THE HITACHI DCASE 2021 TASK 3 SYSTEM: HANDLING DIRECTIVE INTERFERENCE WITH SELF ATTENTION LAYERS

Technical Report

Nelson Yalta, Takashi Sumiyoshi, Yohei Kawaguchi

Hitachi, Ltd. Research & Development Group
Tokyo, Japan
nelson.yalta.wm@hitachi.com

ABSTRACT

This report describes the Hitachi system for the DCASE 2021 Challenge - Task 3. Our proposal relies on a single-stage system that employs the transformer encoder (i.e., self-attention layers) as a core idea. We evaluate the effect of applying different transformer configurations to handle the directive interferences on the presence of multiple sound events. Additionally, the transformer employs residual connections to extract the features from the input streams. We trained the model using specaugment as data augmentation and performed threshold postprocessing for each sound event. Employing the first-order Ambisonic (FOA) signals, the transformer was trained using the activity-coupled Cartesian DOA vector (ACCDOA) representations. This unified training framework showed better performance than training the model for each sub-task independently.

Index Terms— residual connections, transformer, data augmentation, self attention

1. INTRODUCTION

Directional characteristics of sound signals and their detection have an essential role in the sound processing signal. The correct recognition of these features enables the following subprocesses such as sound source separation or speech recognition. This task of jointly detect a given sound event and estimate its direction-of-arrival (DOA) is known as sound event localization and detection (SELD) [1]. It is applied to several areas: robotics [2], meeting transcriptions [3], autonomous driving [4], etc.

The last years have witnessed considerable growth in SELD studies thanks to the DCASE Challenges [5, 6]. The reported studies proposed systems that aim to handle the SELD task jointly (i.e., single-stage) [7] or separately in two sub-tasks: sound event detection (SED) and sound event localization (SEL), i.e., two-stages [8]. The proposed systems aimed to process the location of sound events under several environmental issues, such as noises, room reverberations, moving and overlapping sound events, conditional issues present on on Task 3 of the DCASE 2019 [9] and DCASE 2020 [6] challenges. The Task3 of the DCASE 2021 challenge [10] included a new challenging feature: events outside the target classes with directional interferences.

Using the first-order Ambisonic (FOA) signals provided with the data set, this report proposes using self-attention layers implemented into a single-stage system to handle this new conditional issue. The model employs an activity-couple Cartesian DOA vector (ACCDOA) representation to estimate the event class and its coordinate [7].

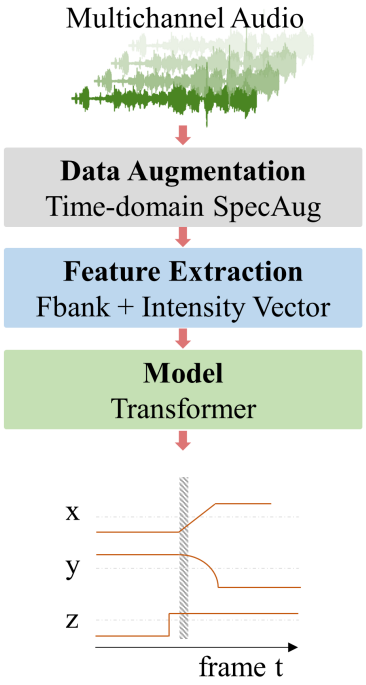

Figure 1: System overview.

2. SYSTEM

In this section, we describe the components of the pipeline for our SELD model 1.

2.1. Data Augmentation

Data augmentation techniques improve the performance of the trained model enabling a better generalization of the given problem. SpecAugment is an augmentation scheme applied to sound processing tasks, such as speech separation or speech recognition [11]. The initial implementation acted directly on the spectrogram of the input signals, requiring a negligible amount of additional computational resources. SpecAugment consists of the policies of frequency

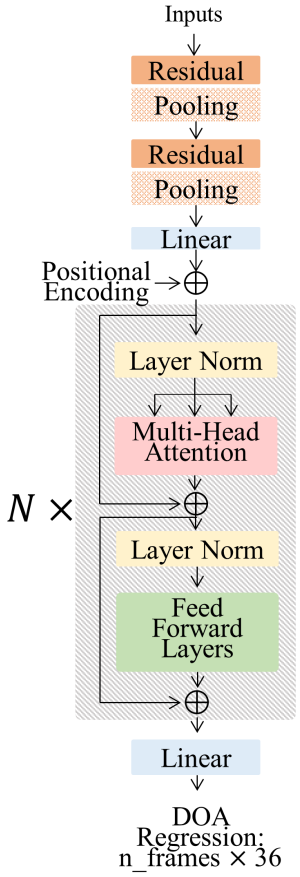

Figure 2: Transformer Encoder for SELD.

masking, time masking, and time warping. We employed the frequency and time masking of a time-domain version of specaugment. Also, speed perturbation was employed as one additional technique for data augmentation [12]. These three techniques were applied to one random channel of the input batch, each technique separately. The corrupted waveform signals are stacked to the initial input batch, obtaining a batch size four times from the initial size.

2.2. Feature Extraction

The TAU-NIGENS Spatial Sound Events 2021 [10] data set comprises two input data formats: a tetrahedral microphone array and a first-order Ambisonic (FOA) signal. For our experiments, we employed the FOA signals.

The FOA input comprises signals from four channels which indicates the Omni-directional component w, and the directional components x, y, and z. From the channels, we extracted the logmel features and the instantaneous sound intensity vector. The intensity vectors carry the acoustical energy direction of a given sound wave. We calculated the intensity vectors in the STFT domain using the process described at [13]. The obtained intensity vectors are then concatenated as additional channels to the logmel features. The resulting input has seven channels: four logmel + three intensity vectors.

2.3. Model Description

We employed the architecture named Transformer for our experiments [14]. The Transformer aims at transforming an input feature sequence into its corresponding output sequence. The Transformer, initially proposed for handling natural language processing issues [15], is widely used in sound applications, such as speech recognition [14, 16]. The Transformer comprises an encoder-decoder architecture, using an input sequence a few times longer than the output one [16]. The Transformer employs a self-attention mechanism with multi-head attention and position-wise feed-forward networks.

For this report, we employed the transformer encoder instead of the whole encode-decoder architecture to model the locations and the classes of the sound events 2.

The multichannel input represented as a signal of F-dim log-Mel filterbank features is subsample by using two CNN/Residual layers with max-pooling layers. The CNN layers use a stride size of 1, a kernel size of 3, and a padding size of 1. The residual layers comprise two CNN layers with kernel size three and one CNN layer as identity shortcut, as described in [17]. Then, a fully connected layer without bias resizes the output from the CNN layers into D-dim vectors. After adding the positional encoding information, the signal is processed by the N-blocks of self-attention layers. Finally, the output of the self-attention is mapped into a $3 \times n_{classes}$-dim vector.

2.4. Post-processing

The audio tracks are split into subsegments with an overlap during inference. Then, the results from the overlapped frames are averaged. Finally, we conduct post-processing of minimum thresholding for each sound event using a hyperparameter search [18]. The threshold is selected when the average of the error accuracy and the effectiveness measure (i.e., $1-F_{score}$). The search is performed along with a uniform distribution between $[T_{min}, 1\rangle$. We observed that the value of T_{min} affects the final result by improving the Localization recall and degrading the other measures.

During training, we used a global threshold for all sound events set to 0.5.

3. EXPERIMENTS

This report summarizes a series of experiments conducted to improve the localization and detection of sound events in the presence of directional interferences.

3.1. Experimental settings

In this report, we summarize the effectiveness of the transformer for SELD tasks on directional interferences. The experiments were performed using the SpeechBrain toolkit [19] implemented on the PyTorch framework.

The models were trained for 2000 epochs using an NVIDIA RTX 3060Ti graphic processing unit. For the optimization, we employed the "AdamW" solver with an initial learning rate of 10^{-3} and a linear learning scheduler with a final learning rate of 10^{-7}. The input batch uses 16 files with six second-length, which after the data-augmentation becomes 64. The mean squared root is employed as a cost function.

We employed the metrics described at [5].

For our experiments, we employed the following architectures:

Table 1: Evaluation results using the FOA array on the development set - valid fold

	$\mathbf{ER}_{20^\circ} \downarrow$	$\mathbf{F}_{20}(\%) \uparrow$	$\mathbf{LE}_{\mathbf{CD}} \downarrow$	$\mathbf{LR}_{\mathbf{CD}}(\%) \uparrow$
Baseline	-	-	-	-
SELDnet	0.72	33.4	30.1	47.8
Residual-GRU	0.68	41.0	27.0	**56.3**
Transformer-1	0.70	37.0	26.2	48.5
Transformer-2	0.70	39.3	24.6	51.3
Transformer-3	0.71	40.9	25.9	55.9
Transformer-4	**0.67**	**42.9**	**22.3**	51.9

Table 2: Evaluation results using the FOA array on the development set - test fold

	$\mathbf{ER}_{20^\circ} \downarrow$	$\mathbf{F}_{20}(\%) \uparrow$	$\mathbf{LE}_{\mathbf{CD}} \downarrow$	$\mathbf{LR}_{\mathbf{CD}}(\%) \uparrow$
Baseline	0.69	33.9	24.1	54.9
SELDnet	0.67	38.8	23.9	51.2
Residual-GRU	0.66	41.1	24.5	54.9
Transformer-1	0.65	42.4	22.1	52.7
Transformer-2	0.68	42.3	20.7	50.5
Transformer-3	0.66	46.5	21.3	**58.1**
Transformer-4	**0.61**	**49.4**	**18.6**	54.8

SELDnet: A convolutional recurrent neural network employed for SELD tasks [20]. We employed the configuration described at [10], using as input the data augmentation described at section 2.1 and the feature extraction of section 2.2

Residual-GRU: Similar to **SELDnet**, replacing the CNN layers with residual connections as decribed at [17].

Transformer-1: This models employes 2 residual blocks with 64 channels. After the residual block a max-pooling layers is stacked. The first max-pooling has a stride of (5, 4) size and the second a size of (1, 4). After the second max-pooling, a linear layer reduce to 128-dims. For this model, we employed a fixed positional encoding. The model has three self-attention blocks with 4 attention heads and 1024 units for the position-wise FF.

Transformer-2: Similar to **Transformer-1**, without positional encoding.

Transformer-3: Similar to **Transformer-1**, with six self-attention blocks and without positional encoding.

Transformer-4: Similar to **Transformer-1**, with a linear layer with 256-dims of outputs, 8 attention heads, 2048 position-wise FF units, and without positional encoding.

3.2. Results

We evaluate our models on the development set of the TAU-NIGENS Spatial Sound Events 2021. Table 1 and 2 show the results for the validation and testing fold of the development set, respectively. The baseline results were obtained from [10]. Without post-processing, the models show a slight improvement over the baseline. In table 1, we also observe that the use of Residual connections in a SELDnet model improves localization recall. However, the other metrics do not show a relevant improvement. Table 3 lists the results of Transformer-4 using the post-processing method described in section 2.4. We observe that the Transformer reaches a higher location recall when using a T_{min} close to zero. However, the performance of the models on the other metrics degrades.

Table 3: Evaluation results using the FOA array on the evaluation set - valid fold, with different minimum threshold

	$\mathbf{ER}_{20^\circ} \downarrow$	$\mathbf{F}_{20}(\%) \uparrow$	$\mathbf{LE}_{\mathbf{CD}} \downarrow$	$\mathbf{LR}_{\mathbf{CD}}(\%) \uparrow$
Baseline	0.69	33.9	24.1	54.9
Transformer-4 (0.05)	0.73	51.0	22.6	**72.4**
Transformer-4 (0.3)	**0.60**	**54.0**	**20.0**	65.3

4. REFERENCES

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *CoRR*, vol. abs/1807.00129, 2018. [Online]. Available: http://arxiv.org/abs/1807.00129

[2] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, vol. 1, 2004, pp. 1033–1038 Vol.1.

[3] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865509001603

[4] M. K. Nandwana and T. Hasan, "Towards smart-cars that can listen: Abnormal acoustic event detection on the road," in *INTERSPEECH*, 2016.

[5] 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*, vol. 29, pp. 684–698, 2020. [Online]. Available: https://https://arxiv.org/abs/2009.02792

[6] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2020)*, November 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[7] K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Sound event localization and detection using activity-coupled cartesian doa vector and rd3net," DCASE2020 Challenge, Tech. Rep., July 2020.

[8] T. N. T. Nguyen, D. L. Jones, and W. S. Gan, "Dcase 2020 task 3: Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking," DCASE2020 Challenge, Tech. Rep., July 2020.

[9] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and uetection," in *Submitted to Detection and Classification of*

Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019. [Online]. Available: https://arxiv.org/abs/1905.08546

[10] 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 preprint arXiv:2106.06999*, 2021. [Online]. Available: https://arxiv.org/abs/2106.06999

[11] 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," *Interspeech 2019*, Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[12] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[13] Y. Cao, T. Iqbal, Q. Kong, Z. Yue, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," DCASE2020 Challenge, Tech. Rep., July 2020.

[14] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5884–5888.

[15] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[16] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. Interspeech 2019*, 2019, pp. 1408–1412. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1938

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[18] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, T. Chen, J. Pan, J. Du, and C.-H. Lee, "The ustc-iflytek system for sound event localization and detection of dcase2020 challenge," DCASE2020 Challenge, Tech. Rep., July 2020.

[19] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[20] 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