SOUND EVENT LOCALIZATION AND DETECTION WITH PRE-TRAINED AUDIO SPECTROGRAM TRANSFORMER AND MULTICHANNEL SEPARATION NETWORK

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

LINE

LINE DCASE2022 Task3 System

Abstract—We describe our system submitted to the DCASE Challenge 2022 Task 3. The system uses features extracted using a fine-tuned Audio Spectrogram Transformer [1] and a pre-trained multichannel separation model [2]. We compare three different ways of incorporating the features in a SELD network.

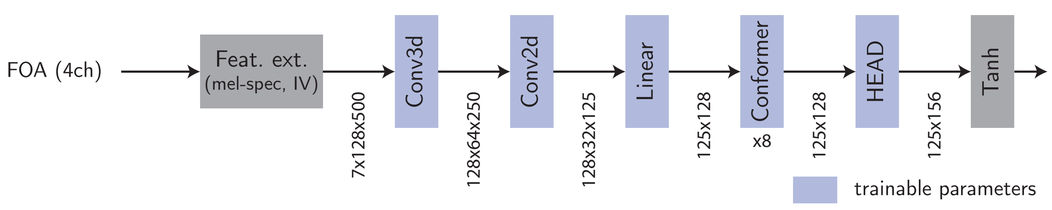
Architecture

Features

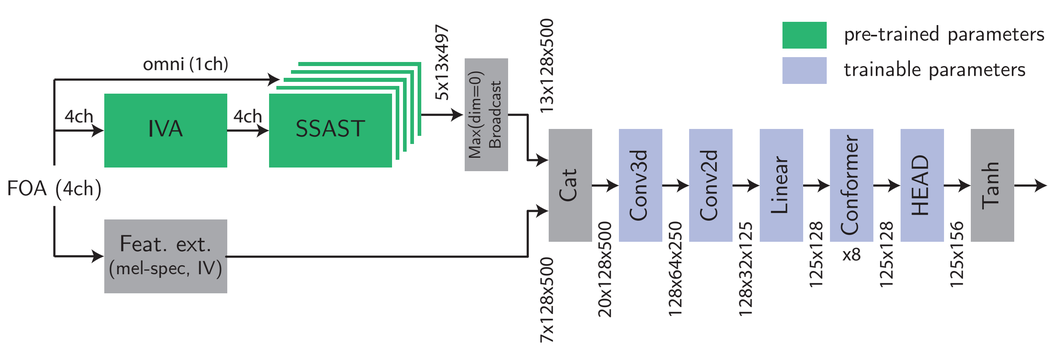
We use the **first-order ambisonics (FOA)** recordings because they are free of spatial aliasing up to 9 kHz. We then compute,

- Log-mel spectrograms (mel-spec) of each channel (4 ch.)
- Intensity vectors (IV, 3 ch.)

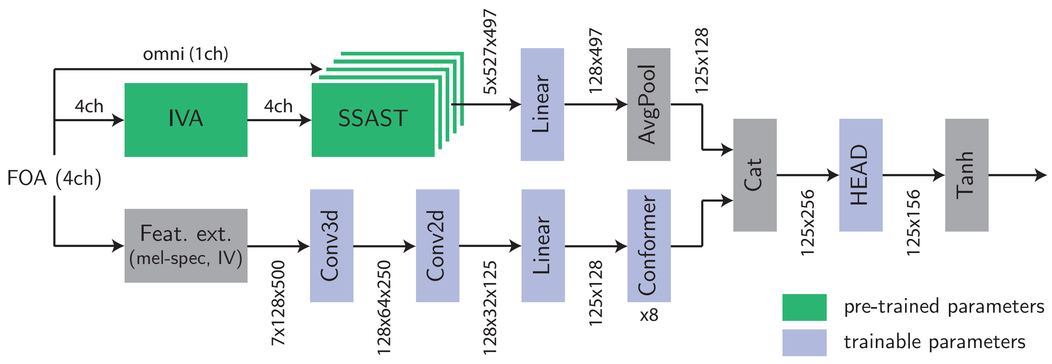
Networks



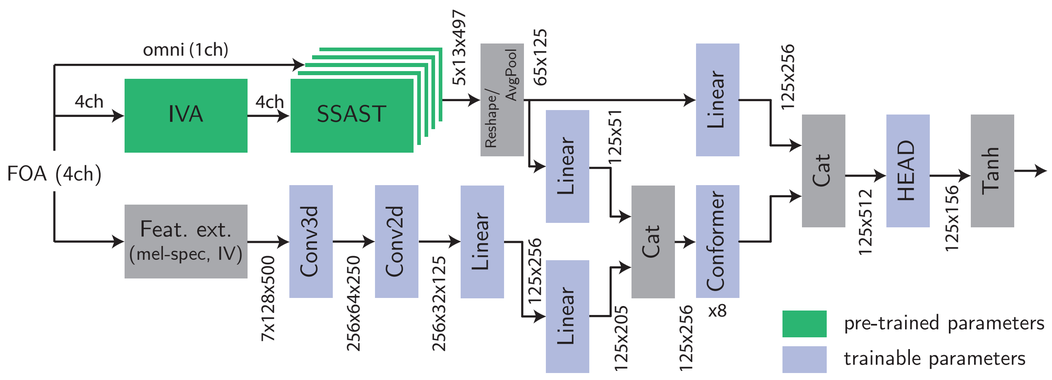
Base architecture (3.8 M param.)



Architecture I (3.8 M param.)



Architecture II (3.9 M param.)



Architecture III (15.3 M param.)

Details

- Loss function: Multi-ADPIT with 4 tracks
- Output HEAD: linear or MLP
- Output frame interval: 40 ms
- FINE-tuning on the STARSS22 dataset
- POST-processing: DOA deduplication, voting, per-class thresholds

Datasets

Name	Ref	Type	Ov.	Inter.	Train	Val.
STARSS22	[5]	Rec.	5	$\sim 4, 5$	2.9 h	2.0 h
Synth1	[6]	Sim.	2	0	20.0 h	—
Synth2		Sim.	4	1	20.0 h	—

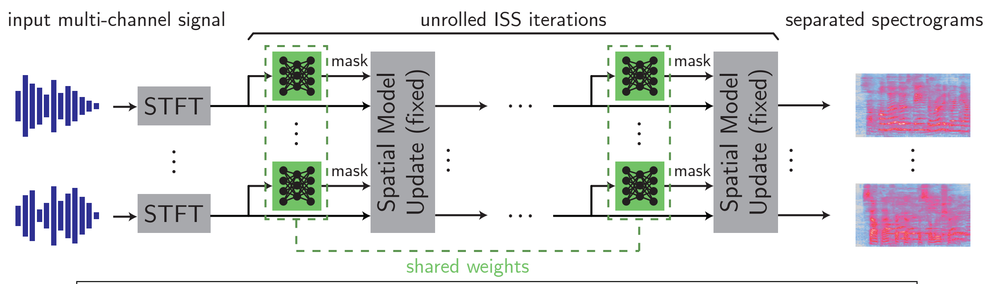
Table 1: Datasets. Ov.: maximum number of overlapping event. Inter.: # of interfering out-of-classes events. Rec.: recorded. Sim.: simulated.

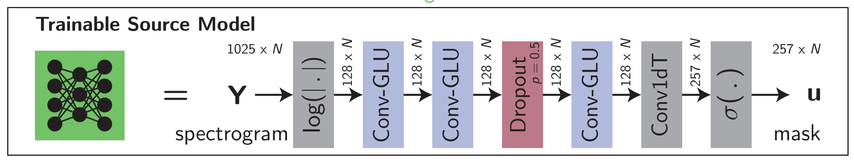
Proposed Features

Independent Vector Analysis

We exploit recent progresses in multichannel separation

- Combination of IVA and DNN source model [2]
- IVA helps sound event detection (SED) [3]
- Spatial loss for IVA requires only DOA of sources [4]





We learn the source model end-to-end from the SELD dataset using a spatial loss [4]. The model has 2.4 M parameters.

Self-supervised Audio Spectrogram Transformer (SSAST)

The SSAST is a general audio classification system pre-trained in a self-supervised manner on AudioSet [1].
We fine-tune it on the official Task 3 dataset (STARSS22 + Synth1).

- Input: spectrogram
- Output: class presence probability vector ($13 \times T$)
- Task: SED, same class events are merged
- Num. param.: 87.2 M

Ablation Study

Model	ER↓	F↑	LE↓	LR↑	SELD ↓
Baseline (FOA) [7]	0.71	0.21	29.3	0.46	0.5507
Base Network					
	0.578	0.421	19.083	0.602	0.4154
+MLP	0.594	0.412	17.015	0.608	0.4174
+FINE	0.561	0.451	16.314	0.563	0.4094
+POST	0.535	0.464	**15.869**	0.562	0.3994
Architecture I					
+AST	0.575	0.423	18.752	0.591	0.4164
+IVA	0.574	0.418	17.809	0.582	0.4182
+MLP	0.584	0.455	17.331	0.606	0.4050
+FINE	0.562	0.469	16.881	0.616	0.3928
+POST	0.519	0.480	16.375	0.598	0.3830
Architecture II					
+AST	0.572	0.424	18.130	0.604	0.4111
+IVA	0.589	0.414	18.016	0.611	0.4160
+MLP	0.592	0.445	18.156	**0.641**	0.4020
+FINE	0.534	0.478	17.163	0.595	0.3891
+POST	0.516	0.497	16.551	0.603	0.3768
Architecture III					
+AST	0.579	0.417	18.785	0.607	0.4147
+IVA	0.572	0.437	17.957	0.621	0.4037
+MLP	0.567	0.460	18.294	0.616	0.3980
+FINE	0.551	0.493	17.505	0.639	0.3792
+POST	**0.500**	**0.514**	17.131	0.624	**0.3644**

References

[1] Gong et al., AAAI, Feb. 2022.
[2] Scheibler and Togami, ICASSP, Jun. 2021.
[3] Scheibler et al., EUSIPCO, Aug. 2021.
[4] Saijo and Scheibler, INTERSPEECH, Sep. 2022.
[5] Politis et al., arXiv:2206.01948, Jun. 2022.
[6] Politis et al., https://doi.org/10.5281/zenodo.6406873, Apr. 2022.
[7] Adavanne , https://github.com/sharathadavanne/seld-dcase2022, 2022.