Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 4, August 2020, pp. 1387~1393
ISSN: 2302-9285, DOI: 10.11591/eei.v9i4.2230

Performance analysis of the convolutional recurrent neural network on acoustic event detection

Suk-Hwan Jung, Yong-Joo Chung
Department of Electronics Engineering, Keimyung University, South Korea

Article Info

Article history:

Received Nov 11, 2019
Revised Feb 3, 2020
Accepted Mar 23, 2020

Keywords:

Acoustic event detection
Convolutional neural network
Convolutional recurrent neural network
Hyper-parameters
Recurrent neural network

ABSTRACT

In this study, we attempted to find the optimal hyper-parameters of the convolutional recurrent neural network (CRNN) by investigating its performance on acoustic event detection. Important hyper-parameters such as the input segment length, learning rate, and criterion for the convergence test, were determined experimentally. Additionally, the effects of batch normalization and dropout on the performance were measured experimentally to obtain their optimal combination. Further, we studied the effects of varying the batch data on every iteration during the training. From the experimental results using the TUT sound events synthetic 2016 database, we obtained optimal performance with a learning rate of 10^{-4}. We found that a longer input segment length aided performance improvement, and batch normalization was far more effective than dropout. Finally, performance improvement was clearly observed by varying the starting points of the batch data for each iteration during the training.

Corresponding Author:

Yong-Joo Chung,
Department of Electronics Engineering,
Keimyung University,
1095 Dalgubul-Daero, Dalseo-Gu, Daegu, South Korea.
Email: yjjung@kuu.ac.kr

1. INTRODUCTION

Recently, there has been increasing research interests in acoustic event detection, where the existence and occurrence times of the various sounds in our daily lives are identified. There are many applications for acoustic event detecton, including surveillance [1, 2], urban sound analysis [3, 4], information retrieval from multimedia contents [5], health care monitoring [6-7] and bird call detection [8, 9]. Deep neural networks (DNNs) have demonstrated superior performance to conventional machine learning techniques in image classification [10-12], speech recognition [13-15], and machine translation [16, 17]. In [18], we see that the feedforward neural network (FNN) now outperforms the Gaussian mixture model (GMM) and support vector machine (SVM), which have traditionally been employed for acoustic event detection. FNN has also been shown to outperform the conventional GMM-HMM-based methods in polyphonic acoustic event detection [19]. Therefore, we can say that current studies on acoustic event detection mainly focuse on DNN-based approaches.

However, due to the fixed connection between the input and hidden layers, FNN is apparently inadequate to overcome the signal distortions in image classification. Similarly, it is apparent that FNN is also insufficient for acoustic event detection as audio signal distortions are frequently encountered in the 2-dimensional time-frequency domain of the signal. Another problem with FNN lies in modeling the correlation between the time-frames of the audio signal. As FNN only concatenates several input frames together to model the time correlation, it often fails to model the long-term time correlations of the audio signal.

Convolutional neural networks (CNNs) can alleviate the limitation of FNN using 2-dimensional filters, whose parameters are shared along the time and frequency shift [20], and it has exhibited superior performance to FNN in various pattern recognition tasks. Particularly, due to its structural characteristics, CNN can efficiently handle image distortions in the 2-dimensional space [10]. Similarly, we expect that the time-frequency domain distortions occurring in the audio signal can be accurately modeled by CNN. Nevertheless, CNN is inefficient for modeling long-term time correlations between audio signal samples.

Recurrent neural networks (RNNs) have been used successfully in speech recognition [13], and are superior to other neural networks in modeling the time-correlation information of the time-series signals such as speech and audio. However, because RNN is unable to tackle 2-dimensional distortions in the time-frequency domain of the audio signal, its performance is usually inferior to CNN when used alone in acoustic event detection. Recently, there have been some approaches to combine CNN and RNN for their combined merits. Among them, convolutional recurrent neural networks (CRNNs) have been used successfully for acoustic event detection [21], speech recognition [22], and music classification [23]. As CRNN is constructed by connecting CNN, RNN, and FNN in series, it is more complex than other neural networks; therefore, the combined effect of the networks is difficult to predict. Moreover, as the use of CRNN on acoustic event detection is in its early stages, there are few research studies on optimizing the various hyper-parameters of CRNN.

Therefore, we attempted to find the optimal hyper-parameters of CRNN for acoustic event detection in this study. Several experiments were performed to identify the optimal hyper-parameters of CRNN, and we used the test results on the validation data to determine the hyper-parameters. Important hyper-parameters, such as the input segment length, learning rate, and criterion for the convergence test were determined from the experiments. Additionally, the effects of batch normalization and dropout on the performance were observed. We also studied the effects of varying the batch data in every iteration during the training. This paper is organized as follows. In Section 2, we introduce the feature extraction method for audio signals as well as the architecture of the CRNN used for acoustic event detection. In Section 3, we present and discuss various experimental results, and conclusions are given in Section 4.

2. FEATURE EXTRACTION AND CRNN ARCHITECTURE

2.1. Feature extraction

In this study, we used log-mel filterbank (LMFB) outputs as input features for CRNN and the entire process of feature extraction is shown in Figure 1. We first computed short-time Fourier transform (STFT) from the 40-ms audio signals, which are sampled at 44.1 KHz. STFTs were computed at every 20 ms with 50% overlap [21]. Further, 40-dimensional mel filterbanks were extracted from the STFTs spanning 0 to 22050 Hz, and they were log-transformed to obtain the LMFBs, which are normalized by subtracting the mean and dividing by standard deviation of the entire training data.

Figure 1. Extraction process of log-mel filterbank (LMFB)

2.2. The architecture of CRNN

Figure 2 presents the architecture of the CRNN used in this paper. CRNN consists of CNNs followed by RNN and FNN in sequence. The CNNs act as audio feature extractors, which are robust against distortions in the time-frequency domain. The RNN utilize the time-correlation information of the audio signals. Finally, the FNN serves as an output layer, which produces the posterior probabilities for each sound class at each time frame.

As CNN takes 40-dimensional LMFBs as input features, the dimension of the input of the CNN is Tx40, where the length of the input segment is set to T. The CNN consists of 3 convolution layers, each of which has 256 feature maps with 5×5 filters. The output of the filter is processed by batch normalization and then passes through ReLU activation function. To maintain the time-domain dimension, max pooling is performed only in the frequency-domain. Further, dropout is applied at a rate of 0.25 after the max pooling layer [10]. The input segment length T is set to 1024 frames (20.48s). We experimented with different values of T to find the one that produced the best result on the validation dataset.

The output of the CNN is input to the RNN, which consists of 256 gated recurrent units (GRUs). The output layer consists of K units with sigmoid activation function, where K represents the number of acoustic event classes. The sigmoid activation function produces the posterior probabilities for each class at each time frame, from which we decide whether an acoustic event is active based on a threshold (0.5).

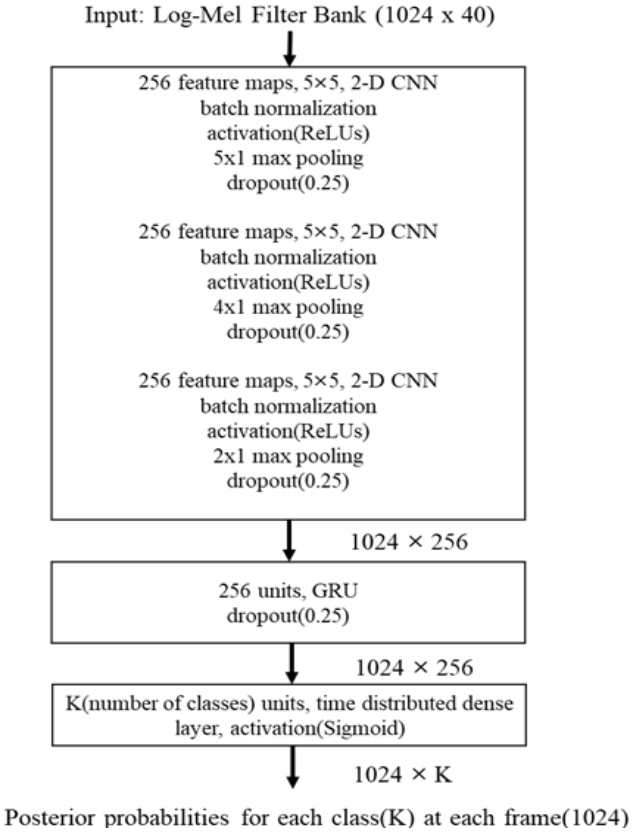

Figure 2. Architecture of the CRNN used for acoustic event detection

3. EXPERIMENTAL RESULTS

3.1. Database and evaluation metric

To evaluate the performance of CRNN on acoustic event detection, we used the TUT sound events synthetic 2016 (TUT-SED Synthetic) database, which is popularly used in this area [24]. TUT-SED Synthetic contains artificially generated audio data, because it is difficult to obtain enough data using only audio data recorded in real environments. Moreover, the subjective labelling error can be mitigated by artificial data. TUT-SED Synthetic was generated by mixing isolated sound events from 16 different classes. The total length of the data is 566 minutes, which were divided into training, testing, and validation data with proportions of 60%, 20%, and 20%, respectively. Segments of length 3-15 seconds were used for the training, testing, and validation, and there were no common acoustic event instances between them. The detailed sound classes and their total duration in the database are shown in Table 1.

For the evaluation metric of the acoustic event detection, we used both error rate (ER) and F-score [25]. We adopted two types of evaluation methods:, segment and event-based. In the segment-based method, the binarized outputs of the CRNN are compared with the ground truth table in every segment of length 1 s. In the event-based method, the output of the CRNN are compared with the label in the ground truth table whenever an acoustic event has been detected by CRNN [25]. We sought the optimal hyper-parameters of the CRNN by applying various conditions during the training. To find the optimal learning rate, we experimented as we changed the learning rate. The results are shown in Table 2.

We applied batch normalization and dropout in all cases, and binary cross entropy was employed as the loss function, which acts as the criterion for the convergence of the weights. The Adam optimizer was used to optime the neural networks. As seen in Table 2, the optimal CRNN performance was achieved at a learning rate 10^{-4}. Generally, as the optimal learning rate for the neural networks is not pre-determined, and varies by both network architecture and amount of training data, we searched for the optimal learning

rate by observing the performance of the validation data set. We can see in Table 2 that the best learning rate for the testing data is same as the learning rate that achieves the best performance on the validation data. This indicates that it is reasonable to determine the optimal learning rate of CRNN by its performance on the validation data. Table 2 also shows that as the learning rate decreases, the number of epochs that shows the best performance increases. For example, the number of epochs is 33 when the learning rate is 10^{-4}, whereas it increases to 191 with the learning rate of 10^{-6}. The larger number of epochs indicates a slower convergence of the CRNN parameters, which causes performance degradation due to the underfitting of the neural networks. In contrast, when the learning increases to 10^{-3}, the number of epochs decreases dramatically to 16, which causes overfitting and results in performance degradation.

To investigate the performance variation with the learning rate further, we show, (as shown in Figure 3), the variation of the loss function and accuracy at the output of the CRNN during the training as the learning rate varied from 10^{-4} to 10^{-7}. When the learning rate is 10^{-4}, it can be seen that the loss function on the validation data reaches its minimum at approximately 30 epochs (33 precisely) and subsequently fluctuates (but never falls below the minimum). However, for the training data, the loss function decreases from the beginning to the end of the training (we set the maximum number of epoch to 200). As it is important for the networks not to be overfitted, we stopped the iteration at 33 epochs by the early stopping algorithm mentioned previosuly. Meanwhile, we see different characteristics when the learning rate is 10^{-5}. The loss function on the validation data decreases for longer and reaches its minimum at 157. The longer iterations contribute to decrease the performance of the CRNN with both validation and test data due to the underfitting problem. This phenomenon becomes more pronounced as we further decrease the learning rate. When the learning rate is 10^{-7}, the loss function does not reach its minimum until the end of the training. A similar trend is observed when we monitor the accuracy instead of the loss function.

We investigated the effect on performance as we changed the input batch data in every epoch of the training. The starting points of each batch data were shifted at each epoch making the input segments at successive epochs differ by the shift-length. For the evaluation, both segment-based and event-based methods were used and the results are shown in F-score and ER. We used ER as the convergence criterion for the training. Further, early stopping was employed where we stopped the training when the convergence criterion did not improve for more than 100 epochs on the validation data. This was to prevent overfitting, and the maximum number of epochs was set to 200. We also investigated the effects of batch normalization (BN) and dropout. BN has been used to mitigate the vanishing and exploding gradient problems in the backpropagation algorithm, and we used BN before the ReLU activation function in the convolution layer. Dropout is a popular regularization method in the neural networks, which is used to exclude the neurons from training at a predefined probability (dropout rate). In this study, dropout is used in all layers of the CNN and RNN (but not the FNN [18].

In Table 3, the results of using the shift of batch data are presented. In the segment-based evaluation, we see an improved average F-score/ER of 63.18%/0.52 using the shift compared to the F-score/ER of 62.11%/0.54 without the shift (non-shift). We also observe slight performance improvement in the event-based evaluation. From these results, we confirm that superior performance is expected by the batch data shift for training the CRNN.

Table 3 also shows the effects of BN and dropout on performance. Expectedly, the best average F-score/ER is obtained when applying both BN and dropout (56.67%/0.71). If we apply only BN without dropout, we see slight performance degradation (54.54%/0.79), which implies that the effect of dropout on performances is not significant. Meanwhile, if we do not apply BN, the performance degrades significantly (regardless of whether we apply dropout), and the poorest result is obtained when we apply neither BN nor dropout (47.15%/0.81). In Table 4, we compare the performances of the CRNN between two convergence criteria (ER and F-score) for training. BN and dropout were applied and the overlap method was used. In the table, we observe slight performance improvementa with ER, but the performance difference appears negligible, and we conclude that the ER and F-scores can be used as the convergence criteria without significantly affecting the performance.

Table 1. Sound classes and their total duration in seconds on the TUT-SED synthetic 2016 databases

Classes	Duration(s)	Classes	Duration(s)
Glass	621	Motor cycle	3691
Gun	534	Foot steps	1173
Cat	941	Claps	3278
Dog	716	Bus	3464
Thunder	3007	Mixer	4020
Bird	2298	Crowd	4825
Horse	1614	Alarm	4405
Crying	2007	Rain	3975

Table 2. Performance of CRNN as learning rate changes

Leaning rate	Validation data (F-score/ER)		Testing data (F-score/ER)		Epoch
	Segment	Event	Segment	Event	
10^{-3}	61.69%/ 0.52	37.69%/0.96	60.61%/0.53	37.05%/0.97	16
10^{-4}	68.75%/ 0.45	43.49%/0.88	64.21%/0.50	40.50%/0.96	33
10^{-5}	66.44%/ 0.49	39.10%/0.96	63.76%/0.52	36.48%/1.04	157
10^{-6}	44.16%/ 0.69	9.83%/1.24	43.38%/0.71	10.82%/1.27	191

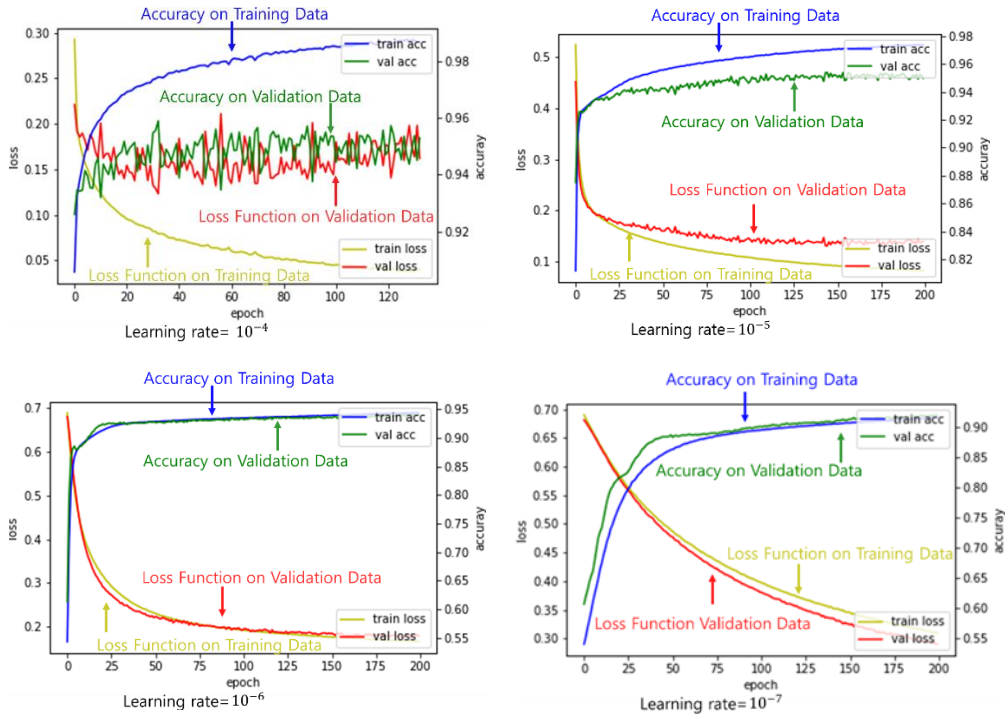

Figure 3. Variation of loss function and accuracy as learning rate changes

Table 3. Performance of CRNN with various training conditions

BN	Drop out	Segment-based (F-score/ER)		Event-based (F-score/ER)		Average
		Shift	Non-shift	Shift	Non-shift	
Yes	No	66.10%/0.48	65.28%/0.54	43.80%/1.01	42.99%/1.12	54.54%/0.79
Yes	Yes	67.24%/0.49	66.62%/0.49	45.93%/0.94	46.88%/0.92	56.67%/0.71
No	No	58.58%/0.57	58.88%/0.58	35.47%/1.06	35.68%/1.04	47.15%/0.81
No	Yes	60.80%/0.54	57.67%/0.56	40.02%/0.97	39.18%/1.00	49.4%/0.77
Average		63.18%/0.52	62.11%/0.54	41.3%/1.0	41.18%/1.02	

Table 4. Performance of CRNN depending on the convergence criterion

Convergence Criterion	Segment-based (F-score/ER)	Event-based (F-score/ER)	Average
ER	67.24%/0.49	45.93%/0.94	56.59%/0.72
F-score	66.45%/0.51	44.97%/0.98	55.71%/0.75

Finally, in Table 5, we show performance variation as we changed the length of the input segment of CRNN. Compared to shorter lengths (2.56s, 5.12s), longer length segments (10.24s, 20.56s) show better performances. This may be due to the fact that many of acoustic events in the *TUT-SED Synthetic* have long lengths and the RNN can efficiently capture the time-correlations in the long segments.

Table 5. Performance of CRNN depending on the length of the input segment

SegmentLength (s)	Segment-based (F-scores/ER)	Event-based (F-scorse/ER)	Average
2.56	66.84%/0.51	42.75%/1.13	54.80%/0.82
5.12	67.47%/0.50	44.27%/1.04	55.87%/0.77
10.24	68.01%/0.47	45.33%/0.97	56.67%/0.72
20.56	67.24%/0.49	45.93%/0.94	56.54%/0.70

4. CONCLUSION

For acoustic event detection, approaches based on deep neural networks have demonstrated superior performances to conventional machine learning methods, such as GMM and SVM. Among them, CRNN is thought to be well suited for acoustic event detection due to its ability to reduce signal distortions in the time-frequency domain as well as in exploiting the temporal-correlation information of the audio signal. In this study, we empolyed CRNN as the classifier for the acoustic event detection, and several of its conditions were tested by extensive experiments to determine its optimal hyper-parameters.

In the experiments, by varying the learning rate, we found that the optimum performance is obtained when the learning rate is set to 10^{-4}. From the results, we could also see that the learning rate that exhibits optimum performance on the validation data also performs best in the testing data. This suggests that it is reasonable to determine the optimal learning rate based on performance tests on the validation data. We further confirmed that BN and dropout contributed to improving the performance of CRNN. Particularly, BN had a larger impact on the performance improvement than dropout.

Instead of using identical batch data at every iteration, we obtained improved performance by changing the batch data in every iteration, which resulted from increasing the number of training samples for CRNN. We further found that the length of the input segments of the CRNN also affects the performance. We obtained better performance using longer segments, as the acoustic event used in this paper had relatively long time-durations.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by Ministry of Education (No. 2018R1A2B6009328).

REFERENCES

[1] M. K. Nandwana, et al., "Robust Unsupervised Detection of Human Screams In Noisy Acoustic Environments," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, pp. 161-165, 2015.
[2] M. Crocco, et al., "Audio Surveillance: A Systematic Review," *ACM Computing Surveys*, no. 52, 2016.
[3] J. Salamon and J. P. Bello, "Feature Learning with Deep Scattering for Urban Sound Analysis," *2015 23rd European Signal Processing Conference (EUSIPCO)*, Nice, pp. 724-728, 2015.
[4] S. Ntalampiras, et al., "On Acoustic Surveillance of Hazardous Situations", *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, pp. 165-168, 2009.
[5] Y. Wang, et al., "Audio-based Multimedia Event Detection Using Deep Recurrent Neural Networks," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, pp. 2742-2746, 2016.
[6] D. Stowell and D. Clayton, "Acoustic Event Detection for Multiple Overlapping Similar Sources *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, pp. 1-5, 2015.
[7] G. Dekkers, et al., "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," *KU Leuven, Tech. Rep.*, 2018.
[8] D. Stowell, et al., "Automatic Acoustic Detection of Birds through Deep Learning: the First Bird Audio Detection Challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp.1-21, 2018.
[9] F. Briggs, et al., "Acoustic Classification of Multiple Simultaneous Bird Species: A Multi-instance Multi-Label Approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp.4640-4640, 2012.
[10] A. Krizhevsky, et al., "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 1097-1105, 2012.
[11] K. He, et al., "Deep Residual Learning for Image Recognition*," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 770-778, 2016.
[12] Olga Russakovsky, et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015.
[13] A. Graves, et al., "Speech Recognition with Deep Recurrent Neural Networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, pp. 6645-6649, 2013.
[14] J. Chorowski, et al., "Attention-Based Models for Speech Recognition," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, vol. 1, pp. 577-585, 2015.
[15] A. Hannun, et al., "Deep Speech: Scaling up End-to-End Speech Recognition," *arXiv: 1412.5567*, 2014.
[16] K. Cho, et al., "Learning Phrase Representations Using Rnn Encoder-Decoder for Statistical Machine Translation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724-1734, 2014.
[17] I. Sutskever, et al., "Sequence to Sequence Learning with Neural Networks", in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pp. 3104-3112, 2014.

[18] S. H. Chung and Y. J. Chung, "Comparison of Audio Event Detection Performance using DNN," *Journal of the Korea Institute of Electronic Communication Sciences*, vol. 13, no. 3, pp. 571-577, 2018.
[19] E. Cakir, et al., "Polyphonic sound event detection using multilabel deep neural networks," *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, pp. 1-7, 2015.
[20] Y. LeCun, et al., "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.
[21] E. Cakir, et al., "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291-1303, 2017.
[22] T. N. Sainath, et al., "Convolutional, Long Short-term Memory, Fully Connected Deep Neural Networks," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, pp. 4580-4584, 2015.
[23] K. Choi, et al., "Convolutional Recurrent Neural Networks for Music Classification," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 2392-2396, 2017.
[24] "TUT-SED Synthetic," 2016. [Online]. Available at: http://www.cs.tut.fi/sgn/arg/taslp2017-crnn-sed/tut-sed-synthetic-2016
[25] A. Mesaros, et al., "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162-178, 2016.

BIOGRAPHIES OF AUTHORS

Sukwhan Jung received his B.Sc. and M.Sc. Degree in Electronics Engineering from Keimyung University, Daegu, South Korea in 2016 and 2018, respectively. He has been with Samju Electroincs Co. since March 2018. His main research interests are audio event detection under noisy environments and deep learning for artificial intelligence.

Yongjoo Chung received his B.Sc. degree in Electronics Engineering from Seoul National University, Seoul, South Korea in 1988. He earned his M.Sc. and PhD degree in Electrical and Electronics Engineering from Korea Advanced Science and Technology, Daejon, South Korean in 1995. He is currently a Professor with the Department of Electronics Engineering at Keimyung University, Daegu, S. Korea. His research interests are in the areas of speech recognition, audio event detection, machine learning and pattern recognition.