JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 10, pp. 33-42, October 2025
https://doi.org/10.9708/jksci.2025.30.10.033

Robust Audio Spectrogram Transformer for Sound Source Localization in Noisy Environments

Won Jun Lee*, Woo Jin Jung**, Hyun-Jong Cha***, Ah Reum Kang****

*M.A. Student, Dept. of Smart ICT Convergence, Pai Chai University, Daejeon, Korea
**Student, Dept. of Cyber Security, Pai Chai University, Daejeon, Korea
***Professor, Dept. of Computer Engineering, Pai Chai University, Daejeon, Korea
****Professor, Dept. of Cyber Security, Pai Chai University, Daejeon, Korea

[Abstract]

Conventional sound source localization methods suffer from significant accuracy degradation in low SNR (Signal-to-Noise Ratio) environments. In this paper, we propose a sound source localization model based on an audio spectrogram transformer, which takes GCC (Generalized Cross Correlation) features extracted from multichannel audio signals as input. The proposed model was evaluated under various indoor environments and SNR conditions, and its performance was compared with conventional GCC-PHAT (Generalized Cross Correlation – Phase Transform) and MUSIC (MUltiple SIgnal Classification) algorithms. Experimental results show that the proposed model achieves superior performance, with a mean angular error of 10.0163°, a mean distance error of 0.1626, and a RMSE (Root Mean Square Error) of 0.89 in a 5 m × 5 m × 5 m environment, even at 0 dB SNR. Additionally, the model demonstrates robust performance under changes in room size and noise conditions. This study demonstrates that transformer-based models can be effectively applied to achieve reliable sound source localization in noisy environments.

▸ **Key words**: Sound Source Localization, Direction of Arrival, Signal-to-Noise Ratio, Audio Sectrogram Transformer, Image Source Method

[요 약]

음원 위치 추정은 로봇 내비게이션, 스마트 홈 제어, 음향 모니터링 등 다양한 응용 분야에서 중요한 역할을 한다. 그러나 기존의 음원 위치 추정 기법들은 SNR(Signal-to-Noise Ratio)이 낮은 환경에서 정확도가 크게 저하되는 한계가 있다. 본 논문에서는 멀티채널 오디오 신호로부터 GCC(Generalized Cross Correlation) 특징을 추출하여 입력하는 오디오 스펙트로그램 트랜스포머 기반 음원 위치 추정 모델을 제안한다. 제안 모델은 다양한 실내 공간과 SNR 조건에서 기존의 GCC-PHAT(Generalized Cross Correlation – Phase Transform) 및 MUSIC(MUltiple SIgnal Classification) 알고리즘과 비교 평가되었다. 실험 결과, 제안 모델은 SNR 0 dB 환경에서도 평균 각도 오차 10.0163°, 평균 거리 오차 0.1626, RMSE(Root Mean Square Error) 0.89(5 m × 5 m × 5 m 기준)로 기존 기법 대비 우수한 성능을 보였다. 또한, 공간 크기 변화와 잡음 환경 변화에도 강인한 일반화 성능을 나타냈다. 본 연구는 트랜스포머 기반 딥러닝 모델이 실내 환경에서 신뢰성 높은 음원 위치 추정에 효과적으로 활용될 수 있음을 실증하였다.

▸ **주제어**: 음원 위치 추정, 도래각 추정, 신호대 잡음 비, 오디오 스펙트로그램 트랜스포머, 이미지 소스 기법

• First Author: Won Jun Lee, Corresponding Author: Ah Reum Kang
*Won Jun Lee (ayjmlo6040@gmail.com), Dept. of Smart ICT Convergence, Pai Chai University
**Woo Jin Jung (2284057@pcu.ac.kr), Dept. of Cyber Security, Pai Chai University
***Hyun-Jong Cha (hjcha@pcu.ac.kr), Dept. of Computer Engineering, Pai Chai University
****Ah Reum Kang (armk@pcu.ac.kr), Dept. of Cyber Security, Pai Chai University
• Received: 2025. 08. 22, Revised: 2025. 09. 14, Accepted: 2025. 10. 14.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

음원 위치 추정은 스마트 기기, 로봇, 회의 시스템, 증강현실 등 다양한 분야에서 핵심적인 역할을 한다. 특히 다중 마이크 배열을 이용해 음원의 방향이나 위치를 파악하는 기술로, 음성 신호의 공간적 특성을 분석하여 사용자와 시스템 간 상호작용의 정확도와 효율성을 높이는 데 필수적이다. 전통적인 음원 위치 추정 알고리즘들은 주로 시간 지연 추정, 빔포밍, 그리고 음향 신호의 공간적 특성을 수학적으로 모델링하는 방법에 기반한다. 그러나 이러한 방법들은 잡음이나 반향이 많은 실제 환경에서 성능 저하가 크며, 특히 SNR(Signal-to-Noise Ratio)가 낮은 상황에서는 정확한 위치 추정이 어려운 한계가 있다[1].

최근에는 딥러닝 기반 기법들이 음원 위치 추정 문제에 활발히 적용되고 있다. 특히, 신경망 모델은 잡음과 반향에 강인한 특징 표현 능력을 통해 기존 알고리즘 대비 더 우수한 성능을 보이고 있으며, 복잡한 음향 환경에서도 효과적인 위치 추정이 가능하다[2].

본 연구는 이러한 배경에서, 잡음과 반향이 심한 환경에서도 강건하게 동작할 수 있는 트랜스포머 음원 위치 추정 모델을 제안한다. 제안하는 모델은 다중 마이크 배열에서 수집된 스펙트로그램의 특징을 효과적으로 학습하여, 기존 수학적 모델링 방법이 갖는 한계를 극복하고 높은 정확도와 안정성을 달성한다. 다양한 잡음 조건과 실제 환경에서 실험을 통해 제안 모델의 우수성을 검증하였으며, 이를 통해 실용적이고 신뢰성 높은 음원 위치 추정 기술 발전에 기여하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 음원 위치 추정을 살펴보고 한계점을 분석한다. 3장에서는 본 연구에서 제안하는 오디오 스펙트로그램 기반 트랜스포머 음원 위치 추정 모델의 구조와 동작 원리를 설명한다. 4장에서는 다양한 잡음 조건 및 실제 환경에서의 실험을 통해 제안 기법의 성능을 기존 방법들과 비교 평가하며, 5장에서는 본 연구의 결과를 요약하고 향후 연구 방향을 제시한다.

II. Related Work

1. Audio Spectrogram Transformer

오디오 신호 분류 및 인식 분야에서는 오랫동안 CNN(Convolutional Neural Network)을 기반으로 한 모델이 주류를 이루어 왔다. CNN은 시간-주파수 스펙트로그램 상의 지역적 특징을 효과적으로 추출하지만, 장기적인 시간적 의존성과 고차원 특징 간 전역적 상호작용을 모델링하는 데 한계가 있다. 이를 극복하기 위해 CNN과 어텐션 메커니즘을 결합한 하이브리드 모델들이 제안되고 있다.

Gong 등은 이미지 분류 분야에서 매우 높은 성능을 보인 Vision Transformer의 구조를 오디오 데이터 처리에 맞게 변형한 AST(Audio Spectrogram Transformer)를 제안하였다[3][4]. AST는 시간-주파수 스펙트로그램을 일정 크기의 패치로 나누어 임베딩하고, 시간 및 주파수 차원에 대한 포지셔널 인코딩을 적용한다. Self-attention을 통해 전역적인 특징 간 상호작용을 학습함으로써 장기적 시간 의존성과 고차원 특징 간 복잡한 관계를 효과적으로 포착한다.

또한 Transformer 기반 구조의 특성상 입력 시퀀스 길이에 제한이 없어, AST는 가변적인 길이의 오디오 신호를 자연스럽게 처리할 수 있다. 이러한 이유로 AST는 기존 CNN 기반 모델보다 뛰어난 성능을 보이며, 음원 위치 추정과 같은 분야에도 적용할 수 있다.

2. Sound Source Localization Methods

2.1 Time Delay-Based Sound Source Localization

마이크 어레이는 기하학적 구조에 따라 동일한 음원이 각 마이크에 서로 다른 시간에 도달하게 되며, 이에 따라 수신 신호 간에 시간 지연이 발생한다. 시간 지연 기반의 음원 위치 추정 방식은 TDOA(Time Difference of Arrival)를 계산하고, 이를 기반으로 음원의 위치를 추정하는 방식이다.

TDOA는 두 마이크에서 수신된 신호 간의 상호 상관(cross-correlation)을 통해 추정된다. 상호 상관은 두 신호 간의 유사성을 평가하기 위해 사용되며, 두 신호 간의 유사도를 측정하는 방법이다. 특히, 상호 상관 함수의 피크는 두 신호 간의 시간 지연을 나타내므로, 이를 통해 TDOA를 계산할 수 있다.

상호 상관 기반 기법 중 TDOA 에서 가장 널리 사용되는 방법은 GCC-PHAT(Generalized Cross Correlation)이다[5]. GCC-PHAT은 신호의 위상 정보만을 이용하여 상관 함수를 계산하게 되며, 신호의 진폭에 의한 왜곡을 억제할 수 있다. 이 방식은 잡음이나 신호 세기의 변동에 덜 민감하여, 다양한 환경에서 안정적인 시간 지연 추정 성능을 제공한다.

마이크 어레이에 포함된 두 개의 마이크로폰에서 각각 수신된 시간 영역 신호를 $x_1(n)$와 $x_2(n)$할 때, GCC-PHAT은 다음과 같이 구해진다[6].

$$R_{x_1x_2}(n) = \frac{1}{2/\pi}\int_{-\infty}^{\infty} W(w)X_1(w)X_2^*(w)e^{iwt}$$

$X_1(w)$와$X_2(w)$는 각각의 신호를 푸리에 변환한 결과이고, $X_2^*(w)$는 $X_2(w)$의 켤레 복소수이다. $W(w)$는 $X_1(w)$와 $X_2^*(w)$ 가중치 함수 PHAT으로, 신호의 위상 정보만을 강조하여 잡음과 반향 환경에서 시간 지연 추정을 안정화한다. 수식적으로는 다음과 같이 표현된다.

$$W(w) = \frac{1}{\left|X_1(w)X_2^*(w)\right|}$$

구해진 $R_{x_1}x_2(n)$의 최댓값을 구함으로써 TDOA값을 추정한다. 수식은 다음과 같이 표현된다.

$$\tau = \arg\max R_{x_1x_2}(n)$$

추정된 TDOA 값을 기반으로 실제 음원의 위치는 최소자승법(Least Squares)을 통해 계산된다. 각 마이크로폰 쌍에서 이론적으로 예상되는 시간 지연과 실제로 측정된 TDOA 값 간의 차이를 최소화하는 방향으로 음원 위치를 추정한다. 이 과정에서는 모든 마이크로폰 쌍에서 발생하는 시간 지연 정보를 종합하여, 음원의 2차원 또는 3차원 위치를 동시에 계산한다. 초기 추정값으로는 마이크로폰 어레이의 평균 위치를 사용하며, 반복적인 최적화 과정을 통해 전체 측정 오차를 최소화한 최종 음원 위치가 결정된다. 하지만 TDOA 기반 방식은 반향이나 잡음 환경에서 상관 피크가 왜곡되어 정확한 지연 추정이 어려우며, 마이크 배열의 구조적 오차나 동기화 문제에도 민감하다는 한계가 있다.

2.2 MUSIC(Multiple Signal Classification)

MUSIC 알고리즘은 고윳값 분해를 기반으로 하는 고해상도 주파수 추정 및 음원 위치 추정 기법이다[7]. 이 알고리즘은 1977년 Schmidt에 의해 처음 개발되었으며, 이후 다양한 신호 처리 분야에서 널리 활용되고 있다. 마이크 어레이로 수신된 신호의 공분산 행렬을 고윳값 분해하여 신호 서브스페이스와 잡음 서브스페이스로 분리하는 것이 핵심 원리이다.

MUSIC 알고리즘은 배열 안테나에서 수신한 신호가 여러 개의 송신 신호와 잡음의 선형 조합으로 이루어진다고 가정한다. 이때 핵심 가정은 신호원의 수가 측정 벡터의 요소 수보다 적어야 한다는 것이다. 이러한 가정하에서 수신 신호의 공분산 행렬을 계산하고, 이를 고윳값 분해하여 신호와 잡음 성분을 분리한다.

신호 벡터의 공분산 행렬은 신호 성분과 잡음 성분으로 구성된다. 이 행렬은 일반적으로 표본 상관 행렬을 통해 추정되며, 이를 통해 신호의 주파수 내용이나 방향을 추정할 수 있다. 공분산 행렬은 에르미트 행렬이므로, 모든 고유벡터는 서로 직교하는 특성을 가진다.

공분산 행렬의 고윳값을 내림차순으로 정렬할 때, 가장 큰 고윳값에 해당하는 고유벡터들은 신호 서브스페이스를 형성한다. 나머지 고유벡터들은 잡음 서브스페이스를 형성하며, 이는 신호 서브스페이스와 직교한다. 이러한 직교성은 MUSIC 알고리즘의 핵심 원리로, 신호 방향을 추정하는 데 활용된다.

신호 서브스페이스에 속하는 임의의 신호 벡터는 잡음 서브스페이스와 직교해야 한다. MUSIC은 이러한 직교성의 정도를 측정하기 위해 제곱 노름을 정의한다. 만약 신호 벡터가 신호 서브스페이스에 속한다면, 잡음 서브스페이스와의 직교 조건에 의해 이 제곱 노름은 0이 된다.

MUSIC의 주파수 추정 함수는 이 제곱 노름의 역수로 정의된다. 이 함수는 신호 주파수에서 뚜렷한 피크를 생성하며, 이를 통해 신호의 방향을 추정할 수 있다. 각 후보 방향에 대해 스티어링 벡터를 정의하고, 이 벡터가 잡음 서브스페이스에 수직일수록 해당 방향에 음원이 존재할 가능성이 높다고 판단한다.

추정 함수의 가장 큰 피크 위치는 신호 성분에 대한 주파수 추정치를 제공한다. 이 제곱 노름 표현식의 역수를 취하면 신호 주파수에서 뚜렷한 피크가 생성되며, 이를 통해 음원의 위치를 추정할 수 있다. MUSIC 알고리즘은 신호 서브스페이스와 잡음 서브스페이스 간의 직교성을 이용하여 높은 분해능의 주파수 추정을 제공한다. 그러나 MUSIC 알고리즘은 계산 복잡도가 높아 실시간 처리에 부적합하며, 환경 변화에 따라 성능이 불안정하다는 한계가 있다.

2.3 Deep Learning-Based Sound Source Localization

전통적인 수학적 모델 기반 음원 위치 추정 기법은 반향, 잡음, 배열 불완전성 등 실제 환경의 변수에 민감하여 성능이 저하되며, 고해상도 공간 스펙트럼 방식은 연산 복잡도가 높아 실시간 처리에 제약이 존재한다. 딥러닝 기반 음원 위치 추정은 이러한 한계를 극복하기 위해 원시 오디오나 스펙트로그램을 입력으로 사용하여 특징 추출과 방향 추정을 end-to-end로 학습하는 방식으로, 잡음과 반향 환경에서 강인하면서도 효율적인 추론이 가능하다.

DOANet에서는 드론에 탑재된 8채널 마이크 배열로부터 입력된 원시 오디오에 1D dilated convolution을 적용하여 방위각과 고도각을 회귀 방식으로 예측한다[8]. 이 모델은 dilated convolution을 이용해 긴 시계열 문맥 정보

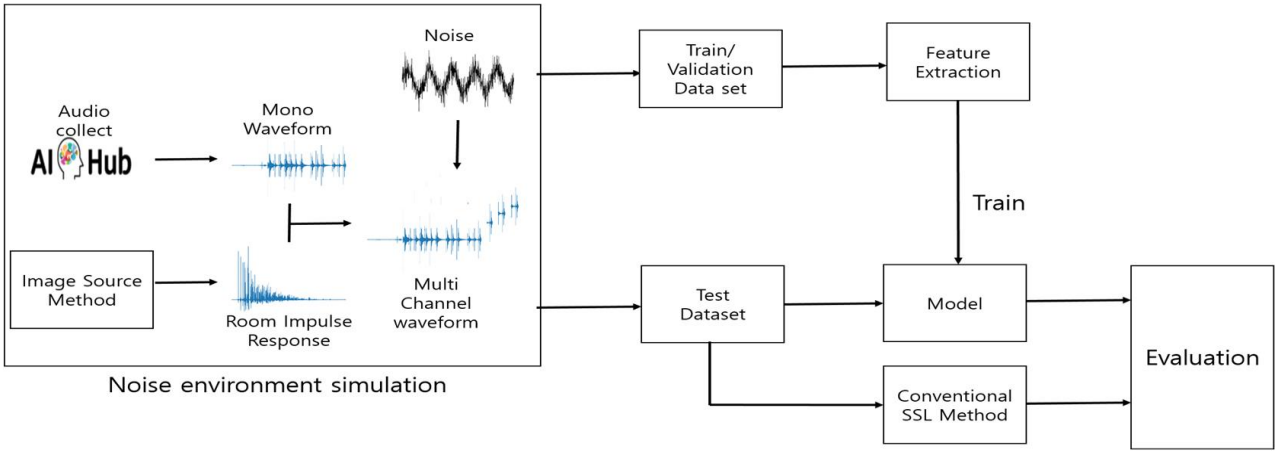

Fig. 1. Proposed system overview

를 포착하며, 별도의 특징 설계 없이도 신호의 공간 정보를 직접 학습한다. 이를 통해 전통적인 GCC-PHAT 및 MVDR 기법 대비 향상된 성능을 보였다. 특히 드론의 심각한 ego-noise 환경에서도 안정적인 음원 위치 추정이 가능함을 입증했다.

Yalta 등은 Deep Residual Networks를 기반으로 하는 구조를 통해 다중 채널 오디오에서 방향 추정을 수행했다[9]. 이들의 연구에서는 반향 환경에서 마이크 어레이를 사용한 음원 위치 추정에 deep residual networks를 적용했다. 아울러 STFT(Short Time Fourier Transform) 기반 전력 정보 처리와 채널별 화이트 노이즈를 활용해 강인한 성능을 보였으며, 입력 각도와 RMS 전력 기반 레이블링으로 도전적인 환경에서도 모델의 정확도를 높이는 데 성공했다.

실제 공간 음향 데이터셋(STARSS23)를 수집 및 활용한 연구에서는 log-mel 스펙트로그램과 intensity vector를 입력으로 CRNN을 활용하여 Sound Event Detection과 DOA 회귀를 동시에 수행한다[10][11]. 이 연구들은 다중 클래스별 Cartesian DOA 예측과 이벤트 활성도 추정이 가능한 Multi-ACCDOA 표현을 사용하며, 실제 환경에서 수집된 다중 채널 오디오 데이터를 통해 실시간 음향 환경에서 강건한 성능을 보인다.

Park 등은 Audio Spectrogram Transformer(AST) 기반의 Many-to-Many Audio Spectrogram Transformer (M2M-AST) 모델을 제안하였다[12]. 이 모델은 다중 채널 오디오 입력을 처리할 수 있도록 설계되었으며, 입력 스펙트로그램을 16×16 패치로 나누어 patch embedding과 다중 classification token을 활용한다. 순수 Transformer encoder 구조를 통해 SED 및 DOA를 예측하며, 다중 classification token 시퀀스를 통해 다양한 출력 해상도를 지원한다. 이 모델은 CRNN 대비 약 15~20% F-score 향상과 15° 이하 DOA 오차를 달성하였다.

3. Image Source Method

ISM(Image Source Method)는 실내 음향 시뮬레이션에서 반향 및 음원의 반사 경로를 모델링하기 위해 널리 사용되는 기법으로, 음원 위치 추정 연구에서 실내 환경의 음향 특성을 시뮬레이션하는 데 중요한 역할을 한다[13]. ISM은 음파가 벽, 천장, 바닥 등 반사면에서 반사되는 과정을 기하학적으로 모델링하여, 음원의 직접 경로와 다양한 반사 경로를 계산한다. 이 방법은 특히 음원 위치 추정 알고리즘의 성능을 평가하기 위한 합성 데이터 생성이나, 실제 실내 환경에서의 음향 특성 예측에 효과적으로 활용된다.

ISM을 기반으로 한 다양한 오픈소스 라이브러리도 개발되어 연구 및 응용에 널리 사용되고 있다. 대표적으로 Pyroomacoustics는 Python 기반의 라이브러리로, 이미지 소스 방법을 포함한 다양한 음향 시뮬레이션 기능을 제공하여 연구자들이 손쉽게 RIR(Room Impulse Response)를 생성하고 실내 음향 환경을 모델링할 수 있도록 지원한다[14]. 그러나 복잡한 구조나 대규모 시뮬레이션 환경에서는 여전히 상당한 계산 시간이 소요될 수 있다. 이를 해결하려는 방안으로는 GPU(Graphics Processing Unit)를 활용한 RIR 생성 기법인 GPU RIR과 같은 연구가 활발히 진행되고 있다[15]. GPU RIR은 병렬처리에 특화된 GPU의 성능을 활용하여 ISM 기반 RIR 계산 속도를 획기적으로 향상하며, 대규모 합성 데이터셋 구축에 필요한 시간을 효과적으로 단축한다.

III. The Proposed Scheme

잡음 환경에서도 강건한 음원 위치 추정을 가능하게 하고자 본 연구에서는 오디오 스펙트로그램 트랜스포머 기반의 새로운 추정 모델을 제안한다. 본 절에서는 제안하는 시스템의 전체적인 구조를 설명하며, 주요 구성 요소는 Fig. 1과 같이 다양한 잡음 환경을 반영한 합성 데이터셋 구축, 스펙트로그램 기반의 특징 추출 과정, 그리고 AST를 기반으로 한 음원 위치 추정 모델이다. 각 구성 요소에 대한 세부적인 내용은 다음 절에서 차례로 설명한다.

1. Dataset

음원 위치 데이터셋을 구성하는 방법에는 두 가지 방법이 존재한다. 첫 번째는 실제 환경에서 마이크 어레이를 이용해 다양한 위치에 음원을 배치하고, 이를 녹음하는 방식이다. 이 방법은 실제 환경에 대해서 데이터를 수집함으로 현실성을 보장하지만, 음원 위치의 정확한 정보를 얻기 위해서는, 고가의 모션 캡처 시스템이나 정밀한 위치 추적 장비가 요구된다.

두 번째 실내 음원 시뮬레이션 기반 합성 데이터 생성 방식은 다양한 음향 조건과 음원 위치를 정밀하게 제어할 수 있으며, 대규모 데이터셋을 비교적 낮은 비용과 시간으로 생성할 수 있다. 본 연구에서는 이러한 장점이 있는 시뮬레이션 기반 합성 데이터셋 접근법을 채택하여, 다양한 반향 조건과 잡음 환경을 포함한 음원 위치 데이터셋을 구성하였다.

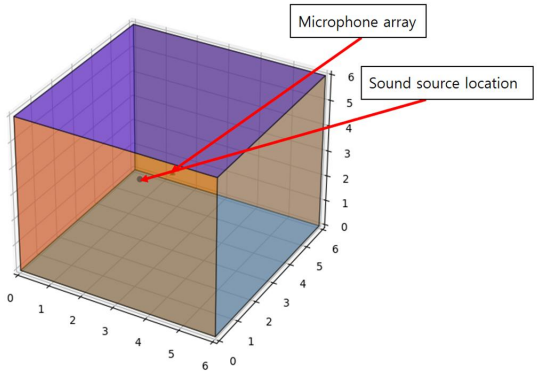

Fig. 2. Pyroomacoustics Simulate

실내 음원 시뮬레이션 환경은 Pyroomacoustics 라이브러리를 활용하여, Fig. 2와 같이 6 m x 6 m x 5m 크기의 직육면체 실내 공간 내에 4개의 마이크가 있는 마이크 어레이를 중심에 배치하게 구성하였다. 마이크 어레이의 구조는 마이크 간의 거리를 약 4.57cm로 설정하여 정사각형 형태로 구성하여 모든 마이크가 같은 평면상에 존재하도록 배치하였다. 이러한 구조를 통해 모든 방위각에 대한 음원 시뮬레이션이 가능하다.

음원은 각 샘플마다 실내 공간 내 임의의 위치에 무작위로 배치되며, 마이크 어레이 중심을 기준으로 방위각을 계산하였다. 각 음원 위치에 대해 잔향 시간은 0.2초에서 0.6초 사이의 무작위로 설정되었으며, SNR는 10 dB에서 40 dB 사이의 값을 갖도록 하여 다양한 잡음 음향 환경을 시뮬레이션하였다.

음원 신호는 AI허브 생활환경소음 데이터셋으로부터 무작위로 선택된 1초 길이의 오디오 클립으로 구성되며, 신호의 에너지 기준을 통해 무음 또는 에너지가 지나치게 낮은 샘플은 제외되었다[16]. 생성된 데이터의 형태는 1초 길이의 4채널 음원이다.

2. Feature Extraction

본 논문에서 사용하는 AST의 경우 입력으로 시계열 형태의 스펙트로그램 기반 특징을 요구한다. 일반적으로 오디오 신호로부터 Spectrogram을 생성할 때는 음원의 에너지 분포를 주파수 영역에서 분석할 수 있는 Mel Spectrogram이 널리 사용된다. Mel 스케일은 사람의 청각 인지 특성을 반영하여 저주파 영역에 더 높은 해상도를 부여하므로, 음성 인식이나 감정 분류 등에서 효과적인 특징으로 활용된다.

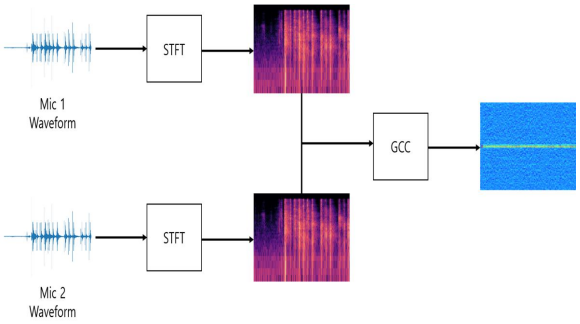

Fig. 3. GCC Feature Extraction Process

그러나 본 논문에서는 단순한 에너지 분포보다는 음원 간 시간 지연 정보가 더 중요하게 작용하는 음원 방향 추정 문제를 다루기 때문에, Mel Spectrogram은 사용하지 않고, 대신 마이크 쌍 간의 시간 차 정보를 담고 있는 GCC 기반의 특징만을 입력으로 사용하였다. GCC의 계산 과정은 Fig. 3과 같다. 입력 데이터는 1초 길이의 4채널 오디오 신호로 구성되며, 샘플링 주파수는 16kHz로 고정된다. 채널별로 25ms 길이의 Hann Window를 적용하고

10ms 간격으로 프레임을 이동시키며 STFT를 수행하였다. 이때 nfft는 512로 설정되었으며, 이는 STFT를 통해 주파수 영역으로 변환 시 사용되는 샘플 수로, 주파수 해상도와 계산 효율성을 조절하는 핵심 매개변수이다.

이와 같이 얻어진 복소수 형태의 스펙트럼을 바탕으로, 모든 마이크 쌍에 대해 크로스 스펙트럼을 계산하고 PHAT 가중치를 적용하여 GCC를 산출한다. 이후 역푸리에 변환을 통해 시간 지연 도메인의 GCC 함수를 도출하며, 각 마이크 쌍별로 128개의 시간 지연 정보를 포함하는 입력 특징으로 정리된다. 최종적으로 모델에 입력되는 특징은 6, 100, 128 차원의 텐서이다.

또한, 모델의 일반화 성능 향상을 위해 학습 과정에서 SpecAugment 기반의 데이터 증강 기법을 적용하였다[17]. 구체적으로 주파수 마스킹과 시간 마스킹을 무작위로 적용하여 특정 시간 또는 주파수 구간의 정보를 일부러 제거함으로써 모델이 특정 영역에 과도하게 의존하지 않도록 하였으며, 이는 잡음과 간섭이 많은 실제 환경에서의 추정 성능 향상에 기여한다.

3. Model

AST는 단일 채널 음성 신호의 스펙트로그램을 입력으로 처리하도록 설계된 모델로, 음성 인식 및 음성 분류 등 단일 채널 오디오 데이터의 특징을 효과적으로 학습하는 데 널리 사용됐다. 그러나 음원 위치 추정 문제에서는 다중 채널 음성 신호 간의 시간차와 공간 정보를 활용하는 것이 필수적이므로, 다중 채널 입력을 처리하기 위해 기존 AST의 입력 구조를 수정할 필요가 있다.

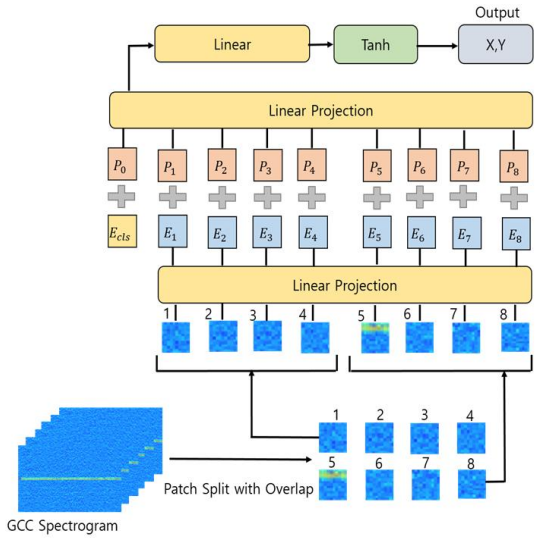

Fig. 4. Modified M2M-AST Model for Single-Source Direction-of-Arrival Estimation

Many-to-Many Audio Spectrogram Transformer (M2M-AST)는 이러한 요구에 맞춰 기존 AST 구조를 확장한 모델로, 여러 개의 분류 토큰을 사용하여 프레임별로 다중 채널 오디오 입력에 대한 방향 추정 및 이벤트 감지를 동시에 수행할 수 있다. 즉, 시간 프레임 단위로 연속적인 예측이 가능하여 다중 방위각 이벤트의 동시 추정을 지원한다.

하지만 M2M-AST는 다중 채널 및 다중 프레임 입력에 최적화된 모델로, 단일 음원 단일 방위각 추정 문제에서는 모델의 복잡성과 계산 비용이 불필요하게 증가할 수 있다. 이에 본 연구에서는 M2M-AST 구조를 Fig. 4와 같이 수정하였다.

패치 임베딩 과정에서는 4채널 음원 신호에서 추출한 GCC 특징을 입력으로 받기 위해 입력 채널 수를 6채널로 설정하였다. M2M-AST가 여러 개의 분류 토큰을 사용하여 프레임별 출력을 생성하는 것과 달리, 본 모델은 단일 방위각 추정을 위해 하나의 분류 토큰만을 사용하였다. 출력층의 수는 2개로 설정하여 방위각을 2차원 좌표 형태로 표현하였으며, Tanh 활성화 함수를 적용하여 출력값 범위를 제한함으로써 음원 위치 추정에서 더욱 정확한 단위 벡터 추정을 할 수 있다.

IV. Experiments and Result

1. Model Training Details

Table 1. Experimental Environment

Item	Value
OS	Ubuntu 22.04.4 LTS
CPU	13th Gen Intel(R) Core(TM) i7-13700KF
GPU	NVIDIA GeForce RTX 4090
Python Version	Python 3.10.9
library	Pytorch, librosa, scipy, Pyroomacustics

본 연구에서 제안한 모델의 학습을 위해 Pyroomacoustics를 사용하여 훈련용 6만 개, 검증용 2만 개의 합성 음원 데이터셋을 제작하였다. 대규모 이미지 데이터셋에서 사전 학습된 deit tiny 224 모델을 기반으로 전이 학습을 수행하였다[18]. 손실 함수로는 MSE(Mean Squared Error)를, 최적화 알고리즘으로는 Adam을 사용하였다.

모델의 패치 추출 과정에서는 시간 및 주파수 축에 대해 스트라이드 10을 적용하여 패치를 중첩되게 추출함으로

써, 공간적·시간적 정보를 효과적으로 반영하였다. 학습률은 0.0001, weight decay는 1e-7로 설정하여 과적합을 방지하였으며, 학습률은 OneCycleLR을 스케줄러를 사용하여 동적으로 조절하였다. 배치 크기는 256으로, 총 50epochs 동안 학습을 진행하였다. 학습은 다음 표 1과 같은 환경에서 수행되었다.

Fig. 5는 모델 학습 과정에서 각 epoch마다의 훈련 데이터 오차와 검증 데이터 오차의 변화를 나타낸다. 초기에는 오차가 빠르게 감소하였으며, 이후 점차 안정화되어 학습이 잘 진행되고 있음을 확인할 수 있다. 두 곡선이 유사한 형태를 보여 과적합 없이 모델이 효과적으로 일반화되고 있음을 알 수 있다.

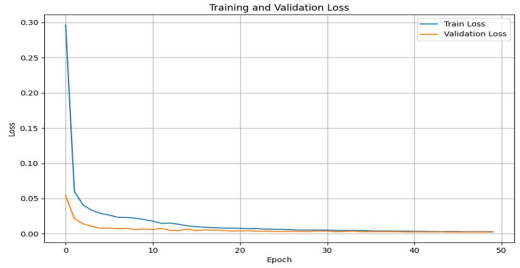

Fig. 5. Training and validation loss curves of the proposed model

2. Evaluation and Analysis

본 연구에서는 모델의 일반화 성능을 평가하기 위해 학습에 사용되지 않은 5 m × 5 m × 5 m 및 7 m × 7 m × 5 m 크기의 실내 공간에서 테스트를 수행하였다. 각 공간에서는 SNR를 각각 0dB과 5dB로 설정하여 총 네 가지 잡음 환경을 구성하였으며, 각 환경에 대해 시뮬레이션 기반의 합성 데이터를 5,000개씩 생성하여 평가에 활용하였다.

또한, 제안 모델의 실용성을 검증하기 위해 계산 복잡도와 실시간 처리 성능을 분석하였다. 모델은 약 6,000,000개의 학습 가능한 파라미터로 구성되어 있으며, NVIDIA GeForce RTX 4090 GPU에서 1초 오디오 샘플의 평균 추론 시간은 약 1.73 ms로 측정되어 실시간 응용에도 적합한 수준임을 확인하였다.

모델 성능은 세 가지 주요 정량적 지표로 평가하였다. 첫 번째는 평균 각도 오차(Mean Angular Error)로, 이는 예측된 음원 방향과 실제 방향 간의 각도 차이를 도(°) 단위로 측정한 것이다. 각도 값의 주기성을 고려하여 오차는 -180°에서 +180° 범위로 정규화되며, 다음과 같이 정의된다.

$$MAE_\theta = \frac{1}{n}\sum_{i=1}^{N}\left|((\hat{\theta}_i - \theta_i + 180)\bmod 360) - 180\right|$$

여기서 $\hat{\theta}_i$ 는 모델이 추정한 방위각, θ_i는 해당 샘플의 실제 방위각을 의미한다.

두 번째 평가지표는 두 벡터 사이의 유클리드 거리를 계산하여 평균화한 값이다. 이는 모델이 예측한 방향 벡터가 실제 음원 방향에 얼마나 근접하는지를 2차원 공간상에서 정량적으로 평가할 수 있도록 해줄 수 있으며. 다음과 같이 정의된다.

$$\frac{1}{N}\sum_{i=0}^{N}\sqrt{(x_{pred}^{(i)} - x_{true}^{(i)})^2 + (y_{pred}^{(i)} - x_{true}^{(i)})^2}$$

해당 값은 방향성 예측의 정밀도를 나타내며, 값이 작을수록 예측이 실제 방향에 가깝다는 것을 의미한다.

세 번째 평가지표는 RMSE(Root Mean Square Error)로, 두 번째 평가지표와 유사하게 2차원 방향 벡터 간 오차를 평가하지만, 오차를 제곱하여 평균한 후 제곱근을 취함으로써 큰 오차에 더 민감하게 반응한다. 따라서 RMSE는 모델이 예측 과정에서 발생하는 큰 방향 오차를 얼마나 효과적으로 줄이는지를 정량적으로 평가할 수 있다. 값이 작을수록 모델이 실제 음원 방향을 보다 정확하게 추정했음을 의미한다.

Table 2. Performance comparison in a 5 m × 5m × 5 m room (Euclidean Distance, angular errors, RMSE).

SNR	Method	Avg. Euclidean Distance Error	Avg. Angular Error (°)	RMSE
0	GCC-Phat	0.2419	15.2975	0.4638
0	MUSIC	0.2986	18.6276	0.5082
0	Proposed Model	0.1626	10.0163	0.2896
5	GCC-Phat	0.1447	8.8749	0.312
5	MUSIC	0.2438	15.0114	0.4326
5	Proposed Model	0.1001	5.8659	0.1841

제안 모델의 성능을 비교해서 기존의 대표적인 음원 위치 추정 기법인 GCC-Phat과 MUSIC을 이용하여 비교 분석하였다.

5 m × 5m × 5m 크기의 실내 공간에서 SNR 0dB 및 5dB 환경에서의 각 모델 성능 측정 결과는 Table 2에 제시되어 있다. 5 m × 5 m × 5m 환경에서 제안하는 모델은 모든 SNR 조건에서 기존의 GCC-Phat 및 MUSIC 기법 대비 우수한 성능을 나타내는 것을 볼 수 있다. SNR 0 dB과 같은 열악한 잡음 환경에서도 제안 모델은 평균 유클리드 거리 오차 0.1626, 평균 각도 오차 10.0163°, RMSE 0.2896를 기록하여, GCC-Phat 및 MUSIC에 비해

현저히 낮은 오차율을 보였다.

이는 제안한 딥러닝 기반 모델이 다양한 잡음 환경에서도 방향성 정보와 공간적 특성을 효과적으로 학습하여, 기존 전통적 알고리즘 대비 뛰어난 잡음 강인성과 일반화 능력을 보유함을 의미한다. 또한, 모든 실험 환경에서 일관되게 낮은 평균 오차를 기록하여 실제 복잡한 실내 환경에서도 신뢰성 높은 음원 위치 추정이 가능함을 확인했다.

Table 3. Performance comparison in a 7 m × 7 m ×5 m room (Euclidean Distance, angular errors, RMSE).

SNR	Method	Avg. Euclidean Distance Error	Avg. Angular Error (°)	RMSE
0	GCC-Phat	0.2457	15.5915	0.4672
0	MUSIC	0.2986	18.6178	0.5054
0	Proposed Model	0.1618	9.8353	0.2897
5	GCC-Phat	0.1456	9.0011	0.317
5	MUSIC	0.2459	15.3569	0.45
5	Proposed Model	0.1037	6.153	0.196

Table 3에 제시된 7 m × 7 m × 5 m 환경에서의 성능 평가 결과를 살펴보면, 5 m × 5 m × 5 m 환경과 유사하게 제안 모델이 기존의 GCC-Phat 및 MUSIC 기법에 비해 일관되게 우수한 성능을 나타냄을 확인할 수 있다. 특히, SNR 0dB 조건에서 제안 모델의 평균 유클리드 거리 오차는 0.1618, 평균 각도 오차는 9.8353°, RMSE는 0.89로 GCC-Phat 및 MUSIC 대비하여 현저히 낮은 오차를 보였다.

또한, SNR 5dB 환경에서도 제안 모델은 평균 유클리드 거리 오차 0.1037, 평균 각도 오차 6.153°, RMSE 0.196로 기존 기법들보다 더욱 향상된 정확도를 기록하였다. 이는 공간의 크기가 커짐에 따라 전통적 알고리즘의 오차가 다소 증가하는 반면, 제안 모델은 공간 변화에도 불구하고 낮은 오차를 안정적으로 유지함을 보여준다. 이러한 결과는 제안한 딥러닝 기반 모델이 다양한 실내 환경과 잡음 조건에서 뛰어난 적응력과 강인성을 갖추고 있음을 의미한다.

이러한 결과들을 종합해 볼 때, 제안하는 모델은 다양한 실내 공간 크기와 잡음 환경에 걸쳐 기존의 음원 위치 추정 기법들을 능가하는 강력한 성능과 뛰어난 일반화 능력을 보유하고 있음이 확인되었다. 특히, 낮은 SNR 조건에서도 안정적인 성능을 보인다는 점은 실제 응용 분야에서의 활용 가능성을 높이는 중요한 이점이다.

V. Conclusion

음원 위치 추정에 있어 잡음과 잔향은 성능에 중대한 영향을 미치는 요소로 작용한다. 특히 낮은 SNR 환경에서는 기존의 전통적인 수학 기반 접근법인 GCC-PHAT 및 MUSIC 알고리즘이 외부 간섭에 민감하게 반응하며, 정확한 위치 추정이 어려워지는 한계가 존재한다. 이러한 문제는 실내 환경에서 더욱 두드러지며, 실제 응용을 위한 기술적 제약으로 이어진다.

이를 해결하고자 본 연구에서는 오디오 스펙트로그램 트랜스포머 모델을 제안하였으며, 멀티채널 오디오로부터 GCC 특징으로 변환하여 모델에 입력함으로써 잡음에 강인한 음원 위치 추정을 가능하게 하였다. 실험 결과, 제안한 오디오 스펙트로그램 트랜스포머 모델은 기존의 GCC-PHAT 및 MUSIC 알고리즘과 비교하여 모든 SNR 조건에서 일관되게 낮은 평균 각도 오차와 거리 오차를 기록하였다. 특히, SNR 0 dB과 같은 극한의 잡음 환경에서도 5 m × 5 m × 5 m 환경에서 평균 각도 오차 10.0163°, 평균 거리 오차 0.1626, RMSE 0.2896를 기록하였으며, 7 m × 7 m × 7 m 환경에서도 평균 각도 오차 9.8353°, 평균 거리 오차 0.1618, RMSE 0.89의 우수한 성능으로 기존 전통 기법 대비 뛰어난 잡음 강인성과 정확도를 입증하였다. 이는 트랜스포머 모델의 전역적 특징 학습 능력이 잡음과 잔향이 심한 환경에서도 효과적으로 작용함을 보여주며, 음원 위치 추정 문제에 대한 실용적이고 신뢰성 높은 해결책임을 의미한다.

향후 연구에서는 2D 평면상의 방위각 추정을 넘어, 고도각 정보까지 포함한 3D 음원 위치 추정으로 모델을 확장하고, 잡음이 있는 복수의 음원이 동시에 존재하는 환경에서 다중 음원 위치 추정을 정밀하게 예측할 수 있는 모델 구조를 설계하여 분석할 예정이다. 또한, 실제 음향 데이터를 기반으로 한 정밀 평가를 통해 실제 환경에서의 모델 신뢰도와 응용 가능성을 더욱 강화하고자 한다.

ACKNOWLEDGEMENT

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2022-00156334)

REFERENCES

[1] Valzolgher, C., Capra, S., Gessa, E., Rosi, T., Giovanelli, E., & Pavani, F., "Sound localization in noisy contexts: performance, metacognitive evaluations and head movements," Cognitive Research: Principles and Implications, Vol. 9, No. 1, pp. 4, January 2024. DOI: 10.1186/s41235-023-00530-w

[2] Grumiaux, P. A., Kitić, S., Girin, L., & Guérin, A., "A survey of sound source localization with deep learning methods," The Journal of the Acoustical Society of America, Vol. 152, No. 1, pp. 107-151, July 2022. DOI: 10.1121/10.0011809

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint, June 2021. DOI: 10.48550/arXiv.2010.11929

[4] Gong, Y., Chung, Y. A., & Glass, J., "Ast: Audio spectrogram transformer," arXiv preprint, July 2021. DOI: 10.48550/arXiv.2104.01778

[5] Knapp, C., & Carter, G., "The generalized correlation method for estimation of time delay," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 24, No. 4, pp. 320-327, Dec. August 1976. DOI: 10.1109/TASSP.1976.1162830

[6] Lim, J. S., Cheong, M., & Kim, S., "Improved generalized cross correlation-phase transform based time delay estimation by frequency domain autocorrelation," The Journal of the Acoustical Society of Korea, Vol. 37, No. 5, pp. 271-275, Oct. September 2018. DOI: 10.7776/ASK.2018.37.5.271

[7] Schmidt, R., "Multiple emitter location and signal parameter estimation," IEEE Transactions on Antennas and Propagation, Vol. 34, No. 3, pp. 276-280, March 1986. DOI: 10.1109/TAP.1986.1143830

[8] Qayyum, A. B. A., Hassan, K. N., Anika, A., Shadiq, M. F., Rahman, M. M., Islam, M. T., ... & Haque, M. A., "DOANet: A deep dilated convolutional neural network approach for search and rescue with drone-embedded sound source localization," EURASIP Journal on Audio, Speech, and Music Processing, Vol. 2020, No. 1, p. 1-18, November 2020. DOI: 10.1186/s13636-020-00184-2

[9] Yalta, N., Nakadai, K., & Ogata, T., "Sound Source Localization Using Deep Learning Models," Journal of Robotics and Mechatronics, Vol. 29, pp. 37-48, February 2017. DOI: 10.20965/jrm.2017.p0037.

[10] Politis, A., Shimada, K., Sudarsanam, P., Adavanne, S., Krause, D., Koyama, Y., ... & Virtanen, T., STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. arXiv preprint, september 2022. DOI: 10.48550/arXiv.2206.01948

[11] Shimada, K., Politis, A., Sudarsanam, P., Krause, D. A., Uchida, K., Adavanne, S., ... & Mitsufuji, Y., STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. Advances in neural information processing systems, Vol. 36, pp. 72931-72957, 2024.

[12] Park, S. Jeong, Y. Lee, T. Many-to-Many Audio Spectrogram Tansformer: Transformer for Sound Event Localization and Detection. In Proceedings of the DCASE, Barcelona, pp. 105–109, Barcelona, Spain, November 2021. DOI: 10.5281/zenodo.5770113

[13] Allen, J. B., & Berkley, D. A., "Image method for efficiently simulating small-room acoustics," The Journal of the Acoustical Society of America, Vol. 65, No. 4, pp. 943-950, 1April 1979. DOI: 10.1121/1.382599

[14] Scheibler, R., Bezzam, E., & Dokmanić, I., Pyroomacoustics: A python package for audio room simulation and array processing algorithms, 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1-5, Calgary, Canada, April 2018. DOI: 10.1109/ICASSP.2018.8461310

[15] Diaz-Guerra, D., Miguel, A., & Beltran, J. R., "gpuRIR: A python library for room impulse response simulation with GPU acceleration," Multimedia Tools and Applications, Vol. 80, No. 4, pp. 5653-5671, February 2021. DOI: 10.1007/s11042-020-09905-3

[16] Aihub, Construction of Data and Civil Service Management Service for AI Learning of Living Environmental Noise, https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71296

[17] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V., “Specaugment: A simple data augmentation method for automatic speech recognition,” arXiv preprint, December 2019. DOI: 10.21437/Interspeech.2019-2680

[18] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H., "Training data-efficient image transformers & distillation through attention," International Conference on Machine Learning, Vol. 139, pp. 10347–10357, July 2021.

Authors

Won Jun Lee received the B.S degrees in Cyber Security from Pai Chai University, South Korea in 2024. He is currently pursuing M.S degree in the Department of Smart ICT Convergence at Pai Chai University. His current research interests include artificial intelligence, cybersecurity, and computer vision.

Woo Jin Jung is currently pursuing the B.S. degree in the Department of Cyber Security at Pai Chai University in Daejeon, South Korea. His current research interests include Artificial Intelligence, Cyber Security and Deepfake Detection.

Hyun-Jong Cha received the M.S. and Ph.D. degree in Computer science and Defense Acquisition Program from Kwangwoon University, South Korea, in 2008 and 2014. He is a professor in the Department of Computer Engineering at Pai Chai University in Daejeon, South Korea. His current research interests include information security, Artificial Intelligence, IoT, and Blockchain.

Ah Reum Kang received the M.S. and Ph.D. degrees in information security from Korea University, South Korea, in 2012 and 2016. She is a professor in the Department of Information Security at Pai Chai University in Daejeon, South Korea. Her current research interests include security, artificial intelligence, malware, medical data analysis, and online game security.