Universidad Autónoma
de Madrid

Multiple feature resolutions for different Polyphonic Sound Detection Score scenarios in DCASE 2021 Task 4

Diego de Benito-Gorrón   Sergio Segovia   Daniel Ramos   Doroteo T. Toledano

AUDIAS Research Group, Universidad Autónoma de Madrid

< audias >

DCASE Challenge 2021 Task 4

Sound Event Detection (and Separation) in Domestic Environments

- Determining the **temporal location** of acoustic events and their category
- Performance measured by **two different PSDS** (Polyphonic Sound Detection Score) **scenarios**

Multi-resolution analysis

Motivation

- Different acoustic events show **different temporal and spectral characteristics**
- Using **multiple time-frequency resolution points** should improve SED performance

Resolution points

Taking the parameters of the Baseline System (BS) as reference, we define **5 resolution points** for Mel-spectrogram feature extraction, from twice better time resolution (T_{++}) to twice better frequency resolution (F_{++}).

Resolution	$\mathbf{T}_{++}$	$\mathbf{T}_{+}$	**BS**	$\mathbf{F}_{+}$	$\mathbf{F}_{++}$
N	1024	2048	2048	4096	4096
L	1024	1536	2048	3072	4096
R	128	192	256	384	512
$\mathbf{n}_{\mathbf{mel}}$	64	96	128	192	256

Table 1. FFT length (N), window length (L), window hop (R) and number of Mel filters (n_{mel}) of resolution points. N, L, and R are reported in samples, using a sample rate $f_s = 16000$ Hz.

Model fusion

1. Train a single-resolution SED system for each resolution point (we use the DCASE 2021 Baseline System, available at `https://github.com/DCASE-REPO/DESED_task`)
2. Ensemble the class-wise score sequences of several resolutions through **average fusion**, obtaining multi-resolution score sequences
3. Process the resulting score sequences (**threshold** and **median filtering**) to obtain PSDS and F_1 results

Single-resolution results

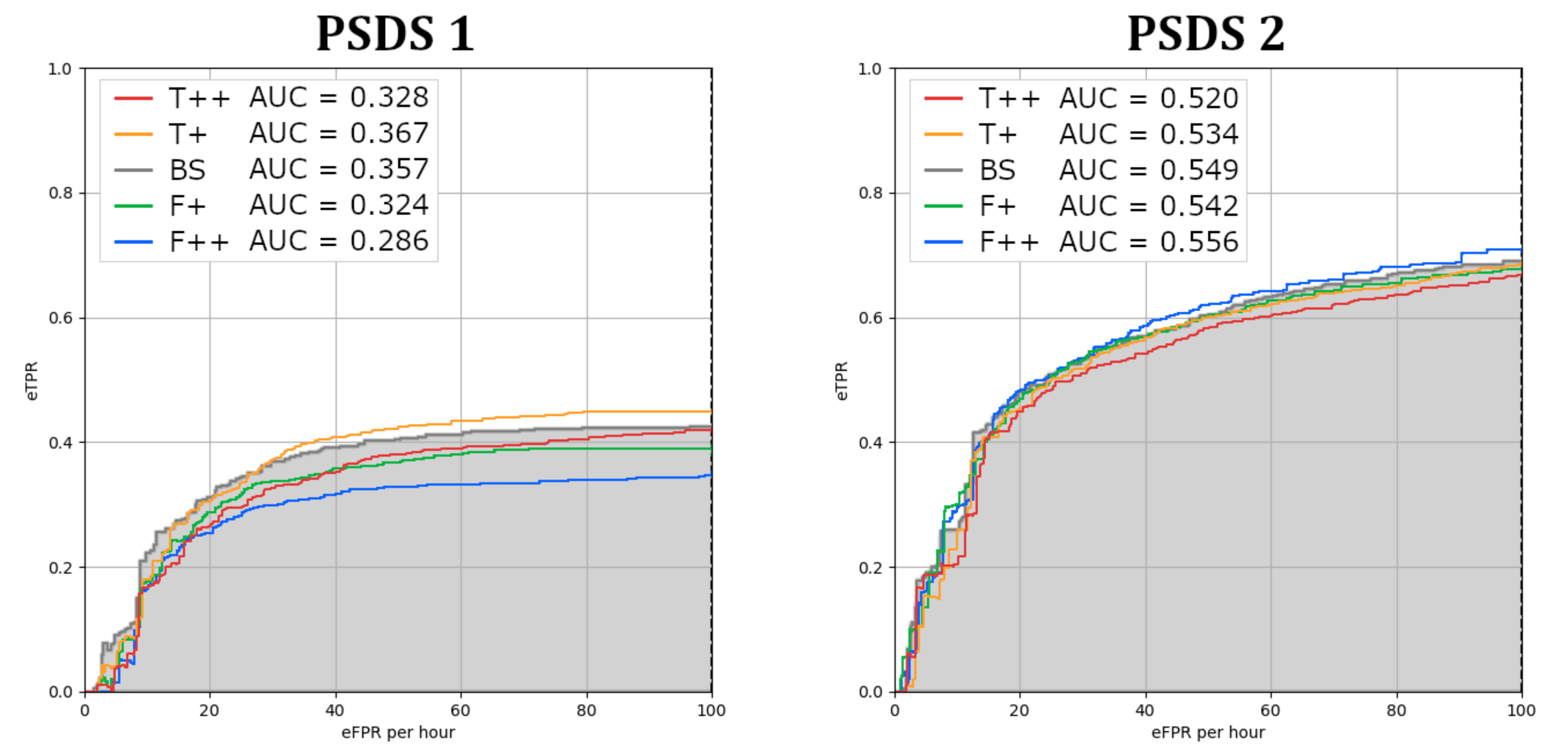

Figure 1. PSDS curves for single-resolution systems in each scenario.

- For **PSDS 1** (focused on precise temporal localization of events), $\mathbf{T}_{+}$ obtained the best AUC, $\mathbf{F}_{++}$ obtained the worst AUC
- For **PSDS 2** (focused on correct event classification), $\mathbf{F}_{++}$ obtained the best AUC, $\mathbf{T}_{++}$ obtained the worst AUC

Higher time resolution benefits precise temporal detection of events, whereas higher frequency resolution helps correct classification.

Multi-resolution results

System	Resolutions	DESED Validation PSDS 1	DESED Validation PSDS 2	DESED Validation $\mathbf{F}_1$(%)	DCASE 2021 Eval PSDS 1	DCASE 2021 Eval PSDS 2	DCASE 2021 Eval $\mathbf{F}_1$(%)
3res	F_{+}, BS, T_{+}	0.380	0.589	45.0	0.343	0.571	42.6
3res-T	BS, T_{+}, T_{++}	**0.386**	0.578	**46.4**	**0.363**	0.574	**43.1**
4res	F_{++}, F_{+}, BS, T_{+}	0.372	**0.600**	45.1	0.345	0.571	42.2
5res	F_{++}, F_{+}, BS, T_{+}, T_{++}	**0.386**	**0.600**	**46.4**	0.361	**0.577**	42.7
Challenge Baseline		0.353	0.553	42.1	0.315	0.547	37.3

Table 2. PSDS and F_1 results of multi-resolution systems over the DESED Validation / Evaluation 2021 sets.

Performance in both PSDS scenarios improves when combining different resolutions.

Class-wise analysis

	PSDS 1					PSDS 2				
	$\mathbf{F}_{++}$	$\mathbf{F}_{+}$	BS	$\mathbf{T}_{+}$	$\mathbf{T}_{++}$	$\mathbf{F}_{++}$	$\mathbf{F}_{+}$	BS	$\mathbf{T}_{+}$	$\mathbf{T}_{++}$
Alarm b./ring.	0.446	0.512	0.556	0.561	**0.567**	**0.855**	0.852	0.836	0.842	0.814
Blender	**0.694**	0.627	0.677	0.652	0.671	**0.851**	0.783	0.799	0.782	0.791
Cat	0.378	0.414	0.411	**0.439**	0.401	**0.717**	0.705	0.661	0.665	0.622
Dishes	0.107	0.132	**0.176**	0.172	0.121	**0.394**	0.376	0.388	0.374	0.389
Dog	0.242	0.272	0.306	**0.316**	0.295	0.666	**0.672**	0.661	0.643	0.604
El.shaver/tooth.	0.787	**0.798**	0.751	0.765	0.687	**0.938**	0.913	0.885	0.912	0.851
Frying	0.582	0.613	0.635	**0.639**	0.607	0.771	0.780	**0.795**	**0.795**	0.759
Running water	0.481	0.510	0.540	0.548	**0.553**	0.714	0.714	0.749	0.750	**0.755**
Speech	0.581	0.603	**0.631**	0.634	0.620	0.830	0.821	**0.834**	0.822	0.813
Vacuum cleaner	0.732	0.769	0.771	0.770	**0.790**	0.892	**0.902**	0.886	0.879	0.873

Table 3. Class-wise PSDS results of single-resolution systems over the DESED Validation set.

	PSDS 1				PSDS 2			
	3res	3res-T	4res	5res	3res	3res-T	4res	5res
Alarm bell/ringing	0.572	**0.584**	0.558	0.577	0.858	0.855	**0.870**	**0.870**
Blender	0.724	0.744	0.746	**0.768**	0.840	0.838	0.853	**0.856**
Cat	0.455	**0.472**	0.435	0.457	0.701	0.667	**0.727**	0.712
Dishes	0.202	0.200	0.197	**0.214**	0.415	0.402	0.435	**0.436**
Dog	0.319	**0.327**	0.312	0.324	0.693	0.681	**0.701**	0.700
Electric shaver/toothbrush	**0.740**	0.695	0.739	0.714	0.902	0.909	**0.918**	0.916
Frying	0.677	**0.682**	0.668	0.674	**0.841**	0.836	0.829	0.833
Running water	0.567	**0.574**	0.562	0.569	0.775	**0.780**	0.771	0.775
Speech	0.661	**0.673**	0.659	0.666	0.851	**0.857**	0.852	0.855
Vacuum cleaner	**0.893**	0.885	0.877	0.890	**0.933**	0.923	0.932	0.932

Table 4. Class-wise PSDS results of multi-resolution systems over the DESED Validation set.

The overall performance pattern (higher time resolution for PSDS 1 and higher frequency resolution for PSDS 2) is not observed for every individual class (e.g: Blender obtains best PSDS 1 with F_{++}, Running water obtains best PSDS 2 with T_{++}).

Conclusions

- Certain resolutions allow to **optimize either PSDS 1** (precise temporal localization of events) or **PSDS 2** (correct event classification)
- Multi-resolution **improves SED performance** for both PSDS settings, and **outperformed the Baseline System** in the DCASE Challenge 2021 Task 4
- Class-wise analysis shows that **each resolution perform better for different event categories**