Two-stage Polyphonic Sound Event Detection Based on Faster R-CNN-LSTM with Multi-token Connectionist Temporal Classification

Inyoung Park[1] *and Hong Kook Kim*[1,2]

[1]AI Graduate School, [2]School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology, Gwangju 61005, Korea
{pinyoung, hongkook}@gist.ac.kr

Abstract

We propose a two-stage sound event detection (SED) model to deal with sound events overlapping in time-frequency. In the first stage which consists of a faster R-CNN and an attention-LSTM, each log-mel spectrogram segment is divided into one or more proposed regions (PRs) according to the coordinates of a region proposal network. To efficiently train polyphonic sound, we take only one PR for each sound event from a bounding box regressor associated with the attention-LSTM. In the second stage, the original input image and the difference image between adjacent segments are separately pooled according to the coordinate of each PR predicted in the first stage. Then, two feature maps using CNNs are concatenated and processed further by LSTM. Finally, CTC-based n-best SED is conducted using the softmax output from the CNN-LSTM, where CTC has two tokens for each event so that the start and ending time frames are accurately detected. Experiments on SED using DCASE 2019 Task 3 show that the proposed two-stage model with multi-token CTC achieves an F1-score of 97.5%, while the first stage alone and the two-stage model with a conventional CTC yield F1-scores of 91.9% and 95.6%, respectively.

Index Terms: polyphonic sound event detection (SED), faster regional convolutional neural network (R-CNN), multi-token connectionist temporal classification (Multi-token CTC), attention long short-term memory (attention-LSTM)

1. Introduction

Sound event detection (SED) for surveillance applications requires high accuracy with real-time constraints [1]. This is due to the contextual characteristic that a single missing detection can lead to a major accident. There have been several studies addressing the challenge of sound source localization or event detection. For example, in the series of IEEE AASP Challenges on Detection and Classification of Acoustic Scenes and Events (DCASE), SED mainly focused on detecting sound events in the home or domestic environments [2,3]. However, for surveillance applications such as a drone-based rescue system, SED model is employed for detecting cries for help from targets (e.g., “help me” or “SOS”) as well as weapon fire or voices from fields. Such sound could occur as a sequence of single events or mixtures of multiple events. The accurate determination of each event can help to effectively distribute scarce human resources to rescue injured people. As such, higher-accuracy SED with a lower processing delay is of great importance for saving lives within the window of opportunity. In practice, multiple sound events occur with some degree of overlap. Moreover, such polyphonic SED should deal with sound events that come from different locations with different distances, azimuths, and elevations.

There have been many research works proposing polyphonic SEDs, such as feedforward neural networks (FNNs) [4], convolutional neural networks (CNNs) [5], and recurrent neural networks (RNNs) [6]. These approaches have tried to predict single or multiple sound events for every audio frame or to predict events for each audio clip. To train neural network models, each audio clip is divided into consecutive segments of a spectrogram. Then, each segment is labeled as one or more target events depending on the overlap of events in polyphonic SED. Therefore, some segments are trained with a single target label but others with multiple target labels. This results in degrading the performance of polyphonic SED rather than monophonic SED [7].

Thus, SED based on connectionist temporal classification (CTC), which is a sequence-to-sequence model, has been proposed to detect polyphonic events [8]. In the CTC-based SED, each sound event is attached with a blank token, thus the total number of tokens is twice the number of sound events, where overlapping sound events are also allowed for each segment [9]. If conventional CTC works for character detection, CTC detects an event where it remains at the corresponding event token until it enters a blank token [10]. This approach shortens the event duration, which can be overcome by setting a hinting tolerance that corresponds to a predefined minimum duration [9]. This method is a simple and powerful solution; however, the event length varies in surveillance environments. The assumption of such a minimum duration may jeopardize the detection of short-term events such as gunshot sounds.

We propose mitigating this problem with a new two-stage polyphonic SED model. The first stage of the proposed SED is composed of a faster regional CNN (R-CNN) [11] and an attention-LSTM. For the faster R-CNN, hand-labeled time intervals are used as the ground truth for the region proposal network (RPN). Thus, each log-mel spectrogram segment is first decomposed in the time-frequency (T-F) domain into proposed regions by the RPN. The second stage of the proposed SED consists of two CNNs combined with one LSTM for feature representation, and a softmax layer combined with CTC for event detection and classification. In this stage, the proposed regions predicted from the first stage are region-of-interest (RoI) pooled [12]. In parallel, the difference spectrogram image between adjacent segments is also pooled according to the predicted regions. Next, two feature maps constructed by CNNs are obtained by using the RoI-pooled static and difference images. After that, these feature maps are concatenated and fed into an LSTM layer.

Finally, CTC-based n-best SED is conducted using the softmax output connected from the CNN-LSTM. Compared to

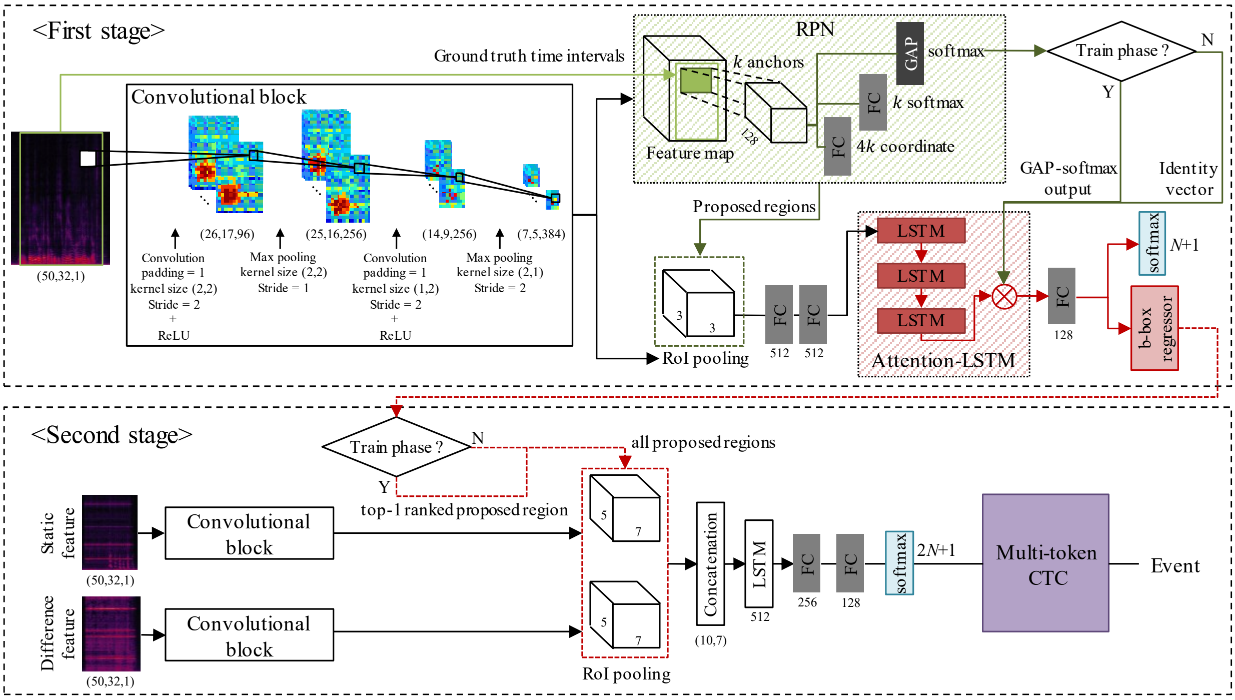


Figure 1: *Block diagram of the proposed two-stage polyphonic SED model based on a faster R-CNN and CNN-LSTM with multi-token CTC.*

the previous approach in [9], we change the CTC by assigning the start and ending token for each event as well as one blank token shared with all the events. Thus, an event is detected when the ending token is passed into the shared blank token after entering the start token of the event.

The remainder of this paper is organized as follows. Section 2 proposes the two-stage polyphonic SED and describes the procedure of each stage in detail. Then, Section 3 describes the experimental setup including datasets and model training. Next, Section 4 evaluates the performance of the proposed SED model on the event detection of DCASE 2019 Task 3 [13] and compares it with those of the first stage and the two-stage SED with a conventional CTC proposed in [9]. Finally, Section 5 concludes the paper.

2. Proposed two-stage polyphonic SED

Figure 1 shows a block diagram of the proposed two-stage SED model. First of all, each audio clip sampled at 16 kHz is divided into consecutive frames of 32 ms (= 512 samples) with 22-ms overlap between the frames. Then, a 50-dimensional mel-filterbank analysis is performed and the consecutive 32 frames are combined into one segmen t of log-mel spectrogram once every frame, which results in one (50×32) image per frame. Detailed network architecture and processing for each stage will be discussed in the following subsections.

2.1. Training phase

2.1.1. First stage: proposed region selection of multiple sound events using faster R-CNN-LSTM

This stage provides a proposed region for each of multiple sound events by using a faster R-CNN and a bounding box (b-box) regressor based on LSTM. To this end, a feature map is first obtained using four convolutional layers with max pooling, where the numbers of kernels for each convolutional layer are 96, 256, 256, and 384 with sizes of (2×2), (2×2), (1×2), and (2×1), respectively, and with stride sizes of 2, 1, 2, and 2, respectively, which is set through exhaustive experiments. Next, this feature map is used as an input to an RPN that has nine different anchor boxes whose shapes are {1:1, 1:2, 2:1} with the three anchor sizes of 1, 2, and 4 [14]. Note here that the hand-labeled time intervals corresponding to sound events are used as the ground truth for the RPN. Thus, we have (7×5×9) anchors in total; we then apply the non-maximum suppression (NMS) [15] to remove highly overlapped anchors with a ground truth box. The 128-dimensional feature of each anchor at each center location of the sliding window is fed to two dense layers to predict the probability of having an event and to encode the coordinates of regions proposals, respectively. In addition, the 128-dimensional features collected for positive anchors are processing by global average pooling (GAP). Note that the min-batch size of training RPN is set to 128 by taking into account the size of the CNN feature map.

Next, the proposed regions predicted by the RPN are RoI-pooled and passed to an attention-LSTM to classify the event using a softmax layer as well as to provide the coordinate of each proposed region. The attention-LSTM is composed of three LSTM layers with 128 hidden nodes each, where an attention mechanism is applied to the last LSTM hidden vector with the output of the GAP layer in the RPN. In particular, we take the top-1 ranked proposed region that shows the highest probability among all the proposed regions from the b-box regressor connected to the attention-LSTM. Thus, we have two proposed regions for the image from the polyphonic sound with a mixture of two sound events.

2.1.2. Second stage: CNN-LSTM-based feature representation and CTC-based event detection

This stage represents a CNN-LSTM-based feature map and detects the time intervals of multiple sound events using CTC. Here, we use a multi-feature combination approach to make the SED robust to the power variation of incoming sound events due to different locations of a sound source. To this end, the original log-mel spectrogram image is first cropped according to the coordinates of each of the proposed regions from the first stage, and then the cropped images are RoI-pooled and concatenated into a size of (50 × 32). Notice that only one cropped image is pooled for the original image corresponding to a monophonic sound. In parallel, the difference spectrogram image between adjacent segments of log-mel spectrogram is computed, and then it is processed by the same procedure as described above.

Next, two CNN feature maps are obtained by applying the convolution block to the RoI-pooled static and difference images, respectively, and then they are concatenated into a (10×7×384) map. After that, an LSTM layer with 512 hidden nodes, two FC layers, and a softmax layer are sequentially applied to the map. Finally, we have CTC that is designed to have (2*N*+1) tokens corresponding to one shared blank token and two tokens for *N* sound events with one start and one ending token each.

2.2. Evaluation phase

In the evaluation phase, each audio clip is also represented by a series of spectrogram images as in the training phase, and each image is processed by the faster R-CNN. Compared to the training phase, there are two main differences in this evaluation phase:

- The attention for LSTM is done using an identity vector for the evaluation phase, while the GAP layer output of the RPN is used for the training phase.
- All the proposed regions from the b-box regressor in the LSTM are taken in the first stage for the evaluation phase, while only one top-1 ranked proposed region is brought to the second stage of the training phase.

Except for them above, the second stage of the evaluation phase is identical to that of the training phase.

2.3. Multi-token connectionist temporal classification

In this paper, the CTC has (2*N*+1) tokens as mentioned in Section 2.1.2, where a blank token is shared by all the events, and the start and ending time for one event are represented by different tokens, which is here referred to as multi-token CTC. The objective function for CTC is based on the total probability of the token sequence by summing over all possible alignments. Let l_i be the *i*-th token element of a (2*N*+1)-dimensional token vector, where l_1 is the shared blank token and l_{2s} and l_{2s+1} are the start and ending token, respectively, for the *s*-th sound event. In addition, let $\alpha_t(i)$ be the total probability of partial paths landed on l_i at time frame t. Then, the forward algorithm in the proposed multi-token CTC is defined as follows:

$$\alpha_1(i) = \begin{cases} y_1(l_i) & if\ i = 1\ or\ even \\ 0 & otherwise \end{cases} \tag{1}$$

$$\alpha_t(i) = \begin{cases} [\alpha_{t-1}(1) + \sum_{i=1}^{N} \alpha_{t-1}(2i+1)]\, y_t(l_i) \\ \qquad\qquad if\ i = 1\ \text{(blank token)} \\ [\alpha_{t-1}(1) + \alpha_{t-1}(i) + \sum_{i=1}^{N} \alpha_{t-1}(2i+1)] y_t(l_i) \\ \qquad\qquad if\ i = even\ (start\ token) \\ [\alpha_{t-1}(i-1) + \alpha_{t-1}(i)]\, y_t(l_i) \\ \qquad\qquad otherwise\ (ending\ token) \end{cases} \tag{2}$$

where $y_t(l_i)$ corresponds to the softmax layer output in the second stage.

In this paper, the proposed multi-token CTC finds the first best path from the end of the audio clip to the beginning, then an event is detected from the time frame when entering the start token of the event, to the time frame when the ending token is passed into the shared blank token. Next, this best path is eliminated for the next-best path search. This is repeated until the subsequent next-best path is composed of all the blank tokens.

3. Experimental setup

3.1. Dataset

The proposed SED model was applied to the DCASE 2019 Challenge Task 3. This task was about sound event localization and detection, but we only took the dataset for SED. Especially, the TAU Spatial Sound Events 2019 - Ambisonic (FOA) dataset was used for the experiment. Such SED task consisted of 11 isolated sound events (clearing throat, coughing, door knock, door slam, drawer, human laughter, keyboard, keys [put on table], page turning, phone ringing, and speech) and synthetic mixtures of the same examples in multiple signal-to-noise ratios (SNRs), event density conditions, and polyphony [16].

To evaluate the performance of the proposed SED model, the FOA dataset was divided into a development set and an evaluation set. The development set consisted of 400 audio clips whose length was one minute long at a sampling rate of 48 kHz. Then, each audio clip was down-mixed into a single channel audio by using Adobe Audition 3.0, which was further down-sampled to 16 kHz. Next, the development set was divided into four subsets for four cross-validations. In addition, the evaluation set consisted of 100 audio clips one minute long each. All the audio clips were preprocessed by down-mixing followed by down-sampling as in the development set.

As performance measures, F1-score and event-based error rate (ER) were used, and they were measured by the sed_eval toolbox [17] that was provided by the Challenge [18]. The F1-score and ER regarding the development set were averaged over four cross-validations.

3.2. Model training

Each model layer shown in Figure 1 was implemented by the deep learning package in Python 3.5.2 with TensorFlow 1.8.0. All the experiments were conducted on an Intel Core i7-7700 workstation with an NVidia GTX 1080ti GPU.

The adaptive moment estimation (Adam) optimization [19] was utilized for the backpropagation algorithm with a learning rate of 0.01. Dropout was applied by a rate of 0.25, and the rectified linear unit (ReLU) was used as an activation function. Consequently, the model size of the first stage of the proposed SED model was around 25 MB, and the model size in total was around 34 MB.

4. Performance evaluation

First, the effect of the attention mechanism applied to the LSTM in the first stage was examined. Table 1 compares F1-scores and ERs of the first stage of the proposed SED when the LSTM was used for the b-box regressor and classifier with/without attention. As shown in the table, the attention mechanism increased F1-score and decreased ER in the first stage. Moreover, the first stage only provided better performance than the baseline of DCASE 2019 Task 3, which will be discussed below.

Next, the proposed two-stage SED model was evaluated according to different types of features, such as static, difference, and combination of static and difference features. Note here that the SED was performed by only using the

Table 1: Performance comparison of F1-score (%) and ER (%) for the first stage of the proposed SED model without or with attention.

Model	Development		Evaluation	
	F1-score	ER	F1-score	ER
First stage w/o attention	87.5	0.19	88.1	0.18
First stage w/ attention	89.3	0.17	91.9	0.14

Table 2: Performance comparison of F1-score (%) and ER (%) according to different types of features in the second stage of the proposed SED model, where CTC is not applied.

Features	Development		Evaluation	
	F1-score	ER	F1-score	ER
Static feature	90.1	0.17	92.8	0.13
Difference feature	90.2	0.17	93.2	0.12
Multi feature combination	91.3	0.15	94.1	0.10

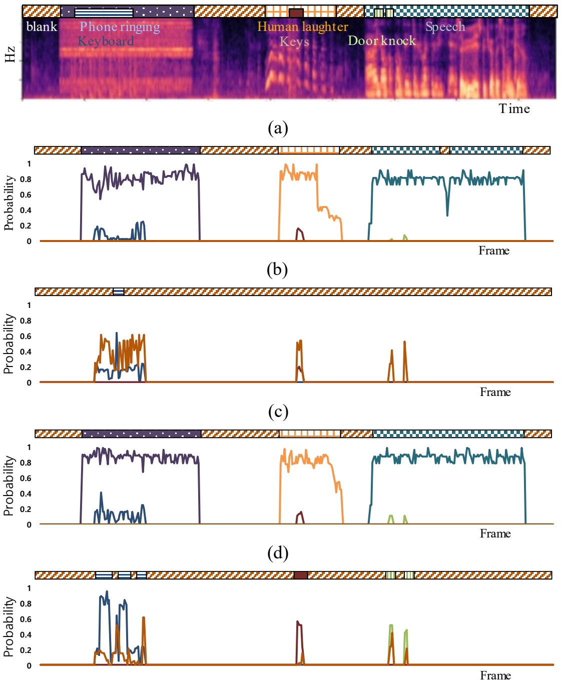
Table 3: Performance comparison of F1-score (%) and ER (%) between the baseline, the top-ranked models from the DCASE 2019 Task 3, and the proposed SED model with a conventional and the proposed multi-token CTC.

Features	Development		Evaluation	
	F1-score	ER	F1-score	ER
DCASE2019 Task 3 baseline [16]	79.9	0.34	85.4	0.28
Challenge Participant: Multi-task learning SED [20]	93.4	0.11	96.3	0.06
Challenge Participant: PKR-based CRNN [21]	91.6	0.14	96.7	0.06
Proposed two-stage SED with conventional CTC (no hinting)	91.6	0.14	94.9	0.10
Proposed two-stage SED with conventional CTC (hinting tolerance k=5)	91.9	0.13	95.6	0.08
Proposed two-stage SED with multi-token CTC	93.2	0.11	97.5	0.05

softmax layer output without any CTC. As shown in Table 2, the proposed SED using the difference feature provided similar to or slightly better performance than that using the difference feature for both development and evaluation dataset. Interestingly, the multi-feature combination achieved the best performance among them.

Finally, the performance of the proposed SED model was evaluated and compared with those of 1) the baseline and 2) the two top-ranked models reported in the DCASE 2019 Challenge Task 3. Also, two different versions of the proposed SED model were implemented according to the CTC used: a conventional CTC employing one token for each event followed by a blank each [9], and the proposed CTC employing multi-token tokens for each event. As shown in Table 3, the proposed two-stage SED model with the conventional CTC achieved similar F1-scores and ERs to the top-ranked challenge SED models. However, among all the comparatives, the proposed two-stage SED model with multi-token CTC significantly increased F1-scores and reduced ERs for development and evaluation datasets.

Figure 2 illustrates the frame-level detection results between the conventional CTC with hinting tolerance of k=5 and the proposed multi-token CTC. For an input log-mel spectrogram of an audio clip given in Figure 2(a), the first path from multi-token CTC (Figure 2(d)) showed similar detection accuracy to that from the conventional CTC (Figure 2(b)). However, the second path results differed between the two CTCs. In particular, as shown in Figure 2(c), the conventional CTC mis-detected "keyboard" as "blank", but the proposed multi-token CTC correctly detected, as shown in Figure 2(e). Consequently, it could be concluded here that the proposed CTC could outperform the conventional one for polyphonic SED.



(a)

(b)

(c)

(d)

(e)

Figure 2: *Comparison of decoding paths between the conventional CTC with hinting tolerance of 5 and proposed multi-token CTC for a given audio log-mel spectrogram (a); (b) the first-best path and (c) the second-best path for the conventional CTC; (d) the first-best path and (e) the second-best path for the multi-token CTC.*

5. Conclusion

In this paper, we proposed a two-stage polyphonic SED model based on a faster R-CNN-LSTM and a CNN-LSTM-CTC for the first and second stages, respectively. In order to efficiently train the second stage, the number of proposed regions from the faster R-CNN-LSTM was restricted to the number of events for each log-mel spectrogram segment. Moreover, the second stage was mainly composed of CNN-LSTM-based feature representation and CTC-based detection, where CTC had multi-tokens for each sound event and the intervals of sound events were obtained by an n-best search. The performance of the proposed SED model was evaluated on the SED dataset of the DCASE 2019 Challenge Task 3. As a result, it was shown that the proposed two-stage SED model with multi-token n-best CTC outperformed the state-of-art SED models reported in the Challenge. Moreover, the multi-token CTC provided higher F1-scores and lower ER than the conventional CTC with hinting tolerance.

As future works, we will try to extend the proposed SED model, especially CTC, to accommodate the duration constraints of polyphonic sounds, and we will also try to incorporate the GAP-softmax layer output of the first stage into the second stage of the proposed SED model.

6. Acknowledgements

This work was supported in part by the IITP grant funded by the Korea government(MSIT) (No. 2019-0-01767, Development of Machine Learning-Based Acoustic Intelligence Technology for Disaster Response Using Drones) and (No. 2019-0-01842, Artificial Intelligence Graduate School (GIST)).

7. References

[1] M. H.-Y. Liao, D.-Y. Chen, C.-W. Sua, and H.-R. Tyan, "Real-time event detection and its application to surveillance systems," in *ISCAS 2006 – IEEE International Symposium on Circuit and Systems, May 21-24, Island of Kos, Greece, Proceedings,* 2006, pp. 4–512.

[2] DCASE 2018 challenge website, Online on: http://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments/ (accessed on 01 May 2020).

[3] DCASE 2019 challenge website, Online on: http://dcase.community/challenge2019/task-sound-event-detection-in-domestic-environments/ (accessed on 01 May 2020).

[4] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *IJCNN 2015 – International Joint Conference on Neural Networks, July 11-16, Killarney, Ireland*, *Proceedings,* 2015, pp. 1–7.

[5] E. Cakir, E. Ozan, and T. Virtanen, "Filterbank learning for deep neural network based polyphonic sound event detection," in *IJCNN 2015 – International Joint Conference on Neural Networks, July 24-29, Vancouver, Canada*, *Proceedings,* 2016, pp. 3399–3406.

[6] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *ICASSP 2016 – IEEE International Conference on Acoustics, Speech and Signal Processing, March 20-25, Shanghai, China, Proceedings*, 2016, pp. 6440–6444.

[7] Z. Xu, I. W. Tsang, Y. Yang, Z. Ma, and A. G. Hauptmann, "Event detection using multi-level relevance labels and multiple features," in *CVPR 2014 – IEEE Conference on Computer Vision and Pattern Recognition, June 24-27, Columbus, Ohio, Proceedings,* 2014, pp. 4321–4328.

[8] T. Matsuyoshi, T. Komatsu, R. Kondo, T. Yamada, and S. Makino, "Weakly labeled learning using BLSTM-CTC for sound event detection," in *APSIPA 2018 Asia-Pacific Signal and Information Processing Association, November 12-15, Honolulu, HI, Proceedings,* 2018, pp. 1918–1923.

[9] Y. Wang and F. Metze, "A first attempt at polyphonic sound event detection using connectionist temporal classification," in *ICASSP 2017 – IEEE International Conference on Acoustics, Speech and Signal Processing, March 5-9, New Orleans, LA, Proceedings,* 2017, pp. 2986–2990.

[10] Y. Wang and F. Metze, "Connectionist temporal localization for sound event detection with sequential labeling, " in *ICASSP 2019 – IEEE International Conference on Acoustics, Speech and Signal Processing, May 12-17, Brighton, UK, Proceedings*, 2019, pp. 745–749.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *NIPS 2015 – Advances in Neural Information Processing Systems 28, December 7-12, Montreal, Canada, Proceedings,* 2015, pp. 1–9.

[12] R. Girshick, "Fast R-CNN," arXiv:1504.08083, 2015.

[13] DCASE 2019 challenge Task 3 Description, Online on: http://dcase.community/challenge2019/task-sound-event-localization-and-detection#description/ (accessed on 20 April 2020).

[14] C.-C Kao, W. Wang, M. Sun, and C. Wang, "R-CRNN: region-based convolutional recurrent neural network for audio event detection," in *Interspeech 2018 – Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings,* 2018, pp. 1358–1362.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR 2014 – IEEE Conference on Computer Vision and Pattern Recognition, June 24-27, Columbus, Ohio, Proceedings,* 2014, pp. 580–587.

[16] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Workshop*, arXiv: 1905.08546, 2019.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 162, p. 1–17, 2016.

[18] Performance evaluation metrics, Online on: http://dcase.community/challenge2017/metrics (accessed on 20 April 2020).

[19] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *ICLR 2015 – 3rd International Conference on Learning Representations, May 7-9, San Diego, USA, Proceedings,* 2015, pp. 1–15.

[20] W. Xue, T. Ying, Z. Chao, and D. Guohong, "Multi-beam and multi-task learning for joint sound event detection and localization," in *DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge*, Online on: http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Xue_91.pdf (accessed on 20 April 2020).

[21] J. Zhang, W. Ding, and L. He, "Data augmentation and prior knowledge-based regularization for sound event localization and detection," in *DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge*, Online on: http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_He_97.pdf (accessed on 20 April 2020).