THREE-STAGE APPROACH FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

*Kyoungjin Noh, Jeong-Hwan Choi, Dongyeoup Jeon, and Joon-Hyuk Chang**

Department of Electronics and Computer Engineering
Hanyang University, Seoul, Republic of Korea
{nkj0318,suwoncjh,dongyeoup,jchang}@hanyang.ac.kr

ABSTRACT

This paper describes our three-stage approach system for sound event localization and detection (SELD) task. The system consists of three parts: sound activity detection (SAD), sound event detection (SED), and sound event localization (SEL). Firstly, we employ the multi-resolution cochleagram (MRCG) from 4-channel audio and convolutional recurrent neural network (CRNN) model to detect sound activity. Secondly, we extract log mel-spectrogram from 4-channel audio, harmonic percussive source separation (HPSS) audio, mono audio, and train another CRNN model. Lastly, we exploit the generalized cross-correlation phase transform (GCC-PHAT) of each microphone pairs as an input feature of the convolutional neural network (CNN) model for the SEL. Then we combine SAD, SED, and SEL results to obtain the final prediction for the SELD task. To augment overlapped frames that degrade overall performance, we randomly select two non-overlapped audio files and mix them. We also average the predictions of several models to improve the result. Experimental results on the four cross-validation splits for the TAU Spatial Sound Events 2019-Microphone Array dataset are error rate: 0.23, F score: 85.91%, DOA error: 3.62°, and frame recall: 88.66%, respectively.

***Index Terms*—** Sound activity detection, sound event detection, direction of arrival estimation, convolutional neural network, recurrent neural network

1. INTRODUCTION

This paper was intended to explain our algorithm for DCASE 2019 challenge's Task 3. The task 3 of the DCASE 2019 challenge is 'Sound Event Localization and Detection (SELD) [1], which purpose is to detect and localize sound events. There are many studies for sound event detection (SED) which is to detect the onset and offset times for each sound event in an audio recording, such as clearing throat, coughing, human laughter, dog bark, and phone ringing. In SED task, there are two kinds of systems depending on whether the maximum number of sounds can be detected. One is called monophonic SED system which can only detect one sound at a time, the other is called polyphonic SED system which can detect multiple overlapping sound events at given time instance. Polyphonic SED task is more challenging than monophonic, but the performance improvements have been made since the multi label deep neural network (DNN) based SED has been first proposed [2]. Especially, recurrent neural network (RNN) and convolutional recurrent neural network (CRNN) based SED which can model the sequential information of the audio recording has shown good performance [3, 4, 5]. Recently, polyphonic SED systems using binaural [6] and spatial [7] features in multi-channel environment was proposed. In multi-channel environment, the SED systems can detect sound events more accurate then single-channel, and also can localize the sound events.

In this paper, we propose a three-stage approach for sound event localization and detection which consists of sound sound activity detection (SAD), sound event detection, sound event localization (SEL).

2. SOUND ACTIVITY DETECTION

To more accurately detect if sounds are present or not, we design the separate SAD model with multi-resolution cochleagram (MRCG) [8] and CRNN. We attempt two model, named SAD-01 and SAD-012. SAD-01 means that model can classify only sounds are present or not, and SAD-012 can classify sounds into three classes; no sounds, one sound, two sounds. In this task, the SAD 012 model is valid because maximum two sounds occur at the same time.

2.1. Data augmentation

Data augmentation is shared between SAD model and SED model. We used two methods, pitch shifting and block mixing on monophonic audio clip.[9]

2.2. Feature extraction

Fig. 1 shows the MRCG features that we extracted. This process was performed on each of the 4-channel audio signals.

2.3. Deep learning model

Fig. 2 shows the CRNN model for sound activity detection and sound event detection. Only the CNN block is different between the SAD and SED models. As shown as Fig. 3 (a), 3 x 1 convolution filter was used for SAD, considering the MRCG features of adjacent frames and max pooling was applied on frequency axis. For SAD-01, the sigmoid function was applied on the last fully connected layer, and softmax function was applied on the last fully connected layer for SAD-012.

3. SOUND EVENT DETECTION

The SED was performed independently, the results of SAD and SED results were separately estimated and merged at the end.

*Corresponding author.

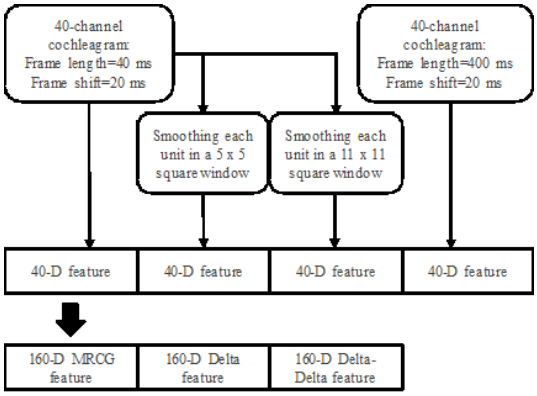

Figure 1: Multi-resolution cochleagram feature.

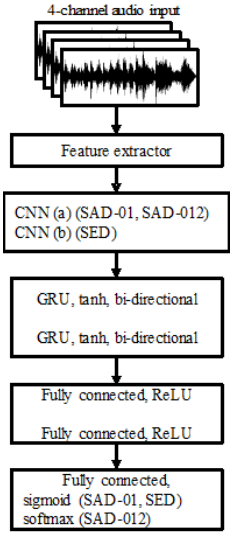

Figure 2: Convolutional recurrent neural network for sound activity detection and sound event detection.

3.1. Data augmentation

Data augmentation is shared between SAD model and SED model. We used two methods, pitch shifting and block mixing on monophonic audio clip.

3.2. Feature extraction

For SED, log mel-spectrograms were used as input feature. The log mel-spectrograms were extracted from total 7-channels audio signals; 4-channel input audio, harmonic and percussive components by harmonic-percussive source separation [10], mono audio averaged from 4-channel input audio.

3.3. Deep learning model

In SED, also the CRNN model was used in Fig. 2, only with different convolution block was used. The convolution block for SED is shown in Fig. 3 (b). Inspiring by VGGNet [11] and Inception V2 [12], 3 x 1 convolution filter was first performing and then 1 x 3

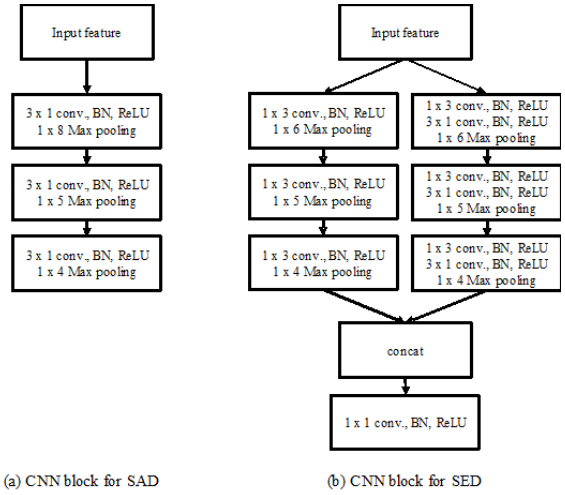

Figure 3: a) Convolution blocks for SAD and b) SED

convolution filter was performing on its output. And the last layer of the convolution block, 1 x 1 convolution was performing for reducing computational complexity and selecting an effective feature maps.

3.4. Combining SAD and SED results and post-processing

To improve system performance, we estimated the SAD and SED results respectively and combined them. First, we changed the SED results to zero when the SAD-01 results were zero for combining SAD-01 and SED. Second, for combining SAD-012 and SED, we extracted as zero, one, and two results from SED results according to the SAD results. Finally, median filter was applied to each class with different filter lengths.

4. SOUND EVENT LOCALIZATION

Recently, deep learning based sound source localization shows better performance than previous signal model-based such as multiple signal classification (MUSIC) [13]. Motivated of earlier works, we also use deep learning approach to estimate the DOA of the sound source. Although SELDnet [1] adapted multi-task learning to localize and detect each sound event jointly, we separate the SEL task from the SELD task to achieve better localization performance. Then, we formulate the DOA estimation problem as a multi-label classification for localizing multiple sound sources.

4.1. Feature extraction for SEL

Since the time-difference-of-arrival (TDOA) is one of the most widely used spatial cues for localizing a sound source, we decide to exploit the TDOA of all combinations of each microphone pair for estimating azimuth and elevation angle. More specifically, we calculate the generalized cross-correlation phase transform (GCC-PHAT) [14], which is the generally used feature for estimating TDOA between two microphones. Let X_i and X_j denote the spectrum of the i-th microphone and the j-th microphone. Then, GCC-

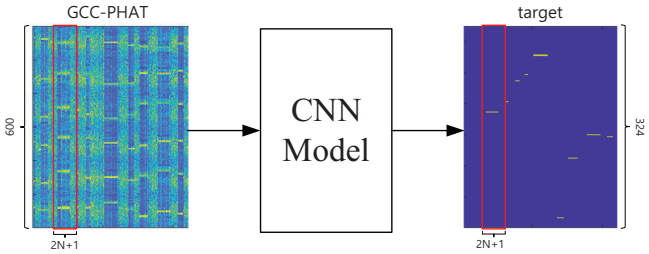

Figure 4: system overview for sound event localization

PHAT for the microphone pair iss calculated as follows:

$$\phi_{i,j}(n,\tau) = \sum_{k=1}^{K} \frac{X_i(k,n)X_j(k,n)}{|X_i(k,n)|\,|X_j(k,n)|} e^{-2j\pi \frac{f_s k}{K}\tau} \quad (1)$$

where k, n, f_s, and τ represent the frequency bin index, time frame index, sampling rate, and lag respectively. The TAU Spatial Sound Events 2019 dataset [15] was synthesized from estimated room impulse response recorded by using Eigenmike[1], which is a rigid spherical microphone array with radius 42mm. Considering the spacing between the microphones, We calculated the GCC-PHAT by dividing the lag from -50 ms to 50 ms by 100 in consideration of possible TDOA range. After calculating the GCC-PHAT for each microphone pairs, we concatenate them into the 600 (100×6) dimension vector, and adapt the min-max normalization for every frame as below:

$$\bar{\boldsymbol{\phi}}(n) = \frac{\boldsymbol{\phi}(n) - \min(\boldsymbol{\phi}(n))}{\max(\boldsymbol{\phi}(n)) - \min(\boldsymbol{\phi}(n))} \quad (2)$$

where $\boldsymbol{\phi}(n)$ is the GCC-PHAT vector. To consider context information, we splice N adjacent frames back and forth, and we experimentally observed that the more splicing gradually decrease the DOA error.

4.2. Model architecture

For localizing sound event, we use the CNN model, which is not only widely used in image classification but also DOA estimation [16, 17, 18]. The detailed structure of the CNN model is summarized in Table 1. The model is designed to classify DOA angles among the 324 (36×9) classes.

4.3. Data augmentation for SEL

In the experiment stage, we observed that higher DOA error in overlapped frames where two sound sources are present at the same time than the non-overlapped frame. Therefore, our augmentation strategy is to make more overlapped frames, and it is quite simple and effective. First, we randomly select two audio files in the train data of each split and add two audio files at random time frame intervals in the short-time Fourier transform (STFT) domain. Finally, we collect actual overlapped frame among the added frames.

[1]https://mhacoustics.com/products

Table 1: Specification of the CNN model.

Layer	Type	Kernel	Padding	Stride	Pooling	Units
1	Conv2D	32@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
2	Conv2D	32@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
3	Conv2D	32@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
4	Conv2D	64@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
5	Conv2D	128@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
6	Conv2D	256@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
7	Conv2D	256@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
8	Conv2D	256@3*3	SAME	-	-	-
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
	Maxpool	-	-	2*1	2*2	-
9	Flatten	-	-	-	-	-
	Dense	-	-	-	-	2048
	Batch norm	-	-	-	-	-
	ReLU	-	-	-	-	-
10	Dense	-	-	-	-	324
	Sigmoid	-	-	-	-	-

5. EXPERIMENTS AND RESULTS

5.1. Experimental setup

5.1.1. SAD

MRCG: MRCG features were extracted as shown as Fig. 1. This process was performed on each of the 4-channel audio signals, so total 480 x 4 dimension MRCG features were used.
Neural network configurations: The input sequence length was determined as 128, and batch size for training was 16. The number of cnn filter was 64 for every convolution layer, and max-pooling size of each layer was 8, 5, and 4, respectively. For two RNN layers, GRU sizes were 128 respectively. Finally, the number of nodes were 256 for two fully connected layers. For regularization, drop-out was applied to fully connected layers with drop rate 0.5. We trained the Model with binary cross-entropy loss function using Adam optimizer with default parameters.

5.1.2. SED

Log mel-spectrogram: We used log mel-spectrogram for SED. The log mel-spectrograms were extracted from total 7-channels audio signals; 4-channel input audio, harmonic and percussive components by harmonic-percussive source separation (HPSS). For log mel-spectrogram and HPSS, librosa which is a Python package was used. Window length and hop length were 40 ms and 20 ms respectively, and the number of mel filters was 240.
Neural network configurations: The input sequence length was de-

termined as 128, and batch size for training was 16. The number of CNN filter was 64 for every convolution layer, and max-pooling size of each layer was 6, 5, and 4, respectively. For two RNN layers, GRU sizes were 128 respectively. Finally, the number of nodes were 256 for two fully connected layers. For regularization, drop-out was applied to fully connected layers with drop rate 0.5. We trained the Model with binary cross-entropy loss function using Adam optimizer with default parameters.
Median filter length: The median filter lengths for clearthroat, cough, doorslam, drawer, keyboard, keysDrop, knock, laughter, pageturn, phone and speech are 17, 19, 5, 17, 31, 5, 11, 25, 21, 37, and 27, respectively. These were determined based on the mean and standard deviation of the sound length of each class.

5.1.3. SEL

To evaluate the localization performance only, we made the oracle SED results for the development set. Afterward we overwrite the estimates of the DOA from CNN model.
GCC-PHAT: We extract the GCC-PHAT feature vector every 20ms frame, and the number of spliced frame, N, was set from 5 to 13.
Model hyper-parameters: Because of the overlapped frames, we used the binary cross-entropy as a loss function, and Adam optimization method with a batch size of 128 frames and a learning rate of 0.001 was used to train the our CNN model.

5.2. Results

The proposed system was evaluated on both MIC (microphone array) and FOA (first-order ambisonic) datasets [15]. Each dataset consists of a pre-defined four cross-validation splits. There are four metrics to evaluate sound event localization and detection. For the SED, error rate (ER) and F-score is calculated in segments of one second [19]. For the SEL, two frame-wise DOA metrics were used to evaluate our localization performance: DOA error and frame recall described in [13].

The results of the integrated system with SAD-01 and with SAD-012 are shown in table 2 and table 3, respectively. In case of the integrated system with SAD-012, the system only achieved higher frame recall performance than system with SAD-01.

Table 2: Final results of the integrated system with SAD-01 on development dataset.

Dataset	MIC					FOA				
Split	1	2	3	4	Aver.	1	2	3	4	Aver.
ER	0.17	0.27	0.19	0.30	0.23	0.22	0.31	0.22	0.28	0.26
F-score	89.6	83.5	89.2	82.5	86.2	87.6	82.2	87.7	83.2	85.1
Frame recall	88.1	85.3	87.8	84.3	86.4	87.2	84.9	87.2	85.9	86.3
DOA error	3.70	3.61	3.88	3.28	3.62	9.43	9.84	9.78	9.10	9.55

6. CONCLUSION

In this paper, we proposed three-stage approach to localize and detect sound event using deep learning. SAD and SED results were estimated separately and combined for final SED results. For post-processing, median filter was applied to each class with different filter lengths. Finally, azimuth and elevation angle is estimated using the CNN model and integrated with the SED results.The model used in the final submission was the ensemble of the 4 models trained by 4 cross validation sets for each of the SAD, SED, and SEL models.

Table 3: Final results of integrated system with SAD-012 on development dataset.

Dataset	MIC					FOA				
Split	1	2	3	4	Aver.	1	2	3	4	Aver.
ER	0.22	0.35	0.22	0.30	0.28	0.25	0.25	0.23	0.28	0.25
F-score	87.3	79.9	87.9	81.9	84.2	85.8	84.1	87.0	83.1	85.0
Frame recall	92.5	90.6	92.4	91.8	91.8	92.9	93.1	92.3	92.9	92.8
DOA error	4.17	4.26	4.11	3.65	4.05	11.0	10.3	10.7	10.3	10.6

7. REFERENCES

[1] 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 2019.

[2] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[3] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6440–6444.

[4] R. Lu and Z. Duan, "Bidirectional gru for sound event detection," *Detection and Classification of Acoustic Scenes and Events*, 2017.

[5] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[6] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," *arXiv preprint arXiv:1710.02997*, 2017.

[7] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," *arXiv preprint arXiv:1706.02293*, 2017.

[8] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[10] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," *IEEE AASP Challenge on DCASE 2018 technical reports*, 2018.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] 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.

[13] 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.

[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[15] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *arXiv preprint arXiv:1905.08546*, 2019.

[16] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 2814–2818.

[17] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[18] S. Chakrabarty and E. A. P. 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)*, Oct 2017, pp. 136–140.

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