The impact of non-target events in synthetic soundscapes for sound event detection

Francesca Ronchini, Romain Serizel, Nicolas Turpault, Samuele Cornell

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Overall goal

Detection and classification of sound events using an heterogeneous dataset.

- Training dataset composed of small amount of labeled data and a bigger set of unlabeled data.
- Domain mismatch between synthetic and recorded audio clips.
- Impact of non-target events on system performance when they are included in the training set.

Problem definition

Deep learning is the main method used but:

- it is data-hungry;
- labeling data is time-consuming and bias-prone.

Alternatives:

- Using a limited amount of labeled data together with a bigger set of unlabeled data.
- Including synthetic soundscapes starting from isolated sound events.

The purpose of this paper is to focus on the impact on the system's performance when non-target events are included in the synthetic soundscapes of the training dataset.

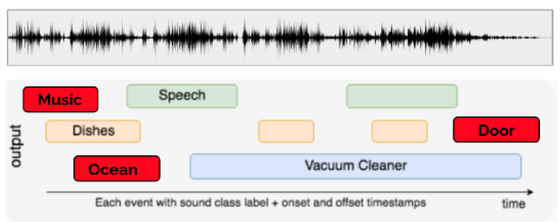

Figure 1:Sound event detection description.

Problem setting

- Exploiting an heterogeneous and unbalanced training dataset.
- DESED dataset composed of:
  - 14412 unlabeled recorded audio clips.
  - 1578 weak labeled recorded audio clips.
  - 10000 strong labeled synthetic audio clips.
- Same SED mean-teacher system released for DCASE 2021 Challenge Task 4.

Dataset Generation

We generated different versions of the synthetic part of the **DESED dataset** in order to investigate their relationship with the target events at training phase of the system.

- Varying TNTSNR at training and validation.
- Using different combinations of dataset at validation and training time.

Using target/non-target events at training

- **Focus:** influence of non-target events at training.
- **Training**: different combinations of the training dataset.
- **Evaluation**: public evaluation set.

Non-target		PSDS1	PSDS2
Train	Val		
✓		33.81 (0.36)	52.62 (0.19)
	✓	35.92 (0.49)	54.85 (0.29)
		34.90 (0.82)	53.07 (1.22)
✓	✓	**36.40**	**58.00**

Table 1:Evaluation results for the **public** set.

- **PSDS1**: performance improved only marginally.
- **PSDS2**: performance improved by a large margin.

- Non-target events confuse the systems at training.
- Help reducing the mismatch with recorded soundscapes at validation.

Evaluating matched/unmatched conditions.

- **Focus:** evaluating mismatched conditions between training and validation phase.
- **Training:** different combinations of the training dataset.
- **Evaluation:** target and target/non-target events.

Non-target		Eval set	PSDS1	PSDS2
Train	Val			
✓		synth_tg_ntg	23.22 (1.33)	36.44 (2.62)
	✓	synth_tg_ntg	20.08 (0.39)	31.33 (1.29)
		synth_tg_ntg	20.13 (0.35)	30.99 (1.07)
✓	✓	synth_tg_ntg	**25.14**	**40.12**
✓		synth_tg	42.82 (2.42)	58.26 (2.08)
	✓	synth_tg	46.92 (1.02)	**62.79 (0.55)**
		synth_tg	**47.73 (0.33)**	62.54 (1.00)
✓	✓	synth_tg	43.22	61.09

Table 2:Evaluation results for the **synth_tg_ntg_eval** set and **synth_tg_eval** set.

Varying TNTSNR at training

- **Focus:** understanding the impact of varying the TNTSNR at training and validation.
- **Training:** different combinations of the training dataset with adjusted TNTSNR.
- **Evaluation:** public evaluation set.

TNTSNR analyzed in this study: **5dB, 10dB and 15dB**.

- **TNTSNR 5dB and 10dB:** the performance changes only marginally between configurations.
- **TNTSNR 15dB:** best performance when validating on **synth_tg_ntg_val**.
- TNTSNR 15dB is a condition closer to that of the recorded soundscapes.
- It allows for selecting models that will be more robust towards non-target events at test time.

Non-target		PSDS1	PSDS2
Train	Val		
Original	15 dB	36.08 (1.13)	57.78 (1.33)
15 dB	Original	**37.37 (0.70)**	**58.64 (1.34)**
15 dB	15 dB	36.10 (0.50)	57.36 (0.89)
Original	Original	36.40	58.00

Table 3:Evaluation results for the second part of the experiment, varying TNTSNR by 15 dB (**synth_15dB** and **synth_15dB_val**). Evaluating with **public** set.

Varying TNTSNR at validation phase.

- **Focus:** investigating the impact of varying the TNTSNR during validation phase.
- **Training:** using only target event.
- **Evaluation:** public evaluation set.

Validation set	PSDS1	PSDS2
synth_5dB_val	38.68 (1.07)	60.57 (0.78)
synth_10dB_val	**39.07 (0.75)**	**60.75 (0.80)**
synth_15dB_val	37.95 (0.53)	59.99 (1.14)

Table 4:Evaluation results of the SED system, training with **synth_tg**, validating with varying TNTNSNR set and evaluating with **public** set.

Evaluating on non-target events only

- **Focus:** understanding if the system gets acoustically confused by a possible similarity in sound between events.
- **Training:** using target and non-target sound events.
- **Evaluation:** using only non-target sound events.

Classes	Nref	Nsys			
		A	B	C	Base
Dog	197	135	126	146	79
Vacuum_cleaner	127	31	42	44	47
Alarm_bell	191	47	50	52	59
Running_water	116	34	41	61	30
Dishes	405	1478	395	1270	305
Blender	100	63	32	55	19
Frying	156	70	41	60	33
Speech	1686	206	181	180	201
Cat	141	99	103	98	73
Electric_shaver	103	21	18	18	7

Table 5:Preliminary evaluation results by classes, evaluating the system with **synth_ntg_eval**. Nsys (A): training with **synth_tg**, validating with **synth_tg_val**; Nsys (B): training with **synth_tg_ntg**, validating with **synth_tg_val**; Nsys (C): training with **synth_tg**, validating with **synth_tg_ntg_val**; Base: baseline using target and non-target events for training and validation.

- Some sound events are detected more than others.
- Discrepancy between the original distribution and the amount of false alarms (for some classes).

Conclusion & Future work

- Using **non-target sound events** can help the SED system to better detect the target sound events, but it is not clear to what extend and what would be the best way to generate the soundscapes.
- SED performance could depend on **mismatches between synthetic and recorded soundscapes**.
- Using non-target events at training decreases the amount of **false alarms** at test.
- Open question: the impact of the **per class distribution** of the sound events (both target and non-target) and their **co-occurrence distribution** on the SED performance.
- Open question: the impact of **non-target events at test time** (induce confusion between classes?)