# 합성 및 실제 데이터를 활용한 트랜스포머 구조 기반의 음향 이벤트 추정 및 탐지

신영서[1], 고창환[1], 장범석[2], 전찬준[1*]

조선대학교[1], 비에스소프트[2]

{ys070400, 20174249}@chosun.kr, bsjang@bs-soft.co.kr, cjchun@chosun.ac.kr

# Transformer-Based Sound Event Localization and Detection Using Synthetic and Real Data

Yeongseo Shin[1], Changhwan Go[1], Bumsuk Jang[2], Chanjun Chun[1*]

Chosun University[1], BS Soft Co., Ltd.[2]

요 약

음향 이벤트 감지(Sound Event Detection, SED)는 발생하는 음향의 종류, 음향의 발생 시점과 소멸 시점을 판단해내는 작업이며 음원 위치 탐지(Sound Source Localization, SSL)는 마이크를 기반으로 음원의 방향이나 위치를 식별하는 작업이다. 음향 이벤트 위치 파악 및 탐지(Sound Event Localization and Detection, SELD)는 SED와 SSL이라는 2가지 문제를 해결해야 한다. 본 논문에서는 DCASE 2022 Task3의 베이스라인과 데이터 세트를 활용하여 보다 나은 성능의 음향 이벤트 추정 및 탐지를 위한 방법을 제안한다. 주어진 데이터 세트는 합성 데이터와 실제 데이터로 이루어져 있으며 실제 데이터는 합성 데이터에 비해 매우 적은 양만 존재한다. 데이터양의 불균형 문제를 해결하기 위해 여러 개의 제너레이터를 활용하는 방법을 적용하였으며 데이터를 증대 방법으로 SpecAugment를 적용하였다. 또한 모델의 구조는 SELD에서 주로 사용되는 합성곱 순환 신경망(Convolutional Recurrent Neural Network, CRNN) 구조에 트랜스포머 인코더 모듈을 적용하였으며 단일 모델로 학습시킨 후 앙상블을 적용하였다. 제안 방법으로 수행한 결과 베이스라인 시스템에 비해 성능이 향상됨을 확인하였다.

## Ⅰ. 서 론

음향 이벤트 위치 파악 및 탐지(Sound Event Localization and Detection, SELD)는 음향 클래스를 식별하고 해당 음향 이벤트의 시작과 끝 지점, 도착 방향(Direction of Arrival, DOA)을 추정하는 작업이 수행된다. SELD는 음향 이벤트 감지(Sound Event Detection, SED)와 음원 위치 탐지(Sound Source Localization, SSL)라는 2가지 문제로 나눠 볼 수 있다. DCASE(Detection and Classification of Acoustic Scenes and Event)에서는 오디오 데이터 기반의 다양한 챌린지가 존재하는데 Task 3에서는 SELD에 대해 수행한다. 본 논문에서는 DCASE 2022 Task 3의 베이스라인과 데이터를 활용하여 실제 공간에서의 음향 장면에서의 SELD를 수행하였다 [1-2]. 데이터 세트로는 실제 공간에서 녹음된 소량의 실제 데이터와 학습을 위한 다량의 합성 데이터를 사용하였으며 1차 엠비소닉 형식의 데이터를 사용하였다. 데이터의 불균형 문제를 해결하기 위해 여러 개의 제너레이터를 활용하였으며 입력 특징으로는 다채널 로그-멜 스펙트로그램과 강도 벡터를 사용하였다. 또한 SED와 SSL을 해결하기 위해 SELDNet 모델을 기반으로 진행하였으며 3차원 공간에서 중첩된 음향 이벤트에 대해 SELD를 수행하기 위해 기존의 합성곱 순환 신경망(Convolution recurrent neural network, CRNN) 모델에서 트랜스포머 인코더 모듈을 사용하여 시간적 정보를 학습하는 방법을 제안한다 [3]. 모델의 출력 형식은 Multi-ACCDOA형식을 사용하였으며 [4], 성능을 증대하기 위해 약간의 다른 조건들로 학습시킨 여러 모델을 앙상블로 결합하여 성능을 확인하였다.

## Ⅱ. 제안 방법

DCASE에서 제공하는 데이터 세트는 1,200개의 합성 데이터와 121개의 실제 데이터로 구성되어 있다. 본 연구에서는 모델의 성능을 향상 시키기 위해 제공된 외부 데이터를 사용하여 총 5,100개의 합성 데이터를 추가 생성하여 학습에 사용하였다. 합성 데이터를 생성하기 위한 하이퍼파라미터로는 최대 동시 이벤트 수는 2, 데이터의 길이는 60초, 샘플링 레이트는 44.1kHz로 설정하였다. 신호 대 잡음비(signal-to-noise ratio, SNR)는 6~31의 값 중 랜덤하게 설정하였고, 동적인 음향 이벤트의 속도는 10.0, 20.0, 40.0으로 설정하였다. 학습용 데이터 세트는 합성 데이터와 실제 데이터 간의 양의 불균형이 존재한다. 이러한 문제를 해결하기 위해 여러 개의 제너레이터를 사용하여 배치마다 데이터의 비율을 설정하여 모든 배치마다 실제 데이터를 사용한 학습을 진행하였다. 마지막으로 데이터의 양이 한정적이기 때문에 모든 학습용 데이터에 증강 기법을 적용하여 학습을 진행하였다. 사용된 데이터 증강 기법으로는 음성 인식에서 널리 사용되는 SpecAugmrnt를 사용하였으며 주파수 마스킹과 시간 마스킹 기법을 사용하였다 [5].

데이터 세트에서 특징을 얻기 위한 방법으로는 제공된 베이스라인과 동일한 방식을 사용하였으며[2] first-order Ambisonic (FOA)형식의 데이터를 사용하였다. 입력 특징으로는 각 오디오 파일마다 4개의 로그-멜 스펙트로그램과 3개의 강도 벡터, 총 7개의 채널을 사용하였다. 스펙트로그램을 추출하는 방법과 추출한 데이터의 정규화 과정은 베이스라인과 동일한 방식을 적용하였다.

모델은 베이스라인의 CRNN 구조를 기반으로 생성되었으며 단일 분기에서 SED와 DOA를 예측하기 위해 Multi-ACCDOA를 적용하였다. 모델의 성능을 향상 시키기 위해 기존의 RNN층을 트랜스포머로 변경해보는 시도를 하였고 최종적으로 제안된 모델은 그림 1과 같다. 여러 개의 제너레이터를 사용하여 일정 비율만큼 데이터를 받아온다. 제너레이터에서 생성된 데이터는 먼저 CNN층을 통과하고 Residual Block의 입력 모양에 맞춰 변경된다. 그다음, 3층의 Residual Block을 통과한다. Residual Block은 2차원 합성곱 층과 풀링 층을 통과하고 마지막에 Dropout이 적용되는 구조로 이루어져 있다. Residual Block을 통과한 후 데이터는 트랜스포머 인코더를 통해 완전 연결층을 통과하고 활성화 함수로 하이퍼볼릭

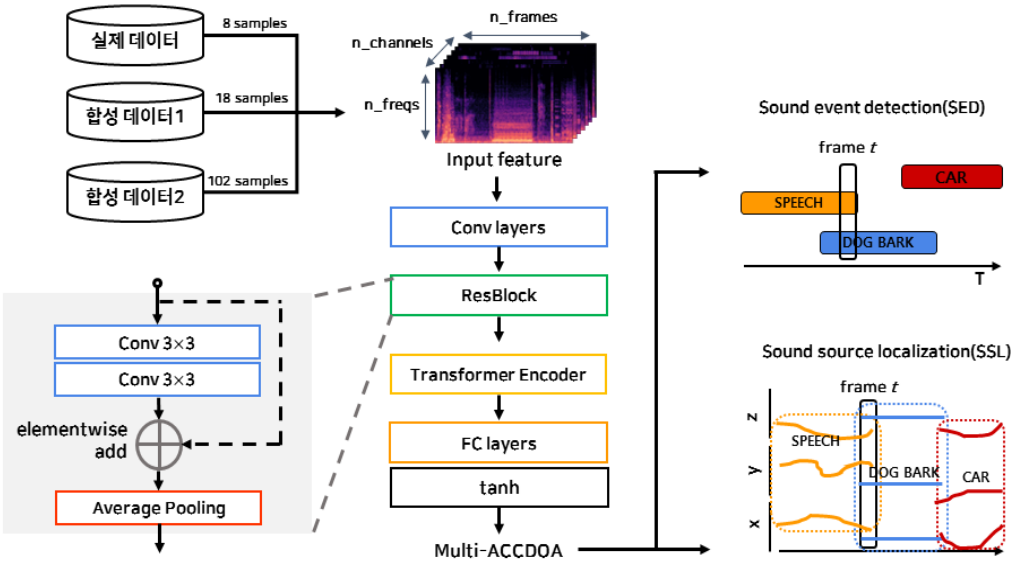


그림 1. 음원 추정 및 탐지를 위한 합성곱 순환 신경망 구조

탄젠트를 적용하여 Multi-ACCDOA 형식의 출력을 생성한다.

학습 파라미터로 옵티마이저는 Adam [6]과 Nesterov Momentum Adam (NAdam) [7]을 사용하였고 dropout은 0.2로, learning rate는 $2\times10^{-3}$과 $10^{-3}$으로 설정하여 각 결과를 비교하였다. 또한 스케줄러를 이용하여 30 epoch 이후 코사인 주기 함수에 따라 learning rate를 변경하는 방법으로 학습을 진행하였다.

학습 조건들을 다르게 적용하여 여러 모델을 학습시켰고 학습 후 7개의 단일 모델을 선택하여 앙상블을 적용하였다. 표 1은 베이스라인, 단일 모델, 앙상블에 대한 결과를 나타낸다. 모델의 평가 지표로 SED에 관해서는 error rate와 F-1 score를 적용하며, SSL에 관해서는 Localization error와 Localization recall을 적용한다 [8]. 테스트 결과 기존의 베이스라인보다 좋은 성능을 보인 것을 확인할 수 있었다.

표 1. 모델에 대한 테스트 결과 (베이스라인, 단일 모델, 앙상블)

| Model | ER | F | LE | LR | SELD |
|---|---|---|---|---|---|
| Baseline | 0.71 | 0.21 | 29.3 | 46.0 | 0.57 |
| Proposed | 0.65 | 0.31 | 26.41 | 0.62 | 0.47 |
| Ensemble | 0.59 | 0.33 | 24.63 | 0.61 | 0.47 |

## Ⅲ. 결 론

본 논문에서는 트랜스포머 구조를 이용한 음향 이벤트 추정 및 탐지 방법을 제안하였으며 실제 장면을 녹음한 오디오에서 성능을 평가하였다. 모델 학습을 위해 실제 데이터와 합성 데이터를 사용하였으며 데이터양의 불균형을 해결하기 위해 학습마다 정해진 비율만큼 각 데이터를 사용하는 방법을 적용하였다. 또한 Residual Block과 트랜스포머 인코더를 적용한 모델을 제안하였고 제안된 모델이 기존의 베이스라인 모델에 앙상블을 적용한 결과보다 성능이 오른 것을 확인하였다. 또한 제안된 모델에 앙상블을 적용하여 성능이 향상된 것을 확인할 수 있었다.

## ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업 (2017-0-00137) 및 한국연구재단의 지원을 받아 수행되었음 (NRF-2019R1C1C1011597)

## 참 고 문 헌

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

[2] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *arXiv preprint arXiv:2206.01948*, Jun. 2022.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, K. Lukasz, and P. Illia, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 30, pp. 6000-6010, Dec. 2017.

[4] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 316-320, May 2022.

[5] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, pp. 2613-2617, Sep. 2019.

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, May 2015.

[7] T. Dozat, "Incorporating nesterov momentum into Adam," in *Proc. International Conference on Learning Representations (ICLR)*, pp. 1-4, Feb. 2016.

[8] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in DCASE 2019", in *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp.684-698, Dec. 2020.