2023년 8월
석사학위논문

잡음 환경에 강건한 음향 이벤트 위치 추정 및 탐지에 대한 연구

조선대학교 대학원

컴퓨터공학과

신 영 서

잡음 환경에 강건한 음향 이벤트 위치 추정 및 탐지에 대한 연구

A Study on Robust Sound Event Localization and Detection in Noisy Environment

2023년 8월 25일

조선대학교 대학원

컴퓨터공학과

신 영 서

잡음 환경에 강건한 음향 이벤트 위치 추정 및 탐지에 대한 연구

지도교수 전 찬 준

이 논문을 공학 석사학위신청 논문으로 제출함

2023년 4월

조선대학교 대학원

컴퓨터공학과

신 영 서

신영서의 석사학위논문을 인준함

위원장 조선대학교 교 수 최 우 열 (인)

위 원 조선대학교 교 수 양 희 덕 (인)

위 원 조선대학교 교 수 전 찬 준 (인)

2023년 5 월

조선대학교 대학원

목 차

표 목 차

그 림 목 차

ABSTRACT

A Study on Robust Sound Event Localization and Detection in Noisy Environment

Shin, Yeongseo
Advisor : Prof. Chun, Chanjun, Ph.D.
Department of Computer Engineering
Graduate School of Chosun University

SELD(sound event localization and detection) aims to classify sound events, detect their onset and offset, and estimate their direction. Recent advancements in deep learning have led to the proposal of various SELD methods based on deep learning techniques.

However, existing deep learning-based SELD methods have mostly been developed and evaluated on audio data with minimal noise. Consequently, applying these methods in environments with high levels of noise often yields suboptimal performance. To address this issue, some approaches preprocess the data by applying noise reduction techniques before performing SELD, but this can be time-consuming.

To overcome these challenges, this study proposes a SELD model that incorporates deep learning-based speech enhancement techniques. The proposed model combines a U-net architecture for speech enhancement and a SELDnet for SELD, enabling end-to-end processing of noise reduction and SELD tasks.

To evaluate the proposed method, experiments were conducted using datasets containing diverse levels of noise. The results demonstrated that the proposed model outperformed existing SELD models in environments with high levels of noise, although it showed slightly lower performance in noise-free

environments.

제 1 장 서론

1절 연구 배경 및 연구 동기

오디오 신호를 활용한 로봇 [1-2], 보안 및 감시 시스템 [3-4], 음성 인식 [5-6] 등 다양한 산업의 발달을 통해 오디오 신호를 활용하기 위한 작업에 대한 중요성이 부각되고 있으며, 이를 위한 작업 중 하나로 음향 이벤트 탐지 및 위치 추정(Sound event localization and detection; SELD)에 대한 연구가 진행 중에 있다.

SELD는 음향 이벤트 탐지(Sound event detection; SED) [7]와 음원 위치 추정(Sound source localization; SSL) [8]이라는 두 가지 하위 task로 구성된다.

음향 이벤트 탐지는 그림 1.1과 같이 입력으로 들어오는 오디오 신호에서 특정 음향 이벤트의 시작 지점 (onset)과 끝 지점(offset)을 파악하여 이벤트의 활성화를 식별하며, 식별된 음향 이벤트를 클래스에 따라 분류를 수행하는 작업이다. 배경 소음, 간섭 음원 발생, 음원 중첩, 동일한 클래스의 다중 음원 발생 등 다양한 환경에서의 음향 이벤트 탐지에 대한 연구가 진행되고 있다 [9-10].

음원 위치 추정은 그림 1.2와 같이 입력으로 들어오는 다채널 오디오 신호를 마이크 기준으로 3차원 공간에서 음향 이벤트의 위치를 추정하는 작업이다. 단순히 방향을 추정하는 작업뿐만 아니라 거리까지 고려하는 작업도 이에 포함된다. 음원 위치 추정은 결과적으로 도착 방향(Direction of Arrival; DOA)을 추정하는 과정이 포함된다.

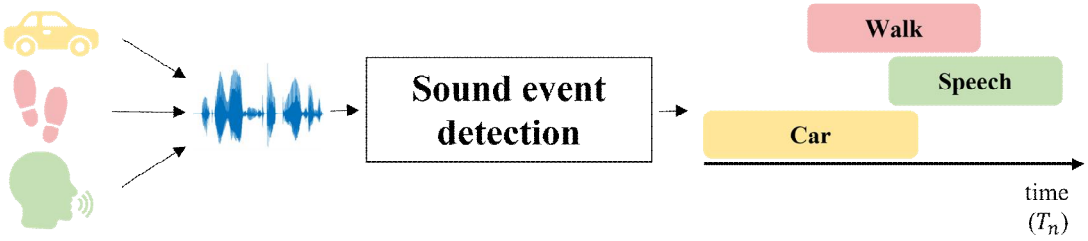


그림 1.1 음향 이벤트 탐지 시스템의 개요

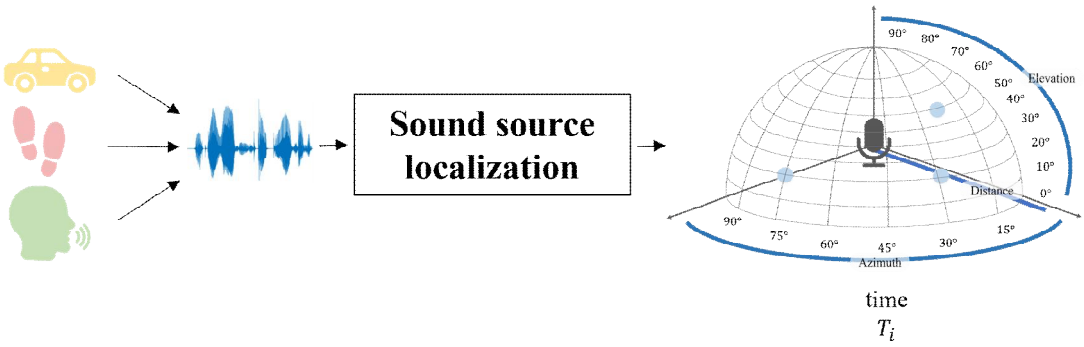


그림 1.2 음원 위치 추정 시스템의 개요

음향 이벤트 탐지 및 위치 추정은 두 가지 작업을 동시에 수행하여 그림 1.3 처럼 오디오 신호에서 음원을 분류하고, 분류된 음원의 시작 지점과 끝 지점을 탐지하며, 위치까지 추정하는 작업을 수행한다. 음향 이벤트 탐지와 음원 위치 추정을 개별적으로 수행할 경우, 인식된 이벤트와 추정된 DOA 사이의 데이터 연관성 문제가 야기 된다는 문제가 있다 [11]. 그러나 음향 이벤트 탐지 및 위치 추정은 두 가지 작업을 동시에 수행함으로써 데이터 연관성 문제를 해결할 수 있으며 상호 보완적인 정보 활용을 통해 더 효율적이고 정확한 결과를 얻는 것이 가능하다. 또한 컴퓨팅 자원과 에너지 절약 또한 가능하다는 장점이 있다.

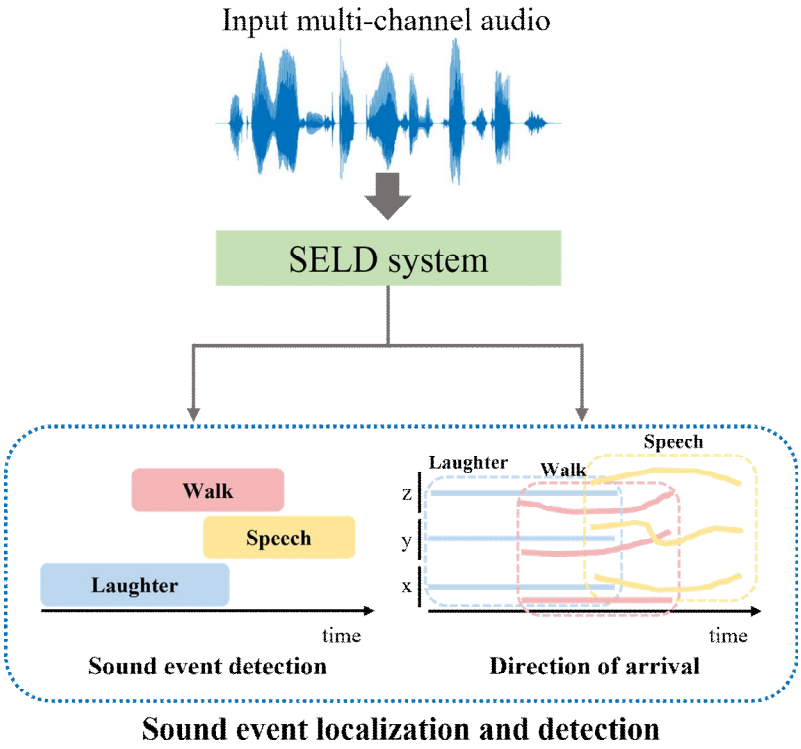


그림 1.3 음향 이벤트 탐지 및 위치 추정 시스템의 개요

딥러닝(deep learning) 기술이 발전함에 따라, 음향 이벤트 탐지 및 위치 추정을 수행하기 위한 다양한 딥러닝 기반의 방법들이 제안되고 연구되고 있다. 그러나 현재 연구되고 있는 연구는 잡음이 없는 환경에서의 성능 향상에 초점을 맞춰 진행되고 있다[12-13]. 그러나 이러한 방법들은 오디오 신호가 심한 잡음의 영향을 받는 경우 성능이 저하되는 문제가 발생한다. 표 1.1은 동일한 딥러닝 모델을 사용하여 잡음의 유무에 따른 성능을 비교한 결과이다. 잡음이 존재할 경우, 모든 지표에서 성능이 저하된 것을 확인할 수 있다. 실생활에서는 다양한 종류의 잡음들이 존재한다. 실생활에서 음향 이벤트 탐지 및 위치 추정 모델을 적용하기 위해서는 잡음 환경에서도 높은 성능을 유지할 수 있도록 개선이 필요하다.

표 1.1 잡음 유무에 따른 신경망 모델의 성능 비교

	Error Rate	F1-score	Localization error	Localization recall	SELD score
잡음이 없는 데이터	0.71	0.30	17.56	0.38	0.53
잡음이 있는 데이터 (SNR -10)	0.84	0.17	23.20	0.24	0.63

잡음은 음향 이벤트 탐지 및 위치 추정 작업뿐만 아니라 오디오 신호를 활용하는 다양한 분야들에서도 성능 저하 문제를 발생시킨다. 이러한 문제를 해결하기 위해 음성 향상(speech enhancement)에 대한 연구가 진행되고 있다[14-15]. 음성 향상은 오디오 신호 내에 불필요하게 섞인 잡음을 제거하고 원하는 음성 신호를 복원 및 향상시키는 작업이다. 최근 다양한 기술에 적용되는 음성을 활용하는 분야에서는 기계가 필요한 신호를 똑바로 인식할 수 있어야 하기 때문에 음성 향상은 이에 필수적인 전처리 과정으로 적용된다. 이러한 음성 향상 작업을 음향 이벤트 탐지 및 위치 추정 모델에 적용하면 잡음 환경에서 성능 저하 문제가 개선될 것으로 보인다.

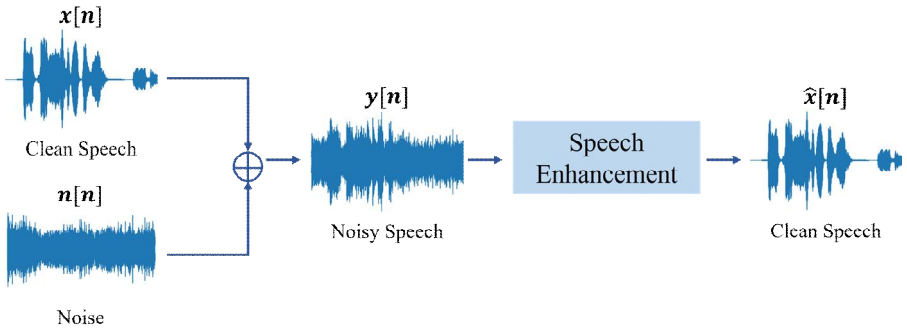


그림 1.4 음성 향상의 개요도

본 연구에서는 잡음 환경에서 기존 음향 이벤트 탐지 및 위치 추정 모델의 성능 저하 문제를 개선하기 위해 음성 향상 기술을 결합한 음향 이벤트 탐지 및 위치 추정 모델인 SELD U-net을 제안한다. 본 연구는 딥러닝 기반의 음성 향상 작업이 우수한 성능을 보이고 있고, 이를 음향 이벤트 탐지 및 위치 추정을 수행하는 딥러닝 모델과 결합을 통해 잡음에 강건성을 가진 모델을 구축할 수 있음을 보인다.

2절 연구 목적

딥러닝 기반의 음향 이벤트 탐지 및 위치 추정 모델은 잡음이 없는 환경에서의 성능 향상에 초점이 맞춰 연구가 진행되고 있다. 이러한 모델들은 오디오 데이터가 잡음의 영향을 받을 경우, 성능이 저하되는 문제가 존재한다. 이러한 문제를 개선하기 위해 음향 이벤트 탐지 및 위치 추정을 수행하기 전, 필터링 기술이나 딥러닝 기술을 활용한 잡음을 제거하는 전처리 작업을 수행하는 방법이 존재한다.

그러나 단순한 필터링 기술을 사용하여 잡음 제거를 수행하는 경우, 잡음이 다양한 주파수 범위에서 변화하면 일부 잡음 성분이 남아있게 되어 효과적인 잡음 제거가 어려울 수 있다는 단점이 있다. 딥러닝을 활용한 잡음 제거를 수행할 경우, 필터링 기술과 달리 잡음이 다양한 주파수 범위에서 변화하더라도 모델 학습 시 이를 고려하여 효과적으로 잡음 제거를 수행할 수 있다.

하지만 딥러닝 기반의 잡음 제거 방식의 수행은 일반적으로 전처리 과정에서 수행된다. 즉, 학습 시간이 오래 걸리는 딥러닝을 전처리와 음향 이벤트 탐지 및 위치 추정 작업에서 별도로 수행되기 때문에 시간적 효율성이 떨어질 수 있다는 단점이 존재하며, 이러한 단점은 실생활에서 적용하기 어려울 수 있다.

본 연구는 잡음 환경에 강건한 딥러닝 기반의 음향 이벤트 탐지 및 위치 추정 모델의 구축을 목표로 하며 음성 향상과 음향 이벤트 탐지 및 위치 추정을 동시에 수행하는 SELD U-net 모델을 제안한다. 제안한 모델은 end-to-end 방식으로, 잡음 제거를 수행하는 U-net과 음향 이벤트 탐지 및 위치 추정을 수행하는 SELDnet으로 구성된다.

제안한 모델을 사용하여 잡음 환경에서 음향 이벤트 탐지 및 위치 추정을 수행할 경우, 기존 방법의 단점을 보완할 수 있다. 첫 번째로, 딥러닝 기반의 잡음 제거를 수행하기 때문에 잡음이 다양한 주파수 범위에서 변화하더라도 모델이 이를 고려하여 잡음 제거를 수행하므로 잡음 성분이 남을 수 있다는 필터링 기반의 단점을 보완할 수 있다. 두 번째로, 단일 모델을 통해 잡음 제거와 음향 이벤트 탐지 및 위치 추정을 동시에 수행할 수 있으므로, 단계가 분리되어 있던 기존의 방식과 비교하여 시간적 효율성이 향상될 수 있다.

제 2 장 배경 이론 및 관련 연구

본 장에서는 본 연구와 관련된 연구들의 배경 이론과 이와 관련된 연구들에 대하여 설명한다.

2.1 절에서는 음향 이벤트 탐지 및 위치 추정, 음성 향상, 그리고 각 작업의 평가 지표에 대하여 설명한다.
2.2절에서는 음향 이벤트 탐지 및 위치 추정, 음성 향상과 관련된 연구를 설명한다.

1절 배경 이론

1. 음향 이벤트 탐지 및 위치 추정

음향 이벤트 탐지 및 위치 추정(sound event localization and detection)는 주어진 음향 신호 속에서 특정한 이벤트를 탐지하고 위치를 추론하는 것을 목표로 한다. 복잡한 음향 환경에서 다양한 소리를 분리하고 식별하는데 중요한 역할을 하는 신호 처리 기술로, 시간적, 공간적 패턴을 파악하고 이를 통해 음향 환경을 더 잘 이해할 수 있도록 한다.

음향 이벤트 탐지 및 위치 추정은 크게 두 가지 작업으로 나눌 수 있는데, 음향 이벤트 탐지(sound event detection)와 음원 위치 추정(sound source localization)이다.

음향 이벤트 탐지는 대부분 각 음향 이벤트 클래스의 프레임 별 활동을 예측하는 다양한 지도학습 분류 방법을 사용하였다. 은닉 마르코프 모델(Hidden

Markov models; HMM)[16], 완전 연결 신경망(fully connected neural networks)[17], 순환 신경망(recurrent neural networks; RNN)[9], 합성곱 신경망(convolutional neural networks; CNN)[18] 등을 사용한 연구들이 존재한다. 최근에 순환 신경망과 합성곱 신경망을 결합한 합성곱 순환 신경망(convolutional recurrent neural networks; CRNN)을 사용한 높은 성능을 얻는 음향 이벤트 탐지 시스템 연구가 진행되었다[19].

음원 위치 추정은 대체로 통계적 방법과 심층 신경망 기반 방법으로 나눌 수 있다. 파라메트릭 방법으로 도착 시간 차이(Time Difference of Arrival; TDOA)[20], SRP(steered-response-power)[21], MUSIC(Multiple Signal Classification)[22], ESPRIT(estimation of signal parameters via rotational invariance technique)[23]을 기반으로 수행되었다. 이후, 파라메트릭 방법의 일부 단점을 극복하기 위한 방법으로 심층 신경망 기반 방법이 사용되었다[24].

그러나 독립적인 접근 방법을 통해 음향 이벤트 탐지 및 위치추정을 수행할 경우, 이벤트 탐지와 위치 추정 사이의 데이터 연관 문제가 발생한다는 단점이 있다[25]. 예를 들어, 음향 이벤트 탐지 과정에서 특정 이벤트가 발견되면 해당 이벤트의 위치를 추정하는 음원 위치 추정에서 유용한 정보가 될 수 있으나 과정이 독립적일 경우 이러한 정보를 활용할 수 없다. 즉, 탐지에 대한 정보와 위치에 대한 정보의 통합과 최적화가 어렵다. 이에 대한 해결책으로 음향 이벤트 탐지와 위치 추정을 동시에 수행하는 것이다. 이를 위해 심층 신경망을 기반으로 동시에 수행하는 방법[26]과 음향 이벤트 탐지와 파라메트릭 방식을 사용한 위치 추정을 결합한 연구[11,27]들이 제안되었다.

딥러닝 기술이 발달함에 따라, 이를 기반으로하는 음향 이벤트 탐지 및 위치 추정에 대한 연구가 가속화되고 있다. 탐지하는 이벤트의 수를 증가하거나, 기존에 제안된 심층 신경망 모델의 단점을 극복하기 위한 방법들이 제시되고 있다. 대표적인 모델로 SELDnet[12]이 있으며, 출력 방식에 대한 연구로 ACCDOA[28], Multi-ACCDOA[29]가 있다.

2. 음성 향상

음성 향상(speech enhancement)은 발화의 품질을 상승시키기 위해 오디오 신호 처리 기술을 사용하여 잡음이 존재하거나 잔향이 존재하는 음원의 품질을 향상시키는 작업이다. 얻고자 하는 오디오 신호, 즉 잡음이 없는 오디오 신호 $x[n]$과 잡음 신호 $n[n]$의 합으로 구성된 신호 $y[n]$이 존재할 때, 음성 향상은 $y[n]$을 $x[n]$과 유사하게 복원시키는 것이 목표이다.

이를 위해 과거 위너 필터(Wiener filter)[30], 매치드 필터(Matched filter)[31], 칼만 필터(Kalman filter)[32]와 같은 통계적 방법들을 적용한 음성 향상 기술을 사용하여 진행되었으나 딥러닝 기술의 발달로 기존 연구보다 뛰어난 성능을 보여주는 연구들이 발표되며 딥러닝 기반의 음성 향상 연구들이 수행되고 있다.

딥러닝을 활용한 시간 도메인에서의 음성 향상 연구는 잡음이 포함된 오디오 신호를 입력 특징(feature)으로 사용하여 이를 인코딩하고, 인코딩된 특징 맵(feature map)을 디코딩하여 잡음이 제거된 오디오 신호를 출력하는 오토 인코더 (auto-encoder) 모델을 주로 사용하였다. 음성 향상에 사용되는 오토 인코더 모델로는 인코더의 중간 특징 맵을 디코더의 특징 맵과 합쳐 추가적인 입력으로 사용하는 U-net[33] 구조가 대표적이다. 이러한 오토인코더 모델은 음성 향상 뿐만 아니라 이미지 향상에도 적용할 수 있다. 이러한 향상 작업을 수행하는 모델로 Wave-U-net[34], Residual U-net[35] 등이 존재한다.

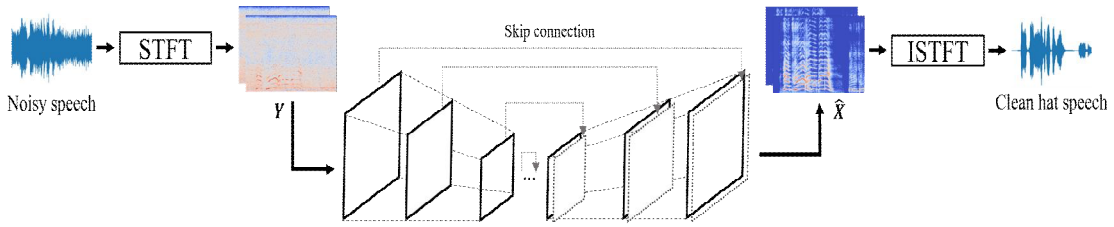


그림 2.1 오토 인코더 모델을 활용한 음성 향상

3. 음향 이벤트 탐지 및 위치 추정 평가 지표

딥러닝 기반의 음향 이벤트 탐지 및 위치 추정에서 사용되는 평가 지표는 음향 이벤트 탐지에 해당하는 평가 지표(Error rate, F1-score)와 음원 위치 추정에 해당하는 평가 지표(Localization error, Localization recall)가 있다.

인공지능 기반 음향 이벤트 및 장면 인식 기술 경진 대회(Detection and Classification of Acoustic Scenes and Events; DCASE)에서 두 작업의 성능을 공동으로 평가하기 위한 평가 지표를 제시하였다[36]. 또한, 최종적으로 4가지 평가 지표를 모두 고려한 평가 지표(SELD score)를 제시하였다. 본 연구에서는 해당 지표를 사용하여 평가를 진행하였으며, 본 절에서는 해당 평가 지표들에 대해 설명한다.

Error rate는 참조(reference)에서 활성화된 음향 이벤트의 수 N에 상대적인 치환(substitutions; S), 삽입(insertions; I), 삭제(deletions; D)의 총 개수를 계산하여 측정된다. 여기서 S는 시스템 출력이 잘못된 라벨 이벤트를 활성화로 표시하는 경우를, I는 N을 제외한 위양성(false positive; FP)을 나타내며, D는 S를 제외한 위음성(false negative; FN)을 의미한다. S, I, D는 다음과 같이 정의되며, 여기서 k는 각각의 1초 세그먼트(segment) 수를 의미한다.

$$S(k) = \min(FN(k), FP(k)), \tag{2.1}$$
$$D(k) = \max(0, FN(k) - FP(k)), \tag{2.2}$$
$$I(k) = \max(0, FP(k) - FN(k)), \tag{2.3}$$

최종적인 error rate는 각 세그먼트를 모두 더한 총 세그먼트 수 K로 나누어 계산되며 다음과 같이 정의된다.

$$Error\ rate = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)}. \tag{2.4}$$

F1-score는 정밀도(Precision; P)와 재현율(Recall; R)을 사용하여 계산되며, P, R, 그리고 F1-score는 다음과 같이 계산된다.

$$P = \frac{\sum TP(k)}{\sum TP(k) + \sum FP(k)}, \tag{2.5}$$

$$R = \frac{\sum TP(k)}{\sum TP(k) + \sum FN(k)}, \tag{2.6}$$

$$F = \frac{2P \cdot R}{P + R}. \tag{2.7}$$

여기서 $TP(k)$, $FP(k)$, $FN(k)$은 각각 k번째 세그먼트에서의 참양성(True positive; TP), 위양성, 위음성을 의미한다. 위치 종속적인 음향 이벤트 탐지 평가를 위해, 시스템의 예측 위치와 실제 위치 사이의 각도 차이가 $20°$ 이내인 경우 TP로 판단하고, $20°$를 초과하는 경우에는 FN으로 평가한다.

음원 위치 추정에 대한 평가 지표로 localization error와 localization recall이 사용되었다. 음향 이벤트 탐지를 고려한 위치 추정을 위해 클래스 종속 평가 지표를 사용한다. class-aware localization을 기반으로 계산되며, class-aware localization error(LE_c), class-aware localization recall(LR_c)은 모든 예측 이벤트와 참조 이벤트 사이에서 계산된다. LE_c와 LR_c은 다음과 같이 계산된다.

$$LE_c = \frac{\| A_c \odot D_c \|_1}{\| A_c \|_1}, \tag{2.8}$$

$$LR_c = \frac{\sum_l \| A_c^{(l)} \|_1}{\sum_l N_c^{(l)}}. \tag{2.9}$$

여기서, $c \in [1, \cdots, C]$는 클래스 인덱스를 나타내며, $l = 1, \cdots, L$은 데이터가 시

간적으로 분할된 경우 각 분할 단위를 지칭한다. D_c는 $M_c \times N_c$ 거리 행렬(distance matrix)을 나타내며, M_c이 클래스 c에 대한 예측의 수를 나타내는 변수, N_c가 참조 이벤트 수를 나타내는 변수이다. A_c는 행렬 D_c에 대한 연관 행렬(association matrix)로 $H(D_c)$로 나타내며, $H(\bullet)$는 헝가리안 알고리즘(Hungarian algorithm)을 나타낸다. 모든 예측 이벤트를 고려한 LE_c와 LR_c를 토대로 같은 클래스에 속할 때만 고려하는 LE_{CD}와 LR_{CD}를 구할 수 있다.

$$LE_{CD} = \frac{1}{C \bullet L} \sum_{c} \sum_{l} \leq {}_{c}^{(l)}, \tag{2.10}$$

$$LR_{CD} = \frac{1}{C} \sum_{c} LR_c. \tag{2.11}$$

마지막으로 음향 이벤트 탐지에 대한 평가 지표(error rate, f1-score)와 음원 위치 추정에 대한 평가 지표(localization error, localization recall)의 성능을 통틀어 전반적인 성능을 나타내기 위한 SELD score라는 지표를 제시하였으며 다음과 같이 계산된다.

$$SELD\ score = \frac{ER\ +\ (1-F-score)\ +\ (LE/180)\ +\ (1-LR)}{4} \tag{2.12}$$

4. 음성 향상 평가 지표

딥러닝 기반의 음성 향상 평가 모델로는 MOSNet[37]이 존재한다. MOSNet은 음성 변환 챌린지에서 청취자에 의해 평가된 샘플들의 MOS 결과를 학습에 사용하여, MOS를 예측하기 위한 딥러닝 기반의 평가 모델로 1과 5사이의 값을 가진다.

그러나, 본 연구에 적용한 딥러닝 기반의 음성 향상은 멜 스펙트로그램(mel-spectrogram)과 강도 벡터(intensity vector)를 입력으로 수행되며, 제안한 모델의 출력 값으로 2차원 이미지를 예측한다. 그렇기에 이미지 향상에 사용하는 평가지표를 사용하여 모델을 평가한다. 이에 대한 평가 지표로 PSNR(Peak Signal-to-Noise Ratio), SSIM(Structural Similarity Index)[38], 그리고 평균

제곱 오차(Mean Squared Error, MSE) 등이 있으며, 본 절에서는 연구를 수행하며 적용한 평가 지표에 대해 설명한다.

PSNR는 이미지나 동영상과 같은 신호에서 재생성 오차를 측정하는데 주로 사용된 평가지표이다. 원본 신호나 이미지와 잡음이 있는 신호나 이미지 사이의 품질 손실 측정에 사용되며, 수치가 높을수록 재생성 오류가 더 적게 발생하며 복원의 품질이 더 높다는 것을 의미한다. 최대 신호 대 잡음비는 다음과 같이 계산되며 MAX_I는 가능한 최대 픽셀 값, MSE는 원본 이미지와 복원된 이미지 사이의 픽셀간 평균 제곱 차이를 의미한다.

$$PSNR = 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE) \tag{2.13}$$

SSIM은 디지털 이미지 품질을 측정하는 방법 중 하나로, 두 이미지의 구조적 유사성을 측정하는 평가지표이다. 두 이미지의 구조적 유사성을 측정하기 위해 밝기, 대조, 구조 세 가지 요소를 고려하여 구조적 변화를 정량적으로 측정한다. SSIM은 다음과 같이 계산된다.

$$SSIM = l(x,y)^{\alpha} c(x,y)^{\beta} s(x,y)^{\gamma} \tag{2.14}$$

여기서 $l(x,y)^{\alpha}$는 두 이미지의 밝기를 비교한 결과, $c(x,y)^{\beta}$는 두 이미지의 대조를 비교한 결과, $s(x,y)^{\gamma}$는 두 이미지의 구조를 비교한 결과를 나타낸다. α, β, γ는 각 요소의 중요도를 조절하는 가중치로 일반적으로 1로 설정한다.

MSE는 예측 값과 실제 값 사이의 차이를 제곱하여 평균을 구한 값으로 회귀 분석, 이미지나 신호 처리 분야에서 활용되는 평가 지표이다. MSE는 각 차이를 제곱하기 때문에, 오차가 클수록 그 값이 증가하며 이상치에 민감하게 반응하며 다음과 같이 표현된다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2 \tag{2.15}$$

여기서 n은 데이터 포인트의 총 개수를 나타내며, Y_i는 실제 값, $\hat{Y}_i$는 예측값을 나타낸다.

2절 관련 연구

본 절에서는 본 연구와 관련된 연구를 소개한다.

2.1 절에서는 음향 이벤트 탐지 및 위치 추정 관련 연구를 다룬다. 음향 이벤트 탐지와 위치 추정이라는 하위 작업으로 구성되기 때문에 각 하위 작업의 관련 연구와 이 두 작업을 결합한 작업의 관련 연구를 다룬다. 또한, 딥러닝 기반 방법에서 소개된 입력 특징이나, 출력 형식에 관하여 서술한다.

2.2 절에서는 위너 필터, 매치드 필터, 칼만 필터와 같은 통계적 방법의 음성 향상 기법과 U-net, Wave-U-net, Residual U-net과 같은 딥러닝 기반의 음성 향상 모델에 대해서 서술한다.

1. 음향 이벤트 탐지 및 위치 추정 연구

음향 이벤트 탐지는 음향 이벤트 탐지와 음원 위치 추정이라는 두 가지 하위 작업으로 구분된다. 음향 이벤트 탐지 작업은 대부분 각 음향 이벤트 클래스의 프레임 별 활동을 예측하는 지도 학습 분류 방법을 주로 사용하였다.

은닉 마르코프 모델은 상태의 순차적 변화를 모델링하는데 유용한 통계적 모델로, 각 상태는 Markov 프로세스에 따라 변하며 각 상태는 관측 가능한 출력을 생성한다. 각 상태는 음향 이벤트를 나타낼 수 있으며, 전이 확률은 한 이벤트에서 다른 이벤트로의 전환을 나타낸다. 시간에 따른 상태 변화를 모델링하므로, 시간에 따른 패턴이 중요한 음향 이벤트 탐지에서 유용하며 불확실성을 직접 처리할 수 있는 확률적 모델이다[16].

서포트 벡터 머신은 분류 문제를 해결하는 머신러닝 알고리즘으로 각 음향 이벤트의 특성을 나타내는 특징 벡터를 입력으로 받아, 이벤트가 발생 여부를 예

측한다. 학습 데이터에서 결정 경계(hyper plane)를 최대화하므로, 일반화 능력이 뛰어나다는 장점이 있다[39].

가우시안 혼합 모델(Gaussian mixture model; GMM)은 여러 개의 가우시안 분포의 가중치 합을 사용하여 데이터를 모델링하는 방법으로, 다양한 음향 특성들을 추출하고, 이러한 특성에 대해 GMM을 학습시키는 방식으로 사용된다[16,40].

음원 위치 추정을 위한 방법으로 도착 시간 차이, MUSIC 등 통계적 방법들과 심층 신경망 기반의 방법들이 제안되었다.

도착 시간 차이는 두 개 이상의 센서를 이용하여, 각 마이크로폰에서 소리가 도착하는 시간 차이를 측정한 후, 이를 바탕으로 삼각측량으로 음원의 위치를 추정하는 방법이다. 신호의 위치를 상당히 정확하게 추정할 수 있으며 복잡한 계산이나 신호 처리 기법을 필요로 하지 않으나 신호가 장애물을 통해 반사하거나 회절하는 경우, 실제 신호의 도착 시간과 측정된 시간 사이에 차이가 생길 수 있어 정확도가 감소할 수 있다[20].

MUSIC 알고리즘은 고유벡터 분해에 기초한 DOA 추정의 한 종류로, 공간 구조에 기반을 두고 있다. 이는 복수의 배열 요소에서 수신된 다중 신호의 상관 행렬을 구하고, 고유 값 분해를 통해 신호 부분과 잡음 부분의 고유 값으로 분류하는 과정을 포함한다. 분류된 고유값 중에서, 잡음에 관련된 고유 값을 이용하여 잡음 공간 고유벡터를 만든다. 이 고유벡터를 이용하여 MUSIC 스펙트럼을 계산하고, 이 과정을 통해 신호의 DOA를 추정하게 된다[22].

통계적 방법들은 복잡한 알고리즘을 가지며, 신호 대 잡음비(Signal to Noise Ratio; SNR) 시나리오에 민감하다는 단점이 있다[41]. 이러한 단점을 극복하기 위해 심층 신경망 기반의 방법들이 사용되었으며[2,42,43], 이 방법은 작업의 확장에도 매끄럽게 통합할 수 있다는 장점이 있다.

신경망 기반의 방법은 그림 2.2와 같이 이벤트 탐지와 위치 추정이라는 두 가지 목표에 대해 두 개의 분기를 사용하는 two-branch 방식[12]과 그림 2.3과 같이 각 작업을 위한 별도의 네트워크를 사용하는 two-stage 방식[43]이 있다. Two-branch 방식은 다중 목표를 단일 네트워크로 해결할 수 있어 네트워크의 크기가 작다는 장점이 있으나, 두 목표에 대한 손실과 가중치가 선형적으로 결합되어 조정이 어려우며 가중치에 따라 성능이 큰 영향을 받는다는 단점이 있다. Two-stage 방식은 각 목표에 해당하는 네트워크를 구별하여 사용하기 때문에 다중 목표 문제를 해결할 수 있다는 장점이 있으나, 시스템의 복잡성과 네트워크의 크기가 증가한다는 단점이 있다.

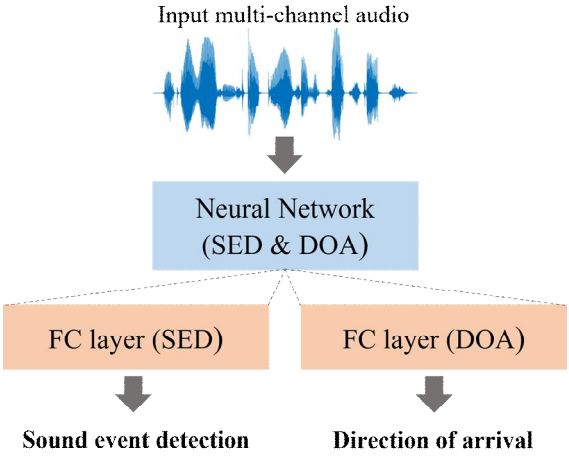


그림 2.2 Two-branch 방식의 모델 구조

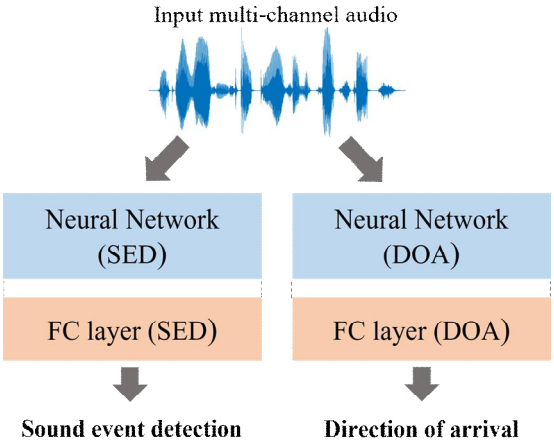


그림 2.3 Two-stage 방식의 모델 구조

이러한 문제를 해결하기 위해, ACTIVITY-COUPLED CARTESIAN DIRECTION OF ARRIVAL (ACCDOA) 출력형식이 제안되었다[28]. ACCDOA 출력형식은 분기가 없는 1개의 모델로 음향 이벤트 탐지 및 위치 추정을 수행하기 위해 제안된 출력형식으로 음원의 위치를 나타내는 Cartesian DOA 벡터에 음원의 활성 여부를 할당하는 방식이다.

ACCDOA는 다음과 같이 표현이 가능하다. 어떤 음향 이벤트가 활성화 되었는지를 나타내는 행렬 a는 수식 2.16과 같이 표현될 수 있으며 각 활동의 참조 값은 이벤트가 활성 되었을 때 1을 가지고 비활성일 때 0이며, 수식 2.17과 같다. 여

기서 C는 클래스, T는 시간 프레임을 의미한다.

$$a \in \mathbb{R}^{C \times T} \tag{2.16}$$

$$a_{ct}^{*} \in \{0,1\} R \tag{2.17}$$

Cartesian DOA 행렬 R은 수식 2.18과 같이 표현되며, 이때 클래스 c의 활성 행렬 a가 활성되어 있을 때, 각 Cartesian DOA 행렬의 길이는 1이다. 여기서 $||\cdot||$는 유클리드 거리를 의미한다. 각 이벤트 클래스 C는 x, y, z축에 대응하는 위치를 나타내는 노드로 표현된다.

$$R \in \mathbb{R}^{3 \times C \times T} \tag{2.18}$$

$$||R_{ct}|| = 1 \tag{2.19}$$

음원의 활동과 Cartesian DOA를 갖는 ACCDOA 행렬 P는 수식 2.21과 같다.

$$P \in \mathbb{R}^{3 \times C \times T} \tag{2.20}$$

$$P_{ct} = a_{ct} R_{ct} \tag{2.21}$$

그림 2.4는 이러한 표현 간의 관계의 예시를 나타낸다. 그림 내의 ACCDOA 표현에서 음향 이벤트가 비활성화 된 경우, x, y, z에 해당하는 값이 0으로 설정되어 있으며, 활성된 경우에만 해당 위치에 대응되는 값을 갖는 것을 확인할 수 있으나 ACCDOA는 각 이벤트 클래스마다 벡터를 만들기 때문에 동일한 클래스의 중복된 음원이 발생하게 되는 경우, 이를 탐지해낼 수 없다는 단점이 있다.

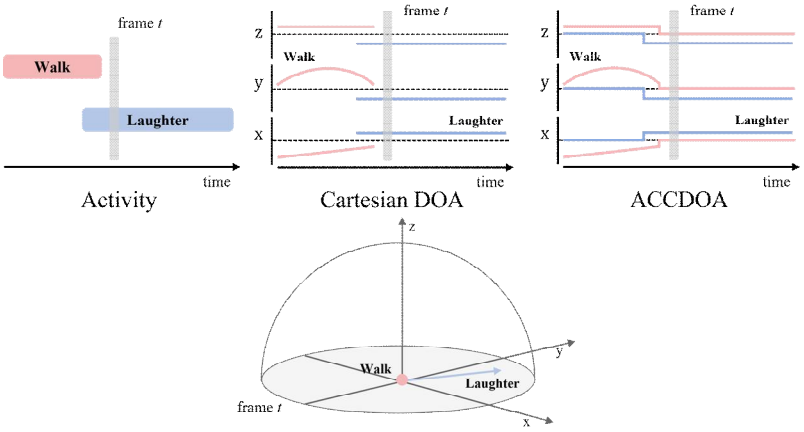


그림 2.4 ACCDOA 출력 방식의 예시

ACCDOA의 문제를 보완하기 위해 Multi-ACCDOA 출력형식이 제안되었다 [30]. 각 이벤트 클래스에 하나의 위치만 할당하는 Class-wise 형식의 장점을 유지하며 동일한 클래스의 중복된 음원 탐지를 위해 제안된 출력형식으로 ACCDOA를 트랙 차원으로 확장된 Class-and track-wise 형식이다. Multi-ACCDOA 형식의 각 트랙은 ACCDOA 형식과 동일하며, Multi-ACCDOA 행렬 P는 수식 2.23과 같이 나타낼 수 있다.

$$P \in \mathbb{R}^{3 \times N \times C \times T} \tag{2.22}$$

$$P_{nct} = a_{nct} R_{nct} \tag{2.23}$$

여기서 N은 트랙, C는 클래스, T는 프레임을 의미한다. 트랙의 각 음향 이벤트 클래스는 x, y, z축에 해당하는 세 개의 노드로 표현된다. 그림 2.4는 Multi-ACCDOA의 예시를 나타낸다. 동일한 시간대에 동일한 클래스의 음원이 다중으로 발생해도 트랙마다 이벤트를 할당하여 음향 이벤트를 탐지할 수 있다.

Multi-ACCDOA에서 track-wise 방식과 유사하게 트랙 순열 문제가 존재한다. 트랙 순열 문제란, 복수의 신호를 동시에 추적하려 할 때 어떤 추정값이 어떤 신호의 음원에 해당하는지 결정하는 문제를 의미한다. Multi-ACCDOA에서는 이러한 문제를 해결하기 위해 학습 과정에 permutation invariant training(PIT)를 적용한다.

PIT는 여러 출력값 사이의 대응 관계가 명확하지 않은 경우, 사용되는 학습 방식으로, 가능한 모든 순열에 대해 손실을 계산하고 가장 작은 손실 값을 갖는 순열을 선택하여 학습을 진행하는 방법이다. Class-wise PIT를 사용할 경우, 각 클래스별로 하나의 활성 이벤트를 하나의 트랙에 할당하며 이벤트가 트랙의 수보다 적을 경우에는 보조 타겟으로 0벡터를 할당한다. 그러나 이 0벡터를 할당할 경우, 해당 트랙은 이벤트가 활성된 단위 벡터를 목표로 학습되는 다른 트랙과 달리, 이벤트가 존재하지 않는다는 목표를 갖고 학습이 진행되기 때문에 최적화를 방해하는 요소가 될 수 있다.

이를 위해 Multi-ACCDOA에서는 모든 트랙이 단위 벡터를 목표로 학습을 수행할 수 있도록 원본 타겟을 복제하여 보조 타겟에 할당하는 auxiliary duplicating permutation invariant training (ADPIT)를 제안한다. N개의 트랙을 갖는 Multi-ACCDOA 형식은 클래스 c와 프레임 t에서 $M_{ct}(\leq N)$개의 원본 타겟과 $N-M_{ct}$보조 타겟을 출력하며 PIT를 적용하여 손실이 적은 최상의 순열을 찾는다. 그림 2.5는 그림 2.6의 예시에 해당하는 모든 순열을 나타낸다.

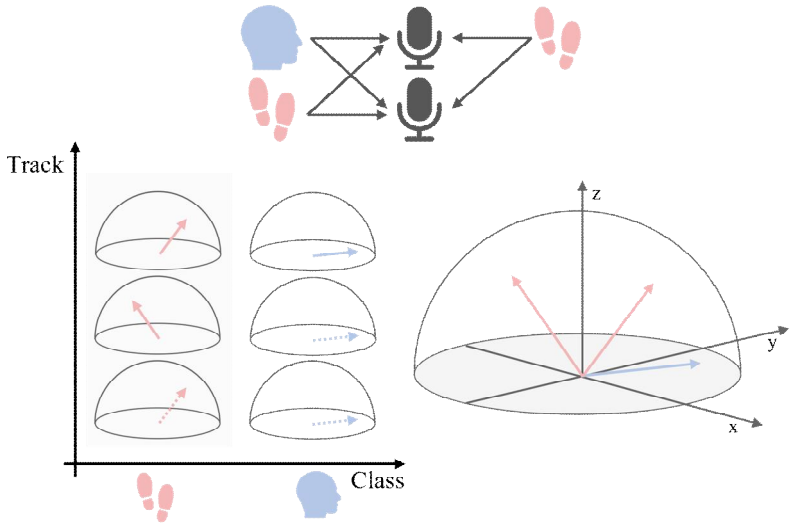


그림 2.5 Multi-ACCDOA 출력 형식 예시

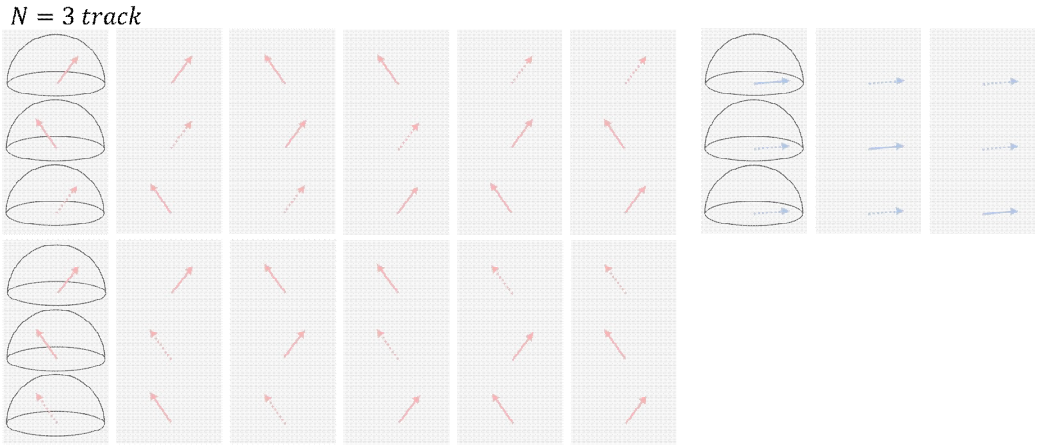


그림 2.6 예시(그림 2.5)에 해당하는 ADPIT에 대한 모든 순열 조합

2. 음성 향상 연구

U-net[33]은 2015년에 발표된 합성곱 신경망 구조 중 하나이다. 입력 이미지를 점차적으로 축소하는 인코딩 경로를 사용하며, 세밀한 위치 정보를 잃게 되는 단점을 보완하기 위한 디코딩 경로를 사용하여 높은 해상도 특성을 복원한다. 이미지 분할(image segmentation) 작업을 위해 고안된 모델이나, 이를 음성 향상에도 적용이 가능하다. 오디오 신호를 인코딩하여 특징을 추출하고, 특징을 기반으로 원래의 신호를 복원하며 원하는 신호 성분만 복원하는 것이 가능하다.

Wave-U-net[34]은 2018년에 발표된 논문이다. 기존 모델은 대부분 진폭 스펙트럼 (manitude spectrum)을 통해 수행되었고 위상(phase) 정보를 고려하지 않고 스펙트럼의 전처리에 의존하는 문제를 해결하기 위해 제안된 모델이다. U-net을 1차원 시간 도메인에 적용하였으며, 다른 시간 척도에서 특징을 계산하고 결합하기 위한 재샘플링(resampling), 출력 아티팩트를 줄이는 컨텍스트 인식 예측 프레임워크(context-aware prediction framework) 등을 도입하였으며, 이를 통해 고정된 스펙트럼 변환의 제한을 극복할 수 있다는 장점을 갖는다.

Residual U-net[35]은 2018년에 발표된 논문으로, 잔차 학습(residual learning)과 U-net의 장점을 결합한 신경망이다. 잔차 유닛(residual unit)을 통해 깊은 네트워크의 학습을 용이하게 하며, 스킵 커넥션(skip connection)을 통해 더 적은 매개변수로 너 나은 성능의 네트워크를 설계할 수 있는 장점을 갖는다.

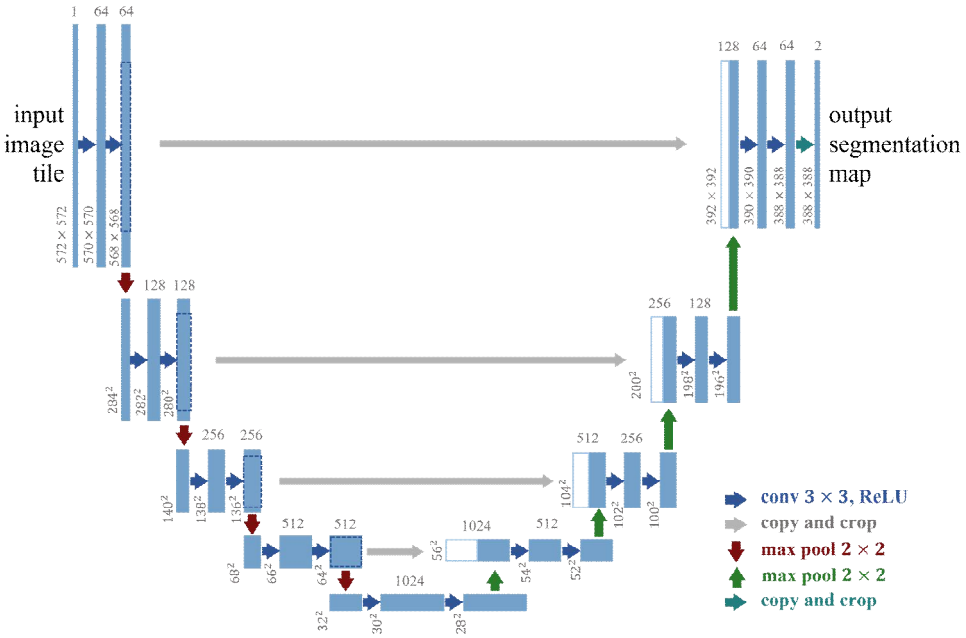


그림 2.7 U-net의 구조

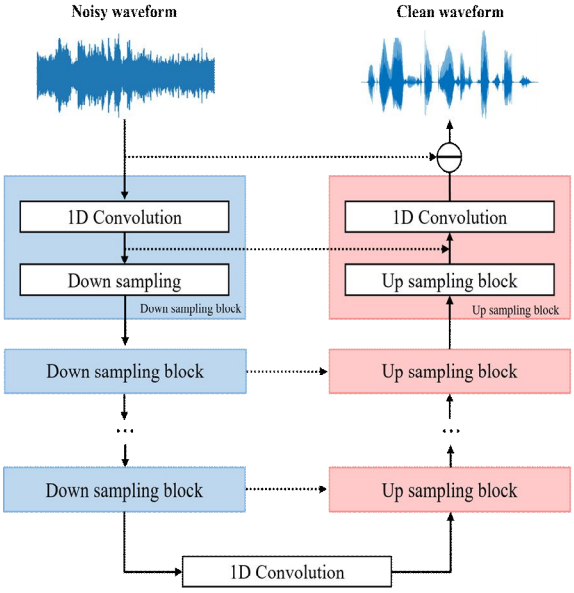


그림 2.8 Wave-U-net 구조

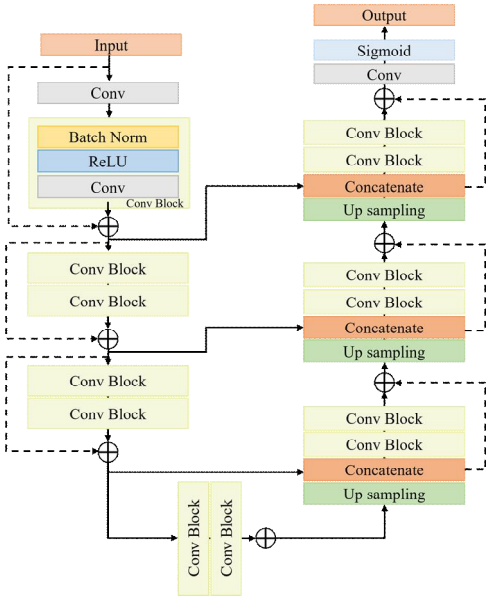


그림 2.9 Residual U-net 구조

제 3 장 제안 방법

본 연구에서는 기존 딥러닝 기반의 SELD 수행 방법이 잡음 환경의 오디오에서는 성능이 저하된다는 점과, 딥러닝 기반의 음성 향상 기술이 높은 성능을 보인다는 점을 통해 잡음 환경의 오디오 신호를 음성 향상 기술을 적용해 SELD를 수행하면 성능 저하 문제가 해결될 것이라고 가정하였다. 또한 Retina U-net[40]의 구조에서 착안하여 enhancement를 수행하는 U-net과 SELD를 수행하는 SELDnet을 결합하여 end-to-end 기반의 딥러닝 모델로 음성 향상과 SELD를 동시에 수행할 수 있을 것이라 가정하였다.

3.1절에서는 음성 향상을 수행하는 U-net과 SELD를 수행하는 SELDnet을 결합하여 전처리 과정과 SELD 수행과정을 분리하지 않고 한 번에 수행하는 SELD U-net을 제안하고 해당 모델의 구조에 대하여 설명한다.

1절 SELD U-net

제안하는 모델인 SELD U-net의 구조는 전단부의 음성 향상을 수행하는 U-net과 SELD를 수행하는 SELDnet을 결합한 구조의 모델로 그림 3.1과 같다. 기존 SELDnet의 입력 특징은 4채널의 멜 스펙트로그램(mel spectrogram)과 3채널의 강도 벡터(intensity vector)를 쌓아 7채널의 입력 특징을 사용하며 이에 관한 내용은 4.1.2절에서 논한다. 오디오 신호의 잡음 제거와 SELD를 별도로 수행할 경우, 잡음 제거 후, SELD의 입력 특징을 추출한 뒤 SELD 모델을 학습시키는 순서로 학습이 진행된다. 이 경우 두 번의 개별적인 모델 학습과 데이터 추출 과정을 거치기 때문에 시간 손실이 발생하게 된다. 이러한 문제를 해결하기 위해 잡음 제거와 SELD를 한 번에 수행하는 U-net과 SELDnet을 결합한 형태의 모델 구조를 제안한다. End-to-end 방식의 모델 구조를 사용함으로써 여러 단계의 처리 과정을 하나의 모델로 통합하여 시간 손실을 최소화할 수 있다.

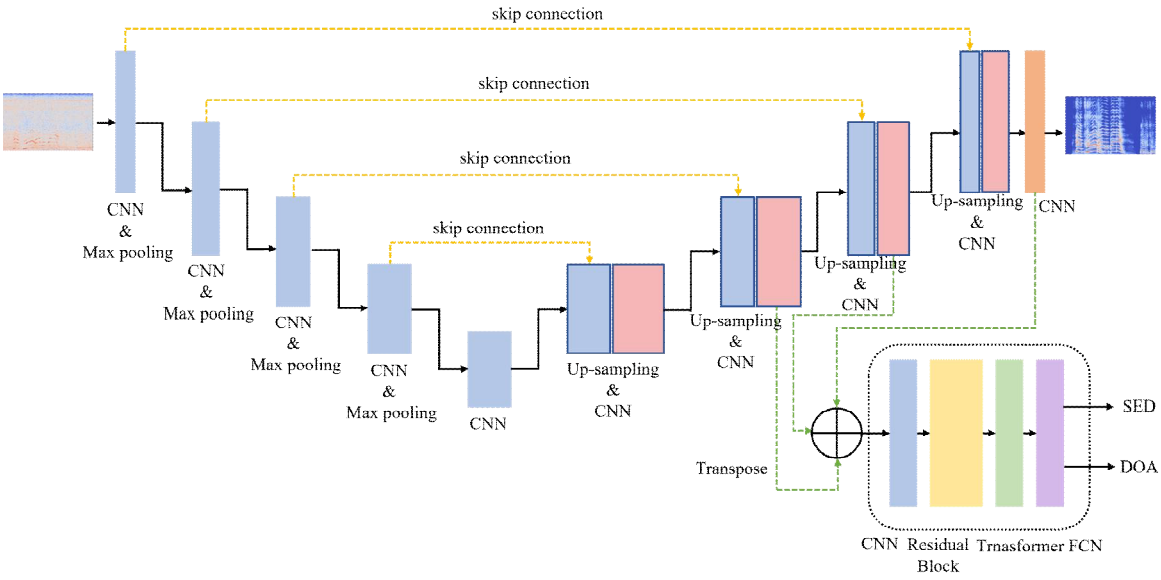


그림 3.1 제안하는 SELD U-net의 개략도

제안하는 모델은 U-net 디코더 부분의 일부 특징 맵을 합쳐 SELDnet의 입력 특징으로 사용하는 구조로 이루어졌다. U-net은 잡음이 환경의 오디오 신호에서 추출된 입력 특징을 인코더 부분을 통해 잡음을 제거하고 중요한 오디오 정보만을 포착하며 디코더에서 이를 고차원 오디오 신호로 복원하는 과정을 통해 잡음이 없는 신호에서 추출한 입력 특징을 추론한다. 이 과정에서 디코더의 특징 맵은 잡음이 제거된 신호의 핵심적인 정보가 포함되어 있는 특징으로 볼 수 있다. 잡음이 없는 오디오 신호의 핵심적인 정보를 사용하여 SELD를 수행하기 위해 디코더의 일부 특징 맵을 추출하여 SELDnet의 입력 특징으로 사용한다.

1. U-net

제안한 모델에 사용된 U-net의 구조로는 기본적인 U-net[33]과 Nested U-net[44], Residual U-net[35] 총 3가지 종류의 U-net이 사용되었다. Nested U-net은 U-net과 달리 skip pathway에 합성곱 층이 존재하여, 인코더와 디코더의 특징 맵 사이의 semantic gap을 연결해준다는 특징이 있다. Residual U-net

은 잔차 유닛을 사용하는 구조로 깊은 네트워크의 학습을 용이하게 한다.

U-net, Nested U-net은 5층의 깊이를 갖고, Residual U-net은 4층의 깊이를 갖는 구조로 설계되었다. 3 종류의 U-net은 동일한 입력 특징을 사용하였다. 잡음이 포함된 오디오에서 추출된 SELDnet의 입력 특징인 4채널의 멜 스펙트로그램과 3채널의 강도 벡터을 입력 특징으로 사용하였으며, 모델이 예측 값으로 잡음이 없는 오디오에서 추출된 4채널 멜스펙트로그램과 3채널의 강도 벡터를 복원하도록 학습이 진행되었다. 입력 특징의 핵심 정보를 갖는 디코더 부분의 일부 특징을 추출하여 SELDnet의 입력 특징으로 사용한다.

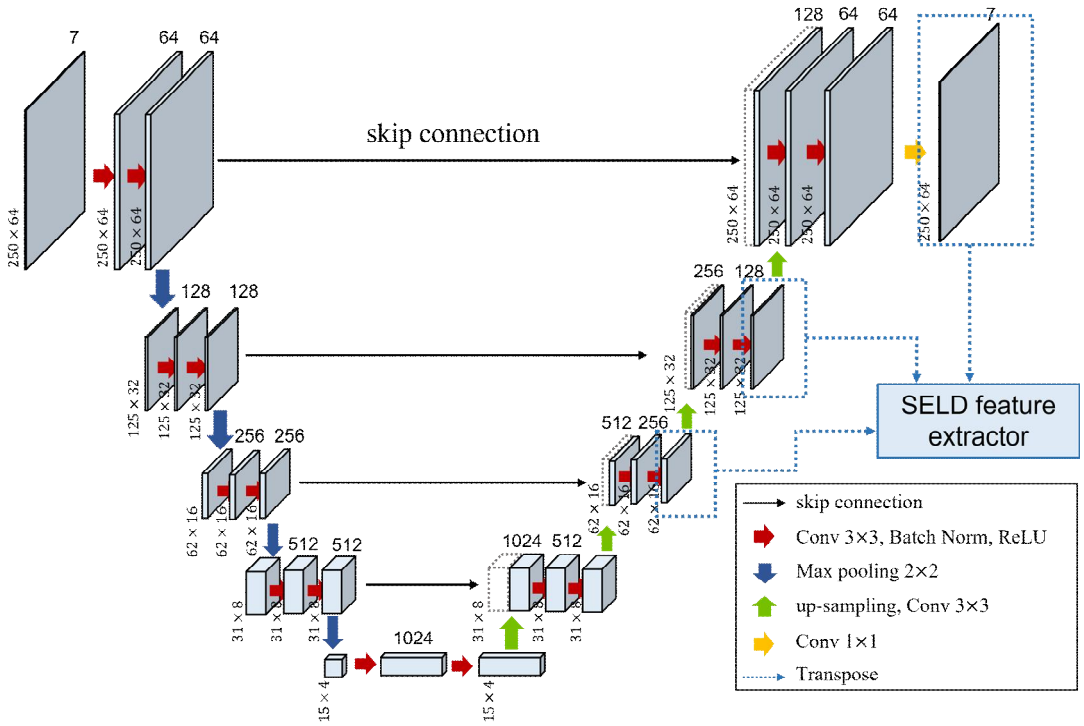


그림 3.2 제안하는 SELD U-net의 세부 구조(U-net)

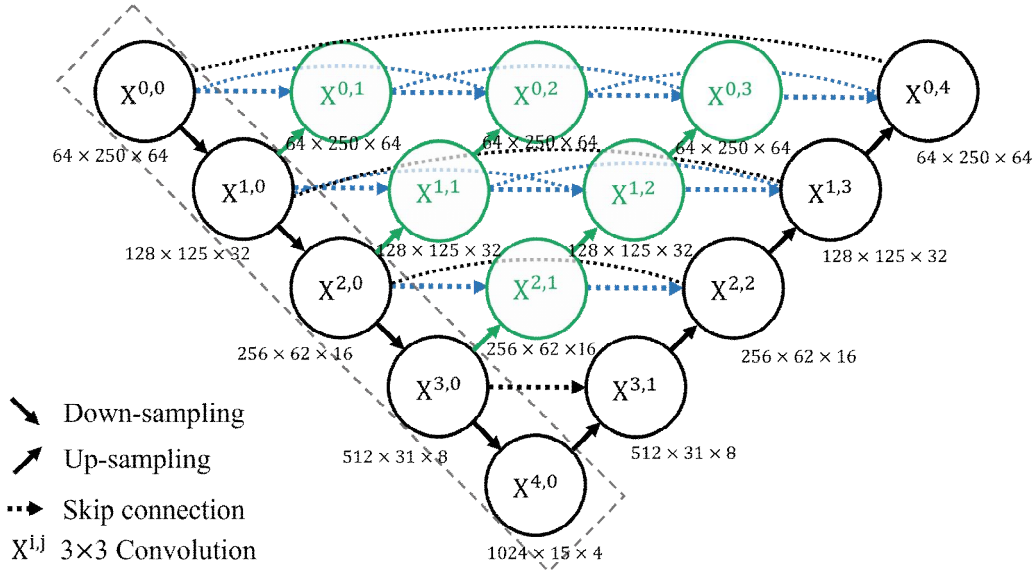


그림 3.3 제안하는 SELD U-net의 세부 구조(Nested U-net)

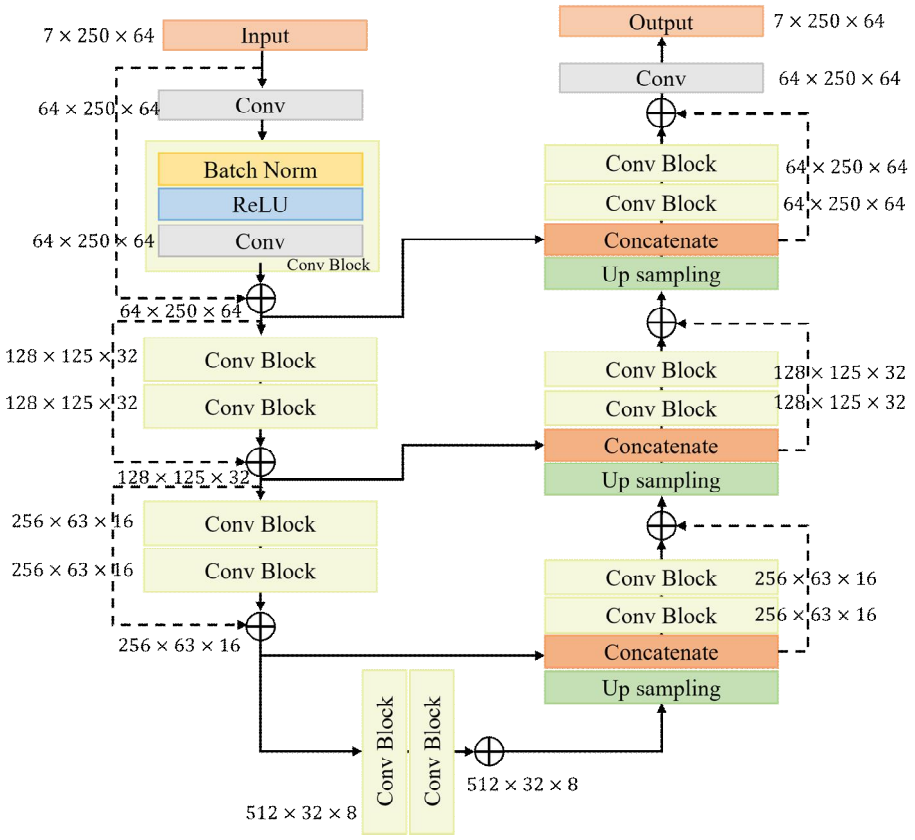


그림 3.4 제안하는 SELD U-net의 세부 구조(Residual U-net)

2. SELDnet

제안하는 SELDnet의 구조는 그림 3.5과 같다. SELDnet은 입력으로 전단부인 U-net의 디코더 부분의 일부 특징맵을 사용한다. 특징 맵을 추출하여 쌓은 후 SELDnet의 입력 특징으로 적용되기 때문에 각 특징 맵에 동일한 모양을 갖게 해주는 합성곱 연산이 수행된다.

사용된 SELDnet은 CRNN 구조의 SELDnet을 기반으로 설계되었다. 기존 CRNN 구조의 SELDnet은 CNN층과 RNN층을 결합한 구조로 CNN을 사용하여 오디오 신호의 특징을 추출하고 RNN을 통해 시간에 따른 정보의 흐름을 모델링하는 구조이다. 제안 모델은 CNN층을 잔차 합성곱 신경망(Residual convolutional neural network; RCNN)이 적용된 블록으로 교체하였으며, RNN층을 트랜스포머 인코더(Transformer encoder)로 교체하였다. 최종적인 예측 값을 생성하기 위해 완전 연결층(Fully connected layer)를 사용하였으며, Multi-ACCDOA 출력형식을 사용하여 음향 이벤트의 탐지와 위치 추정을 수행한다.

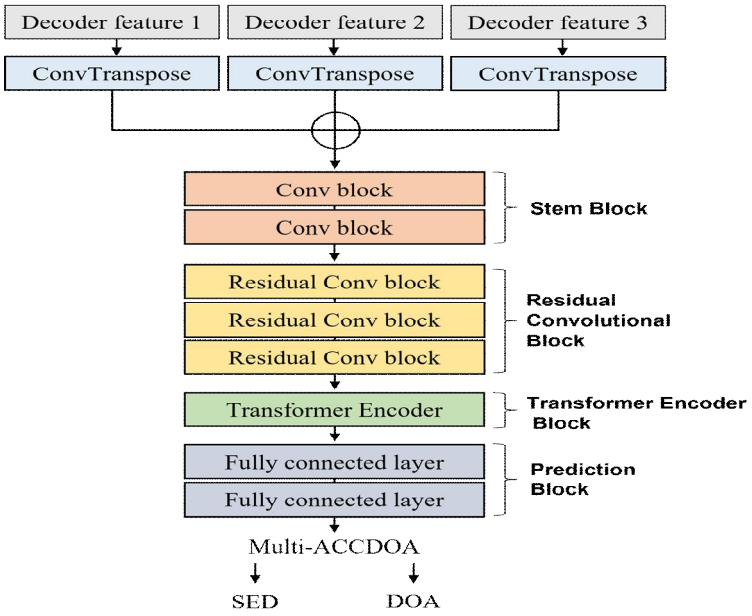


그림 3.5 제안하는 SELD U-net의 세부 구조(SELDnet)

제안하는 SELDnet의 잔차 블록은 그림 3.6의 구조로 구성된다. 모델의 입력 값과 출력 값을 더한 후, 다시 출력 값으로 매핑하는 방식으로 학습이 진행된다. 이러한 구조는 깊은 신경망 구조에서 발생하는 기울기 소멸 문제(vanishing gradient problem)를 해결할 수 있어 깊은 신경망 모델을 구축하는 것이 가능하다. 또한, 이전 층에서 추출된 특징 정보를 다음 층의 입력으로 전달하여 모델 학습 시 빠르게 수렴하는 효과를 기대할 수 있다. 이러한 장점들을 통해 CNN을 사용하는 구조와 비교하였을 때, 더 좋은 성능을 기대할 수 있다.

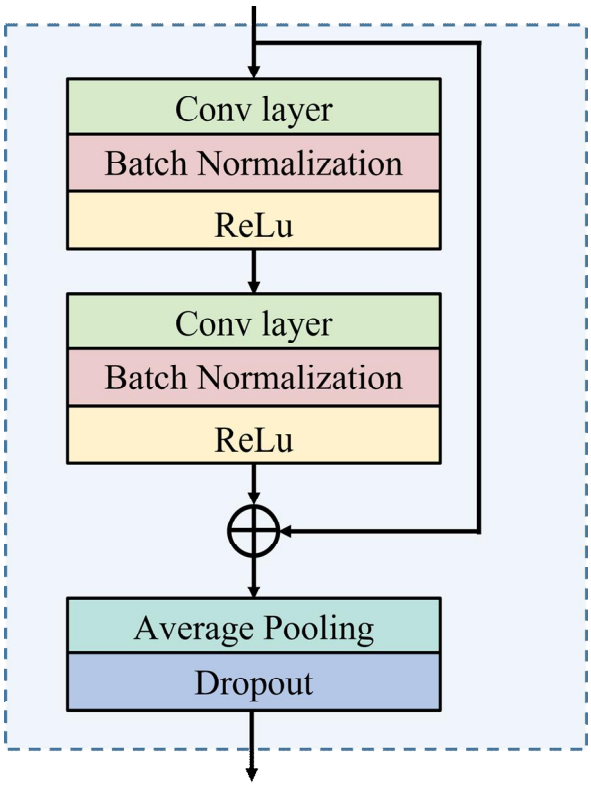


그림 3.6 Residual block의 세부 구조

트랜스포머 인코더는 그림 3.7의 구조와 같이 Positional Encoding 블록, Multi-Head Self-Attention 블록, Feed-Foward 블록으로 구성되며 이를 통해 입력 시퀀스(sequence)를 고차원적인 표현으로 변환하는 것이 가능하다.

Positional Encoding 블록에서는 입력 시퀀스의 각 요소에 시퀀스의 순서 정보를 인지시키며 사인 함수나 코사인 함수와 같은 주기 함수의 패턴을 사용하여 구현된다.

Multi-Head Self-Attention 블록은 입력 데이터의 다양한 표현을 포착하기 위해 설계된 구조로 입력 데이터에 대해 여러 번의 Attention 연산을 수행한다. 각각의 Head는 독립적인 Attention 메커니즘을 가지고 있으며, 이를 통해 각기 다른 가중치 세트를 사용하여 각 입력에 대한 Attention 연산이 수행된다. 각 Head가 생성한 출력 값들을 결합한 값이 최종 출력값으로 도출된다. 해당 연산을 통해 동일한 데이터에 다양한 관점에서의 Attention을 수행하여 복잡하고 다양한 의미를 포착하는 것이 가능하다.

Feed-Forward 블록은 Multi-Head Self-Attention 블록의 출력을 받아 선형 변환을 수행한다. Multi-Head Self-Attention 블록의 출력 값은 각 Head의 출력 값을 독립적 결합한 상태이기 때문에 정보들의 융합이 이루어지는 과정이 필요하며 해당 블록에서 이러한 과정을 수행한다. 선형 변환을 수행하여 각각의 출력 값들을 융합시키는 과정을 통해 모델은 데이터의 패턴을 학습할 수 있다.

시퀀스의 요소를 순차적으로 처리하는 RNN과 비교하여 트랜스포머 인코더는 입력 시퀀스를 동시에 처리하기 때문에 병렬처리가 가능하다는 장점이 존재한다. 또한, Self-Attention 메커니즘을 사용하여 입력 시퀀스의 모든 요소의 관계를 학습하기 때문에 장기의존성 문제(long-term dependency problem)를 해결할 수 있다.

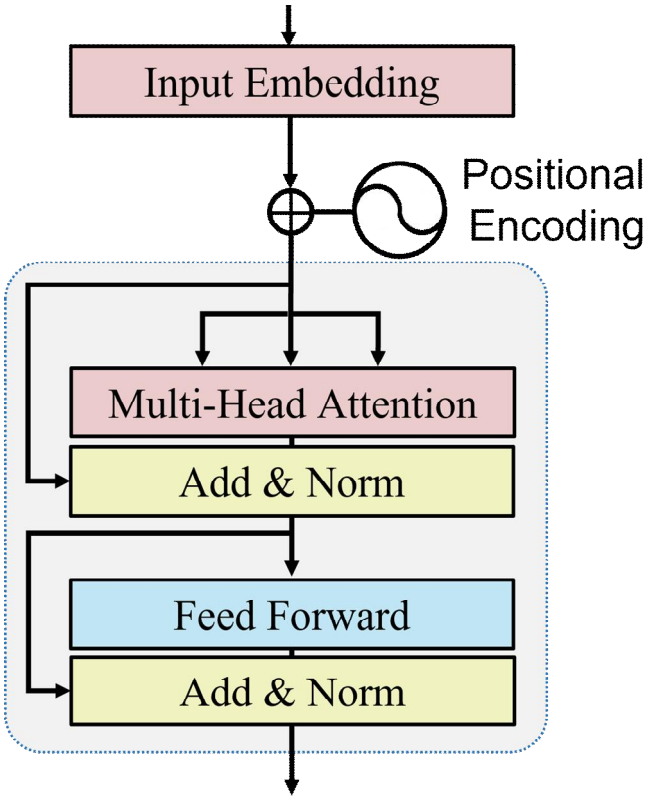


그림 3.7 Transformer encoder 구조

제 4 장 실험 수행

본 장에서는 이전 장에서 제안한 모델의 성능을 평가하기 위한 준비 과정과 결과를 서술한다.

4.1 절에서는 본 연구에서 사용한 데이터 셋과 입력 특징 추출 방법 및 적용한 증강 기법에 대하여 설명한다.

4.2절에서는 제안한 모델과의 성능 비교를 위한 비교 대상이 될 모델 구조에 대하여 설명한다.

4.3절에서는 실험 결과를 통해 제안한 모델의 성능을 파악하고 세부 분석을 수행한다.

1절 데이터셋 및 학습 환경 설정

1. 데이터셋 구축

제안된 모델은 음성 향상 모델 학습을 위해 잡음 없는 오디오 데이터와 동일한 오디오 데이터에 잡음이 추가된 데이터 쌍으로 구성되어야 한다. 이를 위해 음원 합성을 사용하여 데이터셋을 구축하였다. 합성 데이터 생성에 사용된 음원 데이터셋으로 Free-Sound dataset 50k (FSD50K)[45] 데이터셋을 사용하였다. FSD50K는 5만개 이상의 다양한 음원이 포함된 공개 데이터셋이다. 본 연구에서는 각 클래스마다 약 300개 가량의 음원을 사용하여 학습 데이터를 생성하였으며, 학습 데이터에 사용된 음원과 별도로 100개의 음원을 사용하여 평가 데이터셋을 구축하였다.

생성된 데이터셋은 FOA(First-order-Ambisonic)형식의 4채널 형식이며, 13종

류의 음원이 랜덤하게 발생한다. 각 오디오 데이터는 샘플링 레이트는 24kHz로 설정된 1분 길이의 데이터로 구성되며 학습을 위한 데이터 1,200개, 평가용 데이터 300개로 구성된다. 생성된 데이터 셋에 SNR 값을 다르게 설정하여 잡음 데이터를 생성한다. 이때 설정된 SNR 값은 +30, +20, +10, -10, -20, -30으로 이에 해당하는 데이터셋을 구축하였다.

2. 입력 특징 및 증강 기법

제안된 모델의 입력 특징으로는 잡음 환경의 오디오 데이터에서 추출한, 4채널의 멜 스펙트로그램과 3채널의 강도 벡터를 사용한다.

멜 스펙트로그램은 오디오 신호의 스펙트럼에 멜 스케일(mel scale)을 적용하여 2차원 이미지 형태로 시각화한 정보이다. 인간의 청각 특성을 반영한 멜 스케일을 사용하여, 사람의 청각과 가장 유사한 오디오 특성을 제공한다. 시간, 주파수, 에너지에 대한 정보를 모두 제공하며 시간에 따른 이벤트의 지속성과 변동성, 이벤트 주파수 성분 등 오디오 처리 작업에서 유용한 특징들을 제공한다.

강도 벡터는 방향성 오디오 신호의 특징으로, 특정 시간에 음원이 어느 방향에서 발생했는지를 나타내는 정보이다. 오디오 신호에서 방향 정보를 추출하여 3차원 공간에서 소리의 방향을 정량화하며 다채널 오디오 데이터에서 계산되며, 각 채널에서 측정된 신호의 상대적인 시간 지연과 진폭의 차이를 통해 방향을 추정한다. 강도 벡터는 스펙트로그램으로부터 계산되며 이는 수식9와 같이 계산된다[46].

$$\mathrm{I_{f,t}} \propto \mathrm{R}(\mathrm{W}_{\mathrm{f,t}}^{*}, h_{f,t}) = \left[I_{X,f,t}, \mathrm{I_{Y,f,t}}, \mathrm{I_{Z,f,t}}\right]^{\top} \qquad (9)$$

여기서 $f \in 1, \ldots, F$는 주파수 인덱스를 의미하며, $t \in 1, \ldots, T$는 시간 인덱스를 의미한다. $h_{f,t} = \left[X_{f,t}, Y_{f,t}, \mathrm{Z_{f,t}}\right]^{\top}$, $\mathrm{R}(\cdot)$은 복소수의 실수 부분을 나타내며, *는

켤레 복소수를 의미한다. 4채널의 멜-스펙트로그램에서 W채널을 기반으로 X, Y, Z채널에 대해 계산된다.

학습 데이터셋은 1,200개로 구성되어있으며 이는 딥러닝 모델을 학습 시키기에는 적은 양의 데이터일 수 있다. 학습 데이터의 수가 부족할 경우, 과적합 문제가 발생할 수 있으며 이는 모델의 성능 저하로 이어진다. 다양한 학습 시나리오를 사용하여 모델의 과적합을 방지하고 모델의 성능을 향상 시키기 위해 데이터 주파수 마스킹(frequency masking) 증강 기법을 적용하였다. 주파수 마스킹은 스펙트로그램의 일정 주파수 채널을 마스킹하는 방법으로, 본 연구에서 적용된 주파수 마스킹은 랜덤한 영역의 주파수 정보를 완전히 제거하는 방법을 적용하였다.

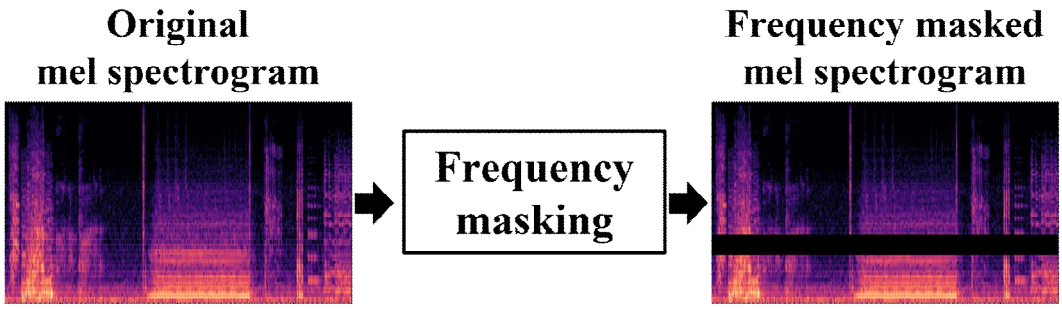


그림 4.1 주파수 마스킹 적용 예시

3. 학습 환경 설정

본 연구에서 제안한 모델과 비교군 모델 모두 NAdam(Nesterov-accelerated adaptive memement Adam)[47] 최적화기(optimizer)를 사용하였으며, 최적화기의 가중치는 scheduler는 step learning rate를 사용하였다. 모델의 학습률과 드롭아웃(dropout)은 각각 10^{-3}, 0.2로 설정하였으며 모든 모델 학습은 PyTorch 라이브러리를 활용하였다.

2절 비교군 모델 구조

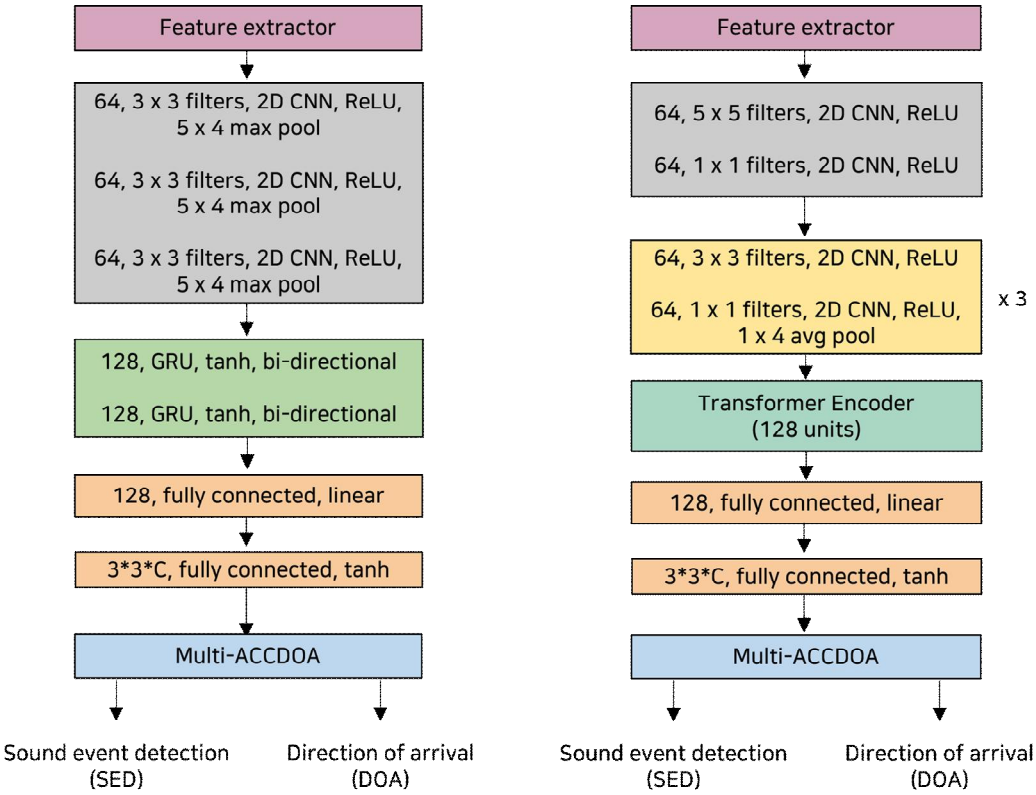
해당 절에서는 본 연구에서 제안한 모델의 성능을 평가하기 위해 사용한 비교군 모델에 대해서 설명한다. 비교군 모델은 CRNN 모델과 Residual block과 트랜스포머를 사용한 모델로 구성된다.

1. CRNN 모델

CRNN은 CNN과 RNN으로 구성된 모델로, 인공지능 기반 음향 이벤트 및 장면 인식 기술 경진 대회(Detection and Classification of Acoustic Scenes and Events; DCASE)의 2022년 음향 이벤트 위치 추정 및 탐지 챌린지의 베이스라인이 된 모델이다. 구체적인 모델의 구조는 그림 4.2 (a)와 같다. 층의 CNN과 2층의 RNN으로 구성되어있으며, RNN으로는 양방향 게이트 순환 유닛(Bidirectional Gated Recurrent Unit; Bi-GRU)가 사용되었다. 최종적인 출력 형식으로는 Multi-ACCDOA가 사용되었다.

2. Residual+Transformer 모델

Residual block과 트랜스포머 인코더로 구성된 모델은 제안하는 SELD U-net의 SELDnet과 동일한 구조를 갖는 모델이다. 구체적인 모델의 구조는 그림 4.2 (b)와 같다. 입력 특징은 CNN층을 거쳐 모양이 변화하게 되고, Residual block을 통과한다. Residual block은 총 3층으로 구성되어 있으며 구조는 그림 3.4와 같다. 이후, 트랜스포머 인코더를 거쳐 Multi-ACCDOA 출력 형식에 따라 예측 값을 출력한다.



(a) CRNN 구조의 모델 (b) Residual + Transformer encoder 구조의 모델

그림 4.2 비교군 모델의 구조

3절 실험 결과

해당 절에서는 3장에서 제안한 모델의 음성 향상 결과 및 음향 이벤트 위치 추정 및 탐지의 결과를 보인다.

3.1절에서는 음성 향상 결과로는 제안한 모델의 결과를 평가 지표를 활용하여 성능 분석을 수행한다.

3.2절에서는 음성 탐지 및 위치 결과의 성능 평가는 제안된 모델과 비교군 모델의 성능을 토대로 분석을 수행한다.

1. 음성 향상 결과

해당 절에서는 3장에서 제시한 U-net, Nested U-net, Residual U-net을 사용한 음성 향상 결과를 나타낸다. 그림 4.3은 U-net을 활용하여 스펙트로그램의 잡음 제거를 수행한 결과이며, 그림 4.4는 강도 벡터의 잡음을 제거한 결과이다. 시각적으로 모델의 추론 결과를 비교하였을 때, U-net, Nested U-net, Residual U-net의 복원 결과가 비슷한 수준인 것을 확인할 수 있다.

신호 대 잡음 비가 +30, +20, +10, -10인 경우 복원된 스펙트로그램을 확인한 결과, 잡음이 없는 상태의 스펙트로그램과 거의 유사하게 복원이 수행된 것을 확인할 수 있다. 그러나, 신호 대 잡음비가 -20, -30인 경우, 진폭의 값이 높은 부분의 특징은 탐지하지만 낮은 부분을 잘 탐지하지 못해 전반적으로 복원이 잘 수행되지 않은 것을 확인할 수 있다.

강도 벡터를 확인한 결과, 스펙트로그램의 복원 결과와 유사하게 +30, +20, +10에서의 복원 결과는 타겟과 유사하게 복원이 전반적으로 잘 수행되었으나, -10, -20, -30의 경우 복원이 잘 수행되지 않은 것을 확인할 수 있다.

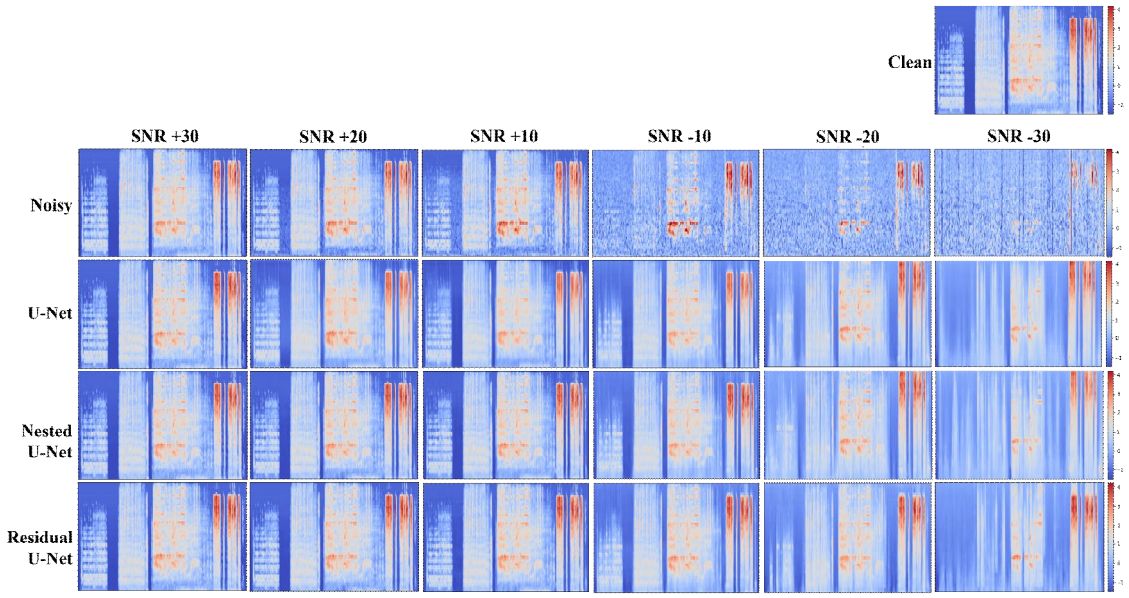


그림 4.3 음성 향상 실험 결과(스펙트로그램)

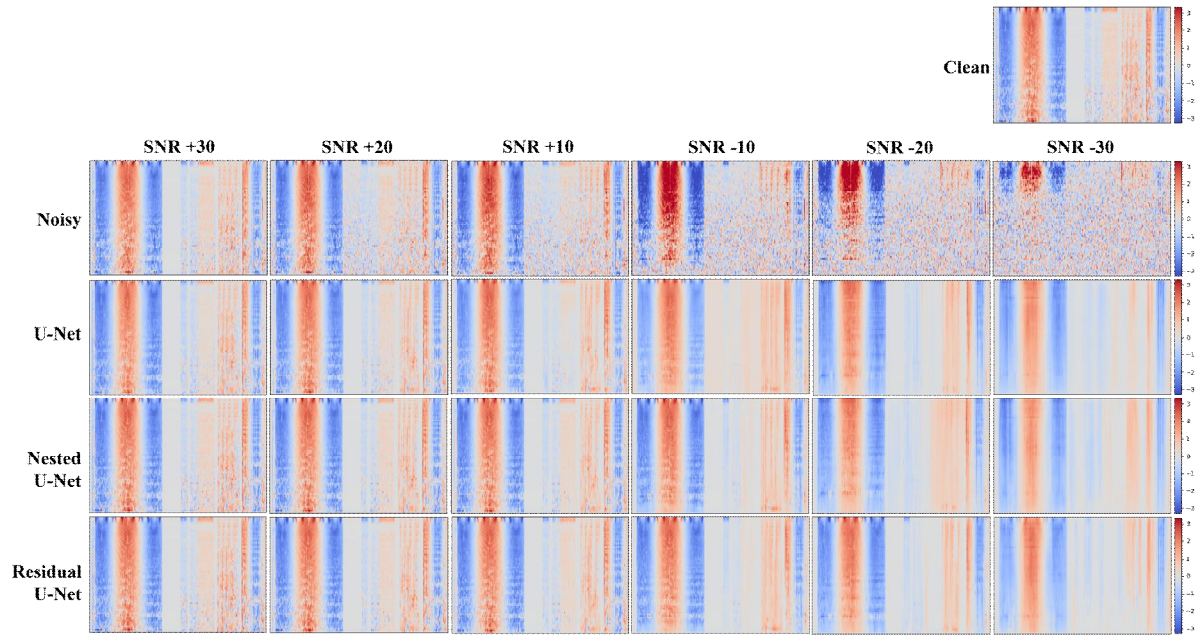


그림 4.4 음성 향상 실험 결과(강도 벡터)

이를 MSE, SSIM, PSNR을 사용하여 수치적으로 성능을 비교한 결과는 표 4.1, 표 4.2와 같다.

신호 대 잡음비가 모두 양수인 경우, 즉, 잡음이 거의 존재하지 않는 경우, 전반적으로 Residual U-net 모델이 스펙트로그램과 강도 벡터의 대부분의 평가 지표에서 가장 좋은 성능을 보인 것을 확인할 수 있다. 그러나, 신호 대 잡음비가 +20인 경우의 강도 벡터의 SSIM과 신호 대 잡음비가 +10인 경우의 스펙트로그램과 강도 벡터의 SSIM 지표에서는 Nested U-net이 더 좋은 성능을 보임을 확인할 수 있었다. U-net의 경우, 3가지 모델들 중, 모든 지표에서 가장 낮은 성능을 달성하였다.

신호 대 잡음비가 음수인 경우를 살펴보면, Residual U-net이 모든 지표에서 가장 좋은 성능을 보이는 것을 확인할 수 있었으며, Nested U-net과 U-net은 비슷한 수준의 성능을 보임을 확인할 수 있었다.

해당 실험을 통해, 잡음이 거의 존재하지 않은 환경에서 Residual U-net의 결과가 가장 우수한 성능을 달성함을 확인할 수 있었다. Nested U-net과 U-net은 Residual U-net과 비교할 경우 상대적으로 낮은 성능을 달성하였으나 잡음 데이터와의 결과를 비교하였을 때, 잡음 제거를 전반적으로 잘 수행하고 있음을 확인할 수 있다. 결론적으로 제안하는 SELD U-net의 전단부의 U-net은 잡음 제거를 수행할 수 있으며 이를 통해 SELD의 성능 향상을 기대할 수 있다.

표 4.1 평가 지표를 통한 음성 향상 실험 결과 1(SNR +30, +20, +10)

Enhancement Results							
		Spectrogram			Intensity vector		
		MSE	SSIM	PSNR	MSE	SSIM	PSNR
SNR +30	Noise	0.0	1.0	inf	0.0	1.0	inf
	U-net	0.041	0.639	62.066	0.043	0.861	61.901
	Nested U-net	0.001	0.979	77.563	0.001	**0.999**	77.170
	Residual U-net	**1e-07**	**0.999**	**114.02**	**6e-07**	**0.999**	**113.41**
SNR +20	Noise	0.047	0.782	61.885	0.040	0.543	62.501
	U-net	0.052	0.642	61.312	0.048	0.631	61.641
	Nested U-net	0.009	0.895	69.111	0.008	**0.722**	69.483
	Residual U-net	**0.008**	**0.907**	**69.609**	**0.007**	0.712	**70.032**
SNR +10	Noise	0.142	0.553	56.983	0.124	0.380	57.549
	U-net	0.066	0.596	60.100	0.065	0.492	60.146
	Nested U-net	0.023	**0.801**	64.867	0.022	**0.570**	65.112
	Residual U-net	**0.021**	0.798	**65.201**	**0.020**	0.563	**65.465**

표 4.2 평가 지표를 통한 음성 향상 실험 결과 2(SNR -10, -20, -30)

Enhancement Results							
		Spectrogram			Intensity vector		
		MSE	SSIM	PSNR	MSE	SSIM	PSNR
SNR -10	Noise	0.732	0.137	49.863	0.663	0.091	50.267
	U-net	0.178	0.438	55.924	0.171	0.192	56.069
	Nested U-net	0.176	0.433	55.951	0.168	0.160	56.109
	Residual U-net	**0.133**	**0.505**	**57.228**	**0.126**	**0.265**	**57.443**
SNR -20	Noise	1.155	0.051	47.879	1.090	0.028	48.123
	U-net	0.344	0.290	53.081	0.333	0.096	53.177
	Nested U-net	0.358	0.267	52.826	0.344	0.071	52.966
	Residual U-net	**0.283**	**0.351**	**53.847**	**0.277**	**0.166**	**53.944**
SNR -30	Noise	1.485	0.013	46.744	1.448	0.005	46.858
	U-net	0.660	0.150	50.267	0.639	0.054	50.357
	Nested U-net	0.650	0.126	50.277	0.640	0.044	50.319
	Residual U-net	**0.584**	**0.198**	**50.742**	**0.583**	**0.106**	**50.753**

2. 음성 탐지 및 위치 추정 결과

본 절에서 제안한 모델과 비교군 모델의 비교를 통한 음향 이벤트 위치 추정 및 탐지의 결과들을 보인다.

표 4.3, 4.4, 4.5는 신호 대 잡음비가 각각 +30, +20, +10인 경우의 성능을 비교한 결과이다. 잡음이 거의 존재하지 않는 환경에서 제안하는 모델들의 성능이 CRNN 모델보다 준수한 성능을 보이나, Residual block과 트랜스포머 인코더를 적용한 모델에는 미치지 못하는 성능을 보임을 확인하였다. 제안한 모델 중에서 신호 대 잡음 비가 +30 +20인 경우, Nested U-net을 사용한 모델이 가장 좋은 성능을 보였으며, 신호 대 잡음 비가 +10인 경우, Residual U-net을 사용한 모델이 가장 좋은 성능을 보였다. SELD 모델은 잡음이 거의 없는 환경에서는 별도의 음성 향상 작업이 없어도 높은 성능을 보인다. 때문에, 잡음 제거 과정을 추가한 제안 모델이 복잡성이 높으며, 데이터를 필요 이상으로 수정하여 오버피팅(overfitting)을 할 가능성을 높이게 된다. 이러한 이유를 통해 제안한 SELD U-net이 잡음 제거 과정이 없는 동일한 SELDnet을 사용한 모델보다 낮은 성능을 보이는 것으로 추정된다.

표 4.6, 4.7, 4.8은 신호 대 잡음비가 각각 -10, -20, -30인 경우의 성능을 비교한 결과이다. 비교적 잡음이 약한 환경(신호 대 잡음비 -10)인 경우, Residual U-net 모델과 Residual block과 트랜스포머 인코더를 적용한 모델이 가장 좋은 성능을 보임을 확인할 수 있었다. 그러나 신호대 잡음 비가 -20, -30인 경우, 제안한 모델들이 기존의 모델보다 좋은 성능을 달성하였다. 이를 통해, 제안한 모델이 잡음이 강한 환경에서 더 강건한 SELD 모델임을 보여주며, 잡음이 강한 환경에서의 강건성이라는 측면에서 유의미한 결과라는 것을 확인할 수 있다.

표 4.3 SNR이 +30인 경우의 음향 이벤트 위치 추정 및 탐지 결과

Model Results (SNR +30)					
	Error Rate	F1-score	Localization Error	Localization Recall	SELD score
CNN + Bi-GRU	0.71	0.30	17.66	0.37	0.54
Residual block + Transformer	**0.66**	**0.40**	**13.51**	**0.51**	**0.46**
SELD U-net (U-net)	0.71	0.32	17.28	0.34	0.51
SELD U-net (Nested U-net)	0.69	0.35	15.64	0.46	0.49
SELD U-net (Residual U-net)	0.70	0.34	16.69	0.46	0.50

표 4.4 SNR이 +20인 경우의 음향 이벤트 위치 추정 및 탐지 결과

Model Results (SNR +20)					
	Error Rate	F1-score	Localization Error	Localization Recall	SELD score
CNN + Bi-GRU	0.72	0.30	17.45	0.36	0.54
Residual block + Transformer	**0.69**	**0.38**	**14.13**	**0.48**	**0.48**
SELD U-net (U-net)	0.73	0.31	17.36	0.43	0.52
SELD U-net (Nested U-net)	0.71	0.32	17.80	0.44	0.51
SELD U-net (Residual U-net)	0.72	0.31	17.99	0.44	0.52

표 4.5 SNR이 +10인 경우의 음향 이벤트 위치 추정 및 탐지 결과

Model Results (SNR +10)					
	Error Rate	F1-score	Localization Error	Localization Recall	SELD score
CNN + Bi-GRU	0.75	0.27	17.18	0.35	0.56
Residual block + Transformer	**0.70**	**0.37**	**15.06**	**0.47**	**0.49**
SELD U-net (U-net)	0.73	0.29	18.35	0.39	0.54
SELD U-net (Nested U-net)	0.72	0.31	17.53	0.41	0.52
SELD U-net (Residual U-net)	0.71	0.34	15.88	0.44	0.50

표 4.6 SNR이 -10인 경우의 음향 이벤트 위치 추정 및 탐지 결과

Model Results (SNR -10)					
	Error Rate	F1-score	Localization Error	Localization Recall	SELD score
CNN + Bi-GRU	0.84	0.17	23.20	0.24	0.63
Residual block + Transformer	**0.78**	**0.27**	**18.14**	**0.35**	**0.57**
SELD U-net (U-net)	0.79	0.25	18.66	0.33	0.58
SELD U-net (Nested U-net)	0.80	0.24	20.11	0.35	0.58
SELD U-net (Residual U-net)	**0.78**	0.26	18.98	0.34	**0.57**

표 4.7 SNR이 -20인 경우의 음향 이벤트 위치 추정 및 탐지 결과

Model Results (SNR -20)					
	Error Rate	F1-score	Localization Error	Localization Recall	SELD score
CNN + Bi-GRU	**0.79**	0.11	27.10	0.17	0.69
Residual block + Transformer	0.87	0.16	22.81	0.22	0.65
SELD U-net (U-net)	0.86	**0.19**	**19.41**	0.24	**0.63**
SELD U-net (Nested U-net)	0.85	**0.19**	22.25	**0.26**	**0.63**
SELD U-net (Residual U-net)	0.85	0.18	21.54	0.24	0.64

표 4.8 SNR이 -30인 경우의 음향 이벤트 위치 추정 및 탐지 결과

Model Results (SNR -30)					
	Error Rate	F1-score	Localization Error	Localization Recall	SELD score
CNN + Bi-GRU	0.97	0.05	43.88	0.10	0.76
Residual block + Transformer	**0.94**	0.07	65.72	**0.13**	0.78
SELD U-net (U-net)	0.95	**0.09**	**26.86**	**0.13**	**0.72**
SELD U-net (Nested U-net)	**0.94**	**0.09**	36.42	0.12	0.73
SELD U-net (Residual U-net)	**0.94**	0.08	27.91	0.12	**0.72**

제 5 장 결론

1절 연구 의의

본 연구에서는 음성 향상 모델과 음향 이벤트 탐지 및 위치 추정 모델을 결합한 잡음 환경에서 강건한 모델을 제안하였다. 이러한 발상은 잡음 환경에서 추출된 입력 특징에 음상 향상 기법을 적용하여 입력 특징의 잡음을 제거하고, 이를 SELD의 입력 특징으로 사용함으로써 잡음 환경에 강건하게 작용할 것이라는 가정하에 이루어졌고, 본 연구에서 이 가정을 입증하고자 하였다. 그리고, 입증한 결과를 기반으로 잡음 제거를 수행하는 음성 향상 모델과 음향 이벤트 위치 추정 및 탐지 모델을 결합하여 잡음 환경에서 강건한 모델을 구성할 수 있음을 보였다. 음향 이벤트 탐지 및 위치 추정 모델은 음향 이벤트 탐지와 소리 원천 위치 추정을 동시에 진행하여 정보 통합과 최적화에 효과적이기 때문에 다양한 분야에서 사용되었다. 그러나 잡음 환경에서 수행되는 음향 이벤트 탐지 및 위치 추정 모델에 관한 연구는 부족하였으며 특히, 별도의 전처리 과정 없이 잡음에 강건한 모델은 존재하지 않았다.

제안하는 모델은 음성 향상을 수행하는 다양한 구조의 U-net과 음향 이벤트 탐지 및 위치 추정을 수행하는 SELDnet과 결합하여 잡음에 강건한 모델을 설계하고자 하였다. U-net 디코더의 일부 특징 맵을 SELDnet의 입력으로 사용하여, 잡음이 제거된 핵심적인 정보를 갖는 입력 특징을 사용하여 잡음 환경에 강건한 성능을 보이는 모델을 설계하였다.

제안된 모델의 성능이 잡음이 없는 환경과 잡음이 적은 환경에서는 잡음 제거를 수행하지 않는 동일한 구조의 모델에 비해 모든 지표에서 낮은 점수를 기록하였다. 그러나, 잡음이 심한 환경에서는 다른 모델과 비교하였을 때, 모든 지표에서 비교 모델보다 좋은 성능을 보이는 것을 확인하였다.

2절 향후 연구

잡음 환경에 강건한 음향 이벤트 탐지 및 위치 추정 모델 구축을 위해 음성 향상에 수행되는 U-net을 SELDnet에 결합한 모델을 제안했다. 잡음 환경의 오디오에서 추출한 입력 특징에서 잡음에 해당하는 성분을 제거하는 것을 목표로 하였다. 제안한 모델을 통해 음성 향상을 수행한 결과, 스펙트로그램의 복원 결과에 비해, 강도 벡터의 복원은 부족한 결과를 보였다. 또한, 잡음이 강한 환경에서 제안 모델이 기존 모델보다 좋은 성능을 보였으나, 잡음이 없는 환경에서의 성능과 비교해 볼 때, 추가적인 성능 향상이 필요해 보인다. 이를 위해, 강도 벡터의 잡음 제거에 적합한 모델 구조에 대한 연구와 제안한 SELDnet보다 우수한 성능을 갖는 모델 연구를 통해 제시한 모델의 한계점을 극복하는 과정이 필요해 보인다.

참고문헌

[1] R. Takeda and K. Komatani, “Sound source localization based on deep neural networks with directional activate function exploiting phase information,” in *Proc.* IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 405-409.

[2] N. Yalta, K. Nakadai, and T. Ogata, “Sound source localization using deep learning models,” J. Robot. Mechatronics, vol. 29, no. 1, pp. 37-48, 2017.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, “Audio surveillance: A systematic review,” ACM Comput. Surv., vol. 48, 2016, Art. no. 52.

[4] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, “Audio surveillance of roads: A system for detecting anomalous sounds,” IEEE Trans. Intell. Transp. Syst., vol. 17, no. 1, pp. 279-288, Jan. 2016.

[5] P. Swietojanski, A. Ghoshal, and S. Renals, “Convolutional neural networks for distant speech recognition,” IEEE Signal Process. Lett., vol. 21, no. 9, pp. 1120-1124, Sep. 2014.

[6] Wolfel and J. McDonough, ˝ Distant Speech Recognition. Hoboken, NJ, USA; Wiley, 2009.

[7] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial", IEEE Signal Processing, vol. 38, no. 5, pp. 67-83, 2021.

[8] M. Risoud, J-H. Hanson, F. Gauvrit, C. Renard, P-E. Lemesre, N-X. Bonne, and C. Vincent, "Sound source localization", European annals of otorhinolaryngology, vol. 135, no. 4, pp. 259-264, 2018.

[9] G. Parascandolo, and H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc.* International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6440-6444, 2016.

[10] T. Zhou, Z. He, R. Zhai, and X. Xu, "Sound Event Detection with Speech Interference Using Convolutional Recurrent Neural Networks," in *Proc.* International Conference on Big Data and Artificial Intelligence (BDAI), pp. 70-74, 2021.

[11] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, “Two-source acoustic event detection and localization: Online implementation in a smart-room,” in *Proc.* Euro. Signal Process. Conf., 2011, pp. 1317-1321.

[12] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 1, pp. 34-48, 2018.

[13] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc.* International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 885-889, 2021.

[14] S. Pascual, A. Bonafonte, J. Serra, "SEGAN: Speech enhancement generative adversarial network," arXiv preprint arXiv:1703.09452, 2017.

[15] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," Interspeech, vol. 2013, pp. 436-440, 2013.

[16] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, “Acoustic event detection in real-life recordings,” in European Signal Processing Conference (EUSIPCO), pp.1267-1271, 2010.

[17] E. C¸ akır, T. Heittola, H. Huttunen, and T. Virtanen, “Polyphonic sound event detection using multi-label deep neural networks,” IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1-7, 2015.

[18] H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), PP. 559-563, 2015.

[19] E. C¸ akır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291-1303, 2017.

[20] Y. Huang, J. Benesty, G. Elko, and R. Mersereati, “Real-time passive source localization: a practical linear-correction least-squares approach,” IEEE Transactions on Speech and Audio Processing, vol. 9, no. 8, pp. 943-956, 2001.

[21] M. S. Brandstein and H. F. Silverman, “A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays,” in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1997.

[22] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Transactions on Antennas and Propagation, vol. 34, no. 3, pp. 276-280, 1986.

[23] R. Roy and T. Kailath, “ESPRIT-estimation of signal parameters via rotational invariance techniques,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 37, no. 7, pp. 984-995, 1989.

[24] W. He, P. Motlicek, and J.-M. Odobez, “Deep neural networks for multiple speaker detection and localization,” in Proc. International Conference on Robotics and Automation (ICRA), pp. 74-79, 2018.

[25] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, “Two-source acoustic event detection and localization: Online implementation in a smart-room,” European Signal Processing Conference (EUSIPCO), pp. 1317-1321, 2011.

[26] T. Hirvonen, “Classification of spatial audio location and content using convolutional neural networks,” Audio Engineering Society Convention 138, 2015.

[27] K. Lopatka, J. Kotus, and A. Czyzewsk, “Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations,” Multimedia Tools and Applications Journal, vol. 75, no. 17, pp. 10407-10439, 2016.

[28] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 915-919, 2021.

[29] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 316-320, 2022.

[30] Dang, Xin and T. Nakai, "Noise reduction using modified phase spectra and Wiener Filter," in Proc. International Workshop on Machine Learning for Signal Processing, pp.1-5, 2011.

[31] J.-S. Hu and C.-H. Yang, "Speech enhancement using transfer function ratio beamformer and matched filter array," in Proc. International Conference on Information and Automation, pp. 1161-1166, June 2009.

[32] Y. Wang, J. An, V. Sethu, and E. Ambikairajah, "Perceptually motivated pre-filter for speech enhancement using Kalman filtering", in Proc. International Conference on Information, Communications & Signal Processing, pp. 1-4, Dec. 2007.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Proc. Medical Image Computing and Computer-Assisted Intervention(MICCAI), pp. 234-241, 2015.

[34] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," arXiv preprint arXiv:1806.03185, 2018.

[35] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749-753, 2018.

[36] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 333-337, 2019.

[37] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning-based objective assessment for voice conversion," Interspeech 2019, Sep 2019.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[39] D. A. Sadlier, and N. E. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 10, pp. 1225-1233, 2005.

[40] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in Proc. Advanced Video and Signal Based Surveillance, pp. 21-26, 2007.

[41] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," Microphone arrays: signal processing techniques and applications. pp. 157-180, 2001.

[42] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in Proc. Workshop on Applications of Signal Processing to Audio and Acoustics

(WASPAA), pp. 136-140, 2017.

[43] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbtley, "Polyphonic sound event detection and localization using a two-stage strategy," in Proc. of DCASE Workshop, 2019.

[44] Z. Zhou, M.M.R Siddiquee, N. Tajbakhsh, and J. Liang, J "A nested U-Net architecture for medical image segmentation," arXiv preprint arXiv:1807.10165, 2018.

[45] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.30, pp. 829-852, Dec. 2021.

[46] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by DNN-based denoising and source separation," in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 651--655, May 2020.

[47] T. Dozat, "Incorporating nesterov momentum into adam," in Proc. International Conference on Learning Representations (ICLR), pp. 1--4, May 2016.