DCASE2021 Workshop on Detection and Classification of Acoustic Scenes and Events

WHAT MAKES SOUND EVENT LOCALIZATION AND DETECTION DIFFICULT? INSIGHTS FROM ERROR ANALYSIS

T. N. T. Nguyen*, K. N. Watcharasupat*, Z. J. Lee, N. K. Nguyen, D. L. Jones†, W. S. Gan* nguyenth003@e.ntu.edu.sg

*School of EEE, Nanyang Technological University, Singapore

†Department of ECE, University of Illinois at Urbana-Champaign, USA

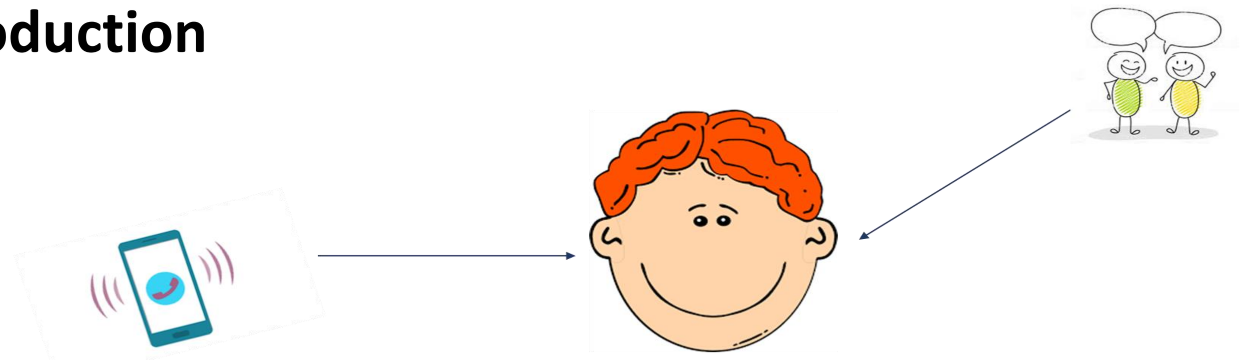
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Electrical and Electronic Engineering
College of Engineering

Acknowledgement

This research was supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2017-T2-2-060.
K. N. Watcharasupat acknowledges the support from the CN Yang Scholars Programme, Nanyang Technological University, Singapore.

Introduction



Problem: Polyphonic sound event detection and localization (SELD)

Challenges of SELD task : Noise, reverberation, interference, polyphony, non-stationarity, association between sound classes and directions of arrival (DOA), etc..

Objective: Understand the major sources of errors in SELD task.

Approach: Error analysis on different SELD models and datasets.

Proposed error analysis method

- Focus: polyphony, moving source, class-location interdependence, and class-wise performance.
- Use two public datasets for SELD
  - TAU-NIGENS Spatial Sound Events 2020 (TNSSE 2020) [1]
  - TAU-NIGENS Spatial Sound Events 2021 (TNSSE 2021) [2]
- Use two SELD systems that ranked second in the team categories of the DCASE 2020 and 2021 SELD challenges, respectively.
- System outputs are divided into segments of 1 second. Ground truth is used to group these segments into different categories such as polyphony (0, 1, 2, and 3 sources), static and moving sources, etc., in order to evaluate the SELD performance in each category.

SELD Datasets

Characteristics	TNSSE 2020	TNSSE 2021
Channel format	FOA	FOA
Moving sources	✓	✓
Ambiance noise	✓	✓
Reverberation	✓	✓
Unknown interferences	×	✓
Maximum degree of polyphony	2	3
Number of target sound classes	14	12
Evaluation split	eval	test

References

[1] A. Politis, S. Adavanne, and T. Virtanen. "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection". In DCASE Workshop, 2020.

[2] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection," arXiv, 2021.

[3] T. N. T. Nguyen, D. L. Jones, and W. Gan, "Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking," In DCASE Workshop, 2020.

[4] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "DCASE 2021 Task 3: Spectrotemporally-aligned Features for Polyphonic Sound Event Localization and Detection," DCASE2021 Challenge, Tech. Rep., 2021.

Evaluation metrics

- Sound event detection (SED): Location-dependent error rate ($ER_{\leq T}$) and F1 score ($F_{\leq T}$). T is DOA threshold (typical value is 20°)
  - $ER_{\leq T} = substitution + deletion + insertion\ error\ rate$
  - $F_{\leq T} = 2\frac{precison * recall}{precison + recall}$
- DOA estimation: Class-dependent localization error (LE_{CD}) (in degrees) and localization recall (LR_{CD})

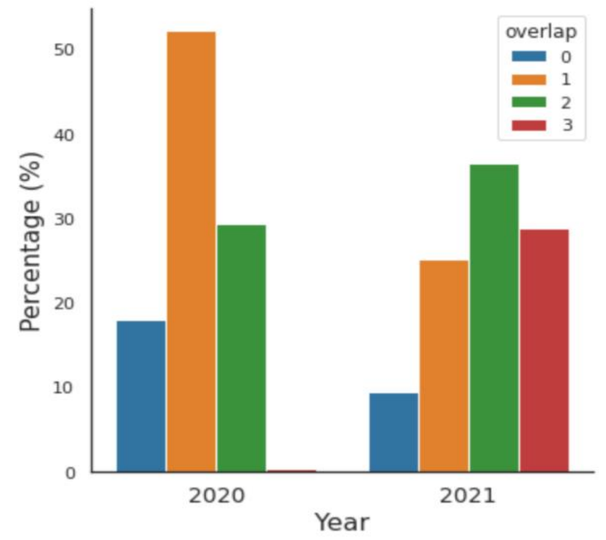
SELD systems

Year	System	$ER_{\leq 20°}$	$F_{\leq 20°}$	LE_{CD}	LR_{CD}
2020 (eval)	Baseline [9]	0.69	0.413	23.1°	0.624
	#1: USTC'20 [25]	**0.20**	**0.849**	**6.0°**	0.885
	#2: NTU'20 [27]	0.23	0.820	9.3°	**0.900**
2021 (test)	Baseline [11]	0.73	0.307	24.5°	0.448
	#1: Sony'21 [26]	0.43	0.699	**11.1°**	0.732
	#2: NTU'21 [23]	**0.37**	**0.737**	11.2°	**0.741**

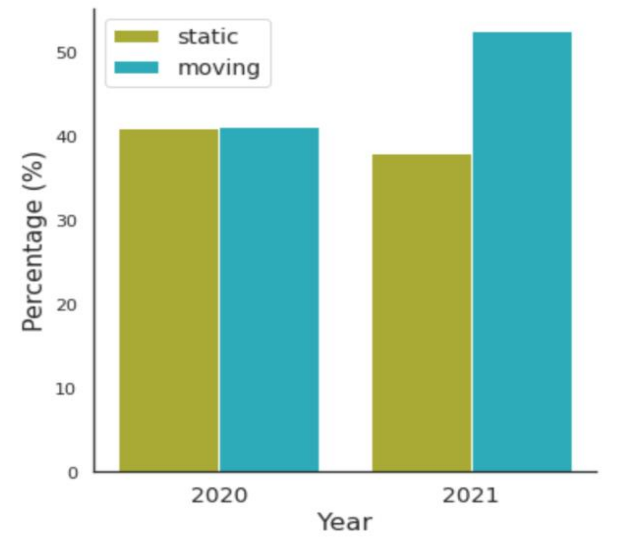
- NTU'20: ensemble of sequence matching networks, evaluated on the evaluation split of the TNSSE2020 dataset [3].
- NTU'21: ensemble of SELDnet-like networks, trained on SALSA features, evaluated on the test split of the TNSSE2021 dataset [4].

Dataset distributions

Segment-wise polyphonic and static distribution of TNSSE2020 and TNSSE2021 datasets



(a) Polyphonic distribution



(b) Static vs moving

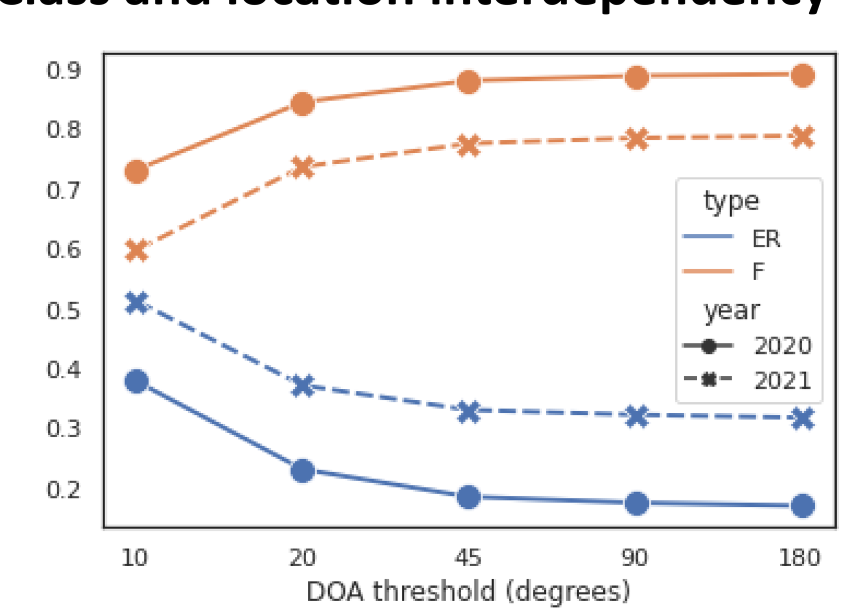
Effect of polyphony

Metrics	2020: 1	2020: 2	2020: All	2021: 1	2021: 2	2021: 3	2021: All
↓ $ER_{\leq 20°}$	**0.108**	0.331	0.232	0.349	**0.338**	0.394	0.372
↓ Substitution	**0.029**	0.072	0.052	**0.093**	0.104	0.129	0.114
↓ Deletion	**0.042**	0.155	0.103	**0.091**	0.137	0.182	0.152
↓ Insertion	**0.038**	0.104	0.078	0.164	0.096	**0.083**	0.105
↑ $F_{\leq 20°}$	**0.930**	0.765	0.845	**0.784**	0.763	0.704	0.737
↑ Precision	**0.932**	0.788	0.875	0.757	**0.780**	0.746	0.756
↑ Recall	**0.928**	0.743	0.833	**0.813**	0.747	0.666	0.719
↓ LE_{CD}	**5.6**	13.4	9.4	**6.8**	10.3	13.5	11.2
↑ LR_{CD}	**0.930**	0.775	0.846	**0.816**	0.764	0.701	0.741

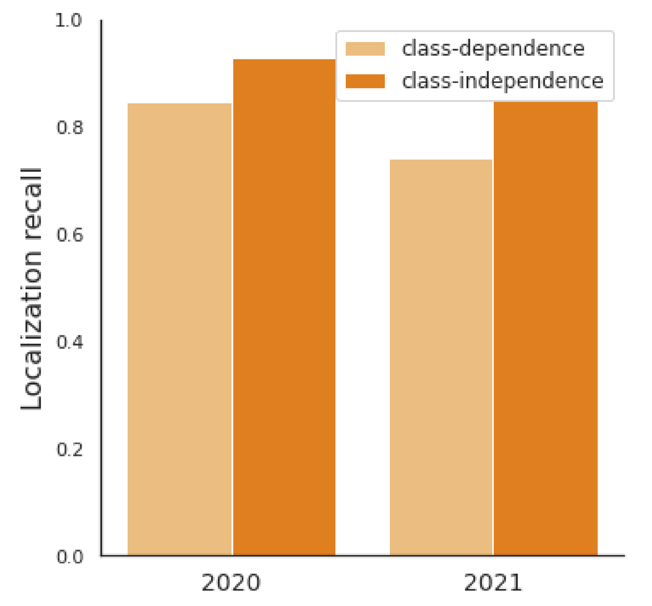
Effect of moving sound sources

Metrics	2020: Static	2020: Moving	2020: All	2021: Static	2021: Moving	2021: All
↓ $ER_{\leq 20°}$	**0.214**	0.239	0.232	0.379	**0.357**	0.372
↑ $F_{\leq 20°}$	**0.854**	0.841	0.845	0.731	**0.745**	0.737
↓ LE_{CD}	**8.7**	10.0	9.4	**10.5**	11.7	11.2
↑ LR_{CD}	**0.847**	0.846	0.846	0.725	**0.751**	0.741
↓ $ER_{\leq 180°}$	**0.166**	0.168	0.171	0.334	**0.298**	0.318
↑ $F_{\leq 180°}$	**0.898**	0.891	0.892	0.778	**0.800**	0.789

Class and location interdependency



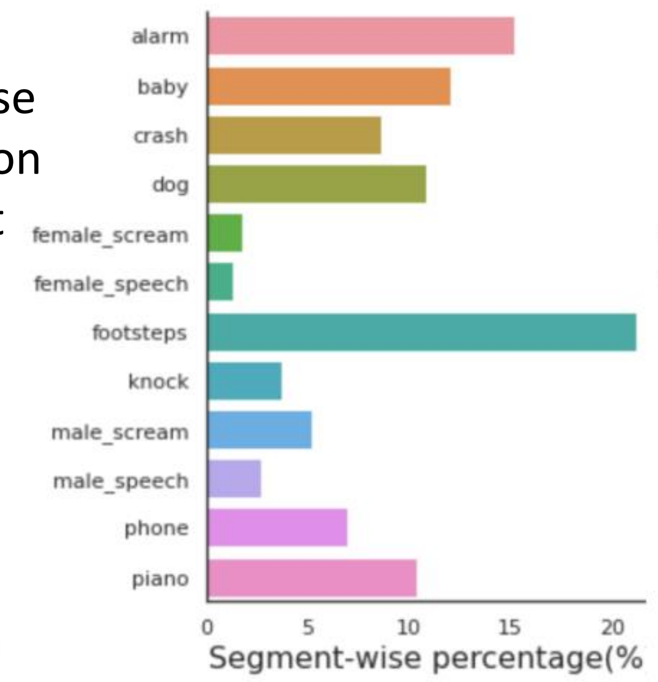
SED performance across different DOA threshold



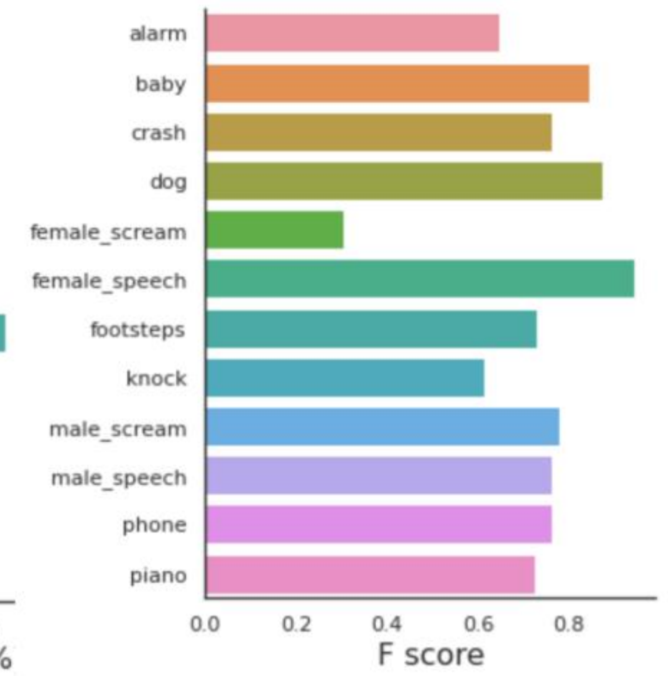
Localization recall by class dependencies

Class-wise performance

a) Segment-wise class distribution of the test split of the TNSSE 2021 dataset
b) Class-wise location-dependent F score of the NTU'21 system



(a) Class distribution



(b) Class-wise $F_{\leq 20°}$

Conclusion

- Polyphony and unknown interferences appear to be the biggest challenges for SELD task as SELD systems struggle to detect all events of interests (low recall and high deletion error).
  - Unknown interferences lead to more substitution errors.
  - $ER_{\leq T}$ is lowest for the polyphonic case that dominates the dataset.
- Moving sources mainly increase the localization errors.
- It is very challenging to further tighten the DOA threshold.
- High segment-wise representation of a class does not necessarily translate to high SED performance.