Sound Event Localization and Detection with pre-trained Audio Spectrogram Transformer and Multichannel Separation Network

Robin Scheibler, Tatsuya Komatsu, Yusuke Fujita,
and Michael Hentschel

November 1, 2022

LINE

Key Ideas of the System

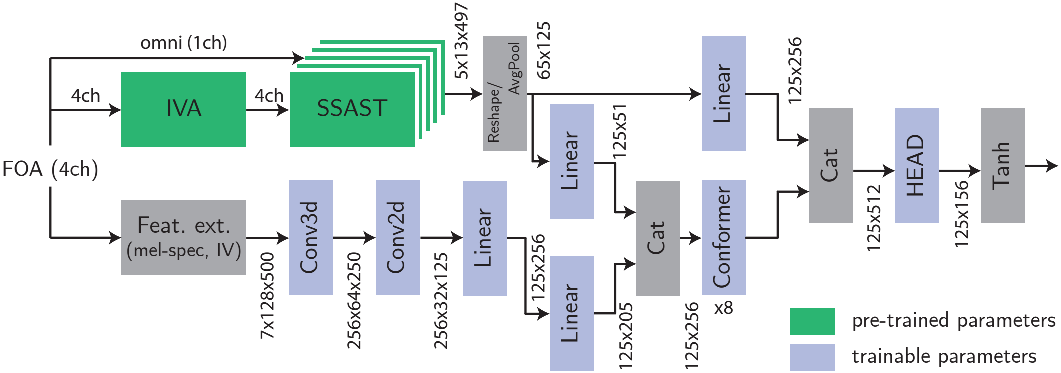
Key Components Evaluated

1. Fine-tune self-supervised audio spectrogram transfomer (AST) for SED
2. Pre-train a multichannel separation network (IVA)
3. MLP output head (vs. linear) (MLP)
4. Fine-tuning on STARSS22 only (FINE)
5. Post-processing with per class thresholds (POST)

Other

- Simulate 20 h of extra data (overlap. 4, with interference)
- Augmentations (rotations, SpecAugment)

Network Architecture



Results of Ablation Study

Model	ER↓	F↑	LE↓	LR↑	SELD ↓
Baseline (FOA) [Adavanne22]					
	0.71	0.21	29.3	0.46	0.5507
Base Network					
	0.58	0.42	19.08	0.60	0.4154
+MLP	0.59	0.41	17.02	0.61	0.4174
+FINE	0.56	0.45	16.31	0.56	0.4094
+POST	0.54	0.46	**15.87**	0.56	0.3994
Architecture III					
+AST	0.58	0.42	18.76	0.61	0.4147
+IVA	0.57	0.44	17.96	0.62	0.4037
+MLP	0.57	0.46	18.30	0.62	0.3980
+FINE	0.55	0.49	17.50	0.64	0.3792
+POST	**0.50**	**0.51**	17.13	**0.62**	**0.3644**