Applied Acoustics 196 (2022) 108882

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance

Keisuke Imoto [a,*], Sakiko Mishima [b], Yumi Arai [b], Reishi Kondo [b]

[a] Doshisha University, Japan
[b] NEC Corporation, Japan

ARTICLE INFO

Article history:
Received 9 December 2021
Received in revised form 13 May 2022
Accepted 14 June 2022
Available online 27 June 2022

Keywords:
Sound event detection
Sound duration
Inactive frame
Data imbalance
Asymmetric focal loss
Focal batch Tversky loss

ABSTRACT

Sound event detection (SED) is a major topic in machine listening research. In many SED methods, a segmented time frame is considered as one data sample for model training. The duration of a sound event depends strongly on the event class, for example, the sound event "fan" is a long-lasting sound, whereas the sound events "mouse clicking" and "glass jingling" are instantaneous sounds. The difference in time duration between sound event classes makes a significant difference in the number of data samples between event classes; therefore, it causes a severe data imbalance problem in SED. Moreover, there are many more inactive time frames of sound events than active frames because most sound events are likely to occur occasionally. This also causes a serious data imbalance problem between active and inactive frames of sound events. In this paper, we study in detail the impact of the sound duration and inactive frames on the detection performance of sound events by introducing five loss functions: simple reweighting loss, inverse frequency loss, class-balanced loss, asymmetric focal loss, and focal batch Tversky loss. Evaluation experiments using the TUT Acoustic Scenes 2016/2017 and Sound Events 2016/2017 datasets show that in SED, inactive frames tend to overwhelm the model training, and the data imbalance problem between active and inactive frames is more severe than that between sound event classes. The evaluation experiments also show that the introduced loss functions can alleviate these data imbalance problems and improve the SED performance considerably.

1. Introduction

Sound event detection (SED) is a major task in environmental sound analysis, and it identifies types of sound event and their onset/offset in an audio recording [1,2]. Recently, many researchers have addressed SED because it has great potential for various applications in the field of artificial intelligence based on sounds, such as life logging, machine monitoring, automatic surveillance, media retrieval, and biomonitoring systems [3–8].

One approach to SED is non-negative matrix factorization (NMF) [9,10]. In the NMF-based SED approach, a sound is decomposed into a product of a basis matrix and activation matrix, where each basis vector and activation vector respectively represent each sound event and its active duration. More recently, many methods using neural networks, such as a convolutional neural network (CNN) [11], recurrent neural network (RNN) [12], convolutional recurrent neural network (CRNN) [13], and a Transformer-based neural network [14,15], have also been widely proposed. In these neural-network-based methods, a sound clip is segmented into short time frames (e.g., 40 ms), and each time frame is regarded as one data sample for model training and evaluation. Fig. 1 is an illustration of durations of sound events and shows that the durations of sound events strongly depending on the sound event class. Table 1 and Fig. 2 show the average duration of each sound event instance with a standard deviation and the total number of time frames covered by sound events in development datasets used for evaluation experiments (TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016 [16,17]), respectively. In these datasets, the number of frames in the sound event "mouse clicking," which has an average length of 0.14 s, is only 1,163 ($\approx 1.16 \times 10^3$), whereas that in the sound event "fan," which has an average length of 29.99 s, reaches 116,837 ($\approx 1.16 \times 10^5$). Therefore, the difference in time duration between sound event classes may cause a serious data imbalance problem in SED. Moreover, Figs. 1 and 2 indicate that there is a much larger difference in the number of time frames between active and inactive frames in sound events. In the development dataset used for evaluation experiments, the total number of active frames is 5.28×10^5 and that of inactive frames is 1.24×10^7; therefore, this difference also causes a serious

* Corresponding author.

https://doi.org/10.1016/j.apacoust.2022.108882
0003-682X/

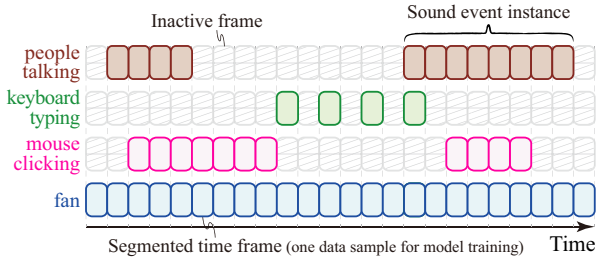

Fig. 1. Examples of active/inactive sound events and number of data samples.

Table 1
Average duration of one sound event instance in training datasets used for evaluation experiments (TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016 [16, 17]).

Sound event	Duration	Dataset
(object) banging	0.78 ± 0.67 s	TUT SE 2016
(object) squeaking	0.74 ± 0.16 s	TUT AS 2016
cupboard	0.65 ± 0.60 s	TUT SE 2016
(object) snapping	0.46 ± 0.30 s	TUT SE 2016
drawer	0.80 ± 0.35 s	TUT SE 2016
breathing	0.43 ± 0.14 s	TUT AS 2016
glass jingling	0.80 ± 0.46 s	TUT SE 2016
mouse wheeling	0.16 ± 0.06 s	TUT AS 2016
cutlery	0.74 ± 0.53 s	TUT SE 2016
mouse clicking	0.14 ± 0.08 s	TUT AS 2016
brakes squeaking	1.65 ± 1.97 s	TUT SE 2017
dishes	1.24 ± 1.12 s	TUT SE 2016
wind blowing	6.09 ± 5.84 s	TUT SE 2016
water tap running	5.92 ± 7.03 s	TUT SE 2016
(object) rustling	2.24 ± 3.40 s	TUT AS 2016, SE 2016
washing dishes	4.15 ± 3.75 s	TUT SE 2016
keyboard typing	0.21 ± 0.22 s	TUT AS 2016
children	6.87 ± 2.06 s	TUT SE 2016/2017
(object) impact	0.35 ± 0.60 s	TUT AS 2016, SE 2016
large vehicle	14.68 ± 7.35 s	TUT SE 2017
people talking	4.09 ± 6.28 s	TUT AS 2016, SE 2016/2017
bird singing	7.63 ± 8.49 s	TUT SE 2016
people walking	6.63 ± 8.78 s	TUT AS 2016, SE 2016/2017
car	6.88 ± 4.72 s	TUT SE 2016/2017
fan	29.99 ± 0.01 s	TUT AS 2016

imbalance problem between active and inactive data samples. To make matters worse, this data imbalance becomes more severe as the number of sound event classes increases. This is because inactive frames are polled over the sound event class, whereas active frames are counted for each sound event class.

To address the imbalance problem in SED, some conventional methods have been proposed [18–20]. For instance, Chen and Jin have proposed a method of detecting rare sound events using data augmentation [18]. Wang et al. have proposed a few-shot SED method based on metric learning [19]. Dinkel and Yu have proposed a method of SED using temporal subsampling within a CRNN [20]. However, the conventional works have not comprehensively investigated the impact of data imbalance caused by the difference in time duration between sound event classes and active/inactive frames on SED performance. To explore the impact of data imbalance, we previously conducted a brief study [21]. In that study, we introduced loss functions, such as simple reweighting loss, asymmetric focal loss, and focal batch Tversky loss, to alleviate the data imbalance problem in SED tasks, and we evaluated the SED performance with the loss functions. The experimental results showed that the overall SED performance was improved using the loss functions. In this paper, in addition to a more detailed discussion of the introduced loss functions, we investigate how the data imbalance impacts the SED performance for each sound event in detail. Moreover, we newly introduce a class-balanced loss function that mitigates the data imbalance considering acoustic properties of active/inactive frames and evaluate in detail the SED performance with the loss function.

The rest of this paper is organized as follows. In Section 2, we discuss the conventional SED method using the sigmoid cross-entropy loss function. In Section 3, we introduce five loss reweighting techniques to relieve the data imbalance problem in SED. In Section 4, we investigate how the imbalance between sound event classes and/or active and inactive frames impacts the SED performance on the basis of experimental results. Finally, in Section 5, we conclude this paper.

2. Sound event detection using binary cross-entropy loss

Let f and $\boldsymbol{\theta}$ denote a SED model and the model parameter, respectively. SED is a task that predicts sound event labels $\hat{\mathbf{Z}}$ in a sound clip as

$$\hat{\mathbf{Z}} = I[f(\mathbf{X}, \boldsymbol{\theta}) \geqslant \phi] \in \{0,1\}^{T\times M}, \tag{1}$$

where $I, \mathbf{X}, \phi, T$, and M are the indicator function, the acoustic feature, the detection threshold, the number of time frames in the sound clip, and the number of sound event classes, respectively. We preliminarily determine the model parameter $\boldsymbol{\theta}$ using the training dataset $\mathscr{D} = \{(\mathbf{X}_1, \mathbf{Z}_1), \ldots, (\mathbf{X}_l, \mathbf{Z}_l), \ldots, (\mathbf{X}_L, \mathbf{Z}_L)\}$. Here, $\mathbf{X}_l$ is the acoustic feature of the lth sound clip in $\mathscr{D}$ and $\mathbf{Z}_l = \{\mathbf{z}_{l,1}, \ldots \mathbf{z}_{l,t}, \ldots, \mathbf{z}_{l,T}\}$ indicates a sequence of multi–hot vectors $\mathbf{z}_{l,t} \in \{0,1\}^M$ in the lth sound clip over the M sound event class. For the acoustic feature $\mathbf{X}_l$, we often use the power spectrogram, mel-band energy, and mel-frequency cepstral coefficients (MFCCs). As the SED model f, the CNN, CRNN, or Transformer-based neural network is usually used. The model parameter $\boldsymbol{\theta}$ is trained by the binary cross-entropy (BCE) loss $\mathscr{L}_{\mathrm{BCE}}(\boldsymbol{\theta})$, which is calculated by summing the cross-entropy loss over all time frames and sound event classes:

$$\begin{aligned}\mathscr{L}_{\mathrm{BCE}}(\boldsymbol{\theta}) &= -\sum_{t=1}^{T}\{\mathbf{z}_t \log(\mathbf{y}_t) + (1-\mathbf{z}_t)\log(1-\mathbf{y}_t)\} \\ &= -\sum_{t,m=1}^{T,M}\{z_{t,m}\log(y_{t,m}) + (1-z_{t,m})\log(1-y_{t,m})\},\end{aligned} \tag{2}$$

where $z_{t,m}$ is the target label in time frame t and is 1 if acoustic event m is active in time frame t and 0 otherwise. $y_{t,m}$ is the prediction of sound event m in time frame t. The sound clip index l is omitted for simplification. Since the sound event duration varies highly depending on the event class, the model parameter estimation based on the BCE loss leads to a data imbalance between sound event classes. Moreover, as shown in Figs. 1 and 2, because the number of inactive frames of sound events is much larger than that of active frames, the model parameter estimation tends to be overwhelmed by the inactive frames. Consequently, active frames tend to be ignored in the model training.

3. Loss function considering data imbalance

In this work, we apply five loss functions that can control the contribution to the model training of short/long sound events and active/inactive frames. One way to address the data imbalance problem is to use an oversampling technique, such as the synthetic minority oversampling technique (SMOTE) [22] or adaptive synthetic sampling (ADASYN) [23]. However, these methods are difficult to apply to polyphonic event detection tasks. Another approach to deal with the data imbalance problem is an undersampling-based technique, which thins out data samples in the major class; however, the undersampling-based technique may lose information in discarded data samples. Thus, we adopt loss-function-based approaches, which can finely control the contribution to the model training of major and minor classes.

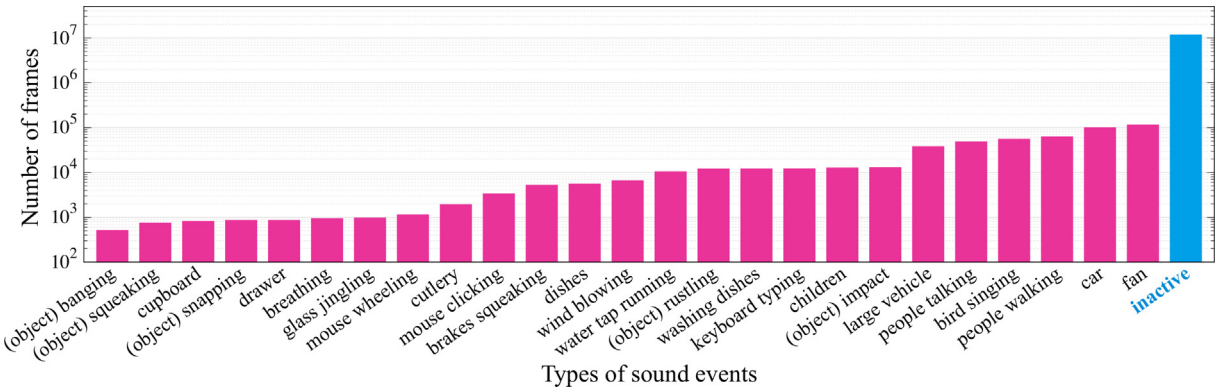

Fig. 2. Numbers of frames of active/inactive sound events in dataset used for evaluation experiments.

3.1. Simple Reweighting Loss (SRL)

To investigate the impact of a very large number of inactive frames on the SED performance, we consider the following simple reweighting loss (SRL):

$$\mathcal{L}_{\mathrm{SRL}}(\boldsymbol{\theta}) = -\sum_{t,m=1}^{T,M} \{\alpha z_{t,m} \log(y_{t,m}) + \beta(1 - z_{t,m}) \log(1 - y_{t,m})\}, \tag{3}$$

where $\alpha \in [0, \infty)$ and $\beta \in [0, \infty)$ are the reweighting factors. In this work, we set α as 1.0.

3.2. Inverse Frequency Loss (IFL)

To investigate the impact of data imbalance between sound event classes, we also consider the following reweighting loss on the basis of the inverse frequency of sound events (IFL):

$$\mathcal{L}_{\mathrm{IFL}}(\boldsymbol{\theta}) = -\sum_{t,m=1}^{T,M} \left\{ \left(\frac{C}{N_m + C} \right)^{\gamma} z_{t,m} \log(y_{t,m}) + (1 - z_{t,m}) \log(1 - y_{t,m}) \right\}, \tag{4}$$

where $\gamma \in [0, \infty)$, N_m, and C are the weighting factor, the number of frames of a sound event m in a training batch, and a constant, respectively. The IFL can reweight the contribution of each sound event to model training in accordance with the frequency and enable a more robust model training with an imbalanced dataset between sound event classes.

3.3. Class-Balanced Loss (CBL)

Since a data sample is often a near duplicate of other samples in the acoustic feature space, a reweighting loss based on the number of time frames may not properly mitigate the data imbalance. To reweight loss functions depending not on the number of time frames but on the "effective number" of samples E_m, which are not overlapped with other samples in the feature space, we apply the following CBL [24] to the SED task:

$$\begin{aligned}\mathcal{L}_{\mathrm{CBL}}(\boldsymbol{\theta}) = -\sum_{t,m=1}^{T,M} \Bigg\{ & \frac{1 - \xi_m}{1 - \xi_m^{N_m}} z_{t,m} \log(y_{t,m}) \\ & + \frac{1 - \xi_m}{1 - \xi_m^{N_{\varnothing m}}} (1 - z_{t,m}) \log(1 - y_{t,m}) \Bigg\},\end{aligned} \tag{5}$$

where $\xi_m = (E_m - 1)/E_m \in [0, 1.0)$. N_m and $N_{\varnothing m}$ are the number of frames in sound event class m and the number of inactive frames in sound event class m, respectively.

3.4. Asymmetric Focal Loss (AFL)

Many long-duration sound events (e.g., "fan" and "car") and inactive frames are stationary, that is, they do not have large variations in their acoustic features for a large number of frames. On the other hand, several short-duration sound events (e.g., "mouse clicking" and "glass jingling") have more than one audio pattern, such as attack, decay, and release parts. This indicates that the model training of long-duration sound events is relatively easy compared with that of instantaneous sounds. To control the training weight of the sound event model in accordance with the progress of model training, the use of focal loss has been proposed [25,26]. In this paper, we introduce the following asymmetric focal loss (AFL), which enables the separate control of the focusing factor of active and inactive frames.

$$\begin{aligned}\mathcal{L}_{\mathrm{AFL}}(\boldsymbol{\theta}) = -\sum_{t,m=1}^{T,M} \Big\{ & (1 - y_{t,m})^{\gamma} z_{t,m} \log(y_{t,m}) \\ & + (y_{t,m})^{\zeta} (1 - z_{t,m}) \log(1 - y_{t,m}) \Big\}\end{aligned} \tag{6}$$

Here, γ and ζ are the weighting parameters that control the focusing weights of active and inactive frames, respectively. When we set large values for the weighting parameters γ and ζ, the loss of active and inactive frames is more down-weighted depending on the prediction error.

3.5. Focal Batch Tversky Loss (FBTL)

As another way to address the data imbalance between active and inactive frames in sound events, we introduce the focal batch Tversky loss (FBTL) $\mathcal{L}_{\mathrm{FBTL}}(\boldsymbol{\theta})$. The FBTL is an extended loss function of the dice loss (DL) [27–29] and Tversky loss (TL) [30,31], which directly optimize the model to maximize the F-score and do not consider the true negative samples as follows.

$$\mathcal{L}_{\mathrm{DL}}(\boldsymbol{\theta}) = 1 - \frac{\sum_{t,m=1}^{T,M} 2 y_{t,m}^{(1)} z_{t,m}^{(1)} + \eta}{\sum_{t,m=1}^{T,M} y_{t,m}^{(1)} + \sum_{t,m=1}^{T,M} z_{t,m}^{(1)} + \eta}, \tag{7}$$

$$\mathcal{L}_{\mathrm{TL}}(\boldsymbol{\theta}) = 1 - \frac{\sum_{t,m=1}^{T,M} y_{t,m}^{(1)} z_{t,m}^{(1)} + \eta}{\sum_{t,m=1}^{T,M} \alpha y_{t,m}^{(1)} + \sum_{t,m=1}^{T,M} \beta z_{t,m}^{(1)} + \eta}, \tag{8}$$

where $y_{t,m}^{(1)}$ and $z_{t,m}^{(1)}$ are the prediction and sound event label for the active frame, respectively. That is, $y_{t,m}^{(1)}$ and $z_{t,m}^{(1)}$ correspond to the precision and recall, respectively. $\alpha \in [0, 1.0]$ and $\beta \in [0, 1.0]$ are the tradeoff

parameters between false negative and false positive samples, where $\alpha + \beta = 1.0$. η is a smoothing parameter. Because DL and TL do not consider the true negative samples, they can both prevent the model training from being overwhelmed by the inactive frames. In this paper, we further introduce the idea of focal loss and batch optimization into TL, and apply the following FBTL to the SED task:

$$\mathscr{L}_{\mathrm{FBTL}}(\boldsymbol{\theta}) = 1 - \frac{\sum_{l,t,m=1}^{B,T,M} (1 - y_{l,t,m}^{(1)})^{\gamma} y_{l,t,m}^{(1)} z_{l,t,m}^{(1)} + \eta}{\sum_{l,t,m=1}^{B,T,M} \alpha (1 - y_{l,t,m}^{(1)})^{\gamma} y_{l,t,m}^{(1)} + \sum_{l,t,m=1}^{B,T,M} \beta z_{l,t,m}^{(1)} + \eta}, \tag{9}$$

where B is the number of sound clips in each batch.

4. Experiments

4.1. Experimental conditions

To evaluate how the data imbalance between sound event classes and active/inactive frames impacts SED performance, we conducted five experiments using various loss functions and network architectures. For the evaluation data, we constructed a dataset composed of TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016/2017 [16,17]. From these four datasets, we selected a total of 266 min of sound clips (192 min for the development set; 74 min for the evaluation set) including the 25 types of sound event listed in Fig. 2. Each sound clip has a length of 10 s with a sampling rate of 16 kHz. Details of the dataset for this experiment are available in [32]. Note that the TUT Sound Events and Acoustic Scenes databases were collected not for rare sound event detection but for the analysis of real-life sounds; therefore, the analysis of seriously imbalanced data is a common problem in SED.

As acoustic features, we extracted the 64-dimensional log mel-band energy, which was calculated every 40 ms with a step size of 20 ms. For the baseline network architecture, we used CNN-BiGRU, which is widely used as a baseline SED system, such as in DCASE2018 challenge task 4 [33]. To train the model parameters, we used a fourfold cross-validation setup for the development set and selected model parameters that achieved the best performance. For each method, the evaluation experiment was conducted 10 times using random initial values for model parameters. The SED performance was evaluated using the frame-based micro-Fscores, macro-Fscores, micro-ROC AUC, and macro-ROC AUC. Other detailed experimental conditions are listed in Table 2.

Table 2
Experimental conditions.

Length of sound clip	10 s
Network for CNN-BiGRU	3 CNN + 1 BiGRU + 1 dense
# channels of CNN layers	128, 128, 128
Filter size	3 × 3, 3 × 3, 3 × 3
Pooling size	1 × 8, 1 × 4, 1 × 2 (max pooling)
# units in GRU layer	32
# units in dense layer	32
Network for Transformer	3 CNN + 2 Transformer encoder + 2 dense
# attention heads	32
Activation function	Leaky ReLU
Optimizer	RAdam [34]
Detection threshold	0.5
Constant number C	500
Smoothing parameter η	1.0

4.2. Experimental results

4.2.1. Impact of inactive frames on event detection performance

Figs. 3 and 4 show the average macro- and micro-Fscores with BCE loss, SRL, and AFL ($\gamma = 0.0$) for various weighting factors. In this experiment, we used CNN-BiGRU as the network structure. To evaluate the impact of inactive frames on the SED performance, distinct from the impact of the imbalance between sound event classes, we set $\gamma = 0$ for AFL. The experimental results show that when we down-weight the loss for inactive frames, both micro- and macro-Fscores tend to improve. This indicates that the inactive frames tend to overwhelm model training, which leads to active sound events being ignored in model training. Fig. 5 shows the average macro- and micro-Fscores using FBTL, which outperforms those of BCE loss when $\gamma < 0.1$. This result also implies that a SED model trained to predict inactive frames accurately is likely to ignore active frames.

4.2.2. Impact of imbalance between sound event classes on detection performance

The average macro- and micro-Fscores with BCE loss, IFL, and AFL ($\zeta = 0.0$) for various weighting factors γ are shown in Fig. 6. To evaluate the impact of the imbalance between sound event classes on the SED performance, distinct from the impact of inactive frames, we set $\zeta = 0$ for AFL. We set the constant C as 500, which is equal to the number of frames in one sound clip. Note that we have conducted the evaluation experiments with other settings of the constant C, with the result that the choice of constant C does not significantly affect the SED performance. The experimental results show that both macro- and micro-Fscores do not improve markedly even when the losses are weighted to be more balanced between sound event classes. This implies that the data imbalance between sound event classes has less effect on the detection performance of sound events than the data imbalance between active and inactive frames. These results imply that, in SED, the data imbalance between active and inactive frames is a more severe problem, which should be addressed preferentially.

4.2.3. Variations in audio patterns in sound events

Fig. 7 shows the average macro- and micro-Fscores with BCE loss and CBL for various ξ_m, where we set the same values for all ξ_m. We also conducted experiments with different effective numbers E_m for each acoustic event class m; however, the results did not improve markedly compared with those shown in Fig. 7. The results indicate that there are about 50 to 100 audio patterns in each sound event, and the different numbers of frames between sound events do not necessarily mean different variations in audio patterns.

4.2.4. Comparison experiments with conventional methods

We conducted evaluation experiments to compare the SED performance of our methods with those using various network architectures and loss functions, such as SED using an α min–max subsampling method within a CRNN [20], batch dice loss–based SED [31,36], the multitask learning of SED and sound activity detection (MTL of SED & ASD) [37], and Transformer-based SED [14,15]. For the α min–max subsampling method within the CRNN and MTL of SED & ASD, we applied the same settings as in [20,37], respectively. For the Transformer-based method, we used three CNN layers with the same structure as the CNN-BiGRU, followed by two Transformer encoder layers and two dense layers with the leaky ReLU activation.

The micro- and macro-Fscores, and micro- and macro-ROC AUC scores for the conventional methods and our methods are presented in Table 3. The results show that down-weighting the loss for inactive frames using SRL, AFL, and FBTL improves both Fscores

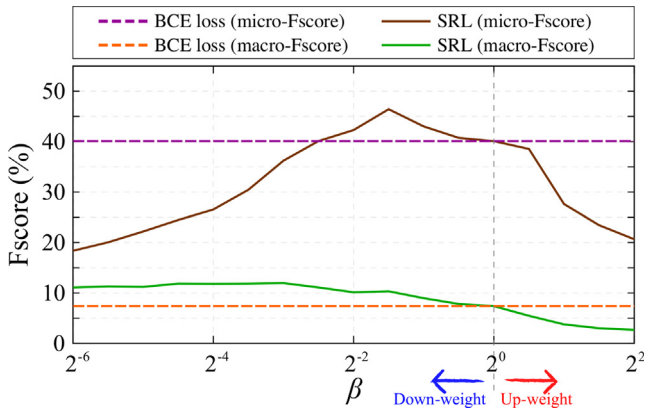

Fig. 3. Average Fscores for SRL and BCE loss with various weighting factors β.

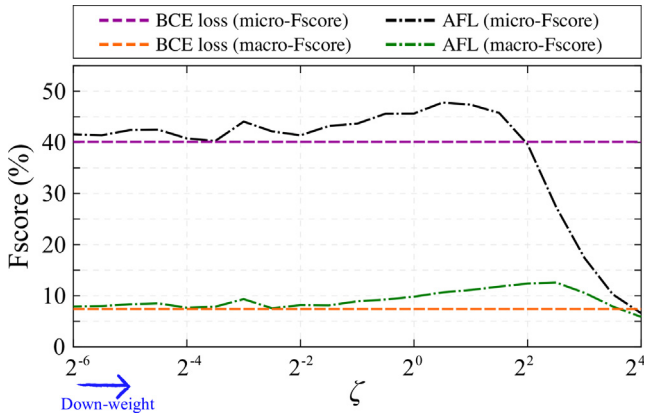

Fig. 4. Average Fscores for AFL and BCE loss with various weighting factors ζ. For AFL, we set $\gamma = 0.0$.

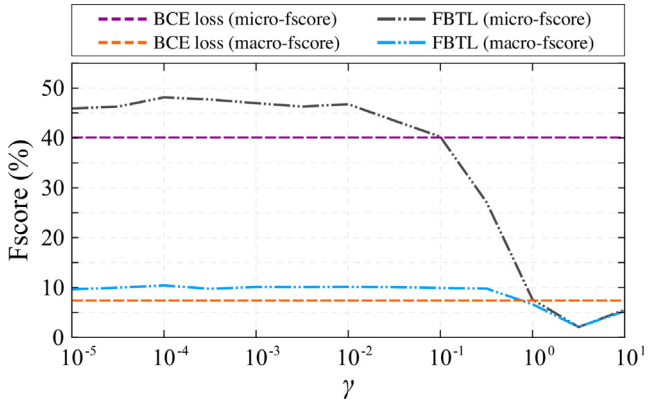

Fig. 5. Average Fscores for FBTL and BCE loss with various weighting factors γ.

and ROC AUC scores to a greater extent than the conventional methods. In particular, even SRL with CNN-BiGRU surpasses the performance of using the BCE loss with the Transformer-based network, which is the state-of-the-art model for SED. On the other hand, FBTL with CNN-BiGRU does not outperform the conventional BCE with CNN-BiGRU in terms of micro- or macro-ROC AUC scores. This is because FBTL does not consider the true negative samples during model training; thus, the false positive rate is likely to be greater than that of the BCE-loss-based methods. By reweighting both types of imbalance using AFL and applying the Transformer-

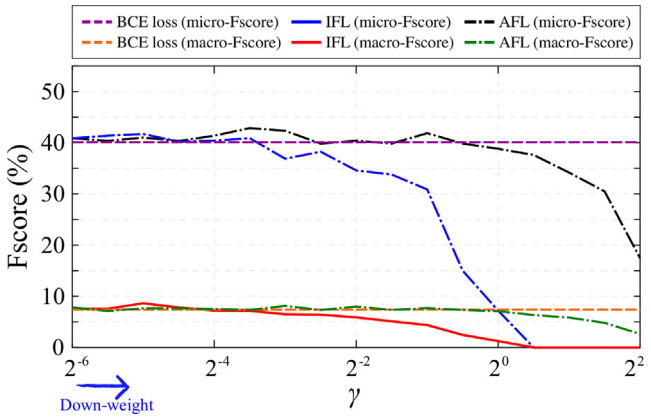

Fig. 6. Average Fscores for BCE loss, IFL, and AFL with various weighting factors γ. For AFL, we set $\zeta = 0.0$.

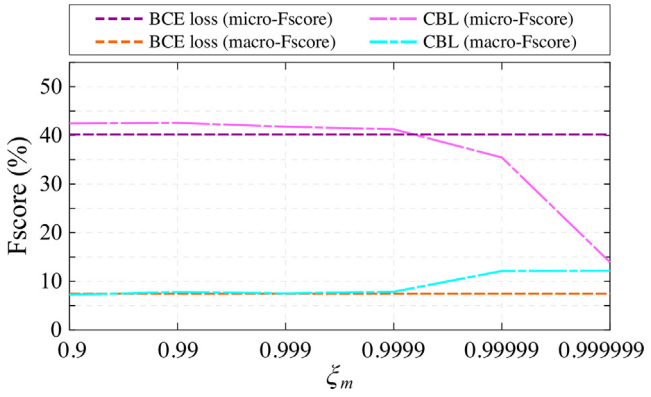

Fig. 7. Average Fscores for CBL with various ξ_m. In this experiment, we set the same value for all ξ_m.

based network with the adaptive thresholding technique [35], the SED performance is finally improved by 10.85 and 5.08 percentage points compared with the baseline system in the macro- and micro-Fscores, respectively.

4.2.5. Detailed detection results for each sound event

Table 4 shows the average Fscores of selected sound events for the conventional methods and our methods. In many sound events, IFL and AFL ($\gamma = 0.125, \zeta = 0.0$) do not considerably improve the SED performance, whereas AFL ($\gamma = 0.0625, \zeta = 1.0$) outperforms the other methods. For example, the detection performance of the sound event "water tap running" is still low when we alleviate the data imbalance between sound event classes by IFL and AFL ($\gamma = 0.125, \zeta = 0.0$), whereas it improves considerably when we weight the imbalances both between sound event classes and between active/inactive frames by AFL ($\gamma = 0.0625, \zeta = 1.0$). This also supports the observation that, in SED, the data imbalance between active and inactive frames is a more serious problem and should be addressed preferentially. On the other hand, Table 4 also shows that even when we apply the weighting methods, the detection performance of sound events with a small number of frames, such as "glass jingling" and "mouse clicking," is only slightly improved. Thus, further study of these sound events is necessary in future works.

To investigate the event detection results in more detail, we show a sample annotation and the detection results in Fig. 8. The results show that the conventional CNN-BiGRU w/ BCE and CNN-BiGRU w/ IFL seriously overlook many sound events. CNN-BiGRU

Table 3
Average SED performance for various loss functions and networks.

Method	Micro-Fscore	Macro-Fscore	Micro-ROC AUC	Macro-ROC AUC
[Conventional methods]				
CNN-BiGRU w/ BCE loss (**Baseline**)	40.10%	7.39%	89.15%	65.85%
CNN-BiGRU w/ α min–max subsampling & BCE loss	44.12%	9.35%	90.27%	67.55%
CNN-BiGRU w/ batch dice loss	45.06%	9.79%	86.99%	63.89%
MTL of SED & SAD w/ BCE loss	43.35%	8.64%	91.40%	70.97%
Transformer w/ BCE loss	45.15%	9.27%	90.32%	66.64%
[Loss weighting between active and inactive frames]				
CNN-BiGRU w/ SRL ($\beta = 0.3535$)	46.44%	10.34%	91.07%	69.31%
CNN-BiGRU w/ AFL ($\gamma = 0.0, \zeta = 1.414$)	47.78%	10.65%	92.35%	76.18%
CNN-BiGRU w/ FBTL ($\alpha = 0.6, \beta = 0.4, \gamma = 0.001$)	46.97%	10.28%	87.95%	65.08%
[Loss weighting between sound event classes]				
CNN-BiGRU w/ IFL (γ=0.03125, $C = 500$)	41.70%	8.64%	89.89%	66.46%
CNN-BiGRU w/ AFL ($\gamma = 0.125, \zeta = 0.0$)	42.33%	8.13%	91.08%	70.46%
[Loss weighting between event classes + active/inactive frames]				
CNN-BiGRU w/ CBL ($\xi = 0.99$)	42.67%	7.79%	91.87%	74.30%
CNN-BiGRU w/ AFL ($\gamma = 0.0625, \zeta = 1.0$)	48.29%	10.46%	92.62%	77.03%
CNN-BiGRU w/ AFL + adaptive thresholding [35]	49.61%	12.40%	92.62%	77.03%
Transformer w/ AFL ($\gamma = 0.0625, \zeta = 1.0$)	**49.14%**	**11.11%**	**92.74%**	**77.49%**
Transformer w/ AFL + adaptive thresholding [35]	**50.95%**	**12.47%**	**92.74%**	**77.49%**

Table 4
Average Fscores for selected sound events.

Method	glass jingling	mouse clicking	wind blowing	water tap running	(object) rustling	washing dishes	bird singing	car	fan
CNN-BiGRU w/ BCE	0.00%	0.00%	0.00%	43.23%	0.13%	0.41%	17.79%	43.85%	68.96%
CNN-BiGRU w/ SRL	0.00%	0.00%	0.00%	69.37%	1.98%	5.09%	32.69%	49.09%	84.27%
CNN-BiGRU w/ AFL ($\gamma = 0.0, \zeta = 1.414$)	0.00%	0.00%	0.00%	73.90%	1.47%	3.94%	34.50%	47.96%	87.46%
CNN-BiGRU w/ FBTL	0.00%	0.00%	0.00%	60.31%	0.00%	3.73%	**45.35%**	48.88%	78.35%
CNN-BiGRU w/ IFL	0.00%	0.00%	0.00%	32.76%	1.00%	0.39%	19.28%	44.42%	81.62%
CNN-BiGRU w/ AFL ($\gamma = 0.125, \zeta = 0.0$)	0.00%	0.00%	0.00%	40.29%	0.00%	0.74%	20.15%	43.35%	85.14%
CNN-BiGRU w/ CBL ($\xi = 0.99$)	0.00%	0.00%	0.00%	32.40%	0.56%	0.96%	18.88%	44.00%	86.11%
CNN-BiGRU w/ AFL ($\gamma = 0.0625, \zeta = 1.0$)	**1.07%**	0.00%	2.21%	74.64%	**9.25%**	10.55%	28.08%	45.71%	85.19%
Transformer w/ AFL ($\gamma = 0.0625, \zeta = 1.0$)	0.00%	**0.06%**	**7.46%**	**79.31%**	0.19%	**26.71%**	18.40%	**49.95%**	**93.07%**

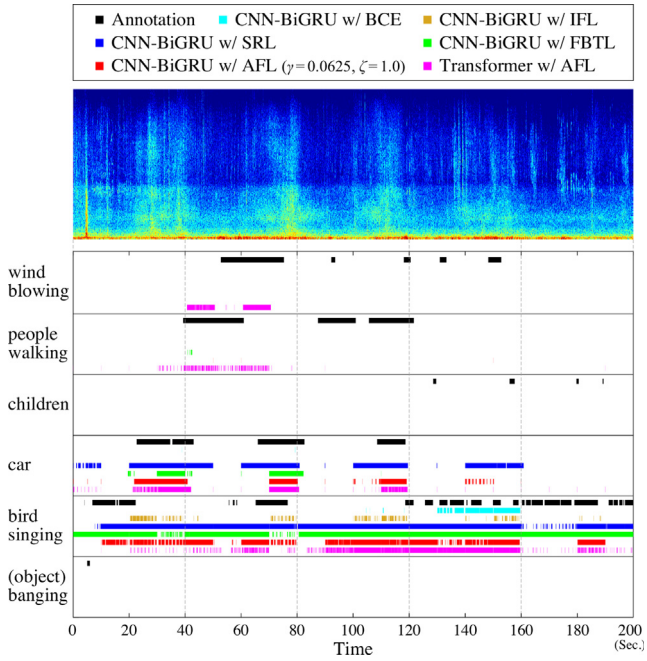

Fig. 8. Sample annotation and event detection results for sounds recorded in residential area. Only sound events occurring in the annotations are depicted.

w/ SRL and CNN-BiGRU w/ FBTL can detect the sound events "car" and "bird singing", and there are many false positive frames. This is because CNN-BiGRU w/ SRL and CNN-BiGRU w/ FBTL tend to depreciate true negatives. On the other hand, Transformer w/ AFL ($\gamma = 0.0625, \zeta = 1.0$) can detect the sound events "people walking" and "wind blowing," and there are fewer false positive frames than in the other methods.

5. Conclusion

In this work, we investigated the impact of the data imbalance between sound event classes and between active/inactive frames on SED performance. To investigate the impact of the data imbalance, we introduced five loss functions, SRL, IFL, CBL, AFL, and FBTL, into SED, which can reweight the losses to alleviate the imbalance of the contribution to model training. The experimental results using TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016/2017 indicated that the inactive frames tend to overwhelm model training, and consequently, the trained model is not likely to detect active sound events. The experimental results also show that the imbalance between sound event classes has less impact on SED performance than that between active and inactive frames. Therefore, in SED, the data imbalance between active and inactive frames is a more serious problem, which should be addressed pref-

erentially. Finally, the results indicate that the SED method with AFL is a promising way to inhibit the adverse impact caused by the data imbalance and considerably improve the SED performance.

CRediT authorship contribution statement

Keisuke Imoto: Conceptualization, Methodology, Software, Writing - original draft. **Sakiko Mishima:** Investigation, Formal analysis, Writing - review & editing. **Yumi Arai:** Formal analysis, Writing - review & editing. **Reishi Kondo:** Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Keisuke Imoto reports financial support was provided by Japan Society for the Promotion of Science.

Acknowledgement

This work was supported by JSPS KAKENHI Grant No. JP19K20304 and the grant-in-aid from Harris Science Research Institute, Doshisha University.

References

[1] Virtanen T, Plumbley M, Ellis D, editors. Computational Analysis of Sound Scenes and Events. Springer; 2017.
[2] Imoto K. Introduction to acoustic event and scene analysis. Acoust Sci Technol 2018;39(3):182–8.
[3] K. Imoto, S. Shimauchi, H. Uematsu, H. Ohmuro, User activity estimation method based on probabilistic generative model of acoustic event sequence with user activity and its subordinate categories, Proc. INTERSPEECH (2013).
[4] Geiger J, Helwani K. Improving event detection for audio surveillance using Gabor filterbank features. Proc European Signal Processing Conference (EUSIPCO) 2015:714–8.
[5] Salamon J, Bello JP, Farnsworth A, Robbins M, Keen S, Klinck H, Kelling S. Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLoS One 2016;11(11).
[6] Y. Okamoto, K. Imoto, N. Tsukahara, K. Nagata, K. Sueda, R. Yamanishi, Y. Yamashita, Crow call detection using gated convolutional recurrent neural network, Proc. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP) (2020) 171–174.
[7] Fan J, Nichols E, Tompkins D, Méndez AEM, Elizalde B, Pasquier P. Multi-label sound event retrieval using a deep learning-based Siamese structure with a pairwise presence matrix, Proc. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 3482–6.
[8] Koizumi Y, Kawaguchi Y, Imoto K, Nakamura T, Nikaido Y, Tanabe R, Purohit H, Suefusa K, Endo T, Yasuda M, Harada N. Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. In: Proc. Detection and Classification of Acoustic Scenes and Events (DCASE). p. 81–5.
[9] Dessein A, Cont A, Lemaitre G. Real-time detection of overlapping sound events with non-negative matrix factorization. Matrix Inform Geometry 2013:341–71.
[10] Komatsu T, Toizumi T, Kondo R, Senda Y. Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries, Proc. In: Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE). p. 45–9.
[11] Hershey S, Chaudhuri S, Ellis DPW, Gemmeke JF, Jansen A, Moore RC, Plakal M, Platt D, Saurous RA, Seybold B, Slaney M, Weiss RJ, Wilson K. CNN architectures for large-scale audio classification. In: Proc IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 131–5.
[12] Hayashi T, Watanabe S, Toda T, Hori T, Roux JL, Takeda K. Duration-controlled LSTM for polyphonic sound event detection. IEEE/ACM Trans Audio, Speech, Language Process 2017;25(11):2059–70.
[13] Çakir E, Parascandolo G, Heittola T, Huttunen H, Virtanen T. Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans Audio, Speech, Language Process 2017;25(6):1291–303.
[14] Miyazaki K, Komatsu T, Hayashi T, Watanabe S, Toda T, Takeda K. Weakly-supervised sound event detection with self-attention. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 66–70.
[15] Kong Q, Xu Y, Wang W, Plumbley MD. Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization. IEEE/ACM Trans Audio, Speech, Language Process 2020;28:2450–60.
[16] Mesaros A, Heittola T, Virtanen T. TUT database for acoustic scene classification and sound event detection. Proc. European Signal Processing Conference (EUSIPCO) 2016:1128–32.
[17] Mesaros A, Heittola T, Diment A, Elizalde B, Shah A, Raj B, Virtanen T. DCASE 2017 challenge setup: Tasks, datasets and baseline system. Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE) 2017;2017:85–92.
[18] Chen Y, Jin H. Rare sound event detection using deep learning and data augmentation. Proc. INTERSPEECH 2019:619–23.
[19] Wang Y, Salamon J, Bryan NJ, Bello JP. Few-shot sound event detection. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 81–5.
[20] Dinkel H, Yu K. Duration robust weakly supervised sound event detection. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 311–5.
[21] Imoto K, Mishima S, Arai Y, Kondo R. Impact of sound duration and inactive frames on sound event detection performance. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 875–9.
[22] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.
[23] He H, Bai Y, Garcia EA, Li S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: Proc. IEEE International Joint Conference on Neural Networks (IJCNN). p. 1322–8.
[24] Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 9268–77.
[25] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proc. IEEE International Conference on Computer Vision (ICCV). p. 2980–8.
[26] Noh K, Chang JH. Joint optimization of deep neural network-based dereverberation and beamforming for sound event detection in multi-channel environments. Sensors 2020;20(7):1–13.
[27] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.
[28] Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. International Conference on 3D Vision (3DV). p. 565–71.
[29] Li X, Sun X, Meng Y, Liang J, Wu F, Li J. Dice loss for data-imbalanced NLP tasks. In: Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL). p. 465–76.
[30] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: Proc. International Workshop on Machine Learning in Medical Imaging (MLMI). p. 379–87.
[31] Kodym O, Spanel M, Herout A. Segmentation of head and neck organs at risk using CNN with batch dice loss. German Conference in Pattern Recognition (GCPR) 2018:105–14.
[32] URL: https://www.ksuke.net/dataset.
[33] Serizel R, Turpault N, Eghbal-Zadeh H, Shah AP. Large-scale weakly labeled semi-supervised sound event detection in domestic environments. In: Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE). p. 19–23.
[34] Liu L, Jiang H, He P, Chen W, Liu X, Gao J, Han J. On the variance of the adaptive learning rate and beyond. In: Proc. International Conference on Learning Representations (ICLR). p. 1–13.
[35] Xu Y, Kong Q, Wang W, Plumbley MD. Surrey-CVSSP system for DCASE2017 challenge task4, Technical report of task 4 of DCASE. Challenge 2017, 2017,:1–3.
[36] S. Park, S. Suh, Y. Jeong, Sound event localization and detection with various loss functions, Technical report of task 3 of DCASE Challenge 2020 (2020) 1–5.
[37] Pankajakshan A, Bear HL, Benetos E. Polyphonic sound event and sound activity detection: A multi-task approach. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). p. 323–7.