MULTI-RESOLUTION CONFORMER FOR SOUND EVENT DETECTION: ANALYSIS AND OPTIMIZATION

Sara Barahona, Diego de Benito-Gorron, Sergio Segovia, Daniel Ramos, Doroteo T. Toledano

AUDIAS Research Group
Universidad Autónoma de Madrid
Calle Francisco Tomás y Valiente, 11, 28049 Madrid, SPAIN
sara.barahona@estudiante.uam.es, diego.benito@uam.es,
sergio.segoviag@estudiante.uam.es, daniel.ramos@uam.es, doroteo.torre@uam.es

ABSTRACT

The Conformer architecture has achieved state-of-the-art results in several tasks, including automatic speech recognition and automatic speaker verification. However, its utilization in sound event detection and in particular in the DCASE Challenge Task 4 has been limited despite winning the 2020 edition. Although the Conformer architecture may not excel in accurately localizing sound events, it shows promising potential in minimizing confusion between different classes. Therefore, in this paper we propose a Conformer optimization to enhance the second Polyphonic Sound Detection Score (PSDS) scenario defined for the DCASE 2023 Task 4A. With the aim of maximizing its classification properties, we have employed recently proposed methods such as Frequency Dynamic Convolutions in addition to our multi-resolution approach, which allow us to analyse its behaviour over different time-frequency resolution points. Furthermore, our Conformer systems are compared with multi-resolution models based on Convolutional Recurrent Neural Networks (CRNNs) to evaluate the respective benefits of each architecture in relation to the two proposed scenarios for the PSDS and the different time-frequency resolution points defined. These systems were submitted as our participation in the DCASE 2023 Task 4A, in which our Conformer system obtained a PSDS2 value of 0.728, achieving one of the highest scores for this scenario among systems trained without external resources.

Index Terms— DCASE 2023, Sound Event Detection, Conformer, PSDS, Multi-resolution, Model fusion

1. INTRODUCTION

Sound Event Detection (SED) is the task that aims to detect and classify different sound events present within an audio clip. Although research in SED has a long history, the last few years have witnessed an increasing interest in the field, motivated in part by the publication of Google Audio Set [1] and the yearly challenges and workshops organized by the DCASE community [2]. This paper is centered in the context of one of these challenges, in particular the DCASE Task 4A: Sound Event Detection with Weak Labels and Synthetic Soundscapes. The goal of this task is to evaluate SED systems by employing both real and synthetic recordings which contain 10 sound event classes that can be found in a domestic environment. Besides, it tackles the issue of employing unlabeled data as well as different types of annotations: strong labels that provide temporal information (timestamps) along with the sound event category, and weak labels which solely indicate the category.

The metric employed for evaluating SED systems in this task is the Polyphonic Sound Detection Score (PSDS) [3], that relies on the intersection between detected and annotated sound events. Considering that it can be tuned for evaluating different properties of a SED system, two PSDS scenarios are proposed for the DCASE Challenge 2023 Task 4A. Whereas the first one (PSDS1) focuses on a fast reaction upon a sound event, requiring highly accurate localization, the second scenario (PSDS2) aims to avoid the confusion between classes, and it is not strict about timing errors.

Over the last few years, different architectures have been proposed to address this task. Since 2018, the baseline is based on a Convolutional Recurrent Neural Network (CRNN) [4], which employs CNNs for extracting local characteristics and RNNs to exploit temporal dependencies. Architectures based on attention mechanisms such as the Transformer [5] or the Conformer (Convolution Augmented Transformer) [6] have also been explored for this task. The Conformer architecture has been successfully employed by recent state-of-the-art models in tasks such as automatic speech recognition (ASR) [7] and automatic speaker verification (ASV) [8]. In the field of sound event detection, it achieved promising results winning the DCASE Challenge Task 4 in 2020 [9]. However, in the subsequent editions it was scarcely used, to the extend that last year we were the only team that submitted systems based on this architecture [10]. Although our experiments revealed a better performance of CRNN-based systems in terms of PSDS1, we observed the potential of the Conformer at classifying sound events. Therefore, in this paper we propose a continuation to our previous research by optimizing the Conformer architecture towards the PSDS2 and analysing its performance following our multi-resolution approach.

For this purpose, we introduce the Conformer architecture and describe the methodologies employed for its optimization in Section 2. The results of our experiments are presented and analysed in Section 3. Finally, Section 4 highlights the salient conclusions derived from this investigation.

2. PROPOSED METHODS

The Conformer (Convolution-Augmented Transformer) was designed with the aim of building an attention-based network capable of extracting both local and global features. For this purpose, a convolution module is added to the Transformer backbone. To solve temporal confusion, the relative positional embedding proposed for the Transformer-XL [11] is added to the global content-based attention mechanism. While this approach initially appeared to be highly promising for addressing the detection and classification of sound events, the Conformer has exhibited limitations in accurately localizing timestamps, resulting in a lower performance in terms of PSDS1 when compared with CRNNs. However, the Conformer has shown a great ability at classifying correctly each sound event, even when two sounds are similar or noise is present in an audio clip.

Considering that the main weakness of the Conformer architecture is the lack of temporal resolution, we propose to optimize a Conformer-based system towards the PSDS2. To accomplish this objective, we employ a multi-resolution approach to assess the system's effectiveness across various time-frequency resolution settings. Considering the pronounced influence of median filtering on the temporal resolution of a SED system's output, we adapt this post-processing technique to the scenario we are targeting. To evaluate the proposed methods in the framework of the DCASE Challenge Task 4A, we compare the performance of our Conformer systems with a multi-resolution version of the official baseline model based on CRNNs.

2.1. Optimized Conformer for PSDS2

Our Conformer model is based on the DCASE 2020 Task 4 winner [9], which consist of a CNN for feature extraction with 4 conformer blocks stacked. Additionally, they employ a tagging token similar to the classification token used in BERT [12] to summarize the weak label predictions through the attention layers.

To improve the PSDS2 value, we perform a hyperparameter tuning setting as objective this metric, leading to an optimal configuration of 7 Conformer blocks with 4 attention heads each and an encoder dimension of 144. Additionally, we substitute the CNN-based feature extractor with a Frequency Dynamic Convolution Neural Network (FDY-CNN) [13] to improve the classification of non-stationary sound events. For the FDY-CNN we employ context gating as the activation function and define a time-resolution reduction of 8 by adding one more average-pooling layer along the temporal dimension. Data augmentation techniques have also been applied to avoid confusion between classes. By this means, we employ both Mixup and FilterAugment [14] with a probability of 50% of applying them to the training data.

As semi-supervised learning, we utilize the mean-teacher method [15] for training both architectures. This method employs two identical models: student and teacher, whose weights are the exponential average weights of the student. By minimizing a consistency cost between the predictions of the student and teacher, the model learns to generate targets from unlabeled data. Generally, the teacher model achieves a more consistent learning trajectory across epochs, leading to a superior performance during testing. Thus, model selection is performed over the teacher network, adjusting the objective metric based on the specific scenario we are targeting. Whereas for the CRNN we employ the one set for the baseline (F1-score based on intersection), our Conformer systems use the PSDS2.

2.2. Multi-resolution analysis

In previous research, we proposed a multi-resolution approach which consist on varying the parameters employed for the extraction of mel-spectrogram features. Our multi-resolution approach has demonstrated the advantages of employing distinct time-frequency resolutions that align with the characteristics of each PSDS scenario or sound event category. Given that the main weakness of the Conformer seems to be the time resolution of its detections, we will explore how the different time-frequency resolutions impact the performance of this architecture.

Resolution	$\mathbf{T}_{++}$	$\mathbf{T}_{+}$	**BS**	$\mathbf{F}_{+}$	$\mathbf{F}_{++}$
N	1024	2048	2048	4096	4096
L	1024	1536	2048	3072	4096
R	128	192	256	384	512
$\mathbf{n}_{\mathbf{mel}}$	64	96	128	192	256

Table 1: FFT length (N), window length (L), window hop (R) and number of Mel filters (n_{mel}) of the five resolution points employed for the feature extraction. N, L, and R are reported in samples, using a sample rate $f_s = 16000$ Hz.

Considering the trade-off between time and frequency resolution of the Short Time Fourier Transform (STFT), we design a total of 5 resolution points such that they span a range from higher frequency resolution to higher time resolution, relative to the original resolution utilized by the baseline system.

As presented in Table 1, we establish the resolution of the baseline system as the intermediate one (referred to as BS). From this one we define four additional resolution points. Among these, two are designed to double the resolution in frequency (F_{++}) and in time (T_{++}), whereas the remaining two are halfway points between BS and F_{++} (F_{+}) or T_{++} (T_{+}).

Single-resolution models are obtained by training each system with one of the points mentioned above. They can be combined into multi-resolution systems by frame-wise averaging the sequences of scores. As this combination is performed frame-wise, the sequences must have the same length. However, the different time resolutions defined in Table 1 lead to different lengths of the score sequences: $T_1, T_2, ...T_N$. To handle this issue we perform a linear interpolation of the sequences to the maximum length, $T_{max} = \max\{T_1, T_2, ...T_N\}$.

2.3. Class-dependent median filtering

Our multi-resolution approach is based on the fact that each sound event class presents different temporal and spectral characteristics. Therefore, smoothing the decoded predictions employing the same median filter for every class would be counter-productive. Additionally, each PSDS scenario can benefit from different window lengths. Whereas shorter median filters can improve the localization of onsets and offsets, longer windows may be advantageous for avoiding potential cross-triggers and, therefore, enhance the PSDS2.

For this purpose, we have employed a class-dependent median filtering in which the optimal lengths of each class are computed based on one of the PSDS scenarios, iterating over a range from 1 to 29 frames on the DESED Validation set.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

For our experimental results we will use the DESED (Domestic Environment Sound Event Detection) dataset [16], which is the data proposed for the DCASE Task 4A. This dataset contains both real recordings, which are obtained from Google AudioSet [1], and synthetically generated audios employing the Scaper library [17]. The training data is composed of a synthetic strongly-labeled set (10,000

PSDS	DTC	GTC	CTTC	$\alpha_{\mathbf{CT}}$	$\alpha_{\mathbf{ST}}$	$\mathbf{e_{max}}$
Scenario 1	0.7	0.7	0.0	-	1.0	100
Scenario 2	0.1	0.1	0.3	0.5	1.0	100

Table 2: Parameter configuration for the PSDS scenarios.

clips), a real weakly-labeled set (1,578 clips) and a real unlabeled set (14,412 clips).

To select the best model during the training procedure, the synthetic validation set (2,500 clips) together with a 10% of the weakly-labeled set is employed. For testing, we use the validation set, which was constructed to match the clip-per-class distribution of the weakly labeled training set. It consists of 1,168 real audio clips annotated with strong labels.

3.2. Evaluation framework

The Polyphonic Sound Detection Score (PSDS) [3] was proposed for the DCASE Challenge 2021 Task 4 to overcome the limitations of event-based metrics, which rely on the overlap of collars and depend on a unique operating point. For this purpose, they define the Detection Tolerance Criterion (DTC) and the Ground Truth Intersection Criterion (GTC), which measure percentages of intersection between ground-truth labels and detected sound events. Additionally, they introduce the Cross-Trigger Tolerance Criterion (CTTC) to consider data bias by distinguishing the subset of false positives that intersect with labeled events, named as *cross-trigger*.

By modifying the threshold of intersection to these criteria, different properties of a SED system can be evaluated. As it is shown in Table 2, PSDS1 is defined with higher values for the DTC and GTC to measure a high intersection between labels and predictions. Conversely, these values are lower for the PSDS2 but in this case, the CTTC is taken into account to penalize the confusion between classes, whose cost is influenced by α_{CT}.

Results are provided for the recently proposed threshold-independent PSDS [18] over the DESED Validation set. Each model has been trained with three different initializations with the aim of estimating the performance's standard deviation. Moreover, we have compared the complexity of individual systems by calculating the Multiply–Accumulate Operations (MACs) for 10 seconds of audio prediction, a metric that was introduced in this year's evaluation.

3.3. Single-resolution results

The performance of both architectures for the different time-frequency resolution points defined is presented in Table 3. It is clearly seen that CRNN-based systems achieve higher PSDS1 results, evidencing the Conformer's limited temporal precision, which is accentuated when employing features that are not temporally enhanced (F_{++}). However, the Conformer system clearly outperforms the CRNN model in terms of PSDS2. Moreover, our Conformer system exhibits a reduced level of complexity in terms of Multiply-Accumulate operations (MACs). This metric is also influenced by the different resolution points, with lower values observed for frequency enhanced points, as they present shorter input lengths. All Conformer results in Table 3 use FDY, which provides enhanced performance as shown in Table 4.

Figure 1a shows a prototypical example highlighting the advantages and limitations of the different architectures. The CRNN accurately predicts the location of each event but confuses the second one by predicting a *Blender* instead of a *Vacuum cleaner*. This detection is considered a cross-trigger and will downgrade the PSDS2 value. In contrast, the Conformer predicts correctly the presence of both sound events in the clip, but it lacks temporal precision, lowering the PSDS1 results. The effect of the low resolution in time of the Conformer is even more visible in Figure 1b, where the prediction of continuous short events such as *Alarm_bell_ringing* is grouped into a single one.

CRNN	**PSDS1**	**PSDS2**	**MACs**
$\mathbf{F}_{++}$	0.316 ± 0.004	0.561 ± 0.012	**0.891G**
$\mathbf{F}_{+}$	0.347 ± 0.015	**0.583 ± 0.022**	0.905G
BS	0.369 ± 0.006	0.579 ± 0.015	0.930G
$\mathbf{T}_{+}$	0.368 ± 0.039	0.550 ± 0.066	1.772G
$\mathbf{T}_{++}$	**0.374 ± 0.003**	0.575 ± 0.015	1.824G
Conformer	**PSDS1**	**PSDS2**	**MACs**
$\mathbf{F}_{++}$	0.194 ± 0.022	0.688 ± 0.015	**0.588G**
$\mathbf{F}_{+}$	0.224 ± 0.030	**0.696 ± 0.030**	0.633G
BS	0.263 ± 0.020	0.688 ± 0.018	0.879G
$\mathbf{T}_{+}$	0.251 ± 0.019	0.682 ± 0.014	1.147G
$\mathbf{T}_{++}$	**0.349 ± 0.029**	0.668 ± 0.015	1.331G

Table 3: Average and standard deviation results of individual CRNN and Conformer systems trained with different resolution points and initialized with diverse seeds over the DESED Validation set. Independent median filter was applied.

Architecture	PSDS1	PSDS2
CNN + Conformer	0.220 ± 0.027	0.607 ± 0.018
FDY-CNN + Conformer	**0.263 ± 0.020**	**0.688 ± 0.018**

Table 4: Effects of employing FDY for the CNN-based feature extractor over the DESED Validation set.

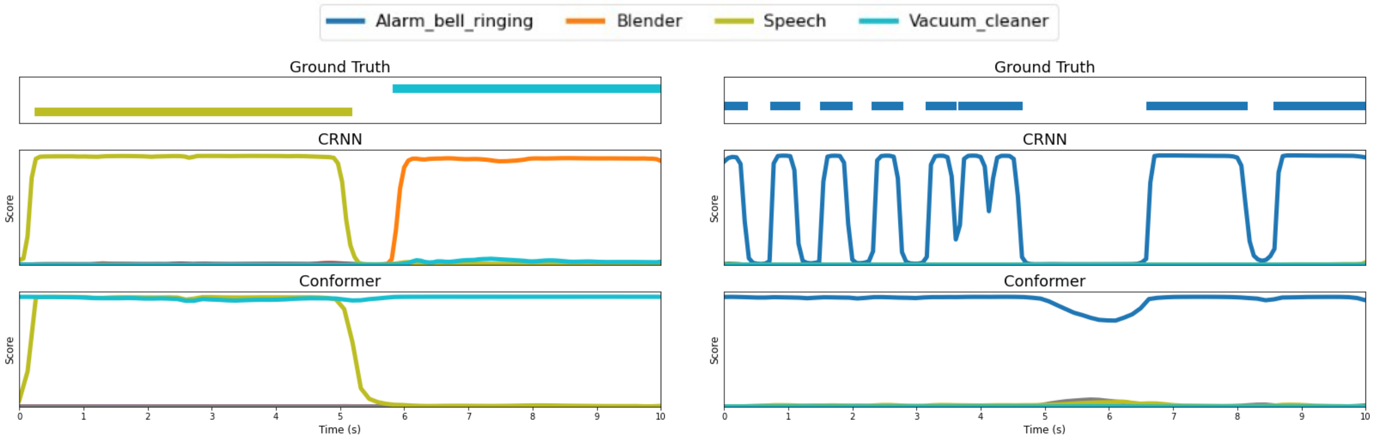
Additionally, results show that each PSDS scenario benefits from a particular resolution point independently of the architecture employed. As expected, PSDS1 benefits from higher temporal resolution, whereas an enhancement in frequency resolution improves the results for PSDS2.

3.4. Multi-resolution results

Single-resolution models are combined following the process described in Section 2.2 in order to obtain multi-resolution systems. In Table 5 the results of six combinations with up to five resolution points are presented individually for CRNNs and Conformers. Multi-resolution not only enhances the performance of single-resolution models, but also evidence that the combination of certain resolution points is more effective for a specific PSDS scenario. For both architectures, the PSDS1 is enhanced when employing a combination of resolutions enhanced in time. Conversely, the PSDS2 benefits from a combination of the five resolution points defined, which is logical as some sound events can be better distinguished by their spectral behaviour while others are better recognized based on their temporal properties.

3.5. Results with task-dependent median filtering

We have experimented with the class-dependent median filtering described in Section 2.3 in our best single-resolution systems (CRNN_T++ and Conformer_F+) and in our two optimal multi-



(a) YG3yk9UXVB9g_20.000_30.000.wav

(b) Y9OqtuFGCCR8_30.000_40.000.wav

Figure 1: Ground truth, CRNN and Conformer predictions for two audio recordings from the DESED Validation set considering the baseline resolution.

		CRNN		**Confomer**	
Resolutions		**PSDS1**	**PSDS2**	**PSDS1**	**PSDS2**
3res	F_{+}, BS, T_{+}	0.397 ± 0.010	0.615 ± 0.012	0.275 ± 0.012	0.719 ± 0.017
3res-F	F_{++}, F_{+}, BS	0.375 ± 0.007	0.617 ± 0.013	0.255 ± 0.015	0.722 ± 0.014
3res-T	BS, T_{+}, T_{++}	0.401 ± 0.007	0.611 ± 0.014	**0.329 ± 0.013**	0.715 ± 0.017
4res-F	F_{++}, F_{+}, BS, T_{+}	0.390 ± 0.007	0.623 ± 0.012	0.268 ± 0.010	0.724 ± 0.015
4res-T	F_{+}, BS, T_{+}, T_{++}	**0.405 ± 0.005**	0.624 ± 0.013	0.309 ± 0.017	0.721 ± 0.016
5res	F_{++}, F_{+}, BS, T_{+}, T_{++}	0.398 ± 0.005	**0.632 ± 0.011**	0.306 ± 0.006	**0.727 ± 0.015**

Table 5: Average and standard deviations results for three initialization seeds of multi-resolution combinations of CRNN and Conformer systems over the DESED Validation set. Fixed median filter was applied.

Obj.	**Model**	**PSDS1**	**PSDS2**
PSDS1	**CRNN_T$_{++}$**	0.387 ± 0.004	0.585 ± 0.012
	CRNN_4res-T	**0.416 ± 0.005**	0.626 ± 0.016
PSDS2	**Conformer_F$_{+}$**	0.164 ± 0.018	0.740 ± 0.033
	Conformer_5res	0.243 ± 0.007	**0.781 ± 0.017**
-	**Baseline**	0.359 ± 0.006	0.562 ± 0.012
	ConformerSED [19]	0.341 ± 0.013	0.576 ± 0.015

Table 6: Effects of employing a class-dependent median filtering on our submitted systems. The Obj. column indicates the objective metric employed to optimize the median filter length of each class. The official baseline and a reproduction of the Miyazaki et al. Conformer system [9] are included for comparison purpose. Results are provided over the DESED Validation set.

resolution systems (CRNN_4res-T and Conformer_5res). Considering that the set of median filters learnt vary depending on which metric is set as objective, we have considered for each system the same PSDS scenario for which it has been designed: PSDS1 for CRNN models and PSDS2 for Conformers.

As we present in Table 6, the systems optimized for PSDS1 improve their results in this metric when the median filters are tuned according the best class-wise PSDS1 performance (from 0.374 to 0.387 in CRNN_T++, and from 0.405 to 0.416 in CRNN_4res-T). Additionally, this criterion is helpful for the PSDS2 as well. When it comes to the systems optimized for the second scenario, their PSDS2 value is also enhanced when the median windows are tuned class-wise (from 0.696 to 0.740 in Conformer_F+, and from 0.727 to 0.781 in Conformer_5res). However, the median filters learnt with this criterion considerably downgrade the performance for PSDS1.

4. CONCLUSIONS

In this paper we presented the benefits of the Conformer architecture for sound event detection by optimizing a system towards the second scenario proposed for the DCASE Challenge 2023 Task 4A. Among the submitted systems without employing external data, our Conformer system achieves one of the best PSDS2 values over the evaluation set (0.729).

Following our previous multi-resolution approach, we were able to analyse its behaviour over different time-frequency resolutions and compare its performance with a CRNN-based system. Additionally, by employing this technique we not only demonstrate that a multi-resolution ensemble can considerably enhance the results, but also revealed that the different PSDS scenarios benefit from features that enhance either time or frequency resolution. Therefore, we obtain the best PSDS1 when combining CRNN systems trained with resolution points enhanced in time, while our best PSDS2 is obtained when combining the five resolutions defined for the Conformer.

5. REFERENCES

[1] J. F. Gemmeke, D. P. W. Ellis, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017.

[2] http://dcase.community.

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

[4] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[6] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[7] J. Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, pp. –, 2022. [Online]. Available: http://dx.doi.org/10.1561/116.00000050

[8] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. yi Lee, and H. M. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *Interspeech*, 2022.

[9] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 100–104.

[10] D. de Benito-Gorron, S. Barahona, S. Segovia, D. Ramos, and T. Doroteo, "Multi-resolution combination of CRNN and conformers for dcase 2022 task 4," DCASE2022 Challenge, Tech. Rep., June 2022.

[11] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," 2019.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[13] 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.

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

[15] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[16] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 2020. [Online]. Available: https://hal.inria.fr/hal-02355573

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

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

[19] K. Miyazaki, "ConformerSED," https://github.com/mkoichi/ConformerSED, 2021.