MULTIPLE FEATURE RESOLUTIONS FOR DIFFERENT POLYPHONIC SOUND DETECTION SCORE SCENARIOS IN DCASE 2021 TASK 4

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

In this paper, we describe our multi-resolution mean teacher systems for 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 introduced 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. Additionally, the combination of several time-frequency resolutions through model fusion is able to improve the PSDS results in both scenarios. Furthermore, a class-wise analysis of the PSDS metrics is provided, indicating that the detection of each event category is optimized with different resolution points or model combinations.

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

1. INTRODUCTION

The development of competitive evaluations such as the DCASE (Detection and Classification of Acoustic Scenes and Events) Challenges, along with the introduction of datasets like Google AudioSet [1] or DESED (Domestic Environment Sound Event Detection) [2, 3], has supported the research in acoustic event detection tasks over the recent years.

DCASE 2021 Challenge 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 the DCASE 2020 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 a system derived from the SED baseline, in terms of both event-based F_1 score and Polyphonic Sound Detection Score (PSDS) [4, 5, 6].

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. Taking that into account, we have applied multi-resolution to the DCASE 2021 SED baseline system, which features the use of mixup [7] for data augmentation, as well as a larger synthetic subset, as main additions to the Mean Teacher [8] convolutional recurrent neural network (CRNN) system of previous years [9].

Our participation for DCASE 2021 Challenge 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.

Work developed under the project DSForSec (RTI2018-098091-B-I00), funded by the Ministry of Science, Innovation and Universities of Spain, and the European Regional Development Fund (ERDF).

2. DATASET

The dataset used for sound event detection in DCASE 2021 Task 4 is DESED, which is composed of real recordings, obtained from Google AudioSet, and synthetic recordings which are generated using the Scaper library [10]. 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 spec-

Resolution	T_{++}	T_{+}	BS	F_{+}	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 proposed time-frequency resolution working points. N, L, and R are reported in samples, using a sample rate $f_s = 16000$ Hz.

PSDS	DTC	GTC	α_{ST}	CTTC	α_{CT}	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 2: 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.

tral characteristics, which vary for each target class. For example, considering the Synthetic training set, some event categories have an average duration shorter than 2 seconds (*Alarm bell/ringing*, *Cat*, *Dishes*, *Dog*, and *Speech*), while other classes are more than 8 seconds long in average (*Electric shaver/toothbrush*, *Frying*, or *Vacuum cleaner*).

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.

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_{+}, and F_{++}) are described in Table 1.

3.2. Model fusion

For each event category i, a binary classification is performed between classes $\theta_{i,0}$, which means "event i not detected", and $\theta_{i,1}$, meaning "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 CRNN detector generates a different score sequence for each detection task i, as a time series with a frame rate that is determined by the resolution point employed. The fusion of K different detectors consists in a combination of their sequences $(s_i^{(1)}, ..., s_i^{(K)})$. This combination is performed as a late integration, using the sigmoid outputs of each CRNN as score sequences. By convention, higher scores indicate a stronger support to the presence of event i ($\theta_{i,1}$). The final score sequence is obtained as the frame-wise average of the K score sequences.

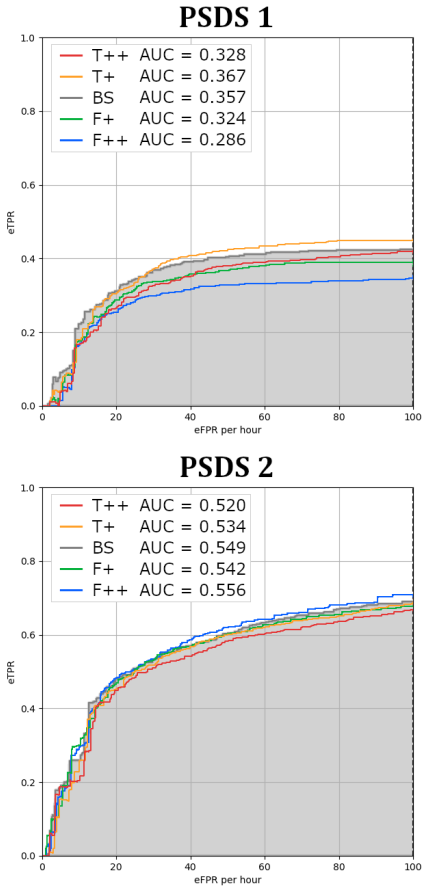

Figure 1: Polyphonic Sound Detection Score (PSDS) curves over the DESED Validation set of the single resolution F_{++}, F_{+}, BS, T_{+}, and T_{++} used to obtain the combined systems submitted to the evaluation.

System	Resolutions	PSDS 1	PSDS 2	$F_1(\%)$
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 3: PSDS and F_1 results of multi-resolution systems over the DESED Validation set.

System	Resolutions	PSDS 1	PSDS 2	$F_1(\%)$
3res	F_{+}, BS, T_{+}	0.343	0.571	42.6
3res-T	BS, T_{+}, T_{++}	**0.363**	0.574	**43.1**
4res	F_{++}, F_{+}, BS, T_{+}	0.345	0.571	42.2
5res	F_{++}, F_{+}, BS, T_{+}, T_{++}	0.361	**0.577**	42.7
Challenge Baseline		0.315	0.547	37.3

Table 4: PSDS and F_1 results of multi-resolution systems over the DESED 2021 Evaluation set.

PSDS 1	$\mathbf{F}_{++}$	$\mathbf{F}_{+}$	BS	$\mathbf{T}_{+}$	$\mathbf{T}_{++}$
Alarm bell/ringing	0.446±0.009	0.512±0.022	0.556±0.015	0.561±0.012	**0.567**±0.007
Blender	**0.694**±0.021	0.627±0.008	0.677±0.018	0.652±0.029	0.671±0.028
Cat	0.378±0.020	0.414±0.004	0.411±0.011	**0.439**±0.004	0.401±0.024
Dishes	0.107±0.008	0.132±0.010	**0.176**±0.010	0.172±0.039	0.121±0.020
Dog	0.242±0.003	0.272±0.008	0.306±0.010	**0.316**±0.005	0.295±0.012
Electric shaver/toothbrush	0.787±0.027	**0.798**±0.021	0.751±0.057	0.765±0.025	0.687±0.050
Frying	0.582±0.018	0.613±0.013	0.635±0.022	**0.639**±0.021	0.607±0.023
Running water	0.481±0.026	0.510±0.006	0.540±0.014	0.548±0.020	**0.553**±0.013
Speech	0.581±0.006	0.603±0.007	**0.631**±0.004	0.634±0.009	0.620±0.006
Vacuum cleaner	0.732±0.041	0.769±0.085	0.771±0.092	0.770±0.086	**0.790**±0.068
Overall PSDS 1	0.290±0.004	0.319±0.005	0.352±0.005	**0.358**±0.015	0.331±0.005

PSDS 2	$\mathbf{F}_{++}$	$\mathbf{F}_{+}$	BS	$\mathbf{T}_{+}$	$\mathbf{T}_{++}$
Alarm bell/ringing	**0.855**±0.003	0.852±0.007	0.836±0.004	0.842±0.004	0.814±0.011
Blender	**0.851**±0.006	0.783±0.016	0.799±0.014	0.782±0.014	0.791±0.016
Cat	**0.717**±0.011	0.705±0.009	0.661±0.015	0.665±0.014	0.622±0.016
Dishes	**0.394**±0.022	0.376±0.019	0.388±0.013	0.374±0.065	0.389±0.021
Dog	0.666±0.014	**0.672**±0.017	0.661±0.007	0.643±0.017	0.604±0.017
Electric shaver/toothbrush	**0.938**±0.020	0.913±0.017	0.885±0.016	0.912±0.015	0.851±0.011
Frying	0.771±0.018	0.780±0.009	**0.795**±0.019	**0.795**±0.012	0.759±0.018
Running water	0.714±0.011	0.714±0.014	0.749±0.012	0.750±0.015	**0.755**±0.015
Speech	0.830±0.007	0.821±0.007	**0.834**±0.007	0.822±0.009	0.813±0.006
Vacuum cleaner	0.892±0.006	**0.902**±0.011	0.886±0.013	0.879±0.014	0.873±0.018
Overall PSDS 2	**0.557**±0.009	0.544±0.013	0.553±0.007	0.544±0.029	0.534±0.012

	$\mathbf{F}_{++}$	$\mathbf{F}_{+}$	BS	$\mathbf{T}_{+}$	$\mathbf{T}_{++}$
Macro F_1 (%)	33.94±0.77	38.26±0.77	**42.58**±0.90	42.20±1.19	41.86±0.79

Table 5: PSDS (scenarios 1 and 2) results for each event category and overall PSDS and F_1 scores of single-resolution systems over the DESED Validation set. Mean and standard deviations are computed across 5 trainings of each system with different random initializations.

	PSDS 1					PSDS 2				
	3res	**3res-T**	**4res**	**5res**	**5×BS**	**3res**	**3res-T**	**4res**	**5res**	**5×BS**
Alarm bell/ringing	0.572	**0.584**	0.558	0.577	0.576	0.858	0.855	**0.870**	**0.870**	0.852
Blender	0.724	0.744	0.746	**0.768**	0.727	0.840	0.838	0.853	**0.856**	0.841
Cat	0.455	**0.472**	0.435	0.457	0.428	0.701	0.667	**0.727**	0.712	0.681
Dishes	0.202	0.200	0.197	**0.214**	0.206	0.415	0.402	0.435	**0.436**	0.419
Dog	0.319	**0.327**	0.312	0.324	0.326	0.693	0.681	**0.701**	0.700	0.689
Electric shaver/toothb.	0.740	0.695	0.739	0.714	**0.783**	0.902	0.909	**0.918**	0.916	0.917
Frying	0.677	**0.682**	0.668	0.674	0.678	**0.841**	0.836	0.829	0.833	0.832
Running water	0.567	**0.574**	0.562	0.569	0.560	0.775	**0.780**	0.771	0.775	0.772
Speech	0.661	**0.673**	0.659	0.666	0.658	0.851	**0.857**	0.852	0.855	0.850
Vacuum cleaner	**0.893**	0.885	0.877	0.890	0.815	**0.933**	0.923	0.932	0.932	0.921
Global PSDS	0.380	**0.386**	0.372	**0.386**	0.380	0.589	0.578	**0.600**	**0.600**	0.585
Macro F_1 (%)	44.97	**46.42**	45.13	**46.42**	45.84	44.97	**46.42**	45.13	**46.42**	45.84

Table 6: PSDS (scenarios 1 and 2) results of combined systems for each event category and overall PSDS and F_1 scores over the DESED Validation set. *3res*, *3res-T*, *4res* and *5res* are the multi-resolution combinations that were submitted to the challenge, whereas *5×BS* is a single-resolution combination of five models trained with the BS resolution point.

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.

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.

PSDS scores are computed applying 50 different thresholds (linearly distributed from 0.01 to 0.99) to the combined score sequences, obtaining binary time series which are then smoothed by means of a median filter.

We report the results in terms of PSDS [11] and event-based, macro-averaged F_1-score [12]. In every case, scores are generated employing the Teacher models obtained from the Mean Teacher training.

To allow the evaluation of SED performance in different conditions, the challenge organization proposes two PSDS configurations. While the PSDS scenario 1 (PSDS 1) gives special importance to the precise temporal localization of events, the PSDS scenario 2 (PSDS 2) focuses on the correct detection of the event categories. The parameters that define these scenarios are described in Table 2.

The PSDS curves obtained with each of the feature resolution points described in 3.1 over the DESED Validation set, as well as their AUC (Area Under Curve) metrics, are shown in Figure 1. 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. This behaviour was expected, taking into account that PSDS 1 is designed to focus on the temporal precision of the systems.

Aiming 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, obtaining the PSDS and macro F_1 results shown in Table 3. The model combinations 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.

The best result for the first PSDS scenario over the Validation set 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 PSDS 2 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 were submitted to the challenge, and their results are presented in Table 4. The best PSDS 1 over the 2021 Evaluation set is achieved by the 3res-T system (0.363), whereas the highest PSDS 2 is obtained by the 5res combination (0.577). Moreover, the performance of the submitted systems over the 2021 Evaluation set is very similar to that observed over the Validation set.

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

4.1. Class-wise results

In previous editions of the DCASE Challenge Task 4, SED systems were evaluated by means of the event-based macro F_1 score. Such metric is an average of the event-based F_1 scores for each target category, thus the scores for each individual class were usually highlighted in the results of the systems. On the other hand, whereas PSDS overcomes several limitations of the F_1 metric [13], the performance for each category is not usually described when reporting the results. For this reason, and considering that the detection of each event class is an independent task with an impact on the global results, we have computed the class-wise PSDS scores in terms of the Area Under Curve (AUC).

The class-wise PSDS results of the single-resolution systems are presented in Table 5. In each scenario, the best performing system in terms of global PSDS provides the largest AUC for several classes: in the first scenario, resolution T_+ holds the best results for *Cat*, *Dog* and *Frying*, whereas in the second scenario resolution F_{++} obtains the highest scores for *Alarm bell/ringing*, *Blender*, *Cat*, *Dishes* and *Electric shaver/toothbrush*. However, the rest of the event categories obtain better results with other resolutions, indicating that, as expected, the optimal resolution point depends not only on the PSDS settings but also on the characteristics of the target class.

Table 6 shows the PSDS results of combined systems. These systems include the multi-resolution fusions that have been submitted to the challenge (*3res*, *3res-T*, *4res*, and *5res*), as well as a combination of five different instances of the *BS* model (*5×BS*) which aims to contrast the performance of a single-resolution combination against the multi-resolution fusions. In most of the event categories, the largest AUC is obtained with a multi-resolution combination rather than with the single-resolution combination *5×BS*, being *Electric shaver/toothbrush* the exception in PSDS 1. Additionally, the *5×BS* achieves better global performance than the individual models. Therefore, it seems that the average fusion provides an improvement in both PSDS scenarios even when combining systems trained with the same resolution point. However, such improvement is larger when the systems to be combined have been trained with different resolutions.

5. CONCLUSIONS

In this work, we present the results of our participation in DCASE 2021 Challenge Task 4. Built upon the baseline provided by the organization, our proposed system combines different time-frequency resolution points of the mel-spectrogram features by averaging the output sequences of several CRNN detectors.

With the described approach, we have been able to outperform the baseline system in both PSDS scenarios and macro F_1 score over the DESED Validation and 2021 Evaluation sets. Moreover, the results indicate that certain resolutions and their combinations allow to optimize either the PSDS 1 (higher time resolutions) or PSDS 2 scenario (higher frequency resolutions), and that model fusion is more beneficial when different resolutions are combined.

Furthermore, the class-wise analysis of PSDS shows that the adequacy of each resolution point for sound event detection is related not only to the evaluation settings but also to the target category.

6. REFERENCES

[1] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[2] 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

[3] 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

[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] 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

[6] 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*, vol. 9, pp. 89 029–89 042, 2021.

[7] 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

[8] 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.

[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] 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.

[11] 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.

[12] 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

[13] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Bilen, and S. Krstulović, "Improving sound event detection metrics: Insights from DCASE 2020," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 631–635.