Sound event localization and detection utilizing overlapping end-to-end learning

Yanke Long[1] Riku Yasuda[2] Yui Sudo[3] Katsutoshi Itoyama[1,3] Kazuhiro Nakadai[1]
Kenji Nishida[1] Hideharu Amano[2]

Abstract: This paper presents efficient end-to-end deep learning for sound event localization and detection by sharing a part of the models called overlapping end-to-end learning, which can be trained with a small amount of data compared to normal end-to-end learning. We demonstrate its superior accuracy compared to traditional cascade integration, achieving a 3.3-point increase in classifying each of the mixed sound sources.

Keywords: sound source localization and detection, end-to-end learning, sound source localization, sound source separation, sound source classification, gated architecture

1. Introduction

In recent years, sound source classification (SSC) has been studied extensively, for example, in DCASE [*1]. Noise robust SSC is necessary for robots, and a common approach to solve it is to combine them with sound source localization (SSL), and sound source separation (SSS) [1]. However, this combination is usually done in a cascade manner, in which the errors in each process accumulate and the total performance is degraded. End-to-end training including SSL, SSS, and SSC can solve this cascade problem [2–4]. However, it is time-consuming to train such a large end-to-end network. This paper addresses these problems based on soft integration using a gated network, which requires a small amount of re-training.

2. Related work

Detection and localization of sound events have been investigated as sound event localization and detection (SELD) [5]. Generally, a deep learning-based approach is adopted for SELD.

In computational auditory scene analysis (CASA), traditionally, sound source localization and sound source separation have been studied separately using microphone array processing [2, 3]. SSL estimates the direction of arrival (DOA) of a sound source from the amplitude and phase difference of the signals from multiple microphones, and SSS separates sound sources arriving from different directions [6]. In their applications, generally SSL and SSS have been integrated in a cascaded manner, and thus, the cumulative errors from the functional blocks degrade the entire system performance.

End-to-end approaches, as opposed to cascade, have been studied [7]. Since it combines these functions as a single system and optimizes the entire system, the problem caused by cumulative errors is relaxed. Compared to the cascade approach, however, end-to-end approaches require huge amounts of data and longer training time.

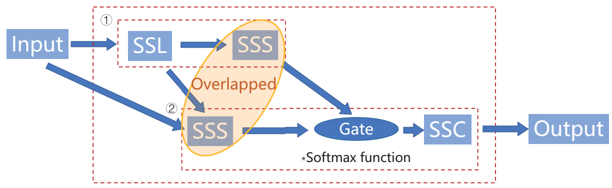

Fig. 1 Structure of the proposed method
SSL: Sound source localization SSS: Sound source separation
SSC: Sound source classification Gate: Combine the outputs

There are the problems with the above approaches:

(1) Cascade leads to large cumulative errors.
(2) A full end-to-end approach requires large datasets.

3. Proposed method

Fig. 1 shows the overall model structure of the proposed method called overlapping end-to-end learning. The three functional blocks of SSL, SSS, and SSC are divided into SSL+SSS and SSS+SSC, respectively, and these two blocks are integrated to overlap in their SSS parts. Practically, the SSL+SSS and SSS+SSC models are combined, but the outputs of the two SSSs are calculated through the Softmax function when sending the outputs of the two SSSs to the SSC block.

The mean square error (MSE) between network outputs and labels was used as a loss function during training. The ADAM optimizer was selected as the optimization function, and the number of epochs was 50 with a learning rate of 0.001 [8]. The single models, *i.e.,* SSL, SSS, and SSC, were trained separately. At this stage, they do not use the output of other models, and they all use the data calculated from the dataset. After each single model was trained, the corresponding parts of the SSL+SSS and SSS+SSC models were replaced with these parameters, and then re-training was performed. The re-trained parameters of SSL+SSS and SSS+SSC were inserted into the all-integrated model, and the model was again re-trained for the final model.

1 Tokyo Institute of Technology
2 Keio University
3 Honda Research Institute Japan Co., Ltd.
*1 https://dcase.community/

ⓒ 2023 Information Processing Society of Japan

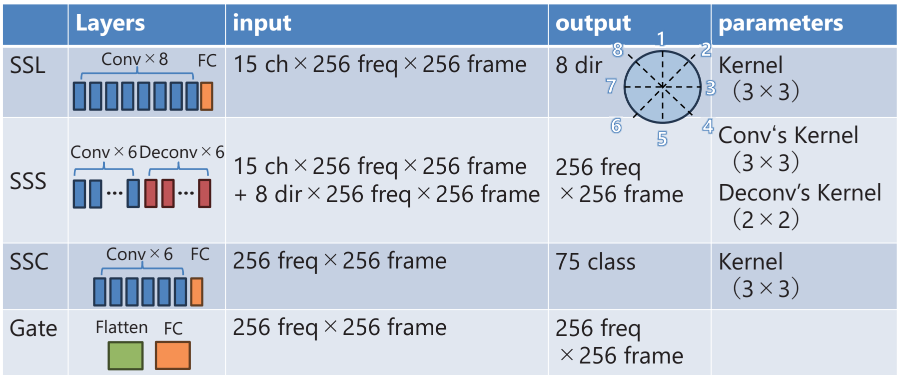

Fig. 2 Structure and parameters of the entire network
Conv: Convolutional layer Deconv: Deconvolutional layer
FC: Fully-connected layer Flatten: Flattened layer

Table 1 SELD results of each model

model	Divided E2E[1]	Ovelapped[2]	results
(SSL)+(SSS)+(SSC)	No	No	69.7%
(SSL+SSS)+SSC	No	No	71.3%
SSL+(SSS+SSC)	No	No	71.0%
(SSL+SSS)+(SSS+SSC)[3]	Yes	Yes	73.0%

[1] Divide the end-to-end network into two small networks.
[2] Let the two networks have the overlapped part.
[3] The proposed method.

4. Evaluations

4.1 Dataset and metrics

For networks that handle SSL, SSS, and SSC, it is necessary to have a pair of mixed sound sources and separated sound sources with known DOA and classes [9, 10]. We used a dataset with 75 classes of single sound sources. Each sample was trimmed to 4.192 seconds long. The signal-to-noise ratio was set to be 15dB on average by adding noise sources recorded in restaurants and halls.

For the metrics, the proposed method infers the class of the single source from the input mixed sound sources. Since there are 1,000 validation data, n/1000 (n: the number of cases where the inferred result matches the class of the label) is used as the index of the number of correct answers.

4.2 Results and Discussion

Table 1 shows the results of SELD. The results show that the accuracy of SELD has improved by the proposed method, which means it can reduce accumulative error. The results in the table 1 show that the results improve in the following order: cascade method, one single model and the integrated model, and the all-integrated model. This may be due to the fact that the number of models used in the inference is reduced so that individual errors do not affect the overall results.

From the above results, we can know that:

(1) The proposed model was well-trained with a small dataset including 10,000 samples.

(2) The proposed model successfully reduced cascading errors with the gated architecture.

5. Conclusions

In this paper, we investigate, implement, and evaluate efficient end-to-end deep learning for sound event localization and detection by sharing a part of the models, and confirm that the learning method that overlaps the SSS task among the SELD tasks (SSL, SSS, and SSC) improves the accuracy of SELD. This implementation was compared to the conventional cascade method for SELD, achieving a 3.3-point increase in the classification of mixed sound sources. Since the effectiveness of the proposed method was demonstrated by comparing it with the cascade method, it is expected to become more practical by applying it to more complex datasets and optimizing the learning in the model for each task. Furthermore, although only the SSS task was overlapped in this study, there is space for improvement in accuracy by examining various overlapping methods in the future.

Acknowledgments This work was supported by JST, CREST Grant No. JPMJCR19K1, Japan.

References

[1] Kazuhiro Nakadai, Hiroshi G. Okuno, and Takeshi Mizumoto. Development, deployment and applications of robot audition open source software HARK. *Journal of Robotics and Mechatronics*, 29:16–25, 02 2017.

[2] Kazuhiro Nakadai, Gökhan Ince, Keisuke Nakamura, and Hirofumi Nakajima. Robot audition for dynamic environments. In *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, pages 125–130, 2012.

[3] D. Gabriel, Ryosuke Kojima, K. Hoshiba, K. Itoyama, Kenji Nishida, and Kazuhiro Nakadai. 2D sound source position estimation using microphone arrays and its application to a VR-based bird song analysis system. *Advanced Robotics*, 33:1–12, 03 2019.

[4] Yui Sudo, Katsutoshi Itoyama, Kenji Nishida, and Kazuhiro Nakadai. Sound event aware environmental sound segmentation with Mask U-Net. *Advanced Robotics*, 34, 10 2020.

[5] Yui Sudo, Katsutoshi Itoyama, Kenji Nishida, and Kazuhiro Nakadai. Improvement of DOA estimation by using quaternion output in sound event localization and detection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pages 244–247, 01 2019.

[6] Yui Sudo, Katsutoshi Itoyama, Kenji Nishida, and Kazuhiro Nakadai. Environmental sound segmentation utilizing Mask U-Net. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5340–5345, 2019.

[7] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, pages 334–340, 06 2018.

[8] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[9] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. A multi-room reverberant dataset for sound event localization and detection, 2019.

[10] Scott Wisdom, Hakan Erdogan, Daniel Ellis, Romain Serizel, and Nicolas Turpault. What's all the FUSS about free universal sound separation data?, 2020.

© 2023 Information Processing Society of Japan