Sound Event Localization and Detection Using Convolutional Recurrent Neural Networks and Gated Linear Units

Tatsuya Komatsu
Research Labs
Line Corporation
Tokyo, Japan

Masahito Togami
Research Labs
Line Corporation
Tokyo, Japan

Tsubasa Takahashi
Research Labs
Line Corporation
Tokyo, Japan

***Abstract*—This paper proposes a sound event localization and detection (SELD) method using a convolutional recurrent neural network (CRNN) with gated linear units (GLUs). The proposed method introduces to employ GLUs with convolutional neural network (CNN) layers of the CRNN to extract adequate spectral features from amplitude and phase spectra. When the CNNs extract features of high-dimensional dependencies of frequency bins, the GLUs weight the extracted features based on the importance of the bins, like attention mechanism. Extracted features from bins where sounds are absent, which is not informative and degrade the SELD performance, are weighted to 0 and ignored by GLUs. Only the features extracted from informative bins are used for the CNN output for better SELD performance. Obtained CNN outputs are fed to consecutive bi-directional gated recurrent units (GRUs), which capture temporal information. Finally, the GRU output are shared by two task-specific layers, which are sound event detection (SED) layers and direction of arrival (DoA) estimation layers, to obtain SELD results. Evaluation results using the TAU Spatial Sound Events 2019 - Ambisonic dataset show the effectiveness of GLUs in the proposed method, and it improves SELD performance up to 0.10 in F1-score, 0.15 in error rate, 16.4° in DoA estimation error comparing to a CRNN baseline method.**

***Index Terms*—Sound Event Localization and Detection, Recurrent Convolutional Neural Network, Gated Linear Unit**

I. Introduction

Understanding the environment using sound is one of the important functions for home monitoring and advanced surveillance system [1]–[7] , and so on. In this field, lots of sound event detection (SED) methods, which identify types of the sound events, have been proposed. However, these methods are classification methods that only identify the class of the sound events. For monitoring and surveillance systems, not only the class of the sound events, but also the direction of Arrival (DoA) of the sound events is an important factor.

Recently, a new task, sound event localization and detection (SELD), has been launched at DCASE challenge 2019 [8], which is a combined task of SED and DoA estimation. More than 20 methods are proposed in this challenge, and the most popular choice of the method is a convolutional recurrent neural network (CRNN) [9]–[13]. [10] combines a CRNN and an additional network which estimates the number of active sound event to estimate DoA of overlapping sound events. [11], [12] employ 2 CRNNs which dedicated to SED and DoA estimation. [13] uses a new GCC-Phat-based feature for the input of the CRNN.

Almost all of these methods have the same CRNN structure as the DCASE challenge baseline method [9]. [9] uses amplitude and phase spectra as input and performs SED and DoA estimation simultaneously. For dealing only with amplitude spectra, the CRNN is one of the promising approaches as is shown that the CRNN-based SED method outperforms other neural network based models [14], [15]. However, for dealing with phase spectra, the CRNN cannot always be effective. It is due to CNN layers in the CRNN, which extract spectral features that represent high-dimensional dependencies of frequency bins. For example, in the case of amplitude spectra, frequency bins where sounds are absent are useful clues for characterizing and identifying sound events. On the other hand, in the case of phase spectra, frequency bins where sounds are absent are physically meaningless and completely unnecessary. In the CNN layers, the meaningless bins also can be treated as same as informative bins, which lead to degradation of the DoA estimation performance. Therefore, using phase spectra for the CRNN input in the same manner as amplitude spectra have an adverse effect. A method to select informative frequency bins for phase spectra is required.

To select informative input, gated linear units (GLUs) has been proposed [16] in the natural language processing field. GLUs are connected in parallel with each CNN layer to control the CNN outputs, like controlling a valve, instead of the activation function. The GLU is made of a CNN layer and trained to weight the connected CNN output by importance of input, like an attention mechanism. The effectiveness of GLUs has been also shown for modeling of amplitude spectra, such as audio source separation [17] and weakly labeled training [18]. However, the effectiveness of GLUs on phase spectra has not been studied and clarified yet.

This paper proposes a CRNN-based SELD method using GLUs. GLUs can learn which frequency bins of amplitude and phase spectra have important information for DoA estimation and SED, and output features extracted only from informative input.

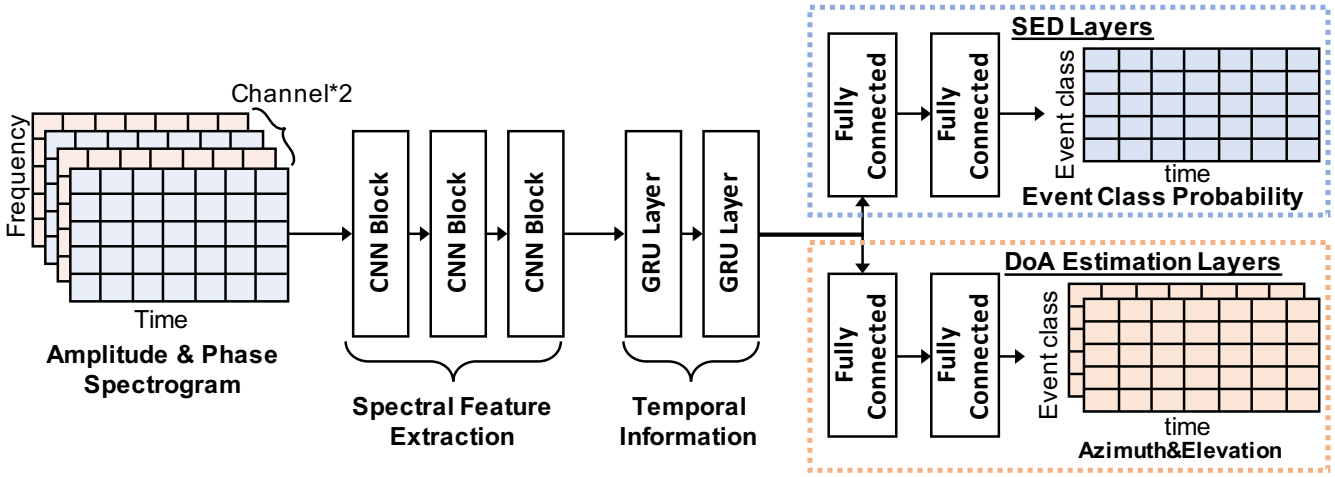


Fig. 1. CRNN-based SELD method. Both the baseline method and the proposed method employ this structure. The only difference is in the CNN blocks, that the baseline method uses CNNs with ReLU activation while the proposed method uses GLUs for them.

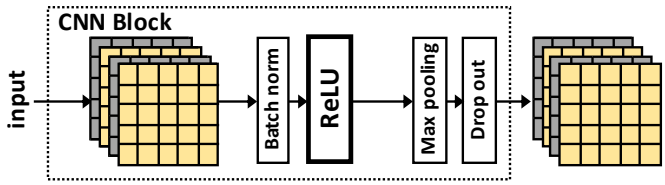


Fig. 2. CNN block in the baseline method

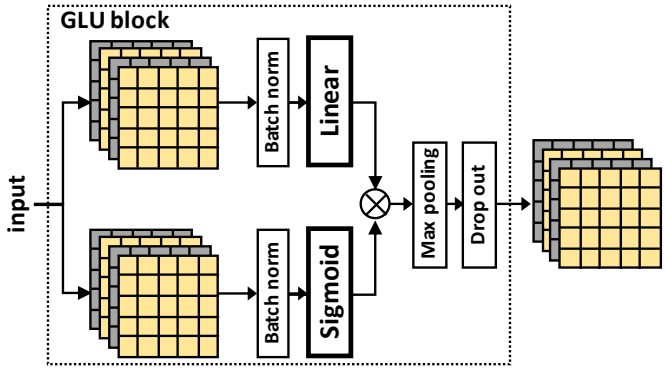


Fig. 3. GLU which replaces the CNN block in the CRNN-based SELD method

II. The Proposed SELD Method with GLUs

In this section, we describe a standard CRNN-based method [9] which is a baseline method of DCASE 2019 SELD task and the proposed CRNN-based SELD method using GLUs. The structures of two methods are almost the same. The only difference is that the baseline method employs standard CNNs with ReLU activation for spectral feature extraction while the proposed method employs CNN layers with GLUs. It should be noted that the structure of CRNN to use GLUs is not limited to this baseline. Any other SELD methods which have CNN layers can be adopted.

A. The CRNN baseline method

The CRNN based baseline method [9] and CNN blocks are illustrated in Fig. 1 and Fig. 2. Input of the method is time-frequency (T-F) representation, which consists of amplitude and phase spectrograms extracted from multi-channel audio recordings. The T-F representations are fed to CNN layers and the CNN layers extract features that represent high-dimensional dependencies of each frequency bins. The extracted features are then sent to bi-directional gated recurrent units (GRUs) to capture temporal context information. The outputs of the recurrent layers are shared as input of consecutive two layers, the SED layers and the DoA estimation layers. The SED layers output probabilities of event classes in each time frame of the input T-F representation. The DoA estimation layer outputs event-class-wise azimuth and elevation angels in each time frame. Training of the network parameters is performed by multi-task learning that minimizes the weighted sum of SED loss and DoA estimation loss. For SED loss, multi-class binary cross entropy between the ground truth and the probability of each event-class, for DoA loss, multi-output regression loss defined by the squared distance on the spherical surface between the estimated angle and the grand truth are used.

It has been found that CNNs, which capture the dependencies of each frequency bins, are effective methods for characterizing structures of amplitude spectra [17], [18]. However, in the case of phase spectra, CNNs may not always be effective. For DoA estimation, frequency bins, in phase spectra, where sounds are absent are completely unnecessary and physically meaningless. Therefore, using the CRNN for DoA estimation in the same manner as acoustic event detection may have an adverse effect.

B. The proposed method

The structure of the proposed method is almost same as the baseline method in Fig. 1. To select informative bins

for spectral feature extraction, we propose to use GLUs in Fig. 3 instead of CNN blocks with ReLU activation. GLUs are defined as:

$$\mathbf{Y} = (\mathbf{W} * \mathbf{X} + b) \odot \sigma(\mathbf{V} * \mathbf{X} + c) \quad (1)$$

where $\mathbf{X}$ and $\mathbf{Y}$ represent input and output to GLUs, respectively. $\mathbf{W}$ and $\mathbf{V}$ are the convolutional filters, b and c are the biases. $\odot$ denotes the element-wise product and $*$ is the convolution operator. $\sigma(\cdot)$ is the sigmoid function which works as a gating function of GLUs.

The sigmoid function works to control information as: If $\mathbf{X}$ is informative input, the sigmoid function makes output close to 1, and hence the extracted feature is used as output. If $\mathbf{X}$ is not informative input, the sigmoid function makes output close to 0, the extracted feature is ignored. In this way, GLUs weight the extracted feature by the importance of corresponding input as an attention mechanism. Using the GLUs, CNNs can extract feature from phase spectra properly. GLUs ignore the frequency bins where sounds are abset, and the CNN feature extractor uses only bins including essential information.

III. Experimental Evaluation

Experimental evaluation is conducted to measure the effectiveness of GLUs of the proposed SELD method. A method for comparison is the baseline method. The only difference between the baseline method and the proposed method is whether to use CNNs with ReLU activation in Fig. 2 or GLUs inFig. 3. All parameters for both methods are set to the same values.

A. Datasets and parameters

The TAU Spatial Sound Events 2019 - Ambisonic dataset [19] is used for the experiment. The summary of the dataset is shown in Table I. The dataset consists of 400 files of four-channels first-order-ambisonics audio recording. The sampling rate for each recording is 48000 Hz and length of each recording is 1 minute. Sound event classes in the dataset is 11 classes of events. The number of azimuth and elevation angles of sound source direction is 36 with $10°$ interval from $-180°$ to $180°$ and 9 with $10°$ interval from $-40°$ to $40°$, respectively. Pre-defined four cross-validation setups (*split 1, split 2, split 3, split 4*) are also provided with audio recordings. Each setup assigns 200 recordings for training, 100 recordings for validation and 100 recordings for testing. In this paper, experiments are conducted based on the provided cross-validation setups.

For input of the neural network, T-F representations are obtained by short time Fourier transform with 2048 sample points, 40 ms window length (1920 sample points) and 20 ms hop length (960 sample points). The network parameters of the proposed method are shown in Table II. The input shape (32, 1024, 128, 8) corresponds to (mini-batch size, frequency bins, time frames, channels (amplitude and phase)). 2D CNN filter = (3,3,6) denotes a 2D CNN layer with (3, 3) filter size and 64 channels. Pool size (1, 8) denotes max pooling operation

TABLE I
TAU Spatial Sound Events 2019 - Ambisonic

Sampling frequency	48,000 [Hz]
number of audio recordings	400 recordings (1 min. per recording)
Event classes	clearthroat, keyboard, pageturn, cough, keysDrop, phone, doorslam, knock, laughter, drawer, speech
azimuth angles	36 angles with $10°$ interval from $-180°$ to $180°$
elevation angles	9 angles with $10°$ interval from $-40°$ to $40°$

TABLE II
Neural network architecture of the proposed method

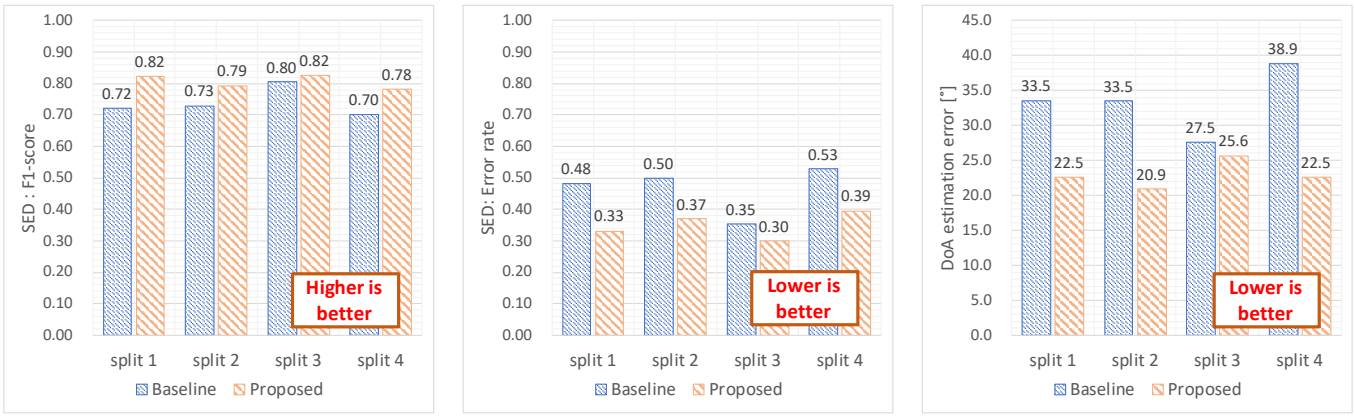
Input: shape = (32, 1024, 128, 8)	
2D CNN filter = (3, 3, 64) Batch normalization	2D CNN filter = (3, 3, 64) Batch normalization
Linear	Sigmoid
Multiply Max pooling: pool size (1, 8)	
2D CNN filter = (3, 3, 64) Batch normalization	2D CNN filter = (3, 3, 64) Batch normalization
Linear	Sigmoid
Multiply Max pooling: pool size (1, 8)	
2D CNN filter = (3, 3, 64) Batch normalization	2D CNN filter = (3, 3, 64) Batch normalization
Linear	Sigmoid
Multiply Max pooling: pool size (1, 4)	
Bi-directional GRU 128 nodes	
Bi-directional GRU 128 nodes	
Fully connected layer 128 nodes	Fully connected layer 128 nodes
Fully connected layer 22 nodes	Fully connected layer 11 nodes
Linear	Sigmoid
DoA estimation output (22, 128)	SED output shape (11, 128)

is applied to 8 bins on frequency axis. The number of epochs is 250. These network parameters including the number of epochs are almost same as the default parameter of the baseline method described in [9]. Adam [20] is used as the stochastic optimization method.

Experiments results are evaluated using the following three metrics; For SED metrics, F1-score and error rate (ER) are calculated in one-second segments [21]. For DoA estimation error, the average angular error in degrees between the predicted DoAs and true DoAs [22] are used.

B. Experimental results

Figs. 4-(a), (b) and (c) show F1-scores and ERs of SED metric and DoA estimation error, respectively. For all cross-validation setups, the proposed method with GLU has been



(a) F1-score of SED (b) Error rate of SED (c) DoA estimation error

Fig. 4. Experimental results

shown superior performance. At the maximum, 0.10 in F1-score, 0.15 in ER, 16.4° in DoA estimation error improvements have been shown. In particular, the higher the improvement in error rate, the higher the improvement in DoA estimation performance. It is considered that GLUs of the proposed method are able to appropriately select informative frequency bins for DoA estimation.

The proposed SELD framework with GLUs can be applied to any kind of CNN-based SELD method, not limited to baseline. Almost all methods proposed in DCASE2019 Challenge including top-ranked methods employ CNNs but they do not have any kind of gated or attention mechanism. So, the proposed SELD framework with GLUs are easily applied to any other CNN-based SELD method and there can be improvement in SED and DoA estimation performance.

IV. CONCLUSION

This paper has proposed a sound event localization and detection (SELD) method using a convolutional recurrent neural network (CRNN) and gated linear units (GLUs). GLUs are connected in parallel with each CNN layer to control the CNN outputs instead of the activation function. When the CNNs extract features of high-dimensional dependencies of frequency bins, the GLUs weight the extracted features based on the importance of the CNN input, like attention mechanism. Frequency bins where sounds are absent are ignored by GLUs and only bins with essentially information are used by the CNN feature extractor. Evaluation results using the TAU Spatial Sound Events 2019 - Ambisonic dataset has shown the effectiveness of GLUs in the proposed method, and it improves SELD performance up to 0.10 in F1-score, 0.15 in error rate, 16.4° in DoA estimation error comparing to a CRNN baseline system.

REFERENCES

[1] K. Imoto, "Introduction to acoustic event and scene analysis," *Acoustical Science and Technology*, vol. 39, no. 3, pp. 182–188, 2018.

[2] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 2009, pp. 1218–1221.

[3] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 158–161.

[4] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 4–pp.

[5] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 165–168.

[6] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 45–49.

[7] T. Komatsu, M. Tani, T. Toizumi, N. Chaitanya, M. Kato, Y. Arai, O. Hoshuyama, Y. Senda, and R. Kondo, "An acoustic monitoring system and its field trials," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1341–1346.

[8] http://dcase.community/workshop2019/.

[9] 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*, pp. 1–1, 2018.

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

[11] Y. Cao, T. Iqbal, Q. Kong, M. Galindo, W. Wang, and M. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," DCASE2019 Challenge, Tech. Rep., June 2019.

[12] J. Zhang, W. Ding, and L. He, "Data augmentation and prior knowledge-based regularization for sound event localization and detection," DCASE2019 Challenge, Tech. Rep., June 2019.

[13] H. C. Maruri, P. L. Meyer, J. Huang, J. A. D. H. Ontiveros, and H. Lu, "Gcc-phat cross-correlation audio features for simultaneous sound event localization and detection (seld) on multiple rooms," DCASE2019 Challenge, Tech. Rep., June 2019.

[14] E. Cakır and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[15] 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.

[16] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 933–941.

[17] L. Li and H. Kameoka, "Deep clustering with gated convolutional networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 16–20.

[18] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 121–125.

[19] 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

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

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

[22] 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.