Multi-resolution and Source Separation for Improved Sound Event Detection based on Deep Neural Networks

PhD Thesis

Diego de Benito Gorrón

AUDIAS – Audio, Data Intelligence and Speech
Escuela Politécnica Superior
Universidad Autónoma de Madrid

October 9, 2023

UAM Universidad Autónoma de Madrid

< audias >

Escuela Politécnica Superior

PhD Candidate's background

Diego de Benito Gorrón

- Ph.D. in Computer Science and Telecommunication (EPS-UAM, 2018–2023)
- Master's Degree in ICT Research and Innovation (EPS-UAM, 2017–2018)
- Bachelor's Degree in Telecommunication Technology and Service Engineering (EPS-UAM, 2013–2017)
    - Special mention to the best academic records

“Multi-resolution and Source Separation for Improved Sound Event Detection based on Deep Neural Networks”

- Thesis elaborated within the AUDIAS research group (EPS-UAM)
- Presented as a compendium of publications
  - Three journal articles and two conference papers as first author
- Mainly funded by a FPI-UAM contract (November 2018 — March 2023)

Escuela Politécnica Superior

“Multi-resolution and Source Separation for Improved Sound Event Detection based on Deep Neural Networks”

- International PhD mention
  - Research stay at Brno University of Technology (Czech Republic)
  - Speech@FIT research group
  - September to December 2021

Overview

Introduction

Introduction — Sound Event Detection (SED)

Introduction — Sound Event Detection (SED)

Speech

- Automatic Speech Recognition
- Speaker Identification
- ...

Audio signal

Smart device

Introduction — Sound Event Detection (SED)

Speech

- Automatic Speech Recognition
- Speaker Identification
- ...

Audio signal

Smart device

Music

- Song/Artist Identification
- Genre Classification
- ...

Audio signal

Smart device

Speech

- Automatic Speech Recognition
- Speaker Identification
- ...

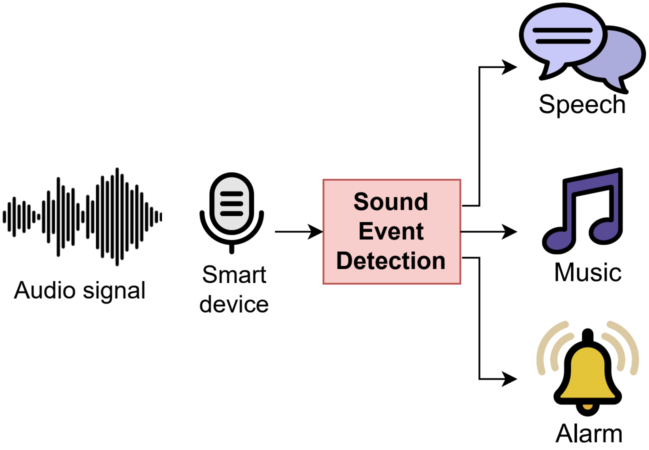
Music

- Song/Artist Identification
- Genre Classification
- ...

Alarm

- Detection of threats and emergencies

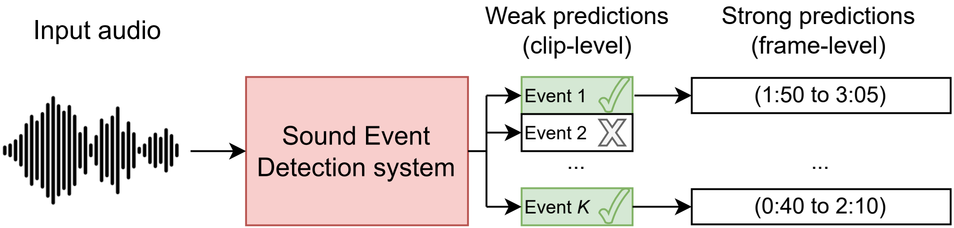
Introduction — Sound Event Detection (SED)



- Automatic Speech Recognition
- Speaker Identification
- ...

- Song/Artist Identification
- Genre Classification
- ...

- Detection of threats and emergencies



Sound Event Detection

- Determine the active sound events in an audio clip, considering a closed set of categories (*weak* SED)
- Additionally, determine the onset and offset times of each active event (*strong* SED)

Introduction — Sound Event Detection (SED)

Applications

- Pre-processing step for event-specific audio tasks
  - *Speech*, *Music*, ...
- Automatic labeling of multimedia contents, home assistance, security or medical diagnosis

Applications

- Pre-processing step for event-specific audio tasks
    - *Speech*, *Music*, ...
- Automatic labeling of multimedia contents, home assistance, security or medical diagnosis

1. Development of *structure-agnostic* methods for neural-network-based SED
   - Independent of the neural network structure
   - Enhancement of input representations: multi-resolution, source separation
2. Validation in standardized benchmarks and competitive evaluations
   - Google AudioSet, DCASE Challenge
3. Analysis and interpretation for different event categories and acoustic conditions

1. Development of *structure-agnostic* methods for neural-network-based SED
   - Independent of the neural network structure
   - Enhancement of input representations: multi-resolution, source separation
2. Validation in standardized benchmarks and competitive evaluations
   - Google AudioSet, DCASE Challenge
3. Analysis and interpretation for different event categories and acoustic conditions

Introduction — Thesis goals

1. Development of *structure-agnostic* methods for neural-network-based SED
   - Independent of the neural network structure
   - Enhancement of input representations: multi-resolution, source separation
2. Validation in standardized benchmarks and competitive evaluations
   - Google AudioSet, DCASE Challenge
3. Analysis and interpretation for different event categories and acoustic conditions

Convolutional and Recurrent Deep Neural Networks for speech and music detection

Convolutional and Recurrent DNNs for speech and music detection

- Initial work of the PhD Thesis (2018–2019), built upon the candidate's Master's Thesis [1]
- Detection of *Speech* and *Music* events in the large-scale dataset Google AudioSet
  - Employing fully-connected DNNs, CNNs, and LSTM
  - Only weak (i.e. clip-level) detection, no time boundaries specified
- The work led to the publication of the first journal article of the Thesis [2]

---

[1] D. de Benito-Gorrón "Detección de voz y música en un corpus a gran escala de eventos de audio" MA thesis, 2018.

[2] D. de Benito-Gorrón et al. "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset" *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Convolutional and Recurrent DNNs for speech and music detection

de Benito-Gorron *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2019) 2019:9
https://doi.org/10.1186/s13636-019-0152-1

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH **Open Access**

Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset

Diego de Benito-Gorron*, Alicia Lozano-Diez, Doroteo T. Toledano and Joaquin Gonzalez-Rodriguez

Abstract

Audio signals represent a wide diversity of acoustic events, from background environmental noise to spoken communication. Machine learning models such as neural networks have already been proposed for audio signal modeling, where recurrent structures can take advantage of temporal dependencies. This work aims to study the implementation of several neural network-based systems for speech and music event detection over a collection of 77,937 10-second audio segments (216 h), selected from the Google AudioSet dataset. These segments belong to YouTube videos and have been represented as mel-spectrograms. We propose and compare two approaches. The first one is the training of two different neural networks, one for speech detection and another for music detection. The second approach consists on training a single neural network to tackle both tasks at the same time. The studied architectures include fully connected, convolutional and LSTM (long short-term memory) recurrent networks. Comparative results are provided in terms of classification performance and model complexity. We would like to highlight the performance of convolutional architectures, specially in combination with an LSTM stage. The hybrid convolutional-LSTM models achieve the best overall results (85% accuracy) in the three proposed tasks. Furthermore, a distractor analysis of the results has been carried out in order to identify which events in the ontology are the most harmful for the performance of the models, showing some difficult scenarios for the detection of music and speech.

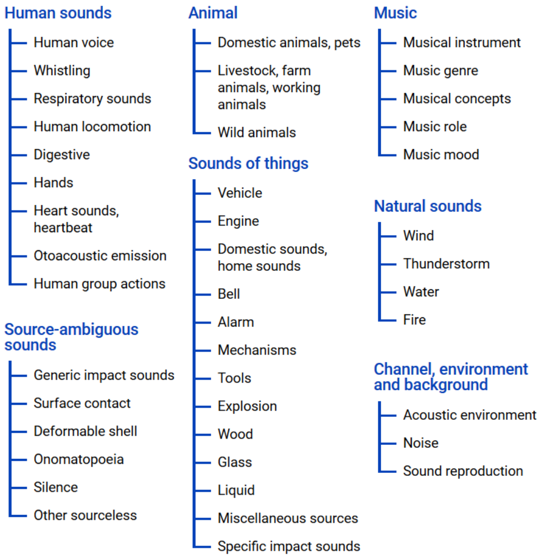
Keywords: Acoustic event detection, Speech activity detection, Music activity detection, Neural networks, Convolutional networks, LSTM

Google AudioSet

- Introduced by Google Research in 2017 [3]
- **Ontology**: Over 600 classes
- **Dataset**: More than 2 million ten-second audio clips from YouTube
  - Including weak labels regarding the classes in the ontology
  - Not balanced across classes

[3] J. F. Gemmeke, D. P. W. Ellis, et al. “Audio Set: An ontology and human-labeled dataset for audio events” *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

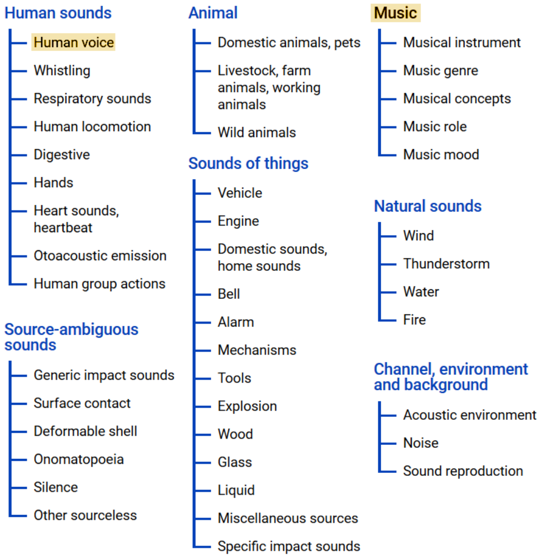
Speech and Music detection in Google AudioSet



- More than 600 classes in 7 groups
- Definition of a **balanced subset** [a] with respect to *Speech* and *Music* events (78000 clips)
    - List of files publicly available

[a] D. de Benito-Gorrón et al. "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset" *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Speech and Music detection in Google AudioSet



- More than 600 classes in 7 groups
- Definition of a **balanced subset** [a] with respect to *Speech* and *Music* events (78000 clips)
  - List of files publicly available

[a] D. de Benito-Gorrón et al. "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset" *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

- Weak Sound Event Detection $\rightarrow$ Clip-level classification
  - Between 2 classes
    - Speech detection: *Speech* or *Non-speech*
    - Music detection: *Music* or *Non-music*
  - Between 4 classes
    - *Speech + Music*
    - *Speech + Non-music*
    - *Non-speech + Music*
    - *Non-speech + Non-music*

Speech and Music detection in Google AudioSet

- Weak Sound Event Detection $\rightarrow$ Clip-level classification
  - Between 2 classes
    - Speech detection: *Speech* or *Non-speech*
    - Music detection: *Music* or *Non-music*
  - Between 4 classes
    - *Speech + Music*
    - *Speech + Non-music*
    - *Non-speech + Music*
    - *Non-speech + Non-music*

Speech and Music detection in Google AudioSet

- Weak Sound Event Detection $\rightarrow$ Clip-level classification
  - Between 2 classes
    - Speech detection: *Speech* or *Non-speech*
    - Music detection: *Music* or *Non-music*
  - Between 4 classes
    - *Speech + Music*
    - *Speech + Non-music*
    - *Non-speech + Music*
    - *Non-speech + Non-music*

Experiments [4]

- Grid search — N° of hidden layers (L) and n° of nodes in each layer (N)
  - Fully-connected (performance baseline)
  - Convolutional Neural Networks (CNN, 3×3 or 7×7 kernels)
  - Long Short-Term Memory (LSTM)
  - Convolutional Recurrent Neural Networks (CNN + LSTM, 1D or 2D convolutions)

[4] D. de Benito-Gorrón et al. "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset" *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Speech and Music detection in Google AudioSet

Baseline network — Fully-connected

Best-performing network — 2D CRNN with $L = 6$ conv. layers and $N = 256$

	Fully-connected	2D CRNN
Speech	75.6%	**83.8%**
Music	72.7%	**84.2%**
4-class problem	55.8%	**71.0%**

Accuracy results over the Test subset (23383 clips)

Speech and Music detection in Google AudioSet

Distractor events — How do other events interfere with detecting *Speech* or *Music*?

- Definition of an objective measure based on conditional probabilities
- Positive distractors (d^+) and Negative distractors (d^-)

$$d^+_{(t,\,dist)} = \frac{\mathrm{N}(y_t = 1,\ \tau_t = 0,\ \tau_{dist} = 1)}{\mu + \mathrm{N}(\tau_t = 0,\ \tau_{dist} = 1)}$$

$$d^-_{(t,\,dist)} = \frac{\mathrm{N}(y_t = 0,\ \tau_t = 1,\ \tau_{dist} = 1)}{\mu + \mathrm{N}(\tau_t = 1,\ \tau_{dist} = 1)}$$

y = System prediction
τ = Ground truth annotation
$\cdot_t$ = Target event
$\cdot_{dist}$ = Distractor event
μ = Aux. term (avg. nº of events)

Speech and Music detection in Google AudioSet

Distractor events — How do other events interfere with detecting *Speech* or *Music*?

- Definition of an objective measure based on conditional probabilities
- Positive distractors (d^+) and Negative distractors (d^-)

$$d^+_{(t,\,dist)} = \frac{\mathrm{N}(y_t = 1,\ \tau_t = 0,\ \tau_{dist} = 1)}{\mu + \mathrm{N}(\tau_t = 0,\ \tau_{dist} = 1)}$$

$$d^-_{(t,\,dist)} = \frac{\mathrm{N}(y_t = 0,\ \tau_t = 1,\ \tau_{dist} = 1)}{\mu + \mathrm{N}(\tau_t = 1,\ \tau_{dist} = 1)}$$

y = System prediction
τ = Ground truth annotation
$\cdot_t$ = Target event
$\cdot_{dist}$ = Distractor event
μ = Aux. term (avg. nº of events)

Speech and Music detection in Google AudioSet

Distractor events — How do other events interfere with detecting *Speech* or *Music*?

- Definition of an objective measure based on conditional probabilities
- Positive distractors (d^+) and Negative distractors (d^-)

$$d^+_{(t,\,dist)} = \frac{\mathrm{N}(y_t = 1,\ \tau_t = 0,\ \tau_{dist} = 1)}{\mu + \mathrm{N}(\tau_t = 0,\ \tau_{dist} = 1)}$$

$$d^-_{(t,\,dist)} = \frac{\mathrm{N}(y_t = 0,\ \tau_t = 1,\ \tau_{dist} = 1)}{\mu + \mathrm{N}(\tau_t = 1,\ \tau_{dist} = 1)}$$

y = System prediction
τ = Ground truth annotation
$\cdot_t$ = Target event
$\cdot_{dist}$ = Distractor event
μ = Aux. term (avg. nº of events)

Speech and Music detection in Google AudioSet

Distractor events [5]

- Positive distractors (d^+) cause False Positives
- Negative distractors (d^-) cause False Negatives

	Speech	Music
d^+	Crowd, Cheering, ...	Percussion, Singing, Organ, ...
d^-	Whispering, Singing, Music, ...	Inside/small room, Outside/rural or natural, Speech, ...

- Semantic similarity
- Difficult conditions
- Masking between events

[5] D. de Benito-Gorrón et al. "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset" *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Distractor events [5]

- Positive distractors (d^+) cause False Positives
- Negative distractors (d^-) cause False Negatives

	Speech	Music
d^+	Crowd, Cheering, ...	Percussion, Singing, Organ, ...
d^-	Whispering, Singing, Music, ...	Inside/small room, Outside/rural or natural, Speech, ...

- Semantic similarity
- Difficult conditions
- Masking between events

[5] D. de Benito-Gorrón et al. "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset" *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Distractor events [5]

- Positive distractors (d^+) cause False Positives
- Negative distractors (d^-) cause False Negatives

	Speech	Music
d^+	Crowd, Cheering, ...	Percussion, Singing, Organ, ...
d^-	Whispering, Singing, Music, ...	Inside/small room, Outside/rural or natural, Speech, ...

- Semantic similarity
- Difficult conditions
- Masking between events

[5] D. de Benito-Gorrón et al. "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset" *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Speech and Music detection in Google AudioSet

Distractor events [5]

- Positive distractors (d^+) cause False Positives
- Negative distractors (d^-) cause False Negatives

	Speech	Music
d^+	Crowd, Cheering, ...	Percussion, Singing, Organ, ...
d^-	Whispering, Singing, Music, ...	Inside/small room, Outside/rural or natural, Speech, ...

- Semantic similarity
- Difficult conditions
- Masking between events

[5] D. de Benito-Gorrón et al. “Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset” *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Multi-resolution for Neural Network-based Sound Event Detection in domestic environments

Multi-resolution for Neural Network-based SED in domestic environments

- Participation in the DCASE Challenge competitive evaluations in 2020, 2021, 2022
  - Detection of ten event categories in domestic environments
- The work has led to the publication of two conference papers [6] [7] and two journal articles [8] [9]

---

[6] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano “A multi-resolution approach to sound event detection in DCASE 2020 task4” *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

[7] D. de Benito-Gorrón et al. “Multiple Feature Resolutions for Different Polyphonic Sound Detection Score Scenarios in DCASE 2021 Task 4” *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021.

[8] 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.

[9] D. de Benito-Gorrón, D. Ramos, and D. T. Toledano “An Analysis of Sound Event Detection under Acoustic Degradation Using Multi-Resolution Systems” *Applied Sciences*, 2021.

Multi-resolution for Neural Network-based SED in domestic environments

Detection and Classification of Acoustic Scenes and Events 2020 — 2–3 November 2020, Tokyo, Japan

A MULTI-RESOLUTION APPROACH TO SOUND EVENT DETECTION IN DCASE 2020 TASK4

Diego de Benito-Gorron, Daniel Ramos, Doroteo T. Toledano

AUDIAS Research Group
Universidad Autónoma de Madrid
Calle Francisco Tomás y Valiente, 11, 28049 Madrid, SPAIN
{diego.benito, daniel.ramos, doroteo.torre}@uam.es

ABSTRACT

In this paper, we propose a multi-resolution analysis for feature extraction in Sound Event Detection. Because of the specific temporal and spectral characteristics of the different acoustic events, we hypothesize that different time-frequency resolutions can be more appropriate to locate each sound category. We carry out our experiments using the DESED dataset in the context of the DCASE 2020 Task 4 challenge, where the combination of up to five different time-frequency resolutions via model fusion is able to outperform the baseline results. In addition, we propose class-specific thresholds for the F_1-score metric, further improving the results over the Validation and Public Evaluation sets.

***Index Terms*—** DCASE 2020 Task 4, CRNN, Mean Teacher, Multi-resolution, Model fusion, Threshold tuning, PSDS

1. INTRODUCTION

Sound Event Detection (SED) systems aim to determine the temporal locations of several categories of acoustic events in a given audio clip. In contrast with the usual single-resolution approach used to train these systems, we propose a multi-resolution analysis of the audio features (mel-spectrograms) in order to take advantage of the diverse temporal and spectral characteristics found in different sound events.

Event	N.	Mean	Std.
Alarm bell / ringing	587	1.10	1.43
Blender	370	2.36	2.04
Cat	731	1.11	0.81
Dishes	1123	0.61	0.49
Dog	824	0.92	0.93
Electric shaver / toothbrush	345	4.61	2.69
Frying	229	5.06	3.07
Running water	270	3.81	2.53
Speech	2760	1.13	0.82
Vacuum cleaner	343	5.87	3.28

Table 1: Number of examples and mean and standard deviation of their durations (in seconds) for each sound category in the Synthetic training set.

The Weakly-labeled, Unlabeled and Synthetic training sets are used to train the neural networks. 20% of the Synthetic training set is reserved for validation. The DESED Validation set is used to tune hyper-parameters and perform model selection. In addition, we provide results over the Public Evaluation set.

3. PROPOSED SOLUTIONS

Multi-resolution for Neural Network-based SED in domestic environments

Detection and Classification of Acoustic Scenes and Events 2021 15–19 November 2021, Online

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) Chal-

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.

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.

Multi-resolution for Neural Network-based SED in domestic environments

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2021, accepted June 2, 2021, date of publication June 14, 2021, date of current version June 28, 2021.
Digital Object Identifier 10.1109/ACCESS.2021.3088949

A Multi-Resolution CRNN-Based Approach for Semi-Supervised Sound Event Detection in DCASE 2020 Challenge

DIEGO DE BENITO-GORRÓN, DANIEL RAMOS, AND DOROTEO T. TOLEDANO
AUDIAS Research Group, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain
Corresponding author: Diego de Benito-Gorrón (diego.benito@uam.es)

This work was supported in part by the Project Deep Speech for Forensics and Security (DSForSec) under Grant RTI2018-098091-B-I00, in part by the Ministry of Science, Innovation and Universities of Spain, and in part by the European Regional Development Fund (ERDF).

ABSTRACT Sound Event Detection is a task with a rising relevance over the recent years in the field of audio signal processing, due to the creation of specific datasets such as Google AudioSet or DESED (Domestic Environment Sound Event Detection) and the introduction of competitive evaluations like the DCASE Challenge (Detection and Classification of Acoustic Scenes and Events). The different categories of acoustic events can present diverse temporal and spectral characteristics. However, most approaches use a fixed time-frequency resolution to represent the audio segments. This work proposes a multi-resolution analysis for feature extraction in Sound Event Detection, hypothesizing that different resolutions can be more adequate for the detection of different sound event categories, and that combining the information provided by multiple resolutions could improve the performance of Sound Event Detection systems. Experiments are carried out over the DESED dataset in the context of the DCASE 2020 Challenge, concluding that the combination of up to 5 resolutions allows a neural network-based system to obtain better results than single-resolution models in terms of event-based F1-score in every event category and in terms of PSDS (Polyphonic Sound Detection Score). Furthermore, we analyze the impact of score thresholding in the computation of F1-score results, finding that the standard value of 0.5 is suboptimal and proposing an alternative strategy based in the use of a specific threshold for each event category, which obtains further

Multi-resolution for Neural Network-based SED in domestic environments

applied sciences

MDPI

Article

An Analysis of Sound Event Detection under Acoustic Degradation Using Multi-Resolution Systems

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

AUDIAS, Electronic and Communication Technology Department, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Av. Francisco Tomás y Valiente, 11, 28049 Madrid, Spain; daniel.ramos@uam.es (D.R.); doroteo.torre@uam.es (D.T.T.)
* Correspondence: diego.benito@uam.es

Abstract: The Sound Event Detection task aims to determine the temporal locations of acoustic events in audio clips. In recent years, the relevance of this field is rising 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 analyze the performance of Sound Event Detection systems under diverse artificial acoustic conditions such as high- or low-pass filtering and clipping or dynamic range compression, as well as under an scenario of high overlap between events. For this purpose, the audio was 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 multiresolution 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 multiresolution approach in different acoustic settings, and compare the performance of the single-resolution systems in the aforementioned scenarios when using different resolutions. Furthermore, we complement the analysis of the performance in the high-overlap scenario by assessing the degree of overlap of each event category in sound event detection datasets.

Citation: de Benito-Gorrón, D.; Ramos, D.; Toledano, D.T. An Analysis of Sound Event Detection under Acoustic Degradation Using Multi-Resolution Systems. *Appl. Sci.* **2021**, *11*, 11561. https://doi.org/10.3390/app112311561

Academic Editors: António Joaquim

DCASE CHALLENGE

- **Task 4** — “Sound Event Detection in Domestic Environments”
- 10 target categories

🔔 Alarm bell/ringing	Blender
🐱 Cat	🍽️ Dishes
🐶 Dog	Electric shaver/toothbrush
🍳 Frying	🚿 Running water
💬 Speech	💨 Vacuum cleaner

- Strong SED $\rightarrow$ Time boundaries (t_{on}, t_{off}) are specified

DCASE (Detection and Classification of Acoustic Scenes and Events)

DCASE CHALLENGE

- **Task 4** — “Sound Event Detection in Domestic Environments”
- 10 target categories

🔔 Alarm bell/ringing	Blender
🐱 Cat	🍽️ Dishes
🐶 Dog	Electric shaver/toothbrush
🍳 Frying	🚿 Running water
💬 Speech	💨 Vacuum cleaner

- Strong SED $\rightarrow$ Time boundaries (t_{on}, t_{off}) are specified

DCASE Challenge Task 4 — Dataset

DESED dataset — Domestic Environment Sound Event Detection [10]

- 10-second audio clips
- “Real” recordings from Google AudioSet
- “Synthetic” recordings produced with the Scaper toolkit [11]
- Mix of weakly-labeled, strongly-labeled and unlabeled data

[10] N. Turpault et al. “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis” *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

[11] J. Salamon et al. “Scaper: A library for soundscape synthesis and augmentation” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

DCASE Challenge Task 4 — Dataset

DESED dataset — Domestic Environment Sound Event Detection [10]

- 10-second audio clips
- "Real" recordings from Google AudioSet
- "Synthetic" recordings produced with the Scaper toolkit [11]
- Mix of weakly-labeled, strongly-labeled and unlabeled data

10 N. Turpault et al. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

11 J. Salamon et al. "Scaper: A library for soundscape synthesis and augmentation" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

DCASE Challenge Task 4 — Dataset

DESED dataset — Domestic Environment Sound Event Detection [10]

- 10-second audio clips

- "Real" recordings from Google AudioSet
- "Synthetic" recordings produced with the Scaper toolkit [11]

- Mix of weakly-labeled, strongly-labeled and unlabeled data

---

10 N. Turpault et al. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

11 J. Salamon et al. "Scaper: A library for soundscape synthesis and augmentation" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

DESED dataset — Domestic Environment Sound Event Detection [12]

	Audio	Labels	Number of clips
Unlabeled Training set	Real	None	14412
Weak Training set	Real	Weak	1578
Synthetic Training set	Synthetic	Strong	12500
Validation set	Real	Strong	1168

[12] N. Turpault et al. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

DCASE Challenge Task 4 — Baseline system

DCASE 2020 Task 4 Baseline system [13]

- Competitive baseline, based on best-performing systems from previous editions
- Convolutional Recurrent Neural Network (CRNN)
- Mel-spectrogram features
- Mean Teacher method [14] for semi-supervised learning
- Additional Baseline system provided for Sound Event Separation and Detection (not considered during this section)

---

13 N. Turpault and R. Serizel "Training Sound Event Detection on a Heterogeneous Dataset" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

14 A. Tarvainen and H. Valpola "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results" *Advances in neural information processing systems*, 2017.

DCASE 2020 Task 4 Baseline system [13]

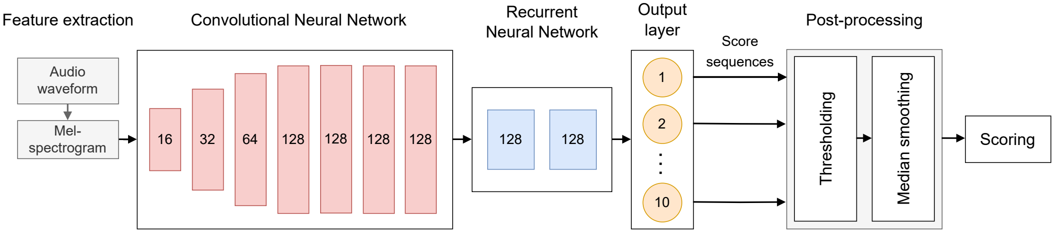
- Competitive baseline, based on best-performing systems from previous editions
- Convolutional Recurrent Neural Network (CRNN)
- Mel-spectrogram features
- Mean Teacher method [14] for semi-supervised learning
- Additional Baseline system provided for Sound Event Separation and Detection (not considered during this section)

---

13 N. Turpault and R. Serizel "Training Sound Event Detection on a Heterogeneous Dataset" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

14 A. Tarvainen and H. Valpola "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results" *Advances in neural information processing systems*, 2017.

DCASE Challenge Task 4 — Baseline system



Convolutional block: Conv 2D (3x3), ReLU activation, Average Pooling 2D

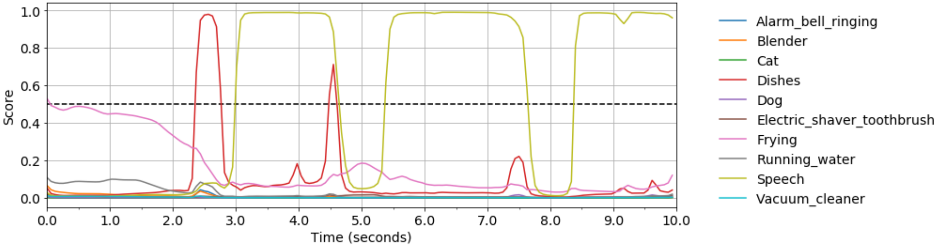
Recurrent block: Bidirectional Gated Recurrent Unit (GRU)

Output unit: Sigmoid activation

- **Training:** 90% of DESED Weak + 80% of DESED Synthetic + DESED Unlabeled
- **Validation:** 10% of DESED Weak + 20% of DESED Synthetic

1. A score sequence $\hat{\mathbf{d}}_k$ is obtained for each class k

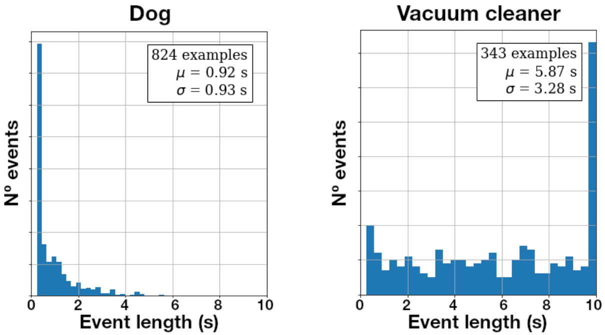
$$f^{(sed)}(\mathbf{x};\boldsymbol{\theta}_{sed}) = \hat{\mathbf{D}} = \langle \hat{\mathbf{d}}_k \rangle, 1 \leq k \leq K$$



2. Time boundaries are estimated from $\hat{\mathbf{d}}_k$ using a threshold $\tau \in (0,1)$
3. Median filtering to smooth boundary predictions

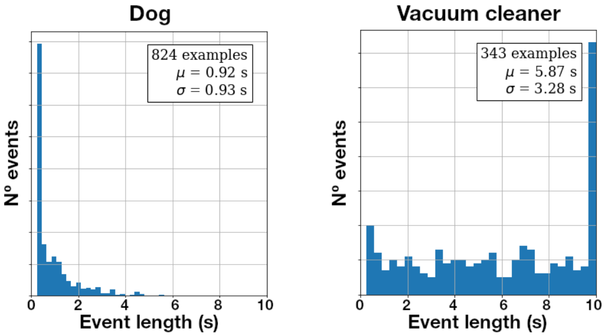
Multi-resolution for Neural Network-based Sound Event Detection

- Acoustic events present different temporal and spectral characteristics
  - E.g. different event lengths



- Mel-spectrogram $\rightarrow$ compromise between time resolution and frequency resolution

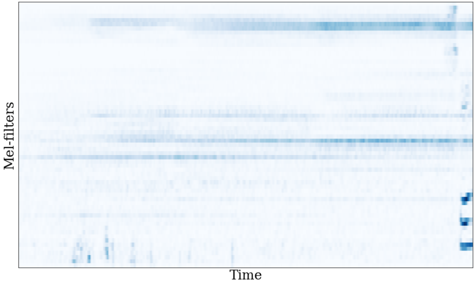
- Acoustic events present different temporal and spectral characteristics
  - E.g. different event lengths



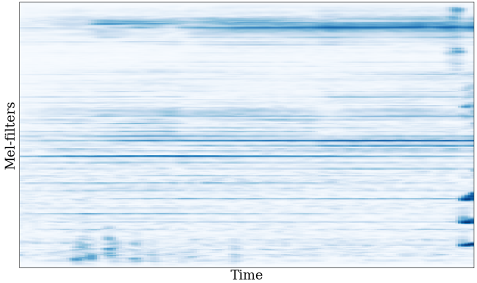
- Mel-spectrogram $\rightarrow$ compromise between time resolution and frequency resolution

Mel-spectrogram — **Electric shaver/toothbrush**

High time resolution

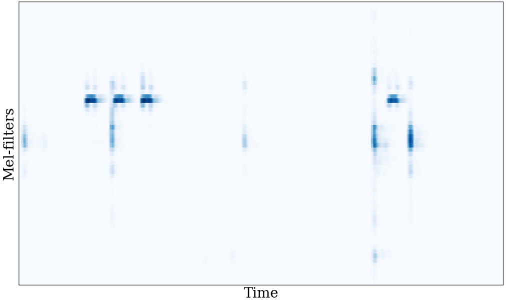


High frequency resolution

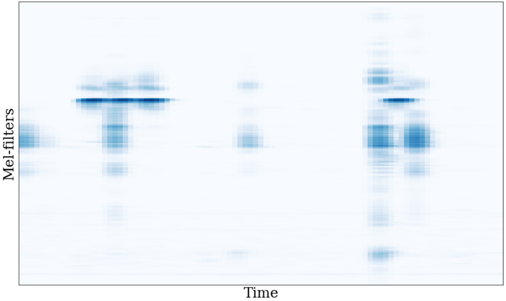


Mel-spectrogram — **Alarm bell/ringing**

High time resolution



High frequency resolution



Time-frequency resolution points

Defined as sets of parameters for mel-spectrogram feature extraction

- Audio sampling frequency (f_s)
- Size of the Discrete Fourier Transform (N)
- Window type, length (L) and hop size (R)
- Number of mel filters (n_{mel})

Mel-spectrogram extraction parameters of the Baseline System

- Sampling frequency: $f_s = 16000Hz$
- Size of the DFT: $N = 2048$ samples
- Hamming window
  - Length: $L = 2048$ samples (128 ms)
  - Hop size: $R = 255$ samples (15.94 ms)
- $n_{mel} = 128$ mel filters

Five resolution points are defined

- Baseline system resolution (BS) as starting point
- Twice better time resolution (T_{++}), and twice better frequency resolution (F_{++})
- Intermediate resolution points T_{+} and F_{+}

	$\mathbf{N}$	$\mathbf{L}$	$\mathbf{R}$	$\mathbf{n_{mel}}$
$\mathbf{T_{++}}$	1024	1024	128	64
$\mathbf{T_{+}}$	2048	1536	192	96
$\mathbf{BS}$	2048	2048	255	128
$\mathbf{F_{+}}$	4096	3072	384	192
$\mathbf{F_{++}}$	4096	4096	512	256

N, L and R reported in samples, using $f_s = 16000 Hz$

Multi-resolution approach

1. Train J single-resolution systems, with features extracted at each resolution point $j = 1 \ldots J$
2. For each event class k, combine the scores of the J single-resolution systems, $\hat{\mathrm{d}}_k^{(j)}$

$$\hat{\mathrm{d}}_k^{(multi)} = \frac{1}{J} \sum_{j=1}^{J} \hat{\mathrm{d}}_k^{(j)}$$

3. Obtain the predictions of onsets and offsets from $\hat{\mathrm{d}}_k^{(multi)}$ and measure the performance

Multi-resolution approach

1. Train J single-resolution systems, with features extracted at each resolution point $j = 1 \ldots J$
2. For each event class k, combine the scores of the J single-resolution systems, $\hat{\mathbf{d}}_k^{(j)}$

$$\hat{\mathbf{d}}_k^{(multi)} = \frac{1}{J}\sum_{j=1}^{J}\hat{\mathbf{d}}_k^{(j)}$$

3. Obtain the predictions of onsets and offsets from $\hat{\mathbf{d}}_k^{(multi)}$ and measure the performance

Multi-resolution approach

1. Train J single-resolution systems, with features extracted at each resolution point $j = 1 \dots J$
2. For each event class k, combine the scores of the J single-resolution systems, $\hat{\mathbf{d}}_k^{(j)}$

$$\hat{\mathbf{d}}_k^{(multi)} = \frac{1}{J} \sum_{j=1}^{J} \hat{\mathbf{d}}_k^{(j)}$$

3. Obtain the predictions of onsets and offsets from $\hat{\mathbf{d}}_k^{(multi)}$ and measure the performance

DCASE 2020 — Single-resolution results

	$\mathbf{T}_{++}$	$\mathbf{T}_{+}$	**BS**	$\mathbf{F}_{+}$	$\mathbf{F}_{++}$
Alarm bell/ringing	42.1	**43.8**	42.0	42.2	41.0
Blender	**32.9**	32.3	27.4	30.0	30.9
Cat	38.4	40.0	**41.0**	39.3	34.7
Dishes	20.8	21.9	20.8	**22.6**	21.0
Dog	15.1	**17.1**	16.5	12.3	12.8
Electric shaver/toothbrush	32.8	35.5	37.2	36.2	**41.1**
Frying	23.5	**23.9**	20.9	**23.9**	22.2
Running water	**31.7**	29.8	30.4	27.6	27.2
Speech	42.7	**47.1**	45.2	46.2	46.3
Vacuum cleaner	40.1	39.9	38.9	**44.5**	40.1
Macro F_1 score	32.0	**33.1**	32.0	32.5	31.7

Event-based F_1-scores (%) over **DESED Validation set**

Different resolutions perform better for different categories

DCASE 2020 — Single-resolution results

	$\mathbf{T}_{++}$	$\mathbf{T}_{+}$	$\mathbf{BS}$	$\mathbf{F}_{+}$	$\mathbf{F}_{++}$
Alarm bell/ringing	42.1	**43.8**	42.0	42.2	41.0
Blender	**32.9**	32.3	27.4	30.0	30.9
Cat	38.4	40.0	**41.0**	39.3	34.7
Dishes	20.8	21.9	20.8	**22.6**	21.0
Dog	15.1	**17.1**	16.5	12.3	12.8
Electric shaver/toothbrush	32.8	35.5	37.2	36.2	**41.1**
Frying	23.5	**23.9**	20.9	**23.9**	22.2
Running water	**31.7**	29.8	30.4	27.6	27.2
Speech	42.7	**47.1**	45.2	46.2	46.3
Vacuum cleaner	40.1	39.9	38.9	**44.5**	40.1
Macro F_1 score	32.0	**33.1**	32.0	32.5	31.7

Event-based F_1-scores (%) over **DESED Validation set**

Different resolutions perform better for different categories

DCASE 2020 — Multi-resolution results (Validation)

- ***Base***: Baseline System results
- ***3res***: Fusion of BS, T_{++} and F_{++} resolutions
- ***5res***: Fusion of all 5 resolution points
- ***5res-thr***: *5res* with adjusted thresholds **(optimistic performance)**

	Base	3res	5res	5res-thr
Alarm bell/ringing	*39.0*			
Blender	*31.6*			
Cat	*45.0*			
Dishes	*25.0*			
Dog	*21.7*			
Electric shaver/toothbrush	*36.0*			
Frying	*24.4*			
Running water	*31.7*			
Speech	*49.0*			
Vacuum cleaner	*44.4*			
Total macro	*34.8*			

Event-based F_1-scores (%) over **DESED Validation set**

DCASE 2020 — Multi-resolution results (Validation)

- ***Base***: Baseline System results
- ***3res***: Fusion of BS, T_{++} and F_{++} resolutions
- ***5res***: Fusion of all 5 resolution points
- ***5res-thr***: *5res* with adjusted thresholds **(optimistic performance)**

	Base	**3res**	**5res**	**5res-thr**
Alarm bell/ringing	*39.0*	46.1		
Blender	*31.6*	46.4		
Cat	*45.0*	42.2		
Dishes	*25.0*	22.1		
Dog	*21.7*	17.7		
Electric shaver/toothbrush	*36.0*	41.8		
Frying	*24.4*	30.0		
Running water	*31.7*	38.2		
Speech	*49.0*	48.0		
Vacuum cleaner	*44.4*	54.8		
Total macro	*34.8*	38.7		

Event-based F_1-scores (%) over **DESED Validation set**

DCASE 2020 — Multi-resolution results (Validation)

- ***Base***: Baseline System results
- ***3res***: Fusion of BS, T_{++} and F_{++} resolutions
- ***5res***: Fusion of all 5 resolution points
- ***5res-thr***: *5res* with adjusted thresholds **(optimistic performance)**

	Base	**3res**	**5res**	**5res-thr**
Alarm bell/ringing	*39.0*	46.1	47.2	
Blender	*31.6*	46.4	49.5	
Cat	*45.0*	42.2	45.2	
Dishes	*25.0*	22.1	23.9	
Dog	*21.7*	17.7	18.6	
Electric shaver/toothbrush	*36.0*	41.8	46.8	
Frying	*24.4*	30.0	29.7	
Running water	*31.7*	38.2	39.6	
Speech	*49.0*	48.0	49.9	
Vacuum cleaner	*44.4*	54.8	58.7	
Total macro	*34.8*	38.7	40.9	

Event-based F_1-scores (%) over **DESED Validation set**

DCASE 2020 — Multi-resolution results (Validation)

- ***Base***: Baseline System results
- ***3res***: Fusion of BS, T_{++} and F_{++} resolutions
- ***5res***: Fusion of all 5 resolution points
- ***5res-thr***: *5res* with adjusted thresholds **(optimistic performance)**

	Base	**3res**	**5res**	**5res-thr**
Alarm bell/ringing	*39.0*	46.1	47.2	**48.2**
Blender	*31.6*	46.4	49.5	**50.0**
Cat	*45.0*	42.2	45.2	**47.3**
Dishes	*25.0*	22.1	23.9	**25.2**
Dog	*21.7*	17.7	18.6	**22.3**
Electric shaver/toothbrush	*36.0*	41.8	46.8	**49.0**
Frying	*24.4*	30.0	29.7	**34.3**
Running water	*31.7*	38.2	39.6	**41.6**
Speech	*49.0*	48.0	49.9	**55.6**
Vacuum cleaner	*44.4*	54.8	58.7	**61.0**
Total macro	*34.8*	38.7	40.9	**43.4**

Event-based F_1-scores (%) over **DESED Validation set**

DCASE 2020 — Multi-resolution results (Evaluation)

- *5res* and *5res-thr* were submitted to DCASE 2020 Task 4
- Both systems outperformed the baseline [15]

	Base	**5res**	**5res-thr**
Alarm bell/ringing	35.9	**40.3**	38.5
Blender	37.0	**42.4**	42.2
Cat	62.6	61.5	**63.1**
Dishes	**26.0**	20.8	22.3
Dog	**27.1**	14.5	21.5
Electric shaver/toothbrush	25.9	**40.9**	36.8
Frying	24.7	28.5	**30.8**
Running water	**24.3**	**24.3**	23.5
Speech	48.2	48.4	**54.0**
Vacuum cleaner	39.0	**60.4**	51.5
Total macro	34.9	37.9	**38.2**

Event-based F_1-scores (%) over **DESED Evaluation set**

[15] http://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments-results

Multi-resolution Sound Event Detection — Overlap analysis

- Overlapped events are particularly difficult for SED systems
- Multi-resolution seems to slightly improve robustness in such scenario [16]
  - Higher Relative improvement (R.I.) of the Recall metric

System	Non-overlapped Recall%	Non-overlapped R.I.%	Overlapped Recall%	Overlapped R.I.%
BS	36.4	-	10.6	-
3res	39.4	8.4	13.5	**27.8**
5res	40.9	12.6	14.8	**40.0**

Results over **DESED Validation set**

- However, conclusions were limited by the scarcity of overlapped data

[16] 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.

Multi-resolution Sound Event Detection — Overlap analysis

- Overlapped events are particularly difficult for SED systems
- Multi-resolution seems to slightly improve robustness in such scenario [16]
  - Higher Relative improvement (R.I.) of the Recall metric

System	Non-overlapped Recall%	R.I.%	Overlapped Recall%	R.I.%
BS	36.4	-	10.6	-
3res	39.4	8.4	13.5	**27.8**
5res	40.9	12.6	14.8	**40.0**

Results over **DESED Validation set**

- However, conclusions were limited by the scarcity of overlapped data

16 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.

Multi-resolution Sound Event Detection — Overlap analysis

- Overlapped events are particularly difficult for SED systems
- Multi-resolution seems to slightly improve robustness in such scenario [16]
  - Higher Relative improvement (R.I.) of the Recall metric

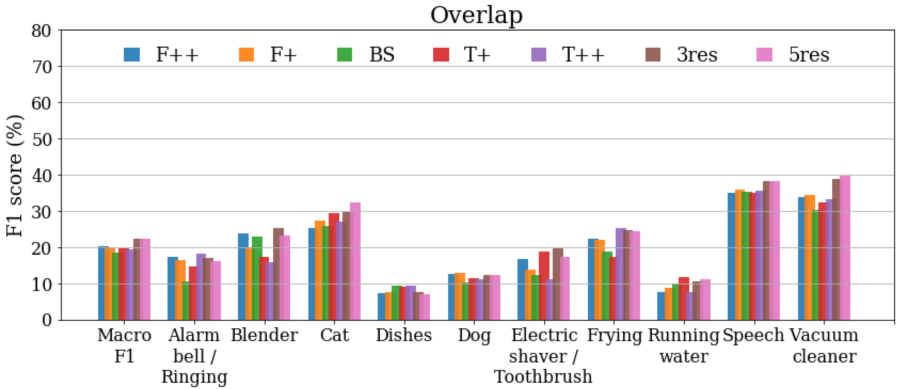
System	Non-overlapped		Overlapped	
	Recall%	R.I.%	Recall%	R.I.%
BS	36.4	-	10.6	-
3res	39.4	8.4	13.5	**27.8**
5res	40.9	12.6	14.8	**40.0**

Results over **DESED Validation set**

- However, conclusions were limited by the scarcity of overlapped data

[16] 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.

- Design of an Overlapped dataset to analyze performance under severe event overlap [17]



17 D. de Benito-Gorrón, D. Ramos, and D. T. Toledano “An Analysis of Sound Event Detection under Acoustic Degradation Using Multi-Resolution Systems” *Applied Sciences*, 2021.

Polyphonic Sound Detection Score [18]
Aims to overcome limitations of F_1 score

- Several operation points considered $\rightarrow$ Area Under Curve (AUC)
  - 50 thresholds, linearly distributed from 0 to 1
- Intersection criterion to enhance robustness
- Adaptable to different application scenarios
  - PSDS1 $\rightarrow$ Accurate temporal detection
  - PSDS2 $\rightarrow$ Accurate classification between events

[18] Ç. Bilen et al. "A framework for the robust evaluation of sound event detection" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Polyphonic Sound Detection Score [18]

Aims to overcome limitations of F_1 score

- Several operation points considered $\rightarrow$ Area Under Curve (AUC)
  - 50 thresholds, linearly distributed from 0 to 1
- Intersection criterion to enhance robustness
- Adaptable to different application scenarios
  - PSDS1 $\rightarrow$ Accurate temporal detection
  - PSDS2 $\rightarrow$ Accurate classification between events

[18] Ç. Bilen et al. "A framework for the robust evaluation of sound event detection" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Polyphonic Sound Detection Score [18]
Aims to overcome limitations of F_1 score

- Several operation points considered $\rightarrow$ Area Under Curve (AUC)
    - 50 thresholds, linearly distributed from 0 to 1
- Intersection criterion to enhance robustness
- Adaptable to different application scenarios
    - PSDS1 $\rightarrow$ Accurate temporal detection
    - PSDS2 $\rightarrow$ Accurate classification between events

[18] Ç. Bilen et al. "A framework for the robust evaluation of sound event detection" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Polyphonic Sound Detection Score [18]

Aims to overcome limitations of F_1 score

- Several operation points considered $\rightarrow$ Area Under Curve (AUC)
  - 50 thresholds, linearly distributed from 0 to 1
- Intersection criterion to enhance robustness
- Adaptable to different application scenarios
  - PSDS1 $\rightarrow$ Accurate temporal detection
  - PSDS2 $\rightarrow$ Accurate classification between events

[18] Ç. Bilen et al. "A framework for the robust evaluation of sound event detection" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

Polyphonic Sound Detection Score [18]

Aims to overcome limitations of F_1 score

- Several operation points considered $\rightarrow$ Area Under Curve (AUC)
  - 50 thresholds, linearly distributed from 0 to 1
- Intersection criterion to enhance robustness
- Adaptable to different application scenarios
  - PSDS1 $\rightarrow$ Accurate temporal detection
  - PSDS2 $\rightarrow$ Accurate classification between events

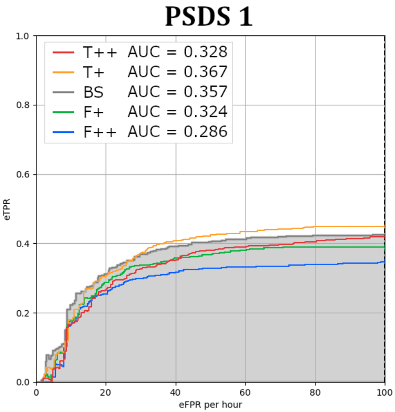
[18] Ç. Bilen et al. "A framework for the robust evaluation of sound event detection" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

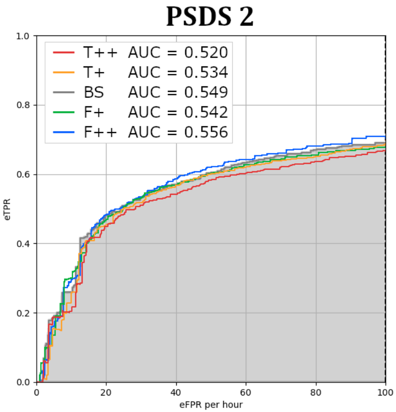
Polyphonic Sound Detection Score [18]

Aims to overcome limitations of F_1 score

- Several operation points considered $\rightarrow$ Area Under Curve (AUC)
  - 50 thresholds, linearly distributed from 0 to 1
- Intersection criterion to enhance robustness
- Adaptable to different application scenarios
  - PSDS1 $\rightarrow$ Accurate temporal detection
  - PSDS2 $\rightarrow$ Accurate classification between events

[18] Ç. Bilen et al. "A framework for the robust evaluation of sound event detection" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.





Results over **DESED Validation set** [19]

19 D. de Benito-Gorrón et al. “Multiple Feature Resolutions for Different Polyphonic Sound Detection Score Scenarios in DCASE 2021 Task 4” *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021.

DCASE 2021 — Multi-resolution results

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

Results over the **DESED Evaluation set**

DCASE 2022 — Multi-resolution CRNN and Conformer

- Multi-resolution analysis applied to Conformer networks [20] (in addition to CRNN)
- Optimization of Mean Teacher model selection strategy [21]

	Resolutions	CRNN PSDS1	CRNN PSDS2	CRNN F_1 (%)	Conformer PSDS1	Conformer PSDS2	Conformer F_1 (%)
3res	T_+, BS, F_+	0.398	0.606	45.8	0.346	0.636	42.6
3res-T	T_{++}, T_+, BS	**0.416**	0.613	47.5	**0.371**	0.633	42.8
4res-T	T_{++}, T_+, BS, F_+	0.414	0.619	**47.8**	0.370	0.647	43.4
5res	T_{++}, T_+, BS, F_+, F_{++}	0.402	**0.625**	47.5	0.366	**0.657**	**44.3**
BS	BS	0.370	0.571	43.5	0.342	0.580	41.9

Results over **DESED Validation set**

[20] A. Gulati et al. *Conformer: Convolution-augmented Transformer for Speech Recognition* arXiv: 2005.08100 (eess.AS), 2020.

[21] D. de Benito-Gorrón et al. *Multi-Resolution Combination of CRNN and Conformers for DCASE 2022 Task 4* tech. rep., 2022.

DCASE 2022 — Multi-resolution CRNN and Conformer

- Multi-resolution analysis applied to Conformer networks [20] (in addition to CRNN)
- Optimization of Mean Teacher model selection strategy [21]

		CRNN			Conformer		
	Resolutions	PSDS1	PSDS2	F_1 (%)	PSDS1	PSDS2	F_1 (%)
3res	T_+, BS, F_+	0.398	0.606	45.8	0.346	0.636	42.6
3res-T	T_{++}, T_+, BS	**0.416**	0.613	47.5	**0.371**	0.633	42.8
4res-T	T_{++}, T_+, BS, F_+	0.414	0.619	**47.8**	0.370	0.647	43.4
5res	T_{++}, T_+, BS, F_+, F_{++}	0.402	**0.625**	47.5	0.366	**0.657**	**44.3**
BS	BS	0.370	0.571	43.5	0.342	0.580	41.9

Results over **DESED Validation set**

[20] A. Gulati et al. *Conformer: Convolution-augmented Transformer for Speech Recognition* arXiv: 2005.08100 (eess.AS), 2020.

[21] D. de Benito-Gorrón et al. *Multi-Resolution Combination of CRNN and Conformers for DCASE 2022 Task 4* tech. rep., 2022.

Multi-resolution CRNN + Conformer [22]

System	Median filtering	PSDS1	PSDS2	$F_1\,(\%)$
7res	Fixed	0.422	0.656	49.2
	Class-wise	**0.428**	0.655	**50.1**
10res	Fixed	0.410	**0.665**	48.9
	Class-wise	0.347	0.663	43.0
Baseline		*0.342*	*0.527*	*40.1*

Results over **DESED Validation set**

22 D. de Benito-Gorrón et al. *Multi-Resolution Combination of CRNN and Conformers for DCASE 2022 Task 4* tech. rep., 2022.

DCASE — Summary of Results

Team	Year	Rank	DESED Validation			DESED Evaluation		
			PSDS1	PSDS2	$F_1\,(\%)$	PSDS1	PSDS2	$F_1\,(\%)$
AUDIAS (de Benito et al.)	2020	13/19	—	—	43.4	—	—	38.2
	2021	9/24	0.386	0.600	46.4	0.363	0.577	43.1
	2022	**7/22**	**0.428**	**0.655**	**50.1**	**0.432**	**0.649**	**46.5**
Baseline systems	2020	17/19	—	—	34.8	—	—	34.9
	2021	14/24	0.342	0.527	40.1	0.315	0.547	37.3
	2022	17/22	0.342	0.527	40.1	0.315	0.543	37.3

Joint Training of Source Separation and Sound Event Detection in domestic environments

Joint Training of Source Separation and SED in domestic environments

- Work derived from the research stay at Brno University of Technology
- Application of Source Separation models as auxiliary modules for Sound Event Detection
- The work led to the publication of a conference paper [23] and the elaboration of a journal article, currently under review [24]

[23] D. de Benito-Gorrón, K. Zmolikova, and D. T. Toledano "Source Separation for Sound Event Detection in Domestic Environments using Jointly Trained Models" *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022.

[24] D. de Benito-Gorrón, K. Zmolikova, and D. T. Toledano "Analysis and Interpretation of Joint Source Separation and Sound Event Detection in Domestic Environments" *Submitted to PLOS ONE*, 2023.

Joint Training of Source Separation and SED in domestic environments

SOURCE SEPARATION FOR SOUND EVENT DETECTION IN DOMESTIC ENVIRONMENTS USING JOINTLY TRAINED MODELS

Diego de Benito-Gorrón[1], *Katerina Zmolikova*[2], *Doroteo T. Toledano*[1]

[1]AUDIAS Research Group, Escuela Politécnica Superior, Universidad Autónoma de Madrid
[2]Brno University of Technology, Faculty of IT, IT4I Centre of Excellence

IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914755

ABSTRACT

Sound Event Detection and Source Separation are closely related tasks: whereas the first aims to find the time boundaries of acoustic events inside a recording, the goal of the latter is to isolate each of the acoustic sources into different signals. This paper presents a Sound Event Detection system formed by two independently pre-trained blocks for Source Separation and Sound Event Detection. We propose a joint-training scheme, where both blocks are trained at the same time, and a two-stage training, where each block trains while the other one is frozen. In addition, we compare the use of supervised and unsupervised pre-training for the Separation block, and two model selection strategies for Sound Event Detection. Our experiments show that the proposed methods are able to outperform the baseline systems of the DCASE 2021 Challenge Task 4.

Index Terms— Sound Event Detection, Source Separation, DCASE, DESED

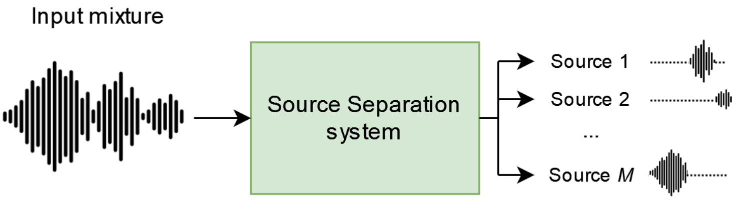
1. INTRODUCTION

An important amount of information about our surrounding environment is provided by sounds. The human ability to recognize them generally gives us an immediate idea of where we are, or what is happening near us. In computational intelligence, several research fields try to automatically retrieve this kind of information from

type (speech, music, background noise, etc.) [7]. Considering a training dataset of audio mixtures for which the original sources are available, deep neural networks can be trained for Source Separation in a classic supervised scheme, using Permutation Invariant Training (PIT) [8]. Moreover, an unsupervised training method for Source Separation called Mixture Invariant Training (MixIT) has recently been introduced [9], allowing to train Source Separation systems when the original sources of the training mixtures are not available.

Recent research has suggested the idea that Source Separation and Sound Event Detection tasks can benefit from each other, for instance, using the predictions of a SED system to guide the separation of events into different sources [10, 11, 12], learning SSep as an intermediate representation for SED [13], or applying SSep as a pre-processing step for SED, either fine-tuning the SED system over automatically separated data [14] or training a SSep network as a front-end stage to a pre-trained SED system [15].

In this paper, we propose a Sound Event Detection system composed of two pre-trained blocks: a Source Separation network and a Sound Event Detection network. In contrast with previous work, we do not limit the training to just one of the two blocks. Instead, aiming for both tasks to learn from each other, we introduce a joint training setting, where the whole system is trained in an end-to-end fashion, and a two-stage training, in which the SED block is fine-tuned first while freezing the SSep block (Stage 1), and then the SSep block is fine-tuned while freezing the SED block (Stage 2). Apart from these training settings and their analysis, our experimental results provide



Source Separation

- Decompose an audio mixture $\mathbf{x}$ into M channels, each one containing a different acoustic source or different types of sounds

Joint Training of Source Separation and SED in domestic environments

Research aims

- Use Source Separation (SSep) to enhance Sound Event Detection (SED)
- Design a model in which both stages (SSep and SED) are trained together
  - Potential mutual benefits between both tasks
- Evaluate in the context of DCASE Challenges
  - PSDS and F_1 score
  - Comparison with DCASE Baseline systems (SED, SSep+SED)

Joint Training of Source Separation and SED in domestic environments

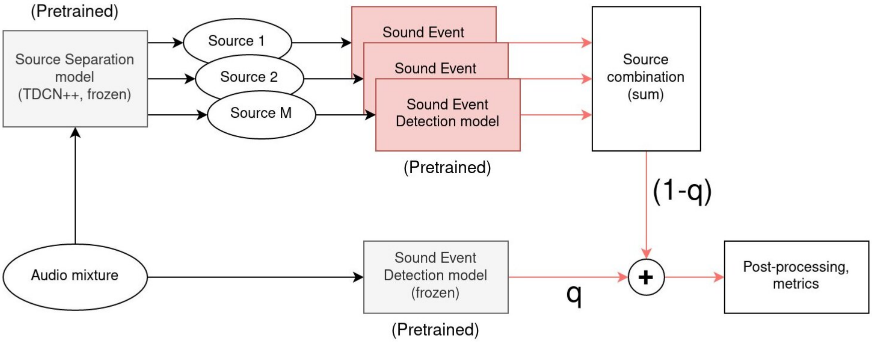
Research aims

- Use Source Separation (SSep) to enhance Sound Event Detection (SED)
- Design a model in which both stages (SSep and SED) are trained together
  - Potential mutual benefits between both tasks
- Evaluate in the context of DCASE Challenges
  - PSDS and F_1 score
  - Comparison with DCASE Baseline systems (SED, SSep+SED)

Research aims

- Use Source Separation (SSep) to enhance Sound Event Detection (SED)
- Design a model in which both stages (SSep and SED) are trained together
  - Potential mutual benefits between both tasks
- Evaluate in the context of DCASE Challenges
  - PSDS and F_1 score
  - Comparison with DCASE Baseline systems (SED, SSep+SED)

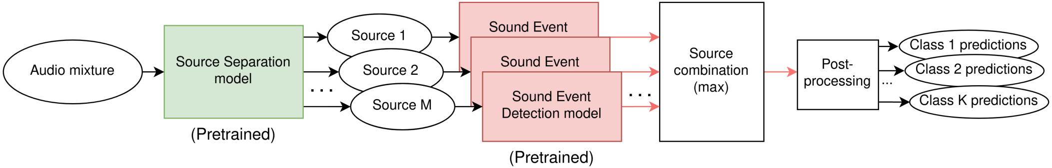
DCASE Baseline system — Sound Event Separation and Detection



- **SSep block** — Improved Time-Domain Convolutional Network (TDCN++) [25]
- **SED block** — CRNN (DCASE SED Baseline)

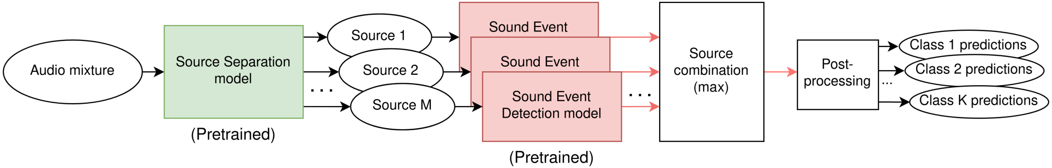
[25] I. Kavalerov, S. Wisdom, et al. "Universal Sound Separation" *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

Joint Source Separation + Sound Event Detection (JSS)



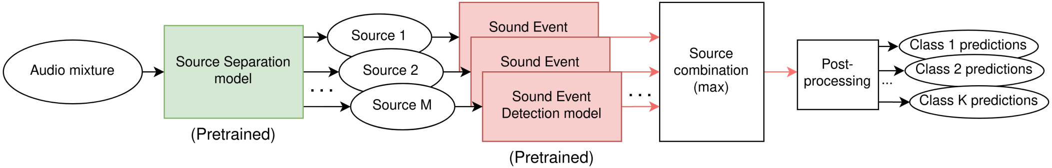
1. Input audio is fed to the **SSep block** $\rightarrow M$ waveforms with estimated sources
2. **SED block** is applied to each source $\rightarrow M$ source-level SED predictions, $\hat{\mathrm{D}}_{1\ldots M}^{(src)}$
3. **Max-pooling** applied to source-level predictions $\rightarrow$ clip-level predictions, $\hat{\mathrm{D}}$

Joint Source Separation + Sound Event Detection (JSS)



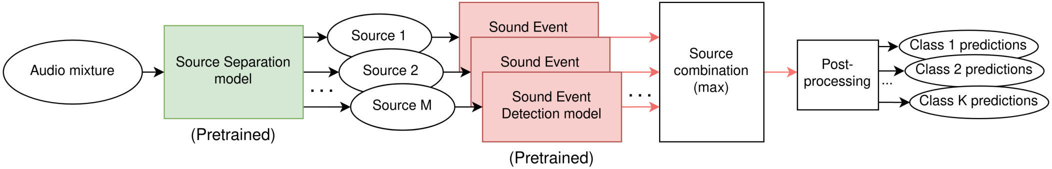
1. Input audio is fed to the **SSep block** $\rightarrow M$ waveforms with estimated sources
2. **SED block** is applied to each source $\rightarrow M$ source-level SED predictions, $\hat{\mathrm{D}}^{(src)}_{1\ldots M}$
3. **Max-pooling** applied to source-level predictions $\rightarrow$ clip-level predictions, $\hat{\mathrm{D}}$

Joint Source Separation + Sound Event Detection (JSS)



1. Input audio is fed to the **SSep block** $\rightarrow M$ waveforms with estimated sources
2. **SED block** is applied to each source $\rightarrow M$ source-level SED predictions, $\hat{\mathbf{D}}^{(src)}_{1 \dots M}$
3. **Max-pooling** applied to source-level predictions $\rightarrow$ clip-level predictions, $\hat{\mathbf{D}}$

Joint Source Separation + Sound Event Detection (JSS)

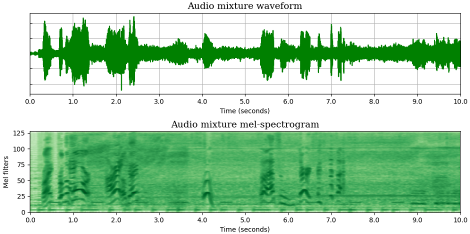


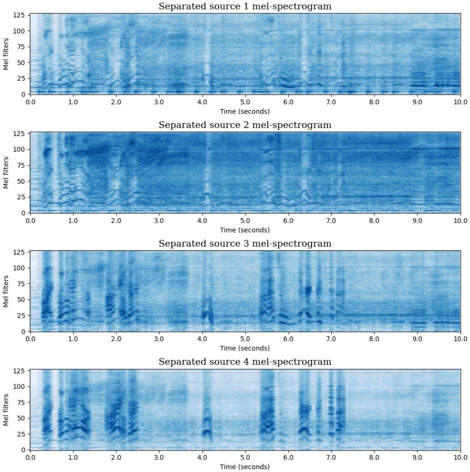
- **SSep block** — Conv-TasNet [26]
- **SED block** — CRNN (DCASE SED Baseline)
- Both blocks are pre-trained for their respective tasks

[26] Y. Luo and N. Mesgarani "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation" *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 2019.

Joint Source Separation + Sound Event Detection

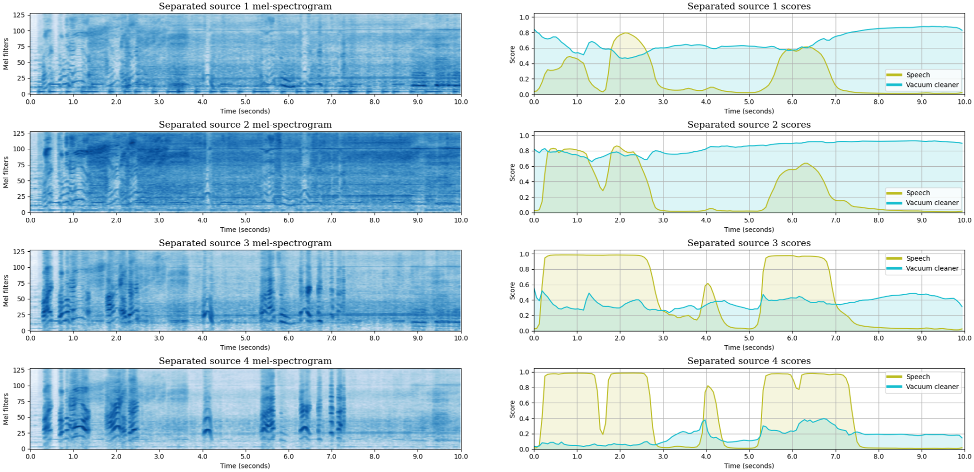
1. Source Separation block → 2. Sound Event Detection block → 3. Source combination





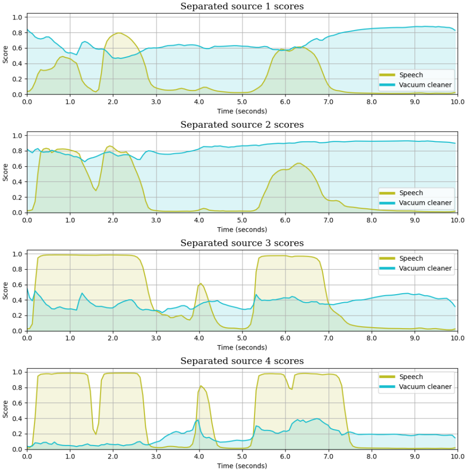
Joint Source Separation + Sound Event Detection

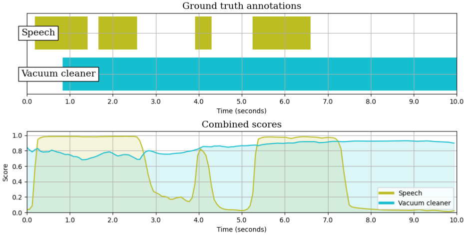
1. Source Separation block → **2. Sound Event Detection block** → 3. Source combination



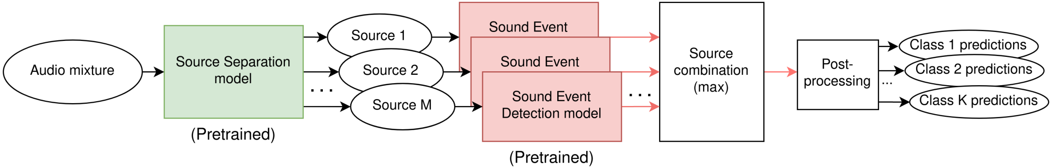
Joint Source Separation + Sound Event Detection

1. Source Separation block $\rightarrow$ 2. Sound Event Detection block $\rightarrow$ **3. Source combination**



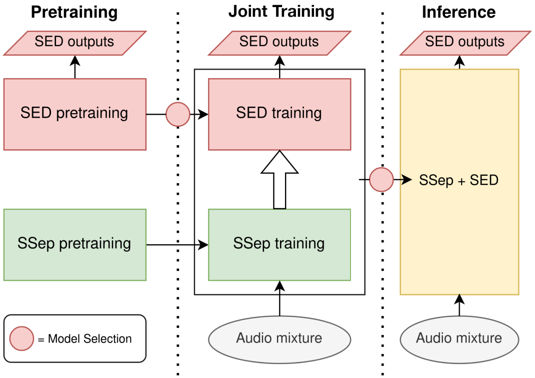


Joint Source Separation + Sound Event Detection (JSS)



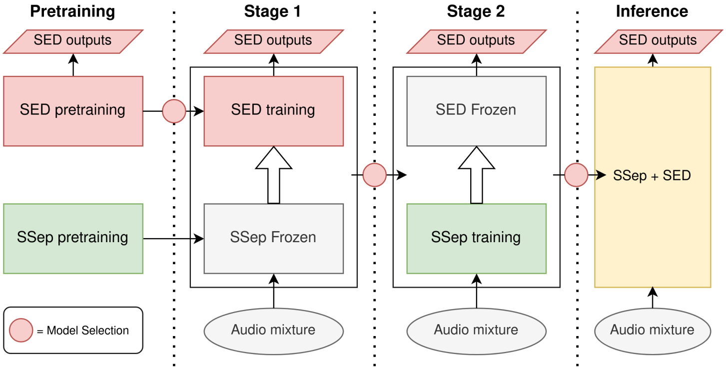
JSS training

- Mean teacher with SED loss function L_{sed}
- Two methods for JSS training
  - (a) Joint Training
  - (b) Two-stage Training



- **Joint Training** optimizes SED and SSep blocks together to minimize L_{sed}

JSS — Two-stage Training



- **Stage 1** updates SED block only, fine-tuning SED on separated data
- **Stage 2** updates SSep block only, back-propagating L_{sed} through the SED block

JSS training methods

1. (a) Joint Training
2. (b) Two-stage Training

Source Separation pre-training

1. (a) Supervised — Permutation Invariant Training (PIT) [27]
2. (b) Unsupervised — Mixture Invariant Training (MixIT) [28]

Mean Teacher model selection

1. (a) Student models (standard in DCASE Baseline systems)
2. (b) Teacher models (proposed)

---

[27] D. Yu et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation" *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[28] S. Wisdom, E. Tzinis, et al. "Unsupervised sound separation using mixture invariant training" *Advances in Neural Information Processing Systems*, 2020.

JSS —Proposed experiments

JSS training methods

- (a) Joint Training
- (b) Two-stage Training

Source Separation pre-training

- (a) Supervised — Permutation Invariant Training (PIT) [27]
- (b) Unsupervised — Mixture Invariant Training (MixIT) [28]

Mean Teacher model selection

- (a) Student models (standard in DCASE Baseline systems)
- (b) Teacher models (proposed)

---

[27] D. Yu et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation" *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[28] S. Wisdom, E. Tzinis, et al. "Unsupervised sound separation using mixture invariant training" *Advances in Neural Information Processing Systems*, 2020.

JSS training methods

a. Joint Training
b. Two-stage Training

Source Separation pre-training

a. Supervised — Permutation Invariant Training (PIT) [27]
b. Unsupervised — Mixture Invariant Training (MixIT) [28]

Mean Teacher model selection

a. Student models (standard in DCASE Baseline systems)
b. Teacher models (proposed)

[27] D. Yu et al. “Permutation invariant training of deep models for speaker-independent multi-talker speech separation” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[28] S. Wisdom, E. Tzinis, et al. “Unsupervised sound separation using mixture invariant training” *Advances in Neural Information Processing Systems*, 2020.

DESED (Domestic Environment Sound Event Detection) [29]

- Used for semi-supervised SED/JSS training and unsupervised SSep pre-training

FUSS (Free Universal Sound Separation) [30]

- Synthetic audio mixtures from 2 to 4 sources
- 20000 mixtures for training, 1000 for validation, 1000 for evaluation
- Used for supervised SSep pre-training

YFCC100M (Yahoo-Flickr Creative Commons 100 Million) [31]

- 0.8 million audio clips from web video (no oracle sources available)
- Used only by the DCASE SSep+SED Baseline system (unsupervised SSep pre-training)

29 N. Turpault et al. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

30 S. Wisdom, H. Erdogan, et al. "What's all the FUSS about Free Universal Sound Separation Data?" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

31 B. Thomee, D. A. Shamma, et al. "YFCC100M: The New Data in Multimedia Research" *Commun. ACM*, 2016.

JSS — Datasets

DESED (Domestic Environment Sound Event Detection) [29]

- Used for semi-supervised SED/JSS training and unsupervised SSep pre-training

FUSS (Free Universal Sound Separation) [30]

- Synthetic audio mixtures from 2 to 4 sources
- 20000 mixtures for training, 1000 for validation, 1000 for evaluation
- Used for supervised SSep pre-training

YFCC100M (Yahoo-Flickr Creative Commons 100 Million) [31]

- 0.8 million audio clips from web video (no oracle sources available)
- Used only by the DCASE SSep+SED Baseline system (unsupervised SSep pre-training)

---

29 N. Turpault et al. “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis” *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

30 S. Wisdom, H. Erdogan, et al. “What’s all the FUSS about Free Universal Sound Separation Data?” *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

31 B. Thomee, D. A. Shamma, et al. “YFCC100M: The New Data in Multimedia Research” *Commun. ACM*, 2016.

DESED (Domestic Environment Sound Event Detection) [29]

- Used for semi-supervised SED/JSS training and unsupervised SSep pre-training

FUSS (Free Universal Sound Separation) [30]

- Synthetic audio mixtures from 2 to 4 sources
- 20000 mixtures for training, 1000 for validation, 1000 for evaluation
- Used for supervised SSep pre-training

YFCC100M (Yahoo-Flickr Creative Commons 100 Million) [31]

- 0.8 million audio clips from web video (no oracle sources available)
- Used only by the DCASE SSep+SED Baseline system (unsupervised SSep pre-training)

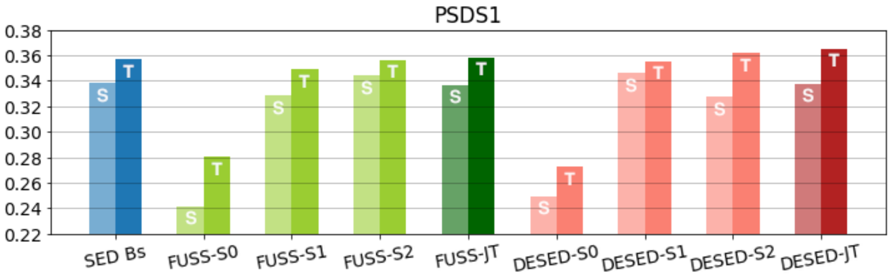
---

29 N. Turpault et al. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019.

30 S. Wisdom, H. Erdogan, et al. "What's all the FUSS about Free Universal Sound Separation Data?" *IEEE International Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.

31 B. Thomee, D. A. Shamma, et al. "YFCC100M: The New Data in Multimedia Research" *Commun. ACM*, 2016.

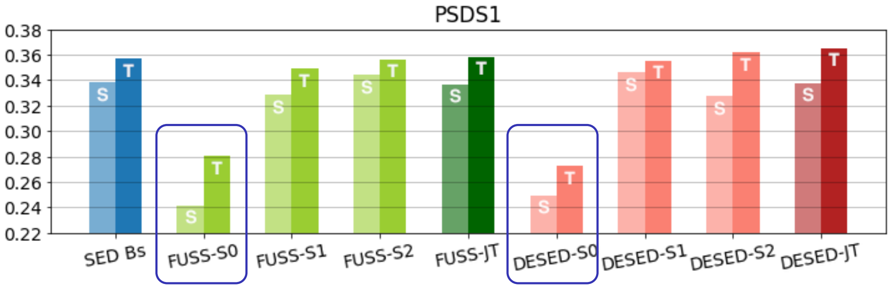
JSS — PSDS1 results



PSDS1 results over DESED Validation set

- Stage 0 (S0): Initial state (pre-training only) → Domain mismatch
- Stage 1 (S1): SED block training → Closer to SED baseline
- Stage 2 (S2): SSep block training → Further improvement
- Joint Training (JT) → Similar to Stage 2

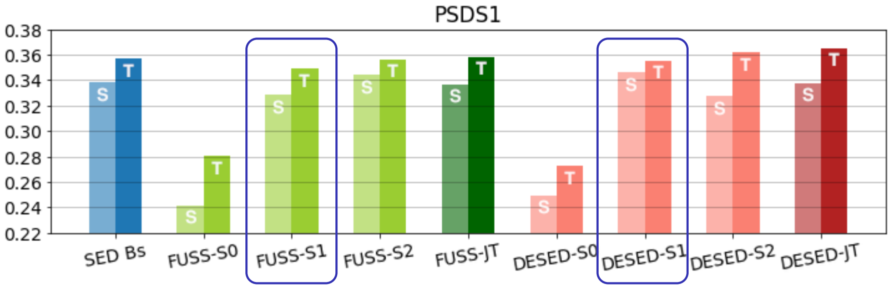
JSS — PSDS1 results



PSDS1 results over DESED Validation set

- Stage 0 (S0): Initial state (pre-training only) $\rightarrow$ Domain mismatch
- Stage 1 (S1): SED block training $\rightarrow$ Closer to SED baseline
- Stage 2 (S2): SSep block training $\rightarrow$ Further improvement
- Joint Training (JT) $\rightarrow$ Similar to Stage 2

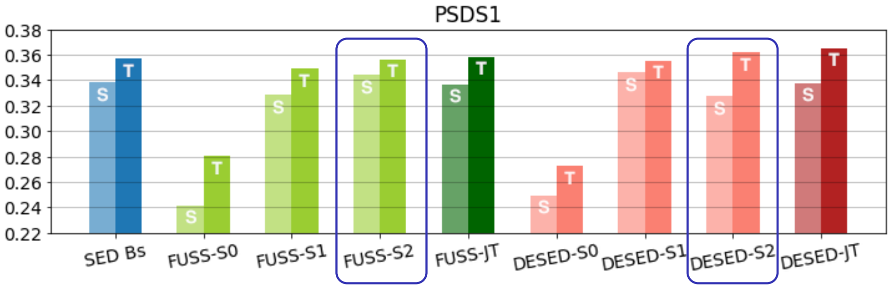
JSS — PSDS1 results



PSDS1 results over DESED Validation set

- Stage 0 (S0): Initial state (pre-training only) $\longrightarrow$ Domain mismatch
- Stage 1 (S1): SED block training $\longrightarrow$ Closer to SED baseline
- Stage 2 (S2): SSep block training $\longrightarrow$ Further improvement
- Joint Training (JT) $\longrightarrow$ Similar to Stage 2

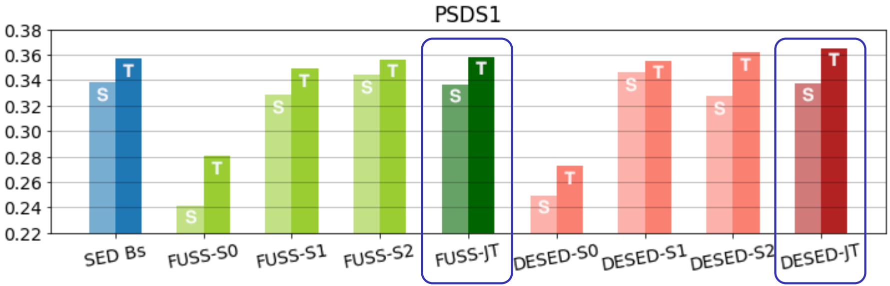
JSS — PSDS1 results



PSDS1 results over DESED Validation set

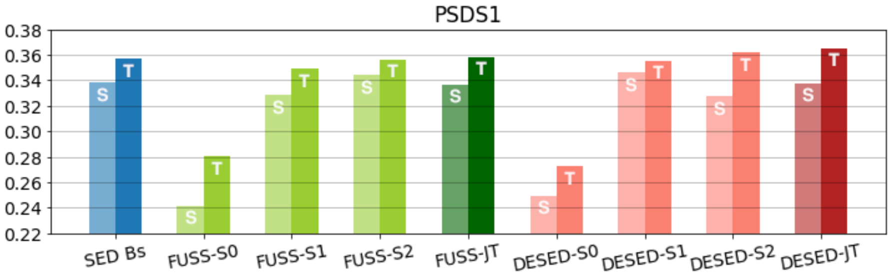
- Stage 0 (S0): Initial state (pre-training only) → Domain mismatch
- Stage 1 (S1): SED block training → Closer to SED baseline
- Stage 2 (S2): SSep block training → Further improvement
- Joint Training (JT) → Similar to Stage 2

JSS — PSDS1 results



PSDS1 results over DESED Validation set

- Stage 0 (S0): Initial state (pre-training only) $\rightarrow$ Domain mismatch
- Stage 1 (S1): SED block training $\rightarrow$ Closer to SED baseline
- Stage 2 (S2): SSep block training $\rightarrow$ Further improvement
- Joint Training (JT) $\rightarrow$ Similar to Stage 2



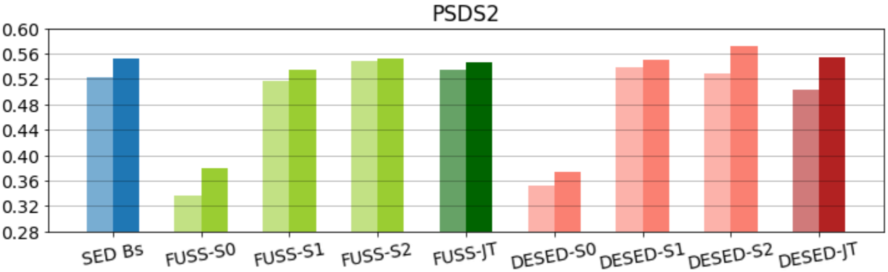
PSDS1 results over DESED Validation set

- Source Separation pre-training

FUSS (supervised, out-of-domain) **<** DESED (unsupervised, in-domain)

- Model selection

Student (bright bars) **<** Teacher (dark bars)

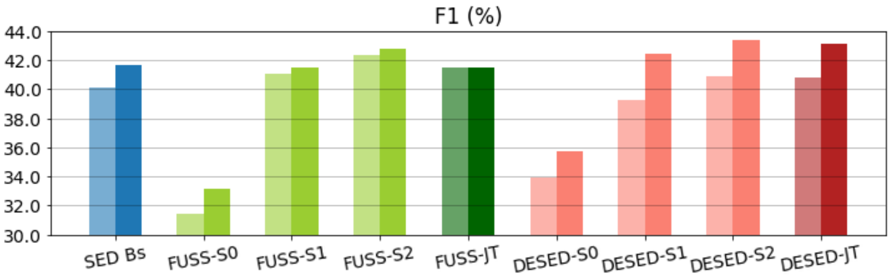


PSDS2 results over DESED Validation set

- Stage 0 → Domain mismatch
- Stage 1 → Closer to SED baseline
- Stage 2 → Further improvement
- Joint Training → Similar to Stage 2

- SSep pre-training: FUSS **<** DESED
- Model selection: Student **<** Teacher

JSS — F1-score results

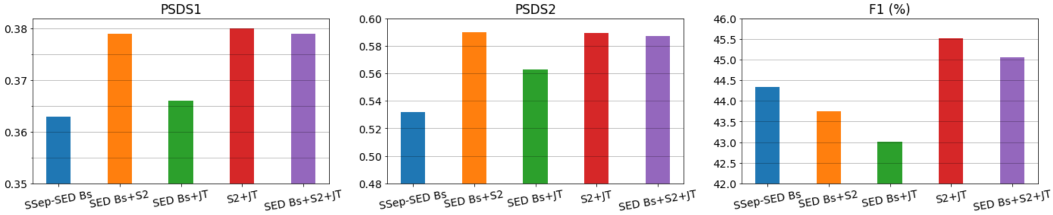


F1-score results over DESED Validation set

- Stage 0 $\rightarrow$ Domain mismatch
- Stage 1 $\rightarrow$ Closer to SED baseline
- Stage 2 $\rightarrow$ Further improvement
- Joint Training $\rightarrow$ Similar to Stage 2

- SSep pre-training: FUSS **<** DESED
- Model selection: Student **<** Teacher

JSS — Model fusion results



Results over **DESED Validation set** (Teacher model sel. and DESED SSep pre-training)

- Comparison with DCASE Sound Event Separation and Detection Baseline
- Score fusion obtained as an average of the network score sequences
- Best performance: Stage 2 + Joint Training (S2+JT)

Conclusions / Ongoing and future work

Conclusions (1/3)

- Three journal articles
  - D. de Benito-Gorrón et al. “Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2019.1 (2019), pp. 1–18
  - 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”. In: *IEEE Access* 9 (2021), pp. 89029–89042. DOI: 10.1109/ACCESS.2021.3088949
  - D. de Benito-Gorrón, D. Ramos, and D. T. Toledano. “An Analysis of Sound Event Detection under Acoustic Degradation Using Multi-Resolution Systems”. In: *Applied Sciences* 11.23 (2021). ISSN: 2076-3417. DOI: 10.3390/app112311561. URL: https://www.mdpi.com/2076-3417/11/23/11561
- Four conference papers
- Three competitive evaluations

Conclusions (1/3)

- Three journal articles
- Four conference papers
  - D. de Benito-Gorrón, D. Ramos, and D. T. Toledano. “A multi-resolution approach to sound event detection in DCASE 2020 task4”. In: *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. Tokyo, Japan, Nov. 2020, pp. 36–40
  - D. de Benito-Gorrón, D. Ramos, and D. T. Toledano. “An Analysis of Sound Event Detection under Acoustic Degradation Using Multi-Resolution Systems”. In: *IberSPEECH 2021*. Valladolid, Spain, Mar. 2021. DOI: 10.21437/IberSPEECH.2021-8
  - D. de Benito-Gorrón et al. “Multiple Feature Resolutions for Different Polyphonic Sound Detection Score Scenarios in DCASE 2021 Task 4”. In: *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. Barcelona, Spain, Nov. 2021, pp. 65–69
  - D. de Benito-Gorrón, K. Zmolikova, and D. T. Toledano. “Source Separation for Sound Event Detection in Domestic Environments using Jointly Trained Models”. In: *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. 2022, pp. 1–5. DOI: 10.1109/IWAENC53105.2022.9914755
- Three competitive evaluations

Conclusions (1/3)

- Three journal articles
- Four conference papers
- Three competitive evaluations
  - D. de Benito-Gorrón et al. *Multi-resolution Mean Teacher for DCASE 2020 Task 4*. Tech. rep. DCASE2020 Challenge, June 2020
  - D. de Benito-Gorrón et al. *Multi-resolution Mean Teacher for DCASE 2021 Task 4*. Tech. rep. DCASE2021 Challenge, June 2021
  - D. de Benito-Gorrón et al. *Multi-Resolution Combination of CRNN and Conformers for DCASE 2022 Task 4*. Tech. rep. DCASE2022 Challenge, June 2022

- Sound Event Detection as a new line of research in AUDIAS
- Participation in the DCASE international challenges with favorable results
  - Performance and ranking position improved each year
  - Multi-resolution approach consistently outperformed reference systems
    - Independence of the neural network
    - Adaptability to different application scenarios

Conclusions (2/3)

- Sound Event Detection as a new line of research in AUDIAS
- Participation in the DCASE international challenges with favorable results
  - Performance and ranking position improved each year
  - Multi-resolution approach consistently outperformed reference systems
    - Independence of the neural network structure (CRNN / Conformer)
    - Adaptability to different application scenarios (PSDS1 / PSDS2)

Conclusions (2/3)

- Sound Event Detection as a new line of research in AUDIAS
- Participation in the DCASE international challenges with favorable results
  - Performance and ranking position improved each year
  - Multi-resolution approach consistently outperformed reference systems
    - Independence of the neural network structure (CRNN / Conformer)
    - Adaptability to different application scenarios (PSDS1 / PSDS2)

Conclusions (3/3)

- Enhancement of SED input representations with Source Separation
  - New proposed method for Joint Training of SED and SSep improved upon the reference system
  - Unsupervised training of SSep allows to improve SED without additional data
- Optimization of model selection strategy in Mean Teacher
  - Motivated by the iterative training methods proposed for JSS
  - Also beneficial for any Mean Teacher model $\rightarrow$ Employed in DCASE 2022

Conclusions (3/3)

- Enhancement of SED input representations with Source Separation
  - New proposed method for Joint Training of SED and SSep improved upon the reference system
  - Unsupervised training of SSep allows to improve SED without additional data
- Optimization of model selection strategy in Mean Teacher
  - Motivated by the iterative training methods proposed for JSS
  - Also beneficial for any Mean Teacher model → Employed in DCASE 2022

Conclusions (3/3)

- Enhancement of SED input representations with Source Separation
  - New proposed method for Joint Training of SED and SSep improved upon the reference system
  - Unsupervised training of SSep allows to improve SED without additional data
- Optimization of model selection strategy in Mean Teacher
  - Motivated by the iterative training methods proposed for JSS
  - Also beneficial for any Mean Teacher model → Employed in DCASE 2022

Conclusions (3/3)

- Enhancement of SED input representations with Source Separation
  - New proposed method for Joint Training of SED and SSep improved upon the reference system
  - Unsupervised training of SSep allows to improve SED without additional data
- Optimization of model selection strategy in Mean Teacher
  - Motivated by the iterative training methods proposed for JSS
  - Also beneficial for any Mean Teacher model $\rightarrow$ Employed in DCASE 2022

Ongoing and future work

- AUDIAS participation for DCASE 2023
  - Multi-resolution CRNN for PSDS1
  - Optimized multi-resolution Conformer for PSDS2
  - Research paper published last month,[32] journal article in progress
- Analysis and interpretation of Joint Source Separation and Sound Event Detection
  - Journal article currently under review [33]
  - Insights on the interaction between Source Separation and Sound Event Detection

[32] S. Barahona Quirós et al. "Multi-resolution Conformer for Sound Event Detection: Analysis and Optimization" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2023.

[33] D. de Benito-Gorrón, K. Zmolikova, and D. T. Toledano "Analysis and Interpretation of Joint Source Separation and Sound Event Detection in Domestic Environments" *Submitted to PLOS ONE*, 2023.

Ongoing and future work

- AUDIAS participation for DCASE 2023
  - Multi-resolution CRNN for PSDS1
  - Optimized multi-resolution Conformer for PSDS2
  - Research paper published last month,[32] journal article in progress
- Analysis and interpretation of Joint Source Separation and Sound Event Detection
  - Journal article currently under review [33]
  - Insights on the interaction between Source Separation and Sound Event Detection

[32] S. Barahona Quirós et al. "Multi-resolution Conformer for Sound Event Detection: Analysis and Optimization" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2023.

[33] D. de Benito-Gorrón, K. Zmolikova, and D. T. Toledano "Analysis and Interpretation of Joint Source Separation and Sound Event Detection in Domestic Environments" *Submitted to PLOS ONE*, 2023.

Ongoing and future work

- AUDIAS participation for DCASE 2023
  - Multi-resolution CRNN for PSDS1
  - Optimized multi-resolution Conformer for PSDS2
  - Research paper published last month,[32] journal article in progress
- Analysis and interpretation of Joint Source Separation and Sound Event Detection
  - Journal article currently under review [33]
  - Insights on the interaction between Source Separation and Sound Event Detection

---

32 S. Barahona Quirós et al. "Multi-resolution Conformer for Sound Event Detection: Analysis and Optimization" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2023.

33 D. de Benito-Gorrón, K. Zmolikova, and D. T. Toledano "Analysis and Interpretation of Joint Source Separation and Sound Event Detection in Domestic Environments" *Submitted to PLOS ONE*, 2023.

Ongoing and future work

- AUDIAS participation for DCASE 2023
  - Multi-resolution CRNN for PSDS1
  - Optimized multi-resolution Conformer for PSDS2
  - Research paper published last month,[32] journal article in progress
- Analysis and interpretation of Joint Source Separation and Sound Event Detection
  - Journal article currently under review [33]
  - Insights on the interaction between Source Separation and Sound Event Detection

---

32 S. Barahona Quirós et al. "Multi-resolution Conformer for Sound Event Detection: Analysis and Optimization" *Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2023.

33 D. de Benito-Gorrón, K. Zmolikova, and D. T. Toledano "Analysis and Interpretation of Joint Source Separation and Sound Event Detection in Domestic Environments" *Submitted to PLOS ONE*, 2023.

Thank you for your attention

Universidad Autónoma de Madrid
Escuela Politécnica Superior