JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Project page

Listening Human Behavior: 3D Human Pose Estimation with Acoustic Signal

WED-PM-092



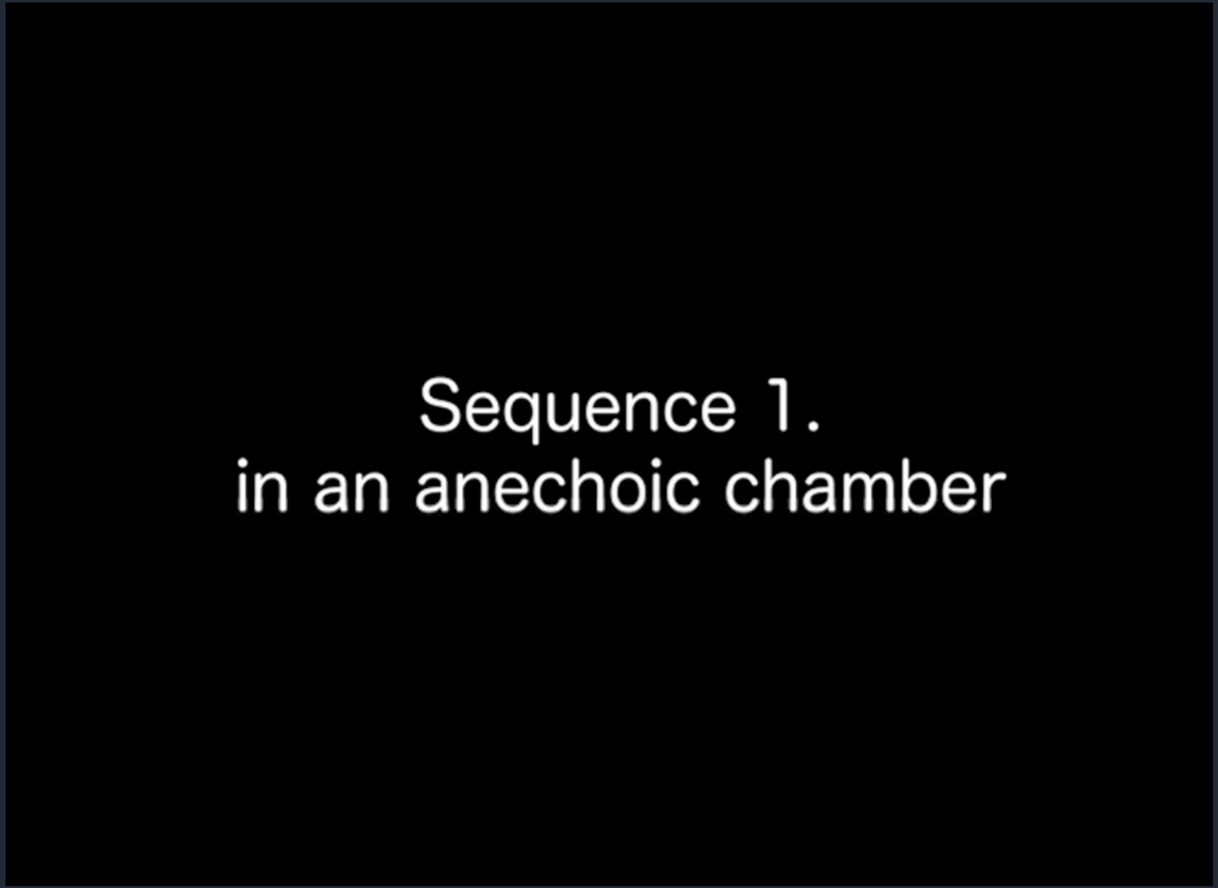


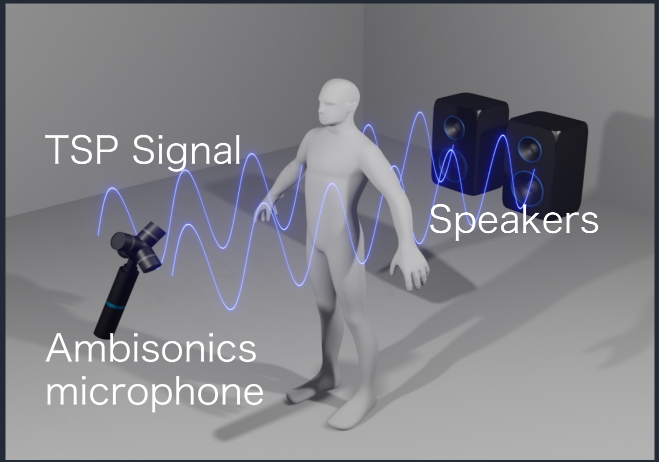

[†] Yuto Shibata, [†] Yutaka Kawashima, [†] Mariko Isogawa, [†‡] Go Irie, [§] Akisato Kimura, [†] Yoshimitsu Aoki

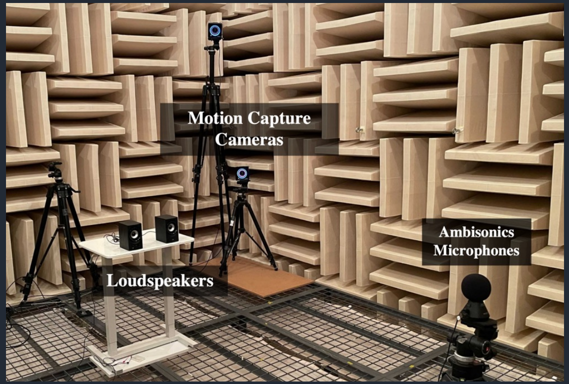
[†]Keio University [‡]Tokyo University of Science [§]NIPPON TELEGRAPH AND TELEPHONE CORPORATION

Summary

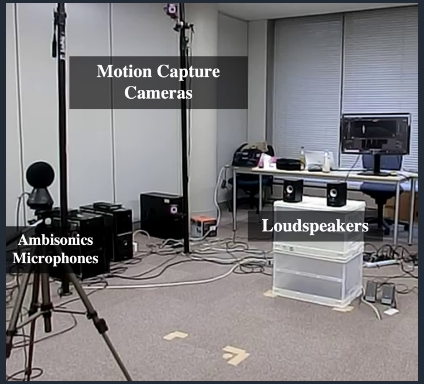
- Tackled 3D pose estimation problem based on active and non-invasive acoustic sensing
- Proposed new framework to map acoustic features into human pose







Anechoic chamber



Classroom

Limitation of existing work

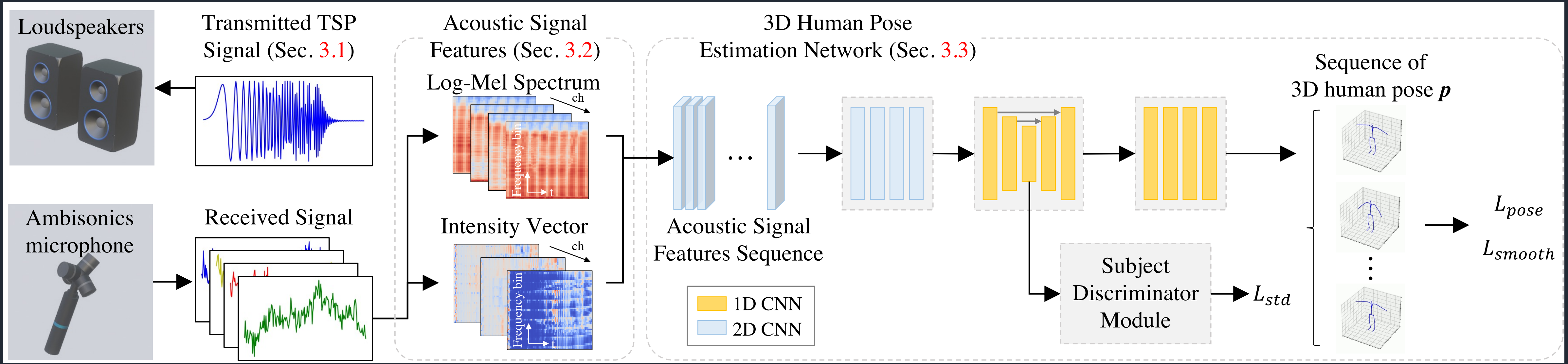
1. RGB image based models
   - Fail in dark environments or scenes with occlusions
   - Privacy issues happen
2. Wifi/RF based models
   - Prohibited use in places with precision instruments exist (e.g., airplanes, hospitals, etc.)
3. Acoustic signal based models
   - Requires invasive sensing with body-mounted devices
   - Requires sound semantics such as speech or instruments

OpenPose [Cao *et al.* TPAMI2019]

RF-Based pose estimation
[Zhao *et al.* CVPR2018]

Technical Contributions

1. Active acoustic sensing with speakers and ambisonics microphone
2. Feature Extraction: Log-mel spectrogram and Intensity Vector
3. 2D CNN and 1D Time-wise Unet
4. Subject Discriminator Module (Adversarial Leaning)

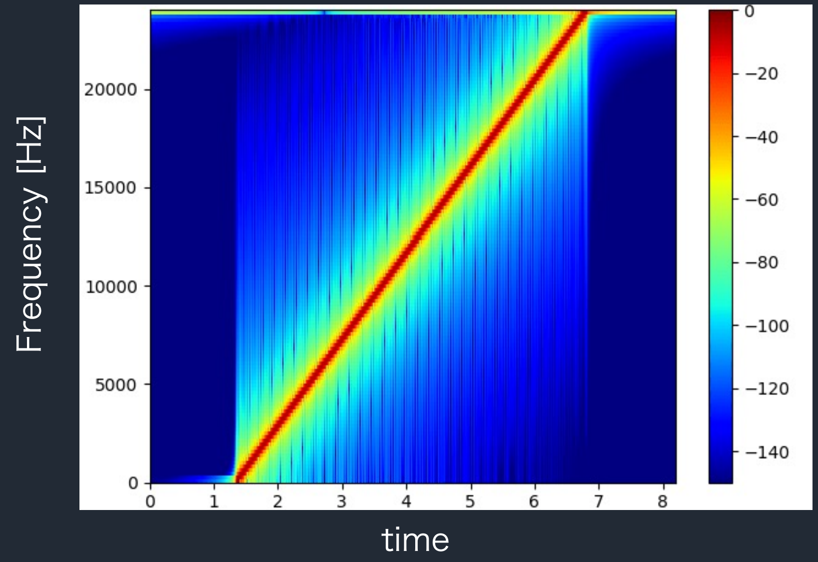


Proposed pipeline

Technical Contributions

1. Active acoustic sensing with speakers and ambisonics microphone
2. Feature Extraction: Log-mel spectrogram and Intensity Vector
3. 2D CNN and 1D Time-wise Unet
4. Subject Discriminator Module (Adversarial Leaning)

- Repeat TSP signal to capture Room Impulse Response at every moment
- Ambisonics microphone is used to capture 360 degree sounds



TSP Signal[1]

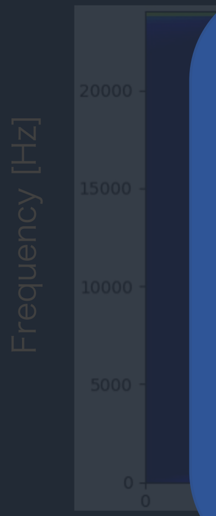
Ambisonics microphone[2]

[1] https://moromisenpy.com/get_impulse_response/
[2] https://zoomcorp.com/ja/jp/handheld-recorders/handheld-recorders/h3-vr-360-audio-recorder/

Technical Contributions

1. Active acoustic sensing with speakers and ambisonics microphone
2. Feature Extraction: Log-mel spectrogram and Intensity Vector
3. 2D CNN and 1D Time-wise Unet
4. Subject Discriminator Module (Adversarial Leaning)

- Repeat TSP signal to capture Room Impulse Response at every moment
- Ambisonics microphone is used to capture 360 degree sounds



Two Features

1. Active sensing

- Our model doesn’t use any sound semantics such as speech or instruments
  → Avoid privacy issues

2. Non-invasive sensing

- Subjects don’t need to put on put on any special devices

TSP Signal[1]

Ambisonics microphone[2]

[1] https://moromisenpy.com/get_impulse_response/
[2] https://zoomcorp.com/ja/jp/handheld-recorders/handheld-recorders/h3-vr-360-audio-recorder/

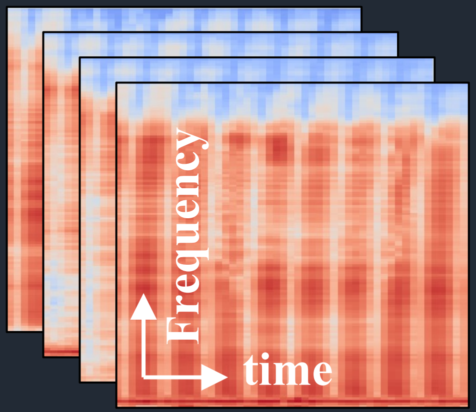
Technical Contributions

1. Active acoustic sensing with speakers and ambisonics microphone
2. **Feature Extraction: Log-mel spectrogram and Intensity Vector**
3. 2D CNN and 1D Time-wise Unet
4. Subject Discriminator Module (Adversarial Leaning)

These two are used to estimate DoA (Direction of Arrival) and distance to subjects

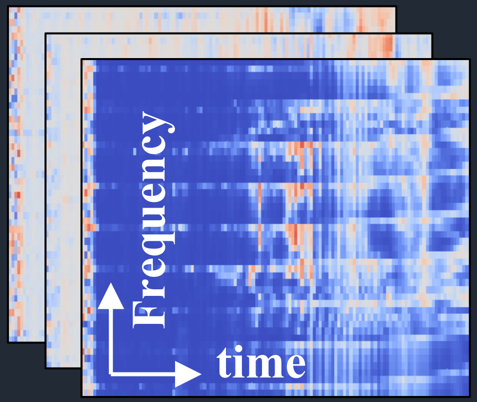
(i) Mel spectrogram

$$I^{mel}(k,t) = H_{mel}(k,f) \cdot F(s(f,t))$$



(ii) Intensity Vector

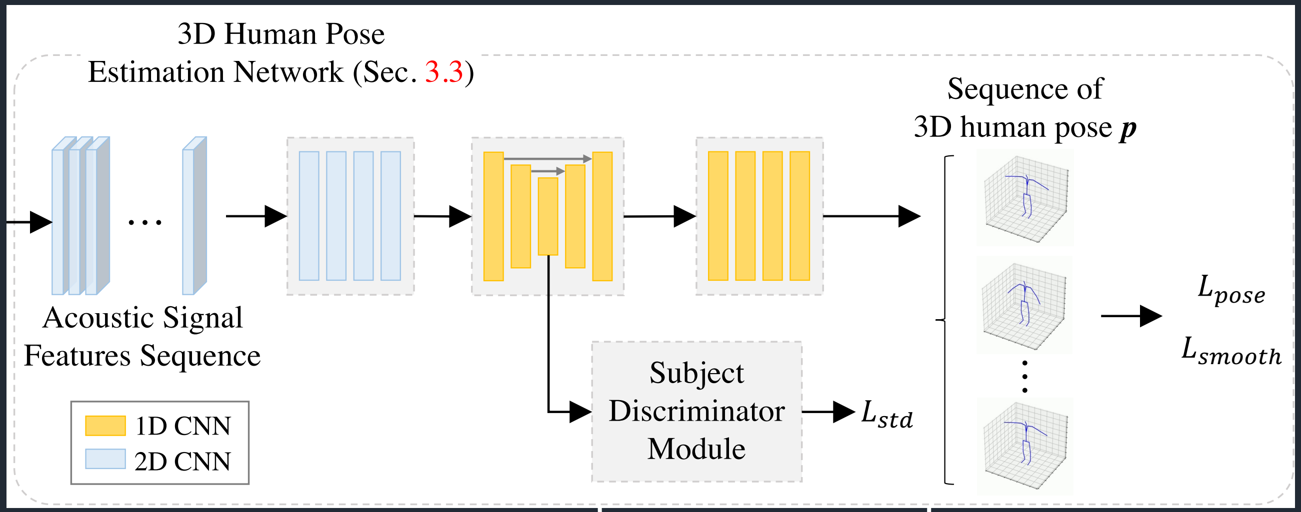
$$I(f,t) = R\left\{ W^{*}(f,t) \cdot \begin{pmatrix} X(f,t) \\ Y(f,t) \\ Z(f,t) \end{pmatrix} \right\}$$



Technical Contributions

1. Active acoustic sensing with speakers and ambisonics microphone
2. Feature Extraction: Log-mel spectrogram and Intensity Vector
3. 2D CNN and 1D Time-wise Unet
4. Subject Discriminator Module (Adversarial Leaning)

- Two CNN to capture temporal consistency-aware features
- Subject Discriminator module to create subject-invariant features



$$L_{std} = \frac{1}{T'} \sum_{n=1}^{T'} \mathrm{STD}(S_n)$$

Experimental Results

- Conducted experiments under two settings for 8 subjects

  (i) Single Subject (ii) Cross subject

- Our proposed model outperformed prior baseline models in almost all settings

Quantitative Results

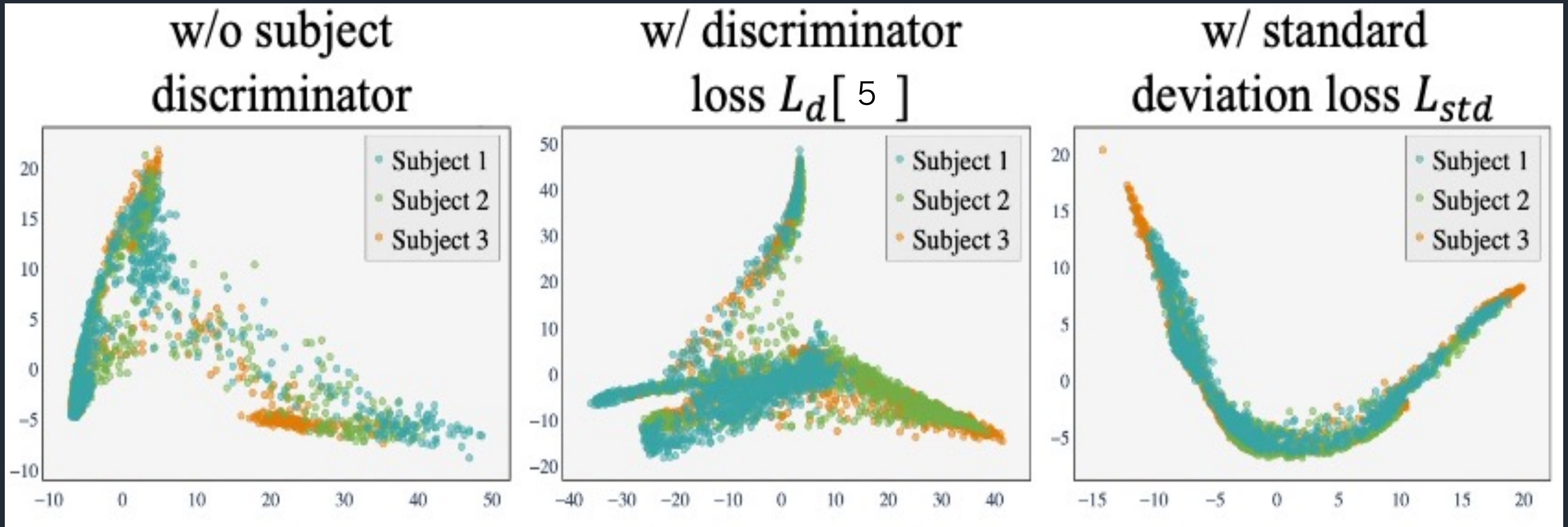
Method	Anechoic Chamber Environment						Classroom Environment					
	Single Subject			Cross Subject			Single Subject			Cross Subject		
	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)
Ginosar *et al.* [3]	0.44	0.23	0.90	0.83	0.51	0.60	0.58	0.30	0.84	0.95	0.56	**0.68**
Jiang *et al.* [4]	0.90	0.44	0.73	0.96	0.55	0.62	0.58	0.34	0.73	1.02	0.63	0.49
Ours (Method's best)	**0.42**	**0.22**	**0.90**	**0.73**	**0.45**	**0.72**	**0.54**	**0.28**	**0.85**	**0.93**	**0.55**	0.67

[3] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. CVPR, pages 3497–3506, 2019. 2, 3, 5, 6
[4] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. Towards 3d human pose construction using wifi. MobiCom, pages 1–14, 2020. 5, 6

Adversarial Leaning Effects

- Has better result than prior cross entropy loss method
- Reduces feature shift among 3 subjects

Method	RMSE (↓)	MAE (↓)	PCKh@0.5 (↑)
Ours (L_d [5])	0.78	0.47	0.68
Ours (L_{std})	**0.73**	**0.45**	**0.72**



[5] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao,Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. Towards environment independent device free human activity recognition. MobiCom, page 289–304, 2018.

Listening Human Behavior: 3D Human Pose Estimation with Acoustic Signal

- Showed for the first time it is possible to obtain human pose with active and non-invasive acoustic sensing

- Proposed new framework to map acoustic features into human pose

- Outperformed previous work with mel spectrogram, intensity vector, and subject discriminator module

- Created new datasets in an anechoic chamber and classroom environment

Code and datasets are available:
https://isogawa.ics.keio.ac.jp/research_project/acoustic_3dpose.html

Keio University CALAMVS GLADIO FORTIOR 1858

NTT

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA