`Micarraylib`: Software For Reproducible Aggregation, Standardization, And Signal Processing Of Microphone Array Datasets

Iran R. Roman[1], Juan Pablo Bello [1,2]

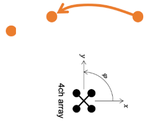
[1]Music And Audio Research Lab, New York University
[2]Center For Urban Science And Progress, New York University

Introduction

Sound Event Localization and Detection

- Machine listening has seen major advances in sound event detection (SED)
  - AST: Audio Spectrogram Transformer (2021)
- In part thanks to large datasets with annotated sound events (like AudioSet)

Large amounts of data will also be needed for Sound Event Localization and Detection (SELD)



- Microphone array hardware and datasets
  - Several useful datasets available online (30+)
  - Very diverse in hardware and content

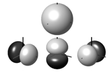

Ambeo®

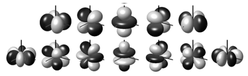

Pyramic (2018)

dataset	no. arr	capsules	length	SELD
DCASE(3) 2019 [5]	1	4	8 Hr	Yes
DCASE(3) 2020 [6]	1	4	13 Hr	Yes
DCASE(3) 2021 [7]	1	4	13 Hr	Yes
LOCATA [8]	4	63	0.5 Hr	Yes
3D-MARCo [9]	7	71	0.2 Hr	No
EigenScape [10]	1	32	11 Hr	No

Datasets with different
- Microphone hardware used
- SELD annotation conventions (or lack of)

Solution: standardization
- Microphone array recordings with B-format (ambisonics)
- SELD annotations with `Soun[D]ata:SpatialEvents`
  - `Label`
  - `Instance No.`
  - `Location`
  - `[start_time, end_time]*N`





`Micarraylib` allows users to access openly-available datasets in standard B-format ambisonics, with `Soun[D]ata:SpatialEvents` annotations

`Soun[D]ata(pre-release)`: https://github.com/soundata/soundata

Under the hood

`Micarraylib`

- After adding the dataset loader to `Soun[D]ata`:
- `Micarraylib`'s dataset object includes the capsule coordinates
- Capsule coordinates are used to compute the spherical harmonics matrix
  - θ = elevation
  - ϕ = azimuth
  - n = order
  - l = degree

$$Y_{n,l}(\theta,\phi) = X_{n,|l|} P_{n,|l|}\cos(\theta) \begin{cases} \sqrt{2}\sin(|l|\phi) & \text{if } l<0 \\ 1 & \text{if } l=0 \\ \sqrt{2}\cos(l\phi) & \text{if } l>0 \end{cases}$$

- The matrix Y is used as B-format encoder (pseudo-inverse matrix):

$$\mathbf{Y}^{\dagger} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$$ (McCormack et al., 2018)

- Datasets included (so far):
  - 2019 DCASE Task 3
  - 2020 DCASE Task 3
  - 2021 DCASE Task 3
  - LOCATA
  - 3D-MARCo
  - Eigenscape
- ~50 hours of data Total
  - ~30hrs with SELD annotations
- Shared audio format (FOA)

```python
import micarraylib as mc

datadir = '~/datasets'

datasets = [
    mc.datasets.dcase19(datadir),
    mc.datasets.dcase20(datadir),
    mc.datasets.dcase21(datadir),
    mc.datasets.locata(datadir),
    mc.datasets.marco(datadir),
    mc.datasets.eigenscape(datadir)
]

for dataset in datasets:
    dataset.load() # using the soundata API [18]

aggregate = mc.aggregators.aggregate_datasets(
    datasets,
    sr=24000,
)
```

Example use-case

Virtual capsule interpolation

- **Can a microphone capsule's recording be reconstructed using neighboring capsules recordings?**
- **Hypothesis**: such reconstruction is possible using the capsule recordings and space coordinates.
- **Data**: EigenMike data from EigenScape, 3D-MARCo, and LOCATA datasets
  - skip the encoding step; keep recordings in raw A-format

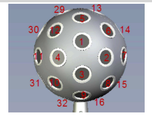
Example use-case (continued)

- **Capturing space info in input data:** intermediate step between A and B format as LeNET input
  - The spherical harmonic matrix multiplies samples recorded by the corresponding capsule.
  - A capsule's samples are missing, and the LeNET architecture is trained to reconstruct it.
- **3 experiments changing the input data:**
  - audio information of a single capsule is missing
  - audio of five neighboring capsules is also missing
  - Input only has samples of 3 capsules resulting in a tetrahedron with respect to the reconstructed capsule
- **Results show:**
  - `micarraylib` is useful and needed.
  - reconstructing a capsule's samples using
  - other capsules' recordings is:
    - possible
    - sensitive to density and physical proximity of recordings available used to reconstruct.

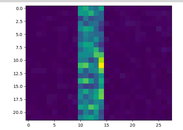
Table of results

Model	MSE (eval)
before training	0.9
exp 1	0.000039
exp 2	0.00013
exp 3	0.0016

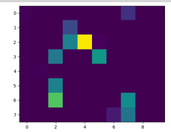
EigenMike capsules



Act in CNN layer 1



Act in CNN layer 2



References

- Gong Y, Chung YA, Glass J. AST: Audio Spectrogram Transformer. arXiv preprint arXiv:2104.01778. 2021 Apr 5.
- Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M. Audio set: An ontology and human-labeled dataset for audio events. In2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 Mar 5 (pp. 776-780). IEEE.
- R. Scheibler, J. Azcarreta, R. Beuchat, C. Ferry, Pyramic: Full Stack Open Microphone Array Architecture and Dataset, IWAENC, Tokyo, 2018.
- Adavanne S, Politis A, Nikunen J, Virtanen T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE Journal of Selected Topics in Signal Processing. 2018 Dec 7;13(1):34-48.
- S. Adavanne, A. Politis, and T. Virtanen, "A multi- room reverberant dataset for sound event localization and uetection," in Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019. [Online]. Available: https://arxiv.org/abs/1905.08546
- A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2020), November 2020. [Online]. Available: https://arxiv.org/abs/2006.01919
- A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Sri- vastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," arXiv preprint arXiv:2106.06999, 2021. [Online]. Available: https://arxiv.org/abs/2106.06999
- H.W.Löllmann,C.Evers,A.Schmidt,H.Mellmann,H.Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in 2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM). IEEE, 2018, pp. 410–414.
- H. Lee and D. Johnson, "An open-access database of 3D microphone array recordings," in Audio Engineering Society Convention 147. Audio Engineering Society, 2019.
- M. C. Green and D. Murphy, "EigenScape: A database of spatial acoustic scene recordings," Applied Sciences, vol. 7, no. 11, p. 1204, 2017.
- Fuentes M, Salamon J, Zinemanas P, Rocamora M, Paja G, Román IR, Bittner R, Miron M, Serra X, Bello JP. Soundata: A Python library for reproducible use of audio datasets. arXiv preprint arXiv:2109.12690. 2021 Sep 26.
- McCormack L, Delikaris-Manias S, Farina A, Pinardi D, Pulkki V. Real-time conversion of sensor array signals into spherical harmonic signals with applications to spatially localized sub-band sound-field analysis. InAudio Engineering Society Convention 144 2018 May 14. Audio Engineering Society.

Acknowledgements

This material is based upon work supported by the National Science Foundation under grant no. IIS-1955357. The authors thank the founding source and their grant collaborators.