NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

A Sequence Matching Network for Polyphonic Sound Event Localization and Detection

Paper: 3583

Session: AUD-L3 Acoustic Event Detection

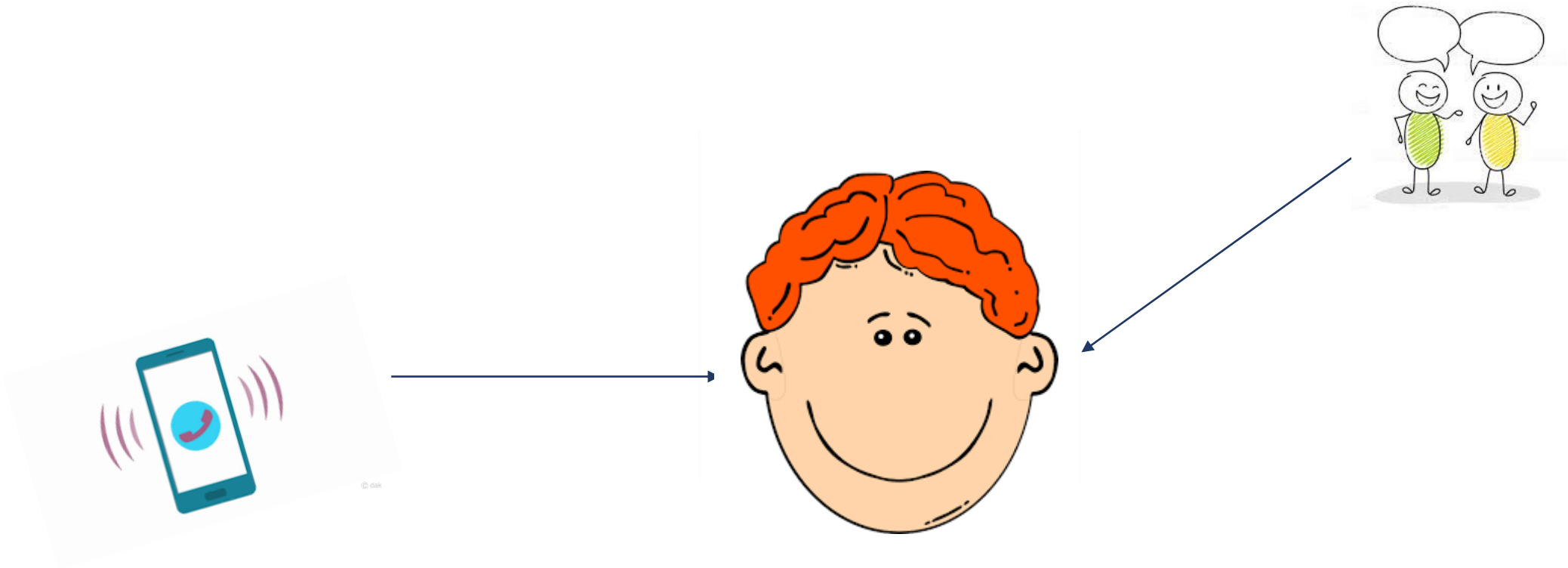
T. N. T. Nguyen*, D. L. Jones†, W. S. Gan*

*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

†Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA

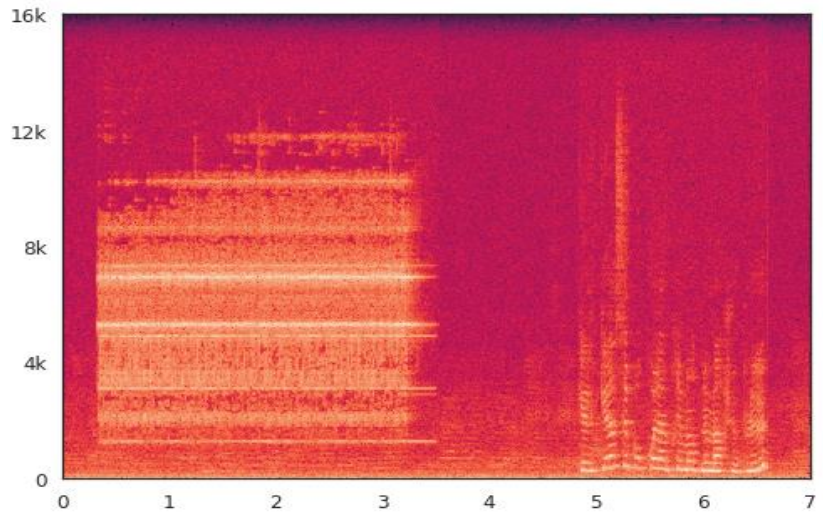
6 May 2020 - ICASSP

Sound event localization and detection

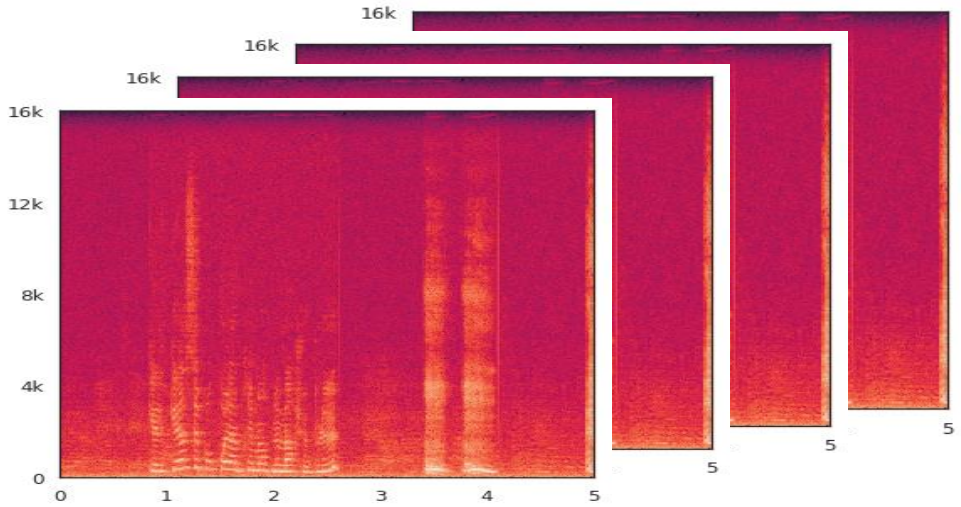


Sound event localization and detection (SELD)

Sound event detection (SED)



Direction-of-arrival (DOA) estimation

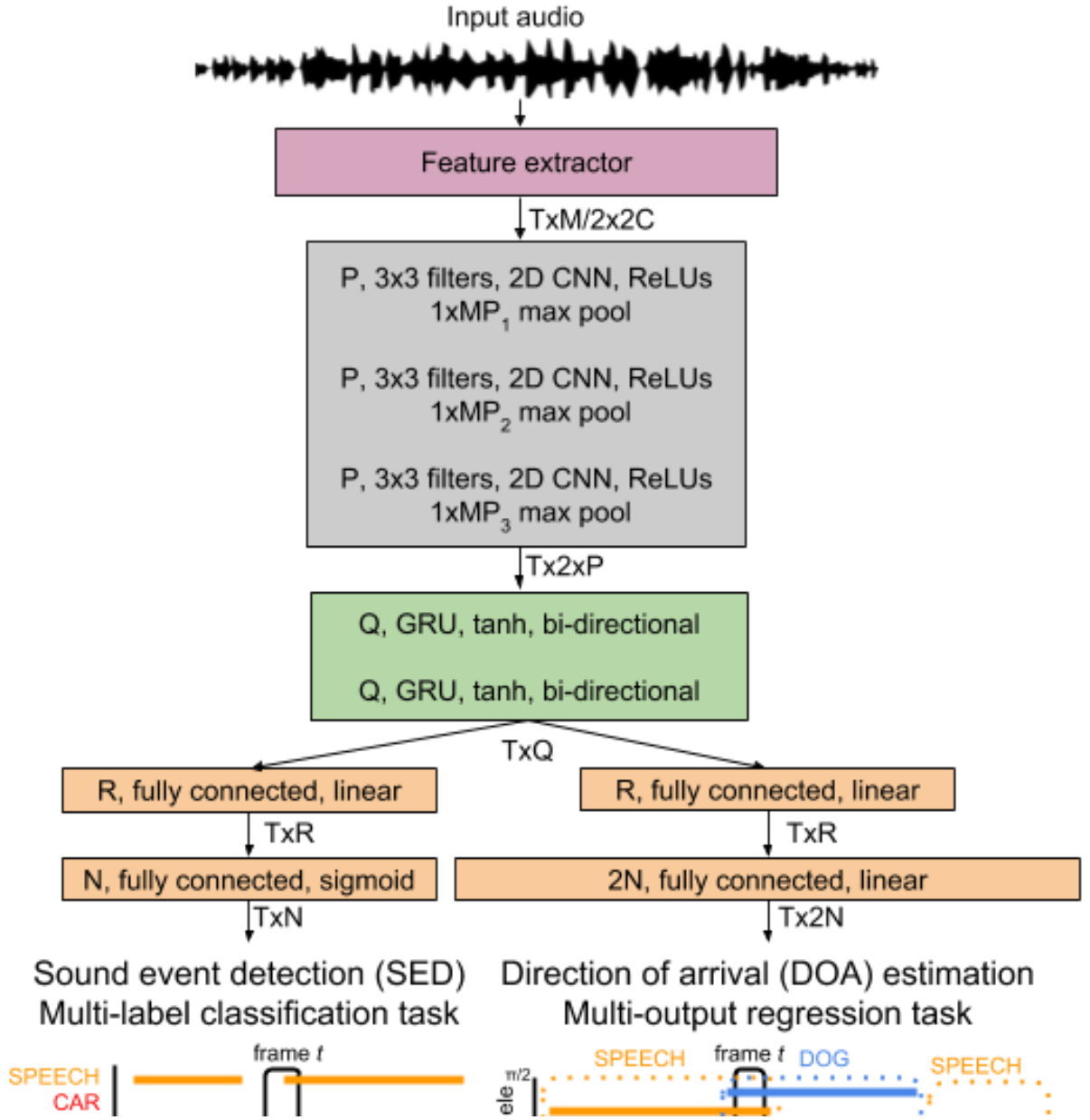


Signal Support

Spatial Filtering

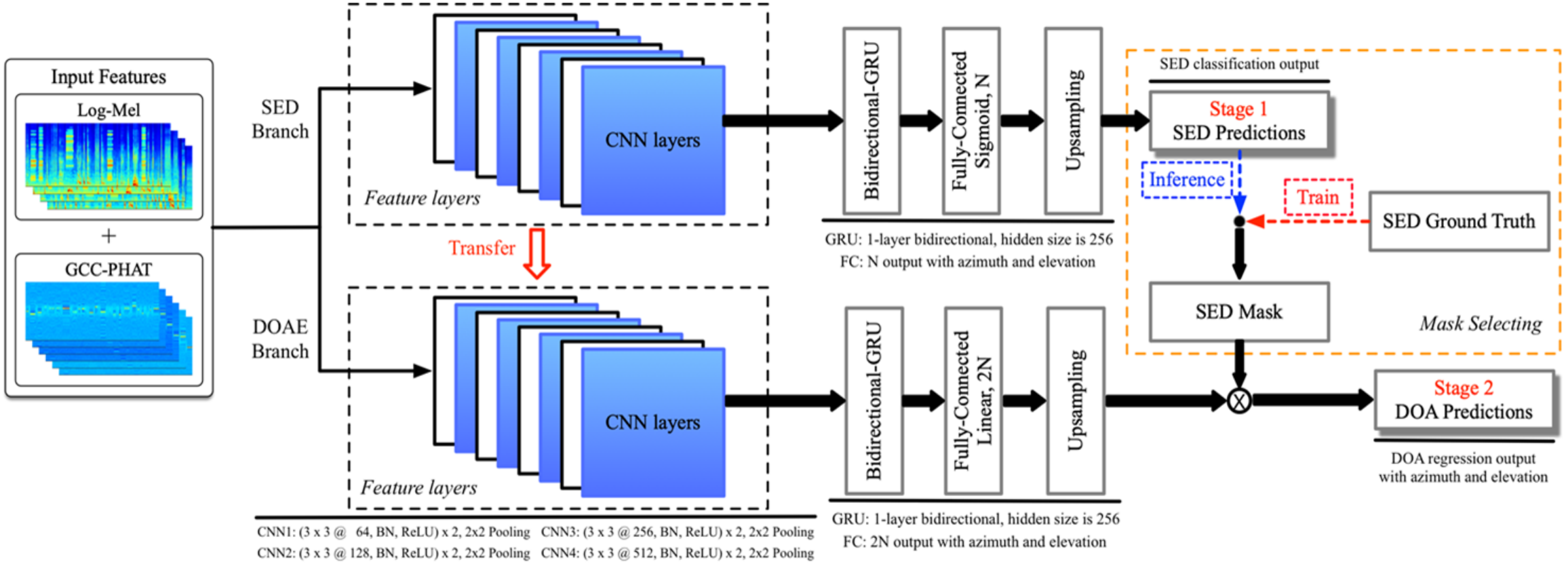
SELDnet: joint SED and DOA estimation

The losses of SED and DOA estimation task are jointly optimized.



S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48,March 2019

Two-stage SELD

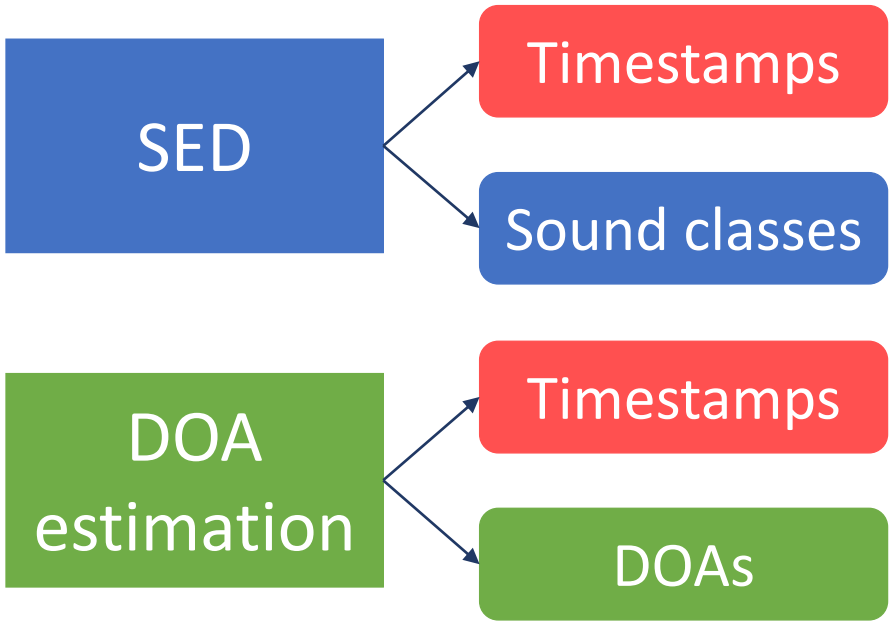


Y. Cao, Q. Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

Observation

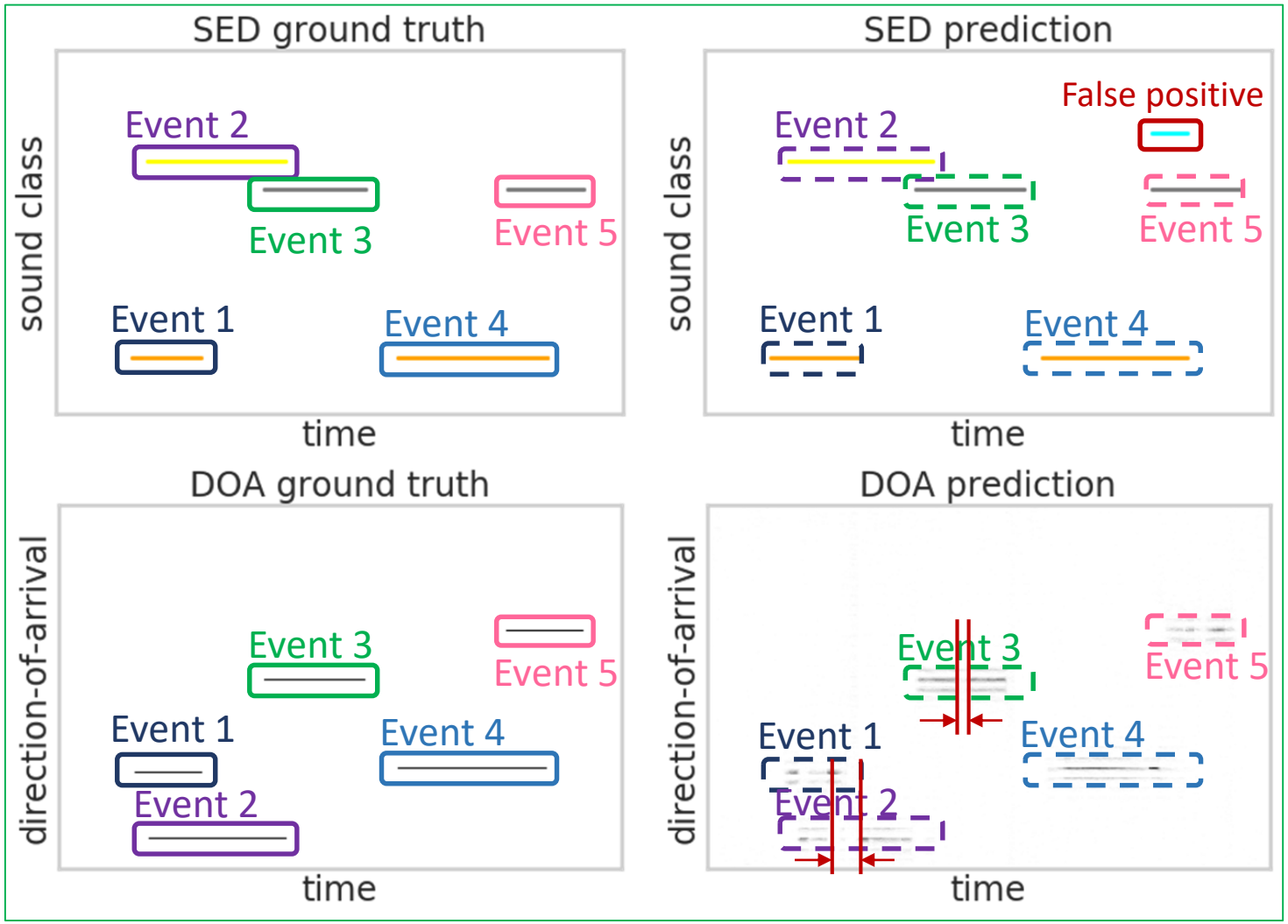
Sound event

1. Timestamp (onset, offset)
2. Sound class
3. DOA

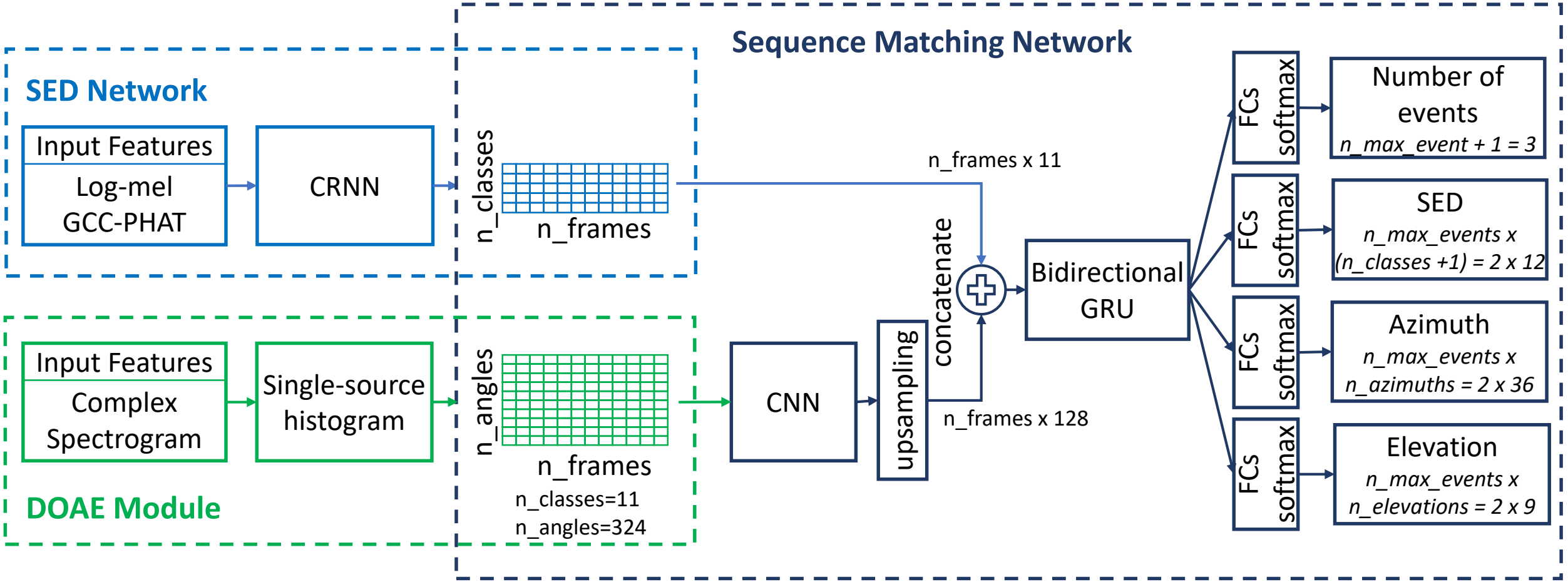


Ground-truth Sequences

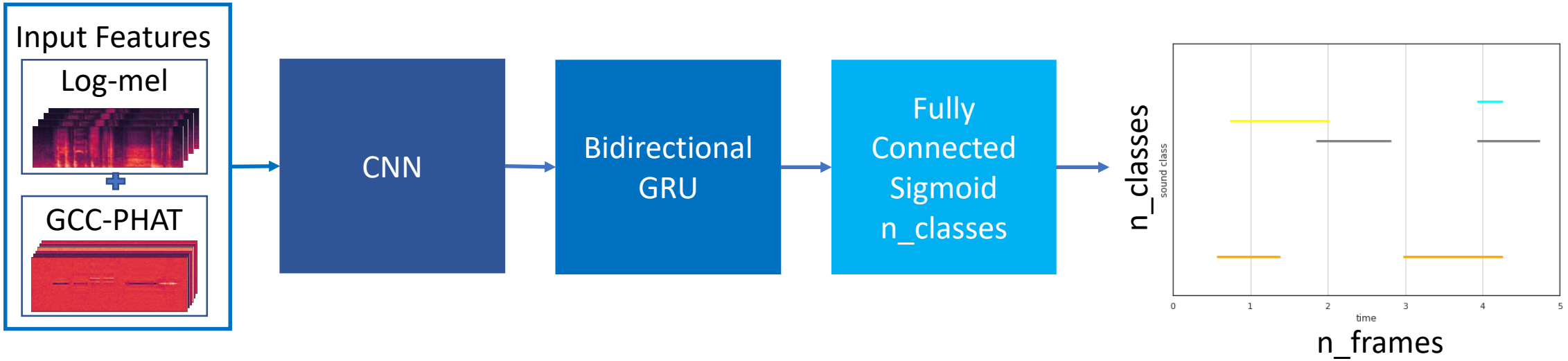
Output Sequences



A sequence matching network (SMN) for SELD



Improved SED network

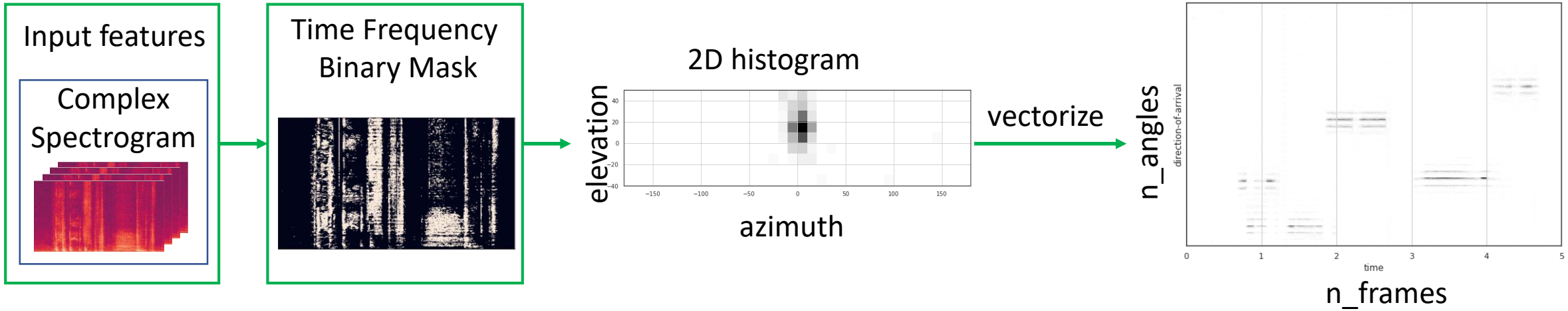


Improvement: data augmentation

Use random cut out with the same mask for all logmel and GCC-PHAT channels

Y. Cao, Q. Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

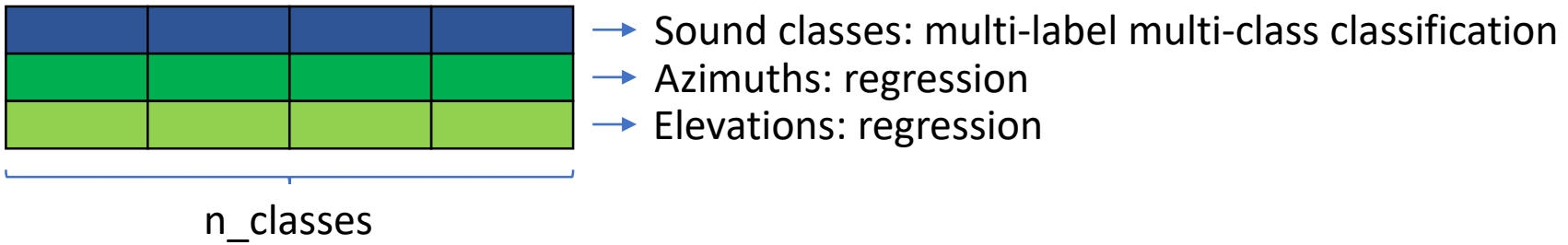
DOA estimation



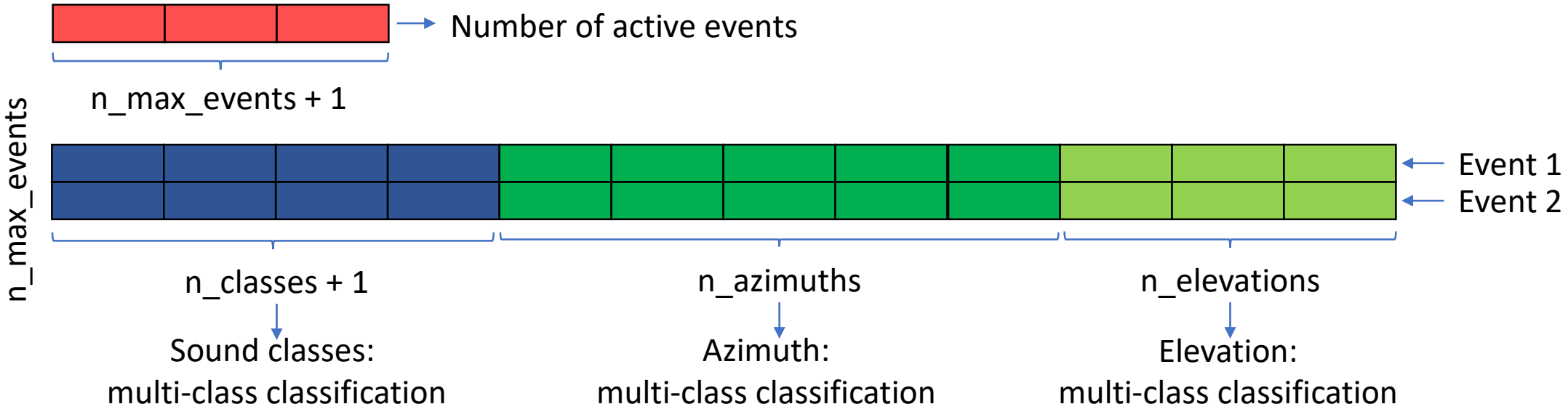
T. N. T. Nguyen, S. K. Zhao, and D. L. Jones, “Robust doa estimation of multiple speech sources,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2287–2291.

Output format

Conventional output format



Proposed output format



Dataset

- TAU Spatial Sound Events 2019 – Ambisonic (DCASE 2019 – task 3)

Development: 400 one-minute recordings	Evaluation: 100 one-minute recording

- Data are synthesized using recorded room impulse responses (RIRs) and clean signals. Maximum 2 overlapping sources in one frame
- SED: 11 indoor sound classes
- DOA: 324 angles
  - Azimuth between [0°, 360°), resolution 10°: 36 angles
  - Elevation between [-40°, 40°], resolution 10°: 9 angles

Evaluation metrics:

SED

- Segment-based error rate
- Segment-based F1 score
- Segment length: 1 second

DOA estimation

- Frame-based DOA error
- Frame-based frame recall
- Frame length: 0.02 second

New evaluation metrics: to account for correct matching of sound classes and DOAs

1. Matching F1 score (frame-based)

SED ground truth	SED prediction: 1	SED prediction: 0
1	Correct DOA estimate: a Incorrect DOA estimate: b	SED false negative: d
0	SED false positive: c	SED true negative: e

$$matching\ precision(mp) = \frac{a}{a+b+c}$$

$$matching\ recall(mr) = \frac{a}{a+b+d}$$

$$matching\ F1 = \frac{2 * mp\ * mr}{mp + mr}$$

2. Same-class matching accuracy (frame-based)

$$matching\ accuracy\ (MA) = \frac{\#\ of\ correctly\ predicted\ frame-based\ events\ that\ have\ same\ sound\ class}{\#\ of\ ground-truth\ frame-based\ events\ that\ have\ same\ sound\ class}$$

Methods for comparison

Group	Methods	Descriptions
Baselines	SELDnet	joint SED and DOAE [1], with log-mel and GCC-PHAT input features [2]
	Two-stage	two-stage SELD [2]
Improved baseline	Two-stage-aug	Two-stage SELD with additional random cut-out augmentation for input features
Inputs to SMNs	SED-net	the SED network of the Two-stage-aug -> SED sequences for SMNs
	DOA-hist	single-source histogram for DOA estimation -> DOA sequences for SMNs [5]
Proposed	**SMN**	SMN with the conventional SELD output format
	SMN-event	SMN with new output format
Top DCASE SELD team ranking	Kapka-en	the consecutive ensemble of CRNN models with heuristics rules; ranked 1 [6]
	Two-stage-en	the ensemble based on two-stage training; ranked 2 [7]

SELD evaluation results

↑: the higher, the better
↓: the lower, the better

Group	Methods	SED error rate ↓	SED F1 score ↑	DOA error ↓	DOA frame rate ↑	Matching F1 score ↑	Same-class matching accuracy ↑
Baselines	SELDnet	0.212	0.880	9.75°	0.851	0.750	0.229
	Two-stage	0.143	0.921	8.28°	0.876	0.786	0.270
Improved baseline	Two-stage-aug	0.108	0.944	8.42°	0.892	0.797	0.270
Inputs to SMNs	SED-net	0.108	0.944	NA	NA	NA	NA
	DOA-hist	NA	NA	4.28°	0.825	NA	NA
Proposed	**SMN**	**0.079**	**0.958**	4.97°	0.913	**0.869**	0.359
	SMN-event	**0.079**	0.957	5.50°	0.924	0.840	**0.649**
Top DCASE team ranking	Kapka-en	**0.08**	0.947	**3.7°**	**0.968**	NA	NA
	Two-stage-en	**0.08**	0.955	5.5°	0.922	NA	NA

Conclusions

- Our proposed sequence matching networks outperformed the state-of-the-art SELDnet and the two-stage method for sound event localization and detection.
- The sequence matching network is modular and hierarchical -> improve the performance while increase the flexibility in designing and optimizing its components.
- The sequence matching networks increase the correct association between the sound classes and the corresponding DOAs in multiple-source cases. The new output format can also handle the cases where multiple sound events of the same class have different DOAs.
- The new evaluation metrics address the problem of matching sound classes and DOAs which was not achievable using the conventional SELD evaluation metrics.

References

1. S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48,March 2019
2. Y. Cao, Q. Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, “Polyphonic sound event detection and localization using a two-stage strategy,” in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.
3. S. Adavanne, A. Politis, and T. Virtanen, “A multi-room reverberant dataset for sound event localization and detection,” in *Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.
4. A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, pp. 162, 2016.
5. T. N. T. Nguyen, S. K. Zhao, and D. L. Jones, “Robust doa estimation of multiple speech sources,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2287–2291.
6. S. Kapka and M. Lewandowski, “Sound source detection, localization and classification using consecutive ensemble of CRNN models,” Tech. Rep., DCASE2019 Challenge, June 2019.
7. Y. Cao, T. Iqbal, Q. Q. Kong, M. Galindo, W. Wang, and M. D Plumbley, “Two-stage sound event localization and detection using intensity vector and generalized cross-correlation,” Tech. Rep., DCASE2019 Challenge, June 2019.

Acknowledgement

This research was conducted at Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU), which is a collaboration between Singapore Telecommunications Limited (Singtel) and Nanyang Technological University (NTU) that is funded by the Singapore Government through the Industry Alignment Fund - Industry Collaboration Projects Grant.


Q
&
A