LOCALIZATION AND DETECTION FOR MOVING SOUND SOURCES USING CONSECUTIVE ENSEMBLE OF 2D-CRNN

Technical Report

Ju-man Song

LG Electronics
Advanced Robotics Research Lab, Seoul, Republic of Korea
juman.song@lge.com

ABSTRACT

This technical report introduces a deep learning strategy for sound event localization and detection in DCASE 2020 Task 3. This strategy is designed to get accurate estimation of both detecting and localizing moving sound events by splitting a task into five sub-tasks. Each sub -task estimates the number of existing sound sources, the number of sound directions, single sound direction, multiple sound directions, and category of events. Thus, each two dimensional convolutional recurrent neural network (2D-CRNN) is focused on each sub-task. In this way, we could improve its robustness to complex conditions. Finally, the consecutive ensemble strategy is performed to achieve high performance with some decision logic. With the proposed strategy, we could get optimal network models for each sub-task. The proposed strategy is evaluated on the development set of TAU-NIGENS Spatial Sound Events 2020, and shows notable improvements.

Index Terms— DCASE 2020, Sound Event Localization and Detection, CRNN

1. INTRODUCTION

Sound Event Localization and Detection (SELD) task consists of two main objectives. One is identifying temporal sound events, and the other is finding where the detected events occur. However, when some additional conditions are considered, this task can be split to several sub-tasks. If there could be one or two sound sources simultaneously, then the SELD task has to distinguish each cases. For the complex task, a method of using consecutive ensemble is suggested [1], and splitting networks into three sub-networks is also suggested in [2]. Those researches considered 2019 data set [3] which includes only static sound sources.

However, 2020 data set [4] includes moving events. Even though, some cases of that data set are even worse than 2019. For example, two different events occur in same direction, and two same events occurs simultaneously. The sample rate of audio is also reduced from 48kHz to 24kHz that means sound information over 24kHz is removed, and the number of event classes increase from 11 to 14. Signal to noise ratio (SNR) setting is changed from average 30 dB to maximum 30 dB. The evaluation metrics also changed by combining event detection results with direction of arrival (DOA) results. Those changes make this 2020 challenge harder than 2019 challenge, and also make the consecutive ensemble method of [1] cannot be used. Furthermore, [1] also cannot be used at real-time applications like smart phone and robots. It takes too much more than 1 minute to get one SELDnet result.

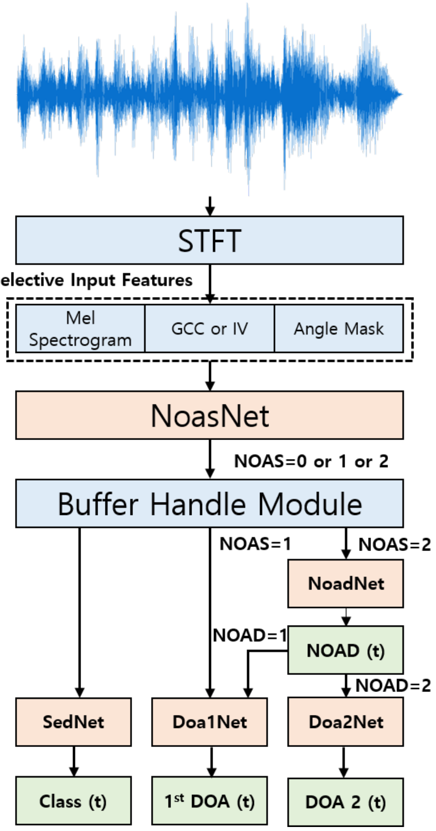

Figure 1: Block diagram of the proposed strategy.

To overcome such obstacles, we suggest a strategy which is developed on consecutive ensemble of five 2D-CRNNs. First network is number of acoustic sources network (NoasNet) which estimates the number of event sources. It uses hole wav data to learning from no event to two events. Second one is number of acoustic direction network (NoadNet) designed to distinguish if two events are overlapped in space simultaneously or not. Thus, NoadNet just uses acoustic data with two sound events. Third network is sound event detection network (SedNet). It predicts events which occur. Fourth and fifth networks are DOA 1 and 2 networks (Doa1Net, Doa2Net).

Table 1: The architecture and parameters of the proposed strategy

Network Name		NoasNet	NoadNet	Doa1Net	Doa2Net	SedNet
Layer Type	Variable	Values				
Input	Input Shape	10 x 300 x 64	10 x 300 x 64	10 x 300 x 64	10 x 300 x 64	4 x 300 x 64
Convolution Block x 3	Pool	(4,4,2) x (5,1,1)				
Permute	Output Shape	60 x 64 x 2				
Reshape	Output Shape	60 x 128				
Recurrent Block x 2	Unit Type	GRU	GRU	LSTM	LSTM	GRU
Time Distributed Dense	Units	128				
Dropout	Rate	0.2				
Time Distributed Dense	Units	128				
Dropout	Rate	0.2				
Time Distributed Dense	Units	1	1	3	6	14
Activation	Function	linear	linear	tanh	tanh	sigmoid
Output	Type	0 or 1 or 2	1 or 2	(x1,y1,z1)	(x1,y1,z1), (x2,y2,z2)	0 to 13
Block Name		Convolution Block				
Conv2D		64 filters, 3 x 3 kernel, same padding				
BatchNorm		-				
Activation		Relu function				
MaxPooling2D		T x P Pooling				
Dropout		0.2 Dropout rate				
Block Name		Recurrent Block				
Bidirectional Network		128 Units				
Activation		tanh Function				
Dropout		0.2 Dropout rate				

Those networks find the location of acoustic sources as Cartesian xyz coordinates.

With the results of five networks, we estimate certain events and location moving sound sources by using consecutive ensemble. NoasNet becomes a on-set detection module with results from zero to two. When the trigger result of NoasNet is one, a chunk of sound data is saved until the trigger level goes to zero or two. Then, the chunk of sound data is passed to Doa1Net to get location. When, the result of NoasNet becomes two, there are more sophisticated process to get location data. A chunk of sound data which has output of NoasNet as two, is transferred to NoadNet. NoadNet determines whether two sound sources has overlapped period or not. According to the result of NoadNet, the chunk is delivered to Doa1Net or Doa2Net. Whether passing Doa1Net or Doa2Net, any chunks are sent to SedNet to identify categories of sound events. Based on this strategy, it is described specific explanation in Section 2.

2. PROPOSED STRATEGY

2.1. Consecutive ensemble

As introduced in chapter 1, we utilize five 2D-CRNN sub-networks using different input data which are depends on the results of NoasNet which is determining the number of events appeared in current frame. To obtain finest result for the challenge, we quantify the result of NoasNet using boundary values. At first, if the output of NoasNet is greater than 1.9 we set the result as two. Else, if the output of NoasNet is greater than 0.4, we set the result as one. Otherwise, all the remained result are set as zero.

With the result of NoasNet, consecutive frames with same result of NoasNet are bound to make a chunk. While audio signal being delivered, NoasNet continues to check whether targeted events are detected or not. When an acoustic event in the SedNet category appears, this system binds frames until the length of chunk full-fills fixed length of feature or the result of NoasNet changes to different values. If the acoustic event ends before the chunk fills fixed length, the chunk is copied to fill the remaining feature length. When the fixed length of feature is filled, but the result of NoasNet still one, then, next chunk is recorded with new audio data after previous chunk. Completed chunk is passed to Doa1Net to estimate the moving location of fixed length. The final DOA estimation results cut by the length of NoasNet results.

If another acoustic event is appeared additionally while creating a chunk for one sound source, then a new chunk with the result of NoasNet as two is generated by same way with that of one. Then, chunks of two acoustic events are transferred to NoadNet to determine whether both events appears at same location or not. If the result of NoadNet is determined as two than that chunk is delivered to Doa2Net, and the other case to Doa1Net. The reason is that, in case of two sound sources are arrived from same location, acoustic channels of each sound sources are theoretically identical, thus mixed sound source could be regarded as single sound source. Moreover, due to Doa2Net is designed to produce two sets of xyz data, the input features of same directional sound sources could be unseen data. Despites of Doa2Net in [1] estimate just one angle information using the other angle information from Doa1Net, we cannot longer use the other angle information, because two acoustic sources can move spatially, in the time of a chunk. As a result, the error rate for the two angles of the two events can be relatively higher than that given in [1].

To reduce DOA error of two different located acoustic events, we compare Doa2Net results with previous frame result by calculating Euclidean distances of two previous and current angles. There could be two choice of previous to current pairs, and we choose the pair which has minimal distances. This could be a very simple tracking method of acoustic signals. Tracking filters like extended

Table 2: The Result and Complexity of Each Submission

Submission	Song_LGE_task3_1	Song_LGE_task3_2	Song_LGE_task3_3	Song_LGE_task3_4	Baseline_{FOA}	Baseline_{MIC}
In. Doa1Net	MIC	MIC	FOA	FOA	FOA	MIC
In. Doa2Net	MIC	FOA	MIC	FOA	FOA	MIC
2020 Metrics	Results on Development Set					
ER_{20°	0.57	0.58	0.57	0.58	0.72	0.78
F_{20°	50.6%	49.5%	50.6%	49.5%	37.4%	31.4%
LE_{CD}	20.1°	21.2°	20.1°	21.2°	22.8°	27.3°
LR_{CD}	64.2%	64.2%	64.2%	64.2%	60.7%	59.0%
2019 Metrics	Results on Development Set					
ER	0.46	0.46	0.46	0.46	0.54	0.56
F	64.1%	64.1%	64.1%	64.1%	60.9%	59.2%
LE	16.2°	17.2°	16.2°	17.2°	20.4°	22.6°
LR	74.6%	74.6%	74.6%	74.6%	66.6%	66.8%
Sub-network	Complexity					
NoasNet	491,409				-	-
NoadNet	491,409				-	-
SedNet	489,630				-	-
Doa1Net	623,251	623,251	622,099	622,099	-	-
Doa2Net	492,054	622,486	492,054	622,486	-	-
Total	2,587,753	2,718,185	2,586,601	2,717,033	-	-

Kalman filter can be used to this section, when the computational complexity is allowed.

Basically, all chunks made from the result of NoasNet have to transferred to SedNet. It means SedNet uses chunks whose NoasNet results are one or two. Though, chunks with no sound event are abandoned for reducing noisy data. From all feature data in all chunks, we just use only mel-spectrogram data for SedNet. Finally, one or two events of higher ranking outputs of SedNet are chosen, according to NoasNet result.

2.2. Feature Extraction

Development set of TAU-NIGENS Spatial Sound Events 2020 [4] 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. The baseline algorithm of the challenge uses mel-spectrogram (64 mel filter banks) feature for both data type. To get phase information of each data set, it uses generalized cross correlation (GCC) from MIC and intensity vector from FOA. We use those features for each data set. For FOA data, we add one more feature, angle mask which has been introduced at [2]. FOA data was used just for Doa1Net and Doa2Net in the second, third and fourth submissions. The first submission uses only MIC data for Doa1Net and Doa2Net. Other four networks use only MIC data. For SedNet, we just use only 4 channel mel-spectrogram features, because it shows better performance than features with phase information.

2.3. Network Architecture

The base architecture of each sub-networks is the baseline SELDnet [5]. Sub-networks build optimal models independently using Adam optimizer with default parameters. To get certain goal of each networks, we modified the baseline networks in various ways. The overall summary of five networks are described in Table 1. It just describes the summary of submission 1 which uses only MIC data set. For other submissions, we use FOA data with 8 channel mel data including mel spectrogram, intensity vectors, and angle mask. This input feature change the input dimension as 8.

3. RESULTS

The proposed strategy were evaluated on the development data set in [4] as suggested validation. We trained our five models on 4 splits (3 to 6) and validated on the second split. Then, we tested our consecutive ensemble of five 2D-CRNN models on the first split. The results are on Table 2. It also shows the result of 2019 metrics which gained better improvement than 2020 metrics. With the 2020 metrics, it can achieve high scores only when the classification and localization get precise results in simultaneous frames. Overall, as shown in Table 2, we got better results than baseline.

4. SUBMISSIONS

We submitted four results by changing input features to Doa1Net and Doa2Net as described in Table 2. Table 2 also explains evaluation metric values for development set. From the development set with same approaches, Doa1Net and Doa2Net with MIC data shows the best performance for the proposed strategy. However, on the validation set of evaluation data, we get better performance on FOA data. Thus, we submitted 4 type results as in Table 2.

5. CONCLUSIONS

In this technical report, we proposed a strategy of consecutive ensemble of five 2D-CRNNs to detect and locate moving acoustic sources. The proposed algorithm is developed to be used on real-time scenario. As time goes by, the streaming audio data could be converted to input feature to NoasNet, and according to the result of NoasNet, we can estimate certain class of events with location data. The data set in [4] has up to two acoustic sources, but we can expand this strategy to three or more sources by adjusting NoasNet and NoadNet from two sources to n sources, and add more DOA networks from Doa2Net to DoanNet. Of course, the computational complexity should be increased as the order of sound sources increases, and more researches also should be done to localize more sound sources. Tracking techniques would be helpful for that. Additionally, each sub-networks can be used separately as target system required. In various ways, we can modify our strategy with independent sub-tasks to various systems.

6. REFERENCES

[1] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," DCASE2019 Challenge, Tech. Rep., June 2019.

[2] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by dnn-based denoising and source separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 651–655.

[3] 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

[4] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv e-prints: 2006.01919*, 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[5] 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