Acoustic-based 3D Human Pose Estimation Robust to Human Position

Paper ID: 135

Yusuke Oumi[1], Yuto Shibata[1], Go Irie[1,2], Akisato Kimura[3], Yoshimitsu Aoki[1], Mariko Isogawa[1,4]
[1]Keio University [2]Tokyo University of Science [3]Nippon Telegraph and Telephone Corporation [4]JST Presto
Contact: youmi@keio.jp (Oumi), mariko.isogawa@keio.jp (Isogawa)

Research Overview

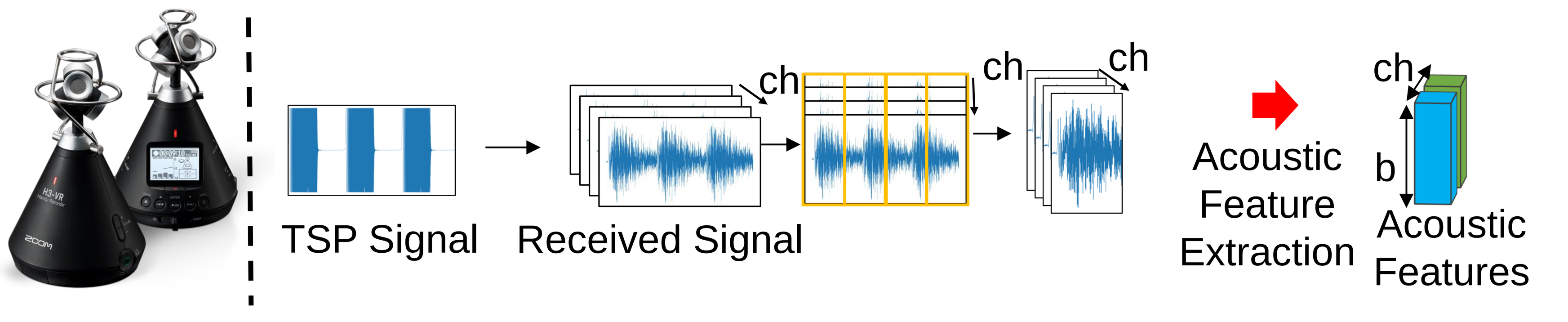
Challenges: Existing methods for estimating human poses face the following difficulties.

- RGB-based method: effects of obstruction; poor performance in low-light conditions; and privacy concerns.
- Wireless signal-based method: limited use in environments with sensitive equipment (e.g. medical facilities and aircraft).
- Existing acoustic signal-based method: estimation affected by human position.

Contribution: Proposing a non-invasive 3D human pose estimation method that uses active acoustic sensing and that is **robust to human position**.

Overview of Acoustic-Based Pose Estimation

1. Time-Stretched Pulse signals are transmitted from speakers.
2. Sound is received by a 4-channel ambisonics microphone.
3. Acoustic features are extracted and input into the model.
4. The pose estimation network estimates multiple frame poses.

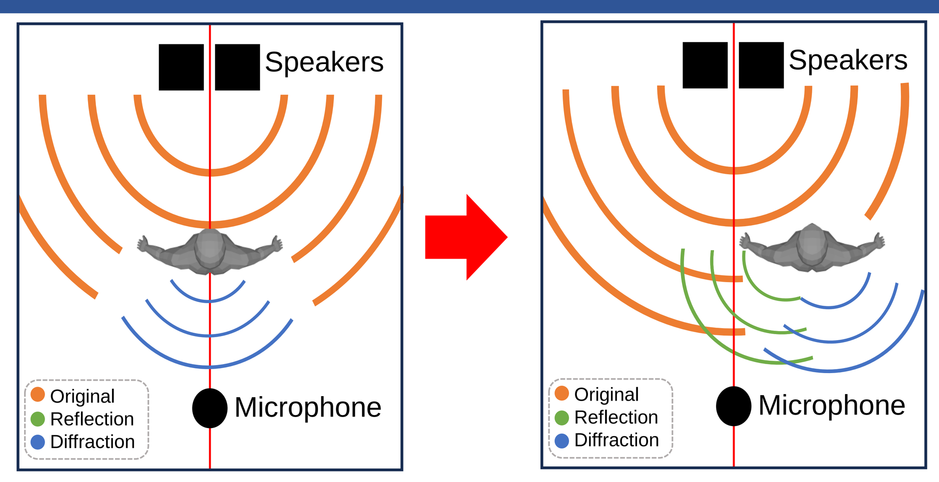


Ambisonics Microphone and Acoustic-signal Preprocessing

Difference from Existing Method

Existing method[1]:
Pose estimation of a subject on the line between speakers and a microphone.

Our method:
Pose estimation of a subject away from the line.



Difference between existing method (left) and our method (right)

Proposed Method

Pose Estimation Module

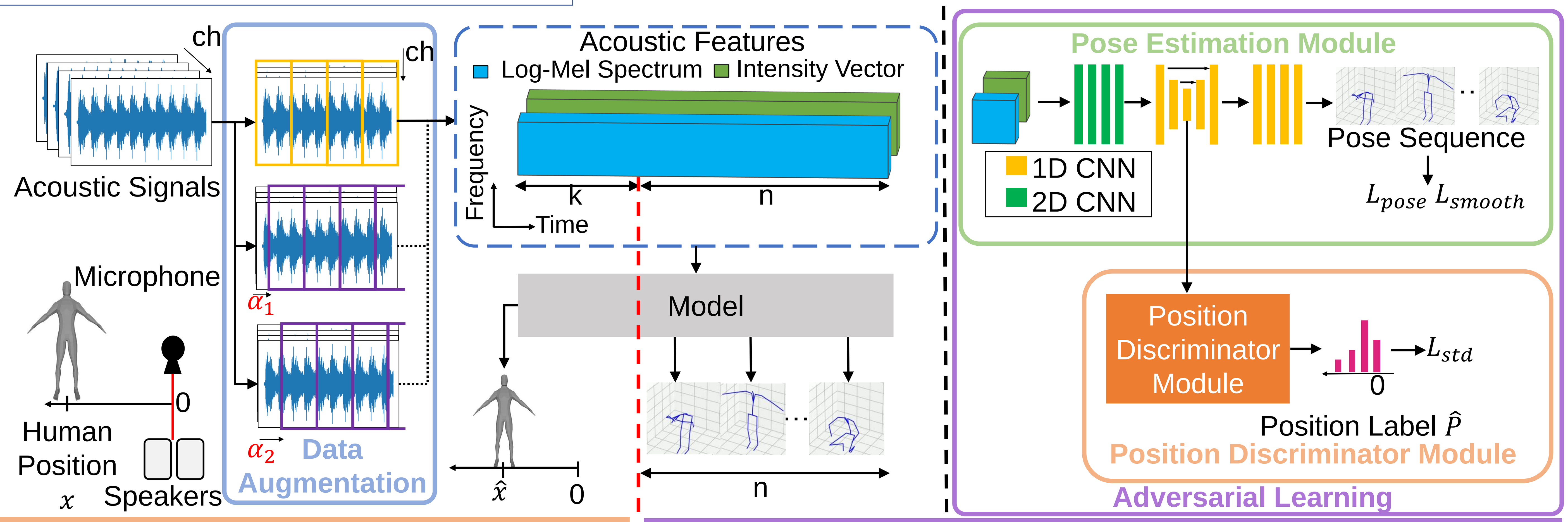
Overview: CNN-based network that estimates n frames of poses simultaneously using $n+k$ frames of acoustic features.

Key Point: Utilize time-series relationships of sound by including acoustic information up to k frames ago.

Data Augmentation

Overview: Shift the phase of the received acoustic signals by α, generating acoustic features from the phase-shifted signals.

Key Point: The phase of TSP signals included in the acoustic features changes before and after shift.

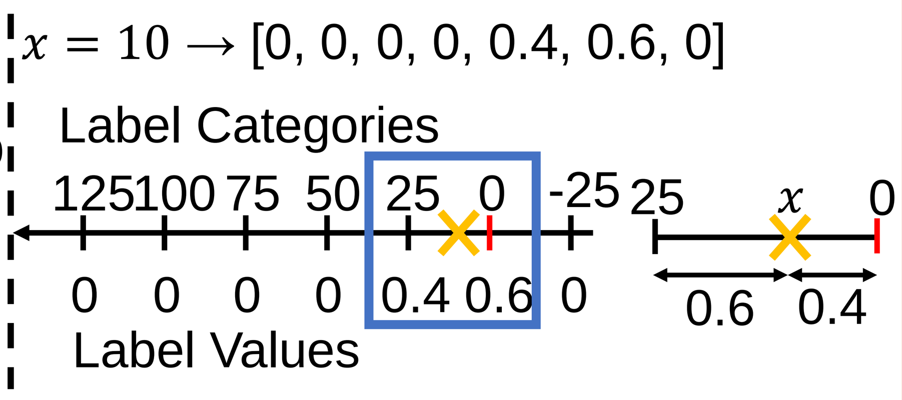


Position Discriminator Module

Overview: Utilize the intermediate layer values from the Pose Estimation Module to estimate the subject's position.

Key Points:

1. Implement the Module as a classifier to prevent gradient explosion.
2. Use soft labels to represent continuous positions.



Adversarial Learning

Overview: Avoid dependence of pose estimation accuracy on subject position.

Key Points:

1. Position Discriminator Module estimates the subject's position.
2. Pose Estimation Module generates features that (a) improve the accuracy of pose estimation but (b) fail to estimate the human position.

Experiment

Dataset:

- Subject: Training with four subjects and testing with another subject.
- Position: Directly on the line, and at 25, 50, 75, and 100 cm away from the line.

Evaluation Metrics: RMSE (↓), MAE (↓), PCK (↑)

Baselines:

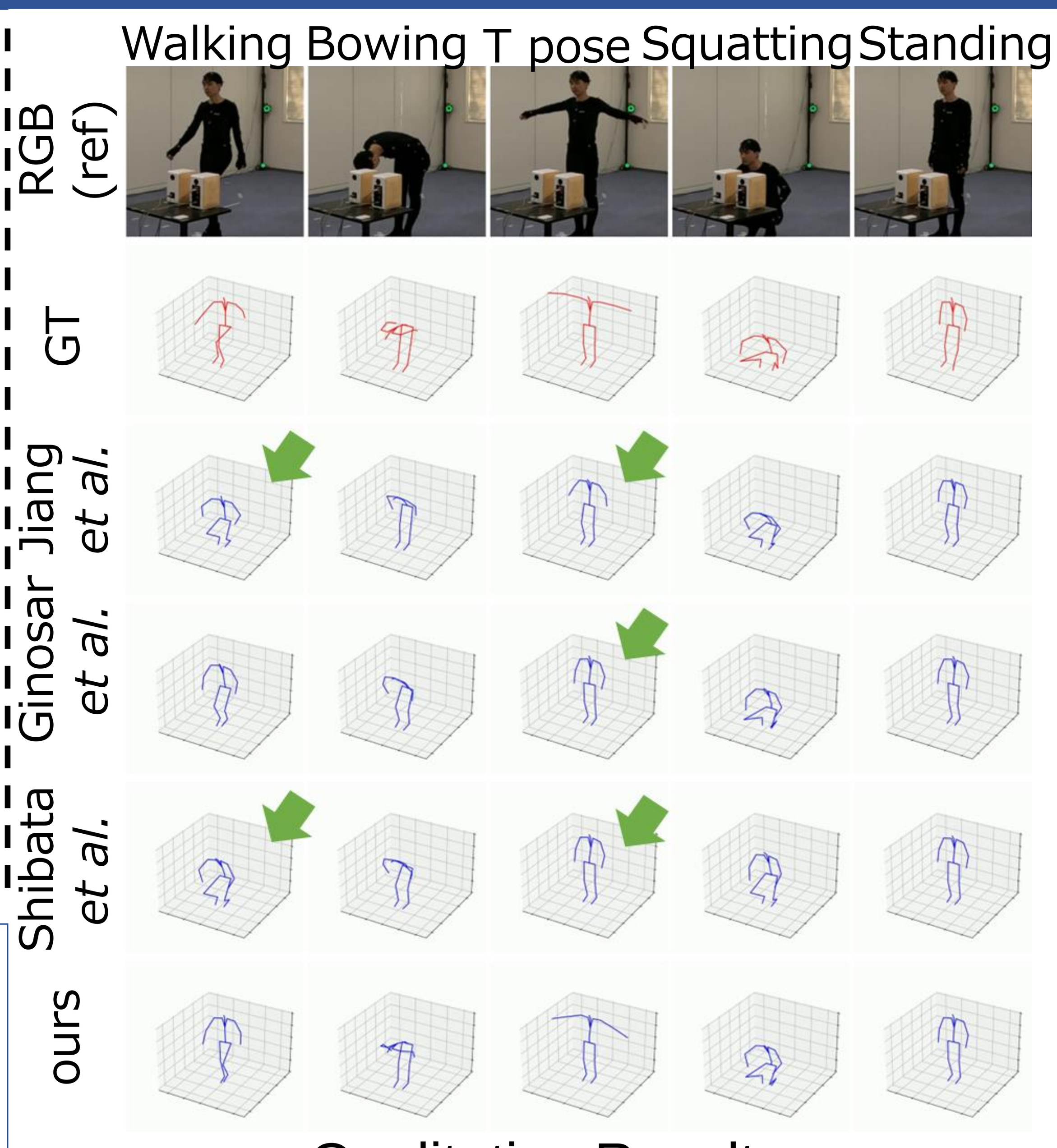
- Jiang *et al.* : Pose estimation using Wi-Fi signals
- Ginosar *et al.* : Gesture estimation using speech sounds
- Shibata *et al.* : Pose estimation using sound for a subject on the line between the speakers and the microphone.

Table1: Comparison against baselines

method	RMSE (↓)	MAE (↓)	PCK (↑)
Jiang *et al.* [2]	0.75	0.40	0.48
Ginosar *et al.*[3]	0.65	0.33	0.55
Shibata *et al.* [1]	0.66	0.35	0.53
Ours	**0.53**	**0.28**	**0.60**

Table 2: Ablation Study

	RMSE (↓)	MAE (↓)	PCK (↑)
Ours w/o Adv	0.55	0.29	0.56
Ours w/o Prior	0.69	0.35	0.55
Ours w/o Aug	0.58	0.31	0.55
Ours	**0.53**	**0.28**	**0.60**



Qualitative Results

References

[1] Shibata *et al.*, "Listening human behavior: 3d human pose estimation with acoustic signals", In CVPR, p.13323–13332, 2023
[2] Jiang *et al.*, "Towards 3d human pose construction using wifi", In MobiCom, p.1-14, 2020
[3] Ginosar *et al.*, "Learning individual styles of conversational gesture", In CVPR, 3497-3506, 2019