Beyond-Voice: Towards Continuous 3D Hand Pose Tracking on Commercial Home Assistant Devices

Yin Li, Rohan Reddy, Cheng Zhang, Rajalakshmi Nandakumar

Cornell University | CORNELL TECH

01

Background

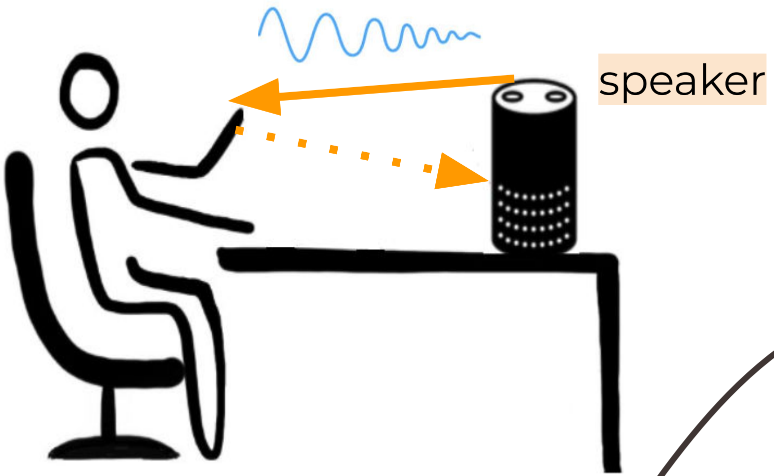
Intro to acoustic sensing

Background
- *Intro to acoustic sensing*

Using **ultrasound for human tracking** has been an active research field in *wireless sensing*.

Acoustic sensing applications

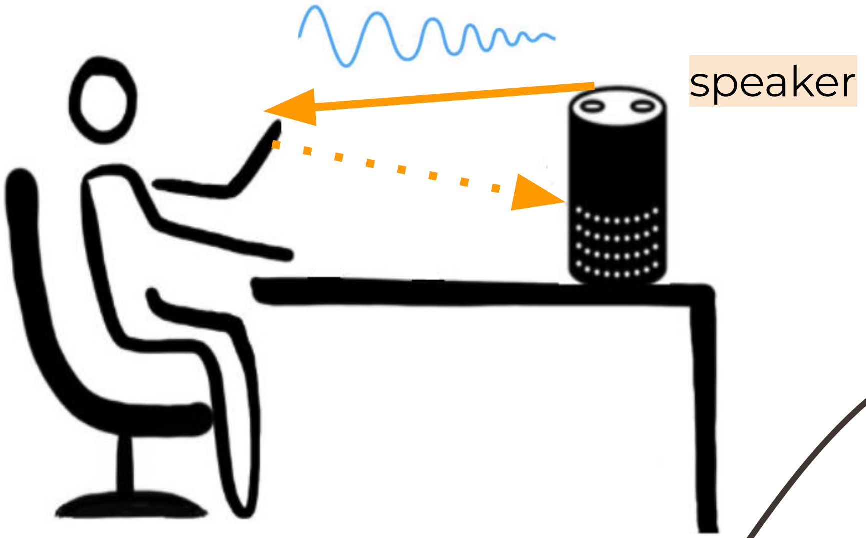
- gesture recognition
- breath rate/heart rate detection
- intrusion detection
- VR/AR interaction



Background
- *Intro to acoustic sensing*

✓ Acoustic sensing:

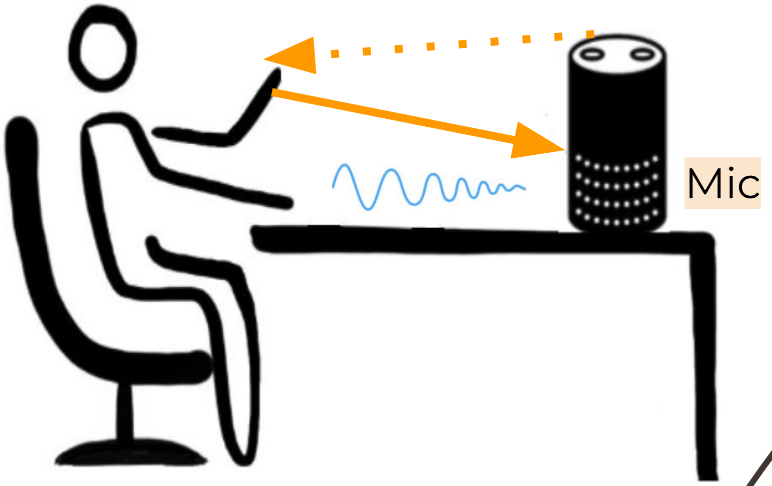
Transform the smart device into a **SONAR** system by leveraging the **speaker** and **microphone on device** for motion tracking.



Background
- *Intro to acoustic sensing*

✓ Acoustic sensing:

Transform the smart device into a **SONAR** system by leveraging the speaker and microphone on device for **motion tracking**.



Background
- *Intro to acoustic sensing*

Acoustic sensing

- ubiquitous

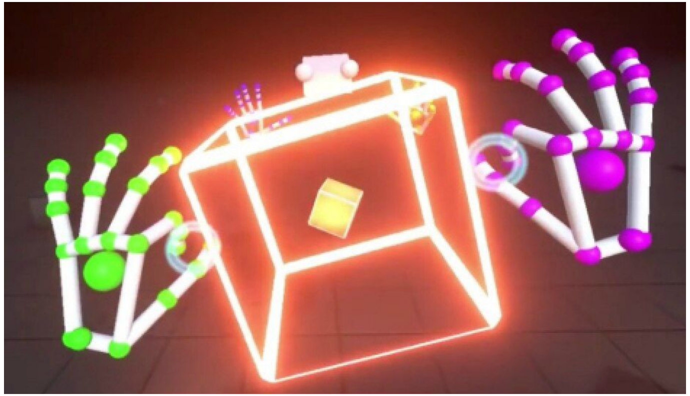
- under-cm resolution
- Speed << RF

- less privacy issue
- non-line-of-sight

Background
- *Problem of acoustic sensing*

Problem: coarse-granularity and usability

- Current acoustic sensing applications are **coarse-grained**
  - Presence detection, gesture classification, single point tracking, etc.



No Continuous & unlimited tracking

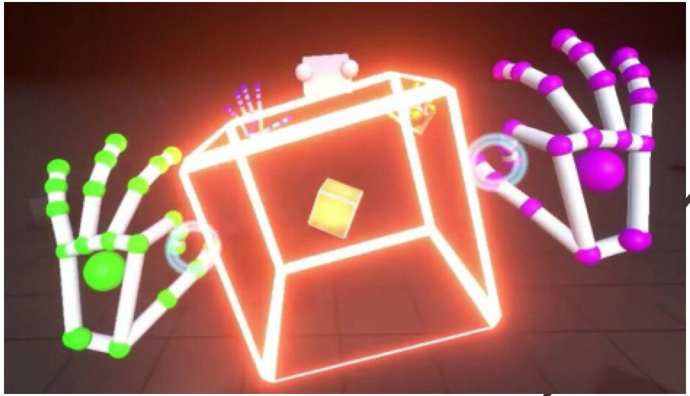
Background
- *Problem of acoustic sensing*

Problem: coarse-granularity and usability

- Current acoustic sensing applications are **coarse-grained**
  - Presence detection, gesture classification, single point tracking, etc.

More applications with the fine-grained:

- HCI: continuous hand tracking input
- Sports analysis
- Rehabilitation
- Semantically better gesture recognition (sign language recognition)



Fine-grained: Continuous & unlimited tracking

Background
- *Problem of acoustic sensing*

Motivation: towards the fine-grained acoustic sensing

- The **weakness** of the **alternatives**
  - **On-body sensors**: intrusive
  - **Cameras**: privacy issue; much worse with occlusion
  - **USRP/WiFi**: large expensive MIMO (multi antenna)

Background
- *What is fine-grained?*

Coarse-grained
Gesture **classification**
[RobuCIR'20, Amaging'22]

Coarse-grained
sub-mm localization of a **single point**: breath detection, writing in the air
[FingerIO'16, LLAP'16, Mao W'19]

Fine-grained
<= 2 **multi-target** hand tracking
[FM-track''20]

Fine-grained
21 hand joints
16 body joints...
[vision/large Wifi/USRP]

→ **usability**

Background
- *Why fine-grained?*

Coarse-grained: limited use cases

Fine-grained: comparable to depth camera
--> various downstream Apps

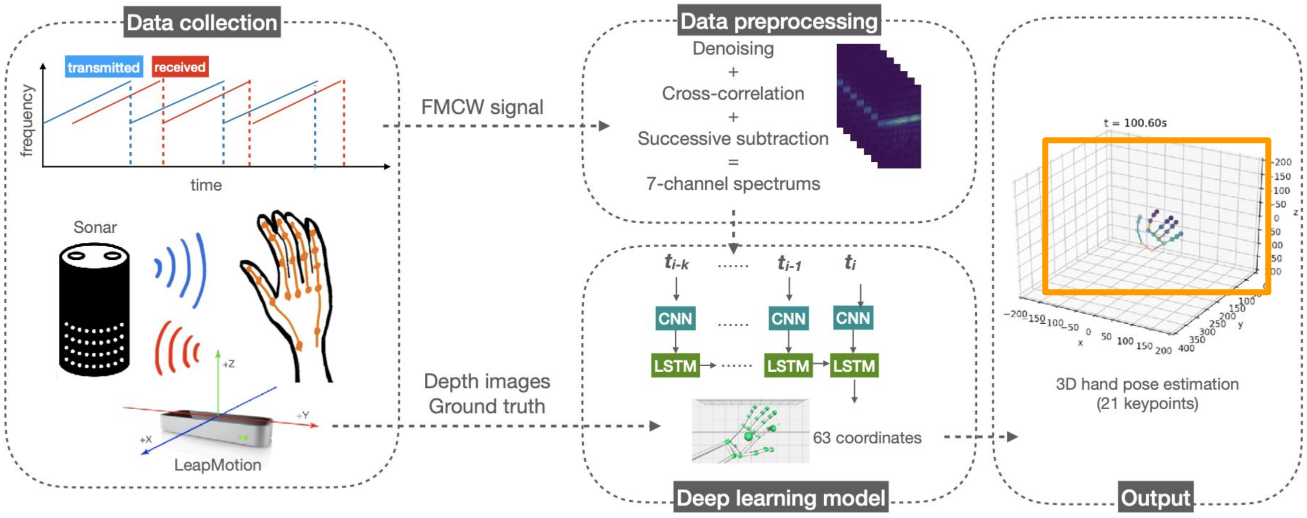
02

Method

Towards fine-grained acoustic sensing

Method
- *Overview*

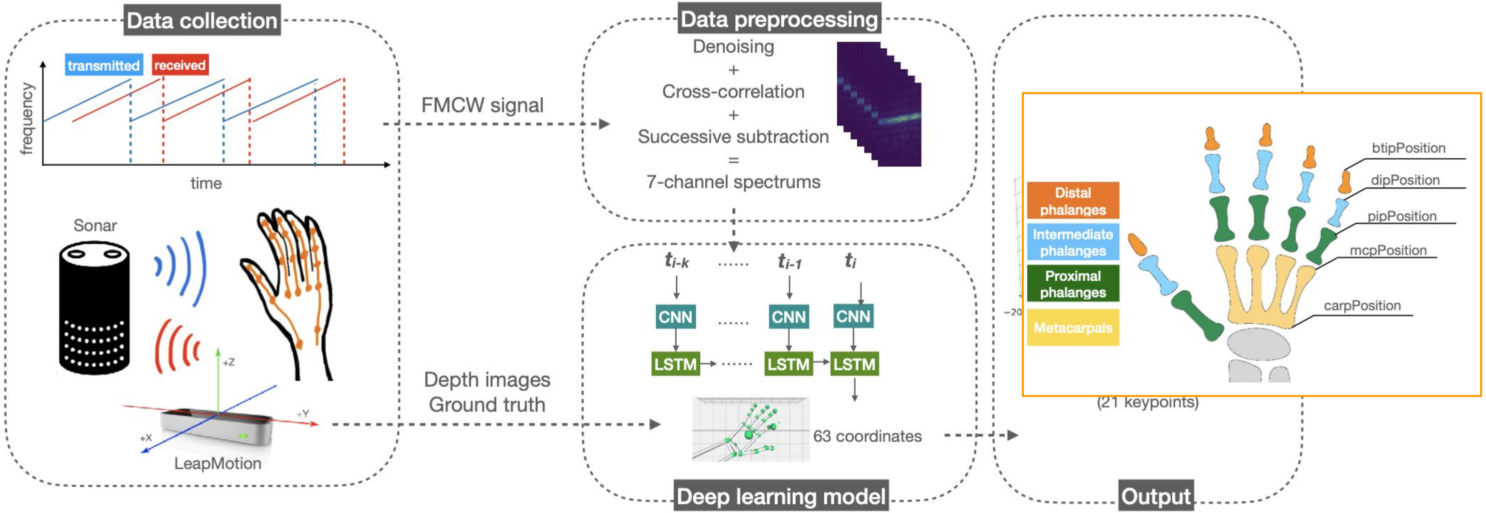
We proposed the first fine-grained acoustic sensing pipeline for hand tracking on a commercial smart speak[1].



[1] Beyond-Voice: Towards Continuous 3D Hand Tracking on Commercial Home Assistant Devices. Y. Li, C. Zhang, R.Nandakumar

Method
- *Overview*

We proposed the first fine-grained acoustic sensing pipeline for hand tracking on a commercial smart speak[1].



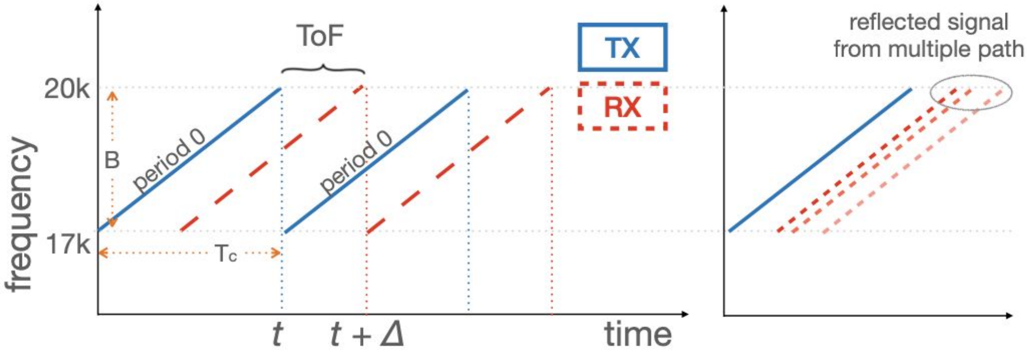
[1] Beyond-Voice: Towards Continuous 3D Hand Tracking on Commercial Home Assistant Devices. Y. Li, C. Zhang, R.Nandakumar

Method

- Frequency-modulated continuous wave

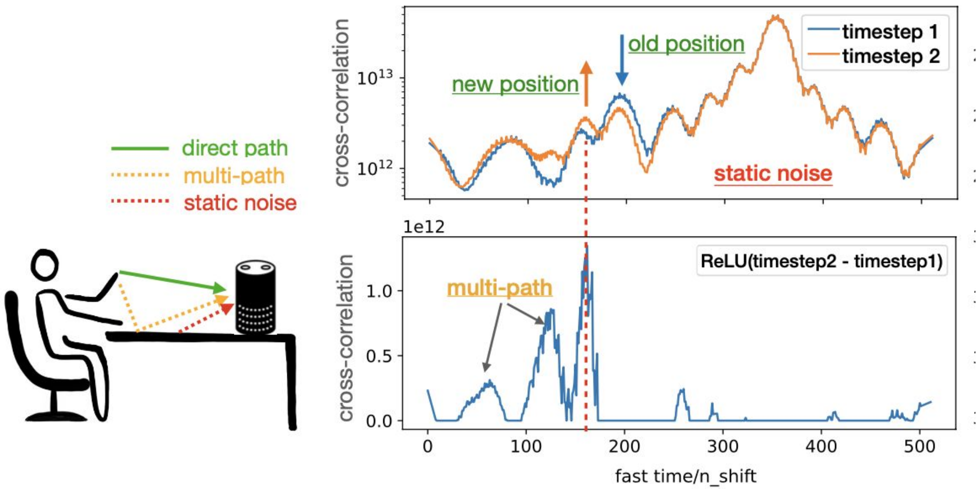
Preliminary exploration: select the right type of modulation

- ~~MUSICDoA~~
- ~~Sine wave~~
- FMCW: each **received** signal is a time-delay version of the **transmitted** signal shifted by a different amount of time proportional to the distance.



Method
- *Successive subtraction for denoising*

Fine-grained range profile

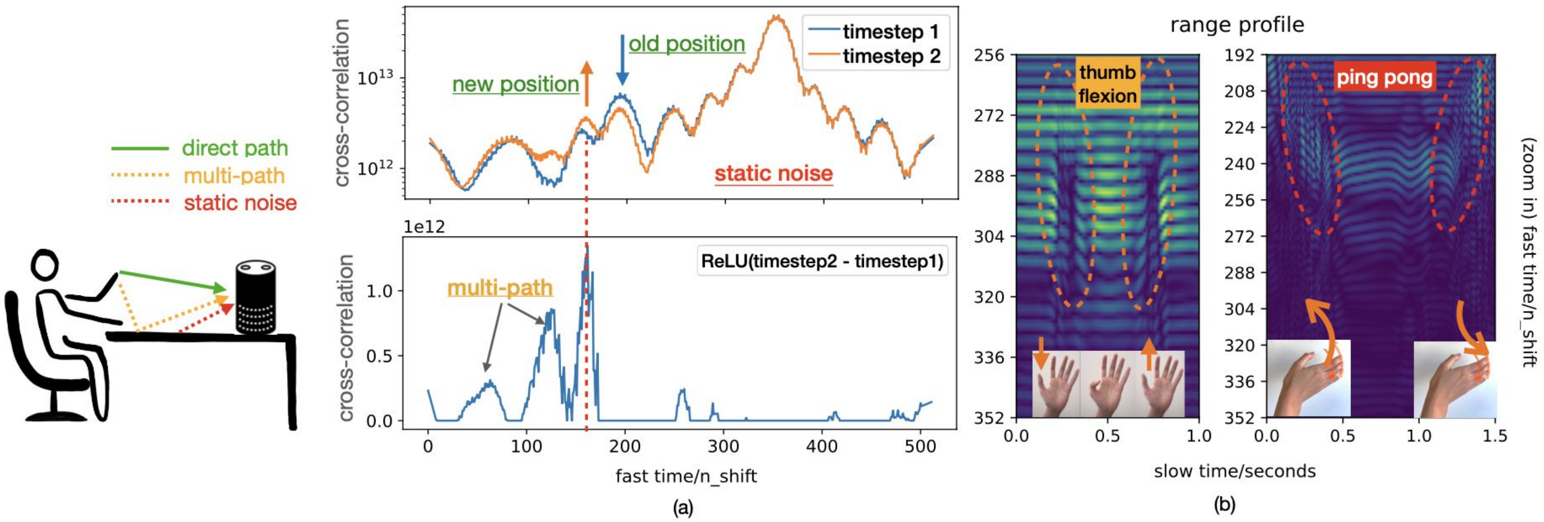


(a)

Method

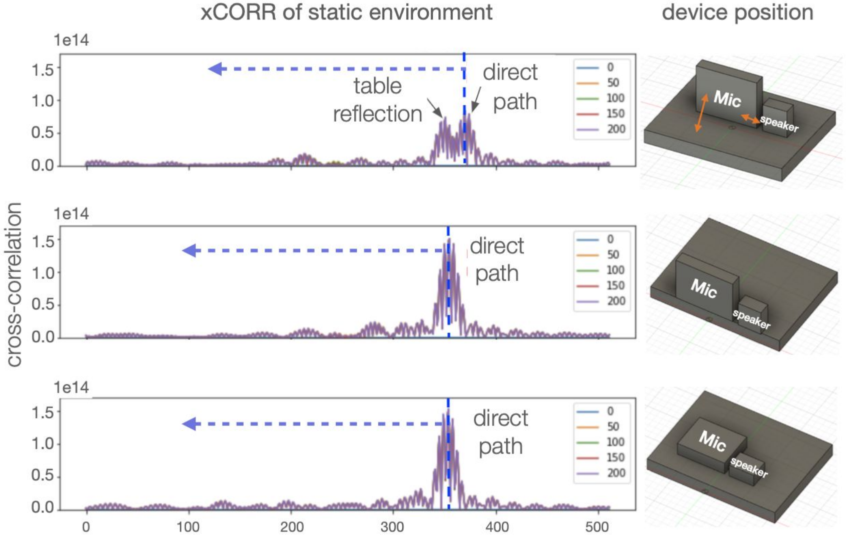
- Successive subtraction for denoising

Fine-grained range profile



Method
- *Starting time error cancelation*

Starting time cancelation



Method

- Dechirping by cross-correlation provides better range resolution

raw signal cross-correlation

—> resolution under *cm*

$$\frac{1}{fs} \times c \times \frac{1}{2} = \frac{343}{48000 \times 2} = 0.00357m = 3.57mm$$

frequency domain cross-correlation

—> efficient computation

Method

- Dechirping by cross-correlation provides better range resolution

raw signal cross-correlation

—> resolution under *cm*

$$\frac{1}{fs} \times c \times \frac{1}{2} = \frac{343}{48000 \times 2} = 0.00357m = 3.57mm$$

frequency domain cross-correlation

—> efficient computation

$$y_k = \sum_{n=0}^{N-1} x_n e^{\frac{-j2\pi nk}{N}} = A_k e^{j\phi_k}$$

$$y'_k = e^{\frac{-j2\pi nk\Delta}{N}} y[k] \ when \ x'_n = x_{n-\Delta}$$

Method

- *Study design: unbounded to pre-defined gestures*

Data collection guiding video

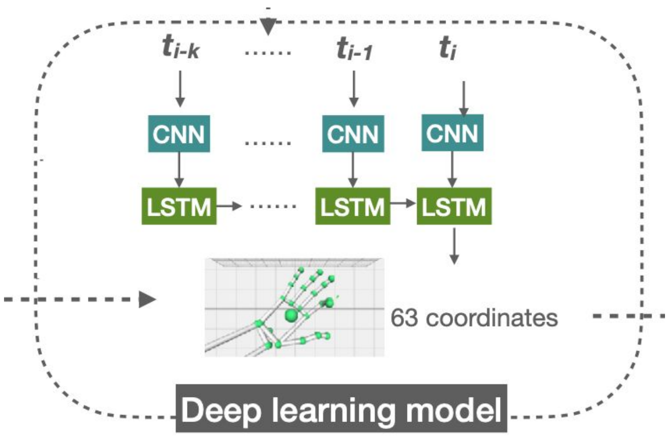
(not gesture classification but **continuous tracking**)

Method
- Training strategies: curriculum learning

Training strategies:

- **Curriculum learning**
- Data augmentation

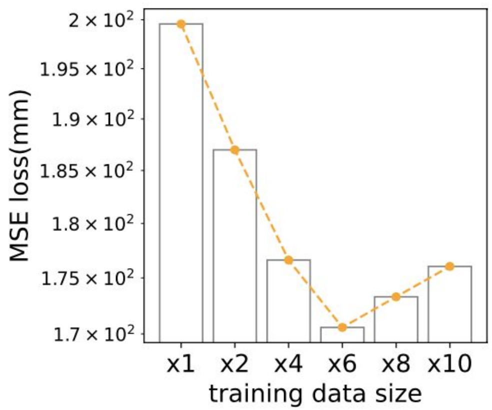
- CL trains the model **hierarchically**
  - from simple gesture sets to complex finger motions;
  - avoid overfitting



Method
- Training strategies: data augmentation

Training strategies:

- Curriculum learning
- **Data augmentation**
-



most sensitive to the change of **y-axis**

↓

shift the starting time cancellation cut-off

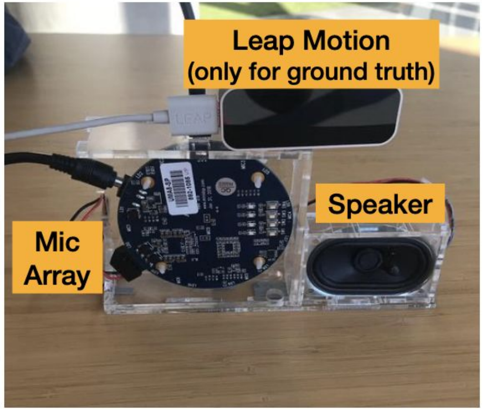
↓

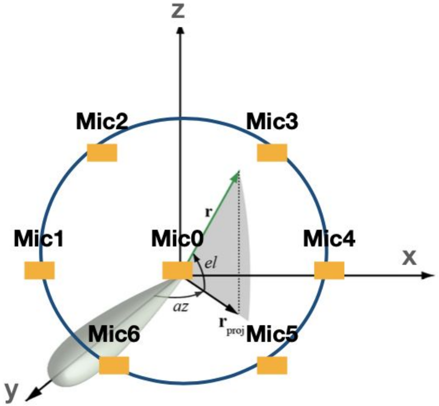
each shift results in +/-3.5mm of ground truth y of all 21 joints

03
Experiments

Experiment
- *setup*

Hardware setup





Same layout and sensitivity as Amazon Echo dot 2nd

Experiment
- user study

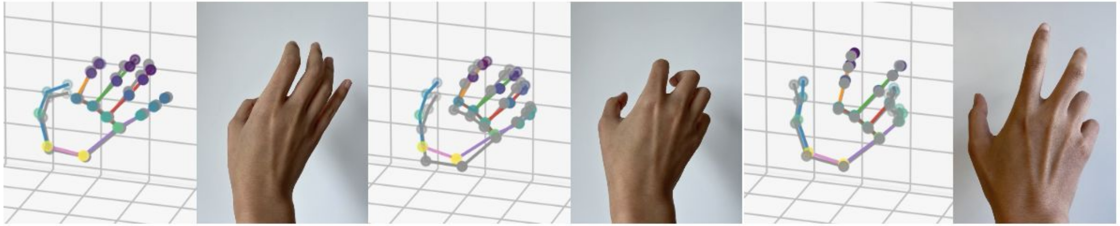
- 11 participants
  - 6 sessions per user, 2min per session, 3 locations
- 2 users for extensive pretraining
- additional validation: extra data collection (10+)

Experiment results
- *cross-user*

Evaluation:

	mean	median	90th percentile
user-independent	16.47	14.57	25.23
user-adaptive	10.36	9.72	18.48
user-dependent	12.49	10.33	21.41

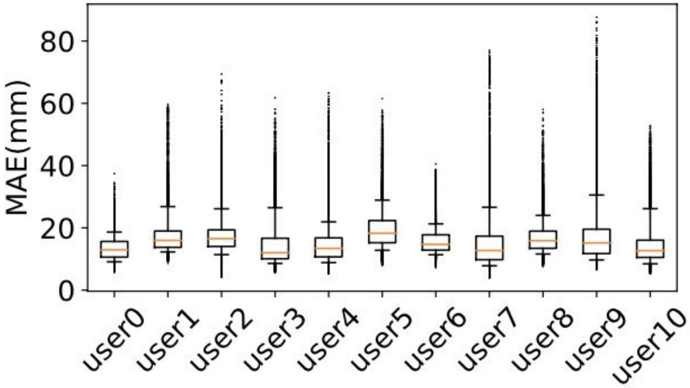
Mean absolute error

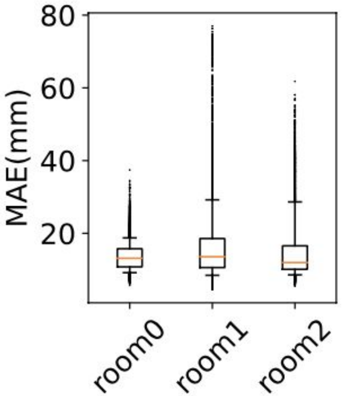


Visualization of sample results. The grey skeleton is ground truth; the cyan is our prediction

Experiment results
- cross-user/environment

Evaluation: **cross user/environment**

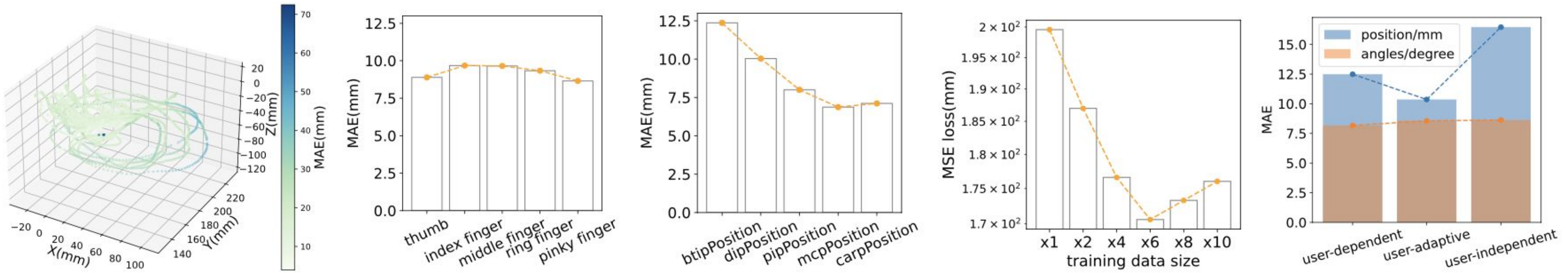




- Study room
- Bedroom
- Open space

User study: the system performance is independent of user and environment.

Experiment results
- error analysis

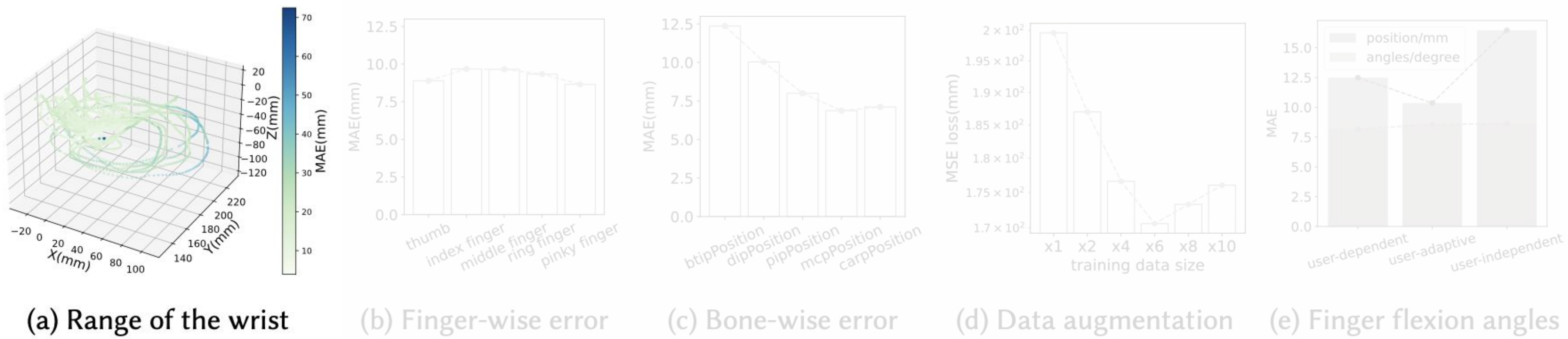


(a) Range of the wrist (b) Finger-wise error (c) Bone-wise error (d) Data augmentation (e) Finger flexion angles

Error analysis from different perspectives

Experiment results
- error analysis

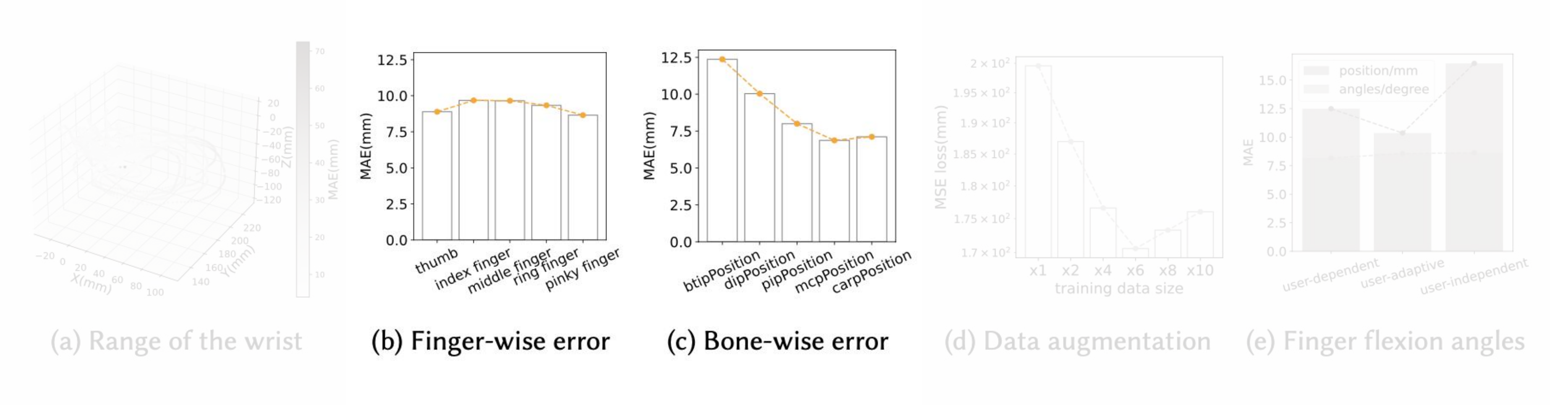
Evaluation: **error analysis**



Error analysis from different perspectives

Experiment results
- *error analysis*

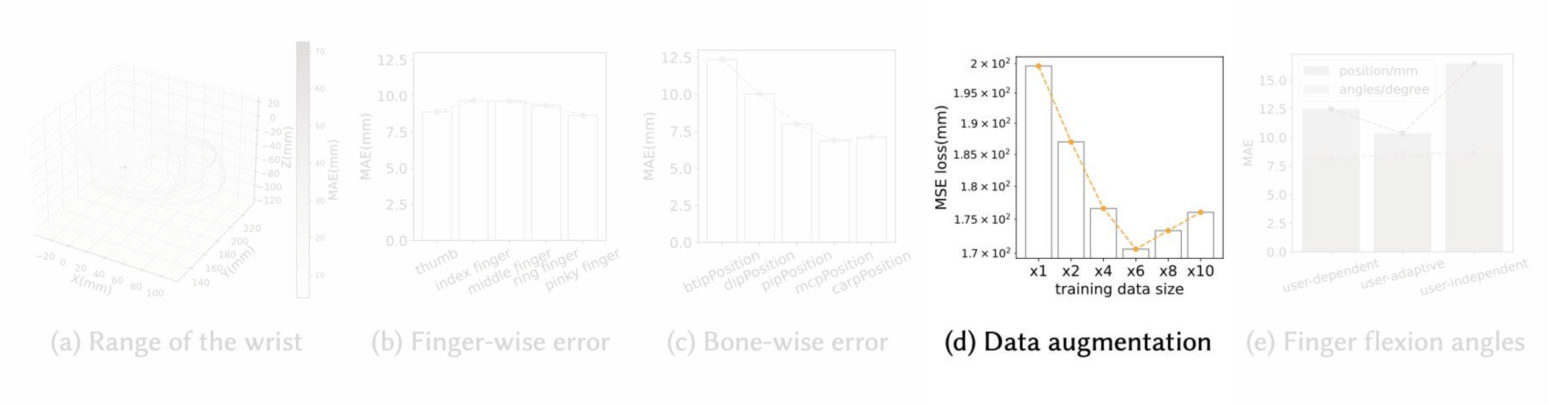
Evaluation: **error analysis**



Error analysis from different perspectives

Experiment results
- error analysis

Evaluation: **error analysis**

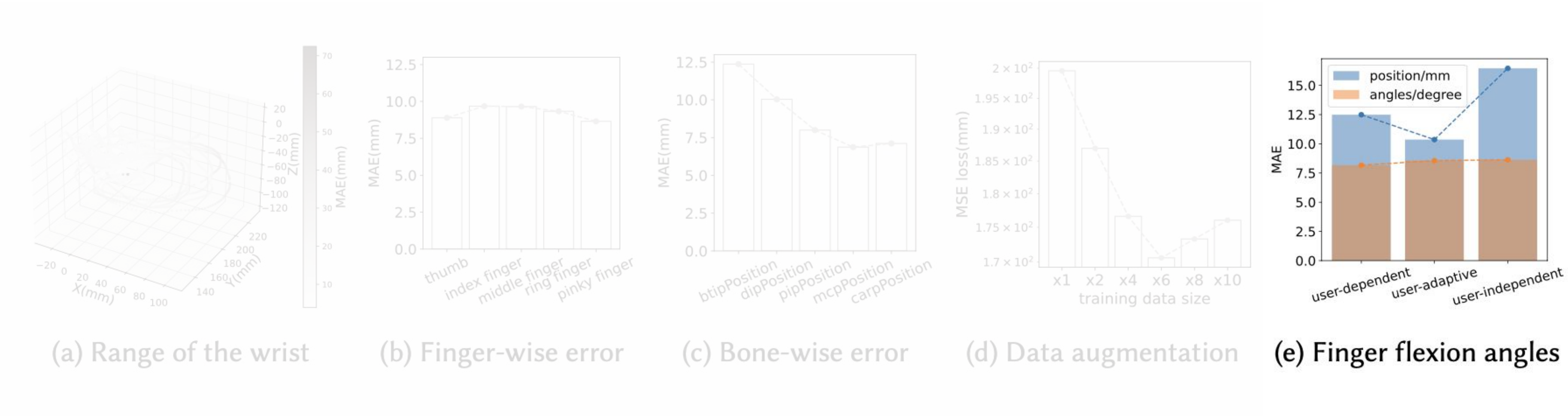


(a) Range of the wrist (b) Finger-wise error (c) Bone-wise error (d) Data augmentation (e) Finger flexion angles

Error analysis from different perspectives

Experiment results
- error analysis

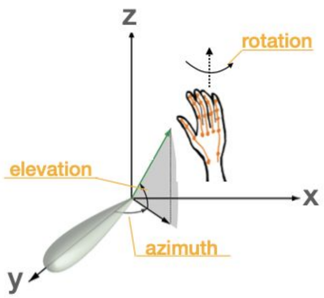
Evaluation: **error analysis**



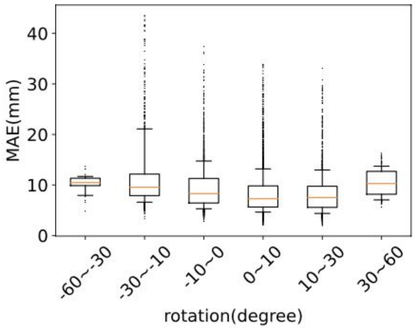
Error analysis from different perspectives

Experiment results
- error analysis

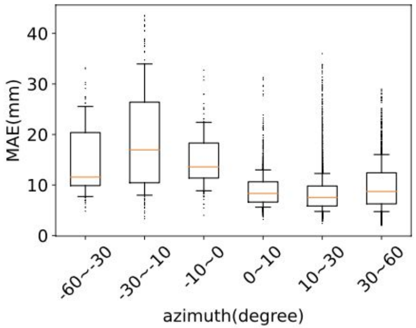
Evaluation: **error analysis**



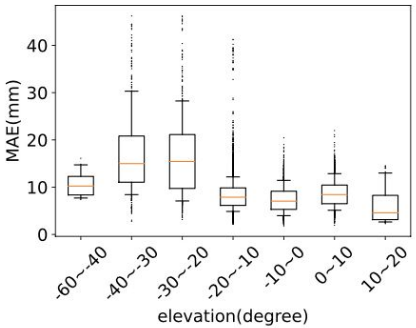
(f) 3 orientations of interest



(g) Palm rotation



(h) Azimuth

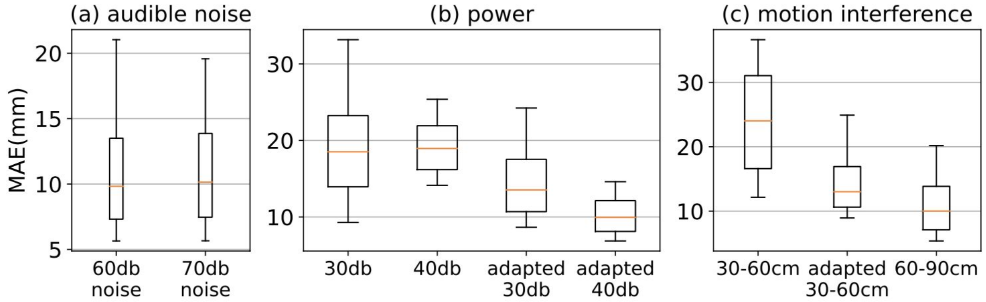


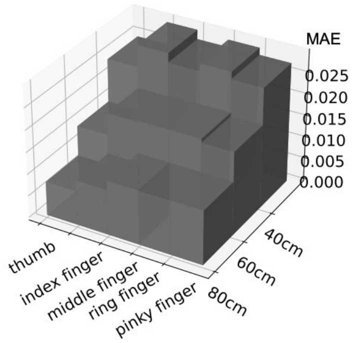
(i) Elevation

Error analysis from different perspectives

Experiment results
- *effect of interference*

Evaluation - effect of **interference**: audible noise, moving objects, and reduce in power





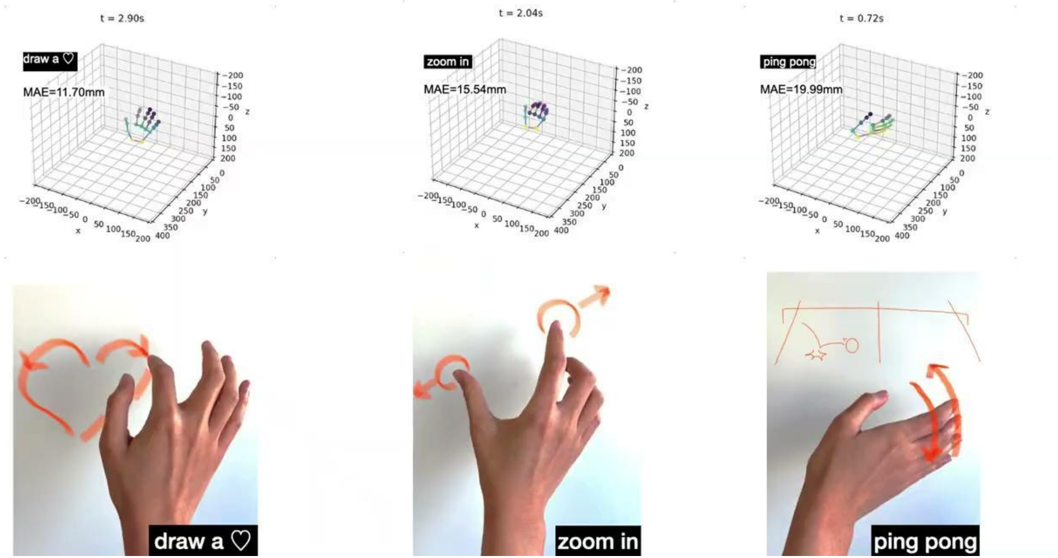
(a) The audible noise does not affect the system performance. (b) The accuracy drops when ultrasound volume is <50db. (c) Nearby motion interferes the accuracy. (b, c) But adaptive training helps.

All fingers' MAE decrease with distance because 2.5D is proportional to image size

Experiment results
- demo applications

Demo applications

The first row is our prediction. The second row illustrate the gesture and its potential application. (We do each individual gesture repeatedly)



Conclusion

- We build a fine-grained acoustic sensing system for hand tracking.
- It continuously tracks 21 joints in 3D
- It leverages on-device speaker and microphone with no hardware modification.
- Results show it work user-independently across environments.

Thanks

Feel free to reach out:

- **[GitHub]** https://github.com/lydhr/**Beyond-Voice**
- 📧 yl3243@cornell.edu
- 🌐 lynneli.xyz

Questions