CoLoC: Conditioned Localizer and Classifier for Sound Event Localization and Detection

Sławomir Kapka, Jakub Tkaczuk
s.kapka@samsung.com, j.tkaczuk@samsung.com

Samsung R&D Institute Poland

SELD problem overview

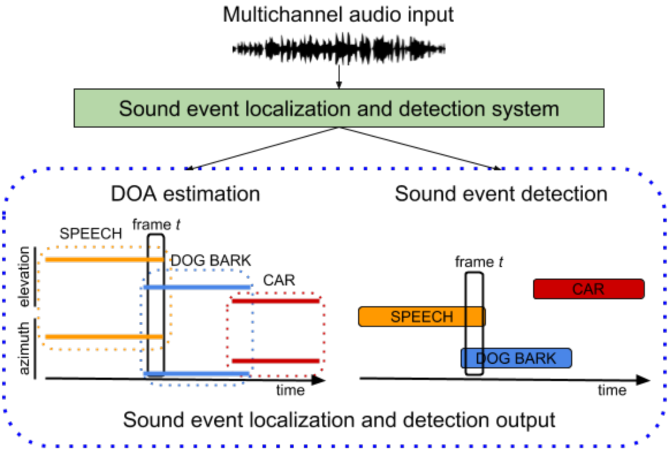


Figure: https://dcase.community/challenge2022/task-sound-event-localization-and-detection-evaluated-in-real-spatial-sound-scenes

SELD is about sets

In SELD we predict **sets** of events

$$P(\{\text{class}_i \wedge \text{location}_i\}_{i=1..k}|\text{audio}),$$

where $k \leq N$ and N is the max number of overlapping events.

Usual approach	Our approach
Permutation Invariant Training (PIT)	**Sequential Set Generation (SSG)**

SSG Localizer $\mathbb{E}(l|X, \{l_i\}_i)$

- Start from empty set $\emptyset$
- Step 1: Get location of first event $l_1 = \mathbb{E}(l|X, \emptyset)$
- Step k: Get location of k'th event by conditioning on previously predicted locations $l_k = \mathbb{E}(l|X, \{l_i\}_{i=1..k-1})$
- Terminate when given a special token $\tau = \mathbb{E}(l|X, \{l_i\}_{i=1..n})$

Location-conditioned classifier $P(c|X, l)$

Based on the output from the localizer we then classify events corresponding to predicted DOAs.

$$P(c_i \wedge l_i|X) = \boldsymbol{P(c_i|X, l_i)} \cdot P(l_i|X).$$

In summary, given:

- SSG localizer $\mathbb{E}(l|X, \{l_i\}_i)$,
- Location-conditioned classifier $P(c|X, l)$,

we can resolve SELD task.

Stacked-Tracks metadata format

Original Metadata

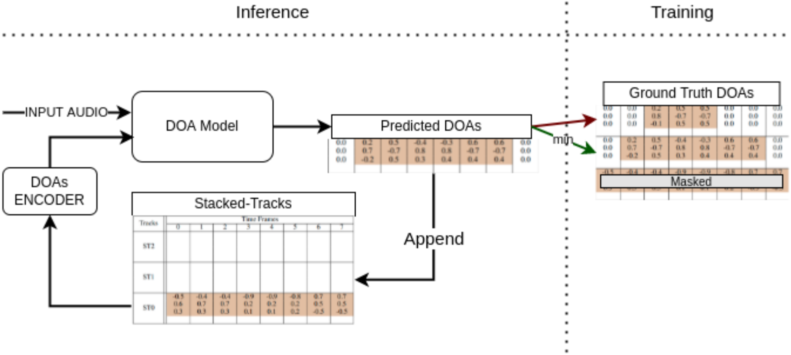
Tracks	Time Frames							
	0	1	2	3	4	5	6	7
T4		0.2 0.7 -0.2 3	0.2 0.8 -0.1 3					
T3			0.5 -0.7 0.5 7	0.5 -0.7 0.5 7	0.5 -0.7 0.5 7	0.6 -0.7 0.4 7	0.6 -0.7 0.4 7	
T2	-0.5 0.6 0.3 3	-0.4 0.7 0.3 3	-0.4 0.7 0.3 3	-0.4 0.8 0.3 3	-0.3 0.8 0.4 3			
T1							0.7 0.5 -0.5 11	0.7 0.5 -0.5 11
T0				-0.9 0.2 0.1 8	-0.9 0.2 0.1 8	-0.8 0.2 0.2 8		

↓↓↓ Stacking ↓↓↓

Stacked-Tracks

Tracks	Time Frames							
	0	1	2	3	4	5	6	7
ST2	0.0 0.0 0.0 13	0.0 0.0 0.0 13	0.2 0.8 -0.1 3	0.5 -0.7 0.5 7	0.5 -0.7 0.5 7	0.0 0.0 0.0 13	0.0 0.0 0.0 13	0.0 0.0 0.0 13
ST1	0.0 0.0 0.0 13	0.2 0.7 -0.2 3	0.5 -0.7 0.5 7	-0.4 0.8 0.3 3	-0.3 0.8 0.4 3	0.6 -0.7 0.4 7	0.6 -0.7 0.4 7	0.0 0.0 0.0 13
ST0	-0.5 0.6 0.3 3	-0.4 0.7 0.3 3	-0.4 0.7 0.3 3	-0.9 0.2 0.1 8	-0.9 0.2 0.1 8	-0.8 0.2 0.2 8	0.7 0.5 -0.5 11	0.7 0.5 -0.5 11

CoLoC: Localizer



Predicted DOAs

0.0 0.0 0.0	0.2 0.7 -0.2	0.5 -0.7 0.5	-0.4 0.8 0.3	-0.3 0.8 0.4	0.6 -0.7 0.4	0.6 -0.7 0.4	0.0 0.0 0.0

Ground Truth DOAs

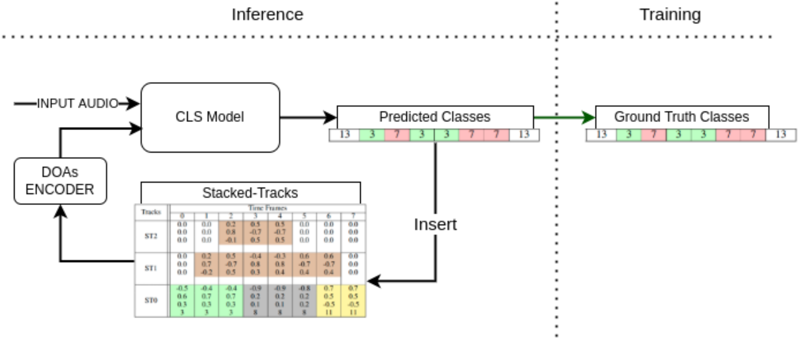
0.0 0.0 0.0	0.0 0.0 0.0	0.2 0.8 -0.1	0.5 -0.7 0.5	0.5 -0.7 0.5	0.0 0.0 0.0	0.0 0.0 0.0	0.0 0.0 0.0
0.0 0.0 0.0	0.2 0.7 -0.2	0.5 -0.7 0.5	-0.4 0.8 0.3	-0.3 0.8 0.4	0.6 -0.7 0.4	0.6 -0.7 0.4	0.0 0.0 0.0
-0.5	-0.4	-0.4	-0.9	-0.9	-0.8	0.7	0.7

Masked

Stacked-Tracks

Tracks	Time Frames							
	0	1	2	3	4	5	6	7
ST2								
ST1								
ST0	-0.5 0.6 0.3	-0.4 0.7 0.3	-0.4 0.7 0.3	-0.9 0.2 0.1	-0.9 0.2 0.1	-0.8 0.2 0.2	0.7 0.5 -0.5	0.7 0.5 -0.5

CoLoC: Classifier



Predicted Classes

13	3	7	3	3	7	7	13

Ground Truth Classes

13	3	7	3	3	7	7	13

Stacked-Tracks

Tracks	Time Frames							
	0	1	2	3	4	5	6	7
ST2	0.0 0.0 0.0	0.0 0.0 0.0	0.2 0.8 -0.1	0.5 -0.7 0.5	0.5 -0.7 0.5	0.0 0.0 0.0	0.0 0.0 0.0	0.0 0.0 0.0
ST1	0.0 0.0 0.0	0.2 0.7 -0.2	0.5 -0.7 0.5	-0.4 0.8 0.3	-0.3 0.8 0.4	0.6 -0.7 0.4	0.6 -0.7 0.4	0.0 0.0 0.0
ST0	-0.5 0.6 0.3 3	-0.4 0.7 0.3 3	-0.4 0.7 0.3 3	-0.9 0.2 0.1 8	-0.9 0.2 0.1 8	-0.8 0.2 0.2 8	0.7 0.5 -0.5 11	0.7 0.5 -0.5 11

Results

We report our results on STARSS22 developement test dataset

Table: Official DCASE metrics; the **boldface** denotes the best scores.

	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
Baseline	**0.71**	21%	29.3°	46%
max-ov3	0.85	32%	24.7°	**51%**
max-ov2	0.76	**33%**	**24.6°**	49%

Thank You!

Questions, comments? Contact me: s.kapka@samsung.com