IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

# An analysis of Sound Event Detection under acoustic degradation using multi-resolution systems

*Diego de Benito-Gorrón, Daniel Ramos, Doroteo T. Toledano*

AUDIAS Research Group
Escuela Politécnica Superior,
Universidad Autónoma de Madrid
(Madrid, Spain)

{diego.benito, daniel.ramos, doroteo.torre}@uam.es

## Abstract

The Sound Event Detection task aims to determine the temporal locations of acoustic events in audio clips. Over the recent years, this field is holding a rising relevance due to the introduction of datasets such as Google AudioSet or DESED (Domestic Environment Sound Event Detection) and competitive evaluations like the DCASE Challenge (Detection and Classification of Acoustic Scenes and Events). In this paper, we analyse the performance of Sound Event Detection systems under diverse acoustic conditions such as high-pass or low-pass filtering, clipping or dynamic range compression. For this purpose, the audio has been obtained from the Evaluation subset of the DESED dataset, whereas the systems were trained in the context of the DCASE Challenge 2020 Task 4. Our systems are based upon the challenge baseline, which consists of a Convolutional-Recurrent Neural Network trained using the Mean Teacher method, and they employ a multi-resolution approach which is able to improve the Sound Event Detection performance through the use of several resolutions during the extraction of Mel-spectrogram features. We provide insights on the benefits of this multi-resolution approach in different acoustic settings. Furthermore, we compare the performance of the single-resolution systems in the aforementioned scenarios when using different resolutions.

**Index Terms**: Sound Event Detection, DCASE Challenge 2020, Multi-resolution, Acoustic degradation.

## 1. Introduction

Sound events can be defined as the acoustic signals that are directly caused by a particular ocurrence in the near environment, so that a human can identify the event by hearing them. Some clear examples of sound events would be an alarm bell, a dog barking or a person speaking.

Aiming to automatically localize and classify the sound events in audio signals, the task of Sound Event Detection (SED) is an ongoing challenge for machine perception. Training deep learning algorithms in order to develop SED systems requires the use of a large amount of annotated data, which is usually costly to obtain. However, several public datasets have been released over the last years, such as Google AudioSet [1], FSD (FreesoundDataset) [2] or DESED (Domestic Environment Sound Event Detection) [3], which are specifically built to train and evaluate SED systems and consist of audio recordings extracted from web sources, such as YouTube[1], Freesound[2] or Vimeo[3]. These audio recordings can be strongly or weakly annotated, depending on whether the temporal location (onset and offset times) of each event is included or not. In addition to a large scale weakly-labeled audio dataset, Google AudioSet introduced an ontology of more than 500 sound event categories in which sound events can be classified, which has been used as well in the other two mentioned datasets.

On the other hand, the recent development of SED systems and techniques has been notably supported by the DCASE (Detection and Classification of Acoustic Scenes and Events) yearly evaluations, which have helped to define benchmarks not only for Sound Event Detection, but also for other related tasks such as Acoustic Scene Classification [4] or Anomalous Sound Detection [5], among others. Regarding Sound Event Detection, the DCASE 2020 Challenge proposed the task called "Sound event detection and separation in domestic environments", which consisted on locating the temporal boundaries of sound events in ten-seconds audio clips and classifying them.

During the DCASE 2020 Challenge, we developed a multi-resolution approach to Sound Event Detection that was able to outperform the evaluation baseline exploiting the use of several time-frequency resolutions in the process of mel-spectrogram feature extraction, combining up to five different resolution points.

In this paper, we offer an analysis of the performance of single-resolution and multi-resolution SED systems when facing adverse acoustic scenarios that critically affect the spectra of the acoustic signals (high-pass and low-pass filtering) or their dynamic range (clipping and dynamic range compression). For this purpose, we process the audio segments of the Public Evaluation set of DESED in order to achieve the mentioned acoustic conditions, then SED metrics are computed over the obtained sets. Through this study, we aim to determine whether the improvement on performance obtained by the multi-resolution approach is robust to the proposed types of acoustic degradation. These adverse settings represent possible scenarios that could be found when applying the detectors in other data, obtained from web sources or from a real life application.

The rest of the paper is organized as follows: Section 2 explains the Sound Event Detection task of the DCASE Challenge 2020, as well as our multi-resolution approach. Section 3 describes the motivation of this analysis and the different acoustic scenarios that we are considering. In Section 4, the results of the experiments are provided and discussed. Finally, the conclusions of this work are highlighted in Section 5.

[1] http://youtube.com/
[2] http://freesound.org/
[3] http://vimeo.com/


10.21437/IberSPEECH.2021-8

## 2. Sound Event Detection in DCASE 2020

### 2.1. DCASE 2020 Challenge: "Sound Event Detection and Separation in Domestic Environments"

In the 2020 edition of the DCASE Challenge, one of the task proposes a Sound Event Detection scenario where systems are trained using the DESED dataset. This dataset includes weakly-labeled data and unlabeled data extracted from Google AudioSet, along with strongly-labeled data which is synthetically generated using the Scaper toolkit [6]. In addition, a subset of AudioSet segments is provided (DESED Validation set) with strong, human-verified annotations, which is used to validate the performance of the systems. An optional pre-processing step based on sound separation is proposed in the task, although our work does not take it into account.

The set of target categories includes ten event categories which are usually found in the acoustic context of a house: *Speech*, *Dog*, *Cat*, *Alarm/bell/ringing*, *Dishes*, *Frying*, *Blender*, *Running water*, *Vacuum cleaner* and *Electric shaver/toothbrush*.

Systems output the predicted onset and offset times of the detected events, along with their category. To define whether a prediction is correct, a collar of 200 ms is considered for the onset times, whereas for the offset times the collar is the maximum between 200 ms and the 20% of the event length, aiming to handle the difficulty to determine the offset times of long events. The system performance is measured by means of the $F_1$ score metric, which is computed as a combination of the True Positive (TP), False Positive (FP) and False Negative (FN) counts [7].

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

First, $F_1$ scores are computed for each event category, then the Macro $F_1$ is obtained by averaging the class-wise $F_1$ scores. Macro $F_1$ is used to measure the global performance of the systems.

The challenge provides a baseline system as a benchmark of SED performance [8]. Such system is based on a Convolutional Recurrent Neural Network trained using the Mean Teacher method [9] for semi-supervised learning. This method allows the network to learn from labeled and unlabeled data. Additionally, an attention module is used to infer the temporal locations of events using weak labels. The system is fed with the mel-spectrogram features of the audio segments.

### 2.2. Multi-resolution analysis

Each sound event category shows different temporal and spectral characteristics. Therefore, in order to improve SED performance, our main idea is that different time-frequency resolutions would be more suited to detect different types of events. Thus, combining the information of several mel-spectrogram features extracted at different resolution points should lead to a better overall performance [10].

Aiming to test this hypothesis, we defined five different time-frequency resolutions, taking as a starting point the resolution used by the baseline system, which we called $BS$. Each resolution point is defined by a set of values for the parameters of feature extraction. It should be noted that, due to the feature extraction process, there is a compromise between temporal resolution and frequency resolution. Hence, we propose a resolution point with twice better time resolution than the baseline, which we call $T_{++}$, and a resolution point with twice better frequency resolution than the baseline, which we call $F_{++}$. In the intermediate points between each of these points and $BS$, we define $T_{+}$ and $F_{+}$, respectively.

In order to obtain multi-resolution systems, first we trained single-resolution systems, which were based on the DCASE Challenge baseline and modified to operate on each of the different resolution points. Then, we performed a model fusion averaging the posterior probabilities given by systems trained with different resolutions. Using this method, we obtained a three-resolution system which combines the $BS$ resolution with $T_{++}$ and $F_{++}$, denoted as *3res* in this paper, and a five-resolution system combining all the mentioned resolutions, denoted as *5res*.

Through the use of the *3res* and *5res* systems, we were able to outperform the single-resolution baseline system in the DCASE 2020 Challenge task 4. The improvement of performance in terms of macro $F_1$ score was observed over the DESED Validation set and the YouTube subset of the DESED 2019 Evaluation set, which is called DESED Public evaluation set. The *5res* system was submitted to the evaluation and outperformed the baseline system over the DESED 2020 Evaluation set.

## 3. Experiments

Both the DESED Validation set and the Public Evaluation set consist of YouTube audio segments drawn from Google AudioSet. Due to the crowdsourced nature of a web resource like YouTube, the audio clips can have very diverse origins and qualities, ranging from mobile recordings to professional studio productions. Therefore, the evaluation of Sound Event Detection on YouTube data requires the systems to be able to handle a variety of acoustic conditions that sometimes may be adverse for the task.

In order to test the performance of Sound Event Detection in a wider range of acoustic settings, we have applied several types of degradations to the DESED Public evaluation set, which contains 692 audio clips. We have computed the $F_1$ scores of single-resolution and multi-resolution systems over the original set and its degraded copies, aiming to analyze to what extent does multi-resolution help to improve performance when the test data is degraded.

The acoustic conditions that we have considered for our experiments are inspired by some of the scenarios described in the DESED Synthetic evaluation set, which has already been used to analyze the performance of state-of-the-art SED systems [11].

### 3.1. Acoustic degradation scenarios

We consider three types of degradations for the audio clips: frequency filtering, dynamic range compression and clipping. We apply each perturbation to the whole dataset, obtaining a total of eight copies of the DESED Public evaluation set:

- **Frequency filtering.** We apply high-pass filtering and low-pass filtering separately. In both cases, the cutoff frequencies are 500 Hz, 1000 Hz and 2000 Hz, leading to a total of six copies of the DESED Public evaluation set.
- **Dynamic range compression.** We apply dynamic range compression with a threshold of -50 dB and a ratio value of 5.
- **Clipping.** To obtain clipping distortion, we multiply the

audio signals, which are bounded to $[-1, 1]$, by a scale factor of 5, limiting the output values again to $[-1, 1]$.

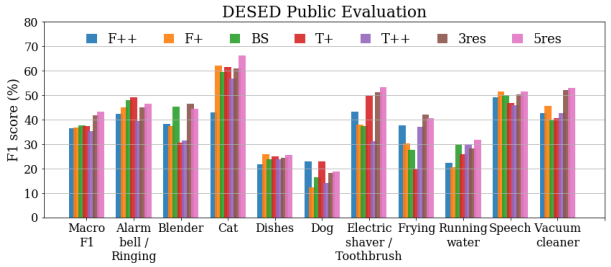


Figure 1: *$F_1$ scores of the single-resolution systems ($F_{++}, F_+, BS, T_+, T_{++}$) and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set. Best viewed in color.*

## 4. Results

All the results are provided in terms of event-based $F_1$ score, considering the same collar settings as in the DCASE 2020 Challenge task 4.

### 4.1. Results over DESED Public evaluation set

The results of the seven systems over the original DESED Public evaluation set are presented in Figure 1. The figure represents the $F_1$ scores of each system in groups of bars, one group for each event category and an additional one for the macro average, which represents the global performance.

It can be observed that, in terms of macro $F_1$, the *3res* and *5res* systems both outperform every single-resolution system. However, this improvement is not applicable to every target class. Whereas most event categories obtain their best performance when using a multi-resolution system, other classes reach their maximum $F_1$ score with a single-resolution system: This is the case of *Alarm bell/ringing* ($T_+$), *Dishes* ($F_+$), *Dog* ($T_+$) and *Speech* ($F+$).

### 4.2. Results under acoustic degradation

#### *4.2.1. High-pass filtering*

The results obtained when applying high-pass filtering to the DESED Public evaluation set are shown in Figure 2. Three separate graphs are presented, one for each cutoff frequency ($f_c$). As expected, the general performance decreases for every class and every system when the cutoff frequency of the high-pass filter increases. In terms of macro $F_1$ score, the multi-resolution systems *3res* and *5res* achieve the best results for $f_c = 500$ Hz, similarly to the clean set. However, for $f_c = 1000$ Hz and $f_c = 2000$ Hz the highest macro $F_1$ scores are obtained with some of the single-resolution systems, $BS$ and $T_+$, respectively.

#### *4.2.2. Low-pass filtering*

Figure 3 shows the results for the DESED Public evaluation set after applying low-pass filtering with $f_c = 2000$ Hz, $f_c = 1000$ Hz and $f_c = 500$ Hz. It can be seen that the performances decrease when lowering the cutoff frequency of the filter, which is the expected behavior. When using a cutoff frequency $f_c =$

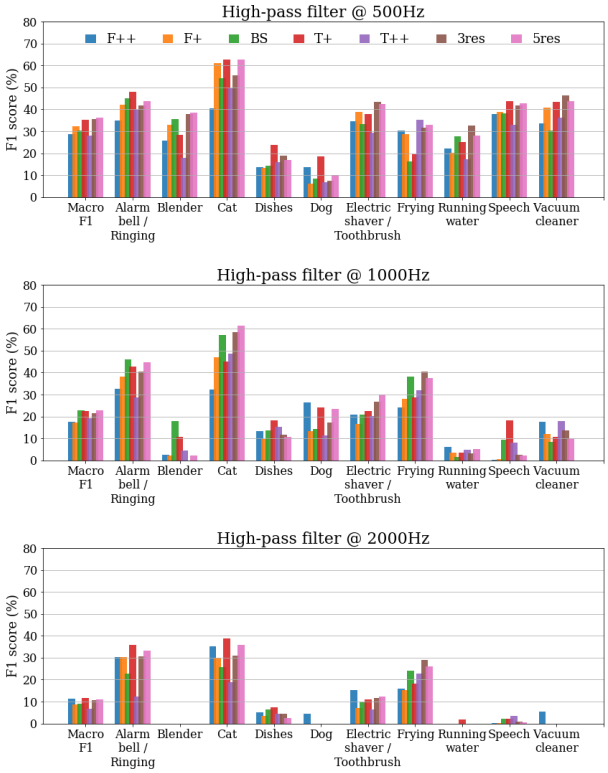


Figure 2: *$F_1$ scores of the single-resolution systems ($F_{++}, F_+, BS, T_+, T_{++}$) and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying a high-pass filter with cutoff frequencies of 500 Hz (top), 1000 Hz (center) and 2000 Hz (bottom). Best viewed in color.*

2000 Hz, the best overall performance is obtained by the multi-resolution system *5res*, whereas for $f_c = 1000$ Hz *3res* and $T_{++}$ both achieve the highest macro $F_1$. When the cutoff frequency is set to $f_c = 500$ Hz, the best macro $F_1$ scores are obtained with the single-resolution systems $T_+$ and $T_{++}$.

#### *4.2.3. Dynamic range compression*

The results obtained after applying dynamic range compression to the DESED Public evaluation set are presented in Figure 4. In this scenario, the best overall performance (macro $F_1$) is obtained by the multi-resolution systems, *3res* and *5res*. However, for some particular classes the best performance is obtained with a single-resolution system, as observed in the clean set results. Such is the case of *Alarm bell/ringing* ($F_{++}$), *Cat* ($F_+$), *Dishes* ($T_{++}$) and *Running water* ($T_{++}$).

#### *4.2.4. Clipping*

Figure 5 presents the results obtained when applying clipping saturation to the Public evaluation set. The best macro $F_1$ performances are achieved by the multi-resolution systems, whereas in some event categories multi-resolution is not able to outperform every single-resolution system. This situation is observed for *Dishes* ($T_+$), *Dog* ($F_{++}$) and *Shaver* ($F_+$).

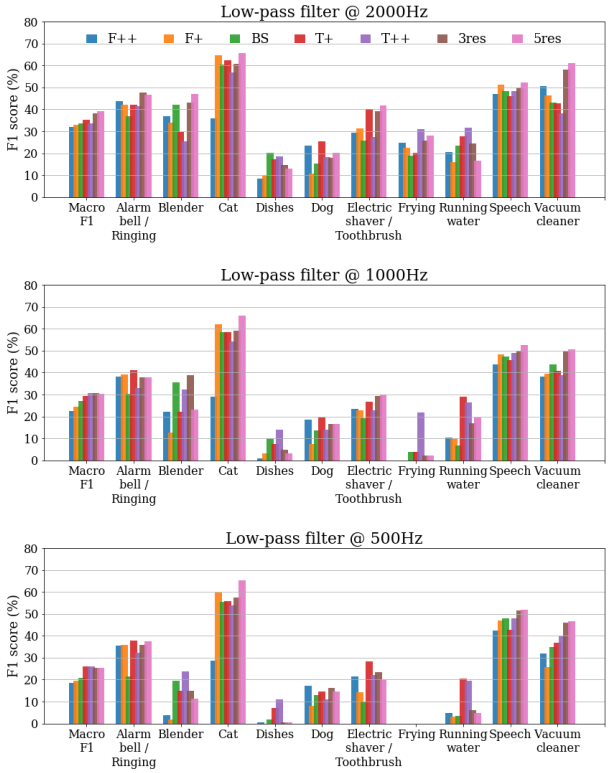


Figure 3: *$F_1$ scores of the single-resolution systems ($F_{++}, F_{+}, BS, T_{+}, T_{++}$) and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying a low-pass filter with cutoff frequencies of 2000 Hz (top), 1000 Hz (center) and 500 Hz (bottom). Best viewed in color.*

### 4.3. Discussion

It has been shown that the proposed adverse settings have a negative impact on the performance of both single-resolution and multi-resolution Sound Event Detection systems. Overall, the most critical scenario is high-pass filtering, especially with $f_c$ values of 1000 Hz and above. This suggests that the information of low frequencies is essential for this task, especially when considering categories like *Blender*, *Speech*, *Running water* or *Vacuum cleaner*. On the other hand, low-pass filtering is the most adverse condition for the class *Frying*, implying that high frequencies are particularly relevant for this event.

The results also show that the improvement on performance obtained when combining several single-resolution systems into a multi-resolution system does not always hold when facing very adverse conditions. Likely, this effect is due to the way in which our multi-resolution systems are obtained. An average fusion of the scores of different models can result in more accurate scores when the individual scores are precise enough. On the other hand, in scenarios where the individual systems perform worse, the average fusion is not able to obtain better results.

Nevertheless, it can be observed that, under these adverse settings, multi-resolution systems have an overall result approximately as good as the best performing resolution in each case, which means that our multi-resolution approach provides an improved robustness against these very adverse distortion scenarios.

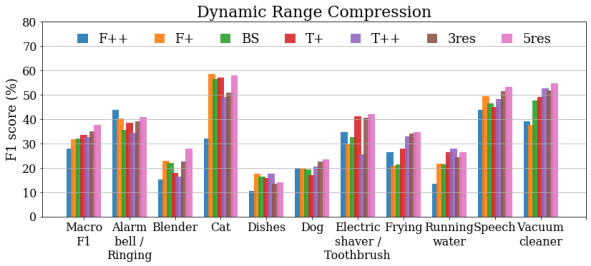


Figure 4: *$F_1$ scores of the single-resolution systems ($F_{++}, F_{+}, BS, T_{+}, T_{++}$) and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying dynamic range compression. Best viewed in color.*

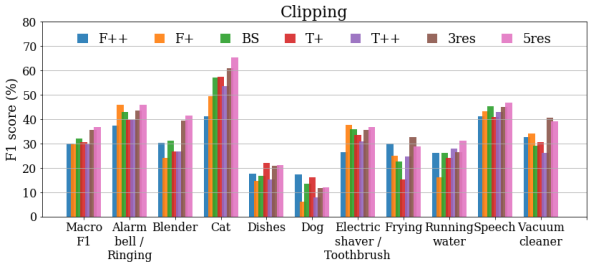


Figure 5: *$F_1$ scores of the single-resolution systems ($F_{++}, F_{+}, BS, T_{+}, T_{++}$) and the multi-resolution systems (3res, 5res) over the DESED Public Evaluation set applying clipping saturation. Best viewed in color.*

## 5. Conclusions

In this paper, we have studied the performance of several Sound Event Detection systems over a public dataset when diverse acoustic perturbations are applied. Five of these systems are convolutional neural networks with a common structure, but employing mel-spectrogram features extracted using different time-frequency resolutions. Two more systems are considered, which combine the previous systems into multi-resolution models by means of an average fusion, increasing the performance over the evaluation subsets of the DESED dataset.

According to the results, the proposed acoustic scenarios have, as expected, a clearly negative impact on the performance of our systems. Although it is shown that our multi-resolution approach is robust to slight degradations, the average fusion is unable to improve performance when facing very adverse conditions. Additionally, an extra robustness against these adverse distortion scenarios is observed when using multiple resolutions. We are currently working on alternative implementations of the multi-resolution approach in which fusion is performed earlier, aiming to improve both performance and robustness.

Furthermore, the data generated and the results obtained through this study will serve as a benchmark to evaluate the performance of future Sound Event Detection approaches and their robustness to diverse acoustic settings.

## 6. Acknowledgements

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

# 7. 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] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, Suzhou, China, 2017, pp. 486–493.

[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] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020, submitted. [Online]. Available: https://arxiv.org/abs/2005.14623

[5] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and Discussion on DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 81–85.

[6] 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.

[7] 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

[8] 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.

[9] 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.

[10] D. de Benito-Gorron, 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.

[11] N. Turpault, R. Serizel, S. Wisdom, H. Erdogan, J. Hershey, E. Fonseca, P. Seetharaman, and J. Salamon, "Sound Event Detection and Separation: a Benchmark on DESED Synthetic Soundscapes," 2020. [Online]. Available: arXiv:2011.00801 [cs.SD]