THE USTC-IFLYTEK SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION OF DCASE2020 CHALLENGE

Technical Report

Qing Wang[1], *Huaxin Wu*[2], *Zijun Jing*[2], *Feng Ma*[2], *Yi Fang*[2],
Yuxuan Wang[1], *Tairan Chen*[1], *Jia Pan*[2], *Jun Du*[1], *Chin-Hui Lee*[3],

[1] University of Science and Technology of China, Hefei, China
{qingwang2, jundu}@ustc.edu.cn, {yxwang1, vea, panjia}@mail.ustc.edu.cn
[2] iFLYTEK, Hefei, China, {hxwu2, zjjing2, fengma, yifang2}@iflytek.com
[3] Georgia Institute of Technology, Atlanta, USA, {chl}@ece.gatech.edu

ABSTRACT

In this report, we present our method for DCASE 2020 challenge: Sound Event Localization and Detection (SELD). We propose an entire technical solution, which consists of data augmentation, network training, model ensemble, and post-processing. First, more training data is generated by applying transformation to both Ambisonic and microphone array signals, and by mixing the non-overlapping samples in the development dataset. And SpecAugment is also used as an augmentation technique to expand the training dataset. Then we train several deep neural network (DNN) architectures to jointly predict the spatial and temporal location of sound events in addition to its type. Besides, for SED estimation, we also use softmax activation function to handle the classification of both non-overlapping and overlapping sound events. With several network architectures, a more robust prediction of SED and directions-of-arrival (DOA) is obtained by model ensemble. At last, we use post-processing to apply different thresholds to different sound events. The proposed system is evaluated on the development set of TAU-NIGENS Spatial Sound Events 2020.

Index Terms— Sound event localization and detection, data augmentation, model ensemble, convolutional recurrent neural network

1. INTRODUCTION

The goal of SELD task is to detect the presence of sound events, and localize them in time and space when active. SELD can be applied in many areas [1]. Environmental types can be recognized and suppressed to improve speech quality during speech communication or to improve performance of robust automatic speech recognition (ASR). In smart cities, audio surveillance system plays an indispensable role.

The SELD task in DCASE 2019 is evaluated with individual metrics for SED and DOA estimation, however joint metrics [2] for SED and DOA, namely location-sensitive detection and calss-sensitive localization, are adopted for the SELD task in DCASE 2020 [3]. It is not appropriate to estimate the SED and DOA separately based on joint metrics. A recently published SELDnet [4, 5] was used as the baseline system. SELDnet uses log-mel spectral coefficients, generalized cross-correlation (GCC), and acoustic intensity vector as input features, and predicts the SED and DOA simultaneously with joint loss [6].

In this report, data augmentation approaches are investigated to expand official dataset. We propose spatial augmentation to simulate new DOA information by voice channel switching. With the same transformation applied to input data and target labels, the representation of the DOA subspace of the official dataset is expanded. Another way to augment DOA information is to process multichannel data simulation. By estimating the room impulse responses (RIRs) of a static sound event and applying the RIRs to other static sound events, more DOA information is generated. Since each recording in the development dataset is marked whether overlapping sound events occur or not, we mix two non-overlapping sound events by adding them in time domain to obtain more training data. SpecAugment [7] is a simple yet effective data augmentation method which is also adopted in the proposed system. We believe a single model can not solve the SELD task perfectly, hence we train several DNN architectures and adopt model ensemble strategy to produce a more accurate SED and DOA estimation. Threshold is important for SED and DOA prediction since joint metrics is used. An optimal threshold is chosen for each sound event on the validation set.

The rest of the report is organized as follows. In Section 2, the proposed method is described in detail, including data augmentation, network training, model ensemble and post-processing. Evaluation results on development dataset is shown in Section 3. Conclusions are summarized in Section 4.

2. PROPOSED METHOD

In the proposed method, several DNN architectures are trained for SELD task based on augmented development dataset. Then model ensemble and post-processing are adopted to get the final sound event detection and localization estimation. We will describe the four parts of the method in detail: the data augmentation, the network training, the model ensemble, and the post-processing.

2.1. Data Augmentation

The official dataset provided by DCASE 2020 consists only 600 recordings. It is not sufficient to train a robust model and may face overfitting problem. So we use data augmentation approaches to overcome the lack of training data.

There are two spatial recording formats for each scene recording in the official dataset, namely FOA format and MIC format. The four microphones extracted from the 32-channel Eigenmike in the MIC format follow the standard tetrahedral structure. Based on the rotation characteristics of tetrahedral microphone arrays, we process voice channel switching for MIC signals to simulate the rotation relationship. And the corresponding FOA signals are processed in a similar way. By applying transformations to MIC channels and FOA channels, we can simulate a new set of DOA labels while keeping SED labels unchanged. Besides, we conduct multichannel data simulation to augment data size by estimating one sound event signal and four RIRs from each static sound event segment. Then one sound event signal and four RIRs randomly selected from different static sound event segments are combined to simulate multichannel data. For the simulated data, the SED labels are the same as the sound event segment used to estimate the sound event signal while the DOA labels are the same as the sound event segment used to estimate the RIRs.

When analysing the experimental results, we found higher DOA error in overlapping segments where two sound events are active at the same time than that in non-overlapping segments. Thus, another data augmentation approach is to generate more overlapping segments. First, we cut the recordings marked with non-overlapping status into small segments where only one kind of sound event is active. Then we randomly select two segments of different sound event classes and add them in the time domain. And the corresponding labels is the union of the two segments. Finally, the mixed segments are concatenated to generate 60 seconds long recordings.

Besides, SpecAugment [7] has been proven to be effective in ASR task. We also adopt the SpecAugment augmentation approach to do time and frequency block masking on the spectrogram. We perform SpecAugment in each mini-batch during training.

2.2. Network Training

In this report, we extract log mel spectra features from multichannel audio of 24 kHz sampling rate using 1024-point discrete Fourier transform from 40 ms length Hanning window and 20 ms hop length. Similar to [6], GCC features are used for the MIC signals, and the acoustic intensity vector are used for the FOA signals. For FOA signals, there are 4 channels of log mel features and 3 channels of intensity vector features, hence up to 7 feature maps. For MIC signals, there are 4 channels of log mel features and 6 channels of GCC features, hence up to 10 feature maps. We use both FOA and MIC signals, so 17 input feature maps are used to train our models.

We adopt several DNN architectures in this task. Generally, the network architecture consists of a high-level feature representation module, a temporal context representation module, and a fully-connected module as shown in Fig. 1. The high-level feature representation module usually contains several convolutional neural network (CNN) blocks, each CNN block usually having a 2D CNN followed by a batch normalization process, a rectified linear unit (ReLU), and a max-pooling operation. ResNet [8] and Xception [9] achieved great performance in image recognition. To solve the specific SELD task, we use the modified version of ResNet and Xception as high-level feature representation modules. Usually recurrent neural network (RNN) is employed to model longer temporal context in the extracted features. In this task, we adopt bidirectional gated recurrent unit (GRU) and factorized time delay neural network (TDNNF) [10] structure for temporal context modeling. The output of GRU or TDNNF is then feed to two parallel branches of fully-connected layers. One branch is for SED estimation and the other for DOA estimation. We try two ways for SED estimation. One way is similar to the baseline where the output layer has N units with N equal to the sound event classes, and a sigmoid activation is used for multiclass-multilabel classification. The second way is to predict $N + 1$ classes with an additional silence class and a softmax activation is used. If overlapping sound events are active, then the probabilities of corresponding labels are set to 0.5. For the DOA estimation, the output layer has $3N$ units corresponding to the Cartesian coordinates (x, y, z) of all sound events, and a tanh activation is used for multioutput regression.

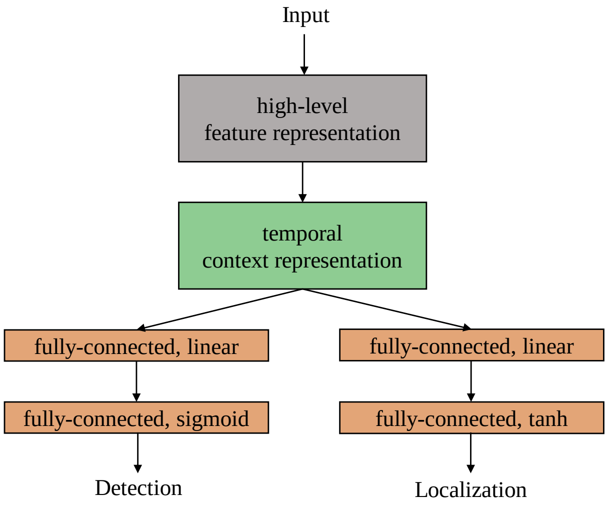

Figure 1: Network architecture for SELD.

In general, we train several DNN architectures with different combination of high-level feature representation modules and temporal context representation modules. Specifically, we train ResNet-GRU, ResNet-TDNNF, Xception-GRU, and Xception-TDNNF.

The optimization criterion for the SED classification and DOA regression are binary cross-entropy and mean square error, respectively. Suggested by second-best performing team [6] in DCASE2019 Task3, DOA regression loss is masked with the ground truth of SED. The SED classification loss and DOA regression loss are combined for joint optimization during training with a weight [1, 10]. The audio clips with a length of 60 seconds are used for training. All DNN architectures are trained with Adam optimizer. The learning rate is set to 0.001 and is decreased by 10% if the SELD score does not improve in 80 consecutive epochs. For single DNN models, if sigmoid activation is used for SED estimation, we adopt a threshold of 0.5. If softmax activation is used for SED estimation, we adopt a threshold of 0.33 to make sure overlapping sound events can be detected. For the ensemble model, we adopt dynamic threshold.

2.3. Model Ensemble

After training all the DNN architectures, model ensemble strategy is adopted to generate the SELD estimation. Here we use the weighted mean of the outputs predicted by different DNN architectures as the ensemble result. Models using the same DNN architecture but trained with different augmented data or models using different DNN architectures but trained with the same augmented data are

Table 1: Evaluation results of the proposed method for development dataset.

	$\mathbf{ER_{20^o}}$	$\mathbf{F_{20^o}(\%)}$	$\mathbf{LE_{CD}}$	$\mathbf{LR_{CD}(\%)}$
Baseline-FOA	0.72	37.4	22.8°	60.7
Baseline-MIC	0.78	31.4	27.3°	59.0
ResNet-GRU	0.29	76.4	9.4°	82.8
ResNet-GRU-Softmax	0.29	76.2	9.1°	81.6
Xception-GRU	0.35	71.4	11.2°	80.0
ResNet-TDNNF	0.35	70.9	12.0°	79.9
Xception-TDNNF	0.37	70.4	12.3°	80.2
Proposed Method	**0.26**	**80.0**	**7.4°**	**84.7**

fused to get the final SELD estimation.

2.4. Post-processing

Instead of using a global threshold for all sound events, we adopt dynamic threshold in this report. An optimal threshold is chosen for each sound event on the validation set.

3. RESULTS ON DEVELOPMENT DATASET

We evaluate our proposed method on the development dataset of TAU-NIGENS Spatial Sound Events 2020. We generate larger training sets with the abovementioned data augmentation approaches, which consists of 112 hours augmented data generated by voice channel switching, multichannel data simulation, and time mixing. Table 1 shows the evaluation results of the proposed method for development dataset. As shown in the table, each proposed single model outperforms the two baseline systems by a large margin. By model ensemble and post-processing, further improvements are achieved for both SED estimation and DOA estimation.

4. CONCLUSION

In this report, we propose an ensemble system to solve the SELD task in DCASE 2020. We first adopt data augmentation approaches to expand the official dataset. Then several DNN architectures are trained to predict SED and DOA simultaneously with multitask learning. Finally model ensemble and post-processing strategies are used to get a more accurate SELD estimation. The evaluation results on the development dataset show that the proposed method outperforms the baseline systems by a significant margin.

5. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[2] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, Oct 2019, accepted.

[3] http://dcase.community/challenge2020/ task-sound-event-localization-and-detection.

[4] 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, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[5] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv preprint arXiv:2006.01919*, 2020.

[6] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 30–34.

[7] 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.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[10] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in *Interspeech*, 2018, pp. 3743–3747.