ICIC Express Letters
Part B: Applications
Volume **14**, Number **2**, February **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **125–132**

SOUND EVENT DETECTION USING SEGMENTED LABELING

Donny Valentinus Chiara[1,*] and Derwin Suhartono[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
dsuhartono@binus.edu
*Corresponding author: donny.chiara@binus.ac.id

Received July 2022; accepted October 2022

Abstract. *Sound Event Detection (SED) is a field of research in computer science that aims to identify events from generated sounds. One of the problems in SED research is that the process of creating strong datasets takes a lot of effort. On the other hand, while a lot of research has been done on the weak dataset, the result still could not compare to the strong dataset. Recent researches open up a new path by creating a new labeling method called point label which is able to combine the easier labeling process while also producing a better result. In this paper, we proposed a new labeling method called "segmented labeling". This labeling method split the sound recording into several segments and used a similar principle as point label to overcome the difficulty of making strong label. The idea of our method is that sound event does not exist throughout the length of the recording. By splitting the recording into several segments, each one will be less likely to produce false positives. The result shows that our proposed model is able to reach F-score = 0.703 and ER = 0.428 which is better than our base model using the strong label.*
Keywords: Mel-filter bank, Acoustic scene classification, Point-labeled dataset, Sound event detection, Synthetic dataset

1. **Introduction.** Sound is a source of information that always exists by our side. By listening to sound, we can identify and understand the environment and the events that are happening there. Sound Event Detection (SED) is a field of research in computer science that aims to identify events from their sound. The main goal of SED is to let computer devices take advantage of the information and understand the sound events that occur [1]. SED can be used in a variety of fields and applications, for example, in contextual indexing and retrieval in multimedia databases, non-disruptive health care monitoring, and surveillance. In addition, the detected sound events can also be used in other research areas such as audio context recognition, automatic tagging, audio segmentation [2], automatic classification of acoustic scenes, and automatic detection and classification of sound events [3].

In the SED research field, there are two main approaches used to train the SED model: fully supervised SED and weakly supervised SED. A fully supervised model has the potential to generate a better result than a weakly supervised SED. However, a fully supervised model requires training using strong labels, which consist of the event label and onsets and offsets for each event. One of the problems in the SED research is that strong labels require manual labeling for the timestamp, which in turn requires more time and cost to produce in large quantities. The weakly supervised SED approach tries to overcome this hindrance by using weak labels, which do not contain information about onsets and offsets. Weak labels took less time and cost to produce; therefore, weakly labeled datasets

DOI: 10.24507/icicelb.14.02.125

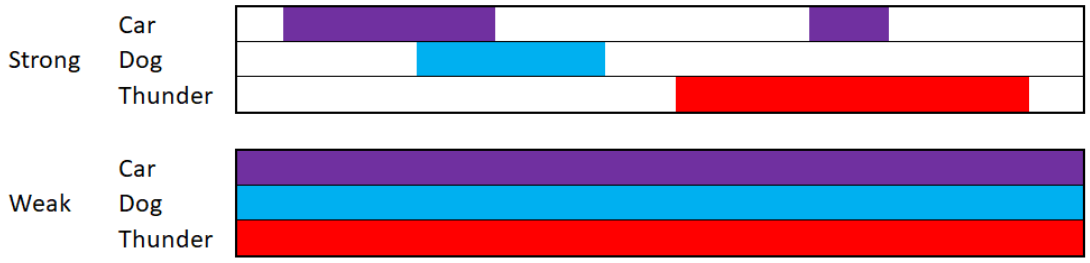

FIGURE 1. Visualization of strong and weak labels

are more often used [4]. A visualization of an audio clip labeled using strong labels and weak labels can be seen in Figure 1.

Although easier to produce, weakly supervised SED still cannot be compared with fully supervised SED. This is the result of removing time information needed for the localization process in SED [5]. While event classification methods tend to be accurate, event localization presents additional challenges, especially when large amounts of labeled data are not available [6]. In 2019, Kim and Pardo [7] proposed a new dataset labeling approach called point labels. This approach contains the event labels and time stamps denoting a point where an event happened in the audio recording, which can be seen in Figure 2. This labeling technique has an advantage where the difficulty level of labeling is comparable to the labeling of weak labels while still maintaining providing some time information.

FIGURE 2. Visualization of point label

Inspired by the problem presented in their paper, we propose a new labeling technique by using the principle of point labels to overcome the difficulty of making strong labels, combined with the principle of weak labels to cover more area. Our proposed label will contain several more points along the length of the sound events presented in the audio clip denoting the existence of a sound event. Then the audio clip will be split into several segments and use the prior point labels to label each segment based on the property of the weak label. The objective of this new labeling technique is to capture more temporal information than point labels while staying cost-effective. The visualization of the segment labels can be seen in Figure 3.

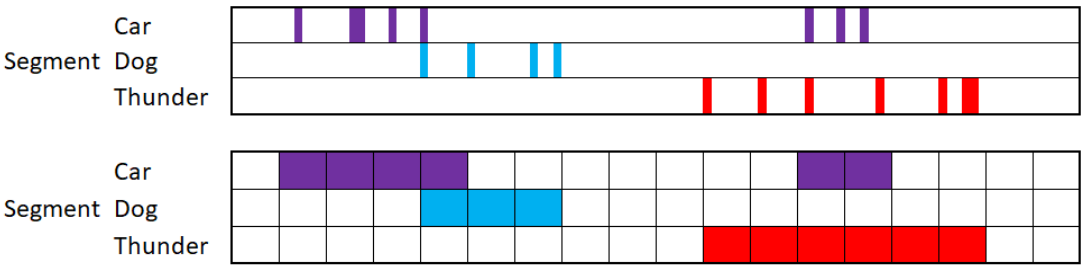

FIGURE 3. Visualization of segment label

The rest of this paper is structured as follows. Section 2 shows the literature review and related works. Section 3 outlines the methodology. Section 4 shows the result and discussion. Lastly, Section 5 presents the conclusion and suggestions for future research.

2. Literature Review.

2.1. Sound event detection.

Sound Event Detection (SED) is a field of research that identifies the class and estimates the starting and end point of each sound event in the sound recording. In practice, the goal of SED is to convert a sound recording in the form of an acoustic signal into a symbolic description of the related sound events contained in the corresponding sound scene. Figure 4 visualized the result of SED, identifying sound events that appeared in the sound recording and pinpointing when they happened.

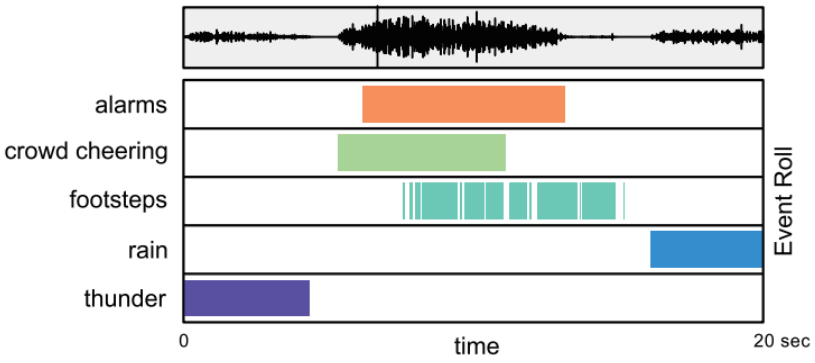

FIGURE 4. Upper panel: sound events in recording waveform; lower panel: sound event class annotation [8]

2.2. Strong model.

Cakir et al. [8] is one of the first SED experiments that focused on the localization of onset and offset rather than only classification. Their work combined the ability of Convolutional Neural Networks (CNN) to extract higher-level features that are invariant to local spectral and temporal variation and the ability of Recurrent Neural Networks (RNN) which are powerful in learning the longer-term temporal context in audio signals, resulting in F-score of 0.664 and error rate of 0.48. Another experiment on SED using the combination of CNN and RNN was done by Lu [9]. They compared the result of using 9-layer CNN, CNN+GRU, and CNN+LSTM on SED. The result of the comparison shows that CNN+LSTM wins against the other networks with an F-score of 0.904 and an error rate of 0.159. A more complex model was used by Nasiri et al. [10] by using a Mask R-CNN combined with a frame-based audio event analyzer to analyze each individual frame in the candidate segments of Mask R-CNN resulting in an F-score of 0.859 and an error rate of 0.28. These works show that strong SED models are able to get better results than weak models even when using a basic CRNN network.

2.3. Weak model.

The difficulty of creating strong datasets encourages researchers to develop the weak model. Some researchers like Lim et al. [11] and Harb and Pernkopf [12] tried to use an already established method, using CRNN as their base model. Lim et al. adopted the CNN and Gated Recurrent Unit (GRU) based Bidirectional Recurrent Neural Network (BiRNN) as their proposed system. Combined with an Inception module to search for the optimal local sparse structure in the convolutional network, their proposed model got 0.293 as the F-score. Harb and Pernkopf used a Gated-CRNN to predict the onset and offset of sound events and used Virtual Adversarial Training (VAT) for regulating the dataset, resulting in an F-score of 0.346 and an error rate of 1.12.

Besides using CRNN, there are also other researches focusing on other parts of the equation, for example on the pooling method. He et al. [13] proposed a hierarchical pooling structure to reduce the number of predicted probabilities for a certain class of sound events. They effectively improved the performance on linear softmax, exponential softmax, and attention. The best F-score was obtained by linear softmax, resulting in an F-score of 0.534 and an error rate of 0.69. McFee et al. [14] have also done an experiment focusing on

the pooling method by making an automatic pooling that smoothly interpolates between common pooling operators. Their proposed solution is able to outperform the standard pooling method on static and dynamic prediction, resulting in an F-score of 0.504 and an error rate of 0.665 on the URBAN-SED dataset, almost reaching their strong model with an F-score of 0.551 and error rate 0.642. Contrary to the strong model, weak SED models need to add more to the table in order to improve their F-score.

2.4. **Point model.** Kim and Pardo proposed a new labeling technique for the datasets of sound event detection, which they named point labeled dataset [7]. Their proposed labeling technique has an advantage over each labeling technique where it is easier to provide than strong labels while also significantly outperforming the weak model. They performed experiments to compare the result for the strong dataset, weak dataset, and point dataset using the same convolutional neural network model for each type. These experiments resulted in an F-score of 0.538 and an error rate of 0.523 for the point dataset which is not far off from the strong dataset with an F-score of 0.639 and an error rate of 0.519.

Based on observations, even though strong models only use the basic CRNN model and weak models use more complex models, there is still a big gap between the result of using a strong dataset and a weak dataset. On the other hand, the point model shows that there is a way to tackle both the problem of the strong dataset and the weak dataset, that is by creating a new labeling method that is easy to make, but also retains temporal information.

3. **Method.** The overall method can be seen in Figure 5. The segmented labeling technique uses a similar principle as the point label where a point somewhere in the range of a sound event is picked and labeled accordingly, the difference for the segmented label is that several points are picked instead of only one. After the points have been labeled, the audio recordings are split into several same-sized segments. Then each segment is labeled based on the number of points that occurred in the respective segment. The basic idea of this technique is that sound event does not exist throughout the length of the recording;

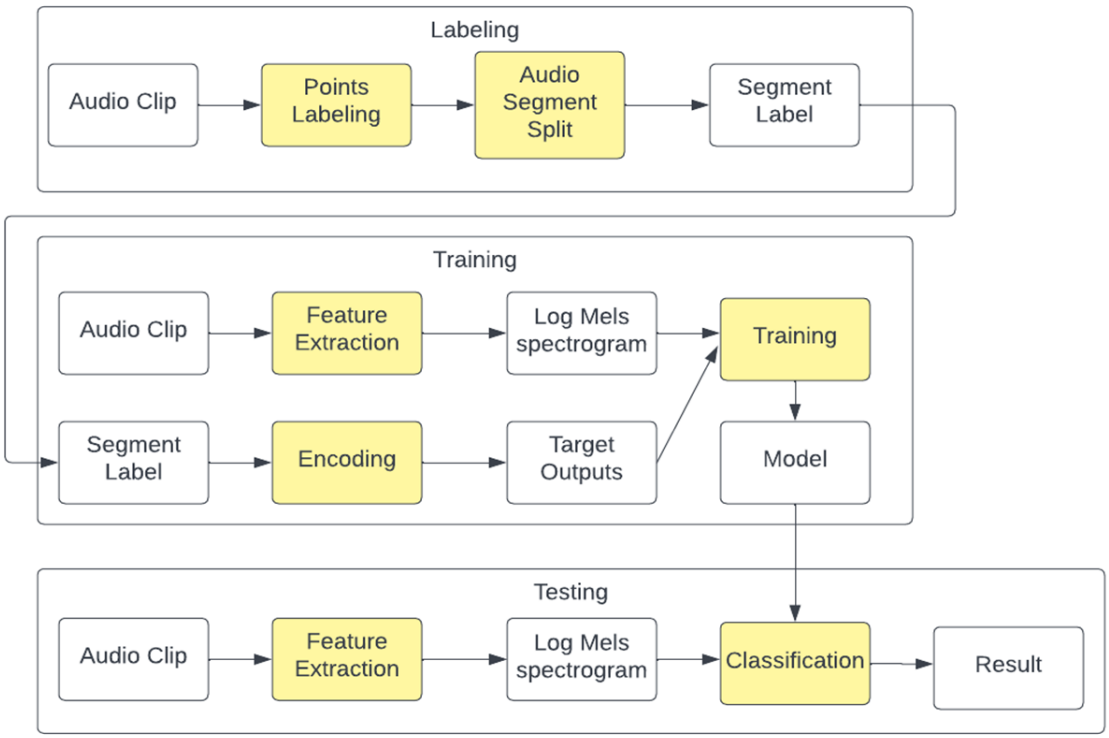

FIGURE 5. Segmented labeling SED methodology

therefore, the weak label generates a lot of false positives. By splitting the recording into several segments, each segment will be more likely to produce the correct result instead of false positives. The next step uses the standard SED method [15] where the feature of the audio clip is extracted into a log Mel-spectrogram and the segment label is processed and transformed to fit the target output for the training. Both the spectrogram and the target output are used for the training to produce a model which can be used for the sound event classification.

3.1. **Dataset.** The dataset used in this experiment is sound data from DESED_synthetic [16]. This dataset consists of 2045 synthetic audio recordings, each 10 s long. Each audio recording contains one or more sound events from the 10 available sound classes in the dataset. The classes are alarm bell ringing, blender, cat, dishes, dog, electric shaver/ toothbrush, frying, running water, speech, and vacuum cleaner. For the training and testing, the dataset is split 80 : 20.

The reason for choosing this dataset is that this dataset already has a strong label that can be used as the basis for the randomly generated point for the point and segment label. The labeling of point labels follows the paper by Kim and Pardo [7] by selecting one random point between the onset and offset of each entry of the strong label to be used as the onset. The offset is picked by moving forward 1 s from the onset. Both the onset and offset are moved accordingly if they went out of the original strong label boundary. For the segmented label, 10 points are picked randomly between the onset and offset of each entry to simulate the real case of a human annotator noting an event that happens during this period. Then the audio clip is split into 0.5-second segments and the number of points is counted for each segment. After counting, segments having points above the threshold will be labeled with the corresponding sound class.

3.2. **Model architecture.** The architecture used for training SED is a CRNN model which can be seen in Table 1. The model takes in the log Mel-spectrogram of the sound clips as the input. The features are then extracted by the three CNN layers, each having a 3×3 kernel size. On each CNN layer, a max pooling operation is performed to reduce the dimension of the model. Each convolutional layer also has a batch normalization to normalize output from the previous step and a dropout layer to prevent overfitting. The RNN part uses two layers of Bidirectional Gated Recurrent Unit (BiGRU) to learn the connection between time frames. The final output is a value in the range of $[0, 1]$ denoting the probability of each sound class happening for each frame, which then is thresholded into binary results.

3.3. **Evaluation model.** The result from the sound event detection will be measured with F-score and error rate for each one-second segment [17]. The F-score can be calculated from the value of True Positive (TP), False Positive (FP), and False Negative (FN) as seen in Equation (1). The error rate can be calculated from the number of insertions (I), deletions (D), and substitutions (S) as seen in Equation (2).

$$F = \frac{2P \cdot R}{P + R}, \text{ where } P = \frac{\sum TP(k)}{\sum TP(k) + \sum FP(k)}, \; R = \frac{\sum TP(k)}{\sum TP(k) + \sum FN(k)} \tag{1}$$

$$ER = \frac{\sum S(k) + \sum D(k) + \sum I(k)}{\sum N(k)} \tag{2}$$

4. **Result and Discussion.** We compared the result from our model using segmented labeling with the model using strong and point labeling. All the models were trained with the same model focusing on minimizing error rate. Table 2 shows the average F-score and ER for each model. As seen in the table, the model using our proposed labeling method is able to get an F-score of 0.703 and an ER of 0.428, beating both the model using strong label and point label both in ER and F-score. In comparison, Table 3 provides the result

TABLE 1. The structure of the proposed convolutional recurrent neural network model

Layer	**Detail**	**Output shape**
Input	Log Mel-Spectrogram	$(2 \times 256 \times 20)$
1st CNN layer	Convolutional 2D	$(2 \times 256 \times 128)$
	Batch Normalization	$(2 \times 256 \times 128)$
	Activation ReLu	$(2 \times 256 \times 128)$
	Max Pooling	$(2 \times 51 \times 128)$
	Dropout	$(2 \times 51 \times 128)$
2nd CNN layer	Convolutional 2D	$(2 \times 51 \times 128)$
	Batch Normalization	$(2 \times 51 \times 128)$
	Activation ReLu	$(2 \times 51 \times 128)$
	Max Pooling	$(2 \times 25 \times 128)$
	Dropout	$(2 \times 25 \times 128)$
3rd CNN layer	Convolutional 2D	$(2 \times 25 \times 128)$
	Batch Normalization	$(2 \times 25 \times 128)$
	Activation ReLu	$(2 \times 25 \times 128)$
	Max Pooling	$(2 \times 12 \times 128)$
	Dropout	$(2 \times 12 \times 128)$
	Permute	$(12 \times 2 \times 128)$
	Reshape	(256×12)
RNN layer	Bi-GRU	(256×32)
	Bi-GRU	(256×32)
	Time Distributed	(256×32)
	Dropout	(256×32)
	Time Distributed	(256×10)
Output	Activation Sigmoid	(256×10)

TABLE 2. F-score and error rate for each model

Data labeling method	F-score	ER
Strong	0.650	0.480
Point	0.616	0.701
Segmented	**0.703**	**0.428**

TABLE 3. F-score and error rate of the previous model

Data labeling method	F-score	ER
Point single (Kim and Pardo [7])	0.612	0.533
Point expanded (Kim and Pardo [7])	0.638	0.523
Strong (McFee et al. [14])	0.551	0.642

of the previous models by Kim and Pardo [7] compared with a recent strong model [14]. Our proposed model is able to get a better F-score and ER than both of their models.

To further examine the detail of the model, Table 4 shows the comparison of the system performances across sound classes present in the dataset. It can be seen that our proposed model is able to outperform the point model in almost every class and also able to perform comparatively well with the strong model. From this experiment, it can be observed that while some classes like 'speech' perform well on all three models, there are also some classes not performing as well on any models such as the 'vacuum cleaner' class.

TABLE 4. Class-wise F-scores

Class	Avg. duration	Strong	Point	**Segment**
Dishes	0.580	0	0.678	**0.005**
Dog	0.984	0.229	0.726	**0.786**
Cat	1.068	0.429	0.481	**0.649**
Alarm bell ringing	1.072	0.769	0.533	**0.736**
Speech	1.160	0.886	0.78	**0.84**
Blender	2.582	0.539	0.348	**0.477**
Running water	3.914	0.218	0	**0.518**
Electric shaver/toothbrush	4.518	0.819	0.136	**0.801**
Frying	5.171	0.701	0	**0.697**
Vacuum cleaner	5.288	0.13	0	**0.235**

One of the factors causing varying results for each class is the duration of the sound event. Some classes like 'frying', 'running water', and 'electric shaver/toothbrush' last longer with an average of four to five seconds per clip, while the other class like 'dishes' only last for half a second. From our observation, longer lasting events risk the over-generalization of the model resulting in a lot more false positives. Longer-lasting classes also tend to get a worse result in the point model because of the loss of temporal information when using only one point of reference. On the other hand, the classes with a shorter duration possess no problem with the point model while they do not do well with the strong and segmented model. We believed that our strong and segmented model is more fitted into medium to long events, resulting in a lot more false-negative in shorter-lasting classes, especially in the 'dishes' class with an average duration of only 0.58 s, which is far shorter than the other 9 classes.

Figure 6 shows the visualization of the class-wise breakdown in the form of a chart. The visualization of the class-wise breakdown shows that the strong model and segmented model are fairly more consistent compared with the point model. The graph also shows a more visible division between classes based on the duration of the event as discussed above.

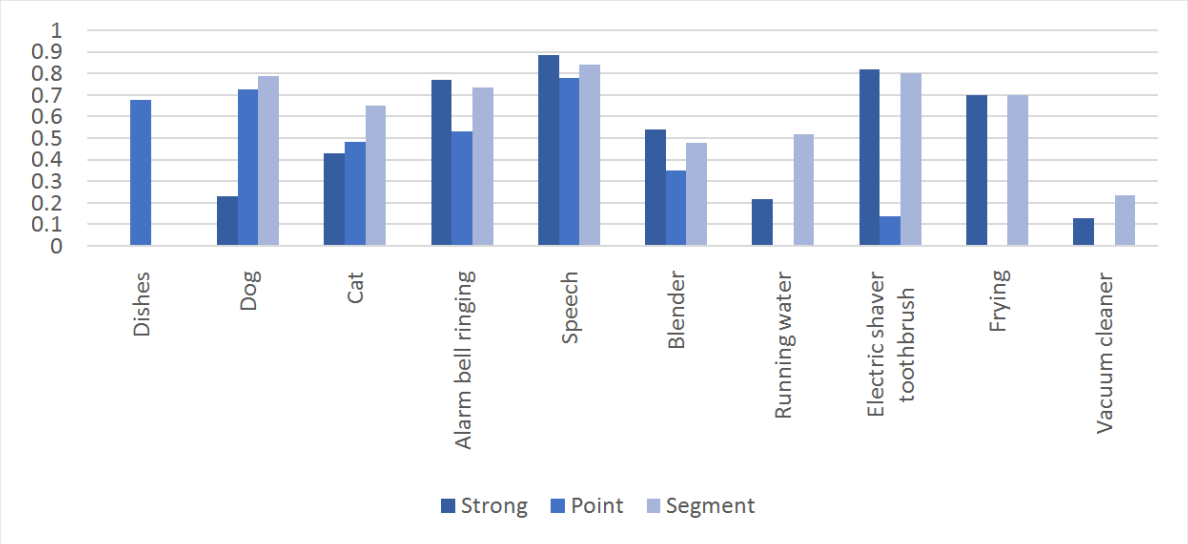

FIGURE 6. Class-wise F-scores

5. **Conclusion.** This paper proposed a new labeling method for sound event detection by combining the principle of point label and weak label to overcome the difficulty of making strong labels while still keeping the temporal information. Our experiment shows that our proposed method is able to produce a prediction with an F-score reaching 0.703 with ER of 0.428 triumphing over the strong and point model while keeping the labeling method easy to perform.

In the future, this paper can be expanded by using the proposed labeling method on other SED models, especially the more advanced and complicated SED models like CRNN-Transformer and automatic threshold optimization used in weak SED [4] to improve the F-score and error rate. Another possible future work is making a real segmented dataset annotated by a human to be compared with the randomly generated dataset used in this experiment.

REFERENCES

[1] Y. Fu, K. Xu, H. Mi, H. Wang, D. Wang and B. Zhu, A mobile application for sound event detection, *Int. Jt. Conf. Artif. Intell. (IJCAI)*, pp.6515-6517, doi: 10.24963/ijcai.2019/941, 2019.

[2] T. Heittola, A. Mesaros, A. Eronen and T. Virtanen, Context-dependent sound event detection, *EURASIP J. Audio, Speech, Music Process.*, vol.2013, no.1, doi: 10.1186/1687-4722-2013-1, 2013.

[3] A. Mesaros et al., Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol.26, no.2, pp.379-393, doi: 10.1109/TASLP.2017.2778423, 2018.

[4] Q. Kong, Y. Xu, W. Wang and M. D. Plumbley, Sound event detection of weakly labelled data with CNN-Transformer and automatic threshold optimization, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol.28, no.c, pp.2450-2460, doi: 10.1109/TASLP.2020.3014737, 2020.

[5] H. Dinkel, M. Wu and K. Yu, Towards duration robust weakly supervised sound event detection, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol.29, no.8, pp.887-900, doi: 10.1109/TASLP.2021.3054313, 2021.

[6] S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe and M. Elhilali, Joint acoustic and class inference for weakly supervised sound event detection, *ICASSP*, pp.36-40, 2019.

[7] B. Kim and B. Pardo, Sound event detection using point-labeled data, *2019 IEEE Work. Appl. Signal Process. to Audio Acoust.*, pp.1-5, 2019.

[8] E. Cakir, 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, doi: 10.1109/TASLP.2017.2690575, 2017.

[9] Z. Lu, *Sound Event Detection and Localization Based on CNN and LSTM*, Technical Report, Corpus ID: 233420598, 2019.

[10] A. Nasiri, Y. Cui, Z. Liu, J. Jin, Y. Zhao and J. Hu, AudioMask: Robust sound event detection using Mask R-CNN and frame-level classifier, *Proc. of Int. Conf. Tools with Artif. Intell. (ICTAI)*, pp.485-492, doi: 10.1109/ICTAI.2019.00074, 2019.

[11] W. Lim, S. Suh and Y. Jeong, Weakly labeled semi-supervised sound event detection using CRNN with inception module, *Detect. Classif. Acoust. Scenes Events 2018*, 2018.

[12] R. Harb and F. Pernkopf, Sound event detection using weakly-labeled semi-supervised data with GCRNNs, VAT and self-adaptive label refinement, *Detect. Classif. Acoust. Scenes Events 2018*, pp.1-5, 2018.

[13] K. X. He, Y. H. Shen and W. Q. Zhang, Hierarchical pooling structure for weakly labeled sound event detection, *Proc. of Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp.3624-3628, doi: 10.21437/Interspeech.2019-2049, 2019.

[14] B. McFee, J. Salamon and J. P. Bello, Adaptive pooling operators for weakly labeled sound event detection, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol.26, no.11, pp.2180-2193, doi: 10.1109/TASLP.2018.2858559, 2018.

[15] A. Mesaros, T. Heittola, T. Virtanen and M. D. Plumbley, Sound event detection: A tutorial, *IEEE Signal Process. Mag.*, vol.38, no.5, pp.67-83, doi: 10.1109/MSP.2021.3090678, 2021.

[16] N. Turpault and R. Serizel, *DESED_synthetic*, doi: 10.5281/zenodo.3550598, 2020.

[17] A. Mesaros, T. Heittola and T. Virtanen, Metrics for polyphonic sound event detection, *Appl. Sci.*, vol.6, no.6, doi: 10.3390/app6060162, 2016.