SOUND EVENT DETECTION BASED ON CURRICULUM LEARNING CONSIDERING LEARNING DIFFICULTY OF EVENTS

Noriyuki Tonami¹, Keisuke Imoto², Yuki Okamoto³, Takahiro Fukumori¹, Yoichi Yamashita¹

¹ Ritsumeikan University, Japan, ² Doshisha University, Japan

ABSTRACT
In conventional sound event detection (SED) models, two types of events, namely, those that are present and those that do not occur in an acoustic scene, are regarded as the same type of events. The conventional SED methods cannot effectively exploit the difference between the two types of events. All time frames of sound events that do not occur in an acoustic scene are easily regarded as inactive in the scene, that is, the events are easy-to-train. The time frames of the events that are present in a scene must be classified as active in addition to inactive in the acoustic scene, that is, the events are difficult-to-train. To take advantage of the training difficulty, we apply curriculum learning into SED, where models are trained from easy- to difficult-to-train events. To utilize the curriculum learning, we propose a new objective function for SED, wherein the events are trained from easy- to difficult-to-train events. Experimental results show that the F-score of the proposed method is improved by 10.09 percentage points compared with that of the conventional binary cross entropy-based SED.

Index Terms— Sound event detection, acoustic scene, curriculum learning

1. INTRODUCTION
The analysis of various environmental sounds in everyday life has become an increasingly important area in signal processing [1]. The automatic analysis of environmental sounds will give rise to various applications, such as anomalous sound detection systems [2], automatic life-logging systems [3], monitoring systems [4], and bird-call detection systems [5].

Sound event detection (SED) is the task of recognizing sound event labels and their timestamp from a recording. In SED, the models need to recognize overlapped multiple sound events in a time frame. Recently, neural-network-based SED models have seen increasingly rapid advances, such as the convolutional neural network (CNN) [6], recurrent neural network (RNN) [7], and convolutional recurrent neural network (CRNN) [8]. CNN is the structure that automatically extracts features and is robust to time and frequency shifts. RNN is good at modeling the time structure in an audio stream. Moreover, some works considering the relationship between sound events and scenes have been proposed. As an example of the relationship, “mouse clicking” occurs indoors such as “office,” whereas, “car” tends to occur outdoor such as “city center.” On the basis of this idea, SED using the information on the acoustic scene [9][11] and the model combining SED and acoustic scene classification (ASC) have been proposed. Heittola et al. [10] have proposed the SED model using the results of the ASC, where the ASC model is trained in the first stage and then the SED model is trained in the second stage with the ASC results. Tonami et al. [13] have proposed the multitask-learning-based models combining SED and ASC.

In the conventional SED methods, two types of events, namely, those that are present and those that do not occur in an acoustic scene, are treated as the same type of the events. The conventional SED methods cannot effectively utilize the difference between the two types of events. The all time frames of events that do not occur in a scene only need to be treated as inactive in the acoustic scene, as shown in Fig. 1 (“elephant” and “birdsong” in “airplane”), i.e., the training of the easy-to-train events is considered as the task of recognizing one class. On the other hand, the time frames of events that are present in an acoustic scene must be classified as active or inactive in the acoustic scene, as shown in Fig. 1 (“footsteps” in “airplane”), i.e., the training of the difficult-to-train events is regarded as the task of binary classification.

To utilize the difference in the difficulty of training between the sound events, we employ curriculum learning [17]. Curriculum learning is a method of learning data effectively utilizing the difficulty of training, in which a model learns progressively from easy- to difficult-to-train data. Recently, some works using the curriculum learning have been carried out [18][20]. Lotfian and Busso [19] have proposed the speech emotion recognition method based on the curriculum learning, where the ambiguity of emotion is considered. In this paper, we propose a SED method using the curriculum learning, in which strong labels are given for the training. In the proposed method, the SED models are trained from the easy- to difficult-to-train events on the basis of the curriculum learning. More specifically, we present a new objective function of SED considering the difficulty of the training of events based on the curriculum learning.
methods, the acoustic features in the time-frequency domain have been studied. In most of the neural-network-based methods, the binary cross-entropy loss is used as follows:

\[
L_{\text{BCE}} = - \sum_{n=1}^{N} \sum_{t=1}^{T} \left\{ z_{n,t} \log (y_{n,t}) + (1-z_{n,t}) \log \left(1 - \sigma(y_{n,t})\right) \right\}, \tag{1}
\]

where \( N \) and \( T \) indicate the numbers of sound event categories and time frames, respectively. \( z_{n,t} \in \{0, 1\} \) is a target label of an event \( n \) at time \( t \). If the event is active, \( z_{n,t} \) is 1; otherwise, \( z_{n,t} \) is 0. \( y_{n,t} \) represents the output of the network of an event \( n \) at time \( t \). \( \sigma(\cdot) \) denotes the sigmoid function.

### 2. CONVENTIONAL METHOD

SED involves sound event labels and their onset/offset from an audio. Recently, many neural-network-based methods have been studied. In most of the neural-network-based methods, the acoustic features in the time-frequency domain are used for the input to the SED models. To optimize the neural-network-based SED models, the binary cross-entropy loss is used as follows:

\[
L_{\text{BCE}} = - \sum_{n=1}^{N} \sum_{t=1}^{T} \left\{ z_{n,t} \log (y_{n,t}) + (1-z_{n,t}) \log \left(1 - \sigma(y_{n,t})\right) \right\}, \tag{1}
\]

where \( N \) and \( T \) indicate the numbers of sound event categories and time frames, respectively. \( z_{n,t} \in \{0, 1\} \) is a target label of an event \( n \) at time \( t \). If the event is active, \( z_{n,t} \) is 1; otherwise, \( z_{n,t} \) is 0. \( y_{n,t} \) represents the output of the network of an event \( n \) at time \( t \). \( \sigma(\cdot) \) denotes the sigmoid function.

### 3. PROPOSED METHOD

#### 3.1. Training difficulty of events considering scenes

In the conventional SED methods, two types of events, namely, those that exist and those that do not occur in an acoustic scene, are treated as the same type of the events. The conventional SED methods cannot effectively employ the difference between the two types of sound events. The all time frames of events that do not occur in an acoustic scene are treated as the same type of the events. The conventional SED methods cannot effectively employ the difference between the two types of sound events. The all time frames of events that do not occur in an acoustic scene are treated as the same type of the events.

The training of the sound events is treated as the task that recognizes one class (inactive), that is, the events are easy-to-train. On the other hand, the time frames of sound events that exist in an acoustic scene need to be classified as active in addition to inactive in the acoustic scene, as shown in Fig. 1 (“elephant” and “birdsong” in “airplane”). The training of the sound events is treated as the task that classifies two classes (active or inactive), that is, the events are difficult-to-train. In short, the sound events that exist in an acoustic scene are hardly trained compared with the events that do not occur in the acoustic scene as shown in Fig. 2.

#### 3.2. Curriculum-learning-based objective function

As mentioned in Sect. 3.1., there are differences in training difficulty between the sound events when the acoustic scenes are considered. In the proposed method, we employ the curriculum learning to take advantage of the difference in the difficulty of training between the sound events when the acoustic scenes are considered. To incorporate the concept of the curriculum learning into the BCE, the following loss function is used instead of Eq. (1)

\[
L_{\text{prop}} = - \sum_{n=1}^{N} \sum_{t=1}^{T} \left\{ z_{n,t} \log (y_{n,t}) + (1-z_{n,t}) \log \left(1 - \sigma(y_{n,t})\right) \right\}, \tag{2}
\]

where \( g_n \) is a gate function that controls the weight of training of two types of events. More specifically, the gate function is calculated as

\[
g_n = \alpha_s f_n + (1 - \alpha_s)(1-f_n), \tag{3}
\]

where \( \alpha_s \) is a progressive parameter, which is changed from 0 to 1 with time-step \( s \) (epoch) during training. \( f_n \) is an event-flag. If an event \( n \) occurs at least once in the acoustic scene of the input audio, the frag is 1; otherwise, it is 0.

As shown in Fig. 2 in the early stage of the training, only the events that do not occur in an acoustic scene are trained. On the other hand, in the late stage of the training, only the events that are present in an acoustic scene are trained. Note that whether an event is difficult- or easy-to-train is determined by each acoustic scene label of an audio clip. For example, a dataset includes a scene \( A \). Events \( a \) and \( b \) occur at least once in the scene \( A \). An event \( c \) does not occur in the scene \( A \). When the scene label of input audio is \( A \), \( a \) and \( b \) are regarded as the difficult-to-train events. \( c \) is regarded as the easy-to-train event.

### 4. EXPERIMENTS

#### 4.1. Experimental conditions

To evaluate the performance of the proposed method, we conducted evaluation experiments using the TUT Sound Events 2016 [21], TUT Sound Events 2017 [22], TUT Acoustic Scenes 2016 [21], and TUT Acoustic Scenes 2017 [22] datasets. From these datasets, we selected sound clips includ-


Table 1. Experimental conditions

| Feature | Log-mel energy (64 dim.) |
|---------|---------------------------|
| Frame length/shift | 40 ms / 20 ms |
| Length of sound clip | 10 s |
| Network architecture | 3 CNN + 1BiGRU + 1 fully con. |
| # channels of CNN layers | 128, 128, 128 |
| Filter size | 3 × 3 |
| Pooling size | 8 × 1, 2 × 1, 2 × 1 (max pooling) |
| # units in GRU layer | 32 |
| # units in fully con. layer | 32 |
| # units in output layer | 25 |
| Threshold | 0.5 |

4.2. Experimental results

Table 2 shows the SED performances in terms of the segment-based F-score and error rate. In Table 2, micro and macro indicate the overall and class-average scores, respectively. The numbers to the right of ± represent standard deviations. “BCE” is the CNN–BiGRU using the BCE loss. “Proposed method” represents the SED performance using Eqs. 2 and 3 with CNN–BiGRU. “Proposed+MTL of SED & SAD” indicates the SED performance using Eqs. 2 and 3 with SAD. “Proposed+MTL of SED & ASC” denotes the multitask-learning-based SED with ASC using the proposed objective function for SED. The results show that the proposed method achieves a more reasonable performance than the conventional BCE. Moreover, when using SAD and the model combining SED and ASC with the proposed objective function, the SED performance is better than those of the conventional MTL of SED & SAD and the MTL of SED &
In this paper, we proposed the curriculum-learning-based objective function for SED. In the proposed method, we applied the training difficulty between sound events considering acoustic scenes to the conventional BCE loss. More specifically, the SED models using the proposed method are trained from the easy-to-train to difficult-to-train events during training. The experimental results indicate that the proposed method improves the F-score of the SED by 10.09 percentage points compared with that of the conventional CNN–BiGRU using the BCE loss. In our future work, we will investigate a more effective method for SED considering the relationship between sound events and acoustic scenes.

### 5. CONCLUSION

### 6. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP19K20304.

### 7. REFERENCES

[1] K. Imoto, “Introduction to acoustic event and scene analysis,” *Acoust. Sci. Tech.*, vol. 39, no. 3, pp. 182–188, 2018.

[2] C. Chan and E. W. M. Yu, “An abnormal sound detection and classification system for surveillance applications,” *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 1851–1855, 2010.

[3] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, “Audio-based human activity recognition using non-Markovian ensemble voting,” *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 509–514, 2012.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, “On acoustic surveillance of hazardous situations,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 165–168, 2009.
[5] Y. Okamoto, K. Imoto, N. Tsukahara, K. Sueda, R. Yamashita, and Y. Yamashita, “Crow call detection using gated convolutional recurrent neural network,” Proc. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP), pp. 171–174, 2020.

[6] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, “CNN architectures for large-scale audio classification,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135, 2017.

[7] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, “Duration-controlled LSTM for polyphonic sound event detection,” IEEE/ACM Trans. on Audio, Speech, and Language Processing, vol. 25, no. 11, pp. 2059–2070, 2017.

[8] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 25, no. 6, pp. 1291–1303, 2017.

[9] A. Mesaros, T. Heittola, and A. Klapuri, “Latent semantic analysis in sound event detection,” Proc. European Signal Processing Conference (EUSIPCO), pp. 1307–1311, 2011.

[10] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, no. 1, pp. 1–13, 2013.

[11] K. Imoto and S. Shimauchi, “Acoustic scene analysis based on hierarchical generative model of acoustic event sequence,” IEICE Trans. Inf. Syst., vol. E99-D, no. 10, pp. 2539–2549, 2016.

[12] H. L. Bear, I. Nolasco, and E. Benetos, “Towards joint sound scene and polyphonic sound event recognition,” Proc. INTERSPEECH, pp. 4594–4598, 2019.

[13] N. Tonami, K. Imoto, M. Niitsuma, R. Yamashimi, and Y. Yamashita, “Joint analysis of acoustic events and scenes based on multitask learning,” Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 333–337, 2019.

[14] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamashizhi, and Y. Yamashita, “Sound event detection by multitask learning of sound events and scenes with soft scene labels,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 621–625, 2020.