pp. 157–170, May 1994.
- [78] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, 2015, pp. 234–241.
- [79] V. Pulkki and C. Faller, “Directional audio coding: Filterbank and STFT-based design,” in *Proc. Audio Eng. Soc. Conv. 120*, Paris, France, 2006.
- [80] Y. Luo and N. Mesgarani, “Real-time single-channel dereverberation and separation with time-domain audio separation network.,” in *Interspeech*, 2018, pp. 342–346.
- [81] J. Su, A. Finkelstein, and Z. Jin, “Perceptually-motivated environment-specific speech enhancement,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 7015–7019.
- [82] Y. Luo and N. Mesgarani, “Tasnet: Time-domain audio separation network for real-time, single-channel speech separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 696–700.
- [83] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [84] Y. A. LeCun *et al.*, “Efficient backprop,” in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48.
- [85] D. Stoller, S. Ewert, and S. Dixon, *Wave-u-net-pytorch*, 2019 (accessed April, 2020). [Online]. Available: <https://github.com/f90/Wave-U-Net-Pytorch>.
- [86] K. He *et al.*, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016, pp. 770–778.

- [87] A. W. Rix *et al.*, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Salt Lake City, Utah, 2001, pp. 749–752.
- [88] C. H. Taal *et al.*, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Dallas, Texas, 2010, pp. 4214–4217.
- [89] J. Ma, Y. Hu, and P. C. Loizou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.
- [90] *EigenStudio User Manual R02C*, mh acoustics LLC, New Jersey, 2017, pp. 26–27. Accessed on: Dec. 21, 2019. [Online]. Available: <https://mhacoustics.com/sites/default/files/EigenStudio%20User%20Manual%20R02C.pdf>.
- [91] J. S. Garofolo *et al.*, “DARPA TIMIT Acoustic-phonetic continuous speech corpus CDROM,” 1993.

## 사 사

2년이라는 시간은 길고도 짧은 시간입니다. 대학원에서 보낸 2년은 그 말이 특허나 잘 어울리는 듯합니다. 미적지근한 열정만으로는 나아가기는커녕 버티기도 쉽지 않았던 시간이었습니다. 도망치고 싶은 순간도 있었지만, 감사하게도 곁에서 응원해주고 믿어주는 분들이 있어 지금까지 올 수 있었습니다.

먼저 그러한 시간을 함께 해주신 최정우 교수님께 감사드립니다. 약한 모습을 비출 때에도 늘 독려하고 나아갈 길을 차근차근 알려주셨습니다. 교수님으로부터 배웠던 가장 큰 점은 연구에 있어 비판적 태도를 항상 유지하는 것입니다. 모든 것에 왜라는 의문을 던지고 어떤 논리에 의해서 결과가 도출되었는가를 파악하는 일은 연구자로서의 기본 덕목이지만, 이를 체화하는 일은 쉽지 않습니다. 교수님과의 토론을 통해 그러한 자세를 기를 수 있었고 앞으로 연구를 하는 동안, 나아가 매 순간에 많은 도움이 되리라 생각합니다.

연구실 생활을 함께한 분들에게도 감사드립니다. 부드러움으로 먼저 다가와 주신 전세운 박사님, 연구를 할 때 늘 의견을 물어봐 주시고 조언을 해주셔서 감사합니다. 다가와 주신만큼 기꺼움을 표현하지 못한 것 같아 죄송한 마음이 듭니다. 나중에도 연락하며 이런저런 이야기들을 나눌 수 있으면 좋겠습니다. 대학원에서의 시간은 어떻게 보내야 할지, 연구 방향은 어떻게 잡는 것이 좋은지 알려주고 가끔은 음악 얘기도 나누었던 수연이형, 고맙습니다. 더 많은 이야기를 나눌 수 있었을 텐데 그러지 못해 아쉽고 미안합니다. 매사에 조언을 아끼지 않으며 어떤 일이든 해보자고 손을 내밀어준 병호형도 고맙습니다. 입학 후 모든 게 낯설었던 제가 연구실에 적응하는 데에 큰 도움을 받았습니다. 연구적으로도 의지가 되어주어 늘 감사했습니다. 종종 고마움을 전했지만 더 말해도 부족할 정민이, 쉽지 않은 연구실 생활을 시작부터 함께한 동기이자 연구 사수에게 감사를 전합니다. 많은 것을 공유하고 배울 수 있어서 참 좋았습니다. 앞으로도 꾸준히 연락하며 오래 잘 지냈으면 합니다. 실없는 이야기도 진중한 대화도 많이 나누었던 견지, 답답한 마음을 터놓을 수 있는 친구가 되어주어 고맙고 즐거웠습니다. 투투거림도 아무렇지 않게 넘겨준 현준이에게는 미안함과 좋은 일이 있기를 바라는 마음을 전하고 싶습니다. 끈기 있는 모습에 늘 감탄하게 만드는 형민이 형, 제가 좀 더 도움을 줄 수 있었다면 하는 아쉬움이 듭니다. 세 명 모두 남은 시간 동안 좋은 결과를 얻기를 항상 응원하겠습니다. 마지막 학기를 함께한 하영이형, 동현이와 지훈이, 짧지만 즐거운 시간이었습니다. 앞으로의 연구실 생활에 무운을 빌며 모든 일을 잘 해쳐 나갈 수 있기를 바랍니다. 마지막으로 연구실에서 함께 했던 동안과 졸업 후에도 꾸준히 소식을 전하며 찾아준 형준이형, 민성이형, 수영이형에게도 고마움을 전합니다.

멀리 있어도 언제나 곁에 있어준 가족들에게도 감사합니다. 항상 걱정하고 사랑한다고 말씀해주시는 할머니와 할아버지, 충고와 응원을 아끼지 않은 부모님, 본인도 힘들지만 격려해주는 동생, 모두들 사랑합니다. 늘 할 수 있다는 믿음을 전해준 친척들에게도 감사하다는 말씀을 드립니다.

언제든 찾아가도 싫은 기색 없이 자리를 내어준 하용이, 서로에게 힘이 되는 말을 나누고 자신감을 북돋아준 성연이, 정말 고맙습니다. 직장 생활로 힘들 텐데도 모이자고 하면 나와 주는 하영누나도, 얼결에 대학원 생활을 하고 있는 혼택이도, 모두 좋은 일이 있기를 바랍니다. 비슷한 때에 힘든 일을 겪었지만 잘 이겨내고 응원의 말을 아끼지 않았던 은영누나, 박사까지 잘 할 수 있을 거라 믿고 항상 응원합니다. 한 달에 한 번씩은 꼭 보는 것 같은 지원누나, 태선이형, 병준이형, 곁에 있어주어 고맙습니다. 아직도 실감이 잘 안 나는 새신랑 재문형과 이제는 종종 볼 수 있으면 하는 영빈이형, 근태형도 고맙습니다. 그리고 학부 때부터 많은 도움을 준 승재형에게도 고마움을 전합니다. 친구처럼 지내는 상수와 은주, 모두 건강하고 행복한 일이 다가오길 바랍니다.

부산에 가면 항상 반갑게 맞아주는 동주와 승훈이, 일이 잘 풀려 목표하는 바를 꼭 이룰 수 있기를 바라고 종종 연락하겠습니다. 시간이 안 맞아 잘 만나지 못했지만 언제나 연락하면 환영해준 보윤이, 다음에는 꼭 볼 수 있도록 하겠습니다. 만날 때마다 즐거운 대화를 나누는 혜승이도 계속해서 지금처럼 연락하고 친하게 지냈으면 합니다. 알게 된 시간은 짧았지만 더 알고 싶은 혼욱이, 앞으로도 친하게 지냅시다. 소소한 일로도 자주 연락해준 희찬이, 이제는 좋은 일과 좋은 사람만이 찾아오길 기도하겠습니다. 자주 연락은 못했지만 기꺼워해준 기현이, 태완이, 기환이도 고맙습니다.

마지막 학기를 함께 술로 스트레스를 풀었던 현수, 승훈이형, 윤재형, 재웅이, 수문이형 모두 고맙습니다. 좋은 친구를 사귀게 되어 정말 즐거웠습니다. 졸업 후에도 종종 함께하러 오겠습니다. 그리고 형이자 친구로서 많은 대화를 나누고 시간을 공유한 룸메이트 성원이형에게도 감사를 전합니다.

다 적지 못하였지만, 2여 년의 시간 동안에 도움을 준 모든 분들에게 감사의 말씀을 드립니다. 옆에서 건네주신 한 번의 인사와 말이 제게는 큰 힘이 되었습니다. 이제는 필요할 때 도움을 드릴 수 있는 사람이 되겠습니다. 감사합니다.

## 약력

이름: 정경석

## 학력

2011. 3. – 2014. 2. 동인고등학교 (3년 수료)

2014. 2. – 2018. 2. 성균관대학교 전자전기공학부 (학사)

2018. 2. – 2020. 8. 한국과학기술원 전기및전자공학부 (硕사)

## 경력

2018. 2. – 2020. 8. 한국과학기술원 전기및전자공학부 일반조교