SOUND EVENT DETECTION GUIDED BY SEMANTIC CONTEXTS OF SCENES

Noriyuki Tonami[1], *Keisuke Imoto*[2], *Ryotaro Nagase*[1],
Yuki Okamoto[1], *Takahiro Fukumori*[1], *Yoichi Yamashita*[1]

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

ABSTRACT

Some studies have revealed that contexts of scenes (e.g., "home," "office," and "cooking") are advantageous for sound event detection (SED). Mobile devices and sensing technologies give useful information on scenes for SED without the use of acoustic signals. However, conventional methods can employ pre-defined contexts in inference stages but not undefined contexts. This is because one-hot representations of pre-defined scenes are exploited as prior contexts for such conventional methods. To alleviate this problem, we propose scene-informed SED where pre-defined scene-agnostic contexts are available for more accurate SED. In the proposed method, pre-trained large-scale language models are utilized, which enables SED models to employ unseen semantic contexts of scenes in inference stages. Moreover, we investigated the extent to which the semantic representation of scene contexts is useful for SED. Experimental results performed with TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016/2017 datasets show that the proposed method improves micro and macro F-scores by 4.34 and 3.13 percentage points compared with conventional Conformer- and CNN–BiGRU-based SED, respectively.

Index Terms— Sound event detection, acoustic scene, semantic embedding

1. INTRODUCTION

The analysis of various environmental sounds in real life has been attracting significant attention [1]. The automatic analysis of various sounds enables many applications, such as anomaly detection systems [2], life-logging systems [3], and monitoring systems [4].

Sound event detection (SED) is the task of estimating sound events (e.g., bird singing, footsteps, and wind blowing) and their time boundaries from acoustic signals. In SED, a number of neural-network-based approaches have been proposed, including the convolutional neural network (CNN) [5], recurrent neural network (RNN) [6], and convolutional recurrent neural network (CRNN) [7]. CNN automatically extracts features and RNN models temporal structures. CRNN is a hybrid of CNN and RNN, which is widely used as a baseline system of SED. More recently, to handle longer sequences, non-autoregressive models, such as Transformer [8, 9] and Conformer [10], have been studied.

Furthermore, some studies have revealed that the contexts of scenes (e.g., "home," "office," and "cooking"), which are defined by locations, activities, and time, help increase the accuracy of SED [11–18]. For example, Heittola *et al*. [12] have proposed a cascade method for SED using results of acoustic scene classification (ASC). Bear *et al*. [13], Tonami *et al*. [16], and Komatsu *et al*. [15] have proposed joint models of SED and ASC to take advantage of the relationships between sound events and scenes; e.g., the sound event "mouse wheeling" tends to occur in the scene "office," whereas the event "car" is likely to occur in the scene "city center." Cartwright *et al*. [19] have proposed audio tagging considering spatiotemporal contexts such as a city block, an hour, and a day.

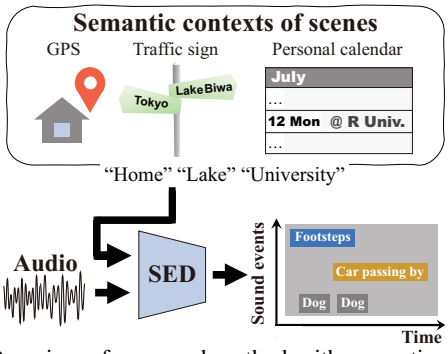

Fig. 1. Overview of proposed method with semantic contexts of scenes

Mobile devices (e.g., smartphones) and sensing technologies (e.g., global positioning system (GPS)) provide useful information on scenes for SED without the use of acoustic signals, as shown in Fig. 1. However, conventional SED methods [12, 15, 19] can utilize pre-defined contexts in inference stages but not undefined contexts. This is because one-hot representations or identifiers of pre-defined scenes are used as prior contexts for such conventional methods.

To mitigate this problem, we propose scene-informed SED where pre-defined scene-agnostic contexts are available for precisely detecting sound events. To this end, pre-trained large-scale language models are utilized, which enables SED models to use unseen semantic contexts of scenes in inference stages. To extract better scene contexts for SED, we further verify the effectiveness of the semantic representation of scene contexts, i.e., representation learning.

2. FRAMEWORK OF SED

In SED, the goal is to predict active (1) or inactive (0) sound events; $\hat{z}_{n,t} \in \{0,1\}^{N\times T}$ for each sound event n and time frame t from given acoustic features $x_{t,f}$. Here, N and T are the total numbers of sound event classes and time frames, respectively. In general, $x_{t,f}$ is an element of acoustic features represented by the time-frequency domain such as log-mel spectrograms, where f is the index of a dimension of the acoustic features. In recent SED, neural-network-based methods have been dominant. In this work, we focus on strongly supervised SED, where ground truths of time stamps are available for the training. The binary cross-entropy is used to optimize neural-network-based SED models as follows:

$$\mathcal{L}_{\mathrm{SED}} = -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\Big\{z_{n,t}\log\sigma(y_{n,t}) + (1-z_{n,t})\log\big(1-\sigma(y_{n,t})\big)\Big\}, \quad (1)$$

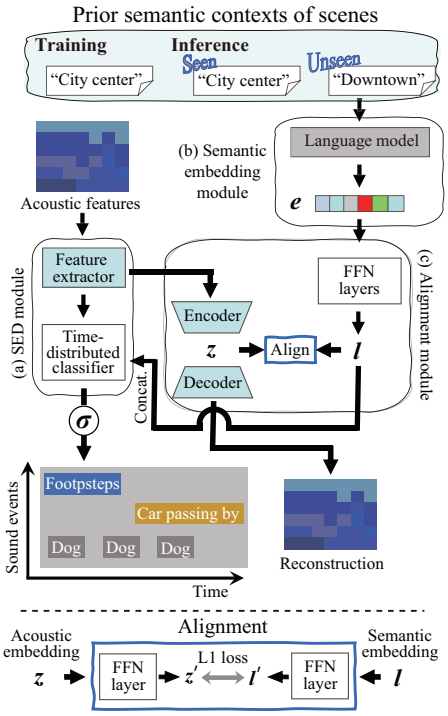

Fig. 2. Network architecture of proposed method

where $z_{n,t} \in \{0, 1\}$ is a target label of a sound event, $y_{n,t} \in [0, 1]$ indicates the output of the SED network, and $\sigma(\cdot)$ denotes the sigmoid function. In inference stages, $y_{n,t}$ is binarized using a pre-defined threshold value to obtain $\hat{z}_{n,t}$.

3. PROPOSED METHOD

The proposed method exploits the semantic contexts of scenes (e.g., "home," "office," and "cooking") to handle the pre-defined scene-agnostic contexts for SED. As previously mentioned, the conventional SED methods cannot employ undefined (unseen) contexts of scenes because one-hot representations or identifiers of pre-defined scenes are utilized as the prior contexts for effective SED. To address this problem, the proposed SED method utilizes the semantic embedding of scenes instead of the one-hot representations or their identifiers. In Fig. 2, an example of the proposed network is depicted. The proposed network is divided into three modules: (a) SED module, (b) semantic embedding module, and (c) alignment module. Module (a) works as a sound event detector, module (b) transforms contexts of scenes into semantic embedding, and module (c) aligns the semantic embedding with an acoustic embedding (representation learning).

The SED module (a) is a general architecture used in conventional SED. In Fig. 2, the acoustic features in the time-frequency domain are input to a feature extractor such as CNN or RNN layers. The output of the feature extractor is input to a time-distributed classifier such as feedforward neural network (FFN) layers.

The semantic embedding module (b) handles the contexts of not only pre-defined (seen) but also unseen scenes. In this work, to extract the semantic contexts of scenes, we utilize label texts of acoustic scenes, e.g., "city center" or "home," using a procedure similar to that in [20, 21]. The label texts of scenes are input to pre-trained language models, then a vector of the semantic embedding $\boldsymbol{e} \in \mathbb{R}^E$ is produced, where the number of dimensions E depends on the pre-trained language models. By leveraging the pre-trained language models, the proposed SED model can utilize unseen semantic contexts of scenes, i.e., data not used for training, in inference stages. Note that not only language models but also models for other modalities can be utilized to employ semantic contexts of the other modalities.

The semantic embedding $\boldsymbol{e}$ might not be appropriate for SED since module (b) comprises the language-model-trained texts. Hence, in the alignment module (c), the agreement between the semantic and acoustic embedding spaces is maximized. More specifically, a transformed semantic embedding $\boldsymbol{l} \in \mathbb{R}^L$, which is passed through two FFN layers ($\mathbb{R}^E \to \mathbb{R}^{E'} \to \mathbb{R}^L$), and an acoustic embedding $\boldsymbol{z} \in \mathbb{R}^A$, which is a bottleneck feature encoded by an autoencoder (AE), are followed by each FFN layer. Here, the input of the AE is the output of the last CNN layer of the feature extractor in module (a). L1 loss between $\boldsymbol{l}' \in \mathbb{R}^S$ and $\boldsymbol{z}' \in \mathbb{R}^S$ is then calculated, which is passed through each of the FFN layers. Finally, the aligned semantic embedding $\boldsymbol{l}$ is concatenated with each sequence of the first layer of the FFN layers before the time-distributed classifier in module (a) to take advantage of the prior semantic contexts of scenes as follows.

In module (c), to optimize the AE similarly to that in [22], the following mean squared error between the input $x_{t,f}$ and the reconstructed output $\hat{x}_{t,f}$ is used:

$$\mathcal{L}_{\mathrm{AE}} = \frac{1}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} (x_{t,f} - \hat{x}_{t,f})^2, \tag{2}$$

where F is the number of dimensions of the acoustic feature. To align the semantic embedding $\boldsymbol{l}'$ and acoustic embedding $\boldsymbol{z}'$, L1 loss is used as follows:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{S} \sum_{s=1}^{S} |l'_s - z'_s|, \tag{3}$$

where l'_s and z'_s represent the elements s of the semantic embedding $\boldsymbol{l}'$ and acoustic embedding $\boldsymbol{z}'$, respectively, and S is the number of dimensions of $\boldsymbol{l}'$ and $\boldsymbol{z}'$. Overall, to optimize the proposed network, the following objective function is used:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SED}} + \alpha \mathcal{L}_{\mathrm{AE}} + \beta \mathcal{L}_{\mathrm{align}}, \tag{4}$$

where hyperparameters α and β are tuned on the development set and set to 0.01 and 1.0, respectively.

4. EXPERIMENTS

4.1. Experimental conditions

We performed experiments to evaluate the performance of the proposed method using the TUT Sound Events 2016 [23], TUT Sound Events 2017 [24], TUT Acoustic Scenes 2016 [23], and TUT Acoustic Scenes 2017 [24] datasets. From these datasets, we chose audio clips comprising four scenes, "home," "residential area" (TUT Sound Events 2016), "city center" (TUT Sound Events 2017, TUT Acoustic Scenes 2017), and "office" (TUT Acoustic Scenes 2016), which include 266 min (development set, 192 min; evaluation set, 74 min) of audio. There were no sound event labels for the scenes "office" in TUT Acoustic Scenes 2016 and "city center" in TUT Acoustic Scenes 2017. We thus manually annotated the audio clips with sound event labels by the procedure described in [23, 24]. The resulting audio clips contained 25 types of sound event label [16]. The event label annotations for our experiments are available in [25].

As acoustic features, we used 64-dimensional log-mel energies computed for every 40 ms temporal frame with 50% overlap, where the clip length was 10 s. The audio clip shorter than 10 s was zero-padded. The sampling rate was 44.1 kHz. This setup is from

Table 1. Overall results for SED. 〔1〕 to 〔6〕 are conventional methods. 〔7〕 to 〔14〕 are proposed methods.

Method	SED module	*How scene is represented*	*How scene is fed to SED module*	Micro F-score	Macro F-score
〔1〕	CNN–BiGRU [7]	-	-	44.04% ±2.11	11.64% ±1.23
〔2〕	MTL of SED & ASC [16]	-	-	44.76% ±1.63	11.70% ±1.13
〔3〕	MTL of SED & SAD [31]	-	-	45.22% ±1.58	11.81% ±1.23
〔4〕	Conformer [10]	-	-	47.23% ±2.79	11.95% ±0.61
〔5〕	CNN–BiGRU	One-hot	Direct concatenation	47.04% ±2.45	13.82% ±1.25
〔6〕	MTL of SED & ASC [15]	One-hot	Direct concatenation	46.92% ±3.62	13.40% ±0.83
〔7〕	CNN–BiGRU	BERT embedding	Direct concatenation	46.91% ±2.13	13.79% ±1.22
〔8〕	CNN–BiGRU	GPT2 embedding	Direct concatenation	46.79% ±2.21	13.34% ±1.06
〔9〕	CNN–BiGRU	One-hot	Aligned concatenation	46.58% ±1.76	14.18% ±0.94
〔10〕	MTL of SED & ASC	One-hot	Aligned concatenation	48.28% ±0.99	13.79% ±0.84
〔11〕	CNN–BiGRU	BERT embedding	Aligned concatenation	48.00% ±1.84	14.16% ±1.02
〔12〕	CNN–BiGRU	GPT2 embedding	Aligned concatenation	48.17% ±1.98	**14.77%** ±0.82
〔13〕	Conformer	BERT embedding	Aligned concatenation	**51.57%** ±2.62	12.76% ±0.74
〔14〕	Conformer	GPT2 embedding	Aligned concatenation	50.03% ±2.25	12.90% ±0.68

Table 2. Experimental conditions

SED module	
Network architecture	3 CNN + 1 BiGRU + 1 FFN
# channels of CNN layers	128, 128, 128
Filter size ($T \times F$)	3×3
Pooling size ($T \times F$)	8×1, 2×1, 2×1 (max pooling)
# units in BiGRU layer	32
# units in FFN layers	128, 48
# units in output layer	25
CNN–AE in alignment module	
Network architecture	1 CNN (encoder) + 3 Deconvolution (decoder)
# channels of CNN layers	64, 128, 128, 128
Filter size ($T \times F$)	3×3, 3×3, 4×3, 4×3
Pooling size ($T \times F$), only encoder	1×25 (max pooling)

Table 3. Seen or unseen scenes for each experiment

	Seen	Unseen
Experiment 1	"city center" "home" "office" "residential area"	-
Experiment 2	"home" "office" "residential area"	"city center" "downtown"
	"city center" "home" "office"	"residential area" "apartment"

the baseline system of DCASE2018 Challenge task4 [26]. We performed the experiments using ten initial values of the model parameters. A segment-based metric [27] was used to evaluate the performance of SED. We set the segment size to the frame length. The threshold value for binarizing $\sigma(y_{n,t})$ was 0.5. As an optimizer, we used AdaBelief [28]. For AdaBelief, smoothing parameters β_1 and β_2, and a small number ϵ were tuned on the development set and set to 0.9, 0.999, and 10^{-3}, respectively. The activation function was Swish [29].

As baseline models, the conventional methods denoted as 〔1〕 to 〔6〕 in Table 1 were used. As one of the baseline models, we used the convolutional neural network and bidirectional gated recurrent unit (CNN–BiGRU) [7]. In addition, to verify the usefulness of the proposed method, we used a model combining SED and sound activity detection (SAD) based on multitask learning (MTL), referred to as "MTL of SED & SAD" [30], and a model combining SED and ASC, referred to as "MTL of SED & ASC" [16]. SAD is the process of estimating all active events in a time frame. The reason for choosing MTL of SED & SAD is that this modern method, in which no scene information is considered, is simple yet effective. MTL of SED & ASC is multitask-learning-based SED with ASC, which exploits scene labels for ASC. Moreover, we used Conformer [10] as the SED module, which achieved the best performance in DCASE2020 Challenge task4. To verify the effectiveness of the semantic embedding of scenes, the one-hot representation of scenes was also evaluated as a substitute for their semantic embedding [15].

In this work, as the language model in semantic embedding module (b) (Fig. 2), we used bidirectional encoder representations from transformers (BERT) [31] and generative pre-trained transformer 2 (GPT2) [32], which are pre-trained models and frozen when the SED module (a) and the alignment module (c) are trained. For GPT2, we used the last sequence of the layer before the final layer to extract the semantic embedding. In module (b), E was 768 for BERT or 1280 for GPT2. In module (c), E', L, A, and S were 256, 64, 64, and 32, respectively, which were tuned on the development set. For the AE, we used a CNN–AE. The scale of the encoder of the CNN–AE was smaller than that of the decoder of the CNN–AE because the output of the CNN layers in module (a) was followed by the CNN–AE. Other experimental conditions are listed in Table 2, including the parameters of the CNN–AE and the baseline of module (a). In Table 2, X × Y denotes a filter size of X along the frequency axis by Y along the time axis.

4.2. Experimental results

We conducted the following two experiments.

- **[Experiment 1]**: We investigated the overall SED performance of the proposed method and the extent to which the semantic representation of scene contexts is beneficial for SED. In the inference stages, as the semantic contexts of the scenes, we employed the scene labels assigned to the training data, i.e., the seen semantic contexts shown in Table 3.
- **[Experiment 2]**: The aim of the experiments was to demonstrate that unseen semantic contexts boost SED under the audio of unseen scenes. We verified the SED performance using SED models learned without the audio and semantic contexts of evaluation target scenes. In the inference stages, SED was performed for the audio of the unseen scenes using the unseen contexts shown in Table 3.

[Experiment 1] Table 1 shows the overall results for SED in terms of F-score, where micro and macro denote overall and classwise scores, respectively. The numbers to the right of ± indicate standard deviations. Each method is tagged with an ID from 〔1〕 to 〔14〕,

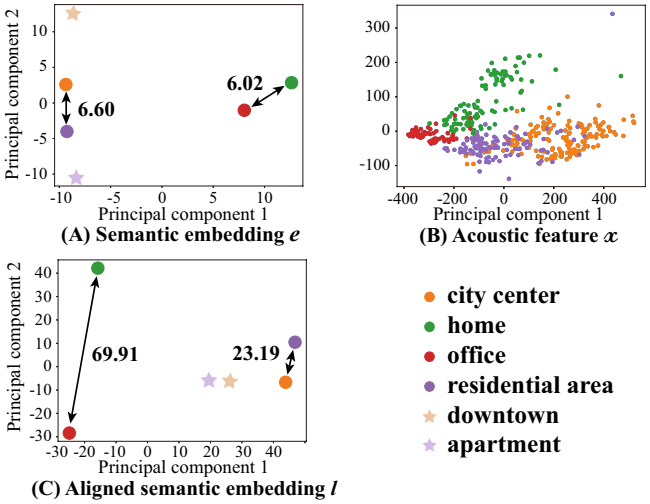

Fig. 3. Visualization of space of semantic embedding and acoustic feature

where 〔1〕 to 〔6〕 are the conventional methods and the others are the proposed methods. "One-hot" means the one-hot encoding of scenes is input to module (b) as a substitute for BERT or GPT2 embedding. "Direct concatenation" indicates that the scene representations (one-hot or BERT or GPT embedding) directly concatenate with the SED module. "Aligned concatenation" indicates that the scene representations concatenate with the SED module after the processing of the alignment by module (c). The results show that the F-score of the proposed methods is higher than that of the conventional methods. In particular, 〔13〕 and 〔12〕 improved the micro and macro F-scores by 4.34 and 3.13 percentage points compared with those of 〔4〕 and 〔1〕, respectively. However, 〔7〕 and 〔8〕 did not improve the F-score compared with that of 〔5〕. This means that the space of the non-aligned semantic embedding of scenes is different from that of the acoustic embedding of scenes. To confirm this, the embedding spaces based on 〔12〕 are visualized by principal component analysis (PCA) in Fig. 3. As shown in Fig. 3 (A), the distance between "city center" and "residential area" (6.60) is almost the same as that between "home" and "office" (6.02). However, in Fig. 3 (B), the distance between the clusters of "city center" and "residential area" is smaller than that between the clusters of "home" and "office." In Fig. 3 (C), the distance between "city center" and "residential area" (69.91) is smaller than that between "home" and "office" (23.19), that is, the embedding $\boldsymbol{l}$ is well aligned with the acoustic feature. "downtown" and "apartment" will be described later.

[Experiment 2] Tables 4 (A) and (B) give the SED performance for each event in the audio of the scenes "city center" and "residential area" using the SED models trained without the audio and contexts of "city center" and "residential area," respectively. In this experiment, the conventional method 〔1〕 and proposed method 〔12〕, which achieved the best macro F-scores, were used. Here, "Unseen context" was input to module (b) in the inference stages for the evaluated unseen scenes. Note that the conventional methods 〔5〕 and 〔6〕 cannot exploit unseen contexts of scenes. To further verify the SED performance using synonyms of the contexts, we added two contexts, "`downtown`" and "`apartment`," which are similar to "`city center`" and "`residential area`," respectively, from [33]. Note that there was no corresponding audio for "`downtown`" and "`apartment`," as shown in Fig, 3. Table 4 (A) shows the results obtained when the SED models were trained

Table 4. SED results for each event and scene, obtained using SED models learned without audio and semantic contexts of evaluation target scenes. 〔1〕 is a conventional method and 〔12〕 is a proposed method.

(A) evaluated scene: city center
trained scenes: home, office, residential area

Method	Unseen context	car	children	large vehicle
〔1〕	-	41.72%	3.17%	14.59%
〔12〕	`city center`	**42.93%**	**4.68%**	**15.88%**
〔12〕	`downtown`	40.29%	3.97%	15.35%

(B) evaluated scene: residential area
trained scenes: city center, home, office

Method	Unseen context	car	brakes squeaking	people talking
〔1〕	-	**41.41%**	0.00%	3.09%
〔12〕	`residential area`	36.80%	**0.68%**	**5.56%**
〔12〕	`apartment`	24.09%	0.00%	4.52%

using the audio of scenes "home," "office," and "residential area" with the corresponding semantic contexts, i.e., scene labels "`home`," "`office`," and "`residential area`." The audio and contexts of "`city center`" and "`downtown`" were not used for the training. The results indicate that the unseen semantic contexts of scenes are effective for more accurate SED under the audio of unseen scenes. Using the unseen context "`city center`," the events are well detected compared with method 〔1〕. In other words, the unseen context indeed boost SED under the audio of the unseen scene. This is because the audio and context of "city center" are similar to those of "residential area" used for training, as shown in Fig. 3. Using the unseen context "`downtown`," which is similar to "`city center`," as shown in Fig. 3 (C), most of the events are well detected compared with method 〔1〕. Thus, the coarse semantic contexts (synonyms) of scenes are also useful for SED even when the fine context is not known. Table 4 (B) shows the results obtained when the SED models were trained using the audio of scenes "city center," "home," and "office" with the corresponding semantic contexts, i.e., scene labels "`city center`," "`home`," and "`office`." Using the unseen context "`apartment`," the SED performance is highly degraded compared with method 〔1〕. This implies that the SED performance of some events is sensitive to even the slightest non-audio-aligned semantic embedding.

5. CONCLUSION

In this paper, we proposed scene-informed SED where pre-defined scene-agnostic contexts are available. In the proposed method, pre-trained large-scale language models were used, which enables SED models to employ contexts of not only pre-defined (seen) scenes but also unseen scenes. We further investigated the extent to which the semantic representation of scene contexts is helpful for SED. The experimental results show that the proposed method improves the micro and macro F-scores by 4.34 and 3.13 percentage points compared with the conventional Conformer- and CNN–BiGRU-based SED, respectively. Moreover, we confirmed that the unseen semantic contexts of the scenes can boost SED under the audio of unseen scenes. In our future work, we will investigate the SED performance guided by the semantic contexts of not only scenes but also sound events.

6. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number JP19K20304.

7. REFERENCES

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

[2] Y. Koizumi et al., "Unsupervised detection of anomalous sound based on deep learning and the neyman-pearson lemma," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 1, pp. 212–224, 2019.

[3] J. A. Stork et al., "Audio-based human activity recognition using non-Markovian ensemble voting," *Proc. IEEE Int. Symp. on Robot and Human Interactive Comm.* (*RO-MAN*), pp. 509–514, 2012.

[4] S. Ntalampiras et al., "On acoustic surveillance of hazardous situations," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* (*ICASSP*), pp. 165–168, 2009.

[5] S. Hershey et al., "CNN architectures for large-scale audio classification," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* (*ICASSP*), pp. 131–135, 2017.

[6] T. Hayashi et al., "Duration-controlled LSTM for polyphonic sound event detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2059–2070, 2017.

[7] E. Çakır et al., "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, 2017.

[8] Q. Kong et al., "Sound event detection of weakly labelled data with cnntransformer and automatic threshold optimization," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2450–2460, 2020.

[9] K. Miyazaki et al., "Weakly-supervised sound event detection with self-attention," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* (*ICASSP*), pp. 66–70, 2020.

[10] K. Miyazaki et al., "Conformer-based sound event detection with semi-supervised learning and data augmentation," *Tech. Rep. DCASE Challenge*, pp. 1–5, 2020.

[11] A. Mesaros et al., "Latent semantic analysis in sound event detection," *Proc. Euro. Signal Process. Conf.* (*EUSIPCO*), pp. 1307–1311, 2011.

[12] T. Heittola et al., "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Process.*, vol. 2013, no. 1, pp. 1–13, 2013.

[13] H. L. Bear et al., "Towards joint sound scene and polyphonic sound event recognition," *Proc. INTERSPEECH*, pp. 4594–4598, 2019.

[14] K. Imoto et al., "Sound event detection by multitask learning of sound events and scenes with soft scene labels," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* (*ICASSP*), pp. 621–625, 2020.

[15] T. Komatsu et al., "Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* (*ICASSP*), pp. 646–650, 2020.

[16] N. Tonami et al., "Joint analysis of sound events and acoustic scenes using multitask learning," *IEICE Trans. Inf. Syst.*, vol. E104-D, no. 02, pp. 294–301, 2021.

[17] J. Jung et al., "DCASENet: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* (*ICASSP*), pp. 621–625, 2021.

[18] N. Tonami et al., "Sound event detection based on curriculum learning considering learning difficulty of events," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.* (*ICASSP*), pp. 875–879, 2021.

[19] M. Cartwright et al., "SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context," *Proc. Workshop on Detect. and Classif. of Acoust. Scenes and Events* (*DCASE*), pp. 16–20, 2020.

[20] H. Hu et al., "Relational teacher student learning with neural label embedding for device adaptation in acoustic scene classification," *Proc. INTERSPEECH*, pp. 1196–1200, 2020.

[21] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1233–1242, 2021.

[22] S. Deshmukh et al., "Improving weakly supervised sound event detection with self-supervised auxiliary tasks," *Proc. INTERSPEECH*, pp. 596–600, 2021.

[23] A. Mesaros et al., "TUT database for acoustic scene classification and sound event detection," *Proc. Euro. Signal Process. Conf.* (*EUSIPCO*), pp. 1128–1132, 2016.

[24] A. Mesaros et al., "DCASE 2017 Challenge setup: Tasks, datasets and baseline system," *Proc. Workshop on Detect. and Classif. of Acoust. Scenes and Events* (*DCASE*), pp. 85–92, 2017.

[25] `https://www.ksuke.net/dataset`, Last accessed: 06/10/2022.

[26] R. Serizel et al., "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *Proc. Workshop on Detect. and Classif. of Acoust. Scenes and Events* (*DCASE*), pp. 19–23, 2018.

[27] A. Mesaros et al., "Metrics for polyphonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, pp. 1–17, 2016.

[28] Z. Juntang et al., "AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients," *Proc. Conf. on Neural Inf. Process. Syst.* (*NeurIPS*), pp. 1–12, 2020.

[29] P. Ramachandran et al., "Searching for activation functions," *arXiv*, *arXiv:1710.05941,* 2017.

[30] A. Pankajakshan et al., "Polyphonic sound event and sound activity detection: A multi-task approach," *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust.* (*WASPAA*), pp. 323–327, 2019.

[31] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proc. Annual Conf. of the North American Chap. of the Assoc. for Computational Linguistics: Human Lang. Tech.* (*NAACL-HLT*), pp. 4171–4186, 2019.

[32] A. Radford et al., "Language models are unsupervised multitask learners," *Tech. Rep., OpenAI*, 2019.

[33] `https://www.thesaurus.com`, Last accessed: 06/10/2022.