MODEL TRAINING THAT PRIORITIZES RARE OVERLAPPED LABELS FOR POLYPHONIC SOUND EVENT DETECTION

Rie Koga[1], *Sawa Takamuku*[1], *Keisuke Imoto*[2], *Naotake Natori*[1]

[1] AISIN CORPORATION, 1-1-20 Aomi, Koto-ku, Tokyo, Japan,
{rie.yamada, sawa.takamuku, naotake.natori}@aisin.co.jp
[2] Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe, Kyoto, Japan,
keisuke.imoto@ieee.org

ABSTRACT

In this study, we propose a model training method for polyphonic sound event detection (polyphonic SED) that prioritizes rare event label frames during multiple overlapping sound events. Multi-label classification typically utilized in polyphonic SED often fails to recognize such events. To overcome this problem, the proposed method is designed to represent event overlaps of rare labels easily without a complicated network structure. During model training, we periodically apply either binary cross-entropy loss (BCE) for multi-label classification or softmax cross-entropy loss (Softmax-CE) for multi-class classification. When multi-class classification is performed using Softmax-CE, the labels of the overlapping frames are reconstructed from the target labels to include the rarest ones and exclude the frequent ones. The model was evaluated on strongly labeled AudioSet data, from which only human voice segments were extracted. The proposed method achieves an improvement of 0.23 percentage points over the baseline, which only used the BCE, in terms of the mean average precision. In particular, the proposed method outperforms the baseline with respect to rare labels, with an average precision of 1.18 percentage points. The experimental results also demonstrate the effectiveness of the proposed method for both overlap of sound events and rare labels.

Index Terms— Polyphonic sound event detection, multi-label classification, multi-class classification

1. INTRODUCTION

Due to the advancements in deep learning, sound event detection (SED), which is a technique used for estimating the type and interval (onset and offset times) of sound events present in an acoustic signal, has recently attracted attention. Additionally, shared mobility services have become ubiquitous in many cities worldwide. For safety, they require surveillance of both the drivers and passengers inside the vehicles [1, 2]. In an in-vehicle surveillance system, various sound events must be detected to understand what is occurring inside the vehicle. Therefore, this study focused on human voice SED for an in-vehicle surveillance system based on human voice signals.

Some DCASE competitions [3, 4] have previously dealt with the sound of human speech or crying babies, where target sounds are often overlapped, whereas real-world data often suffer from extreme imbalances between classes as well as overlapping sound events. For example, as shown in Fig. 1, in the strongly labeled AudioSet [5] dataset annotated using real-world data, despite focusing on the top seven classes with the highest number of event frames among the 14 "human voice" classes, the number of frames among the classes is imbalanced, with a ratio of approximately 100 to 1 between the most common and the rarest.

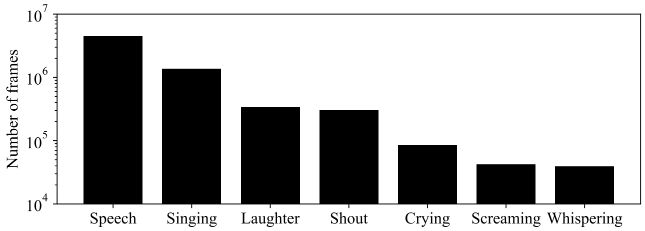


Figure 1: Number of frames in training data.

Many studies have treated SED as a multi-label classification problem for handling overlapping events that often cause detection errors [6, 7, 8]. [9] represented event overlaps by linking a bivariate probability distribution based on time and frequency with class-wise hidden Markov models. In [10], a non-negative matrix decomposition-based method that jointly trained a dictionary and a multinomial logistic regression classifier was used to manage the overlap of sounds. In [11], an event independent network for SED and localization was developed with a track-wise output. In polyphonic sound event detection (polyphonic SED) with deep learning, [12] performed multi-class classification by considering all possible overlapping event combinations as classes. However, the model architecture requires significant modification to manage a multi-class multi-tasking problem. Therefore, the conventional method [12] cannot be applied to the current polyphonic SED system without any modification of the network architecture.

Binary cross-entropy (BCE) loss is often employed as the loss function of multi-label classification in polyphonic SED. However, SED using BCE often falls into imbalance between sound event classes when training an SED model. Therefore, when applied to real-world data, the accuracy of rare class event detection decreases. Specifically, accurately detecting anomalous or rare sounds such as "Screaming" is more important than detecting common sounds such as "Speech," as shown in Fig. 1. Several loss functions that are effective for imbalanced data have been proposed in polyphonic SED. [13] proposed asymmetric focal loss and focal batch Tversky loss; however, these mainly address the imbalance problem between negative and positive samples. [14] proposed time-balanced focal loss, which is highly dependent on the dataset because the class weights used in the loss function are adjusted as hyperparameters.

Therefore, without modifying the original model architecture or preparing the class weights, we propose a method that periodically uses multi-label classification based on BCE and multi-class classification to prioritize rare classes as target labels when sound events overlap.

The contributions of this study can be summarized as follows:

- We propose a new model training method for detecting overlapped and rare sound events. The proposed method combines multi-class classification, in which rare classes are preferentially learned as target labels, along with multi-label classification. We then confirm the efficacy of this method.
- We reconstructed a strongly labeled AudioSet using seven sound event classes with "Human voice" at the upper level. We conducted a baseline evaluation for an SED task covering multiple types of human vocalization with these classes.

2. DATASET

Based on AudioSet's strong labels [5], we created a new dataset comprising 10 seconds of audio taken from the soundtrack of a YouTube video, with approximately 67,000 clips for training and 18,000 clips for evaluation. The strongly labeled AudioSet ontology is a hierarchy of 356 sound event classes. The sound classes selected for this study were the following seven event classes within the "Human voice" class: "Speech," "Singing," "Laughter," "Shout," "Crying, sobbing (Crying)," "Screaming," and "Whispering." In cases where the selected classes have subclasses, the subclasses are merged into the superclass. For example, subclasses "Baby cry, infant cry" and "Whimper" are merged into a superclass "Crying." Sound clips with other sound events in "Human voice," such as "Humming" or "Yawn," were not used in the dataset because there were few events in each class.

When sound clips contain other sound events from the category non-"Human voice," such as "Music" or "Hands," the clips were still used. However, these sound events were only background noise, that is, they were not used as target labels. After extracting the dataset to contain the selected sound class for each audio clip, the dataset contained 50,650 sound clips for training and 8,747 sound clips for evaluation. Note that in this study, rare labels ("Screaming" and "Whispering") were defined as appearing with approximately 1% of the frequency of the most frequent label.

3. PROPOSED METHOD

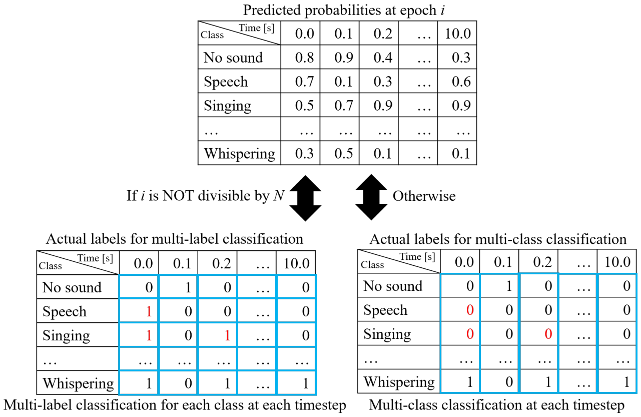
In this section, we first describe the loss functions used in this study for the multi-label and multi-class classification tasks. Next, we discuss a new model training method combining those loss functions. Finally, we describe a method of label selection for multi-class classification using polyphonic SED.

3.1. Loss function

Generally, a sigmoid activation function-based BCE is employed in training polyphonic SED models.

$$f(s)_{i,j} = \frac{1}{1+e^{-s_{i,j}}} \tag{1}$$

$$\mathcal{L}_{\mathrm{BCE}} = -\sum_{i}^{C}\sum_{j}^{T}\{y_{i,j}\log(f(s)_{i,j}) + (1-y_{i,j})\log(1-f(s)_{i,j})\}, \tag{2}$$



Predicted probabilities at epoch *i*

Time [s] Class	0.0	0.1	0.2	…	10.0
No sound	0.8	0.9	0.4	…	0.3
Speech	0.7	0.1	0.3	…	0.6
Singing	0.5	0.7	0.9	…	0.9
…	…	…	…	…	…
Whispering	0.3	0.5	0.1	…	0.1

Actual labels for multi-label classification

Time [s] Class	0.0	0.1	0.2	…	10.0
No sound	0	1	0	…	0
Speech	1	0	0	…	0
Singing	1	0	1	…	0
…	…	…	…	…	…
Whispering	1	0	1	…	1

Multi-label classification for each class at each timestep

Actual labels for multi-class classification

Time [s] Class	0.0	0.1	0.2	…	10.0
No sound	0	1	0	…	0
Speech	0	0	0	…	0
Singing	0	0	0	…	0
…	…	…	…	…	…
Whispering	1	0	1	…	1

Multi-class classification at each timestep

Figure 2: Overview of the proposed model training method.

where f is the sigmoid function, $s_{i,j}$ is the ith class's jth time frame logit, $y_{i,j}$ is the ith class's jth time frame's target label, C is the total number of classes, and T is the total number of time frames.

Conversely, a softmax activation function-based cross-entropy (Softmax-CE) loss for a multi-class classification is employed in monophonic SEDs, to choose one event from multiple sound event classes.

$$g(s)_{i,j} = \frac{e^{s_{i,j}}}{\sum_{k}^{C} e^{s_{k,j}}} \tag{3}$$

$$\mathcal{L}_{\mathrm{Softmax-CE}} = -\sum_{i}^{C}\sum_{j}^{T} y_{i,j}\log(g(s)_{i,j}), \tag{4}$$

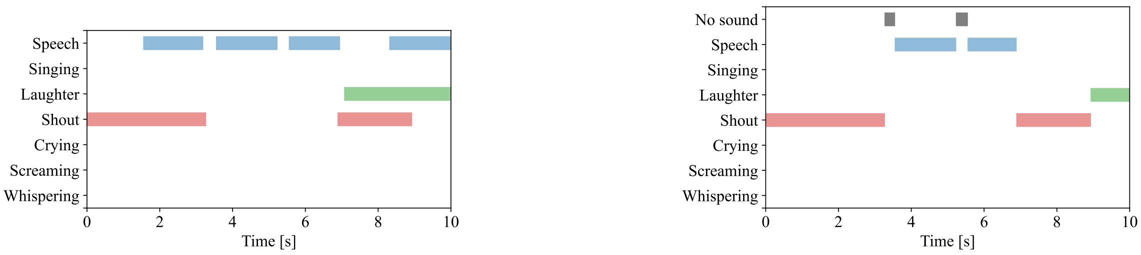
where g is the softmax function.

3.2. Training process

In this study, we applied alternately either BCE or Softmax-CE within a defined period, as shown as Fig. 2. Specifically, Softmax-CE was applied while using multi-class classification every N epoch. When the model was trained by Softmax-CE, it was trained by BCE and saved at the next epoch. Then, the model for evaluation was only used at the epoch with the minimum validation loss. For example, when multi-class classification was performed at every third epoch, i.e., at the 3, 6, 9, ..., 3*i epochs, the validation loss was monitored at the 4, 7, 10, ..., (3*i+1) epochs, with i being a positive integer.

3.3. Label selection for multi-class classification

Two problems are often encountered when performing multi-class classification for polyphonic SED because multi-class classification always requires one target label during loss computation. The first problem is determining which label to allocate when multiple sound events occur simultaneously. The second problem is determining which label to allocate when none of the target sound events occur. To solve these problems, we propose a new method of label selection using the sparsity of sound events. Specifically, we prioritize the rarest label for multiple events and define a new class label for no events. The details are discussed below.



(a) multi-label classification.

(b) multi-class classification.

Figure 3: Overview of the target labels. (a) multi-label classification and (b) multi-class classification.

Table 1: Number of frames for each sound event.

	The number of frames	
Event class	Multi-label	Multi-class
Speech	9,083,336	8,627,958
Singing	2,781,533	2,690,509
Laughter	683,036	666,186
Shout	612,058	607,208
Crying	174,120	173,731
Screaming	85,578	85,578
Whispering	80,194	80,194

Multiple events This section describes the method to allocate sound event labels when multiple overlapping sound events occur concurrently during a single clip, as shown in Fig. 3. In Fig. 3, multiple sound events overlap as follows.

- 1.6 - 3.2 seconds : Speech and **Shout**
- 7.1 - 8.3 seconds : Laughter and **Shout**
- 8.3 - 8.9 seconds : Speech, Laughter, and **Shout**
- 8.9 - 10.0 seconds : Speech and **Laughter**

We adopted the rarest label for each time frame in the clip. In the example in Fig. 3, when multi-class classification was performed, the labels in bold were used. The number of frames for each label in the training data when changing from the labels used in multi-label classification to those used in multi-class classification is shown in Table 1. Frequent labels such as "Speech" and "Singing" show a large decrease in the number of frames when compared with the number of rare labels.

No events In multi-label classification, BCE originally includes calculation of inactive frames. However, in multi-class classification, even when none of the seven target labels exist in a frame, one class must be set as the target label. Therefore, a new class "No sound" was created and allocated to time frames containing no target class. Fig. 3(b) shows the "No sound" class with thick black lines between 3.3 - 3.5 seconds and 5.2 - 5.5 seconds.

4. EXPERIMENT

4.1. Experimental setups

The AudioSet sound clips were downloaded from YouTube. These sounds were mostly monaural. The left and right sides of the stereophonic sounds were averaged to produce monaural sounds. The sounds were resampled to 44.1 kHz, as previously configured [7]. The sounds were then converted into a logmel scale of $F = 64$ filters calculated every 40 milliseconds with a hop size of 20 milliseconds.

Table 2: Model architecture. The kernel sizes of the convolutional and pooling layer are denoted as "Conv (kernel size)" and "Max Pooling (kernel size)," respectively. The number of attention heads is denoted as "Transformer Encoder (number of attention heads)."

Conv3	RB
Log-mel spectrogram 500 frames × 64 mel bins	
Conv (3 × 3) BN, ReLU, Dr Max Pooling (8 × 1)	Conv (3 × 3) BN, ReLU ResBlock
Conv (3 × 3) BN, ReLU, Dr Max Pooling (4 × 1)	ResBlock
Conv (3 × 3) BN, ReLU, Dr Max Pooling (2 × 1)	ResBlock
(Transformer Encoder (32)) × 2	
FC, Sigmoid	

Inspired by [13], a convolutional neural network (CNN)-transformer-based network was used as the model architecture. This architecture performs better than the CNN-biGRU-based network, which is widely used in SED [7, 15, 16]. The model architecture is shown in Table 2. The system has two types of CNN backbones: one with three Convolution layers (Conv3) and the other with three ResBlocks (RB). The parameters of the convolutional layers in RB are the same as those of Conv3.

The models were trained using the RAdam optimizer [17] with a learning rate of 0.001. Early stopping was implemented after 50 epochs if no improvement on validation loss was noted.

As evaluation metrics, we used the mean average precision (mAP), the micro average precision (micro-AP), and the frame-based macro- and micro-Fscores with a threshold for prediction of 0.5. Even when the proposed method was deployed with eight event classes including "No sound," we evaluated them using only the original seven classes. In this study, we used eight classes when performing both the multi-label and multi-class classification using the proposed method.

Table 3: Average SED performance for two backbones.

	Conv3				RB			
Method	mAP	macro-Fscore	micro-AP	micro-Fscore	mAP	macro-Fscore	micro-AP	micro-Fscore
baseline	51.96%	38.16%	83.39%	73.83%	58.04%	50.06%	85.48%	77.11%
AFL	50.92%	36.92%	82.90%	73.19%	59.05%	52.15%	85.66%	77.56%
e0	51.60%	39.15%	83.25%	72.72%	58.22%	49.71%	85.35%	76.82%
e1	51.51%	37.03%	79.59%	50.88%	56.28%	32.97%	80.39%	47.36%
e2 (proposed)	**52.19%**	**40.41%**	**83.55%**	**74.65%**	**60.32%**	**53.40%**	**86.65%**	78.52%
e3 (proposed)	51.63%	38.20%	83.11%	73.49%	59.51%	50.96%	86.47%	78.38%
e4 (proposed)	51.54%	38.59%	83.27%	73.48%	59.89%	52.70%	86.59%	**78.58%**
e5 (proposed)	51.28%	38.25%	83.14%	73.49%	59.66%	51.35%	86.36%	78.28%

Table 4: Average SED performance of the rare-event labels for the two backbones.

		Conv3		RB	
	Method	AP	Fscore	AP	Fscore
Screaming	baseline	17.29%	3.57%	27.04%	**17.03%**
	AFL	16.46%	2.92%	28.08%	15.14%
	e2 (proposed)	**18.73%**	**4.01%**	**28.37%**	13.43%
Whispering	baseline	55.86%	42.91%	64.68%	61.64%
	AFL	55.03%	40.55%	65.67%	60.81%
	e2 (proposed)	**56.79%**	**45.19%**	**67.34%**	**65.64%**
Avg.	baseline	36.58%	23.24%	45.86%	39.34%
	AFL	35.74%	21.73%	46.87%	37.98%
	e2 (proposed)	**37.76%**	**24.60%**	**47.85%**	**39.54%**

4.2. Experimental results

Table 3 shows the results for the baseline using only multi-label classification and the proposed method that performed multi-label classification while using multi-class classification every N epochs. Each result is the average of five iterations. The value of N in eN represents frequency of switching to Softmax-CE. Here, e0 was trained using multi-task learning, where multi-label classification and multi-class classification were performed simultaneously every epoch. Conversely, e1 was trained using Softmax-CE every epoch and evaluated as a multi-label classification. The baseline method was performed on the original seven-class multi-label classification without the "No sound" class, and the proposed method was performed on the eight-class multi-label classification and multi-class classification including the "No sound" class. When e2, i.e., multi-label classification and multi-class classification, was used independently of the backbone for every other epoch, it demonstrated the best performance on several metrics. The proposed method improved mAP by 0.23 percentage points and 2.28 percentage points for Conv3 and RB, respectively. This result demonstrates that the proposed method improves the performance of rare events by using Softmax-CE and retains the performance of frequent events by using BCE. Meanwhile, the simultaneous use of Softmax-CE and BCE in e0 could have prevented the influence of Softmax-CE. With an increase in N, the performance of several metrics gradually decreases, and at e5, the performance is comparable to the baseline values. When Softmax-CE is used less frequently, it becomes less effective.

Table 4 shows the results when focusing on rare labels ("Screaming" and "Whispering"). Comparisons were made with asymmetric focal loss (AFL), which has an effect on the imbalanced data [13]. There is a significant difference between "Screaming" and "Whispering." "Screaming," which is similar to "Shout" and "Singing," is more likely to occur in noisy environments, whereas "Whispering" is a special type of sound event where other sound events are unlikely to occur concurrently. As with the overall performance, e2 performed highest on many measures, but the Fscore for "Screaming" dropped significantly. "Screaming" had a lower Fscore, which was based on the threshold, because underfitting caused by a rare label reduced the predicted probability of "Screaming." However, because e2 performed the highest, when comparing under the same conditions, the intrinsic prediction performance of e2 is superior and is effective for rare labels. Unlike the original experimental dataset used for evaluating AFL, the number of data classes (seven) was limited due to an imbalance of approximately 1 to 100 in a class. This may have contributed to the e2 performance being superior to that of AFL.

5. CONCLUSION

We proposed a method of polyphonic SED by periodically using either multi-label or multi-class classification. Based on the sparsity of sound events, multi-class classification was used to strongly train rare sound event labels, in which the rarest sound event was selected as the label representing each frame. The proposed method was evaluated on a human voice dataset extracted from the strongly labeled AudioSet data. Our approach was found to be most effective when the two loss functions were alternately applied. For the imbalanced data, regarding both the overall metrics and for rare labels, this method significantly outperformed the conventional methods.

6. REFERENCES

[1] G. McKenzie, "Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services," *Computers, Environment and Urban Systems*, vol. 79, p. 101418, 2020.

[2] A. Mishra, S. Lee, D. Kim, and S. Kim, "In-cabin monitoring system for autonomous vehicles," *Sensors*, vol. 22, no. 12, 2022.

[3] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[4] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 165–169.

[5] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. Channing Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 366–370.

[6] T. N. T. Nguyen, K. N. Watcharasupat, Z. J. Lee, N. K. Nguyen, D. L. Jones, and W. S. Gan, "What makes sound event localization and detection difficult? insights from error analysis," in *Proc. 6th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2021, pp. 120–124.

[7] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[8] H. Phan, O. Y. Chén, P. Koch, L. Pham, I. McLoughlin, A. Mertins, and M. D. Vos, "Unifying isolated and overlapping audio event detection with multi-label multi-task convolutional recurrent neural networks," in *Proc. 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 51–55.

[9] E. Benetos, G. Lafay, M. Lagrange, and M. D. Plumbley, "Detection of overlapping acoustic events using a temporally-constrained probabilistic model," in *Proc. 41th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6450–6454.

[10] V. Bisot, S. Essid, and G. Richard, "Overlapping sound event detection with supervised nonnegative matrix factorization," in *Proc. 42th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 31–35.

[11] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 885–889.

[12] H. Phan, T. N. T. Nguyen, P. Koch, and A. Mertins, "Polyphonic audio event detection: Multi-label or multi-class multi-task classification problem?" in *Proc. 47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8877–8881.

[13] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of sound duration and inactive frames on sound event detection performance," in *Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 860–864.

[14] S. Park and M. Elhilali, "Time-balanced focal loss for audio event detection," in *Proc. 47th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 311–315.

[15] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proc. 3rd Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2018, pp. 19–23.

[16] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano, "An analysis of sound event detection under acoustic degradation using multi-resolution systems," *Applied Sciences*, vol. 11, no. 23, 2021.

[17] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. 8th International Conference for Learning Representations (ICLR)*, 2020.