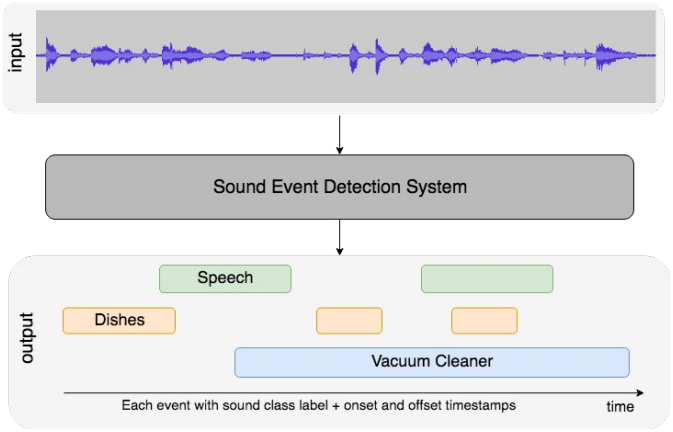
Sound Event Detection in Domestic Environments

DCASE 2022 Challenge Task4

DY(이동윤)

1. Introduction and final results
2. PSDS metric
3. Network architecture
4. AudioSet Data Selection
5. Loss function
6. Data augmentation pipeline
7. Post-processing
8. System results

SED in Domestic Environments



Multi-Label Classification problem

(Not multi-class classification)

- Alarm/bell/ringing
- Blender
- Cat
- Dog
- Dishes
- Electric shaver/toothbrush
- Frying
- Running water
- Speech
- Vacuum cleaner

What we have to solve is to find out which class of sound events occurred and where the event occurred in the audio.

Purpose of the task

- Improve our VAD system by…
  - using modeling techniques and training methods applied to the SED model
  - using augmentation pipeline applied to the SED model
- Have a model that can be used immediately…
  - when providing SED-related services in the future
  - when there is an acoustic event we want to find in the future
- Acquire additional datasets

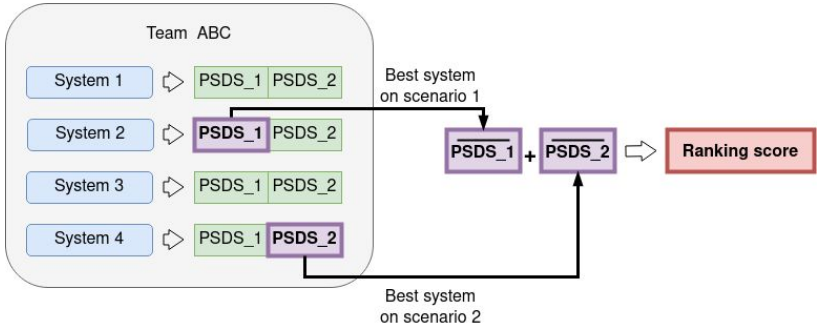
Experiments

testset	model	trainset	student/psds1	student/psds2	teacher/psds1	teacher/psds2	misc
validation	apply-afl/version_1/last		0.3388	0.5561	0.3476	0.5791	
validation	default_realval/1/last		0.3594	0.5599	0.3679	0.5758	
validation	apply-filter-aug/version_0/e195		0.3527	0.5742	0.3623	0.5787	
validation	apply-filter-aug/version_0/last		0.3537	0.5553	0.3650	0.5815	
validation	apply-filter-aug/version_1/e161		0.3858	0.6287	0.3965	0.6401	change hyperparameters of filter aug, use Context Gating for CN
validation	apply-filter-aug/version_1/last		0.3997	0.6339	0.4002	0.6458	change hyperparameters of filter aug, use Context Gating for CN
validation	fdcrnn-weakw/0/e179		0.3930	0.6317	0.4056	0.6419	
validation	fdcrnnn-filt_aug-weakw/0/e174		0.4013	0.6380	0.4068	0.6526	
validation	fdcrnn-mask-filt_aug-weakw/0/e194		0.4212	0.6321	0.4341	0.6542	
public_eval	fdcrnn-mask-filt_aug-weakw/0/e194		0.4496	0.6620	0.4631	0.6832	
validation	fdcrnn-shift-filt_aug-weakw/0/e154		0.4181	0.6615	0.4212	0.6576	
public_eval	fdcrnn-shift-filt_aug-weakw/0/e154		0.4351	0.6516	0.4513	0.6715	
validation	fdcrnn-shift-mask-filt_aug-weakw/0/e189		0.4307	0.6611	0.4456	0.6662	
public_eval	fdcrnn-shift-mask-filt_aug-weakw/0/e189		0.4579	0.6826	0.4649	0.6851	
validation	fdcrnn_filt_aug_a/version_0/e194	dev+real+syn	0.4209	0.6543	0.4216	0.6605	
validation	fdcrnn_filt_aug_a_doubled_filters_no_synth	dev+real	0.4225	0.6815	0.4295	0.6820	fdcrnn_filt_aug_a_doubled_filters/version_0/e199 랑 동일한 config
validation	fdcrnn_filt_aug_a_doubled_filters/version_0	dev+real+syn	0.4399	0.6851	0.4420	0.6882	double model width, RNN cell: 128
validation	fdcrnn_filt_aug_b/version_0/e164	dev+real+syn	0.4063	0.6400	0.4193	0.6602	
validation	fdcrnn_filt_aug_b_doubled_filters/version_0	dev+real+syn	0.3903	0.6509	0.3984	0.6453	double model width
validation	fdcrnn_256gru-shift-mask-filt_aug-weakw/0/	dev+real+syn	0.4340	0.6676	0.4361	0.6460	
public_eval	fdcrnn_256gru-shift-mask-filt_aug-weakw/0/	dev+real+syn	0.4503	0.6681	0.4833	0.6886	
validation	fdcrnn_256gru-shift-mask-filt_aug-weakw-ict/0/e154		0.4000	0.5952	0.4014	0.6132	
validation	fdcrnn-no_more_synth/0/e169	dev+real	0.3675	0.5521	0.3621	0.5304	
validation	main_fdcrnn_double_filters/version_0/e194	dev+real+syn	0.4412	0.6739	0.4505	0.6799	double model width, RNN cell: 256
validation	baseline-fdcrnn-no_realnsynth/0/184	dev	0.4200	0.6470	0.4223	0.6357	
public_eval	baseline-fdcrnn-no_realnsynth/0/184	dev	0.4077	0.6416	0.4110	0.6312	
validation	weak_pred_masking/version_0/e164	dev+real+syn	0.4426	0.6601	0.4410	0.6696	main_fdcrnn_double_filters/version_0와 동일한 config + decode_v
validation	weak_pred_masking/version_1/e139	dev+real+syn	0.4484	0.6831	0.4511	0.6779	fdcrnn_filt_aug_a_doubled_filters/version_0 와 동일한 config + dec
validation	weak_pred_masking/version_1/e139		0.4517	0.6869	0.4540	0.6825	weak_pred_masking/version_1/e139 에 n_test_thresholds 100 적
validation	weak_pred_masking/version_1/e139		0.4514	0.6903	0.4545	0.6858	weak_pred_masking/version_1/e139 에 n_test_thresholds 100 적
validation	weak_pred_masking/version_1/e139						weak_pred_masking/version_1/e139 에 n_test_thresholds 100 적
validation	weak_pred_masking/version_1/e139						weak_pred_masking/version_1/e139 에 n_test_thresholds 100 적
validation	weak_pred_masking/version_2/e169	dev+synth	0.4286	0.6486	0.4325	0.6509	weak_pred_masking/version_1 에서 --strong_real False
validation	weak_SED/version_0/e189	dev+real+syn	0.0566	0.7959	0.0563	0.8010	main_fdcrnn_double_filters/version_0와 동일한 config + decode_v
validation	weak_SED/version_1/e139	dev+real+syn	0.0633	0.8138	0.0578	0.8027	fdcrnn_filt_aug_a_doubled_filters/version_0 와 동일한 config + dec
validation	weak_SED/version_2/e134	dev+synth	0.0582	0.7900	0.0547	0.7926	weak_SED/version_1 에서 --strong_real False

testset	타임스탬프	model	trainset	PSDS1	PSDS2	CB-F1	IB-F1
Validation	220612	e1000-adjust_batch (teach)	dev+real+synth_val	0.465	0.6787	55.6	79.31
Validation	220612	refac_augment (teach)	dev+real+synth_val	0.4559	0.6787	55.1	80.07
Validation	220612	add_music/v1 (teach)	dev+real+synth_val	0.4538	0.7053	55.87	80.01
Validation	220612	add_specaug/v0 (stud)	dev+real+synth_val	0.4455	0.7129	56.98	80.49
Validation	220612	add_specaug/v1 (teach)	dev+real+synth_val	0.4584	0.6956	55.27	79.23
Validation	220612	add_specaug-256gru (stud)	dev+real+synth_val	0.4478	0.7142	54.19	78.71
Validation	220612	add_asp (stud)	dev+real+synth_val	0.4032	0.6948	51.65	73.37
Validation	220612	weak_SED (stud)	dev+real+synth_val	0.0632	0.8138	19.27	52.2
Validation	220612	lrdecay/0/e214 (teach)	dev+real+synth_val	0.4642	0.7186	57.7	77.67
Validation	220613	lrdecay/0/e214 (teach) (+postprocess)	dev+real+synth_val	0.4729	0.7228	56.91%	78.71%
Validation	220614	lrdecay/0/e214 (teach) (T=3)	dev+real+synth_val	0.4656	0.7190	57.71%	77.67%
Validation	220614	lrdecay/0/e214 (teach) (T=5)	dev+real+synth_val	0.4574	0.7111	57.71%	77.67%
Validation	220614	lrdecay/0/e214 (teach) (T=10)	dev+real+synth_val	0.4435	0.6908	57.71%	77.67%
public_eval	220614	lrdecay/0/e214 (teach) (T=1)	dev+real+synth_val	0.4894	0.7395	58.89%	79.43%
public_eval	220614	lrdecay/0/e214 (teach) (T=3)	dev+real+synth_val	0.4876	0.7447	58.89%	79.43%
public_eval	220614	lrdecay/0/e214 (teach) (T=5)	dev+real+synth_val	0.4814	0.7394	58.89%	79.43%
Validation	220614	lrdecay/0/last (stud) (T=1)	dev+real+synth_val	0.4524	0.7014	57.81%	77.15%
Validation	220614	lrdecay/0/last (stud) (T=3)	dev+real+synth_val	0.4526	0.7023	57.81%	77.15%
Validation	220614	lrdecay/0/last (stud) (T=5)	dev+real+synth_val	0.4459	0.6951	57.81%	77.15%
Validation	220614	afl/0/e354 (teach)	dev+real+synth_val	0.4306	0.6852	51.55%	75.33%
Validation	220614	afl-T3/0/e274 (teach)	dev+real+synth_val	0.4518	0.6905	51.97%	77.17%
Validation	220614	afl_00652_1_T1/0/e379 (teach)	dev+real+synth_val	0.4577	0.6962	57.04%	77.04%
Validation	220614	asl_00625_1_T3/0/e149 (teach)	dev+real+synth_val	0.4598	0.6898	57.60%	77.87%
Validation	220614	afl_00625_1_T5/0/e269 (teach)	dev+real+synth_val	0.4500	0.6820	54.92%	75.89%
Validation	2022. 7. 10 오후	fdcrnn_subsample8/1/e179 (stud)	dev+real+synth_val	0.3847	0.7492	50.06%	75.92%
Validation	2022. 7. 10 오후	audioset-afl/3/e174 (teach)	dev+real+synth_val	0.4399	0.6918	56.10%	76.34%
Validation	2022. 7. 10 오후	audioset-afl/2/e274 (teach)	dev+real+synth_val	0.4203	0.6855	52.05% (stud)	76.27% (stud)
Validation	2022. 7. 10 오후	afl_T3-lrdecay/0/e249 (stud)	dev+real+synth_val	0.4594	0.7018	58.12%	79.68%
Validation	2022. 7. 10 오후	lrdecay-weak_w1/0/e319 (teach)	dev+real+synth_val	0.4264	0.6929	52.05%	74.60%
Validation	2022. 7. 10 오후	lrdecay/0/e214 (teach) + wp_mask1	dev+real+synth_val	0.4674	0.7230	57.98%	77.60%
Validation	2022. 7. 10 오후	fdcrnn_subsample8/2/e214 (teach)	dev+real+synth_val	0.3834	0.7361	49.68%	74.41%
Validation	2022. 7. 10 오후	audioset-temperature-3/0/e199 (stud)	dev+real+synth_val	0.41225	0.6422	54.76%	75.23%
Validation	2022. 7. 10 오후	audioset-afl-lrdecay/0/e439 (stud)	dev+real+synth_val	0.4455	0.7045	54.50%	75.44%
public_eval	2022. 7. 10 오후	system1/0/e499 (stud)	dev+synth_val+validati	0.4184	0.6590	51.66%	75.28%
public_eval	2022. 7. 10 오후	system2-seed0/0/e499	dev+real+synth_val+v	0.4738	0.7473	58.44%	77.81%
public_eval	2022. 7. 10 오후	system3/0/e499 (stud)	dev+real+synth_val+v	0.3763	0.7842	51.16%	78.05%

About 110 experiments!

Task Ranking



- Submit up to 4 different systems.
- Should submit at least one system not using external resouces (allowed pre-trained models, allowed external datasets).
- not allowed to use the public evaluation dataset and synthetic evaluation dataset to train their systems or tune hyper-parameters.

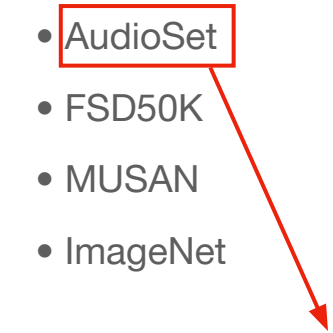
External Resources

Pre-trained Models

- YAMNet
- PANNs
- OpenL3
- COLA
- BYOL-A
- AST
- BYOL-A

Allowed Datasets

- SINS
- AudioSet
- FSD50K
- MUSAN
- ImageNet

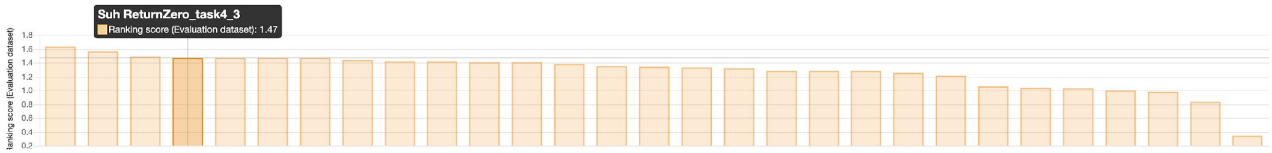


Got good result using AudioSet dataset with Asymmetric Focal Loss

Results

Teams ranking

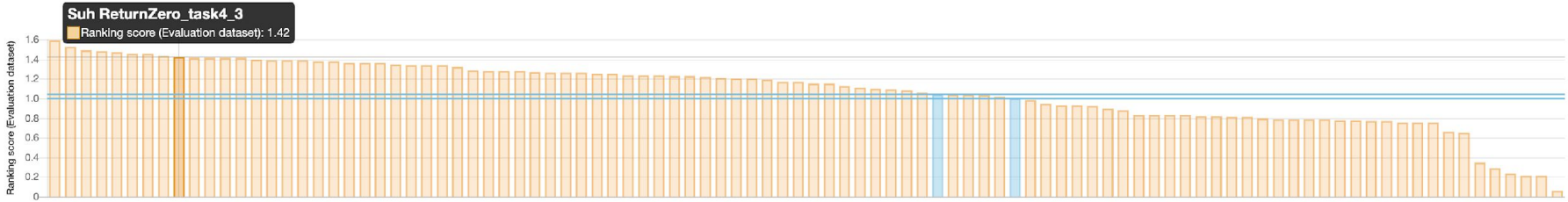
Table including only the best ranking score per submitting team.



Rank	Submission code (PSDS 1)	Submission name (PSDS 1)	Submission code (PSDS 2)	Submission name (PSDS 2)	Technical Report	Ranking score (Evaluation dataset)	PSDS 1 (Evaluation dataset)	PSDS 2 (Evaluation dataset)
1	Ebbers_UPB_task4_1	CRNN ensemble	Ebbers_UPB_task4_2	FBCRNN ensemble		1.63	0.552	0.824
2	He_BYTEDANCE_task4_3	DCASE2022 SED mean teacher system 3	He_BYTEDANCE_task4_4	DCASE2022 SED mean teacher system 4		1.57	0.525	0.810
3	Zhang_UCAS_task4_2	DCASE2022 pretrained system 2	Zhang_UCAS_task4_4	DCASE2022 weak_pred system		1.49	0.484	0.784
4	Suh_ReturnZero_task4_3	rtzr_audioset	Suh_ReturnZero_task4_4	rtzr_weak-SED		1.47	0.478	0.774
4	Dinkel_XiaoRice_task4_3	PRECISE	Dinkel_XiaoRice_task4_4	TAG		1.47	0.451	0.824
4	Kim_GIST_task4_1	Kim_GIST_task4_1	Kim_GIST_task4_1	Kim_GIST_task4_1		1.47	0.514	0.713
4	Xu_SRCB-BIT_task4_2	PANNs-FDY-CRNN-wTCL system 2	Xu_SRCB-BIT_task4_3	PANNs-FDY-CRNN-weak train		1.47	0.482	0.774
5	Li_USTC_task4_SED_1	Mean teacher Pseudo labeling system 1	Li_USTC_task4_SED_2	Mean teacher Pseudo labeling system 2		1.44	0.480	0.740
6	Kim_LGE_task4_2	DCASE2022 Kim system 2	Kim_LGE_task4_3	DCASE2022 Kim system 3		1.42	0.444	0.781
6	Mizobuchi_PCO_task4_2	PCO_task4_SED_B	Mizobuchi_PCO_task4_4	PCO_task4_SED_D		1.42	0.439	0.787
7	Li_WU_task4_4	ATST-RCT SED system ATST ensemble	Li_WU_task4_4	ATST-RCT SED system ATST ensemble		1.41	0.486	0.894
7	Xie_UESTC_task4_4	CBAM-T CRNN 2	Xie_UESTC_task4_2	CNN14 FC		1.41	0.426	0.800
8	Liu_SRCN_task4_3	DCASE2022 task4 AudioSet strong	Liu_SRCN_task4_1	DCASE2022 task4 Pre-Trained 1		1.38	0.425	0.777
9	Li_ICT-TOSHIBA_task4_1	Hybrid system of SEDT and frame-wise model	Li_ICT-TOSHIBA_task4_2	Hybrid system of SEDT and frame-wise model		1.35	0.439	0.709
10	Hao_UNISOC_task4_1	SUBMISSION FOR DCASE2022 TASK4	Hao_UNISOC_task4_2	SUBMISSION FOR DCASE2022 TASK4		1.34	0.425	0.723
11	Nam_KAIST_task4_SED_2	SED_2	Nam_KAIST_task4_SED_3	SED_3		1.33	0.409	0.747
12	Khandelwal_FMSG-NTU_task4_3	FMSG-NTU DCASE2022 SED Model-1	Khandelwal_FMSG-NTU_task4_2	FMSG-NTU DCASE2022 SED Model-1		1.32	0.410	0.731
13	Huang_NSYSU_task4_1	DCASE2022 KDmt SED	Huang_NSYSU_task4_1	DCASE2022 KDmt SED		1.28	0.434	0.650
13	KIM_HYU_task4_4	single2	KIM_HYU_task4_2	single1		1.28	0.423	0.664

Ranked 4th out of 30 teams.

Systems ranking



Bar | Scatter | Off

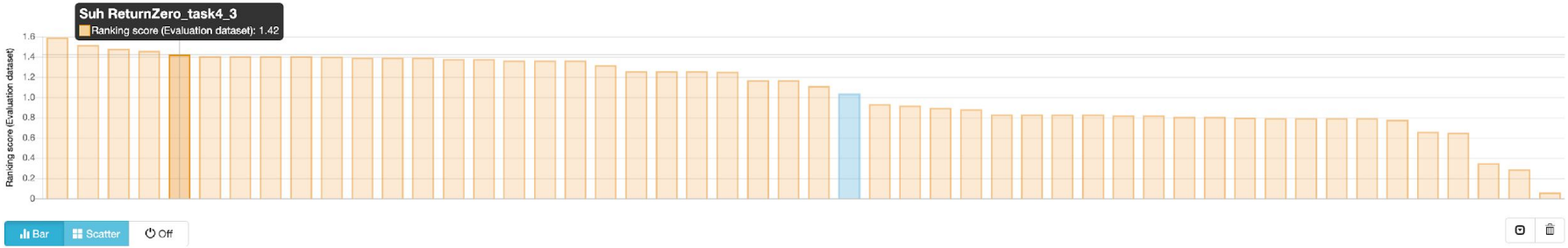
Rank	Submission code	Submission name	Technical Report	Ranking score (Evaluation dataset)	PSDS 1 (Evaluation dataset)	PSDS 2 (Evaluation dataset)	PSDS 1 (Development dataset)	PSDS 2 (Development dataset)
1	Ebbers_UPB_task4_1	CRNN ensemble		1.59	0.552	0.786	0.512	0.772
2	He_BYTEDANCE_task4_3	DCASE2022 SED mean teacher system 3		1.52	0.525	0.748	0.533	0.762
3	Ebbers_UPB_task4_4	CRNN ensemble w/o external data		1.49	0.509	0.742	0.492	0.721
4	He_BYTEDANCE_task4_2	DCASE2022 SED mean teacher system 2		1.48	0.503	0.749	0.521	0.771
5	Kim_GIST_task4_1	Kim_GIST_task4_1		1.47	0.514	0.713	0.458	0.688
6	Kim_GIST_task4_2	Kim_GIST_task4_2		1.46	0.510	0.711	0.456	0.685
6	Ebbers_UPB_task4_3	tag-conditioned CRNN ensemble		1.46	0.527	0.679	0.483	0.713
7	Kim_GIST_task4_3	Kim_GIST_task4_3		1.43	0.500	0.695	0.452	0.682
8	Suh_ReturnZero_task4_3	rtzr_audioset		1.42	0.478	0.719	0.445	0.704
9	Zhang_UCAS_task4_2	DCASE2022 pretrained system 2		1.41	0.484	0.697	0.481	0.694

Showing 1 to 10 of 98 rows 10 rows per page

‹ 1 2 3 4 5 ... 10 ›

Ranked 8th out of 101 systems.

With external resources



Rank	Submission code	Submission name	Technical Report	Ranking score (Evaluation dataset)	PSDS 1 (Evaluation dataset)	PSDS 2 (Evaluation dataset)	PSDS 1 (Development dataset)	PSDS 2 (Development dataset)
1	Ebbers_UPB_task4_1	CRNN ensemble		1.59	0.552	0.786	0.512	0.772
2	He_BYTEDANCE_task4_3	DCASE2022 SED mean teacher system 3		1.52	0.525	0.748	0.533	0.762
3	He_BYTEDANCE_task4_2	DCASE2022 SED mean teacher system 2		1.48	0.503	0.749	0.521	0.771
4	Ebbers_UPB_task4_3	tag-conditioned CRNN ensemble		1.46	0.527	0.679	0.483	0.713
5	Suh_ReturnZero_task4_3	rtzr_audioset		1.42	0.478	0.719	0.445	0.704
6	Zhang_UCAS_task4_2	DCASE2022 pretrained system 2		1.41	0.484	0.697	0.481	0.694
6	Li_WU_task4_4	ATST-RCT SED system ATST ensemble		1.41	0.486	0.694	0.477	0.734
6	Xu_SRCB-BIT_task4_2	PANNs-FDY-CRNN-wrTCL system 2		1.41	0.482	0.702	0.485	0.725
6	Li_USTC_task4_SED_1	Mean teacher Pseudo labeling system 1		1.41	0.480	0.713	0.479	0.735
7	Li_WU_task4_3	ATST-RCT SED system ATST base		1.40	0.482	0.693	0.468	0.702

Showing 1 to 10 of 50 rows 10 rows per page

‹ 1 2 3 4 5 ›

Ranked 5th out of 50 systems that used external resources.

Annotations

Weak Annotation

- Y-BJNMHMZDcU_50.000_60.000.wav Alarm_bell_ringing,Dog
- Without time informations (onset, offset)

Strong Annotation

- YOTsn73eqbfc_10.000_20.000.wav 0.163 0.665 Alarm_bell_ringing
- With time informations (onset, offset)
- Synthesized soundscapes (with Scaper) + Recorded soundescapes

The minimum length for an event is 250ms

The minimum duration of the pause between two events from the same class is 150ms

Datasets (DESED)

Training Set

- Synthetic audio (strongly labeld) - 10000 clips
- AudioSet (strongly labeled) - 3470 clips (considered as external resources)
- Weakly labeled set -1578 clips
- Unlabeld in domain set - 14412 clips

Validation Set

- Strongly labeld set - 3668 clips (synthetic audio - 2500 clips)

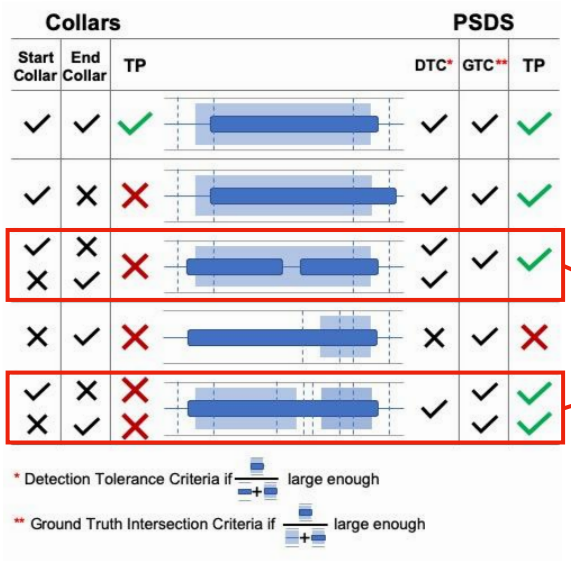
Evaluation Set

- Total evaluation set consists of 10 seconds and 5 minutes audio clips - 14862 clips
- Public evaluation set from Youtube (strongly labeld) - 692 clips (is subset of total evaluation set)
- Not allowed to use to train the model.

PSDS (Polyphonic Sound Detection Score)

- **PSDS Scenario 1**
  - needs to react fast upon an event detection.
  - the localization of the sound event is really important
- **PSDS Scenario 2**
  - must avoid confusing between classes.
  - the reaction time is less crucial than in the first scenario.

Collar based vs DTC and GTC based



It makes sense that the system detected well for these cases where a very short silence duration exists between events of the same class. For example, dog barking, 멍! 멍!. PSDS count these as true positivie, whereas collar-based metrics do not.

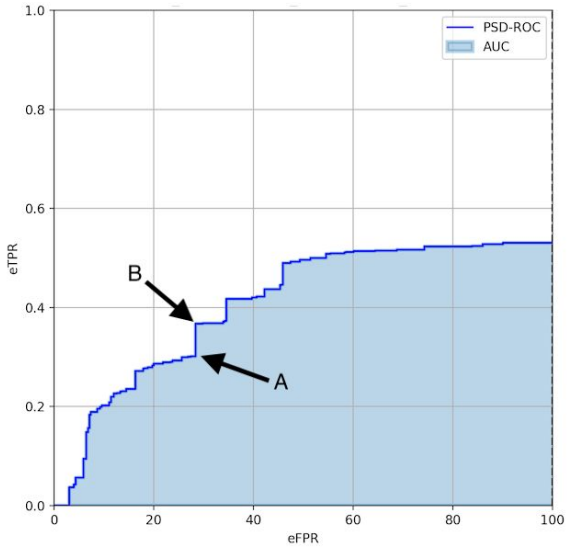
PSDS

- PSDS calculates the area under the Polyphonic Sound Detection Receiver Operating Characteristic Curve (AUC of PSD-ROC Curve).
- Create PSD-ROC Curve by normalizing each class's ROC Curves.

Definition 5 (Polyphonic Sound Detection Score) *Given a dataset with a set of ground truth events, $\mathcal{Y}$, and the set of evaluation parameters, $(\rho_{DTC}, \rho_{GTC}, \rho_{CTTC}, \alpha_{CT}, \alpha_{ST})$, a SED system's PSDS is:*

$$PSDS \triangleq \frac{1}{e_{max}} \int_0^{e_{max}} r(e)\, de \tag{10}$$

where e_{max} is the maximum eFPR *value of interest for the SED application under evaluation.*



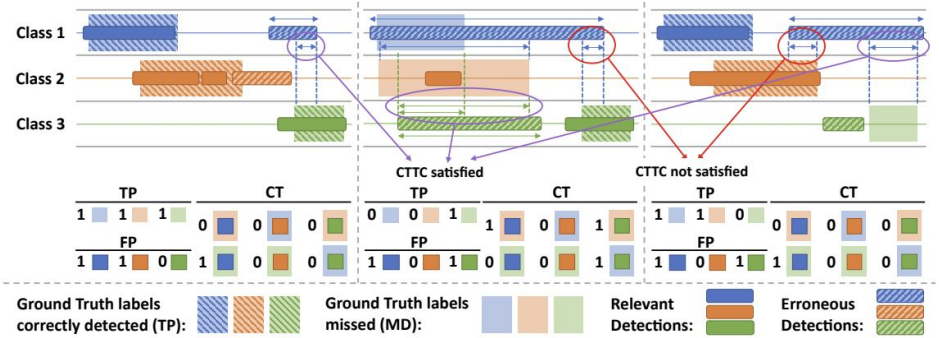
PSDS parameters

PSDS Scenario 1

- Detection Tolerance criterion (DTC): 0.7
- Ground Truth intersection criterion (GTC): 0.7
- Cost of instability across class (α_{ST}): 1
- Cost of CTs on user experience (α_{CT}): 0
- Maximum eFPR (e_max): 100

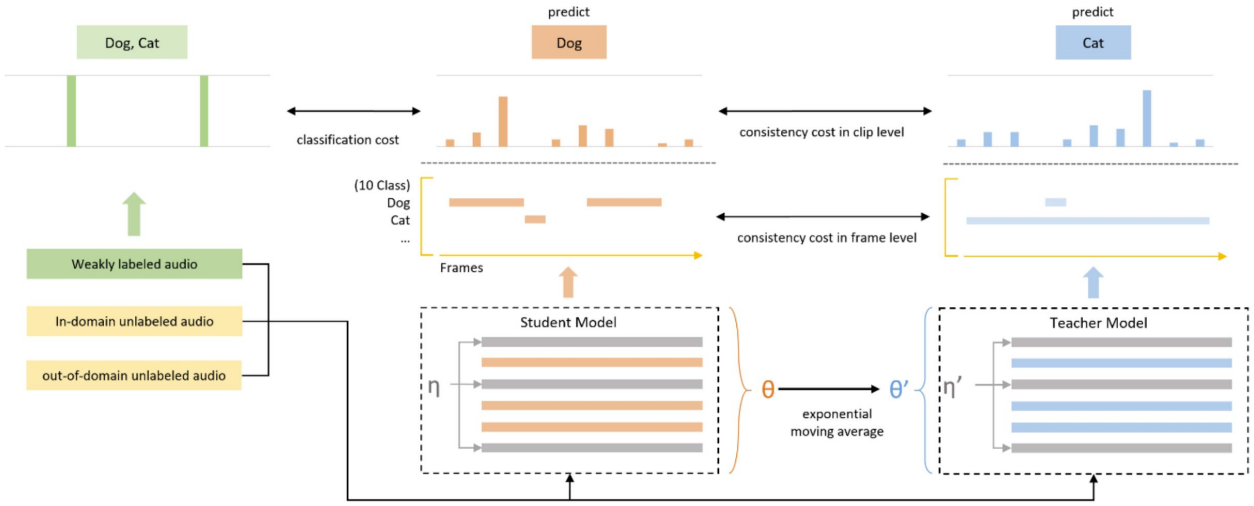
PSDS Scenario 2

- Detection Tolerance criterion (DTC): 0.1
- Ground Truth intersection criterion (GTC): 0.1
- Cost of instability across class (α_{ST}): 1
- Cross-Trigger Tolerance criterion (cttc): 0.3
- Cost of CTs on user experience (α_{CT}): 0.5
- Maximum eFPR (e_max): 100



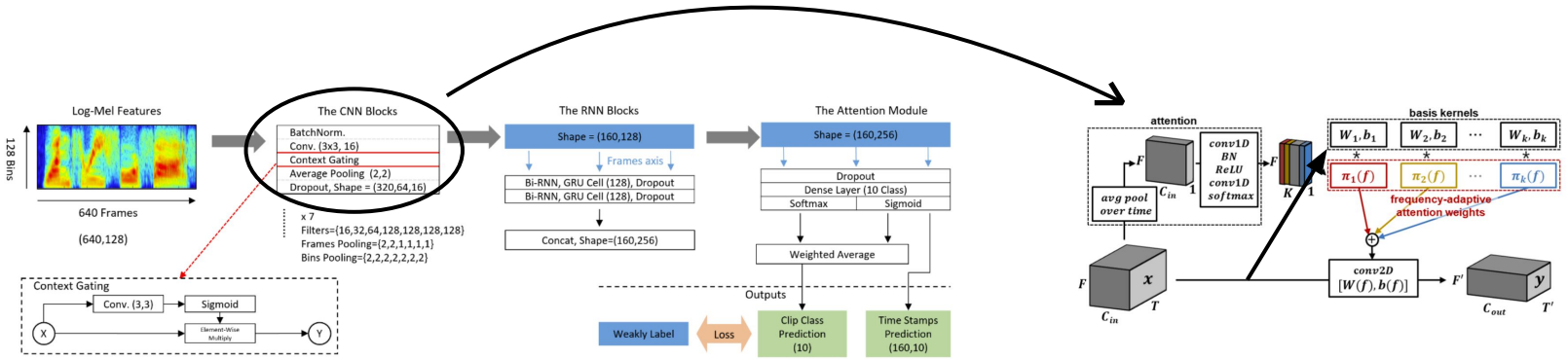
(b) TPs, FPs and CTs in the proposed method for a 3-class problem. *CTTC* effects are pointed at by the arrows.

Mean-Teacher method



- The teacher model helps the student model to train better while training.
- The teacher model's parameters are updated by the exponential moving average of the student model's parameters.

Frequency Dynamic CRNN (FDY-CRNN)



Baseline CRNN architecture

The paper claims that applying FDY-Conv improves performance by solving the problem of not being shift-invariant on the frequency axis in the spectrogram, unlike the image data.

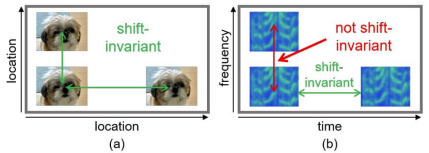
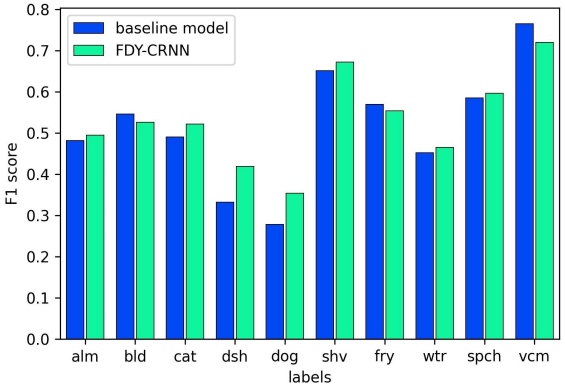
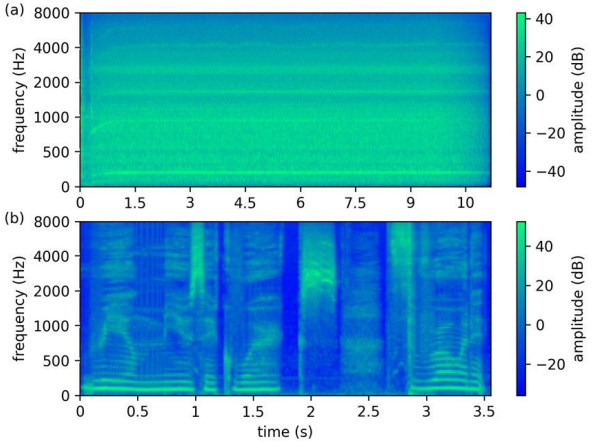


Figure 1: *An illustration of shift-invariance of (a) location in 2D image data, (b) time and frequency in 2D audio data (log Mel spectrogram).*



class-wise event-based (collar-based) F1 scores for the baseline CRNN model and FDY-CRNN



(a) Quasi-stationary sound events spectrogram (eg. blender, frying, vacuum cleaner)
(b) Non-stationary sound events spectrogram (eg. alarm/bell/ringing, cat, dishes)

- Frequency Dynamic Convolution is especially effective on non-stationary sound events, whereas less effective on quasi-stationary sound events.
- The paper claims that this result proves that FDY-CRNN predicts well for frequency-dependent sound events.

AudioSet Data Selection

	AudioSet	DESED	pred	n_samples
20	Beep, bleep	Alarm_bell_ringing	0.362647	2338
21	Alert	Alarm_bell_ringing	0.360047	27
22	Wind chime	Alarm_bell_ringing	0.546944	205
23	Sine wave	Alarm_bell_ringing	0.370052	385
24	Telephone dialing, DTMF	Alarm_bell_ringing	0.341993	208
25	Ringtone	Alarm_bell_ringing	0.554473	296
26	Busy signal	Alarm_bell_ringing	0.510701	199
27	Smoke detector, smoke alarm	Alarm_bell_ringing	0.722010	87
28	Dial tone	Alarm_bell_ringing	0.555186	203
29	Chainsaw	Blender	0.366771	379
30	Blender, food processor	Blender	0.552312	309
31	Cat	Cat	0.319360	444
32	Caterwaul	Cat	0.428744	164
33	Meow	Cat	0.576541	353
34	Snap	Dishes	0.320098	8
35	Dishes, pots, and pans	Dishes	0.380647	558
36	Hammer	Dishes	0.356717	383
37	Cupboard open or close	Dishes	0.318000	29
38	Dog	Dog	0.303968	1164
39	Gobble	Dog	0.314338	150
40	Yip	Dog	0.563824	201

Table 1: Number of selected AudioSet samples

Class	samples
Alm	500
Bld	373
Cat	500
Dish	500
Dog	500
Shv	369
Fry	500
Wtr	359
Spch	500
Vcm	294
Total	4395

- Defined the class of AudioSet that can be grouped into the DESED class.
- Using the model trained during the experiment, only classes with an average prediction score of 0.3 or higher were relabeled as the DESED class.

Loss Function

- Classification Loss (Supervised Loss): Binary Cross Entropy -> **Asymmetric Focal Loss**
- Consistency Loss (Self-supervised Loss): Mean Squared Error

$$L = -yL_+ - (1-y)L_-$$

General Binary Loss Form

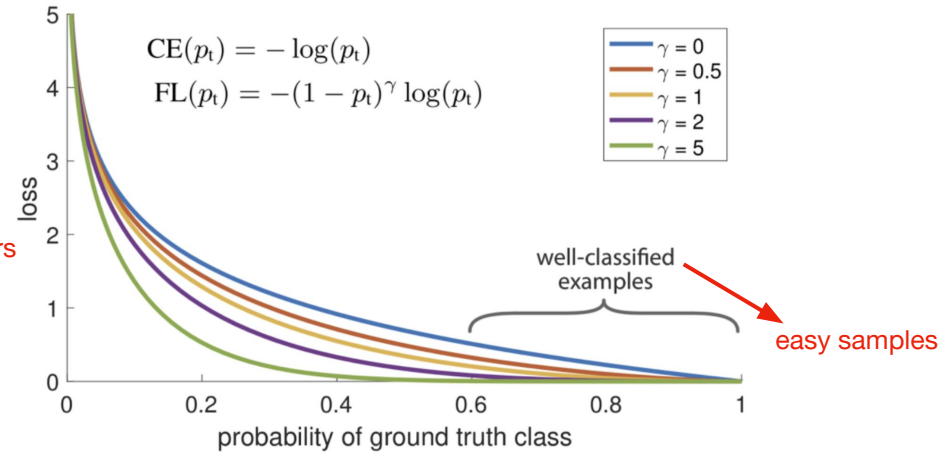
$$\begin{cases} L_+ = (1-p)^{\gamma} \log(p) \\ L_- = p^{\gamma} \log(1-p) \end{cases}$$

Focal Loss

$$\begin{cases} L_+ = (1-p)^{\gamma_+} \log(p) \\ L_- = p^{\gamma_-} \log(1-p) \end{cases}$$

Asymmetric Focal Loss

focusing parameters



Asymmetric Focal Loss

$$L = -yL_+ - (1-y)L_-$$

General Binary Loss Form

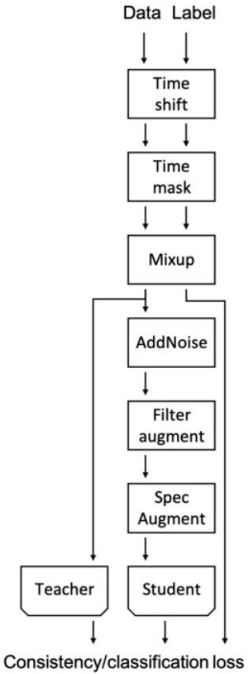
$$\begin{cases} L_+ = (1-p)^{\gamma_+}\log(p) \\ L_- = p^{\gamma_-}\log(1-p) \end{cases}$$

Asymmetric Focal Loss

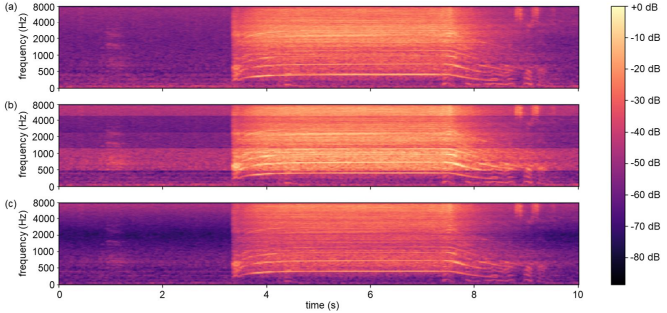
- Worked well with training AudioSet dataset.
- Consider easy and hard sampels using focusing parameters.
- Solve class imbalance problem by setting γ^- greater than γ^+, because most classes have more inactive(negative) frames than active(positive) frames.
  - Then why not just apply linear weighting to each positive and negative loss considering samples distribution? -> The paper claims that it's insufficient to solve in multi-label classification problem.

Data Augmentation Pipeline

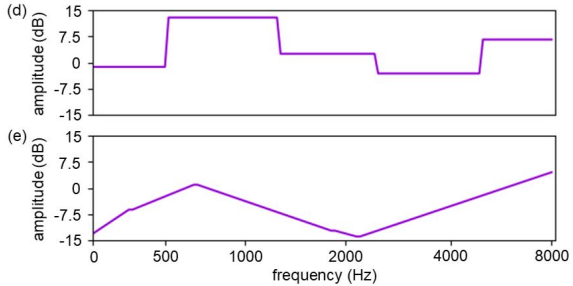
- Create datasets of various patterns applying time shifting, time masking, and mixup to teacher model input.
- Additionally apply noise augmentations such as noise adding, filter augment, and frequency masking to student model input to improve the efficiency of the mean-teacher training method.
- Especially FilterAugment improved the performance of the SED model a lot.



FilterAugment



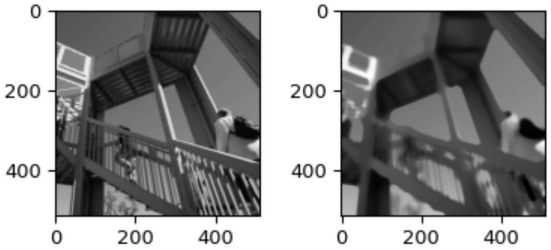
(a) original log mel spectrogram
(b) step type FilterAugment applied log mel spectrogram
(c) linear type FilterAugment applied log mel spectrogram



(d) step type filter
(e) linear type filter

- FilterAugment applies weights randomly to the random frequency bands to increase or decrease the amplitude.
- Similar to Frequency masking, but Frequency masking removes whole specific band.
  - It can be a problem if the band has a significant acoustic feature to help the model train better.
- We applied a linear type filter through many experiments.

Class-wise Median Filtering

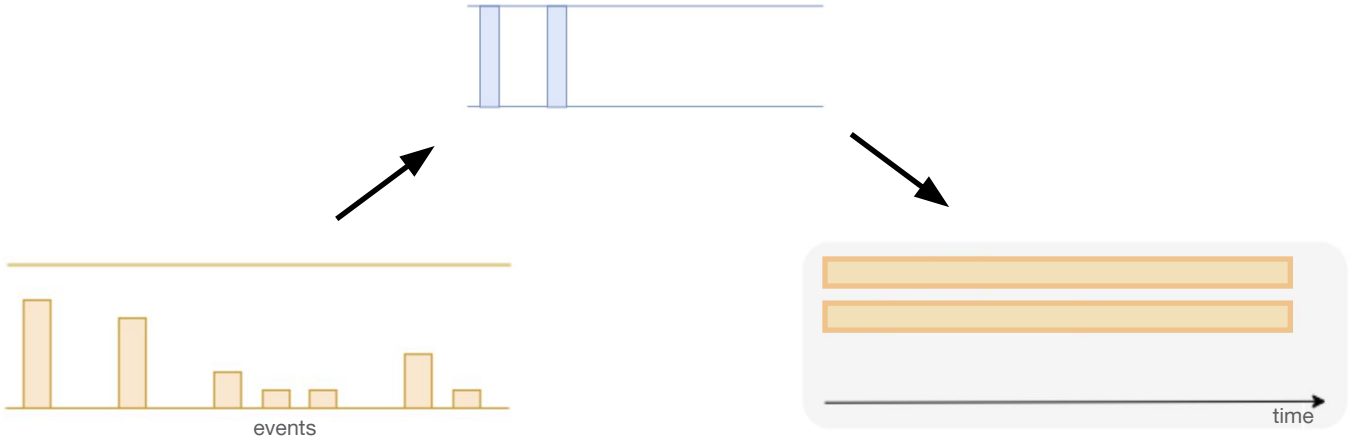


Class	Occur-rences	Mean in s	Medi-an in s
Alarm/bell/ringing	755	1.07	0.38
Blender	540	2.58	1.62
Cat	547	1.07	0.88
Dishes	814	0.58	0.37
Dog	516	0.98	0.48
Electric shaver/toothbrush	230	4.52	4.07
Frying	137	5.17	5.13
Running water	157	3.91	3.60
Speech	2132	1.16	0.89
Vacuum cleaner	204	5.29	5.30

Median filter size per class (above table order): [7, 8, 5, 4, 5, 41, 48, 40, 11, 55]

- Applied median filtering to smooth the prediction results of each class.
- Applied median-filtering class-wisely because average event duration varies between classes as shown in the above table.

Weak SED method



weak prediction

strong prediction made by only weak prediction result

original strong prediction

- Set onset/offset as the total audio duration for each class using weak prediction results.
- This method includes a premise that weak prediction is more accurate than strong prediction in predicting which class has occurred in the audio.
- Maximize the performance of the PSDS2 metric.

Performance of submitted systems

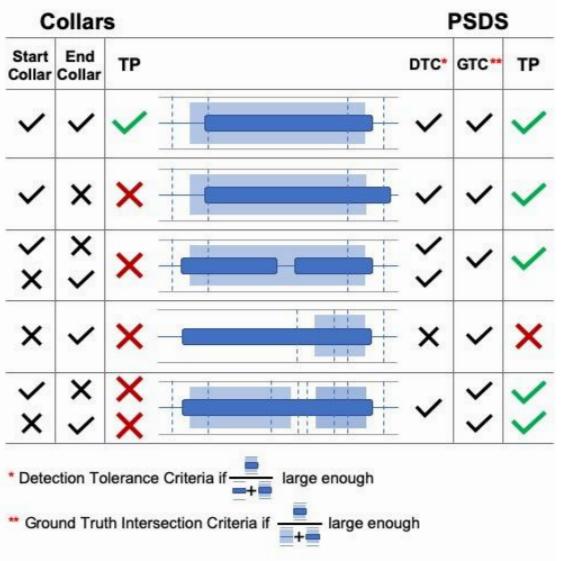
System	PSDS1(eval)	PSDS2(eval)	PSDS1(dev)	PSDS2(dev)
Baseline			0,336	0,536
Baseline (AudioSet strong)			0,351	0,552
Baseline (AST)			0,313	0,722
Zheng et al.			0,454	0,671
FDY-SED			0,452	0,67
rtzr_dev-only	0,393	0,650	0,424	0,649
rtzr_strong-real	0,458	0,721	**0,473**	0,723
rtzr_audioset	**0,478**	0,719	0,445	0,704
rtzr_weak-SED	0,062	**0,774**	0,063	**0,814**

Thank you!

[Appendix]

DTC and GTC

- $\mathcal{C}$ *being a set of sound classes,*
- $\mathcal{Y} = \bigcup_{c \in \mathcal{C}} \mathcal{Y}_c$ *being a dataset which is the union of subsets of ground truth labels for each class* $c \in \mathcal{C}$*, defined as* $\mathcal{Y}_c = \{y_i = (t_{s,i}, t_{e,i}, c_i) \colon c_i = c\}$*, where each ground truth label* y_i *is defined by its class* c_i*, start time* $t_{s,i}$ *and end time* $t_{e,i}$*,*
- $\mathcal{X}^* = \bigcup_{c \in \mathcal{C}} \mathcal{X}^*_c$ *being a set of detections which is the union of subsets of detections for each class* $c \in \mathcal{C}$*, defined as* $\mathcal{X}^*_c = \{x_j = (t_{s,j}, t_{e,j}, c_j) : c_j = c\}$*, where each detection* x_j *is defined by its class* c_j*, start time* $t_{s,j}$ *and end time* $t_{e,j}$*, and where the starred notation* $()^*$ *indicates dependency on operating point parameters* τ_c*,*



Definition 2 (Detection Tolerance Criterion - DTC) *Given a set of ground truth labels* $\mathcal{Y}$*, the DTC filters a set of detections* $\mathcal{X}^*$ *to create the set of relevant detections* $\mathcal{X}^*_{DTC} = \bigcup_{c \in \mathcal{C}} \mathcal{X}^*_{DTC,c} \subset \mathcal{X}^*$*, based on a detection tolerance parameter* $0 \leq \rho_{DTC} \leq 1$ *such that:*

$$\mathcal{X}^*_{DTC,c} \triangleq \left\{ x_j \in \mathcal{X}^*_c \colon \frac{\sum_{y_i \in \mathcal{Y}_c} t_{y_i \cap x_j}}{(t_{e,j} - t_{s,j})} \geq \rho_{DTC} \,\middle|\, \mathcal{Y}_c, \rho_{DTC} \right\} \quad (2)$$

where $t_{y_i \cap x_j}$ *represents the duration of the intersection of the ground truth labels* y_i *and detected events* x_j*. As a corollary to the above definition, the set of FPs is defined as* $\overline{\mathcal{X}}^*_{DTC} = \bigcup_{c \in \mathcal{C}} \overline{\mathcal{X}}^*_{DTC,c}$ *with* $\overline{\mathcal{X}}^*_{DTC,c} \triangleq \mathcal{X}^*_c \setminus \mathcal{X}^*_{DTC,c}$ *and the number of FPs per class is defined as the cardinality of this set:* $N^*_{FP,c} = \left| \overline{\mathcal{X}}^*_{DTC,c} \right|$*.*

Definition 3 (Ground Truth intersection Criterion - GTC) *The GTC creates the set of correctly detected ground truth events,* $\mathcal{Y}_{TP} = \bigcup_{c \in \mathcal{C}} \mathcal{Y}_{TP,c} \subset \mathcal{Y}$*, given the set of relevant detections* $\mathcal{X}^*_{DTC,c} \forall c \in \mathcal{C}$ *and a ground truth tolerance parameter* $0 \leq \rho_{GTC} \leq 1$ *such that:*

$$\mathcal{Y}^*_{GTC,c} \triangleq \left\{ y_i \in \mathcal{Y}_c \colon \frac{\sum_{x_j \in \mathcal{X}^*_{DTC,c}} t_{y_i \cap x_j}}{(t_{e,i} - t_{s,i})} \geq \rho_{GTC} \,\middle|\, \mathcal{X}^*_{DTC,c}, \rho_{GTC} \right\} \quad (3)$$

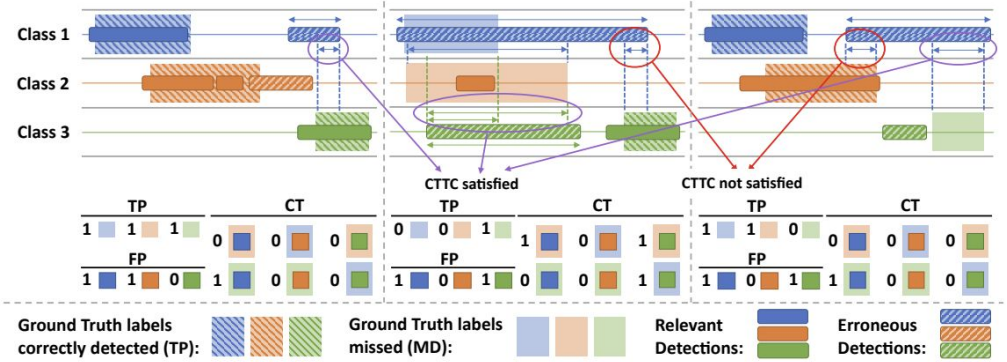
The number of TPs is then defined as the cardinality of the GTC sets across all classes: $N^*_{TP} = \sum_{c \in \mathcal{C}} N^*_{TP,c}$ *where* $N^*_{TP,c} = \left| \mathcal{Y}^*_{GTC,c} \right|$*.*

CTTC

Definition 4 (Cross-Trigger Tolerance Criterion - CTTC) *Given a set of ground truth events $\mathcal{Y}$, the cross-trigger tolerance criterion counts the CTs as $N^*_{CT} = \sum_{c\in\mathcal{C}} \sum_{\substack{\hat{c}\in\mathcal{C}\\ \hat{c}\neq c}} N^*_{CT,c,\hat{c}}$, given FP set $\overline{\mathcal{X}}^*_{DTC}$ and a cross-trigger tolerance parameter $0 \leq \rho_{CTTC} \leq 1$ such that:*

$$N^*_{CT,c,\hat{c}} = \sum_{x_j \in \overline{\mathcal{X}}^*_{DTC,c}} \sum_{\mathcal{Y}_{\hat{c}}:\substack{\hat{c}\in\mathcal{C}\\ \hat{c}\neq c}} \left[\frac{\sum_{y_i \in \mathcal{Y}_{\hat{c}}} t_{y_i \cap x_j}}{(t_{e,j} - t_{s,j})} \geq \rho_{CTTC} \right] \quad (4)$$

where sets $\mathcal{Y}_{\hat{c}}$ select the ground truth for each class $\hat{c} \neq c$. Notation $[\cdot]$ represents the Iverson bracket, which denotes 1 when the enclosed condition is satisfied and 0 otherwise.



eFPR and eTPR

$$\textbf{TP Ratio: } r^*_{\mathrm{TP},c} = \frac{N^*_{\mathrm{TP},c}}{|\mathcal{Y}_c|} \qquad \textbf{FP Rate: } R^*_{\mathrm{FP},c} = \frac{N^*_{\mathrm{FP},c}}{T_{\mathcal{Y}}}$$

$$\textbf{CT Rate: } R^*_{\mathrm{CT},c,\hat{c}} = \frac{N^*_{\mathrm{CT},c,\hat{c}}}{\sum_{y_i \in \mathcal{Y}_{\hat{c}}} (t_{e,i} - t_{s,i})} \tag{5}$$

where $T_{\mathcal{Y}}$ is the total duration of dataset $\mathcal{Y}$. Thus, TP performance is measured as a proportion of detected events, whereas FP and CT performances are rates per unit of time, consistently with keyword spotting evaluation practice [12]. $R^*_{\mathrm{FP},c}$ relates to the total duration of the dataset, whereas $R^*_{\mathrm{CT},c,\hat{c}}$ is only relevant to target class labels.

Moreover, cross-triggers against identified sound classes may trigger more negative user experience than less identifiable FPs [11], thus justifying the definition of the *effective FP rate (eFPR)* as:

$$\textbf{eFPR: } e^*_c \triangleq R^*_{\mathrm{FP},c} + \alpha_{\mathrm{CT}} \frac{1}{|\mathcal{C}| - 1} \sum_{\substack{\hat{c} \in \mathcal{C} \\ \hat{c} \neq c}} R^*_{\mathrm{CT},c,\hat{c}} \tag{6}$$

where weighting parameter α_{CT} represents the cost of CTs on user experience in the SED application under evaluation.

However, the *stability of performance across classes* is of interest for evaluation: systems with much smaller variations in the TP ratio across classes may be preferable due to having better performance for the worst performing (or most difficult) class. For this reason, the *effective TP ratio (eTPR)* is defined using both the mean and the standard deviation of TP ratios across classes, such that:

$$\mu_{\mathrm{TP}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} r_{\mathrm{TP},c} \quad \sigma_{\mathrm{TP}} = \sqrt{\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left(r_{\mathrm{TP},c} - \mu_{\mathrm{TP}}\right)^2} \tag{8}$$

$$\textbf{eTPR: } \quad r(e) \triangleq \mu_{\mathrm{TP}}(e) - \alpha_{ST} * \sigma_{\mathrm{TP}}(e) \tag{9}$$

where parameter α_{ST} adjusts the *cost of instability across classes* for the SED task under evaluation. In principle, the mean and standard