EVALUATION METRIC OF SOUND EVENT DETECTION CONSIDERING SEVERE MISDETECTIONS BY SCENES

Noriyuki Tonami[1], *Keisuke Imoto*[2], *Takahiro Fukumori*[1], *Yoichi Yamashita*[1],

[1] Ritsumeikan University, Japan, [2] Doshisha University, Japan.

ABSTRACT

In this paper, we propose a new evaluation metric for sound event detection (SED) and discuss a problem frequently encountered in conventional metrics. In conventional evaluation metrics, misdetected sound events are treated equally, e.g., the misdetected sound event "birdsong" in the acoustic scenes "airplane" and "park" are treated as the same type of misdetection. However, the misdetected event "birdsong" in "airplane" is a severe mistake compared with its misdetection in "park." The event "birdsong" rarely occurs in the "airplane." SED systems that are evaluated using conventional metrics may cause severe/catastrophic problems and lead to confusion in practice owing to lack of consideration of the relationship between sound events and scenes. Our evaluation metric for SED considers severe misdetections on the basis of the relationship between sound events and scenes. We demonstrate the utility of our proposed method by comparing it with the conventional evaluation metrics on two datasets with events and scenes. Experimental results show that the proposed metric can accurately evaluate whether SED systems appropriately consider the relationship between sound events and scenes.

Index Terms— sound event detection, evaluation metrics, acoustic scene

1. INTRODUCTION

There has been increased interest in the automatic analysis of various environmental sounds within human everyday life [1]. The automatic analysis of environmental sounds will lead to many applications, such as anomalous sound detection systems [2], automatic life-logging systems [3], monitoring systems [4], bird-call identification [5], and hearing-impaired support systems [6].

Sound event detection (SED) is the task of identifying sound event labels and their boundaries from a recording. SED has been studied by machine learning methods [7–10] and several types of evaluation framework have been proposed [11–14]. Bilen *et al.* [13] have proposed a new event-based evaluation metric, which is more robust to the subjectivity of annotation. Baumann *et al.* [14] have proposed a new evaluation framework for rare sound event detection, in which a more realistic construction of data is considered.

In our everyday life, sound events may occur in various soundscapes, so-called "acoustic scenes." For example, the sound event "people walking" can occur in many scenes such as "office," "home," and "supermarket". On the other hand, the sound event "car" tends to occur only outdoors such as in the scene "street". The sound events and scenes are thus strongly related to each other. On the basis of this concept, SED utilizing the results of acoustic scene classification (ASC) [15–17], ASC using information on sound events [18, 19], joint models of SED and ASC [20–23], and a public dataset [20] have been proposed. In those works, conventional metrics were used, e.g. the F-score, which disregard the relationship between sound events and scenes. The misdetected event "birdsong" in the scenes "park" and that in "airplane" are treated as the same type of misdetection (e.g., false positive) when using conventional metrics such as the F-score. However, in our everyday life, the meaning of the sound event depends on the acoustic scene. For example, the sound event "birdsong" in the acoustic scene "park" is a normal sound because "birdsong" often occurs in "park". In contrast, the event "birdsong" rarely occurs in the scene "airplane." Assuming that a SED system detects "birdsong" in the "airplane," this can indicate a severe error compared with its detection in the scene "park." Such a system may encounter a catastrophic problem and lead to confusion people in practice. To avoid this problem, we must accurately evaluate the performance of SED systems.

In this paper, we propose a new evaluation metric for SED utilizing the relationship between sound events and scenes. Moreover, we discuss the problem plaguing the conventional metrics (F-score), in which the relationship between sound events and scenes is not considered. The new evaluation metric employs the severity of misdetection using the incidence relationship between sound events and scenes (which event occurs in which scene). The new evaluation metric is expected to enable precise evaluation of the performance of SED systems and avoid severe or catastrophic problems in practice. In addition, the proposed metric is helpful when choosing/building more realistic systems that do not cause severe problems or confusion.

2. CONVENTIONAL EVALUATION METRIC FOR SOUND EVENT DETECTION

2.1. Segment-based F-score for SED

In SED tasks, segment and event-based measurements [12] have been used to evaluate the performance of SED systems. In this paper, we focus only on the segment-based F-score although the concept of our proposed method can be utilized in both segment- and event-based measurements. To calculate the segment-based F-score, the output of the network is first binarized, such as by the following thresholding:

$$\hat{y}_{m,t} = \begin{cases} 1 & y_{m,t} > \tau \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where $y_{m,t}$ and τ denote the output of the networks for event m in the time frame t and a threshold value, respectively.

We then calculate true positive (TP), false positive (FP), false negative (FN), and true negative (TN) between $\hat{y}_{m,t}$ and the ground truth. To calculate the segment-based F-score, the TPs, FPs, and FNs are aggregated as follows:

$$\mathbf{Fscore} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} . \tag{2}$$

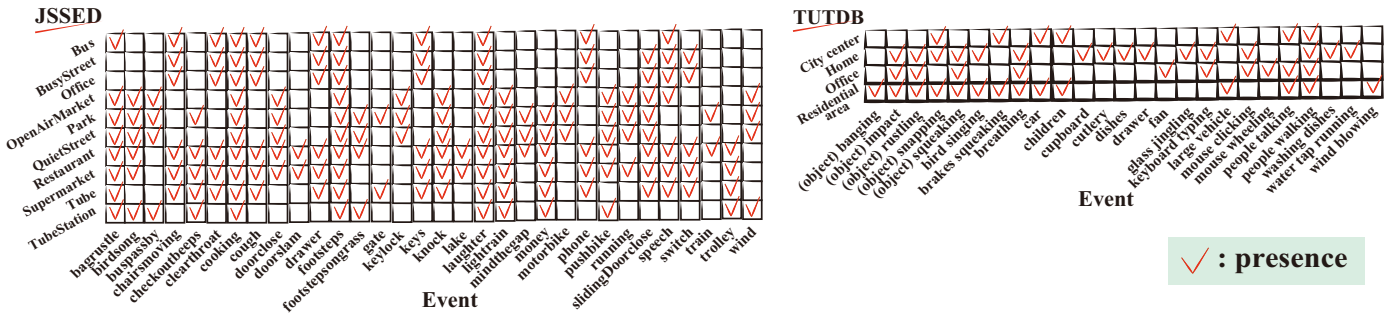

Figure 1: Incidence relationships between sound events and scenes (which event occurs in which scene)

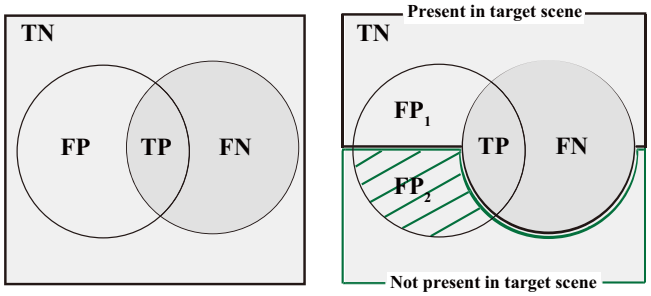

Figure 2: Venn diagram representations: conventional (left) and proposed (right) metrics

Table 1: SED results of two models for event "washing dishes" using toy dataset

Metric	*Model 1*		*Model 2*	
	bank	home	bank	home
# TPs	0	200	0	200
# FPs	**200**	**0**	**0**	**200**
# FNs	0	50	0	50
# TNs	250	250	250	250
F-score	61.54%		61.54%	
IoU	44.44%		44.44%	
HM	**47.06%**		61.54%	

2.2. Problem of evaluation metric (F-score) for SED

None of the conventional evaluation metrics, such as the F-score, for SED take account of the relationship between sound events and scenes. However, the sound events and acoustic scenes are closely related to each other. For example, in the acoustic scene "home," the sound event "washing dihses" often occurs. On the other hand, in the scene "bank," the event "washing dishes" rarely occurs. If the event "washing dishes" occurs in the scene "bank," this leads to confusion.

Table 1 shows the SED results of the event "washing dishes" in two acoustic scenes, "bank" and "home," using the toy dataset. The results are for two models. The intersection over union (IoU) and hybrid metric (HM) are described later (in sec. 3). The difference between the two models is seen only in the number of the FPs in each scene. *Model 1* misdetects the sound event "washing dishes" as FP only in the scene "bank." In contrast, the *model 2* misdetects "washing dishses" as FP only in "home." The performance of *model 1* is equal to that of *model 2* in terms of the F-score. However, the misdetection of "washing dishes" in "bank" as FP is a more severe mistake than the misdetection in "home." For this reason, *model 2* may give rise to serious trouble or confusion in practice. This problem is caused by disregarding the relationship between sound events and scenes.

3. EVALUATION METRIC CONSIDERING RELATIONSHIPS BETWEEN SOUND EVENTS AND SCENES

When using the conventional metric, the F-score, the meanings of the misdetected sound event "birdsong" in the scenes "airplane" and "park" are the same (e.g., FP). However the sound event "birdsong" rarely occurs in the scene "airplane." If the sound event "birdsong" occurs in the scene "airplane," people will be confused. The model in which the relationship between sound events and scenes is ignored may give rise to severe or catastrophic problems in practice. To tackle this problem, we propose a new evaluation metric that considers the relationship between sound events and scenes.

As the relationship, we use an incidence relationship between sound events and scenes (which event occurs in which scene). In particular, we incorporate the following two points into the proposed metric:

- misdetected sound events that are present in target scenes are evaluated as mild mistakes.
- misdetected sound events that are **not** present in target scenes are evaluated as severe mistakes.

Here we review the conventional evaluation metrics. The F-score is used for detection tasks such as SED and information retrieval. In other detection tasks such as object detection, IoU is utilized to evaluate the performance of systems. IoU is calculated as

$$\mathbf{IoU} = \frac{2\mathrm{TP}}{2\mathrm{TP} + 2\mathrm{FP} + 2\mathrm{FN}}, \tag{3}$$

where the difference between the F-score (Eq. 2) and IoU (Eq. 3) is only in the weights of mistakes, FP and FN. For IoU, the weights of mistakes are larger than those of the F-score. This indicates that IoU treats the mistakes severely compared with the F-score.

To apply the relationship between sound events and scenes to the proposed metric, we take advantage of the difference between the F-score and IoU with respect to the mistake. Our proposed evaluation metric is calculated as

$$\mathbf{HM} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP}_1 + 2\mathrm{FP}_2 + \mathrm{FN}}, \tag{4}$$

where FP_1 and FP_2 denote the numbers of segments of the FPs that are present in target scenes and those that are **not** present in target scenes, respectively. In this way, the proposed evaluation metric utilizes the misdetection weight of the F-score for a mild mistake and that of IoU for a severe mistake. By switching two misdetection weights between two types of the FP, we incorporate the relationship between the sound events and scenes into the proposed method.

Table 2: Experimental conditions

JSSED	
Development: 1,200 min	Evaluation: 300 min
TUTDB	
Development: 192 min	Evaluation: 74 min
Shared layers	
Network	3CNN
# channels	128, 128, 128
Filter size	3×3
Pooling size	8×1, 2×1, 2×1 (max pooling)
Scene layers	
Network	2CNN & 1FC
# channels	256, 256
Filter size	3×3
Pooling size	1×25, 1×20 (max pooling)
# units in FC layer	32
Event layers	
Network	1BiGRU & 1FC
# units in GRU layer	32
# units in FC layer	32

Table 3: Overall performance of sound event detection

Metric	micro		macro	
	CRNN	MTL	CRNN	MTL
JSSED				
F-score	27.88%	**28.39%**	17.25%	**17.42%**
IoU	16.20%	**16.55%**	7.73%	**7.83%**
HM	27.83%	**28.36%**	13.47%	**13.60%**
TUTDB				
F-score	45.47%	**46.00%**	**15.28%**	14.40%
IoU	29.42%	**29.87%**	**10.10%**	9.48%
HM	42.04%	**42.95%**	**14.35%**	13.60%

Moreover, the magnitude relation among the F-score, IoU, and the proposed method HM for the same datasets and trained models is

$$\mathbf{IoU} \leq \mathbf{HM} \leq \mathbf{Fscore}\,. \tag{5}$$

Table 1 shows the SED results in terms of IoU and HM in addition to the previously mentioned results in terms of F-score (in sec. 2.2). The performance of *model 1* is equal to that of *model 2* in terms of the F-score and IoU. The ranking of the two models changes when using the proposed metric HM. This is because the misdetected event "washing dishes" in the scene "bank" is treated as a severe mistake. Using the proposed metric is helpful for choosing/building systems that do not confuse people or cause severe problems in practice. In this work, the incidence relationship between sound events and scenes (which event occurs in which scene) on the development set, as shown in Fig. 1, is employed. Note that if an event occurs at least once in a scene, it is regarded as "presence." Fig. 2 shows the conventional (left) and proposed (right) Venn diagram representations. In each venn diagram, the outer frame indicates the set of all the detection results. The left and right inner circles represent the sets of the predicted and actual classes being positive, respectively. In the proposed metric, FP is divided into two types, FP_1 and FP_2. Note that the FNs occur only in the target scenes, as can be seen in Fig. 2.

4. EXPERIMENTS

4.1. Experimental conditions

To evaluate our proposed metric, we used two datasets with sound event and scene labels. As the first dataset, we used the joint sound scene and event dataset (JSSED) [20] consisting of synthesized audio clips with the 32 events and 10 scenes listed in Fig. 1. As the second dataset, we aggregated TUT Sound Events 2016 [24] and 2017 [25] and TUT Acoustic Scenes 2016 [24], and 2017 [25]. Hereafter, those datasets [24, 25] are referred to as TUTDB. We selected audio clips within acoustic scenes "city center," "home," "office," and "residential area," in which the 25 events listed in Fig. 1 are included. Furthermore, we manually annotated the sound event labels within the scene "office" in accordance with the procedure described in [24, 25]. The sound event labels annotated in this work are available in [26].

As the input of the networks, we used the 64-dimensional log mel-band energies, which have a 40 ms window with a 20 ms hop size. In this work, FP_1 and FP_2 are defined in Fig. 1 as the incidence relationship between sound events and scenes (which event occurs in which scene) on the development sets.

For comparison, we used two methods, CNN-BiGRU (CRNN) [9] and the multitask-learning-based model of SED and ASC (MTL) [21], which aims to train the sound events and scenes effectively. The joint model has a shared part of networks of SED and ASC, "shared layers," as shown in Table 2. The joint method has a hyperparameter, β which is a weight of loss function for ASC. In this experiment, we used $\beta = 10^{-6}$ (JSSED) and 10^{-4} (TUTDB) tuned using only the development set. Other experimental conditions are shown in Table 2

4.2. Experimental results

As previously described, for the toy dataset, we verified the utility of the proposed metric by comparing it with conventional metrics. In this section, we demonstrate that the utility of the proposed metric in providing more realistic evaluation of datasets with sound events and scenes. More specifically, we demonstrate how severe misdetections affect the performances of SED systems. The proposed metric is expected to provide a more realistic evaluation of events in scenes e.g., the event "birdsong" in the scene "airplane" is unrealistic. If the HM of a SED system is improved compared with other systems, this indicates that the SED system yield accurate detections considering the relationship between sound events and scenes. Since direct comparison of metrics is difficult, we observed the amount of changes between the two systems in terms of the F-score, IoU, and HM.

Table 3 shows the overall performances of SED in terms of the F-score, IoU, and HM. Note again that the MTL aims to train the relationship between sound events and scenes. If the HM of the MTL is better than the CRNN, the HM is useful in evaluating whether SED systems consider the relationship between sound events and scenes. The results indicate that the MTL achieves better detection performances than the CRNN in terms of the F-score and IoU. Moreover, since the HM of the MTL was improved compared with that of the CRNN, the proposed metric is useful in evaluating whether SED systems consider the relationship between sound events and scenes.

Table 4 shows the experimental results of SED for each event. Only events for which the performance of SED changed significantly are listed. The numbers to the right of arrows indicate the amount of change between the CRNN and MTL in terms of each metric. The results show that the MTL achieves a reasonable per-

Table 4: Performance of SED for each event. Numbers to the right of arrows indicate amount of change.

Event		JSSED				TUTDB			
		buspassby		phone		car		fan	
F-score	CRNN	30.57%	↓0.51	35.14%	↑1.35	51.44%	↓0.48	71.29%	↑0.96
	MTL	30.08%		36.49%		50.96%		72.25%	
IoU	CRNN	18.04%	↓0.34	21.32%	↑0.99	34.63%	↓0.43	55.38%	↑1.28
	MTL	17.70%		22.31%		34.20%		56.56%	
HM	CRNN	30.53%	**↓0.50**	35.00%	**↑1.36**	50.86%	**↓0.73**	56.96%	**↑3.24**
	MTL	30.03%		36.36%		50.13%		60.20%	

Table 5: SED results of "fan" in terms of TP, FP, and FN for each scene. Green and gray cells represent FP_1 and FP_2, respectively.

Scene		city center	home	office	residential area
CRNN	# TPs	0	0	354481	0
	# FPs	1880	220022	419	28257
	# FNs	0	0	34979	0
MTL	# TPs	0	0	321127	0
	# FPs	3188	**149271**	318	**25556**
	# FNs	0	0	68333	0

Table 6: SED results of "car" in terms of TP, FP, and FN for each scene. Green and gray cells represent FP_1 and FP_2, respectively.

Scene		city center	home	office	residential area
CRNN	# TPs	214816	0	0	97417
	# FPs	449257	13942	2	99671
	# FNs	11094	0	0	15483
MTL	# TPs	204515	0	0	92348
	# FPs	420216	**19460**	0	89623
	# FNs	21395	0	0	20552

formance compared with the CRNN. The SED performance of the listed events showed larger changes between the CRNN and MTL in terms of the HM than the F-score, even though $\mathbf{HM} \leq \mathbf{F}$ (Eq. 5). This is because the misdetected events absent from target scenes are penalized severely. This indicates that the HM is useful for considering the relationship between sound events and scenes. In particular, the HM showed a significant change for the event "fan" compared with the other metrics. This indicates that the event "fan" was detected more accurately when using the MTL than when using the CRNN. In other words, the event "fan" was trained effectively considering the relationship between the events and scenes. This is because "fan" is closely related to "office," that is, it accounts for over 20% of the total number of active time-frames, as shown in Fig. 3. We could say that the detection of "fan" was more realistic when using the MTL than when using the CRNN.

To investigate the utility of the proposed metric in detail, we listed the results of two events, "fan" and "car," for each scene, as shown in Tables 5 and 6, respectively. In the two tables, green and ash gray cell colors represent the FP events that are absent from target scenes, FP_2, and present in target scenes, FP_1, respectively. The result in Table 5 shows that the number of severely misdetected events was smaller when using the MTL than when using the CRNN. Tables 4 and 5 show that the HM enables accurate evaluation with the MTL owing to consideration of the relationship between sound events and scenes. On the other hand, the results in Table 6 show that the number of severely misdetected events was significantly larger in the scene "home" when using the MTL than when using the CRNN. In a previous work [21], it was reported that the relationship between sound events and scenes can be trained more effectively by using the MTL than by using the CRNN. However, the results show that some of the events can be misdetected severely using the MTL. Moreover, the results indicate that the severely misdetected events, which correspond to large numbers of time frames (e.g., "car" accounts for roughly 20% of the total number of active time-frames, as can be seen in Fig. 3), affect the overall macro-F-score, IoU, and HM.

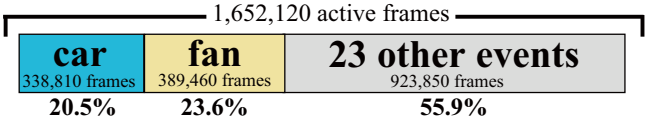

Figure 3: Numbers of active time-frames of events "car" and "fan."

The results of all experiments show that the proposed metric can accurately evaluate whether SED systems consider the relationship between sound events and scenes. Severely misdetected events can cause the severe problems in practice. In our proposed metric, severe misdetections were penalized severely to avoid confusing people or causing severe problems in practice. In these experiments, to define severe misdetections, the incidence relationship between sound events and scenes on the development set is used. When there is a large discrepancy in the incidence relationship between the development and evaluation sets, the HM may not be able to assess the performance of SED correctly in these experiments. For example, on the development set, "children" does not occur in "home," as shown in Fig. 1. In these experiments, "children" in "home" is treated as a severe misdetection. This is not always true. To avoid such results, dataset-independent relationship between sound events and scenes must be considered.

5. CONCLUSION

In this paper, we proposed a new evaluation metric for SED. In the proposed metric, we used two types of misdetection (mild and severe misdetections) on the basis of incidence relationship between sound events and scenes (which event occurs in which scene). The results revealed that the proposed metric can accurately evaluate whether SED systems consider the relationship between sound events and scenes. In future work, to define severe misdetections, prior knowledge of people should be employed as dataset-independent relationships between the sound events and scenes.

6. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number JP19K20304 and NVIDIA GPU Grant Program.

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. Yamanishi, 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] Y. T. Peng, C. Y. Lin, M. T. Sun, and K. C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," *Proc. IEEE International Conference on Multimedia and Expo* (*ICME*), pp. 1218–1221, 2009.

[7] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach to audio event detection," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 1–4, 2013.

[8] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 45–49, 2016.

[9] 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.

[10] 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 Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 992–1006, 2019.

[11] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," *Proc. European Signal Processing Conference* (*EUSIPCO*), pp. 1–5, 2013.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, pp. 1–17, 2016.

[13] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 61–65, 2020.

[14] J. Baumann, T. Lohrenz, A. Roy, and T. Fingscheidt, "Beyond the DCASE 2017 challenge on rare sound event detection: A proposal for a more realistic training and test framework," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 611–615, 2020.

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

[16] 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.

[17] 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.

[18] K. Imoto and N. Ono, "Acoustic topic model for scene analysis with intermittently missing observations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 2, pp. 367–382, 2019.

[19] J. Jung, H. Shin, J. Kim, S. Kim, and H. Yu, "Acoustic scene classification using audio tagging," *arXiv*, *arXiv:2003.09164,* 2020.

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

[21] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, 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.

[22] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, 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.

[23] T. Komatsu, K. Imoto, and M. Togami, "Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 646–650, 2020.

[24] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," *Proc. European Signal Processing Conference* (*EUSIPCO*), pp. 1128–1132, 2016.

[25] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 Challenge setup: Tasks, datasets and baseline system," *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), pp. 85–92, 2017.

[26] https://www.ksuke.net/dataset.