MULTI-RESOLUTION COMBINATION OF CRNN AND CONFORMERS FOR DCASE 2022 TASK 4

Technical Report

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

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

ABSTRACT

This technical report describes our submission to DCASE 2022 Task 4: Sound event detection in domestic environments. We follow a multi-resolution approach consisting on a late fusion of systems that are trained with different feature extraction parameters, aiming to leverage the characteristics of different event categories in time and frequency. Our systems are built upon the Convolutional-Recurrent Neural Network (CRNN) proposed by the baseline system and the Conformer structure proposed by the winners of the 2020 challenge.

Index Terms— DCASE 2022, CRNN, Conformer, Mean Teacher, Multi-resolution, Model fusion

1. INTRODUCTION

The aim of DCASE Task 4 is the detection and classification of 10 different sound event categories. These categories describe sounds commonly found in domestic environments, and each category presents different temporal and spectral properties. In previous editions, we had already explored the idea of employing multiple time-frequency resolution points during the feature extraction process, aiming to exploit these differences, and finding that the combination of different time-frequency resolutions is beneficial for SED performance in terms of F_1 score [1, 2] and different scenarios of the Polyphonic Sound Detection Score (PSDS) [3].

This paper describes our submission to DCASE 2022 Task 4, which is based on the same multi-resolution approach. In addition to the multi-resolution CRNN proposed in previous challenges [3, 4], this year we extend the multi-resolution approach to Conformer networks [5]. Moreover, a different model selection strategy is applied to the CRNN mean teacher networks, monitoring the performance of the teacher model, instead of the student model.

2. DATASET

The dataset used in DCASE 2022 Task 4 is DESED (Domestic Environment Sound Event Detection) [6, 7]. DESED is composed of real recordings, obtained from Google AudioSet [8], and synthetic recordings which are generated using the Scaper library [9]. Real recordings include the Weakly-labeled training set (1578 clips), the Unlabeled training set (14412 clips) and the Validation set (1168 clips). Additionally, the Synthetic set contains 10000 strongly-labeled, synthetic clips, generated such that the event distribution is similar to that of the Validation set.

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 proposed resolution points for CRNN. N, L, and R are reported in samples, using a sample rate $f_s = 16000$ Hz.

3. PROPOSED SOLUTIONS

3.1. Multi-resolution analysis

Our approach to multi-resolution consists on the definition of several sets of parameters for the feature extraction process, each one of them determining a resolution point. The extraction of mel-spectrogram features, based in the Fast Fourier Transform (FFT), implies a compromise between time resolution and frequency resolution. Taking this fact into account, we design the resolution points for a given system so that they cover a range from higher frequency resolution to higher time resolution, with respect to the original resolution used by the system.

For each system, we consider its original resolution (which we call baseline resolution, or BS) and define four additional resolution points: two of them are designed to obtain double resolution in frequency (F_{++}) or in time (T_{++}), and the other two are the intermediate points between BS and F_{++} (F_{+}) or T_{++} (T_{+}). In every case, Hamming windows are used, and audio signals are sampled at $f_s = 16000$ Hz.

Work developed under project DSForSec (RTI2018-098091-B-I00), funded by the Ministry of Science, Innovation and Universities of Spain and FEDER

Resolution	$\mathbf{T}_{++}$	$\mathbf{T}_{+}$	**BS**	$\mathbf{F}_{+}$	$\mathbf{F}_{++}$
$\mathbf{N}$	512	1024	1024	2048	2048
$\mathbf{L}$	512	768	1024	1536	2048
$\mathbf{R}$	160	244	323	488	646
$\mathbf{n_{mel}}$	32	48	64	96	128

Table 2: FFT length (N), window length (L), window hop (R) and number of Mel filters (n_{mel}) of the five proposed resolution points for Conformer networks. N, L, and R are reported in samples, using a sample rate $f_s = 16000$ Hz.

3.2. Convolutional-Recurrent Neural Networks

Our CRNN models follow the structure and configuration of the DCASE Task 4 baseline system [10]. The data distribution proposed by the baseline system uses the Weakly-labeled, Unlabeled and Synthetic sets to train the neural networks, reserving 10% of the Weakly-labeled set, together with 2500 additional synthetic clips, as a validation set for model selection.

Mean teacher [11] is used for semi-supervised learning. In contrast with the baseline system, we performed model selection monitoring the teacher model, instead of the student model. This is motivated by the observation that the teacher models usually present better performance in validation and test.

The parameters of the five resolution points for the CRNN models can be found in Table 1.

3.3. Conformer Networks

Our Conformer models follow the setup of the best submission of the DCASE 2020 Task 4 challenge [12]. Such setup uses the Weakly-labeled, Unlabeled and Synthetic sets for training, whereas the DESED Validation set is used for model selection. Mean teacher is used for semi-supervised learning, with no change to the original system settings.

The parameters of the five resolution points for the Conformer models can be found in Table 2.

3.4. Model combination

In order to use the information provided by the different resolutions, a simple model combination method is proposed, based on training individual models with each of the resolution points, and then frame-wise averaging the sequences of scores obtained with different resolutions. 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, for event class k and time frame t:

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

Usually, different time resolutions in the features lead to different lengths of the score sequences: $T_1, T_2, ...T_N$. In order to compute the frame-wise average, the sequences $s^{(1)}, ..., s^{(N)}$ must have the same length. We handle this by linearly interpolating the sequences to the maximum length, $T_{max} = \max\{T_1, T_2, ...T_N\}$.

3.5. Task-dependent median filtering

Once the combined score sequences $s^{(comb)}$ are obtained, a decoding process is required in order to determine the predicted onsets and offsets of the events. In the first place, this process involves thresholding the scores to obtain binary sequences. After thresholding, it is a common practice to smooth the resulting sequences by means of a median filter.

CRNN	**PSDS1**	**PSDS2**	**Ev-F$_1$** (%)	**Int-F$_1$** (%)
$\mathbf{F}_{++}$	0.2887	0.5720	34.85	65.17
$\mathbf{F}_{+}$	0.3411	0.5566	43.11	66.68
BS	0.3696	0.5706	43.45	66.07
$\mathbf{T}_{+}$	0.3811	0.5633	**43.65**	**66.90**
$\mathbf{T}_{++}$	**0.3820**	**0.5743**	42.86	66.15
Conformer	**PSDS1**	**PSDS2**	**Ev-F$_1$** (%)	**Int-F$_1$** (%)
$\mathbf{F}_{++}$	0.2939	0.5362	40.09	65.05
$\mathbf{F}_{+}$	0.2783	0.5606	35.41	63.96
BS	0.3418	0.5802	**41.86**	**65.27**
$\mathbf{T}_{+}$	0.3356	0.5805	39.40	65.20
$\mathbf{T}_{++}$	**0.3492**	**0.5816**	39.51	63.66

Table 3: Results of individual CRNN and Conformer systems trained with different resolution points over the DESED Validation set.

Although by default we have used the fixed median filter length (450ms) proposed by the baseline system, we have also computed the optimal length of the median filter window for each class and for each PSDS scenario (PSDS1 and PSDS2), considering a range from 220ms to 1.5s.

Whereas the precise detection of event classes that tend to present shorter durations needs shorter median filter windows, longer events benefit from stronger smoothing with longer windows. In an analogous way, the PSDS1 scenario can benefit from shorter median filters, given that it aims for precise detections in time. In contrast, PSDS2 is not so strict about time boundaries, therefore it can benefit from longer median filters. The optimal values are searched over the DESED Validation set.

3.6. PSDS without class-instability penalty

Arguing that it might be desirable for a system to hold similar performances for each class, PSDS introduces the parameter α_{ST} (cost of instability across classes) [13]. This parameter weights the penalty of the class-wise performance variability in the final PSDS.

In both PSDS1 and PSDS2, the instability cost is $\alpha_{ST} = 1$. Therefore, a class-dependent optimization (e.g. class-wise median filtering) could result in a lower global PSDS even when improving the performance of every class, if the variance of performance across classes increases.

Considering this situation, we propose a version of each PSDS scenario with $\alpha_{ST} = 0$. On the experimental side, these scenarios allow to measure the impact of class instability in the final score, whereas in practice, they could be useful for an application in which the stability across classes is not a relevant factor. Moreover, the penalty due to class-instability can be computed directly as the difference between the PSDS score with $\alpha_{ST} = 0$ and with $\alpha_{ST} = 1$:

$$\text{cost}_{ST} = \text{PSDS}(\alpha_{ST} = 0) - \text{PSDS}(\alpha_{ST} = 1) \tag{2}$$

4. RESULTS

The provided results consider 1153 clips of the DESED Validation set, leaving out the ones that do not have annotations, as recently

CRNN	**Resolutions**	**PSDS1**	**PSDS2**	**Ev-F_1** (%)	**Int-F_1** (%)
3res	F_{+}, BS, T_{+}	0.3979	0.6063	45.81	69.59
3res-F	F_{++}, F_{+}, BS	0.3729	0.6030	45.31	68.68
3res-T	BS, T_{+}, T_{++}	0.4164	0.6131	47.47	70.05
4res-F	F_{++}, F_{+}, BS, T_{+}	0.3930	0.6130	47.43	70.22
4res-T	F_{+}, BS, T_{+}, T_{++}	**0.4135**	0.6190	**47.81**	70.28
5res	F_{++}, F_{+}, BS, T_{+}, T_{++}	0.4022	**0.6250**	47.50	**71.18**
Conformer	**Resolutions**	**PSDS1**	**PSDS2**	**Ev-F_1** (%)	**Int-F_1** (%)
3res	F_{+}, BS, T_{+}	0.3460	0.6357	42.58	68.57
3res-F	F_{++}, F_{+}, BS	0.3494	0.6307	43.68	68.91
3res-T	BS, T_{+}, T_{++}	**0.3708**	0.6330	42.77	68.37
4res-F	F_{++}, F_{+}, BS, T_{+}	0.3468	0.6452	43.46	68.53
4res-T	F_{+}, BS, T_{+}, T_{++}	0.3696	0.6470	43.36	68.49
5res	F_{++}, F_{+}, BS, T_{+}, T_{++}	0.3664	**0.6565**	**44.31**	**69.65**
CRNN+Conformer	**Resolutions**	**PSDS1**	**PSDS2**	**Ev-F_1** (%)	**Int-F_1** (%)
7res	CRNN 4res-T, Conformer 3res-T	**0.4218**	0.6559	**49.23**	**72.18**
10res	CRNN 5res, Conformer 5res	0.4101	**0.6652**	48.94	72.00

Table 4: Results of multi-resolution combinations of CRNN and Conformer systems over the DESED Validation set.

System	**Objective**	**PSDS1**	**PSDS2**	**Ev-F_1** (%)	**Int-F_1** (%)
7res	None	0.4218	0.6559	49.23	72.18
	PSDS1	**0.4279**	0.6554	**50.12**	**72.85**
	PSDS2	0.3962	0.6640	45.79	71.95
10res	None	0.4101	**0.6652**	48.94	72.00
	PSDS1	0.4172	0.6626	49.34	72.34
	PSDS2	0.3473	0.6633	43.00	69.68

Table 5: Results of multi-resolution CRNN and Conformer systems over the DESED Validation set using task-dependent median filtering. The Objective column indicates whether the objective criterion for the median filter length of each class is PSDS1, PSDS2, or none (fixed median filtering).

	PSDS1		**PSDS2**	
Class	**7res**	**7res-m**	**10res**	**10res-m**
Alarm/bell/ringing	0.5723	0.5783	0.8817	0.8825
Blender	0.7765	0.7810	0.8853	0.9018
Cat	0.5124	0.5086	0.7687	0.7875
Dishes	0.2409	0.2487	0.5308	0.5355
Dog	0.3593	0.3686	0.7586	0.7686
Electric shaver/tooth.	0.8186	0.8204	0.9678	0.9725
Frying	0.7243	0.7421	0.9016	0.9142
Running water	0.6110	0.6141	0.8135	0.8251
Speech	0.6930	0.6954	0.8947	0.9059
Vacuum cleaner	0.8839	0.8853	0.9443	0.9498
Global, $\alpha_{\mathbf{ST}} = 1$	0.4218	**0.4279**	**0.6652**	0.6633
Global, $\alpha_{\mathbf{ST}} = 0$	0.6192	**0.6243**	0.7972	**0.8040**
cost$_{\mathbf{ST}}$	0.1974	0.1964	0.1320	0.1407

Table 6: Class-wise PSDS results of our submitted systems over the DESED Validation set. PSDS1 is computed with the 7res combination, and PSDS2 with the 10res combination. Task-dependent median filtering is applied in 7res-m and 10res-m, taking PSDS1 and PSDS2, respectively, as objective. In addition to the default PSDS settings ($\alpha_{ST} = 1$), PSDS results with $\alpha_{ST} = 0$ are provided, as well as the class-instability penalties (cost$_{ST}$).

Class	7res-m	10res-m
Alarm/bell/ringing	0.35	0.54
Blender	1.25	1.50
Cat	0.54	1.00
Dishes	0.35	1.50
Dog	0.22	1.44
Electric shaver/tooth.	1.50	1.38
Frying	1.50	1.50
Running water	1.18	1.06
Speech	0.61	1.00
Vacuum cleaner	1.50	0.86

Table 7: Best median filter lengths, in seconds, obtained for each class over the DESED Validation set. In 7res-m, the lengths are optimized for PSDS1, whereas in 10res-m, the lengths are optimized for PSDS2. A range from 0.22s to 1.50s is considered.

suggested by the organization. All of them have been computed using `psds-eval` 0.5.0.

4.1. Single-resolution results

In the first place, we trained a total of 10 single resolution systems, 5 CRNNs and 5 Conformers, each one using a different resolution point for feature extraction. Their results are presented in Table 3. It is to be noted that each resolution point is obtained with different parameters in CRNN and Conformer, as shown in Tables 1 and 2.

4.2. Multi-resolution results

We have defined several combinations of the single-resolution systems, considering only CRNNs, only Conformers, or both. The combinations are performed following the process described in Section 3.4, and their results are presented in Table 4. The median filter length considered is fixed to 450ms.

Six different combinations of up to five resolution points are evaluated for CRNNs and for Conformers. Afterwards, two additional combinations are proposed, joining the best CRNN and Conformer combinations for either PSDS1 or PSDS2. These combinations are formed by 7 resolutions (7res) and 10 resolutions (10res), respectively.

4.3. Results with task-dependent median filtering and custom PSDS scenarios

Using the best combined systems obtained for PSDS1 (7res) and PSDS2 (10res), we have applied the task-dependent median filtering described in Section 3.5. For a given system, a different set of median filter lengths is learnt when setting either PSDS1 or PSDS2 as the objective criterion.

The PSDS1 performance improves when the median filters are tuned according to best class-wise PSDS1 performance (from 0.4218 to 0.4279 in 7res, and from 0.4101 to 0.4172 in 10res). Additionally, this criterion is helpful for F_1-based performance as well.

When tuning the median filters to maximize class-wise PSDS2, the PSDS2 performance in 7res improves (from 0.6559 to 0.6640). However, for the 10res system, the obtained PSDS2 is lower (from 0.6652 to 0.6633). The median filters learnt with this criterion are noticeably less useful for PSDS1 or F_1-based metrics.

Our submission includes the 7res and 10res systems with fixed (450ms) median filtering, the 7res system with task-dependent median filtering with PSDS1 objective (7res-m), and the 10res system with task-dependent median filtering with PSDS2 objective (10res-m). In this way, we have two systems optimized for the PSDS1 scenario (7res and 7res-m), and two optimized for PSDS2 (10res and 10res-m).

Considering these four systems, we have studied the class-wise performance and the PSDS performance without the class-instability penalty, described in Section 3.6. The results, presented in Table 6, assert that task-dependent median filtering reaches a better class-wise performance for each individual, but achieves a lower global PSDS due to the instability penalty. In contrast, if class-instability is not taken into account ($\alpha_{ST} = 0$), the task-dependent median filtering holds better results than the fixed median filtering. The median filters used by 7res-m and 10res-m are described in Table 7.

5. CONCLUSIONS

In this technical report, our submission for DCASE 2022 Task 4 is described. Following the multi-resolution approach that we had used in previous editions, we have trained SED systems using different resolution settings for feature extraction, and then we have computed their combinations as a frame-wise average of their score sequences.

In addition to our previous participations, this year we have also applied multi-resolution to Conformer networks. We maintain the mean teacher CRNNs, but using a different model selection strategy, in which the teacher model, instead of the student, is monitored for model selection. The best results are obtained when combining together both CRNN and Conformer systems.

Moreover, we have applied a task-dependent class-wise median filtering, searching for each class the median filter length that maximizes either the PSDS1 or the PSDS2 scenario. This process allows to improve the class-wise performances of a system in a particular PSDS setting.

Furthermore, we propose two custom PSDS scenarios, obtained by setting the cost of instability across classes (α_{ST}) to zero in PSDS1 and PSDS2. These settings have provided an interpretation of the impact of class-wise performance optimization on the final PSDS score.

6. REFERENCES

[1] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano, “A multi-resolution CRNN-based approach for semi-supervised Sound Event Detection in DCASE 2020 Challenge,” *IEEE Access*, 2021 (early access).

[2] ——, “An analysis of sound event detection under acoustic degradation using multi-resolution systems,” *Applied Sciences*, vol. 11, no. 23, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/23/11561

[3] 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.

[4] 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.

[5] 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.

[6] 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

[7] 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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.