AUXILIARY DECODER-BASED LEARNING OF SOUND EVENT DETECTION USING MULTI-CHANNEL FEATURES AND MAXIMUM PROBABILITY AGGREGATION

Sang Won Son[1], Jongyeon Park[1], Hong Kook Kim[1,2]

[1] AI Graduate School
[2] School of EECS
Gwangju Institute of Science and Technology
Gwangju 61005, Korea
{{ssw970519, jypark3737}@gm., hongkook@}gist.ac.kr

Sulaiman Vesal[3]

[3]AI Lab., Innovation Center
Hanwha Vision
Teaneck, NJ 07666, USA
s.vesal@hanwha.com

Jeong Eun Lim[4]

[4]AI Lab., R&D Center
Hanwha Vision
Seongnam-si, Gyeonggi-do 13488, Korea
je04.lim@hanwha.com

ABSTRACT

This paper proposes a sound event detection (SED) model operating on heterogeneous labeled and/or unlabeled datasets, such as the DESED and MAESTRO datasets. The proposed SED model is based on a frequency dynamic convolution (FDY)–large kernel attention (LKA)-convolutional recurrent neural network (CRNN), and it is trained via mean-teacher-based semi-supervised learning to handle unlabeled data. The FDY–LKA-CRNN model incorporates bidirectional encoder representation from audio transformer (BEATs) embeddings to improve high-level semantic representation. However, the contribution of the BEATs encoder to the performance of the combined SED model is over-emphasized relative to that of the FDY–LKA-CRNN, which limits the overall performance of the SED model. To prevent this problem, an auxiliary decoder is applied to train the SED model with BEATs embeddings. Additionally, to accommodate the different recording characteristics of sound events in the two datasets, multi-channel log-mel features are concatenated in a channel-wise manner. Finally, a maximum probability aggregation (MPA) approach is proposed to address the different labeling time intervals of the two datasets. The performance of the proposed SED model is evaluated on the validation dataset for the DCASE 2024 Challenge Task 4, in terms of class-score-based polyphonic sound detection score (PSDS) and macro-average partial area under the receiver operating characteristic curve (MpAUC). The results show that the proposed model performs better than the baseline. In addition, the proposed SED model employing the multi-channel log-mel feature, auxiliary decoder, and MPA outperforms the baseline model. Ensembling several versions of the proposed SED model improves PSDS and MpAUC, scoring 0.038 higher in the sum of PSDS and MpAUC compared to the baseline model.

Index Terms— Sound event detection (SED), semi-supervised learning, auxiliary decoder, multi-channel log-mel feature, maximum probability aggregation

* This work was supported in part by Hanhwa Vision Co. Ltd., the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korean government (MSIT) (No.2022-0-00963), and the "Practical Research and Development support program supervised by the GTI" grant funded by the GIST in 2024.

1. INTRODUCTION

Sound event detection (SED) aims to localize and classify individual sound events originating from acoustic signals, along with their corresponding timestamps. In recent years, the use of deep learning for SED has been widely researched [1]. While the performance of SED is satisfactory in some applications, such as [2, 3], a major challenge for developing deep learning-based SED models still remains in view of the preparation of label audio data with timestamps, which is expensive and time-consuming. This has prompted the development of weakly supervised and semi-supervised learning techniques [4] based on weakly labeled and unlabeled datasets [5]. Recently, a soft label–based dataset, called the Multi-Annotator Estimated STROng labels (MAESTRO) dataset [6], has also been employed to reduce the overall cost of annotating strong labels while maintaining the timestamps of sound events.

However, the use of mixtures of differently labeled data for SED yields a time misalignment problem that an inconsistency arises in the time recording units between the heterogeneously labeled datasets. In other words, soft labels contain label information over 1 s recording unit, whereas weakly labeled and unlabeled datasets, e.g., the Domestic Environment Sound Event Detection (DESED) dataset, contain sound events recorded over shorter units than 1 s. In addition to this time misalignment problem, there is another mismatch problem in the recording characteristics of sound events in the different datasets.

Thus, this paper proposes a maximum probability aggregation (MPA) approach for SED to address the time misalignment between the DESED and MAESTRO datasets. In addition, to accommodate time-frequency patterns according to different recording characteristics, a multi-channel log-mel feature is extracted to help the SED model capture sound events from two different datasets.

The proposed MPA and multi-channel log-mel feature are applied to an SED model, named a frequency dynamic convolution (FDY) [7]–large kernel attention (LKA) [8]-convolutional recurrent neural network (CRNN) model, which was developed for the DCASE 2023 Challenge Task 4A [9]. The FDY–LKA-CRNN

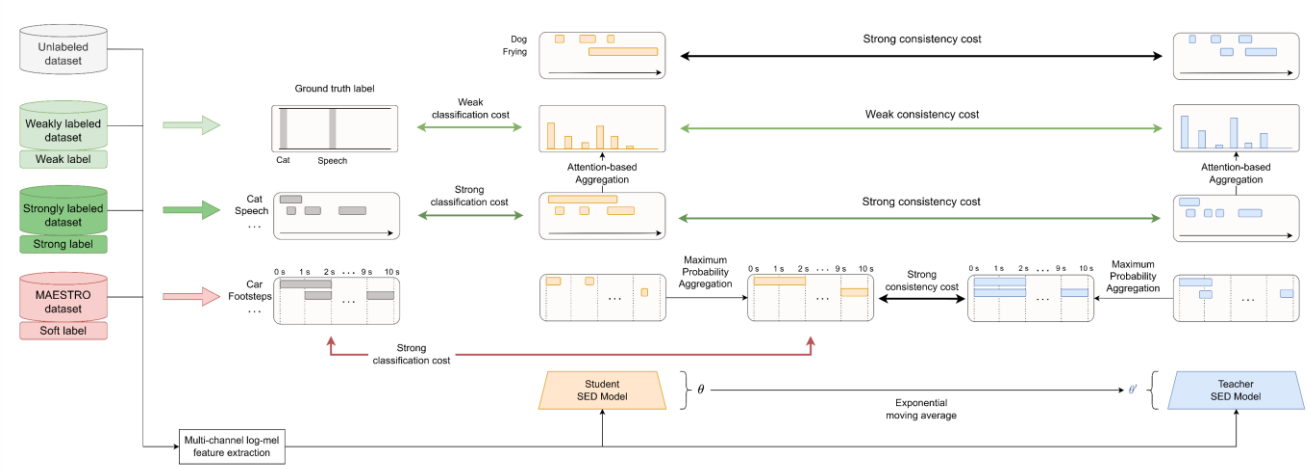

Figure 1. Illustration of the proposed SED model training procedure, focused on maximum probability aggregation.

model is trained via mean-teacher-based semi-supervised learning to handle unlabeled data, and it incorporates bidirectional encoder representation from audio transformer (BEATs) [10] embeddings to improve high-level semantic representation. However, the contribution of the BEATs encoder to the performance of the combined SED model is over-emphasized relative to that of the FDY–LKA-CRNN. To further improve the overall performance of the SED model, an auxiliary decoder [11] is applied to train the SED model with BEATs embeddings.

Our contributions can be summarized as follows:

- To deal with the time misalignment issue between the DESED and MAESTRO datasets, we propose MPA, which effectively aligns the time intervals between the predicted strong labels of the SED model and the soft labels in the MAESTRO dataset, thereby improving the overall performance of the SED model.
- To extract the heterogeneous time-frequency patterns of the sound events between the two datasets, we propose a multi-channel log-mel feature extraction method. Especially the feature improves a metric about MAESTRO dataset.
- Finally, we incorporate an auxiliary decoder to balance the contributions of the convolutional block and pretrained model by providing additional loss weighting during training. Consequently, the proposed auxiliary decoder-based training improves SED performance in both datasets.

The remainder of this paper is organized as follows: Section 2 describes the dataset and input features of the SED model developed in this study. Section 3 proposes a multi-channel log-mel feature and MPA, and also incorporates the auxiliary decoder for SED model training. Section 4 evaluates the performance of the developed SED model on the DCASE 2024 Task 4 validation dataset and compares the SED performance according to different combinations of the proposed approaches. Finally, Section 5 concludes this paper.

2. DATASET

Unlike in 2023, the database for the DCASE 2024 Challenge Task 4 comprises the DESED and MAESTRO datasets. The DESED dataset, which is identical to that for the last year's DCASE Challenge, contains several types of data such as weakly labeled data, unlabeled in-domain training data, strongly labeled synthetic data, and strongly labeled real data. All the audio clips span 10 seconds each. The weakly labeled dataset is composed of 1,578 clips with only class labels. The unlabeled in-domain training dataset contains 14,412 audio clips. Finally, the strongly labeled real and synthetic datasets contain 3,470 and 10,000 clips, respectively, where the strongly labeled synthetic dataset is created using Scraper [12]. Note that the number of audio event classes is 10 in this dataset.

The original MAESTRO dataset contains audio clips longer than 180 seconds. However, to balance the length of audio clips in this dataset with that in the DESED dataset, the audio clips are cropped to 10 s, allowing a 9 s overlap between consecutively cropped audio clips. Each cropped audio clip is softly labeled into 10 vectors, where each vector is assigned to every segment of 1 s with a dimension of 19 for representing 19 audio event classes. Notice that the event classes in the DESED dataset are different from those in the MAESTRO dataset, except for two classes, e.g., "Speech" in DESED and "People Talking" in MAESTRO, and "Dishes" in DESED and "Cutlery and dishes" in MAESTRO. After merging the similar two classes, there are 27 classes in total.

The mono-channel signals in the two datasets are first resampled from 44.1 to 16 kHz to extract audio features. Then, the audio signals are segmented into frames of 2,048 samples with a hop length of 160 samples. A 2,048-point fast Fourier transform is applied to each frame, followed by a 128-dimensional mel-filterbank analysis. Each 10 s audio clip comprises 1,001 frames. Hence, the input feature dimensions are 1001×128. The retrieved mel-spectrogram features are then normalized based on the mean and standard deviation for all training audio samples. When extracting the multi-channel log-mel feature, we use identical parameters for preprocessing.

3. PROPOSED METHOD

The SED model is based on the FDY–LKA-CRNN architecture that was proposed in [9], and it is trained via semi-supervised learning in a mean-teacher framework. Fig. 1 shows the proposed

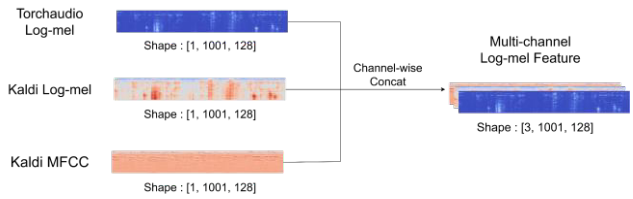

Figure 2. Illustration of the proposed multi-channel log-mel feature extraction procedure for obtaining the heterogeneous time-frequency patterns of the sound events.

SED model training procedure, where the newly proposed approaches, such as MPA and the multi-channel log-mel feature, are exaggerated. In addition to MPA and the multi-channel feature, the auxiliary decoder is intrinsically used for training the student and teacher models shown at the bottom of the figure. The following subsections sequentially describe MPA, the multi-channel feature, and the auxiliary decoder in detail.

3.1. Multi-channel log-mel feature

As mentioned in Section 2, there are different recording environments between the DESED and MAESTRO datasets, which are recorded in almost clean and noise conditions, respectively. To capture the diverse acoustic properties of the two datasets, we extract the multi-channel log-mel feature composed of 1) a log-mel spectrogram extracted using the Torchaudio framework, 2) a log-mel spectrogram extracted using Kaldi within the Torchaudio framework, and 3) the mel-frequency cepstral coefficient (MFCC) feature extracted using Kaldi within the Torchaudio framework.

Fig. 2 illustrates the proposed multi-channel log-mel feature extraction procedure for obtaining the heterogeneous time-frequency patterns of the sound events. First, three different feature vectors, as described above, are extracted and then concatenated channel-wise to create a multi-channel log-mel feature. This concatenated feature vector is input to the SED model during both training and inference. By leveraging multiple configurations to extract the log-mel features, it is expected that we create a robust input representation that effectively bridges the gap between the DESED and MAESTRO datasets.

3.2. Length-adjustable maximum probability aggregation

The FDY–LKA-CRNN-based SED model was developed for the DESED dataset, where audio data labels were assigned in segments less than 1 s. To accommodate different labels for sound events as in the MAESTRO dataset, we need to incorporate new techniques into the SED model. This is because the difference in labeling presents a significant challenge due to the mismatch in time intervals between the label information of the MAESTRO dataset and DESED dataset.

To deal with such a time misalignment problem, we propose the MPA. Compared to the labels in the DESED dataset, the soft labels in the MAESTRO dataset do not guarantee that a sound event entirely exists within each 1 s segment. The output of the SED model consists of predictions for 25 frames, which corresponds to a duration of 1 s. As shown in Figure 3, we select the highest probability value among these 25 frames and use this value as the class probability for the corresponding 1 s segment. This approach ensures that the time interval for the MAESTRO dataset would be aligned with the soft labels. This MPA is performed only during the training step.

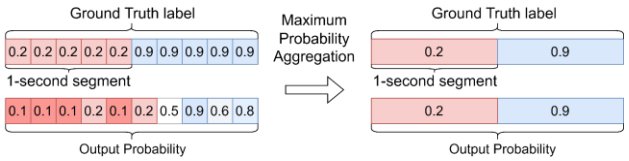

Figure 3. Example of implementing maximum probability aggregation, which is applied only to the MAESTRO dataset.

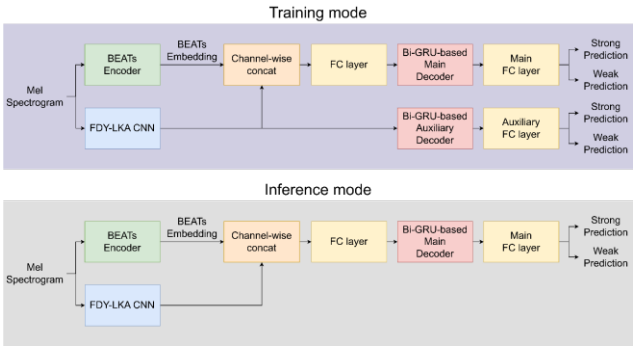

Figure 4. Network architecture of the proposed auxiliary decoder applied to train the FDY-LKA-CNN-based SED model with BEATs embeddings.

3.3. Auxiliary decoder

The BEATs encoder can extract the embedding corresponding to high-level semantic information, resulting in providing improved SED performance [9]. However, the contribution of the BEATs encoder to the performance of the combined SED model is overemphasized relative to that of the FDY–LKA-CRNN. Thus, we incorporate an auxiliary decoder to balance the contributions between the convolutional block and BEATs encoder by providing additional loss weighting during training.

Fig. 4 shows the network architecture of the proposed auxiliary decoder applied to train the FDY-LKA-CNN-based SED model with BEATs embeddings. The proposed auxiliary decoder mirrors the structure of the main decoder, consisting of two bidirectional gated recurrent units (Bi-GRUs) designed to capture temporal context information, followed by a fully connected (FC) classifier that uses a sigmoid function to calculate class probabilities. The auxiliary decoder does not share weights with the main decoder. Also, it is activated only during the training step, and a higher weight is assigned to the auxiliary loss in the initial training steps than the main loss. This guides the learning process so that the convolutional blocks are well-trained compared to without using the auxiliary decoder. During inference, the main decoder is only operated to generate the output of the SED model.

4. EXPERIMENTAL RESULTS

4.1. Model training

The parameters of the FDY–LKA-CRNN-based SED model were initialized through Xavier initialization [13]. The minibatch-wise adaptive moment estimation optimization technique [14] was employed, which involved decoupling the weight decay from the gradient-based updates. In addition, a dropout method [15] was applied to the FDY–LKA-CRNN model at a rate of 0.5. The learn-

Table 1: Performance comparison of the baseline and various versions of the proposed SED model on the validation dataset of DCASE 2024 Challenge Task 4.

Model	Auxiliary decoder	Maximum probability aggregation	Multi-channel log-mel feature	Ensemble	Validation Dataset		
					Class-score-based PSDS	MpAUC	Sum of metrics
Baseline: CRNN-based mean-teacher model [22]	–	–	–	–	0.49 ± 0.004	0.73 ± 0.007	1.22
FDY–LKA-CRNN	–	–	–	–	0.4799	0.665	1.144
FDY–LKA-CRNN–A	√	–	–	–	0.4922	0.673	1.164
FDY–LKA-CRNN–M	–	√	–	–	0.4959	0.692	1.187
FDY–LKA-CRNN–C	–	–	√	–	0.4663	0.709	1.175
FDY–LKA-CRNN–AM	√	√	–	–	0.5092	0.709	1.218
FDY–LKA-CRNN–MC	–	√	√	–	0.4832	0.733	1.216
FDY–LKA-CRNN–AC	√	–	√	–	0.4795	0.712	1.191
FDY–LKA-CRNN–AMC	√	√	√	–	0.5018	0.740	1.241
FDY–LKA-CRNN–AMC(E)	√	√	√	√	0.5162	0.742	1.258

ing rate was set based on the ramp-up strategy [4], with the maximum value reaching 0.001 after 50 epochs. Several augmentation techniques were applied to the train data, including time-frequency shift [16], time mask [17], mix-up [18], and filter augmentation [19].

4.2. Discussion

The performance of the proposed SED model was evaluated using the measures defined in the DCASE 2024 Challenge Task 4 [20]: class-score-based polyphonic sound detection score (PSDS) [21] and macro-average partial area under the receiver operating characteristic curve (MpAUC).

Table 1 compares the performance of the baseline with those of various versions of the proposed SED model on the validation dataset of the DCASE 2024 Challenge Task 4. As shown in the table, there are nine different versions in this study. The FDY–LKA-CRNN is the SED model identical to that in [9], which was developed in the DCASE 2023 Challenge. Then, we applied each of the three proposed approaches, such as auxiliary decoder, MPA, and multi-channel log-mel feature that are abbreviated as A, M, and C, respectively. For example, FDY–LKA-CRNN–A means the FDY–LKA-CRNN-based SED model trained using the proposed auxiliary decoder. The FDY–LKA-CRNN–AMC(E) means an ensemble model combined with the FDY–LKA-CRNN–AMCs obtained from 16 different checkpoints.

First of all, we observed the performance of FDY–LKA-CRNN SED model was degraded compared to that of the baseline model. This was because FDY–LKA-CRNN model was optimized to the labeling of the DESED dataset, as mentioned earlier. Then, we applied each of the three proposed approaches (A, M, and C) to FDY–LKA-CRNN. As shown from the third to fifth row in the table, any FDY–LKA-CRNN–X improved MpAUC compared to FDY–LKA-CRNN, while FDY–LKA-CRNN–C provided a little lower class-score-based PSDS than FDY–LKA-CRNN. However, combining any two out of three approaches achieved higher or comparable class-score-based PSDS and MpAUC to FDY–LKA-CRNN.

Next, we combined all the three approaches to construct FDY–LKA-CRNN–AMC. Then, it was revealed that FDY–LKA-CRNN–AMC yielded better than FDY–LKA-CRNN as well as the baseline model.

Finally, we constructed an ensemble model, FDY–LKA-CRNN–AMC(E), and compared its performance with the baseline and FDY–LKA-CRNN-based single models. As shown in the table, this ensemble model outperformed the baseline as well as the other single models. This superior performance was ascribed to the inherent advantages of ensemble modeling, such as reduced overfitting and improved model robustness.

5. CONCLUSIONS

In this paper, we proposed maximum probability aggregation and a multi-channel log-mel feature to improve SED performance when the training datasets were heterogeneously recorded and labeled. In addition, the auxiliary decoder-based training approach was proposed to balance the contributions of different representations prior to a classifier. In particular, our baseline model was FDY–LKA-CRNN with BEATs embeddings; thus, the auxiliary decoder could help the classifier get balanced information between the CNN block and the BEATs encoder. In summary, the auxiliary decoder enhanced the performance of the convolutional block, enabling it to extract semantics. MPA was applied to the MAESTRO dataset to match the time alignment between the output of the SED model and the soft labels. The multi-channel log-mel feature could help the SED model accommodate the various time-frequency patterns from the two different datasets used in this challenge. We constructed the SED model according to the rules of the DCASE 2024 Challenge Task 4. The experimental results showed that the SED model trained with the multi-channel log-mel feature, MPA, and auxiliary decoder increased the PSDS and MpAUC by 0.0118 and 0.01, respectively, compared to the baseline SED model. An ensemble model derived from the model checkpoints also improved the sum of PSDS and MpAUC by 0.038 over the baseline model.

In future work, we will investigate the effectiveness of the proposed approaches according to different neural architectures of SED models.

6. REFERENCES

[1] T. Khandelwal, R. K. Das, and E. S. Chng, "Sound event detection: A journey through DCASE Challenge series," *APSIPA Trans. Signal Inf. Process*., vol. 13, 2024.

[2] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Visual object detector for cow sound event detection," *IEEE Access*, vol. 8. pp. 162625–162633, 2020.

[3] S. Mohmmad, and S. K. Sanampudi. "Exploring current research trends in sound event detection: A systematic literature review," *Multimedia Tools and Applications*, pp. 1–43, 2024.

[4] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. International Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 1195–1204.

[5] N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[6] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Trans. Audio Speech Lang. Process*., vol. 31, pp. 902–914, 2023.

[7] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *arXiv preprint*, arXiv:2203.15296, 2022.

[8] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *arXiv preprint*, arXiv:2202.09741, 2022.

[9] J. W. Kim, S. W. Son, Y. Song, H. K. Kim, I. H. Song, and J. E. Lim, "Label filtering-based self-learning for sound event detection using frequency dynamic convolution with large kernel attention," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events*, 2023.

[10] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," *arXiv preprint*, arXiv:2212.09058, 2022.

[11] Y Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint,* arXiv:1904.06037, 2019.

[12] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[13] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, 2014.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.

[16] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4," *Tech. Rep. in DCASE 2019 Challenge*, 2019.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint*, arXiv:1904.08779, 2019.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv preprint*, arXiv:1710.09412, 2017.

[19] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4308–4312.

[20] https://dcase.community/challenge2024/task-sound-event-detection-with-heterogeneous-training-dataset-and-potentially-missing-labels.

[21] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1021–1025.

[22] https://github.com/DCASE-REPO/DESED_task/tree/master/recipes/dcase2024_task4_baseline.