SOUND EVENT LOCALIZATION AND DETECTION BASED ON CRNN USING ADAPTIVE HYBRID CONVOLUTION AND MULTI-SCALE FEATURE EXTRACTOR

Technical Report

Xinghao Sun[1,3], *Xiujuan Zhu*[1,3*], *Ying Hu*[1,3], *Yadong Chen*[1,3], *Wenbo Qiu*[1,3], *Yuwu Tang*[1,3], *Liang He*[1,2] *Minqiang Xu*[4]

[1] School of Information Science and Engineering, Xinjiang University, Urumqi, China, {xh_sun2019}@stu.xju.edu.cn
[2] Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China
[3] Key Laboratory of Signal Detection and Processing in Xinjiang, China
[4]SpeakIn Technology

ABSTRACT

In this report, we present our method for Detection and Classification of Acoustic Scenes and Events (DCASE) 2021 challenge task 3: Sound Event Localization and Detection with Directional Interference (SELDDI). In this paper, we propose a method based on Adaptive Hybrid Convolution (AHConv) and multi-scale feature extractor. The square convolution shares the weight in each T-F bin of the fixed area in feature map, that is limited. In order to address this problem, we propose a AHConv mechanism instead of square convolution to obtain time and frequency dependencies. We also explored multi-scale feature extractor which can integrate information from very local to exponentially large receptive field within the block. In order to adaptive recalibrate the feature maps after convolutional operation, we designed an adaptive attention block which are largely embodied in the AHConv. On TAU-NIGENS Spatial Sound Events 2021 development dataset, our systems demonstrate a significant improvement over the baseline system. Only the first-order Ambisonics (FOA) dataset was considered in this experiment.

Index Terms— DCASE2021, Sound source localization, Sound event detection, Adaptive hybrid convolution

1. INTRODUCTION

Sound Event Localization and Detection with Directional Interference (SELDDI) is a combined task of recognizing individual sound events of specific classes, detecting their temporal activity, and estimating their location during it, in the presence of interfering directional events not belonging to the target classes and spatial ambient noise. In realistic aural environments, there are numerous co-occurring different sounds emitted from the sources distributed in space. Even humans cannot all correctly identify and locate multiple sources of sound, so it is very challenging for machines. To solve the SELDDI problem, two key issues denoted as sound event detection (SED) [1–5] and sound source localization (SSL) [6–13] have to be addressed.

The methodology proposed in this paper is based on the SELD-Net proposed by Adavanne et al [14]. A convolutional recurrent neural network (CRNN) model was proposed for joint SSL and SED of multiple overlapping sound events in three-dimensional (3D) space. The phase and magnitude of spectrogram were calculated separately on each audio channel as input features. In order to learn both inter-channel and intra-channel features, the input was fed through three consecutive convolutional blocks. Bidirectional Gate Recurrent Unit (BiGRU) was used for temporal context information learning. The output of the BiGRU is fed to two parallel branches of fully-connected blocks. The classes for all sound events would be output on each time-frame, and the sound source would be located in the three-dimensional Cartesian coordinate system. Making as a multi-output regression task can help to estimate in a continuous space.

Compared with DCASE2020 challenge task 3, The main difference is the emulation of scene recordings with a more natural temporal distribution of target events and, more importantly, the inclusion of directional interferences, meaning sound events out of the target classes that are also point-like in nature. For each reverberant environment and every emulated recording, Interferences are spatialized in the same way as the target events, resulting in recordings that are more challenging and closer to real-life conditions. The other difference is the elimination of the dedicated event classification output branch, by adopting the ACCDOA training target which unifies the localization and classification losses in a homogenous regression vector loss, pioneered by Shimada et al [15].

In this paper, We also propose a CRNN framework based on SELD-Net architecture. We adopt Adaptive Hybrid Convolution (AHConv) mechanism and multi-scale feature extractor to handle feature learning insufficiently. The logmel spectrogram and normalized acoustic intensity vector are extracted as input features. Instead of conventional symmetric convolution, the AHConv structure is design to process more and richer spatial features and increase feature diversity by asymmetric convolution. we adopt a multi-scale feature extracting strategy, in which the strategy was designed to capture the longer temporal context information than the conventional convolutions. Moreover, the parallel structure is applied in Adaptive attention block, which adaptive mitigates interference between the channel-wise and time-frequency-wise by exploring two different branches. Additionally, the Adaptive attention block can also promote the robustness when a single branch is disturbed by

*Equal Contribution

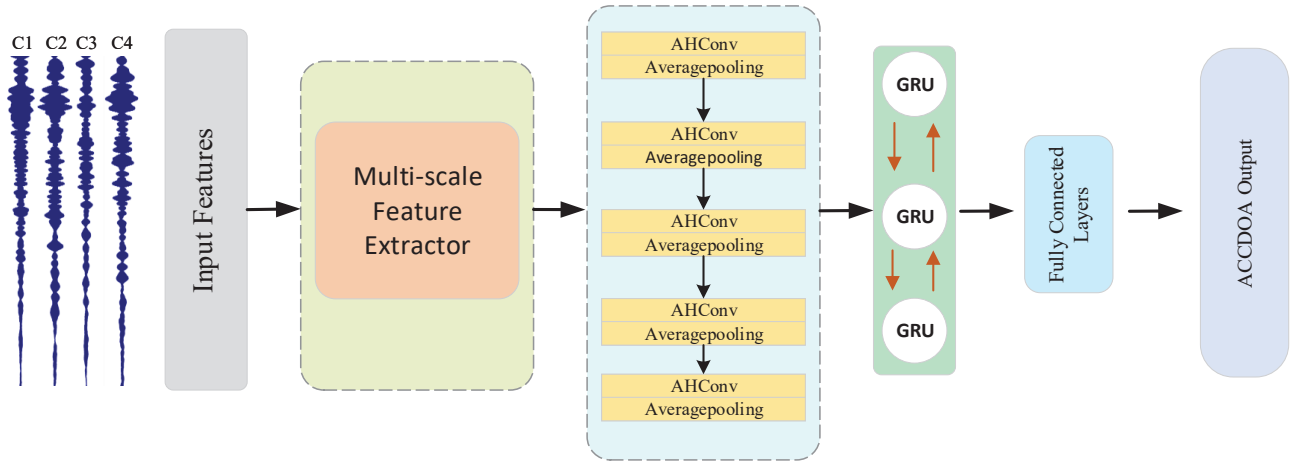

Figure 1: The overall of our proposed method.

the ambient noise without the presence of sound events. Furthermore, we conduct experiments on development dataset to verify the effectiveness of our proposed method.

This paper is organized as follow: we will introduce the proposed method in Section II. The experiment setup will be stated in Section III. The development results compared with the baseline method will be described in Section IV. Finally, we draw a conclusions and future work in Section V.

2. PROPOSED METHOD

We proposed a method with Adaptive Hybrid Convolution (AHConv) and multi-scale feature extractor which achieves great performance to deal with SELDDI in the noisy and reverberant scenes. The proposed network can predict the sound event classes active for each of the input frames along with their respective spatial location, and produce the temporal activity and DOA trajectory for each sound event class. The network diagram is presented in Fig.1. The input to the method is the multichannel audio. The logmel spectrogram and sound intensity vector (SIV) are extracted as the input features of the network. The multi-scale feature extractor as depicted in 4, then followed five AHConv blocks and five average pooling layers. After that, the time dimension is downsampled 5 times, and the frequency dimension is downbsampled 32 times. Then, Bidirectional Gated Recurrent Unit (Bi-GRU) is used to learn the temporal context information. This is followed by fully connected layers.

2.1. Multi-scale Feature Extractor

Among the various CNN architectures, if the network contains shorter connections between layers close to the input and those close to the output, it can be substantially deeper, more accurate, and efficient to train, to further improve the information flow between layers [16]. In this work, we combine the advantages of DenseNet and dilated convolution, and propose a extractor called multi-scale feature extractor. To properly combine DenseNet with the dilated convolution [17], we propose a multi-scale feature extractor that has a multiple dilation factor within a single layer. The dilation factor depends on which skip connection the channels come from, as shown in Fig. 2. The output of each dilated layer are fed into a Adaptive attention block. The Adaptive attention block reweigh the information of channel-wise and of spatial-wise dimension. That can enhance the important features and weaken the less important features. The outputs of the lth layer x_l receives the feature-maps of all preceding layers express as:

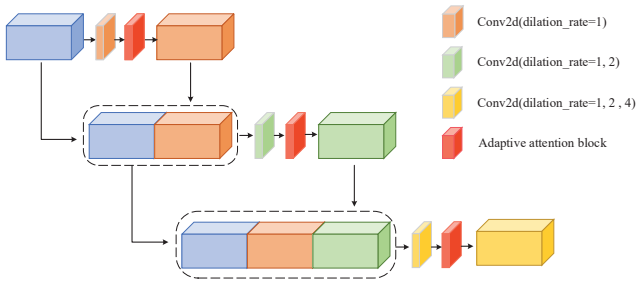

Figure 2: Multi-scale feature extractor

$$x_l = \psi([x_0, x_1, x_2, ...x_{l-1}] \circledast k_l^{d=1,2,...2^{l-1}}) \quad (1)$$

where $[x_0, x_1, x_2, ...x_{l-1}$ denotes the concatenation of the feature maps produced in layers $0, ...l-1$, ψ is a nonlinear transformation consisting of batch normalization (BN) followed by ReLU and dilated convolution with the k_l kernel, $\circledast$ denotes convolution operation and d is the dilated rate in each layer.

2.2. Adaptive Hybrid Convolution

Some of the prior works [18, 19] has shown that a standard square convolutional layer with a filter size of k $\times$ k can be factorized as a sequence of two layers with k $\times$ 1 and 1 $\times$ k filters to reduce network complexity and lighten the computational burden. This asymmetrical convolutional structure is better than a square convolutional structure in processing more and richer spatial features and increasing feature diversity. In addition, asymmetric convolution

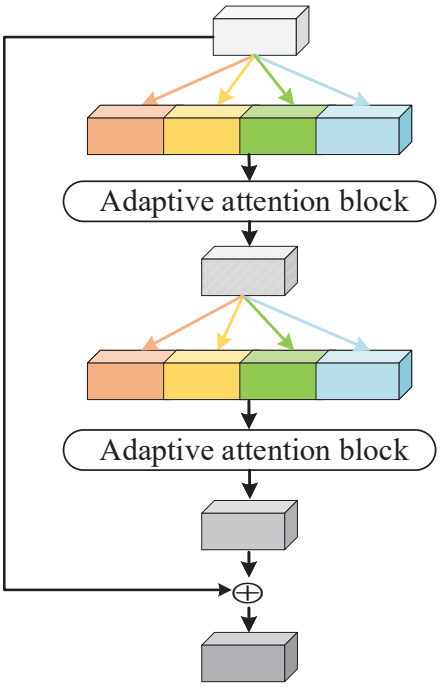

Figure 3: Adaptive Hybrid Convolution (AHConv). Each color represents a different convolution kernel, and the squares represent the convolution graph.

can obtain faster calculation speed and smaller parameter amount while ensuring performance. The weight learning of the square convolution relies on the network, but is limited by the size of filter. Therefore, the square convolution are not capture fine-grained time-frequency features. In order to address this problem, we propose a hybrid convolution mechanism based on the asymmetric convolutional structure, as shown in Fig. 3. That is, a parallel structure is composed of a filter size 1×3 and 1×5 for time frames, and a filter size 3×1 and 5×1 for frequency bin, thus the time dependency and frequency dependency are capture respectively.

2.3. Adaptive Attention block

We design an adaptive attention block as seen in Fig 4. The up half part denotes the path of channel attention (CA) [20], and the lower half part the time-frequency attention (TFA) [21]. In adaptive attention block, different weights are applied to the channel and the time-frequency (TF) domain, which can guide the network to pay different attention to the characteristics of channel-wise and time-frequency-wise. The features of each part will undergo a two-dimensional convolution with a (1×1) kernel size. The convolution will learn the weight of each part and add adaptively.

3. EXPERIMENT SETUP

3.1. Dataset

Development set of TAU-NIGENS Spatial Sound Events 2021 has two types of data, one is 4 channel directional microphone array (MIC) from tetrahedral array and the other one is first-order ambisonic (FOA) data. We used the FOA format for the challenge. The SELD development dataset consists of 600 one-minute audio clips divided into training, validation, and test set of size 400, 100, and 100 clips, respectively. The development dataset are distributed between 12 classes of alarm, crying baby, crash, barking dog, footsteps, knocking on door, female speech, male speech, female scream, male scream, ringing phone and piano. Additionally, dry recordings of disparate sounds not belonging to any of those classes are also spatialized in the same way to serve as directional interference. The sounds are sourced from the running engine, burning fire, and general classes of NIGENS database.

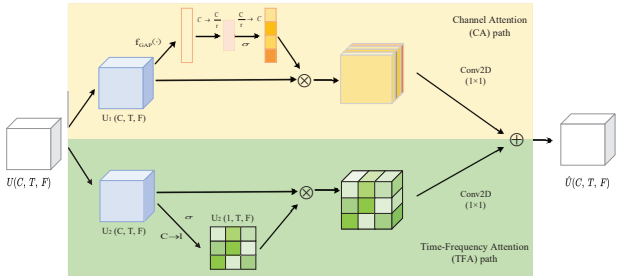

Figure 4: Adaptive attention block.

3.2. Evaluation metrics

The performance of our proposed model is evaluated by the individual metrics for SED task and SSL task. Standard polyphonic SED metrics, F-score (F1) and error rate (ER) across segments of one second without overlapping are utilized [22]. The DOA estimation in SSL task was evaluated using frame-wise metrics [23] of DOA error (DE) and frame recall (FR). Considering that a TP is predicted only when the spatial error for the detected event is within the given threshold of 20° deviate from the reference, ER and F replaced with ER_{20° and F_{20°. Classification-dependent localization metrics are computed only across each class, instead of across all outputs, DE and FR replaced with LE_{CD} and LR_{CD}. A more detailed description can be obtained in [23, 24].

3.3. Training procedure

The sampling frequency was used at 24 kHz in our method. STFT was applied with configurations of 20 ms frame length and 10 ms frame hop. The input frame length is 1,024 frames. We use a batch size of 64. Moreover, to ensure a fair comparison, all models were trained for 500 epochs with the Adam optimizer of the same initialized parameters. An early stopping mechanism is used to avoid overfitting during training, where the training is stopped if no improvements on validation split for 50 epochs.

3.4. Our challenge submissions

- **Sun_AIAL-XJU_task3_1**: Proposed method trained using the same training splits as the baseline method.
- **Sun_AIAL-XJU_task3_2**: Proposed method trained using the five splits development dataset.

Table 1: The performance comparison for different methods on the development dataset

Method	ER_{20°	$F_{20^\circ}(\%)$	LE_{CD}	$LR_{CD}(\%)$
DCASE2021baseline	0.69	33.9	24.1	43.9
Sun_AIAL-XJU_task3_1	**0.57**	**52.6**	**19.6**	**58.1**
Sun_AIAL-XJU_task3_2	**0.52**	**55.3**	**19.1**	**60.9**

4. RESULT AND DISCUSSION

Our proposed model result outperform the DCASE 2021 baseline model, **Sun_AIAL-XJU_task3_1** achieve the improvement of 0.12, 18.7%, 4.5 and 14.2% respectively, and **Sun_AIAL-XJU_task3_2** achieve the improvement of 0.17, 21.4%, 5 and 17% respectively. **Sun_AIAL-XJU_task3_2** had 100 more training data than **Sun_AIAL-XJU_task3_1**, and improved performance 0.05, 2.7%, 0.5, 2.8% respectively. This proves that our model still has great potential to grow with the increase of training data.

5. CONCLUSIONS

In this paper, we propose a SELDDI method based on Adaptive Hybrid Convolution (AHConv) and multi-scale feature extractor. AHConv was design to capture the time and the frequency dependencies. Multi-scale feature extractor was designed to extract the multi-scale feature maps. We also proposed an adaptive attention block embodied in AHConv. The results on the development dataset show that our proposed method outperforms the baseline method on four evaluation metrics. Next we will introduce data augmentation method to improve the performance of the model.

6. REFERENCES

[1] Y. Li, M. Liu, K. Drossos, and T. Virtanen, "Sound event detection via dilated convolutional recurrent neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 286–290.

[2] J. Yan, Y. Song, L.-R. Dai, and I. McLoughlin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 326–330.

[3] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1466–1478, 2020.

[4] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental sound classification with parallel temporal-spectral attention," *Proceedings of INTERSPEECH 2020*, 2020.

[5] X. Zheng, Y. Song, J. Yan, L.-R. Dai, I. McLoughlin, and L. Liu, "An effective perturbation based semi-supervised learning method for sound event detection," *Proc. Interspeech 2020*, pp. 841–845, 2020.

[6] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 136–140.

[7] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, 2018.

[8] 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.

[9] T. N. T. Nguyen, W.-S. Gan, R. Ranjan, and D. L. Jones, "Robust source counting and doa estimation using spatial pseudo-spectrum and convolutional neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2626–2637, 2020.

[10] Z. Tang, J. D. Kanu, K. Hogan, and D. Manocha, "Regression and classification for direction-of-arrival estimation with convolutional recurrent neural networks," *Proc. Interspeech 2019*, pp. 654–658, 2019.

[11] H. Sundar, W. Wang, M. Sun, and C. Wang, "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4642–4646.

[12] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[13] W. He, P. Motlicek, and J.-M. Odobez, "Adaptation of multiple sound source localization neural networks with weak supervision and domain-adversarial training," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 770–774.

[14] 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, 2018.

[15] 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[18] X. Ding, Y. Guo, G. Ding, and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1911–1920.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[21] W. Xia and K. Koishida, "Sound event detection in multichannel audio using convolutional time-frequency-channel squeeze and excitation," *Proc. Interspeech 2019*, pp. 3629–3633, 2019.

[22] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[23] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 333–337.

[24] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.