SOUND EVENT LOCALIZATION AND DETECTION BASED ON CROSS-MODAL ATTENTION AND SOURCE SEPARATION

Technical Report

Jin-Young Park[1], *Do-Hui Kim*[1], *Bon Hyeok Ku*[1], *Jun Hyung Kim*[1],
Jaehun Kim[2], *Kisung Kim*[2], *Hyungchan Yoo*[2], *Kisik Chang*[2], *Hyung-Min Park*[1]

[1]Dept. of Electronic Engineering, Sogang University, Seoul 04107, South Korea
[2]AI Lab, IVS Inc., Seoul 04799, South Korea
{wlsdud7907, a20151411, nine4409, imalbert}@sogang.ac.kr
{jhkim2, kskim, hycyoo, Honors}@ivstech.co.kr, hpark@sogang.ac.kr

ABSTRACT

Sound event localization and detection (SELD) is a task that combines sound event detection (SED) and direction-of-arrival(DOA) estimation (DOAE). This year's SELD task focuses on evaluation on real spatial scene, raising the difficulty for two reasons: 1) increase in overlapped events 2) noise-like events combined with real noises. In order to overcome this, we applied source separation and improved data synthesis logic to our basic (DCMA-SELD) model that utilizes dual cross-modal attention (DCMA) and soft parameter sharing of SED and DOAE streams to simultaneously detect and localize sound events. In order to improve the SELD performance of male/female speech that accounts for a large portion of input sounds, the source separation in our method was performed to separate speech signals from other sounds. Regarding the data synthesis logic, sound events that occur in real life may have some regularity, such as a laugh event that occurs in people's conversations or background music that has a long duration. Instead of data synthesis by mixing random sound events at random times, therefore, we added several rules to simulate more natural data that can learn the context of the events. Experimental results on validation data showed that our proposed approach successfully improved the performance of the task focusing on real spatial scene.

Index Terms— DCASE2022 Task3, sound event localization and detection, source separation, dual cross-modal attention, data augmentation, data synthesis

1. INTRODUCTION

Sound event localization and detection (SELD) is a challenging task that simultaneously requires estimation of direction-of-arrival(DOA) and detection and classification of sound event types (SED). With advances of deep learning on acoustic analysis field, complex acoustic tasks that even human auditory system cannot perform perfectly, such as SELD and source separation, has been recently considered. Cakir *et al.* [1] presented an approach for SED that utilized convolutional neural networks (CNNs) to extract acoustic features in a variety of environments, and recurrent neural networks (RNNs) to analyze long-term temporal features. On the other hand, DOA has been estimated by either approaches based on signal processing like maximum likelihood estimation [2] or deep learning [3]. Also, in the DCASE2020 challenge, Cao *et al.* [4] proposed a two-level network named event independent network V2 (EINV2) that achieved SELD based on track-wise estimation.

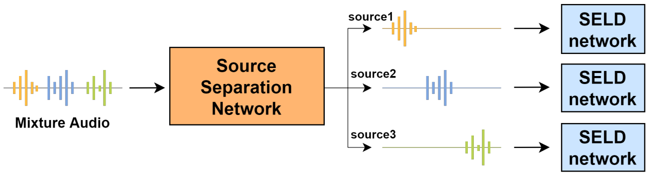

Figure 1: SELD with source separation.

Since 2019 when the SELD task first started in the DCASE challenge, various approaches have been proposed, such as ensemble of many models [5], feature adaptation named SALSA [6], network scaling approach using neural architecture search (NAS) [7], and cross-modal attention with parameter sharing [8].

This year's SELD task especially focuses on the real spatial sound scene, while past challenges used synthesized data for both training and evaluation. The primary difference due to real data evaluation is more overlapped events. Even though synthesized data can be controlled strictly while composing noise and events, real data have inevitably much complicated event structure, which has up to five overlapping events. Also, real data have real environmental noises and event sequences that have different properties from synthesized data. Real data may have various types of distortions in the noises due to echo noise, room characteristic reverberation, and device noise. Also, real data have eventual contexts as these data are usually made with a scenario to give more plausible contexts to the data, such as most of laughing and clapping that occur between male or female speech.

In order to address more overlapped events, we apply a source separation network to separate male/female speech that accounts for a large portion of input sounds before SELD as shown in Fig. 1. There have been similar attempts to adopt source separation to separately interpret overlapped sound events in the DCASE2020 Task4, "sound event detection and separation in domestic environments" [9]. We try to apply a similar approach to the SELD task to separate human speech from other events. Regarding the data synthesis logic, sound events that occur in real life may have some regularity, such as a laugh event that occurs in people's conversations or background music that has a long duration. Instead of data synthesis by mixing random sound events at random times, therefore, we added

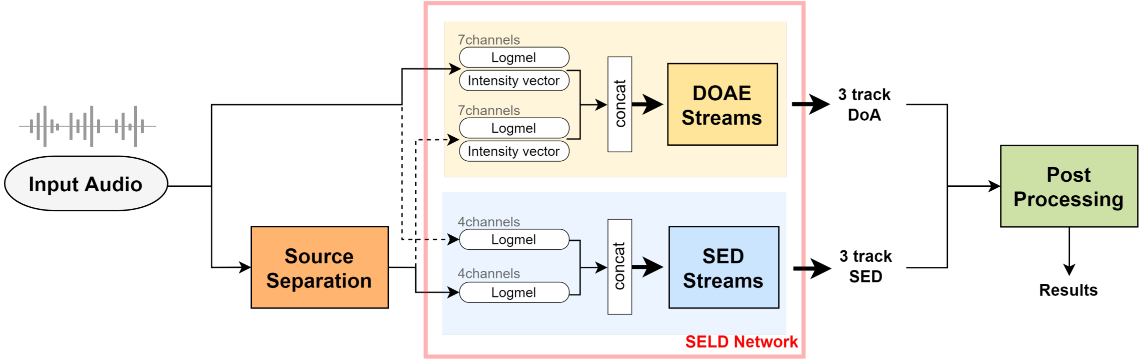

Figure 2: Overall system architecture of our proposed method.

several rules to simulate more natural data that can learn the context of the events.

2. DATABASES

The Sony-TAU realistic spatial soundscapes 2022 (STARSS-22) database [10] is provided as a DCASE2022 Task3 dataset. The data was collected in real spatial sound scenes, and the fold3 and fold4 are used as training and validation, respectively. However, due to its limited quantity, synthesized data, the DCASE 2022 simulated data for baseline training, were provided from the TAU spatial room impulse response database (TAU-SRIR DB). These databases are all in a sampling rate of 24 kHz and provide four-channel microphone raw data and FOA-format data. In this paper, we only used the FOA-format data as an input.

Two major differences of the DCASE2022 Task3 from the DCASE2021 Task3 are that the data are recorded in real spatial sound scene and have 13 different SED target classes, increasing by one event class from last year. Furthermore, the STARSS-22 database has up to five simultaneous events, while up to three overlapped events were occurred last year.

Also, Audioset [11] and FSD-50K [12] are used as sound events to create new synthesized data. Other databases were needed to create mixtures to train our source separation network, so we used CSS 10 - Japanese [13] and JSUT [14] for Japanese data, while using Voxceleb1 [15] for English data.

3. PROPOSED METHODS

Our SELD system adopts a source separation network and then separated sounds are input to the SELD network. As you can see in Fig. 2, four-channel FOA-format input data pass through the source separation network to obtain a speech component from each channel input. Then, the resulting four speech components are concatenated with their original four-channel input data. The concatenated input signals are input to our basic SELD network that utilizes dual cross-modal attention (DCMA) and soft parameter sharing of SED and DOAE streams. Finally, we post-process prediction results by replacing their outliers with the mean value.

3.1. Data Synthesis

We created two datasets for two tasks, source separation and SELD. Each dataset was used to train the source separation and SELD networks described in Sections 3.3 and 3.4, respectively.

3.1.1. Source separation data synthesis

Since we applied source separation to the SELD task, a dataset was created to successfully train the source separation network. In particular, our source separation model aimed to separate speech, that accounts for a large portion of sound events, from other sound events, so we combined female/male speech with other events. To obtain sound events without containing speech at all, we used the FSD-50k and AudioSet. However, we found out that noisy events as well as Japanese speech and female speech were not sufficient to train the network when we extracted sources from the databases. Therefore, background music and noise were added to speech with a signal-to-noise ratio (SNR) of 15 dB and 10 dB, respectively. At the same time, Japanese speech data from the CSS 10 and JSUT and English speech data from the Voxceleb1 were added.

3.1.2. SELD data synthesis

In case of the SELD database, synthesis of more plausible data is additionally applied to overcome the mismatch of training in synthesized data and evaluating in real data. Since most of the real-world data have contextual events or correlation between the events that the model should learn, we tried to give some handcrafted rules on data synthesis to overcome the mismatch.

3.2. Data Augmentation

In case of data augmentation, many papers tried to overcome the data limitation with various approaches. We adopted two data augmentation methods that critically supported the synthesized data to learn real spatial scene. First, we use the mixup technique to give a variety to audio clips by adding up weighted audios. Secondly, we applied channel rotation on the FOA-format data without losing the physical relationships between steering vectors and observations by the reflection and rotations for the elevation or azimuth [16]. In addition, we applied the SpecAugment [17], but could not obtain improved performance.

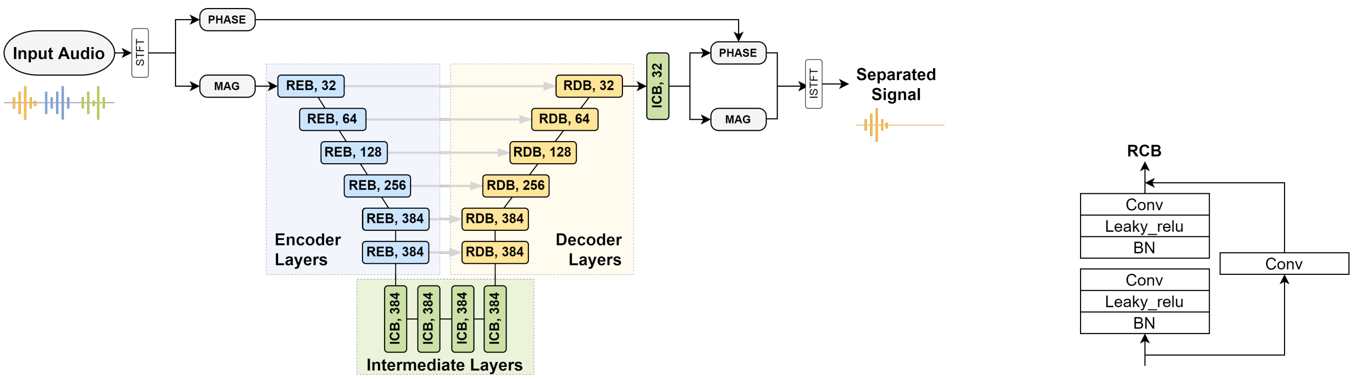

Figure 3: Overall architecture of the source separation network and its RCB.

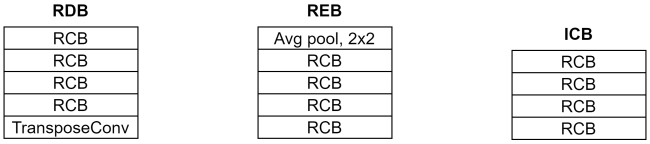

Figure 4: Detailed structures of REB, RDB, and RCB.

3.3. Source separation

To properly separate speech components from mixtures, the source separation network based on 143-layered residual UNet (ResUNet) architecture [18] is used. The model efficiently decoupled magnitude and phase estimation for more successful source separation, having the state-of-the-art music source separation (MSS) result on the MUSDB18 database.

Figure 3 shows an overview of the source separation network and the smallest block used in the architecture, residual convolutional block (RCB). Throughout the system, input audio is divided into phase and magnitude features and combined together after the ResUNet structure, allowing better estimation of complex ideal ratio masks (cIRMs). The magnitude part passes through the ResUNet, which is a UNet-like architecture where their features are contracted in the encoder part and expanded in the decoder part, connected by the skip connection. In Fig. 3, the residual encoder block (REB) works as a feature encoder that derives resolution of the magnitude spectra to be contracted, and the residual decoder block (RDB) works as a feature decoder, while intermediate convolutional blocks (ICBs) exist to maximize representation ability of the architecture. The REB, RDB, and ICB are described in Fig. 4 while the RCB are represented in Fig. 3. The results of the ResUNet, magnitude and phase features are combined to estimate a cIRM.

Since a speech component can be separated from each channel by the source separation network, we exploit four-channel separated outputs for better estimation of DOA.

3.4. DCMA for SELD

After source separation, we tried to apply these features to SED and DOA estimation (DOAE). Figure 5 displays the overall architecture of the SELD network. There are two streams for SED and DOAE, and the only difference in the two streams is that the SED stream uses only log-mel spectrogram, while the DOAE stream utilizes intensity vector as well. Therefore, the overall input of the SED stream would be eight channels, composed with four channels from original FOA inputs and four channels from source separation results, while the input of the DOAE stream would be 14 channels, including four channels of log-mel spectrogram and three channels of intensity vectors from each.

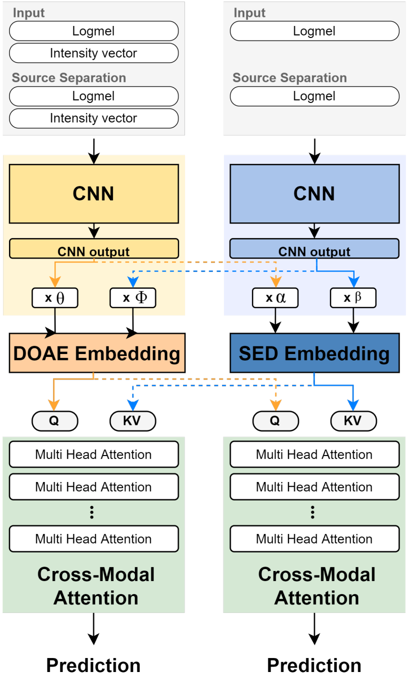

Figure 5: SELD network to predict sound event classes and estimates of DOA.

The CNN-based encoder with soft parameter sharing exchanges intermediate features in the CNN layers for the SED and DOAE, while key and value vectors for either SED or DOAE stream of the DCMA in the decoder part were given from the other stream to efficiently learn the association between SED and DOAE features [19].

Table 1: Experiment results on validation data. “ER”, “LE”, and “LR” represent the error rate, localization error, and localization recall, respectively.

Num	Track	Input ch	Add. data	ER(20°)	F-score(20°, %)	LE (°)	LR (%)
0	Baseline	Baseline	w/o add. data	0.71	21.0	29.3	46.0
1	5 tracks	7/4ch	w/o add. data	0.664	41.9	26.712	74.7
2	3 tracks	7/4ch	w/o add. data	0.652	43.5	26.931	74.9
3	3 tracks	8/4ch	with add. data	0.643	45.5	25.513	77.3
4	3 tracks	14/8ch	with add. data	**0.615**	**46.2**	**23.973**	**78.2**

3.5. Post-process

After outliers were obtained from the distribution of the estimates of DOA, we post-process prediction results by replacing the outlying estimates with the mean value.

4. EXPERIMENTS

Through experiments on validation data, we evaluated the performance according to 1) the numbers of tracks, 2) inputs with and without source separation results, 3) data synthesis with and without additional datasets. Table 1 summarizes the results. The “Track” column represents the number of tracks predicted at a frame. In the “Input ch” column, ”7/4ch” denotes four channels of FOA data for the SED stream, and three additional channels of intensity vector for the DOAE stream. ”8/4ch” represents 7/4ch added by a single channel output from the source separation network, while ”14/8ch” means that all four channels from the source separation network were used as additional four channels for the SED stream and additional seven channels (four of log-mel spectrograms and three of intensity vectors) for the DOA stream. The “Add. data” column indicates whether data from Audioset and FSD-50K datasets was additionally used for training the SELD network.

While up to five predictions (corresponding to the maximum overlapped events) were obtained at each frame in experiment 1, we also tried up to three predictions in experiment 2. Since experiment 1 provided too much predictions, experiment 2 resulted in improved performance.

As shown in experiment 3, adding a single channel output from the source separation network provided better results than those of experiment 2. Moreover, all four channels from the source separation network were exploited to provide additional four channels for the SED stream and additional seven channels for the DOA stream in experiment 4, which achieved the best performance.

5. CONCLUSION

In this paper, we presented an SELD network based on DCMA and soft parameter sharing of SED and DOAE streams incorporating source separation to simultaneously detect and localize sound events in real spatial scene. In addition to the basic model based on DCMA and soft parameter sharing, separated speech components from the source separation network and the synthesis logic to simulate natural real data improved the performance in the SELD task for the real spatial scene.

For a future work, we will study on better pre-processing and post-processing techniques to improve the performance. Also, our separation model can be improved to better extract respective features of overlapped events.

6. REFERENCES

[1] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, jun 2017. [Online]. Available: https://doi.org/10.1109%2Ftaslp.2017.2690575

[2] P. Stoica and K. Sharman, “Maximum likelihood methods for direction-of-arrival estimation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1132–1143, 1990.

[3] Z.-M. Liu, C. Zhang, and P. S. Yu, “Direction-of-arrival estimation based on deep neural networks with robustness to array imperfections,” *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, pp. 7315–7327, 2018.

[4] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, “An improved event-independent network for polyphonic sound event localization and detection,” 2020. [Online]. Available: https://arxiv.org/abs/2010.13092

[5] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, “Ensemble of accdoa- and einv2-based systems with d3nets and impulse response simulation for sound event localization and detection,” DCASE2021 Challenge, Tech. Rep., November 2021.

[6] T. N. T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W. S. Gan, “Dcase 2021 task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection,” DCASE2021 Challenge, Tech. Rep., November 2021.

[7] P. Emmanuel, N. Parrish, and M. Horton, “Multi-scale network for sound event localization and detection,” DCASE2021 Challenge, Tech. Rep., November 2021.

[8] S.-H. Lee, J.-W. Hwang, S.-B. Seo, and H.-M. Park, “Sound event localization and detection using cross-modal attention and parameter sharing for dcase2021 challenge,” DCASE2021 Challenge, Tech. Rep., November 2021.

[9] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, “Improving sound event detection in domestic environments using sound separation,” 2020. [Online]. Available: https://arxiv.org/abs/2007.03932

[10] 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," 2022. [Online]. Available: https://arxiv.org/abs/2206.01948

[11] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[12] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[13] K. Park and T. Mulc, "Css10: A collection of single speaker speech datasets for 10 languages," *arXiv preprint arXiv:1903.11269*, 2019.

[14] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[15] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[16] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using foa domain spatial augmentation," in *Proc. of the 4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.

[17] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[18] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation," 2021. [Online]. Available: https://arxiv.org/abs/2109.05418

[19] S.-H. Lee, J.-W. Hwang, M.-H. Song, and H.-M. Park, "A method based on dual cross-modal attention and parameter sharing for polyphonic sound event localization and detection," *Applied Sciences*, vol. 12, no. 10, p. 5075, 2022.