DATA AUGMENTATION AND PRIOR KNOWLEDGE-BASED REGULARIZATION FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Jingyang Zhang[1,*], *Wenhao Ding*[1,*], *Liang He*[1]

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
jingyang15@mails.tsinghua.edu.cn, dingwenhao95@gmail.com, heliang@mail.tsinghua.edu.cn

ABSTRACT

The goal of sound event localization and detection (SELD) is detecting the presence of polyphonic sound events and identifying the sources of those events at the same time. In this paper, we propose an entire pipeline, which contains data augmentation, network prediction and post-processing stage, to deal with the SELD task. In data augmentation part, we expand the official dataset with SpecAugment [1]. In network prediction part, we train the event detection network and the localization network separately, and utilize the prediction of events to output localization prediction for active frames. In post-processing part, we propose a prior knowledge-based regularization (PKR), which calculates the average value of the localization prediction of each event segment and replace the prediction of this event with this average value. We theoretically prove that this technique could reduce mean square error (MSE). After evaluating our system on DCASE 2019 Challenge Task 3 Development Dataset, we approximately achieve a 59% reduction in SED error rate (ER) and a 13% reduction in directions-of-arrival (DOA) error over the baseline system (on Ambisonic dataset).

***Index Terms*—** Sound event localization and detection, data augmentation, convolution recurrent neural network

1. INTRODUCTION

Since 2016, the community of Detection and Classification of Acoustic Scenes and Events (DCASE) has been focusing on the detection of sound event and holds sound event detection (SED) task every year. In DCASE 2019, this task is extended to a more complex one that requires the participant to predict both the position of the event and the direction of the event source at the same time [2]. Although these two subtasks has been studied for many years, it is claimed that jointly optimizing these two tasks benefits each other [3]. Estimating the SED and DOA separately will suffer the data association problem between the recognized sound events and the DOA [4]. However, after several experiments we come to a different conclusion that the goal and suitable features for SED and DOA are quite different, which means a separated system will achieve better performance.

Solving SELD problems has great impacts on human's daily life. For example, audio surveillance can be achieved in smart cities or smart homes, the automatic speech recognition can also be enhanced with the information about the direction of the source, and the speaker diarization system performance in meeting room will be greatly improved.

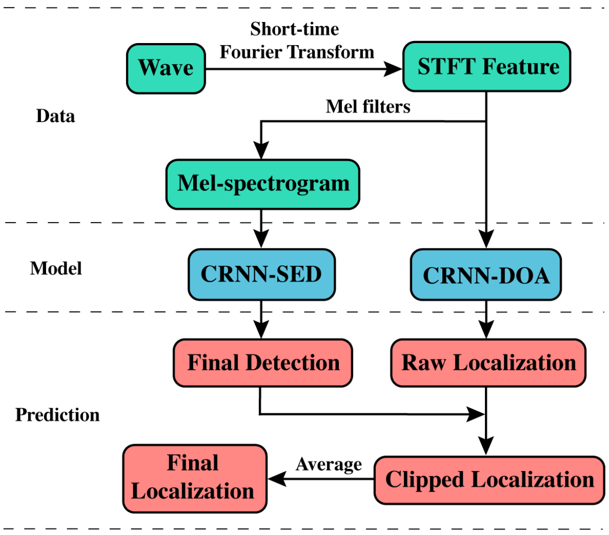

Figure 1: Scheme of our proposed pipeline. Three different colors represent three parts. The green part is about pre-processing the data and extracting raw acoustic features. The blue part is our neural networks for two sub-tasks. The red part is the post-processing stage where we utilize SED detection to generate active localization predictions and perform our PKR technique.

In recent years, the combination of convolution neural network (CNN) and recurrent neural network (RNN), which is called convolution recurrent neural network (CRNN), has displayed great success in processing audio signals. Its effectiveness mainly comes from the capability of local feature extraction of CNN and the capability of processing time sequence of RNN.

Data augmentation is quite widely used when the training data is rare or unbalanced. For audio or acoustic data, there are many methods to augment the dataset like adding white noise, pitch shifting, and time stretching. Recently, there is a simple but effective method called SpecAugment [1]. The idea is to randomly drop several consecutive frames or frequency bins of the Mel-spectrogram, and the result is claimed to be better than other methods.

2. PROPOSED METHOD

In this section, we first give an overview of our system, and then analyze three parts of our entire pipeline: the data augmentation,

[*] Jingyang Zhang and Wenhao Ding are joint first author.

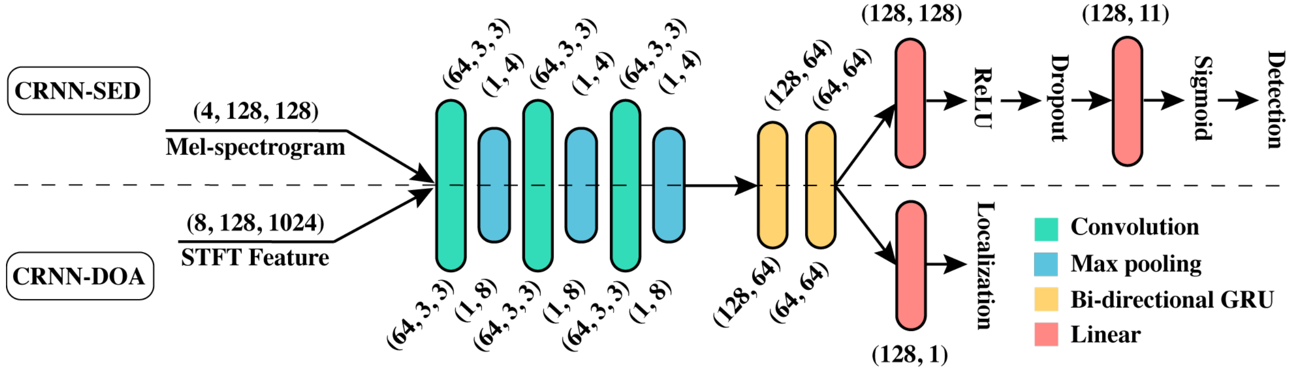

Figure 2: The structure of our CRNN network. The top part describes the details of our CRNN-SED network, including the dimension of network parameters, while the bottom part is our CRNN-DOA network. Different modules are displayed in different colors and the annotation is on the right-bottom of this figure. Some batch normalization and dropout layers are not shown for brevity. Annotations for dimensions represent (C, T, F) or (T, F), where C and T denotes the number of channels and temporal length, respectively, and F represents frequential length or the number of output nodes.

the CRNN structure, and the post-processing method.

2.1. System Overview

The overview of our system is shown in Fig. 1, in which the green part is about raw feature extraction, the blue part consists of two separate CRNN models for SED and DOA, and the red part shows the post-processing of the predictions.

The system can be divided into two threads, since we use two different kinds of features for two tasks. We use the magnitude and phase information of short-time Fourier transform (STFT) feature to deal with the DOA task, and the Mel-spectrogram for SED task. The reason is twofold: (1) SED task requires more information about the events and the detection is mainly based on magnitude, and Mel filters can provide features similar to human's hearing. (2) DOA task requires the direction and angle information that can be directly extracted from the phase of STFT.

In our experiments, we have tried the joint training of SED and DOA with either STFT or Mel-spectrogram. We also tried using STFT to train SED task and Mel-spectrogram to train DOA task. However, all above attempts do not outperform our proposed training method.

2.2. Data Augmentation

The official dataset provided by DCASE 2019 is actually not sufficient since it only consists four isolated sound events for each split. Thus, it is very straightforward and necessary to use data augmentation for this dataset. We finally adopted the SpecAugment [1] in our pipeline because it has been proved to be very efficient when dealing with Mel-spectrogram in automatic speech recognition (ASR). Before the training stage, we generate three types of data that is shown in Fig. 3. The first and second type randomly remove several rows (frames) or columns (frequency bins), respectively, and the third type removes several rows and columns at the same time.

Since the data has 4 channels, we conduct this augmentation for each channel of every sample. The original dataset has 4 splits to build the 4-fold cross-validation, thus we generate **split**$\{$**5, 6, 7, 8**$\}$ using type 1 method, **split**$\{$**9, 10, 11, 12**$\}$ using type 2 method, and **split**$\{$**13, 14, 15, 16**$\}$ using type 3 method. There are 200 samples in each generated split, which means there are two augmented samples for every original sample with one type of augmentation.

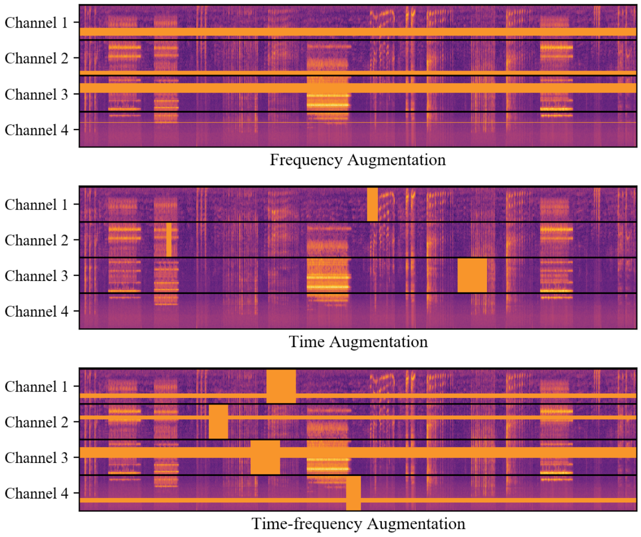

Figure 3: One example of SpecAugment augmentation method on Mel-spectrogram. Each row represents one type of augmentation method: randomly removing frequency bins, removing frames, and removing both of them at the same time.

2.3. CRNN Structure

We use CRNN structure as our basic model to deal with both SED and DOA, and the structure is shown in Fig. 2. Actually, we use the same structure for both tasks despite their differences between the linear layers. In CRNN-SED model, we use two linear layers and use Sigmoid as the last activation function, while in CRNN-DOA model, we only use one linear layer without any activation functions since it is a regression model and introducing activation like ReLU or tanh actually degrades the performance according to

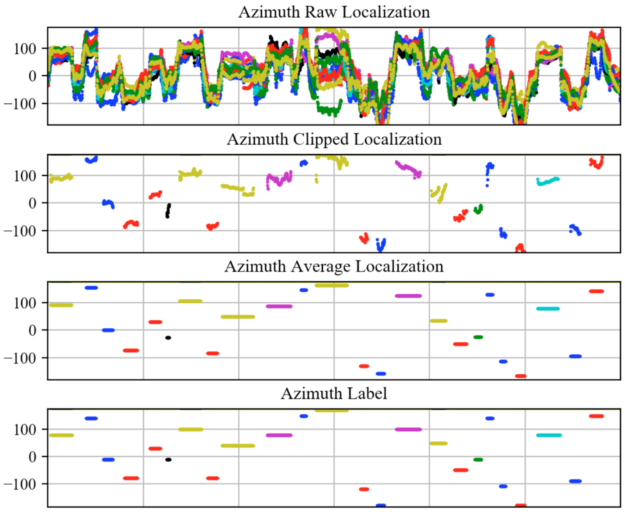

Figure 4: One example of PKR on localization output (Azimuth). The first row shows the raw prediction of DOA model, the second row shows the active frames collected by SED output (not SED label), the third row presents the prediction after PKR, and the last row is the DOA label.

our experiments. It is also worthy to note that we tried to take DOA as a classification task but didn't get satisfying results, which might be due to the severe imbalance between the number of positive and negative nodes at the output layer.

The dimension of all layers is also shown in Fig. 2 and the differences between SED and DOA model are all displayed. We use 1024-point STFT and 128 Mel filters, thus these two models will take inputs of different sizes. As for the loss function, we use binary cross entropy (BCE) for SED task and mean square error (MSE) for DOA task. One thing to note is that we only use active frames and discard inactive frames when calculating the loss of CRNN-DOA, which is implemented by using the label of SED as a mask. We make this move because we suspect the DOA output of inactive frames may not provide useful information for the model, and we do observe a little performance improvement by adopting this strategy.

2.4. Post-processing Method

In post-processing stage, we first use the prediction of CRNN-SED to extract active frames of CRNN-DOA prediction to get the clipped localization. Then we use our prior knowledge-based regularization (PKR) which is to calculate the average value of every segment and use this average value as the final localization prediction. These predictions are all displayed in Fig. 4 with an example of the azimuth of **split1_ir0_ov1_3**.

The reason why we add an PKR to the DOA output is that the prediction of DOA is usually unstable as is shown in Fig. 4, especially in the beginning and the endding part. However, we actually have prior knowledge, which is also reasonable in real-world scenarios, that the position of the sound source is fixed, thus the output should be the same value within each segment. We can also prove that the averaged output will provide a more accurate prediction in terms of MSE loss. Taking azimuth for example, we define the label of one event segment as y, the clipped prediction of frame i as x_i, and the average value of all x_i as $\overline{x}$. Then the original MSE between the raw localization output and reference is:

$$MSE_{clip} = \sum_{i=1}^{N}(x_i - y)^2 = \sum_{i=1}^{N} {x_i}^2 + Ny^2 - 2y\sum_{i=1}^{N} x_i \quad (1)$$

where N notes the length of one event segment. The MSE after we use PKR can be expressed as:

$$\begin{aligned} MSE_{avg} &= N(\overline{x} - y)^2 \\ &= N\overline{x}^2 + Ny^2 - 2yN\overline{x} \\ &= \left(\sum_{i=1}^{N} x_i\right)^2 / N + Ny^2 - 2y\sum_{i=1}^{N} x_i \end{aligned} \quad (2)$$

Then, let $b_i = 1$ in Cauchy-Schwarz inequality:

$$\sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i^2 = N\sum_{i=1}^{N} a_i^2 \geq \left(\sum_{i=1}^{N} a_i\right)^2 = \left(\sum_{i=1}^{N} a_i b_i\right)^2 \quad (3)$$

Then we will get the relationship between MSE_{clip} and MSE_{avg}:

$$MSE(x)_{clip} \geq MSE(x)_{avg} \quad (4)$$

which proves that our PKR can indeed provide a better prediction with MSE metric. However, the metric used in validation and test stage is not MSE but the great-circle distance, resulting in a metric mismatch between the training and testing stage. This should be further studied to get more results.

Table 1: Important Hyper-parameters

Notation	Value	Description
lr	0.001	Learning Rate
B_{sed}	128	Batch size for SED
B_{doa}	64	Batch size for DOA
p	0.3	Dropout rate
τ	128	Length of one segment
τ_{step}	128	Step size between two segments
n_{fft}	2048	Number of the FFT
n_{mel}	128	Number of the Mel filters

2.5. Training details

In this part, we will provide some details during our training stage. Some important hyper-parameters are shown in Table 1. For the RNN part, we use a bi-directional GRU model that has two hidden layers and the dimension of hidden layer is 64. During the training stage, we input 128 frames into the network as one sample, while during the validation and test stages, we input the entire 3000 frames into our model to get the prediction for the whole audio clip.

3. EVALUATION

In this section, we will report experimental results of our proposed pipeline on development dataset, as well as some ablation experiments that show some of our explorations for this SELD task.

Table 2: Results on Development Dataset

Method	Error Rate (ER)		F1-score (%)		DOA Error (°)		Frame Recall (%)		SELD Score	
	Ambisonic	Mic	Ambisonic	Mic	Ambisonic	Mic	Ambisonic	Mic	Ambisonic	Mic
Baseline	0.34	0.35	79.9	80.0	28.5	30.8	85.4	84.0	0.21	0.22
Average	**0.14**	**0.15**	**91.6**	**91.3**	**24.8**	**25.8**	**90.8**	**89.5**	**0.11**	**0.12**
Split 1	0.11	0.11	93.8	93.4	23.8	25.9	93.4	92.6	0.09	0.10
Split 2	0.16	0.17	89.9	89.7	25.0	25.7	91.4	90.0	0.12	0.13
Split 3	0.10	0.12	94.6	93.2	24.8	25.5	93.5	91.9	0.09	0.10
Split 4	0.19	0.19	88.2	89.0	22.7	23.6	89.6	88.3	0.13	0.14

Table 3: Influence of Feature on SED and DOA (Split 1)

Task	Feature	ER	F1-score	DOA	Frame recall
SED	Mel	**0.18**	**89.0**	/	/
SED	STFT	0.27	84.1	/	/
DOA	Mel	/	/	89.1	90.8
DOA	STFT	/	/	**24.8**	**90.8**

Table 4: Influence of Augmentation Strategies on SED (Split 1)

Strategy	Number of files	ER	F1-score
None	100	0.18 ± 0.01	89.0 ± 0.9
F	300	0.12 ± 0.01	92.8 ± 0.9
T	300	0.15 ± 0.02	91.1 ± 1.3
F-T	300	0.12 ± 0.02	92.4 ± 1.3
F, T	500	0.12 ± 0.01	92.8 ± 0.4
F, T, F-T	**700**	**0.10 ± 0.01**	**93.8 ± 0.7**

* F represents frequency and T represents time.

3.1. Results on Development Dataset

The results on development dataset are shown in Table 2. Our model outperforms the baseline in both Ambisonic and Microphone array dataset in terms of all metrics. In order to give more details, we also show results for every split. One thing to note is that, we choose the best model according to the performance on validation files and only choose one model for each split without fusing multiple models.

3.2. Ablation Experiments

In order to provide more information about the performance of our system, we did some ablation experiments. These results can be found in Table 3 and Tabel 4.

We compare the performance of using different features for SED and DOA task. The results in Table 3 demonstrate that Mel-spectrum is much more beneficial for SED task than STFT. However, the phase information in STFT is quite crucial for DOA task, since the DOA error will dramatically increase when Mel-spectrum is used.

We adopt the aforementioned data augmentation strategy based on a pilot study whose results are presented in Table 4. Comparing with no expansion of the dataset, one can clearly observe an improvement in error rate when performing data augmentation. Meanwhile, increasing the number of augmented files and incorporating more types of augmentation can further reduce the error rate. The bottom row of Table 4 corresponds to the strategy that we use in this challenge.

While the SpecAugment [1] technique focuses on Mel-spectrogram, we also attempted to apply it to STFT features for DOA task. Unfortunately, this augmentation brings minor performance gain. Therefore, augmentation skills for DOA and STFT features need to be further studied in the future.

3.3. Submission

We applied the same system of the development dataset to the evaluation dataset. In the final submission, we fused the provided 4 splits and picked out 25 samples from each split to form a validation dataset, then we followed our pipeline to perform data augmentation, train the neural network, and post-process raw output. For SED task, we generated four different groups of augmented dataset (two for Ambisonic dataset and two for Microphone Array dataset) and selected four best model trained with them. For DOA task, since we did not use data augmentation, we only selected the best model from the original dataset.

Finally, we got 4 evaluation results with the combination of our 4 SED models and 1 DOA model and we named them **He_THU_task3_{1, 2, 3, 4}**, where the first two submissions are for Ambisonic dataset and the rest two submissions are for Microphone Array dataset.

4. CONCLUSION

In this paper, we propose a new system to tackle with SELD task. This system consists of three parts: SpecAugment for data augmentation, CRNN architecture for model prediction, and PKR for post-processing the output. Since the data is very insufficient in the challenge setting, we claim that the data augmentation is necessary to improve the entire system. In addition, we believe that the targets of SED task and DOA task are not consistent, and the input feature should also be different according to their different characteristics. This conclusion has been proved by our experiments on development dataset. Compared with DCASE 2019 Challenge Task 3 baseline, we have a 59% reduction in SED ER and a 13% reduction in DOA error (on Ambisonic dataset).

5. REFERENCES

[1] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.

[2] S. Adavanne, A. Politis, and T. Virtanen, “Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network,” in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[3] T. Hirvonen, “Classification of spatial audio location and content using convolutional neural networks,,” in *AES 138th Convention, At Warsaw, Poland*, 2015.

[4] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, “Two-source acoustic event detection and localization: Online implementation in a smart-room,” in *2011 19th European Signal Processing Conference*, Aug 2011, pp. 1317–1321.