# 다성 환경에서의 사운드 이벤트 검출을 위한 맘바 기반 이중 브랜치 구조

여예린, 김정현
세종대학교
yerinee0949@sju.ac.kr, j.kim@sejong.ac.kr

# A Dual-Branch Architecture with Mamba for Polyphonic Sound Event Detection

Yerin Yeo, Junghyun Kim
Sejong Univ.

요 약

본 논문에서는 다성 (polyphonic) 환경에서의 사운드 이벤트 검출을 위해 상태 공간 모델 Mamba를 활용한 이중 브랜치 구조를 제안한다. 제안 모델은 멜-스펙트로그램을 입력 받아 CNN으로 특징을 추출한 뒤, 이를 BiGRU와 Mamba 브랜치에 각각 병렬로 전달하고 두 브랜치의 출력을 더하여 이벤트 간의 상관관계와 시간적 변화를 효과적으로 반영한다. 실험 결과, 제안 모델은 기존 모델 대비 파라미터 수가 소폭 증가했지만 ER 성능과 F1-score 성능이 모두 향상되었다.

## Ⅰ. 서 론

사운드 이벤트 검출 (Sound Event Detection, SED)은 다양한 환경에서 발생하는 사운드 이벤트의 종류와 시점을 식별하는 과제로, 스마트시티, 보안, 상황인지 시스템 등 여러 분야에서 활용되고 있다 [1]. 최근에는 CRNN 기반 모델들이 높은 성능을 보이고 있으며, 이를 변형한 다양한 구조도 제안되었다 [2 - 3]. 그럼에도 불구하고, 다성 (polyphonic) 환경에서 동시에 발생하는 사운드 이벤트는 여전히 개별 이벤트 검출 성능을 낮추는 주요 요인으로 남아있다.

이러한 문제를 해결하기 위한 방법으로 오디오 소스 분리 기반 SED 연구가 활발히 진행되고 있으며, 특히 텍스트 쿼리에 따라 사운드 이벤트를 분리할 수 있는 언어 쿼리 기반 오디오 소스 분리 (Language-Queried Audio Source Separation, LASS) 모델이 주목받고 있다 [4]. LASS와 Convolutional Recurrent Neural Network (CRNN)을 결합한 프레임워크 [5]는 다성 환경에서도 높은 검출 성능을 보여주었지만, 단순한 CRNN 구조가 다성 환경의 이벤트 별 상관관계와 복잡한 시간적 변화를 충분히 반영하지 못한다는 한계가 존재한다. 본 논문에서는 이러한 한계를 해결하고자 Mamba를 활용한 이중 브랜치 구조를 제안하며, 실험을 통해 기존 모델보다 우수한 성능을 달성함을 확인하였다.

## Ⅱ. 본론

본 논문에서는 사전 학습된 LASS 모델을 사용하여 여러 이벤트가 혼합된 오디오로부터 11개 클래스의 멜-스펙트로그램을 추출하였다. 이렇게 얻은 스펙트로그램은 제안하는 이중 브랜치 구조의 입력으로 사용되며, 각 시간 프레임에서 발생하는 이벤트를 검출하는 데 활용된다. LASS 모델은 1069시간 규모의 오디오 - 텍스트 페어 데이터셋으로 학습되었으며, 텍스트 쿼리에 따라 특정 이벤트 신호를 분리할 수 있다.

제안 모델은 기존 CRNN 기반 검출 모델의 한계를 보완하기 위해 상태 공간 모델(State Space Model, SSM)인 Mamba를 활용한 이중 브랜치 구조로 설계하였다. 먼저, 입력된 멜-스펙트로그램에 3층 합성곱 블록을 적용하여 특징을 추출하였다. 시간 정보를 보존하기 위해 풀링은 주파수 축에만 적용하였으며, 크기는 차례로 5, 2, 2를 사용하였다. 이후 추출된 특징은 이벤트 간 상관관계와 시간적 변화를 학습하기 위해 두 개의 병렬 브랜치에 입력하였다. 첫 번째 브랜치는 시간적 맥락을 학습하기 위해 BiGRU를 사용하였다. BiGRU는 과거와 미래 정보를 모두 반영할 수 있지만, RNN 계열 특성상 순차적으로 계산되어 긴 시퀀스를 효율적으로 처리하기 어렵다. 이러한 한계를 보완하기 위해 두 번째 브랜치에는 상태 공간 모델 Mamba를 적용하였다. Mamba는 선형 복잡도로 동작하며 병렬 처리가 가능해 긴 시퀀스를 효율적으로 처리하고, 복잡한 다성 환경에서도 특징을 안정적으로 학습할 수 있다. 이후 두 브랜치의 출력을 합산하여 선형 층에 전달하고, 출력된 이벤트 발생 확률을 기반으로 각 시간 프레임마다의 이벤트 발생 여부를 검출하였다. 제안 모델의 전체 구조는 그림 1에 나타나 있다.

실험에는 실제 다성 환경을 반영하고 있는 MAESTRO-Real 데이터셋 [6]을 사용하였다. 이는 카페, 도심, 지하철역 등 5가지 장소에서 녹음된 오디오로 구성되며 총 11개 이벤트 클래스를 포함하고 있다. 우리는 사전 학습된 LASS 모델을 통해 클래스별 사운드 이벤트를 분리하고, 각 분리 신호로부터 64차원의 멜-스펙트로그램을 추출하였다. 이 과정을 통해 얻은 분리된 멜-스펙트로그램 특징을 제안하는 검출 모델에 입력하여 학습과 평가를 진행하였다.

학습과 평가는 5-fold 교차검증 방식으로 진행하였다. 전체 데이터를 다섯 부분으로 나눈 뒤, 각 단계에서 네 부분은 학습에, 한 부분은 검증에 사용하였다. 이 과정을 다섯 번 반복하여 모든 데이터가 검증에 한 번씩 포함되도록 하였으며, 최종 성능은 다섯 번의 평균으로 계산하였다. 이를 통해 데이터 편향을 줄이고 모델의 일반화 성능을 안정적으로 평가할 수 있었다. 학습 과정에서는 Adam 옵티마이저를 사용하였으며 초기 학습률은 0.001로 설정하였다. 또한 배치 크기는 32로 두었고, 과적합을 방지하기 위해 dropout 비율을 0.2로 적용하였다.

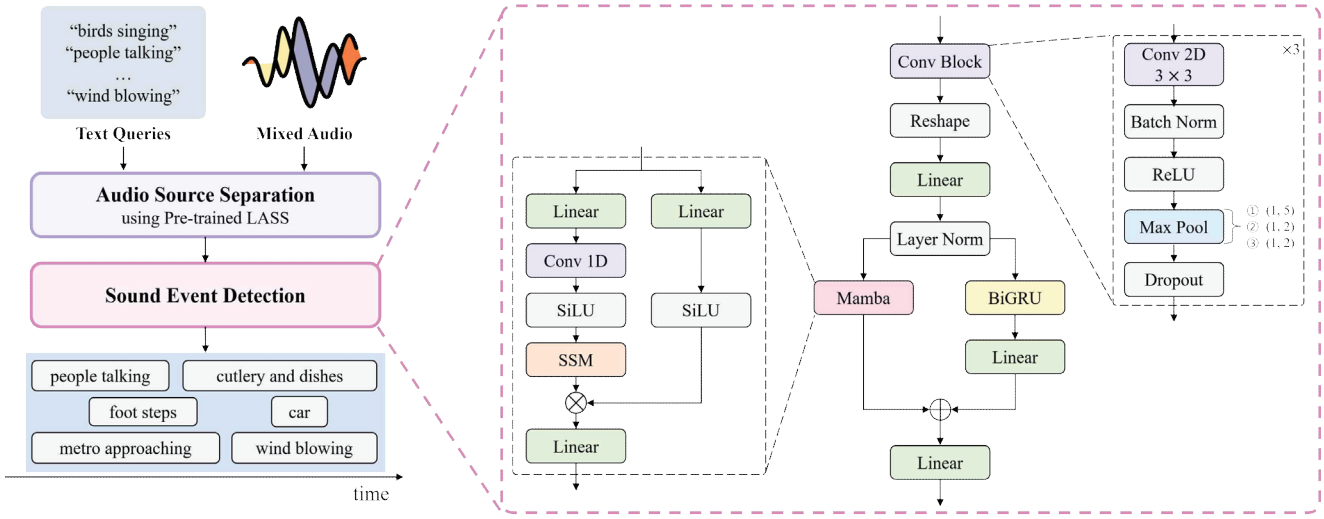


그림 1. 제안하는 사운드 이벤트 검출 모델 구조

본 논문에서는 성능 평가 지표로 Micro-average Error Rate (ER)와 Macro-average F1-score를 사용하였다. ER은 실제 발생한 이벤트를 검출하지 못한 경우와 존재하지 않는 이벤트를 잘못 검출한 경우를 모두 고려하여 모델의 이벤트 검출 성능을 평가한다. 이는 모든 클래스의 예측 결과를 통합해 계산하는 micro-average 방식으로 계산되며, 값이 낮을수록 검출의 안정성과 신뢰도가 높음을 의미한다. F1-score는 정밀도와 재현율의 조화평균으로 정의되며, 실제 이벤트가 발생했을 때 이를 얼마나 정확하고 빠짐없이 검출했는지를 평가한다. 클래스 불균형의 영향을 줄이기 위해 각 클래스의 F1-score를 평균하는 macro-average 방식을 적용하였다. SED에서는 이벤트 클래스마다 발생 빈도와 지속 시간이 달라, 단일 임계값으로 출력 확률을 이진 판정할 경우 일부 클래스의 성능이 왜곡되는 문제가 발생한다. 이를 보완하기 위해 F1-score를 계산하는 과정에서 기존 연구 [5]를 따라 클래스별 최적 임계값을 적용하고, 각 이벤트의 검출 확률 분포를 기반으로 기준 값을 조정하였다. 이러한 절차를 통해 클래스 불균형을 완화하고, 다성 환경에서도 개별 이벤트 검출 결과를 보다 신뢰성 있게 평가할 수 있다.

실험 결과, 제안 모델은 ER 0.413과 F1-score 51.62를 달성하여 기존 모델 [5]의 ER 0.425와 F1-score 50.97 보다 개선된 성능을 보였다. 제안 모델은 기존 모델 대비 파라미터 수가 소폭 증가하였지만, ER 성능과 F1-score 성능이 모두 향상되었다. 이에 대한 자세한 결과는 표 1에 제시되어 있다.

표 1. 기존 모델과 제안 모델의 사운드 이벤트 검출 성능 및 복잡도 비교

| Model | 기존 모델 [5] | 제안 모델 |
|---|---|---|
| ER | 0.425 | 0.413 |
| F1-score | 50.97 | 51.62 |
| 파라미터 수 | 4,175,435 | 4,513,355 |

## Ⅲ. 결론

본 논문에서는 다성 환경에서의 사운드 이벤트 검출 성능을 향상시키기 위해 상태 공간 모델 Mamba를 활용한 이중 브랜치 구조를 제안하였다. 제안 모델은 사전 학습된 LASS 모델을 사용하여 이벤트별 음성을 분리한 뒤, 분리된 멜-스펙트로그램을 입력으로 받아 CNN으로 특징을 추출한다. 이후 이 특징을 BiGRU와 Mamba 브랜치에서 병렬로 처리하고, 두 출력을 결합하여 시간적 정보와 특징 표현을 동시에 학습하였다. 실험 결과, 제안 모델은 기존 모델 대비 파라미터 수가 소폭 증가했지만 ER 성능과 F1-score 성능이 모두 향상되었다. 향후에는 사전 학습된 오디오 - 텍스트 모델 LASS를 이벤트별 쿼리와 실제 소음 환경 데이터에 맞게 미세 조정하고, 이를 제안 모델과 공동 학습 방식으로 결합하여 정밀한 검출 성능을 강화하고 실제 환경에서도 강건한 일반화 능력을 확보하고자 한다.

## ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원)사업 연구 결과로 수행되었음 (IITP-2023-RS-2023-00254529).

## 참 고 문 헌

[1] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 887 - 900, Mar. 2021.

[2] E. Cakır, G. Parascandolo, T. Heittola, et al., "Convolutional recurrent neural networks for polyphonic sound event detection," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 6, pp. 1291 - 1303, Jun. 2017.

[3] S. Adavanne, A. Politis, J. Nikunen, et al., "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," IEEE J. Sel. Topics Signal Process., vol. 13, no. 1, pp. 34 - 48, Jan. 2018.

[4] X. Liu, H. Liu, Q. Kong, et al., "Separate what you describe: Language-queried audio source separation," in Proc. Interspeech, Sep. 2022, pp. 1801 - 1805.

[5] H. Yin, H. Han, J. Chen, et al., "Exploring text-queried sound event detection with audio source separation," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2025, pp. 1 - 5.

[6] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 31, pp. 902 - 914, Apr. 2023.