OPTIMIZING MULTI-RESOLUTION CONFORMER AND CRNN MODELS FOR DIFFERENT PSDS SCENARIOS IN DCASE CHALLENGE 2023 TASK 4A

Technical Report

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

In this technical report we describe our submission to DCASE 2023 Task 4A: Sound Event Detection with Weak Labels and Synthetic Soundscapes. Considering that the different scenarios proposed for the Polyphonic Sound Event Score (PSDS) highlight diverse properties of a Sound Event Detection (SED) system, we have employed two different architectures for optimizing each scenario. Whereas we exploit the temporal benefits of Convolution Recurrent Neural Networks (CRNNs) for maximizing the PSDS1, we employ a Conformer network for improving sound events classification and therefore enhancing PSDS2. Additionally, we follow the multi-resolution approach successfully employed in previous DCASE editions to take advantage of the temporal and spectral disparities among the different sound event categories.

***Index Terms*—** DCASE 2023, Conformer, CRNN, Mean Teacher, Multi-resolution, Model fusion

1. INTRODUCTION

This technical report describes our submission to DCASE 2023 Task 4A, whose goal is the detection and classification of 10 different sound event categories. In order to evaluate the systems' performance, two scenarios are defined for the Polyphonic Sound Event Detection Score (PSDS) [1]. Whereas the first one is focused on a fast reaction upon sound events detection, the second one is defined in order to penalize the confusion between classes. Therefore, taking into account the diverse objectives of each scenario we decide to implement two different systems to improve each PSDS separately:

- For the scenario 1 we employ a convolutional recurrent neural network (CRNN) to exploit the temporal variations of each category with the aim of improving the localisation of events.
- For the scenario 2 we employ a Conformer (convolutional-augmented transformer) [2] which shows a better performance when dealing with sound events classification.

As in previous editions we followed a multi-resolution approach [3, 4, 5], which shows that combining different time-frequency configurations, also known as *resolution points*, during the feature extraction process can enhance the performance of a SED system due to the variability among the diverse sound classes.

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
$\mathbf{n_{mel}}$	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.

Besides, we further improve the results by employing a class-wise median filter for post-processing.

2. DATASET

The data proposed for the DCASE Task 4A is the DESED (Domestic Environment Sound Event Detection) dataset [6]. This dataset contains both real recordings, which are obtained from Google AudioSet [7], and synthetically generated audios employing the Scaper library [8]. The training data is composed of a synthetic strongly-labeled set (10000 clips), a real weakly-labeled set (1578 clips) and a real unlabeled set (14412 clips).

For selecting the best model during the training procedure, the synthetic validation set (2500 clips) together with a 10% of the weakly-labeled set is employed. For testing, we employ the validation set, which was constructed to match the clip-per-class distribution of the weakly labeled training set. It is composed of 1168 real audio clips annotated with strong labels.

3. PROPOSED SOLUTIONS

3.1. Multi-resolution analysis

As in previous editions, we follow a multi-resolution approach which consist on varying the parameters employed for the extraction of mel-spectrogram features. 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_{+}).

3.2. Convolutional-Recurrent Neural Networks

Our CRNN models follow the structure and configuration of the DCASE Task 4A baseline system [9]. Mean teacher [10] is used for semi-supervised learning. In contrast with the baseline system, we perform model selection monitoring the teacher model instead of the student model. This is motivated by the observation that the teacher models usually exhibit superior performance in both validation and test set.

3.3. Conformer Networks

The Conformer system is based on the 2020 DCASE Task 4 winner [11]. The audios employed for training this network have been normalized to -3dbFS and a high pass filter of 10 Hz has been applied to remove the continuous component.

We perform a hyperparameter tuning with the objective of enhancing the PSDS2 value, leading to an optimal configuration of 7 Conformer blocks with 4 attention heads each one and a encoder dimension of 144.

Additionally, we substitute the CNN-based feature extractor for a Frequency Dynamic Convolution [12] to improve the classification of non-stationary sound events. For this 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 confusions between classes. We employ both Mixup and FilterAugment [13] with a probability of 50% of applying them to the training data. Mean teacher is used for semi-supervised learning and as well as with our CRNN models, the best model is selected by monitoring the teacher model. However, for the Conformer models the objective metric employed for model selection is PSDS2.

3.4. Model combination

Single resolution models can be combined by frame-wise averaging the sequence of scores obtained after training them separately. For a given input, the models output a different score sequence for each class by means of a sigmoid layer, thus the scores are bound between 0 and 1. Therefore, the combination of N resolution points for event class k and time frame t can be defined as follows:

$$s_{k,t}^{(comb)} = \frac{1}{N}\sum_{n=1}^{N} s_{k,t}^{(n)} \quad (1)$$

As this combination is performed frame-wise, the sequences $s^{(1)}, ..., s^{(N)}$ must have the same length. However, the different time resolutions defined in Table 1 lead to different lengths of the score sequences: $T_1, T_2, ...T_N$. For handling this issue we perform a linear interpolation of the sequences to the maximum length, $T_{max} = \max\{T_1, T_2, ...T_N\}$

CRNN	**PSDS1**	**PSDS2**	**Ev-F$_1$** (%)
$\mathbf{F}_{++}$	0.316 ± 0.004	0.561 ± 0.012	37.90 ± 0.60
$\mathbf{F}_{+}$	0.347 ± 0.015	**0.583 ± 0.022**	41.84 ± 1.56
BS	0.369 ± 0.006	0.579 ± 0.015	**43.18 ± 0.56**
$\mathbf{T}_{+}$	0.368 ± 0.039	0.550 ± 0.066	42.42 ± 2.49
$\mathbf{T}_{++}$	**0.374 ± 0.003**	0.575 ± 0.015	42.86 ± 0.15
Conformer	**PSDS1**	**PSDS2**	**Ev-F$_1$** (%)
$\mathbf{F}_{++}$	0.194 ± 0.022	0.688 ± 0.015	21.03 ± 1.60
$\mathbf{F}_{+}$	0.224 ± 0.030	**0.696 ± 0.030**	22.72 ± 0.82
BS	0.263 ± 0.020	0.688 ± 0.018	26.16 ± 0.95
$\mathbf{T}_{+}$	0.251 ± 0.019	0.682 ± 0.014	25.78 ± 1.59
$\mathbf{T}_{++}$	**0.349 ± 0.029**	0.668 ± 0.015	**34.30 ± 1.29**

Table 2: 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.

3.5. Class-dependent median filtering

The scores obtained as the output of a system require a decoding process to obtain the onsets and offsets of each sound event prediction. In a first phase, thresholding is employed for obtaining binary sequences. Then, a common procedure for smoothing the predictions is employing a median filter. This can be performed employing a fixed value for the length of the filter or setting a specific length to each class considering its properties. To evaluate the benefits of each post-processing technique, we will apply both of them to our final systems.

When it comes to the fixed post-processing, we employ a median filter length of 450 ms for our CRNN models, whose value in frames will vary depending on the resolution point. However, as the Conformer models output shorter sequences, we employ for all the resolution points a fixed value of 7 frames.

In order to further optimize the objective metric of each system, we have developed 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 over the DESED Validation set. Considering that the first scenario benefits from the precise detection of sound events, shorter median filter windows will improve the localisation. Conversely, the PSDS2 imposes penalties for class confusion, and thus, longer median filters may be advantageous for avoiding potential cross-triggers.

4. RESULTS

Results are provided for the recently proposed threshold-independent PSDS [14] over the 1168 audio clips that compose the DESED Validation set. Each model has been trained with three different initializations with the aim of estimating the performance's standard deviation. Additionally, the systems are evaluated with the event-based F1-score, which is computed using `psds-eval` 0.5.0.

4.1. Single-resolution results

As a first step, we train both the CRNN and the Conformer models with the parameters presented in Table 1, obtaining for each one a total of 5 single resolution systems. In Table 2 we present the results obtained for each system employing the fixed post-processing mentioned in Section 3.5.

CRNN	Resolutions	PSDS1	PSDS2	Ev-$\mathbf{F_1}$ (%)
3res	F_+, BS, T_+	0.397 ± 0.010	0.615 ± 0.012	45.47 ± 0.25
3res-F	F_{++}, F_+, BS	0.375 ± 0.007	0.617 ± 0.013	44.68 ± 1.29
3res-T	BS, T_+, T_{++}	0.401 ± 0.007	0.611 ± 0.014	45.80 ± 1.04
4res-F	F_{++}, F_+, BS, T_+	0.390 ± 0.007	0.623 ± 0.012	45.53 ± 1.76
4res-T	F_+, BS, T_+, T_{++}	$\mathbf{0.405 \pm 0.005}$	0.624 ± 0.013	$\mathbf{46.01} \pm 1.00$
5res	F_{++}, F_+, BS, T_+, T_{++}	0.398 ± 0.005	$\mathbf{0.632 \pm 0.011}$	45.87 ± 1.24
Conformer	**Resolutions**	**PSDS1**	**PSDS2**	**Ev-$\mathbf{F_1}$ (%)**
3res	F_+, BS, T_+	0.275 ± 0.012	0.719 ± 0.017	26.27 ± 0.94
3res-F	F_{++}, F_+, BS	0.255 ± 0.015	0.722 ± 0.014	24.58 ± 1.13
3res-T	BS, T_+, T_{++}	$\mathbf{0.329 \pm 0.013}$	0.715 ± 0.017	$\mathbf{30.46 \pm 0.09}$
4res-F	F_{++}, F_+, BS, T_+	0.268 ± 0.010	0.724 ± 0.015	25.85 ± 0.09
4res-T	F_+, BS, T_+, T_{++}	0.309 ± 0.017	0.721 ± 0.016	29.84 ± 1.77
5res	F_{++}, F_+, BS, T_+, T_{++}	0.306 ± 0.006	$\mathbf{0.727 \pm 0.015}$	28.11 ± 0.23

Table 3: Average and standard deviations results for multiple initialization seeds of multi-resolution combinations of CRNN and Conformer systems over the DESED Validation set. Independent median filter was applied.

Objective	Model	PSDS1	PSDS2	Ev-$\mathbf{F_1}$ (%)
PSDS1	**CRNN_$\mathbf{T_{++}}$**	0.387 ± 0.004	0.585 ± 0.012	44.00 ± 0.05
	CRNN_4res-T	$\mathbf{0.416 \pm 0.005}$	0.626 ± 0.016	$\mathbf{47.02 \pm 0.95}$
PSDS2	**Conformer_$\mathbf{F_+}$**	0.164 ± 0.018	0.740 ± 0.033	22.01 ± 0.82
	Conformer_5res	0.243 ± 0.007	$\mathbf{0.781 \pm 0.017}$	25.22 ± 0.54

Table 4: Average and standard deviations results for multiple initialization seeds of our submitted systems over the DESED validation set employing class-dependent median filtering. The Objective column indicates the objective metric employed for optimizing the median filter length of each class.

4.2. Multi-resolution results

Single-resolution models are combined following the process described in Section 3.4 in order to obtain multi-resolution systems. In Table 3 the results of six combinations up to five resolution points are presented individually for CRNNs and Conformers. For both architectures, the results obtained combining different resolutions show an enhancement with respect to the ones obtained employing a unique resolution point.

Additionally, we combine different resolution points of both architectures. However, this doesn't seem to benefit any of the two scenarios proposed for the PSDS. Therefore, we don't include these systems and their results in the technical report.

4.3. Results with task-dependent median filtering

We have experimented with the class-dependent median filtering described in Section 3.5 in our submitted models. In light of the requirement of submitting a non-ensemble system we have chosen the two optimal single-resolution systems (CRNN_T++ and Conformer_F+) and the two optimal multi-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 Conformer ones.

The systems optimized for PSDS1 improve their results in this metric when the median filter 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 and the F_1-based performance as well.

When it comes to the PSDS2, the systems optimized for this scenario improve this metric when the median windows are tuned class-wise (from 0.696 to 0.740 in CRNN_T++, and from 0.727 to 0.781 in CRNN_4res-T). However, the median filters learnt with this criterion do not improve the other metrics.

5. CONCLUSIONS

This technical report describes our submission for DCASE 2023 Task 4A. This year we have optimized each PSDS scenario separately, employing a CRNN for optimizing PSDS1 and a Confomer system for PSDS2. We have employed mean teacher for semi-supervised learning but selecting the best model monitoring the teacher network. Besides, in the case of the Conformer systems, this model selection is applied over the PSDS2.

Following our previous multi-resolution approach, we have trained SED systems employing different resolution settings for feature extraction. We then compute their fusion by averaging frame-wise their score sequences. While the PSDS1 obtains the best results when combining CRNN systems trained with resolution points enhanced in time, the PSDS2 achieves its highest combining all the resolutions defined for the Conformer.

Furthermore, we implement a class-wise median filtering to further improve the results. By searching for each class the median filter length that maximizes either the PSDS1 or the PSDS2 scenario, we obtained an enhancement in the performance of our submitted systems.

6. REFERENCES

[1] 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.

[2] 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.

[3] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano, "A multi-resolution approach to sound event detection in dcase 2020 task4," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 36–40.

[4] D. de Benito-Gorrón, S. Segovia, D. Ramos, and D. T. Toledano, "Multiple feature resolutions for different polyphonic sound detection score scenarios in dcase 2021 task 4," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 65–69.

[5] 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.

[6] 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

[7] 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.

[8] 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.

[9] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 200–204.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.