ICASSP 2022 L3DAS22 CHALLENGE: ENSEMBLE OF RESNET-CONFORMERS WITH AMBISONICS DATA AUGMENTATION FOR SOUND EVENT LOCALIZATION AND DETECTION

Paper ID: 9336

Yongjian Mao, Ying Zeng, Hongqing Liu, Wenbin Zhu, and Yi Zhou

School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing, China

Context

- **SELD Task:** detect the occurrences of sound events and localize them even when multiple events overlap both temporally and spatially

- **Motivation:**
  - ✓ We can also get the location of a sound event when we hear it
  - ✓ Sound event detection (SED) and sound source localization (SSL) have consistency in a labeled dataset
- **Applications**:
  Machine listening
  acoustic scene analysis
  audio surveillance in intelligent homes and cities

Implementation

Dataset and augmentation

- **L3DAS22 dataset** [1]:
  7.5 hours of B-format first-order Ambisonics recordings
  14 transient classes are to be detected
- **ACS [2]:** Sound field transformation corresponds to sign inversion, channel swapping, or both of FOA audio, the dataset can be expanded **eightfold**

$$\mathbf{Y}(\theta,\phi)=\frac{1}{4\pi}\begin{bmatrix}1\\ \sqrt{3}\sin\phi\cos\theta\\ \sqrt{3}\sin\theta\\ \sqrt{3}\cos\phi\cos\theta\end{bmatrix} \xrightarrow[\text{Sound field transformation}]{(\phi=\phi+\pi/2,\ \theta=-\theta)} \mathbf{Y}_{rot}(\theta,\phi)=\frac{1}{4\pi}\begin{bmatrix}1\\ \sqrt{3}\cos\phi\cos\theta\\ -\sqrt{3}\sin\theta\\ \sqrt{3}\sin\phi\cos\theta\end{bmatrix}$$

- **TFM:** randomly masks consecutive time frames or frequency bands of input features

Loss function

- Binary cross entropy (BCE) loss function for SED-RCnet
- Mean square error (MSE) loss function for SSL-RCnet
- Weighted loss function of both for SELD-RCnet

Model ensemble

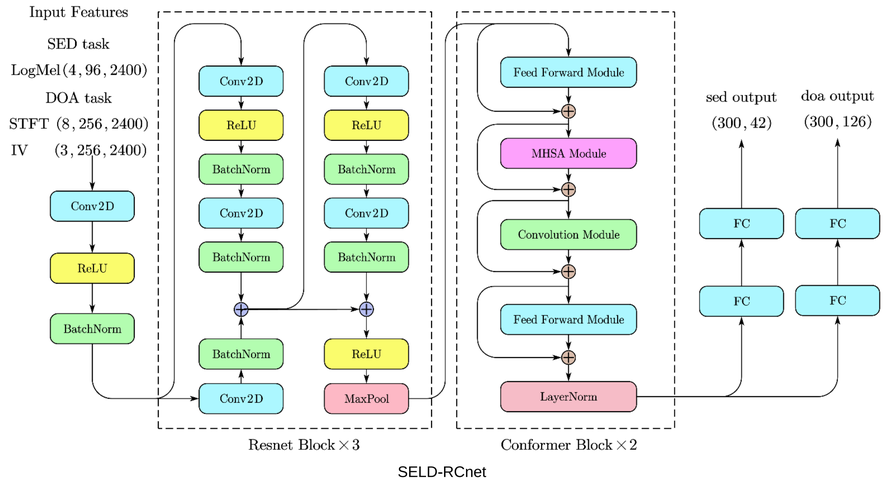
Simple average of the outputs predicted by different models

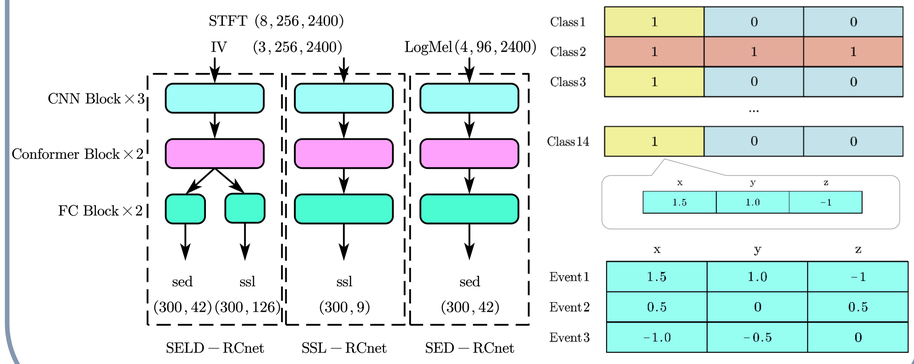
Table 1. Ensemble of Models

Model	*Output*	*LossWeights*[1]	*DataAug*
SELD-RCnet	SED1	(1,5)	ACS;TFM
SELD-RCnet	SED2	(1,5)	TFM
SED-RCnet	SED3	(1,0)	TFM
SSL-RCnet	SSL	(0,1)	ACS;TFM

[1]SED and SSL loss weights in training.

Proposed System





Evaluation and Results

T2 metric

- Location-sensitive detection error and computed on each time frame
- Measures the Cartesian distances between the predicted and true events with the same label, and counts a true positive only when its label is correct and its location is within a threshold from its reference location
- In the range [0, 1], where the higher the value is, the better results the system produces

Result

- Our method outperforms the baseline [1][3] by an overall **0.25** T2 metric
- Our submitted system won **the second place** in the L3DAS22 challenge [1]
  L3DAS22 Results: https://www.l3das.com/icassp2022/results.html

Table 2. Subjective evaluation on the L3DAS22 Challenge

Methods	*T2 Metric*	*Precision*	*Recall*
SELDnet(baseline)	0.343	0.423	0.289
Proposed system	**0.592**	**0.600**	**0.584**

System Description

SELD-RCnet

- **Input features:**
  STFT magnitude and phase spectrograms + Intensity vectors
- **Resnet blocks:** catch local fine-grained features, extract high dimension information, and improve the performance of the system
- **Conformer blocks:** learn both the local features and temporal context information and output a feature, inspired by [2]
- **Fully connected layers:** map the features into final SED and SSL representations.

SSL-RCnet

- Only predicts SSL target with a **special SSL representation**

SED-RCnet

- Only predicts SED target
- Take Log-Mel spectrograms as input features

Output representation

- **SED target:** the model predicts a matrix of the shape (300, 42) with the value in the range of [0,1], which represents the status of 14 different events in 300 frames. These values are thresholded to map the SED output to true or false values which indicates the sound event is active or inactive at the frame, respectively
- **SSL target:** the model predicts a separate location for all possible sounds events.
- **Especially, SSL-RCnet** only predicts three sets of Cartesian coordinates in the order of active events

Findings

- SELD-RCnet's outputs usually produce too many small and DOA-invalid predictions which confuse the DOA estimation
- The accuracy is dependent on SED results strongly when masking invalid DOA predictions with zeros based on SED predictions
- DOA-invalid values close to zeros can guide estimation of SED in each position of an event in a joint SELD network, though introducing the difficulties of estimating DOA

References

[1] E. Guizzo, C. Marinoni, M. Pennese, X. Ren, X. Zheng, C. Zhang, B. Masiero, and D. Comminiello, "L3DAS22 challenge: Learning 3D Audio Sources in a Real Office Environment," in 2022 IEEE International Conference on Acoucstic, Speech, and Signal Processing (ICASSP), 2022, pp. 1–6.
[2] Qing Wang, Jun Du, Hua-Xin Wu, Jia Pan, Feng Ma, and Chin-Hui Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," 2021.
[3] 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, 2019.