실시간 다중 이동 음원 정위를 위한 직접 경로 채널 간 위상 차이 추정 모델 경량화 연구

최유선, 김현승, 신종원*
광주과학기술원

newsun0130@gm.gist.ac.kr, kimhs355@gm.gist.ac.kr, *jwshin@gist.ac.kr

A Lightweight Direct-Path IPD Estimation Model for Real-Time Multiple Moving Sound Source Localization

Yuseon Choi, Hyeonseung Kim, Jong Won Shin*
Gwangju Institute of Science and Technology (GIST)

요 약

본 논문은 잡음 및 잔향이 존재하는 환경에서 다중 이동 음원 정위를 수행하기 위해, 전대역(Full-band) 및 협대역(Narrow-band) 정보를 융합한 LSTM 기반 신경망의 경량화 기법을 제안한다. 연산 효율성 향상을 위해, U-Net 인코더 구조를 차용하여, 시간 축을 다운샘플링(down-sampling)하고 압축된 특징(feature)을 추출하였다. 시뮬레이션 데이터셋 검증 결과, 제안한 모델은 모든 평가지표에서 기존 모델 대비 연산량을 크게 줄이면서 성능을 유지하였다.

Ⅰ. 서 론

음원 위치 추정(Sound Source Localization, SSL)은 마이크로폰 배열로부터 하나 이상의 음원 방향(Direction of Arrival, DoA)을 추정하는 것을 목표로 한다. 음원 정위에 사용되는 중요한 공간 특성(spatial feature)인 채널 간 위상 차이(Interchannel Phase Difference, IPD)는 음원이 다수의 마이크로폰에 도달하는 시간 차이를 주파수 영역에서의 위상 차이로 나타낸 것으로, 음원의 상대적인 방향 정보를 정밀하게 반영할 수 있는 중요한 단서이다. 특히, 직접 경로 채널 간 위상 차이(Direct-Path Interchannel Phase Difference, DP-IPD)는 반사나 잔향 없이 직접 마이크로폰으로 전파된 음원의 위상 차이를 의미하며, 마이크 배열 구조가 알려져 있을 경우, 이론적으로 DP-IPD 로부터 음원의 DoA (Direction of Arrival)를 추정할 수 있다.

본 연구에서 활용하는 베이스라인 모델인 IPDnet[1]은 다채널 마이크로폰 신호의 단기간 푸리에 변환(Short-time Fourier Transform, STFT) 결과의 실수부 및 허수부를 입력으로 받아, 전대역 및 협대역 융합 신경망을 통해, DP-IPD 를 예측한다. 추론 시, 한 프레임에서 예측된 DP-IPD 벡터와 모든 후보 방향의 이론적 DP-IPD 벡터 간 내적을 계산하여, 가장 큰 값을 가지는 방향을 음원의 DoA 로 추정한다.

모델 경량화에서 파라미터 수는 주로 메모리 사용량에 영향을 미치지만, 실제 추론 시간과 에너지 소비는 연산량에 의해 더욱 크게 좌우된다. 연산량이 적을수록 모델 실행 속도는 빨라지고 전력 소모는 줄어들어 모바일 및 임베디드 환경에서의 활용 가능성이 높아진다. 이에 따라, 본 연구는 경량화의 핵심 목표를 파라미터 수 감소보다는 연산량 절감에 두고 모델 구조를 설계하였다.

베이스라인 모델은 시간 프레임과 주파수 대역을 전 범위에 걸쳐 처리하는 구조로 인해, 높은 연산량을 요구하는 한계를 지닌다. 이를 완화하기 위해, 제안한 모델에서는 합성곱 계층(Convolutional layer)으로 구성된 다운샘플링 블록을 도입하여, 시간 축의 해상도를 점진적으로 축소함으로써 연산 효율을 개선하였다.

다중 음원 정위를 위해, 베이스라인 모델은 예측한 DP-IPD 벡터와 정답 벡터 간 가능한 모든 순열을 비교한 후, 최소 오차 조합을 선택하는 프레임 단위 순열 불변 학습(Permutation Invariant Training, PIT)에 기반한 손실함수를 사용한다. 시뮬레이션 데이터셋에서의 실험 결과, 제안한 모델은 기존 모델 대비 연산량을 51% 절감하면서도, 음원 정위 성능은 유사한 수준으로 유지하였다.

Ⅱ. 본론

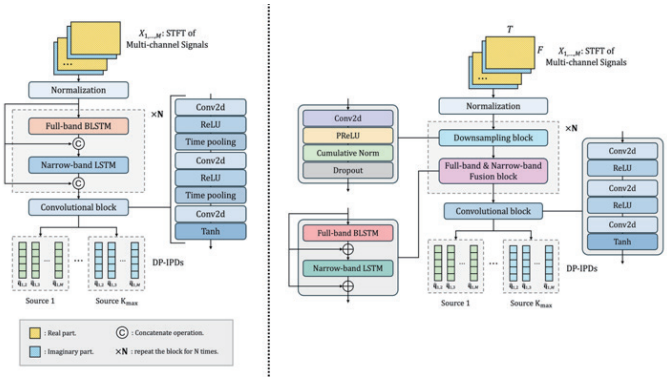

그림 1. 베이스라인 모델(왼쪽)과 경량화 모델(오른쪽) 구조도

Method	# Params. [M]	FLOPs [G/s]	Gross Accuracy [%]	Fine Error [°]
전·협대역 블록 1 개	0.47	11.88	88.54	2.34
전·협대역 블록 2 개 (IPDnet)	0.71	19.35	90.91	**2.01**
경량화 모델	1.49	9.4	**91.31**	2.18

표 1. 시뮬레이션 데이터셋에서 베이스라인 모델과 경량화 모델의 음원 정위 성능 비교

1. IPDnet 모델 및 경량화 구조 설계

그림 1에서 베이스라인 모델인 IPDnet의 전대역 양방향 LSTM 계층은 주파수 축을 따라 시간 프레임을 독립적으로 처리하여, DP-IPD의 주파수 간 상관관계를 효과적으로 포착한다. 반면, 협대역 LSTM 계층은 시간 축을 따라 각 주파수를 독립적으로 처리함으로써, 시간적 특징 정보를 학습한다. 두 계층은 스킵 연결(Skip connection)을 통해, 특징 추출 과정에서의 정보 손실을 최소화한다. 이후 합성곱 블록을 통해, 각 음원의 DP-IPD 벡터를 개별적으로 추출한다. 이 합성곱 블록은 인과 합성곱 계층(Causal convolutional layer)과 시간 평균 풀링 계층(Temporal average pooling layer)로 구성되며, 전자는 마이크 쌍 및 음원 간의 지역적 특성(local feature)을 포착하고, 후자는 시간 프레임 단위의 평균을 통해 시간 해상도(temporal resolution)를 압축한다. IPDnet은 전·협대역 블록 이후, 합성곱 블록의 시간 평균 풀링 계층을 통해 프레임 간 중첩 없이 12개 프레임의 평균값을 출력함으로써, 마지막 계층에서 시간 축을 압축하였다.

모델의 연산 효율성을 보다 향상시키기 위해, 그림 1의 경량화 모델에서는 U-Net 기반 인코더 구조[2]를 차용하였으며, 2D 합성곱 계층을 활용해 시간 축을 점진적으로 다운샘플링함으로써, 압축된 특징을 추출하였다. 또한, 베이스라인의 스킵 연결을 결합(concatenation)에서 합산(summation)하는 방식으로 변경하여, 연산 효율성을 추가적으로 개선하였다.

2. 실험 설정

본 논문에서는 180 도 방위각 범위에서 음원 위치를 추정하기 위해 두 개의 마이크를 사용한다. 마이크 신호는 음성 신호와 RIR(Room Impulse Responses)의 컨볼루션을 통해 생성된다. 음성 신호로는 LibriSpeech 데이터를 사용하였고, RIR 은 gpuRIR toolbox 를 통해 생성하였다. 잔향 시간(Reverberation Time, RT60)은 0.2 초에서 1.3 초까지, 방의 크기는 6×6×2.5 m 에서 10×8×6 m 사이에서 무작위로 설정하였다. 음성 신호의 이동 경로는 고정된 높이를 갖도록 구성하였다. 두 마이크는 8cm 간격으로 방 안에 무작위로 배치되며, 음원과 같은 수평면에 위치한다. 또한, NOISEX-92 데이터베이스에서 추출한 white, babble, factory 잡음을 사용하여 확산 음장(diffuse sound field)을 생성하고, 이를 음성 신호와 -5 dB 에서 15 dB 범위의 SNR 로 혼합하여 mixture 를 구성하였다.

3. 실험 결과

본 연구에서는 성능 평가지표[3]로, 오차 허용 범위 10 도를 기준으로 계산한 Gross Accuracy 와 Fine Error 를 사용하였다. 표 1 에서, 제안한 모델은 베이스라인 모델의 전·협대역 블록 1 개 구성 대비 파라미터 수는 증가하였으나, 연산량은 약 21% 절감하였고, 모든 평가지표에서 향상된 성능을 나타냈다. 전·협대역 블록 2 개 구성과 비교했을 때에도, 연산량을 약 51% 감소시키면서 유사한 수준의 성능을 유지하였다.

Ⅲ. 결론

본 연구에서는 두 개의 마이크로폰을 활용한 음원 정위 모델의 연산 효율성을 향상시키기 위해 경량화 모델을 제안하였다. 기존 모델은 시간 및 주파수 전 구간에 대해 탐색하는 구조적 특성으로 인해, 많은 연산량을 요구하는 한계가 있었다. 이를 개선하기 위해, 합성곱 계층으로 구성된 다운샘플링 블록을 통해 시간 축을 점진적으로 축소하여 연산량을 효과적으로 절감하도록 경량화 모델을 설계하였다.

실험 결과, 제안한 모델은 베이스라인 대비 연산량을 최대 51%까지 줄이면서도 음원 정위 성능 지표인 Gross Accuracy 와 Fine Error 에서 유사하거나 더 우수한 성능을 나타냈다. 연산량이 적은 환경에서도 일정 수준 이상의 정위 성능을 유지할 수 있음을 확인함으로써, 제안한 모델이 모바일 및 임베디드와 같이 연산 자원이 제한된 응용 분야에 적합함을 검증하였다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로

정보통신기획평가원-대학 ICT 연구센터(ITRC)의 지원을 받아

수행된 연구임(IITP-2025-RS-2021-II211835)

참 고 문 헌

[1] Y. Wang, B. Yang, and X. Li, "IPDnet: A universal direct-path IPD estimation network for sound source localization." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[2] 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).

[3] S. S. Battula, H. Taherian, A. Pandey, D. Wong, B. Xu and D. Wang, "Robust Frame-level Speaker Localization in Reverberant and Noisy Environments by Exploiting Phase Difference Losses," *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Hyderabad, India, 2025, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10890098.