WEAKLY LABELED LEARNING USING BLSTM-CTC FOR SOUND EVENT DETECTION

Taiki Matsuyoshi*, Tatsuya Komatsu†, Reishi Kondo†, Takeshi Yamada*, and Shoji Makino*
* University of Tsukuba, Japan
† NEC Corporation, Japan
E-mail: matsuyoshi@mmlab.cs.tsukuba.ac.jp

***Abstract*—In this paper, we propose a method of weakly labeled learning of bidirectional long short-term memory (BLSTM) using connectionist temporal classification (BLSTM-CTC) to reduce the hand-labeling cost of learning samples. BLSTM-CTC enables us to update the parameters of BLSTM by loss calculation using CTC, instead of the exact error calculation that cannot be conducted when using weakly labeled samples, which have only the event class of each individual sound event. In the proposed method, we first conduct strongly labeled learning of BLSTM using a small amount of strongly labeled samples, which have the timestamps of the beginning and end of each individual sound event and its event class, as initial learning. We then conduct weakly labeled learning based on BLSTM-CTC using a large amount of weakly labeled samples as additional learning. To evaluate the performance of the proposed method, we conducted a sound event detection experiment using the dataset provided by Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Task 2. As a result, the proposed method improved the segment-based F1 score by 1.9% compared with the initial learning mentioned above. Furthermore, it succeeded in reducing the labeling cost by 95%, although the F1 score was degraded by 1.3%, comparing with additional learning using a large amount of strongly labeled samples. This result confirms that our weakly labeled learning is effective for learning BLSTM with a low hand-labeling cost.**

I. Introduction

Sound event detection (SED) is important for realizing security and autolabeling systems for movie contents by understanding various sounds. The objective of SED is to detect the beginning and end of each individual sound event and to identify its event class. For example, Fig. 1 indicates that the SED system outputs a label which represents the beginning and end of phone ringing when a sound of phone ringing is input to the SED system.

One of the conventional approaches to SED is to use non-negative matrix factorization (NMF) [1][2]. In the NMF approach, the detection model can be trained using a small amount of learning data by linear processing. However, the detection model cannot achieve good performance in practical tasks. Another approach is to use deep neural networks (DNNs) [3]. The detection model using nonlinear processing achieves good performance for unknown real data. In particular, a method using bidirectional long short-term memory (BLSTM) [4] achieved good performance for the SED task in Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 [5]. However, the detection model requires a large amount of learning data to optimize its parameters. For learning DNN including BLSTM, a large amount of "strongly labeled" samples, in which the beginning and end of each individual sound event and one of the predefined sound event classes are hand-labeled as shown in Fig. 2, are needed. Preparing this hand-labeled learning data is labor intensive and time-consuming, so a method of learning using "weakly labeled" samples, in which only event sound class is labeled as shown in Fig. 2, is necessary.

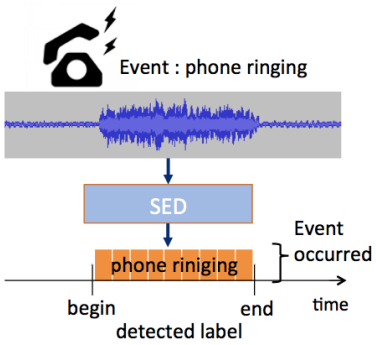


Fig. 1. Sound event detection.

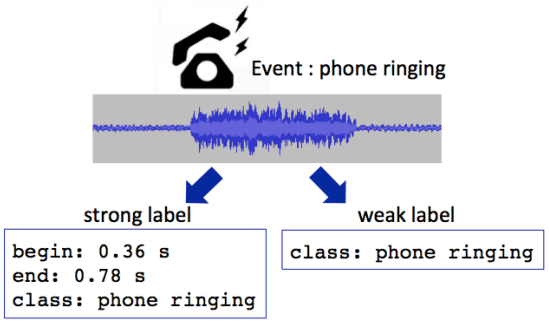


Fig. 2. Strong label and weak label.

In this paper, we describe the weakly labeled learning of BLSTM. To realize it, we focus on the connectionist temporal classification (CTC) [6]. Since the exact error calculation to update the parameters of BLSTM cannot be conducted when using weakly labeled data, we introduce loss calculation using CTC instead of it. Thus, we propose a weakly labeled learning

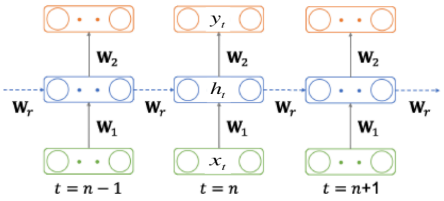


Fig. 3. Recurrent neural network.

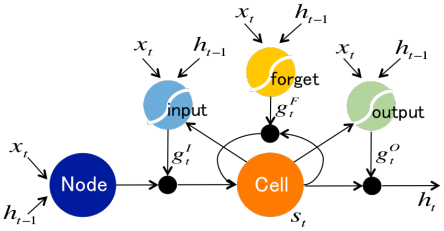


Fig. 4. Long short-term memory.

method of BLSTM using connectionist temporal classification (BLSTM-CTC). In the proposed method, we first conduct strongly labeled learning of BLSTM using a small amount of strongly labeled samples as initial learning. Since it is difficult to learn BLSTM using only weakly labeled samples, we use strongly labeled learning before weakly labeled learning. We then conduct weakly labeled learning of BLSTM based on CTC using a large amount of weakly labeled samples as additional learning. BLSTM-CTC updates the parameters of BLSTM by the loss calculation of CTC. This enables us to improve the SED accuracy of the initially learned BLSTM. To evaluate the performance of the proposed method, we conducted an SED experiment using the dataset provided by DCASE 2016 Task 2.

II. Recurrent Neural Network

A. Overview of recurrent neural network

The detection target of SED is the beginning and end of each sound event and the class of these sounds. Because it includes time information, a recurrent neural network (RNN) [7] that can deal with time-series data is often used in SED. RNN consists of an input layer, a hidden layer and, an output layer. Fig. 3 shows that an output in time-frame $t = n$ is determined by using information of an input in $t = n$ and a hidden layer in $t = n - 1$. The following formula describes this process.

$$\mathbf{h_t} = f(\mathbf{W}_1\mathbf{x}_t + \mathbf{W}_r\mathbf{h}_{t-1} + \mathbf{b}_1) \quad (1)$$

$$\mathbf{y}_t = g(\mathbf{W}_2\mathbf{h}_t + \mathbf{b}_2) \quad (2)$$

$\mathbf{x_t}$ denotes an input sequence of feature vectors. Moreover, $\mathbf{h}_t$ represents a hidden layer vector, and $\mathbf{y}_t$ represents an output layer vector. $\mathbf{W}_i$ and $\mathbf{b_i}$ respectively denote the input weight matrix and bias of the i-th layer. $\mathbf{W}_r$ represents a recurrent weight matrix. f and g represent activation functions of the hidden layer and output layer, respectively. By this process, it can determine outputs by considering the information of the correlation among the previous time frames.

B. Bidirectional long short-term memory

RNN has a problem that good performance cannot be obtained when applying a long time-series data, because of the vanishing gradient problem [8]. To solve this problem, long short-term memory (LSTM) [9], which is the RNN incorporated into the functions of cells to adapt to long time-series data, was proposed. As shown in Fig. 4, LSTM consists of a memory cell $\mathbf{s}_t$ and three gates, namely, an input gate $\mathbf{g}_t^I$, a forget gate $\mathbf{g}_t^F$, and an output gate $\mathbf{g}_t^O$. Each gate $\mathbf{g}^*$ has a value between 0 and 1. The value 0 means that the gate is closed, and the value 1 means that the gate is open. The hidden layer output $\mathbf{h}_t$ in Eq. (1) is replaced by Eqs. (3) to (7).

$$\mathbf{g}_t^I = \sigma(\mathbf{W}^I\mathbf{x}_t + \mathbf{W}_r^I\mathbf{h}_{t-1} + \mathbf{s}_{t-1}) \quad (3)$$

$$\mathbf{g}_t^F = \sigma(\mathbf{W}^F\mathbf{x}_t + \mathbf{W}_r^F\mathbf{h}_{t-1} + \mathbf{s}_{t-1}) \quad (4)$$

$$\mathbf{s}_t = \mathbf{g}_t^I \odot f(\mathbf{W}_1\mathbf{x}_t + \mathbf{W}_r\mathbf{h}_{t-1} + \mathbf{b}_1) + \mathbf{g}_t^F \odot \mathbf{s}_{t-1} \quad (5)$$

$$\mathbf{g}_t^O = \sigma(\mathbf{W}^O\mathbf{x}_t + \mathbf{W}_r^O\mathbf{h}_{t-1} + \mathbf{s}_{t-1}) \quad (6)$$

$$\mathbf{h}_t = \mathbf{g}_t^O \odot tanh(\mathbf{s}_t) \quad (7)$$

In these equations, $\mathbf{W}$ and $\mathbf{W}_r$ denote input weight matrices and recurrent weight matrices, and the superscripts I, F, and O denote the input, forget, and output gates, respectively. $\odot$ means point-wise multiplication, and σ represents a logistic sigmoid function.

In this paper, we use bidirectional LSTM (BLSTM) [10], which is a kind of LSTM and contains both forward hidden layers and backward hidden layers, because SED using BLSTM [4] achieved good performance in DCASE 2016.

C. Problem when applying BLSTM to SED

A large amount of strongly labeled data are needed to learn BLSTM. Preparing them is labor intensive and time-consuming, since the hand-labeling is hard. Therefore, we apply weakly labeled learning, which enables us to learn BLSTM with a low hand-labeling cost.

III. Proposed Method

A. Overview of SED using BLSTM

Fig. 5 shows the processing of SED using BLSTM in our proposed method. First, we divide an input signal into 25 ms windows with a 40% overlap, and calculate a log filterbank feature for each time frame. We then input the feature into BLSTM. The input information of the feature is propagated to the hidden layer, and event presence probability is output at every frame for each event class. A label of the beginning and end of a sound event and its sound event class are determined by a threshold-based method for each event probability sequence. This model is constructed with reference to the method using BLSTM in DCASE 2016 [4].

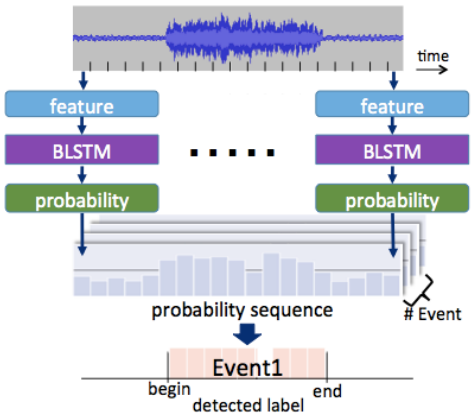


Fig. 5. SED using BLSTM in the proposed method.

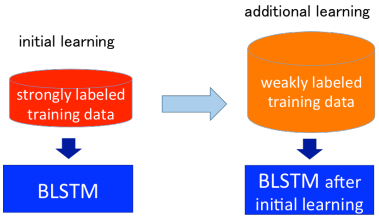


Fig. 6. Training processing of the proposed method.

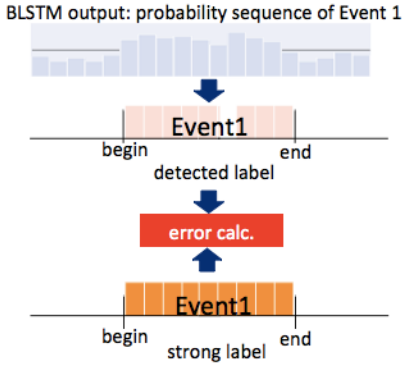


Fig. 7. Error calculation using strongly labeled data.

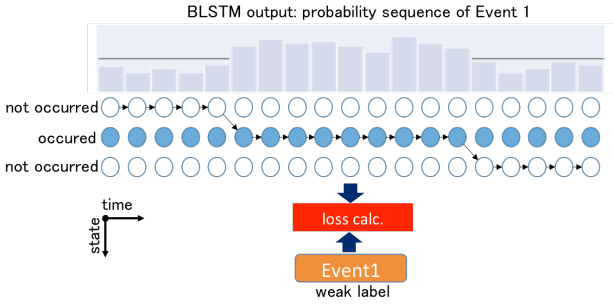


Fig. 8. Loss calculation by CTC using weakly labeled data.

B. Proposed training method of BLSTM

In the proposed method, the BLSTM is learned by using the training data generated from a small amount of strongly labeled samples as initial learning. After that, it is learned by using those generated from a large amount of weakly labeled samples as additional learning, as shown in Fig. 6. The detailed explanation is given as follows.

1) Initial training using strongly labeled data: An overview of initial learning using strongly labeled data is shown in Fig. 7. This figure represents the error calculation between a strong label and a label detected by BLSTM. The error is calculated on the basis of softmax-cross entropy, and the parameters of the BLSTM are updated so that the error is reduced.

2) Additional training using weakly labeled data: An overview of additional training using weakly labeled data is shown in Fig. 8. This figure shows the loss calculation between a weak label and the event probability sequence output by the BLSTM. The loss is calculated by using CTC, and the parameters of the BLSTM are updated so that the loss is reduced.

We explain about loss calculation using CTC. In Fig. 8, the horizontal and vertical axes represent the time and state of a sound event, respectively. The states consist of "occurred" and "not occurred" for Event 1 and "blank". The state of "blank" is regarded as a wild card, which means that whether the event occurs or not is not determined. We omitted the state of "blank" in Fig. 8 for simplicity, although we actually used these three states in the calculation. Loss calculation using CTC is conducted by using the probability sequence of Event 1 output from BLSTM. Each arrow indicates a transition of the state of event in each frame. The loss calculation using CTC is represented by Eqs. (8) and (9).

$$p(\boldsymbol{\pi}|\mathbf{x}) = \prod_{t=1}^{T} \mathbf{y}_{\boldsymbol{\pi}_t}^{t} \tag{8}$$

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\boldsymbol{\pi} \in B^{-1}(\mathbf{l})} p(\boldsymbol{\pi}|\mathbf{x}) \tag{9}$$

The probability $p(\boldsymbol{\pi}|\mathbf{x})$ of the path $\boldsymbol{\pi}$ when the input of BLSTM is $\mathbf{x}$ is calculated using Eq. (8). $p(\boldsymbol{\pi}|\mathbf{x})$ is obtained as a product of $\mathbf{y}_{\boldsymbol{\pi}_t}^{\mathbf{t}}$ that denotes a probability of the state $\boldsymbol{\pi}_t$ on the path $\boldsymbol{\pi}$ at the frame t. The probability is obtained from the event probability sequence output by BLSTM. In Fig. 8, when $t = 2$, the state $\boldsymbol{\pi}_2$ on the path represents "not occurred", and $\mathbf{y}_{\boldsymbol{\pi}_2}^{2}$ represents a probability of "not occurred" in the frame $t = 2$. The probability $p(\mathbf{l}|\mathbf{x})$ of the state sequence $\mathbf{l}$ obtained from the weak label when the input of BLSTM is $\mathbf{x}$ is calculated using Eq. (9). $p(\mathbf{l}|\mathbf{x})$ is obtained as a sum of $p(\boldsymbol{\pi}|\mathbf{x})$. $B^{-1}(\mathbf{l})$ is a set of the paths that become the same as $\mathbf{l}$ by removing the blank and the contiguous state from each path. In Fig. 8, the state sequence of "not occurred", "occurred", and "not occurred" is obtained from the path.

TABLE I
EXPERIMENTAL CONDITIONS (BLSTM)

Learning rate	strongly lebeled learning: 0.0005 weakly labeled learning : 0.00001
Gradient clipping norm	strong labeled learning: 5 weakly labeled learning : 1
Batch size	strongly labeled learning: 50 weakly labeled learning : 1
Epoch	strongly labeled learning: 20 weakly labeled learning : 5
Hidden layer size	400
# of hidden layers	2

TABLE II
EXPERIMENTAL CONDITIONS (AUDIO DATA)

Sampling rate	44100 Hz
SNR	$-6, 0, 6$ dB
feature	39 Mel-filter bank outputs
Frame size	25 ms
# of event class	11
# of learning data for initial learning	12600 (=11 sound events $\times$ 5 samples $\times$ 220 patterns)
# of learning data for additional learning	2200 (=11 sound events $\times$ 20 samples $\times$ 10 patterns)
Generated data length	5 s
# of evaluation data	54 samples
Evaluation data length	120 s

When the loss $-\log p(\mathbf{l}|\mathbf{x})$ becomes small, the state sequence obtained from the event probability sequence output by the BLSTM corresponds to the state sequence $\mathbf{l}$. The parameters of the BLSTM are updated by minimizing $-\log p(\mathbf{l}|\mathbf{x})$ by using the back propagation algorithm. By this loss calculation using CTC, BLSTM can be learned using weakly labeled data.

IV. EXPERIMENT

A. Experimental conditions

An Experiment was conducted to evaluate the effectiveness of our proposed method using weakly labeled data. We used development and evaluation datasets provided by DCASE 2016 Task 2. There are only 20 clean samples per sound event in the development dataset. Because this dataset is insufficient to learn BLSTM, we artificially generated our own training data from the development dataset with reference of the method using BLSTM [4].

The detailed conditions of the experiment are summarized in Tables I and II. We randomly divide the 20 samples into 4 subsets. The learning data are then generated for each subset as follows: 1) randomly extract the 5 s length background noise from the development dataset, 2) randomly select a clean sound sample from the development dataset, 3) add the selected sample to the extracted background noise at the signal to noise ratio (SNR) that is the same as that for the evaluation dataset, and 4) repeat Steps. 2) and 3) at once. By this processing, we obtained 48400 (11 sound events $\times$ 20 samples $\times$ 220 addition patterns) learning data that each includes two sound events. The strong and weak labels of these data are automatically given by the labels of the 20 samples, therefore the labeling cost of the 20 samples becomes important.

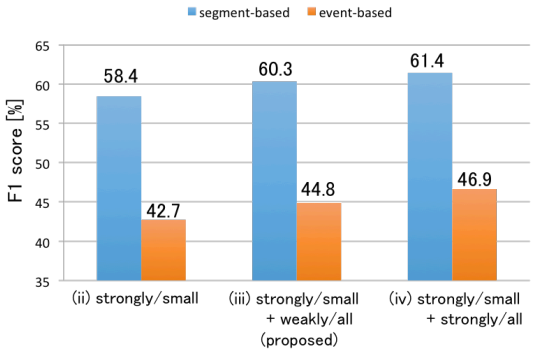


Fig. 9. F1 score for each learning method.

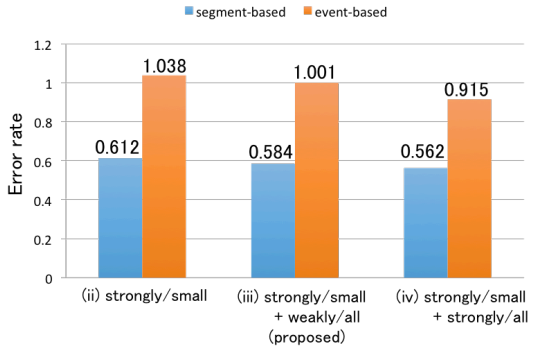


Fig. 10. Error rate for each learning method.

In the experiment, we compare the four BLSTM learning methods.

(i) **Strongly/all**: we conduct strongly labeled learning of BLSTM by using the 48400 data generated from all (20) the samples.

(ii) **Strongly/small**: we conduct strongly labeled learning of BLSTM by using the 12100 data generated from the small (5) samples. This corresponds to initial learning of the proposed method.

(iii) **(Proposed method) strongly/small + weakly/all**: for the BLSTM learned by (ii), we additionally conduct weakly labeled learning using 2200 data (11 sound events $\times$ all 20 samples $\times$ 10 addition patterns) randomly selected among the 48400 data.

(iv) **Strongly/small + strongly/all**: for the BLSTM learned by (ii), we also additionally conduct strongly labeled learning using the same 2200 data.

The evaluation dataset provided by DCASE 2016 Task 2 has 120 s length audio data. In the evaluation, we divide one audio data into 24 audio data with 5 s lengths. The evaluation is event-based and segment-based, where F1 score and error

rate are utilized as evaluation criteria [11].

B. Results of experiment and analysis

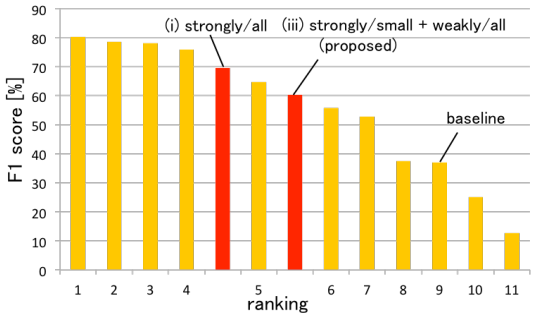
We first show the result of the learning method (i). The segment-based (Seg.) and event-based (Ev.) F1 score are 69.6% and 60.6%, respectively. The segment-based and event-based error rate are 0.4638 and 0.6717, respectively. This performance is the best possible performance of the SED using BLSTM described in Sec. 3.A and is ranked the fourth (Ev.) or fifth (Seg.) among the methods submitted to DCASE 2016 as shown in Fig 11.

We next compare the learning methods (iii) with (ii) and (iv) in Figs. 9 and 10. All results are averaged over 4 subsets. Figs. 9 and 10 show that the proposed learning method (iii) achieved better performance than the learning method (ii) for all evaluation criteria. It means that weakly labeled learning is effective for learning BLSTM. On the other hand, the performance of the proposed learning method (iii) degrades compared with the learning method (iv). It means that it is difficult for weakly labeled learning to achieve as good performance as that of strongly labeled learning with the same number of samples. We furthermore compared the performance of the proposed method with the methods submitted to DCASE 2016. Fig. 11 shows the performance ranking for each evaluation criterion. The figure confirms that the proposed method is located the fifth (Ev.) or the sixth (Seg.) in the ranking. This is reasonable because we used a relatively simple BLSTM architecture and weakly labeled learning.

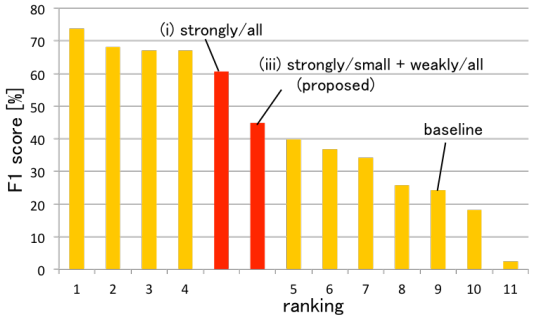
Finally, we examine the hand-labeling cost for the learning methods (ii), (iii), and (iv). A weak label with only its event sound class can be easily attached by listening to each sequence at once. If an event is randomly located in a signal of 5 s long, we need an average of 2.5 s to label it. On the other hand, to obtain a strong label with the beginning, end and class of each sound event, it is necessary to listen to the each sequence repeatedly. Thus, it takes much longer time than the weak label. For example, when we actually performed strong and weak labeling of the data used in [12][13], The strong labeling took 20 times as long as the weak labeling. According to this fact, when the labeling cost (time) of the method (ii) is normalized to 1, the labeling cost of the proposed method (iii) is 1.15 ($= 1 + 3/20$), whereas that of the method (iv) is 4 ($= 1 + 3$). Fig. 12. represents the relationship between the segment-based F1 score and the labeling cost for each method. By comparing, the F1 score of the proposed method (iii) is improved by 1.9% with 15% higher labeling cost over the method (ii).With the method (iv), we confirm that the labeling cost of the proposed method (iii) is reduced by 95% ($= (1 - 0.15/3) \times 100$), although the F1 score is degraded by 1.3%. This result confirms that weakly labeled learning is effective for learning BLSTM with reduction of the hand-labeling cost by 95%.

V. Conclusion

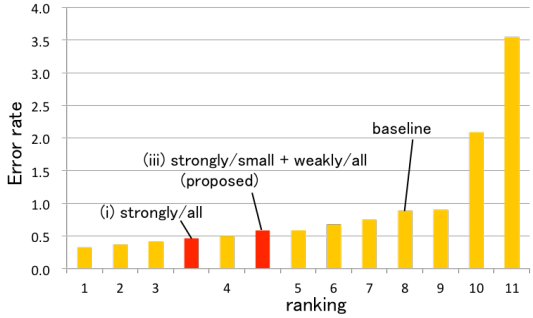
In this paper, we proposed weakly labeled learning of BLSTM-CTC to reduce the hand-labeling cost of learning



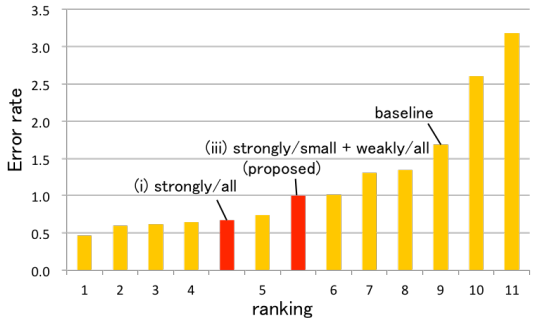
(a) Segment-based F1 score.



(b) Event-based F1 score.



(c) Segment-based error rate.



(d) Event-based error rate.

Fig. 11. Comparison of the performance of the proposed method with the methods submitted to DCASE 2016 for each evaluation criterion.

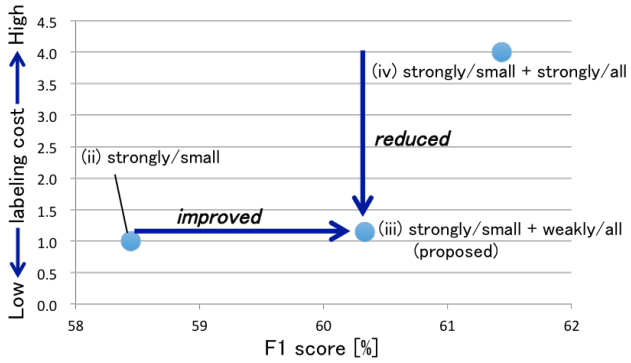


Fig. 12. Relationship between segment-based F1-score and labeling cost.

samples. BLSTM-CTC enables us to update the parameters of BLSTM by loss calculation using CTC, instead of the exact error calculation that cannot be conducted when using weakly labeled samples. We conducted the experiment using the development and evaluation datasets provided by DCASE 2016 Task 2. As a result, the proposed method improved the segment-based F1 score by 1.9% compared with the initial learning mentioned above. Furthermore, it succeeded in reducing the labeling cost by 95%, although the F1 score was degraded by 1.3%, comparing with additional learning using a large amount of strongly labeled samples. This result confirms that our weakly labeled learning is effective for learning BLSTM with a low hand-labeling cost.

REFERENCES

[1] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," Proc. workshop on machine listening in multisource environments, pp. 36–40, 2011.
[2] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Accouctic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries," DCASE2016 Challenge, Tech. Rep., 2016.
[3] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," Proc. IEEE IJCNN, pp. 1–7, 2015.
[4] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Roux and K. Takeda, "Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection," DCASE2016 Challenge, Tech. Rep., 2016.
[5] DCASE 2016, http://www.cs.tut.fi/sgn/arg/DCASE 2016/.
[6] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," Proc. ICML, pp. 369–376, 2006.
[7] C. L. Giles, G. M. Kuhn, and R. J. Williams, "Dynamic recurrent neural networks: Theory and applications," IEEE trans. neural networks, Vol. 5, No. 2, pp. 153–156, 1994.
[8] Y.Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient decent is difficult," IEEE trans. neural networks, Vol. 5, No. 2, pp. 157–166, 1994.
[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, Vol. 8, No. 9, pp. 1735–1780, 1997.
[10] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," Proc. ICANN, Vol. 2 pp. 799–804, 2005.
[11] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," Applied Sciences, pp.1–17, 2016.
[12] T. Komatsu and R. Kondo, "Detection of anomaly acoustic scenes based on a temporal dissimilarity model," Proc. ICASSP, pp.376–380, 2017.
[13] T. Komatsu, Tani, Takahiro Toizumi, Narisetty Chaitanya, Masanori Kato, Yumi Arai, Osamu Hoshuyama, Yuzo Senda and Reishi Kondo, "An acoustic monitoring system and its field trials," Proc. APSIPA, pp. 1–6, 2017.