TRACK-WISE ENSEMBLE OF CRNN MODELS WITH MULTI-TASK ADPIT FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Sang-Ick Kang , Myungchul Keum, Kyongil Cho, Yeonseok Park

KT Corporation, South Korea
{sangick.kang, mc.keum, cho.kyongil, yeonseok.park}@kt.com

ABSTRACT

This report describes our systems submitted to the DCASE2022 challenge task 3: Sound Event Localization and Detection (SELD) with directional interference. Locating and detecting sound events consists of two subtasks: detecting sound events and estimating the direction of arrival simultaneously. Therefore, it is often difficult to jointly optimize these two subtasks at the same time.

We propose track-wise ensemble model which is combined with a multi-task-based auxiliary duplicating permutation invariant training (ADPIT) model and multi-ACCDOA-based model. Specifically, we propose a novel method to ensemble CRNN multi-task models, an event independent network v2 (EINV2)-based multi-task models and CRNN multi-ACCDOA models.

Experimental results on the DCASE2022 dataset for sound event localization and directional interference detection show that the deep learning-based model trained on this new function significantly outperforms the DCASE challenge baseline.

Index Terms— DCASE2022, Sound Event Localization and Detection, Model ensemble

1. INTRODUCTION

The Sound Event Localization and Detection (SELD) is to detect sound events belonging to specific target classes, track their temporal activity, and estimate their directions-of-arrival (DOA) or positions during it. The SELD system provides important operations for autonomous vehicles or robots with multi-channel audio input receivers. For example, self-driving cars can distinguish between car horns and pedestrian steps on the road, and indoor mobile robots can estimate the domestic sounds such as vacuum cleaners, mechanical fans, and door sounds in a house situation. In particular, the serving robot has many customers in the restaurant and many sound events occur, so the solution to analyze multi-channel audio data and estimate the corresponding source in a specific direction is reasonable.

Neural network based SELD methods can be classified into class-wise output formats and track-wise formats [1-10]. The track-wise format allows the model to detect the same event class in multiple locations, whereas the track-wise format detects only one event class and its location on each track. On the other hand, the class-wise format still has problems detecting overlapping events in the same class because it only assigns one location to each event class. To overcome these problems, a multiple ACCDOA method using auxiliary duplicating permutation invariant training (ADPIT) has been proposed [11].

In this study, we perform a model ensemble of multiple systems trained with different conditions and model architectures. A spatial cue-augmented log-spectrogram (SALSA)-lite-based system and an event independent network v2 (EINV2)-based multi-task system and multi-ACCDOA-based model are applied [8, 11, 12]. Each system derives SED and DOA results by track and applying ADPIT. Using the track-wise ensemble method, the output of each single learned system is combined to make a final judgment after learning. To increase the training data, we perform data augmentation such as SpecAugment [13], cutout, frequency shift and rotation of the signal [14]. Experiments on the development dataset showed that our system improved significantly over the baseline system.

2. DATASETS

The DCASE2022 Task3 provides STARS22 Development dataset and DCASE2022 SELD Synthetic dataset [15]. The datasets contains 13 target sound event classes. The STARS22 dataset was actually recorded and manually annotated by real sound scenes. Occurrences of up to 3 simultaneous events are quite usual and up to 5 overlapping events can sometimes happen. The DCASE2022 SELD Synthetic dataset is generated through convolution of isolated sound samples with real spatial room impulse responses (SRIRs) captured in various spaces of Tampere University. Sound samples for the target classes were sourced from the FSD50K [16]. Maximum polyphony of sound samples is 2 track with possible same-class events overlapping.

In additional, we have generated more complex dataset. The synthetic method is basically equal with DCASE2022 SELD Synthetic dataset. However, synthesized sound samples were joined from the AudioSet dataset [17]. The samples were manually selected based on their labels having only one of the classes of interest. The generated dataset consists of 1200 audio files is 1-minute long mixed with the sound events. A maximum polyphony is 5 events without directional interference and does not overlap more than 3 for the same target classes.

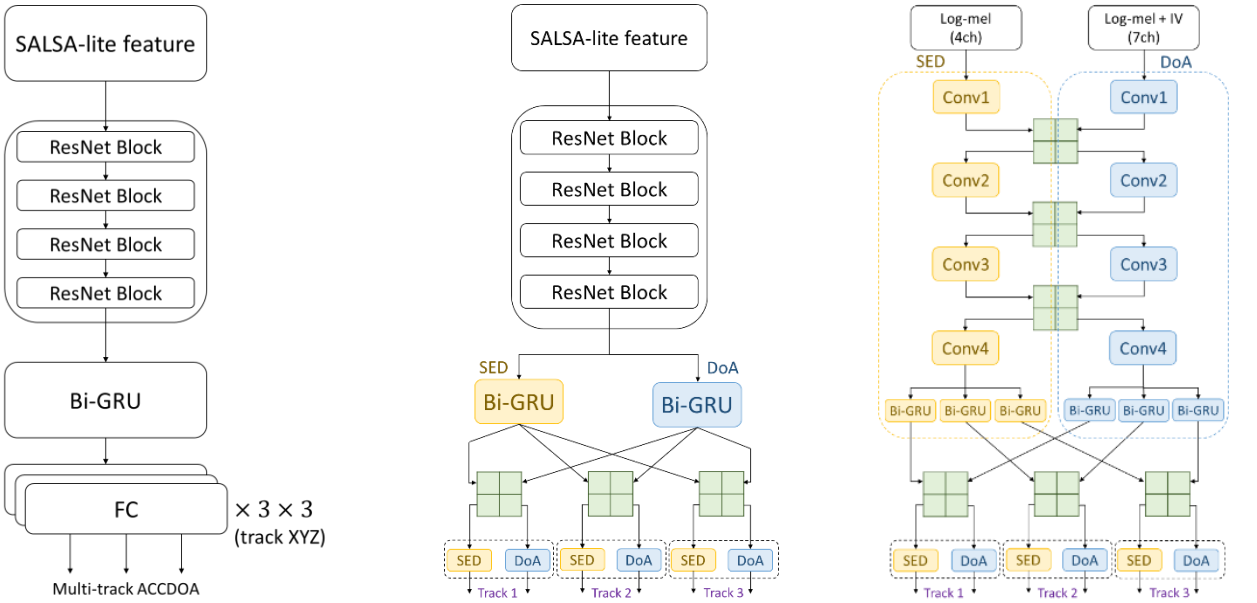

Figure 1. Illustration of three different single SELD architectures.
(a) CRNN multi-ACCDOA network, (b) CRNN multi-task network, (c)EINV2-based multi-task network

3. PROPOSED APPROACH

3.1 Multi-task ADPIT

A PIT framework with a replication target is called ADPIT. Applying ADPIT allows the model to maintain performance in the absence of duplicates of the same class by duplicating the original target with secondary targets on different tracks instead of zero vectors to ensure that all tracks are trained with a single target and the same target [11].
We extended ADPIT to handle multi-task problems. By using Multi-task ADPIT, we can deal multi-track SED and DOA as separated task.

3.2 Features

In this study, two types of features are used for ensemble models. SALSA-Lite feature from MIC is used as the first set of features. Log-mel spectrograms and intensity vector (IV) in log-mel space from FOA are used as the second set of features. We extract features for both MIC and FOA as input features. Log-mel spectrograms are first used for SED, while IV in log-mel space is used for DOA estimation in EINV2-based multi-task system. SALSA-Lite feature is composed of two major components: linear-frequency log-power spectrogram and frequency-normalized inter-channel phase difference.

3.3 Network architecture

In this study, to increase the diversity of the model ensemble, we consider variants of multi-task architecture and multi-ACCDOA based architecture. Those architectures are illustrated in Fig. 1 (a)–(c) and described in Table 1.

The first and second variant of multi-task method is inspired by the SALSA-Lite architecture [12]. For Multi-ACCDOA model in Table 1 (a), we increased number of decoder FC block from 4 to 9, to get 3 track ACCDOA output. For multi-task model in Table 1 (b), soft parameter-sharing using cross-stitch and track-wise decoder are incorporated.

As variant of EINV2 network in Table 1 (c) [8], we replaced the MHSA block with a bidirectional GRU block. As shown in Fig. 1 (c), EINV2-based multi-task network consists of two parts, which log-mel spectrograms features with and without IV are fed respectively.

Table 1. Description of three different single SELD architectures.

System	Methods	Input	Features	Output
(a)	CRNN Multi-ACCDOA	Mic	SALSA-Lite	Multi-ACCDOA
(b)	CRNN multi-task	Mic	SALSA-Lite	Multi-task (SED, DOA)
(c)	EINV2-based	FOA	Log-mel + IV	Multi-task (SED, DOA)

3.4 Model Ensemble

For the track-wise output format, predictions of sound events can be randomly assigned to tracks. Averaging or weighted ensembles cannot predict across different tracks, so these methods cannot be applied to a track-wise output format. To solve these problem, track-wise ensemble model has been proposed [18].

The ensemble model architecture is shown in Fig. 2. The input to an ensemble model is the output of a single, distinct model.

It has same structure with CRNN multi-task network decoder, but it accepts multiple SELD output as input. It handles SED and DOA separately. To improve per track result, soft parameter-sharing is used. The ensemble model predicts the outcome in the form of a track-wise output.

4. EXPRERIMENTAL RESULTS

4.1 Experimental settings

The sampling frequency is set to 24 kHz. The STFT is used with a 20 ms frame length and 10 ms frame hop. 128 mel bins for log-mel spectrograms and IV features. The frame length of input to the networks is 1,000 frames. We use a batch size of 128. Each training sample is generated on-the-fly. We gradually increase the learning rate to 0.0001 with 20,000 iterations. After the warmup, the learning rate is decreased by 10% if the SELD score of the validation do not improve in 40,000 consecutive iterations. We use the AdamW optimizer with a weight decay of 10^{-6}.

Table 2. Ensemble configuration.

Ens.	Base system
Ensemble #1	CRNN multi-ACCODA CRNN multi-task × 2 EINV2-based multi-task × 3
Ensemble #2	CRNN multi-task × 2 EINV2-based multi-task × 2
Ensemble #3	CRNN multi-ACCDOA CRNN multi-task model × 2 EINV2-based multi-task × 5
Ensemble #4	CRNN multi-task model × 2 EINV2-based multi-task × 4

4.2 Experimental Results

A single model was trained based on 3 type of SELD architectures, and the results of single model were used as input for the ensemble model. The models used for the ensemble are listed in Table 2.

Table 3 shows the performance for single SELD network and model ensemble from them. The different models are compared for performance on the evaluation dataset.

From the experimental results, in the case of a single model, EINV2 has the best overall performance. In the case of CRNN-based multi-task model, ER_{20° showed higher performance than other single models. The results of ensemble of heterogeneous models with different characteristics showed lower compared to single EINV2-based multi task model, but improved performance in other metrics. Among the four ensemble models, the ensemble #1 model combined with CRNN multi-task model, EINV2-based multi-task model and multi-ACCDOA-based model showed the best overall performance.

Table 3. SELD performance of our systems.

Methods	ER_{20°	F_{20°(macro)	LE_{CD}	LR_{CD}
Baseline (FOA)	0.71	21%	29.3°	46%
Baseline (MIC)	0.71	18%	32.2°	47%
CRNN multi-ACCDOA	**0.53**	47%	17.5°	61%
CRNN multi-task	0.55	46%	18.0°	67%
EINV2-based multi-task	0.62	**52%**	**17.0°**	**70%**
Ensemble #1	0.49	**53%**	**15.8°**	**68%**
Ensemble #2	**0.48**	51%	16.4°	67%
Ensemble #3	0.49	52%	**15.8°**	66%
Ensemble #4	**0.48**	52%	16.3°	65%

5. CONCLUSION

Our approach present DCASE2022 task 3: sound event localization and detection (SELD) with directional interference. In this report, we proposed track-wise ensemble method to combine outputs of each single learned system. The single learned system consisted of CRNN multi-task network using MIC data and EINV2-based multi-task network using FOA data. Moreover Multi-track ADPIT was applied to derive sound event detection (SED) and DOA estimation. To improve the network model, we further use generated multi-channel signals by convolving spatial room impulse responses (SRIRs) with source signals manually extracted from the sound sample database. Experiments show that the proposed network achieves improved performance when compared to the baseline model for the DCASE2022 challenge task3.

6. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE JSTSP*, vol. 13, no. 1, pp. 34–48, 2018.

[2] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. DCASE Workshop*, 2020.

[3] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *arXiv:2101.02919*, 2021.

[4] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. DCASE Workshop*, 2019.

[5] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *Proc. IEEE ICASSP*, 2021, pp. 915–919.

[6] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant

sound scenes with directional interferers for sound event localization and detection," *arXiv:2106.06999*, 2021.

[7] P. Emmanuel, N. Parrish, and M. Horton, "Multi-scale network for sound event localization and detection," in *Tech. report of DCASE Challenge*, 2021.

[8] 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," in *Proc. IEEE ICASSP*, 2021, pp. 885–889.

[9] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, "A sequence matching network for polyphonic sound event localization and detection," in *Proc. IEEE ICASSP*, 2020, pp. 71–75.

[10] Y. He, N. Trigoni, and A. Markham, "SoundDet: polyphonic moving sound event detection and localization from raw waveform," in *Proc. ICML*, 2021.

[11] 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 ICASSP*, 2022, pp. 316-320.

[12] T. N. Tho Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan and W. -S. Gan, "SALSA-Lite: A Fast and Effective Feature for Polyphonic Sound Event Localization and Detection with Microphone Arrays," in *Proc. IEEE ICASSP*, 2022, pp. 716-720.

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

[14] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in Proc. DCASE 2019 Workshop, 2019, pp. 154–158.

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

[16] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50k: an open dataset of human-labeled sound events," *arXiv:2010.00475*, 2020.

[17] J. F. Gemmeke, D. PW 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 *Proc. IEEE ICASSP*, 2017, pp. 776–780.

[18] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, J. Yang, "A Track-Wise Ensemble Event Independent Network for Polyphonic Sound Event Localization and Detection," in *Proc. IEEE ICASSP*, 2022, pp. 9196-9200.