MULTI-RESOLUTION MEAN TEACHER FOR DCASE 2021 TASK 4

Technical Report

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
diego.benito@uam.es, sergio.segoviag@estudiante.uam.es,
daniel.ramos@uam.es, doroteo.torre@uam.es

ABSTRACT

This technical report describes our participation in DCASE 2021 Task 4: Sound event detection and separation in domestic environments. Aiming to take advantage of the different lengths and spectral characteristics of each target category, we follow the multi-resolution feature extraction approach that we proposed for last year's edition. It is found that each one of the proposed Polyphonic Sound Detection Score (PSDS) scenarios benefits from either a higher temporal resolution or a higher frequency resolution. Furthermore, combining several time-frequency resolutions via model fusion is able to improve the PSDS results in both scenarios.

Index Terms— DCASE 2021, CRNN, Mean Teacher, Multi-resolution, Model fusion, PSDS

1. INTRODUCTION

This paper describes our submission to DCASE 2021 Task 4. Our participation is based on the provided baseline system and follows the scenario of sound event detection (SED) without source separation pre-processing. We propose a multi-resolution analysis of the audio features (mel-spectrograms) used to train the neural network, in contrast with the single-resolution approach of the baseline.

DCASE Task 4 consists in the detection and classification of 10 different sound events. These sound events belong to domestic environments, and each category shows its own temporal and spectral properties. During last year challenge, we 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 was beneficial for the performance of the SED baseline system, in terms of both event-based F_1 score and Polyphonic Sound Detection Score (PSDS) [1, 2, 3].

One of the advantages of our multi-resolution approach is that it is, in principle, complementary to other improvements in the model, such as a different topology of the neural network or additional training data. For that reason, we have applied it to this year's SED baseline system, which features the use of mixup [4] for data augmentation, as well as a larger synthetic subset, as main additions to the Mean Teacher [5] convolutional recurrent neural network (CRNN) system of previous years [6].

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

2. DATASET

The dataset used for sound event detection in DCASE 2021 Task 4 is DESED (Domestic Environment Sound Event Detection) [7, 8]. DESED is composed of real recordings, obtained from Google AudioSet, 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 12500 strongly-labeled, synthetic clips, generated such that the event distribution is similar to that of the Validation set.

The Weakly-labeled, Unlabeled and Synthetic sets are used to train the neural networks. 10% of the Weakly-labeled set and 20% of the Synthetic set are reserved for validation. The DESED Validation set is used to tune hyper-parameters and perform model selection.

3. PROPOSED SOLUTIONS

3.1. Multi-resolution analysis

The baseline system employs mel-spectrogram features, a two-dimensional representation of audio signals based on the Fast Fourier Transform (FFT) and the Mel scale. Thus, the audio segments are transformed into 2-D images that are processed through the CRNN. The process of mel-spectrogram extraction depends on several parameters: the sampling frequency of the audio (f_s), the number of points of the FFT (N), the number of mel filters (n_{mel}), the analysis window function, and its hop and length (R, L). Given a set of values for these parameters, a time-frequency resolution working point is defined.

A particular time-frequency resolution can be more or less fitted to detect a sound event category depending on its temporal and spectral characteristics, which vary for each target class. For example, it is particularly easy to show that the different event classes have different lengths by analyzing the mean and standard deviation of the duration of the ten categories in the Synthetic training set, as presented in Table 1.

Using different mel-spectrogram configurations, we defined five different time-frequency resolution working points. For each one of them, we replicated the baseline, modifying it to handle the corresponding time-frequency resolution. Finally, we combined the frame-level estimation of the class posterior probabilities provided by each resolution, obtaining a multi-resolution system.

	N.	Mean	Std.
Alarm_bell_ringing	1886	1.42	1.97
Blender	1062	4.39	3.80
Cat	1910	1.34	1.74
Dishes	4353	0.68	0.59
Dog	2320	1.16	1.19
Electric_shaver_toothbrush	1074	8.74	2.24
Frying	1395	9.38	1.37
Running_water	1206	6.58	2.99
Speech	15967	1.53	1.18
Vacuum_cleaner	1045	9.35	1.78

Table 1: Number of examples and mean and standard deviation of their durations (in seconds) for each sound category in the Synthetic training 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
$\mathbf{n_{mel}}$	64	96	128	192	256

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

The reference for time-frequency resolution is the set of parameters used by the baseline system for the feature extraction process, which will be referred as BS. We maintain the sampling frequency at $f_s = 16000$ Hz and the use of a Hamming window. The rest of the parameters (N, L, R, n_{mel}) are modified to increase time or frequency resolution in each case. The resulting resolution points (T_{++}, T_{+}, BS, F_{+}, F_{++}) are described in Table 2.

3.2. Model fusion

For a given event category i, a binary classification is performed between classes $\{\theta_{i,0}; \theta_{i,1}\}$, where $\theta_{i,0}$ means "event i not detected" and $\theta_{i,1}$ means "event i detected". This classification task is considered independent of other event categories, and we will call it a detection task.

Given an audio clip, a different score sequence is generated by each CRNN detector for each detection task i, as a time series with a frame rate that is determined by the resolution point employed. In order to compute the fusion of K different detectors, a final score s_i must be computed for each event in this unit of time, in order to make decisions, taking into account the sequences obtained from each individual detector, namely $(s_i^{(1)}, ..., s_i^{(K)})$. This combination is performed as a late integration, using the sigmoid outputs of each CRNN as score sequences, and before thresholding. By convention, higher scores indicate a stronger support to the presence of event i ($\theta_{i,1}$). The combined score is obtained as the average of the scores in this way:

$$s_i = \frac{1}{K} \sum_{j=1}^{K} s_i^{(j)} \tag{1}$$

In order to compute PSDS scores, 50 different thresholds (linearly distributed from 0.01 to 0.99) are applied to the combined scores s_i, obtaining binary time series which are then smoothed by means of a median filter.

Res.	$\mathbf{n_{mel}}$	**Pooling sizes [time, mel]**
$\mathbf{T}_{++}$	64	[2, 1], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]
$\mathbf{T}_{+}$	96	[2, 1], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 3]
BS	128	[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 2]
$\mathbf{F}_{+}$	192	[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 3]
$\mathbf{F}_{++}$	256	[2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2], [1, 4]

Table 3: Dimensions of the max-pooling layers in the convolutional stage, adapted for each resolution point. There is a total of seven max-pooling layers in the model, one after each convolutional layer. The total pooling factor is always $[4, n_{mel}]$.

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

Table 4: Parameter configuration for the PSDS scenarios. DTC = Detection Tolerance Criterion. GTC = Ground Truth intersection Criterion. α_{ST} = Cost of instability across classes. CTTC = Cross-Trigger Tolerance Criterion. α_{CT} = Cost of Cross Triggers. e_{max} = Maximum False Positive Rate.

4. EXPERIMENTS AND RESULTS

Our experiments are based upon the 2021 baseline system[1] released by the DCASE Team. The only modification applied to the structure of the CRNN is the adaptation of the max-pooling layers of the convolutional stage to the number of mel-filters employed by each resolution point. Namely, we adjust the sizes of the pooling operations in the mel-frequency axis so that the input to the RNN stage is a time series for each event category. The pooling sizes for each resolution point are described in Table 3.

In the first place, we trained the baseline system using each one of the resolution points for feature extraction, leading to five single-resolution systems. Afterwards, following the method described in Section 3.2, several sets of resolution points were combined, obtaining multi-resolution systems.

We report the results of single-resolution and multi-resolution systems over the DESED Validation set in terms of PSDS (Polyphonic Sound Detection Score) [10] and event-based, macro-averaged F_1-score [11]. In every case, the Teacher model obtained from the Mean Teacher training is employed to generate predictions.

In order to evaluate the performance SED systems in different conditions, two PSDS configurations are proposed. While the PSDS scenario 1 (or PSDS-1) gives special importance to the precise temporal localization of events, the PSDS scenario 2 (or PSDS-2) focuses on the correct detection of the event categories. The parameters that define these scenarios are described in Table 4.

4.1. Single-resolution results

Table 5 shows the results obtained with each of the feature resolution points described in 3.1 over the DESED Validation set. Their PSDS curves are shown in the top plots of Figure 1.

[1] https://github.com/DCASE-REPO/DESED_task

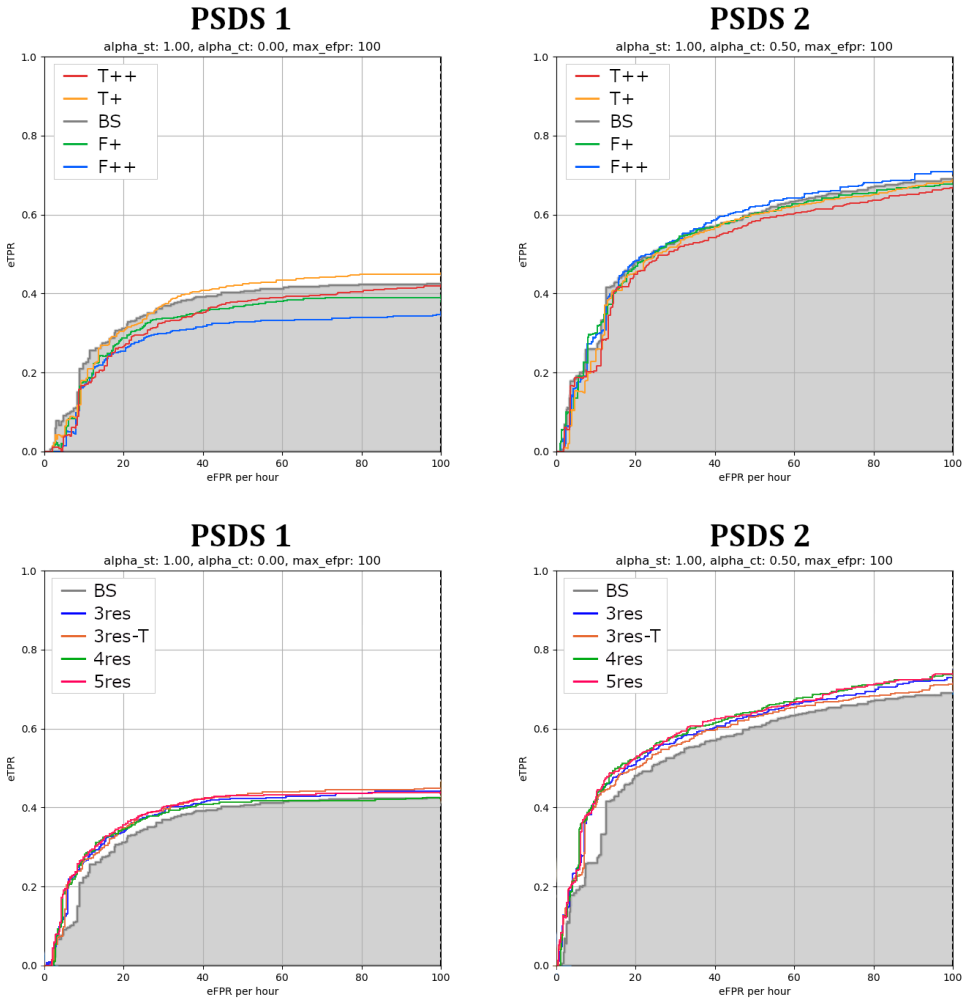

Figure 1: Polyphonic Sound Detection Score (PSDS) curves of the Baseline resolution (BS), the resolution points F_{++}, F_{+}, T_{+}, and T_{++} (top), and our submitted systems *3res, 3res-T, 4res,* and *5res* (bottom) over the DESED Validation set.

Resolution	**PSDS1**	**PSDS2**	F_1-score (%)
$\mathbf{F}_{++}$	0.286	**0.556**	32.8
$\mathbf{F}_{+}$	0.324	0.542	38.9
BS	0.357	0.549	**43.5**
$\mathbf{T}_{+}$	**0.367**	0.534	42.3
$\mathbf{T}_{++}$	0.328	0.520	41.3

Table 5: PSDS and F_1 results of single-resolution systems.

System	**Resolutions**	**PSDS1**	**PSDS2**	F_1(%)
3res	F_{+}, BS, T_{+}	0.380	0.589	45.0
3res-F	F_{++}, F_{+}, BS	0.361	0.589	45.1
3res-T	BS, T_{+}, T_{++}	**0.386**	0.578	**46.4**
4res	F_{++}, F_{+}, BS, T_{+}	0.372	**0.600**	45.1
5res	F_{++}, F_{+}, BS, T_{+}, T_{++}	**0.386**	**0.600**	**46.4**

Table 6: PSDS and F_1 results of multi-resolution systems.

According to the results, it seems that a higher time resolution is beneficial for PSDS-1, while PSDS-2 is optimized using finer frequency resolutions.

4.2. Multi-resolution results

In order to include information from different resolution points in the SED system, networks trained with different feature resolutions have been combined as described in Section 3.2.

Table 6 shows PSDS and event-based macro-averaged F_1 results for several model combinations. These fusions include the Baseline resolution (BS) along with some of the resolution points we have proposed. Combining models trained with different feature resolutions outperforms the baseline and other single-resolution models in both PSDS scenarios, as well as in terms of F_1-score.

Whereas two of the combined models (*3res* and *5res*) were already used in our last participation, this year we have defined other model fusions that give more importance to either time resolution

(*3res-T*) or frequency resolution (*3res-F*, *4res*). With this approach, we aimed to find combinations that optimize each of the PSDS scenarios separately.

The best result for the first PSDS scenario is achieved by the *3res-T* and the *5res* combinations, both of them achieving an area under curve (AUC) of 0.386. On the other hand, the best results for the second PSDS scenario are obtained with *4res* and *5res*, both of them reaching AUCs of 0.600. Thus, although each scenario is optimized by combining either higher time resolutions or higher frequency resolutions, the fusion of the five resolution points (*5res*) seems to optimize both of them at the same time.

The *3res*, *3res-T*, *4res* and *5res* combinations, described in Table 6, have been submitted to the challenge. Their PSDS curves are depicted in the bottom plots of Figure 1.

5. CONCLUSIONS

In this technical report, we describe our participation for the Task 4 of the DCASE 2021 Challenge. Built upon the baseline provided by the organization, and following the scenario of SED without source separation, our system combines different time-frequency resolution points of the mel-spectrogram features by averaging the output sequences of several CRNN detectors.

With this approach, we have been able to outperform the baseline system in both PSDS scenarios over the DESED Validation set. Moreover, we have found that certain resolutions and their combinations allow to optimize either the PSDS-1 (higher time resolutions) or PSDS-2 scenario (higher frequency resolutions), while a combination of five resolution points is able to optimize both scenarios at the same time.

6. REFERENCES

[1] 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.

[2] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano, "An analysis of Sound Event Detection under acoustic degradation using multi-resolution systems," in *Proc. IberSPEECH 2021*, 2021, pp. 36–40. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2021-8

[3] 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).

[4] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[5] 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.

[6] 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.

[7] 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

[8] 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

[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] 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.

[11] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, May 2016. [Online]. Available: http://dx.doi.org/10.3390/app6060162