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

Yin Li, Rohan Reddy, Cheng Zhang, Rajalakshmi Nandakumar

Cornell University

Abstract

Motivation: The voice user interface (VUI) of home assistants has *accessibility and usability* issues; some latest ones are equipped with additional cameras and displays, but are costly and raise privacy concerns.

These concerns jointly motivate Beyond-Voice: a novel high-fidelity acoustic sensing system that allows *commodity home assistant devices to track and reconstruct hand poses continuously.*

- It transforms the home assistant into an active sonar system using its **existing onboard microphones and speakers.**
- We feed a high-resolution range profile to the deep learning model that can analyze the motions of multiple body parts and predict the **3D positions of 21 finger joints**, bringing the granularity for acoustic hand tracking to the next level.
- It operates **across different environments and users** without the need for personalized training data.
- A **user study** with 11 participants in 3 different environments shows that Beyond-Voice can track joints with an average mean absolute error of **16.47mm** without any training data provided by the testing subject.

System overview

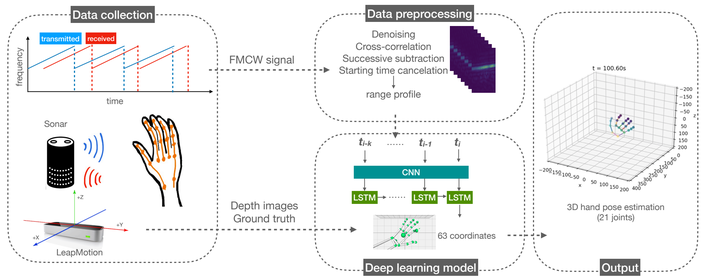


Figure 1. System Overview

- In data collection:
  - Ultrasound signal: we transmit inaudible FMCW.
  - Ground truth: simultaneously, Leap Motion depth camera collects the ground truth(only required in training but not in final use).
- In data preprocessing:
  - High-pass filter eliminates the audible noise.
  - Time-domain cross-correlation of FMCW yields high-resolution time-of-flight.
  - Accelerate the computation by frequency-domain cross-correlation.
  - Successive subtraction removes the reflections from the static environment. (Fig. 2)
  - Hardware starting time cancellation.
- Deep learning model:
  - the data from multi-microphones merge into a multi-channel feature map
  - along with the ground truth, they are input into a CNN+LSTM model.
- At the inference phase:
  - the output is a 3D skeleton
  - a camera is not required.

Method

The recipe of our effective system is attributed to the signal processing design and training strategies.

Specifically, following the cross-correlation-based dechirping pipeline as shown in Fig. 2, it yields a **high-resolution range profile**, providing a range resolution of $\Delta d = \frac{1}{fs} \times c \times \frac{1}{2} = 343/48000 \times 2 = 0.00357m = 3.57mm$, which is much better than the traditional method of $\Delta d = \frac{c}{2B} = 343/2/3000 = 0.05717m = 57.17mm$

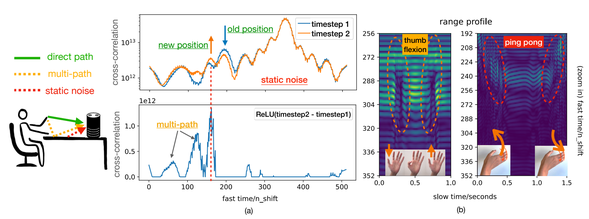


Figure 2. Signal processing: (a) Eliminate static noise by successive subtraction on cross-correlation windows. Each peak represents the strength of reflections y from objects at distance x. (b) Align fast time along slow time to get a range profile. The exemplary spectrums show that continuous motions result in slashes + side lobes.

Training strategies: data augmentation & Curriculum Learning as pre-training

Overfitting is a common issue in pose estimation systems, since the search space is large. So, our system uses two strategies in training the deep learning model:

- Data augmentation by shifting the range profile horizontally increases the training data size.
- Pre-train the model with curriculum learning(CL): CL trains the model hierarchically from simple gesture sets to complex finger motions as shown in Fig. 3; otherwise, we observed that the model might converge at a static pose sometimes.

Figure 3. Poses from single flexions to multiple flexions, for hierarchical Curriculum Learning.

Implementation

Our system consists of (1) a development microphone array board whose layout and sensitivity are the same as **Amazon Echo 2** Home assistant, (2) a speaker and (3) a Leap Motion infrared camera which is only for collecting ground truth in training.

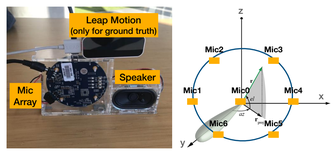


Figure 4. Hardware setup and the layout of the microphone array.

Experiment results

Cross-environment: we conducted a user study with 11 participants across an office, a bedroom, and a small study room. The average leave-one-room-out MAE is 15.73mm.

Cross-user: then we evaluate our system performance and its generalizability across users, i.e. varies the amount of user's training data, as shown in Table. 1.

	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

Table 1. Mean absolute error(mm).

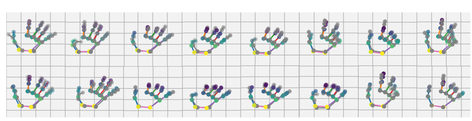
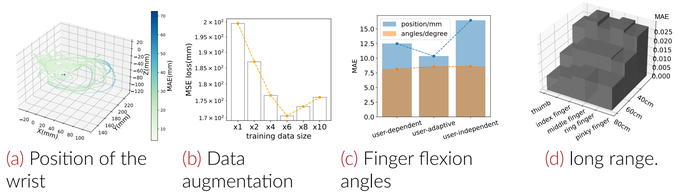


Figure 5. Sample results visualization: where the grey ground truth skeleton and cyan prediction skeleton mostly overlapped together.

Benchmark the performance of the system for individual factors such as range, flexion angles, data augmentation efficacy, etc, as shown in Fig. 6. Note that the long-range result is normalized to image size.



(a) Position of the wrist (b) Data augmentation (c) Finger flexion angles (d) long range.

Figure 6. Error analysis from different perspectives.

Validation experiments under various inferences and observe that

- The audible noise does not affect the system performance.
- The accuracy drops when ultrasound volume is <50db.
- Nearby motion interferes the accuracy. But adaptive training helps.

Demo applications

To better understand the usability of Beyond-Voice, we test certain intuitive applications that need continuous and absolute-range hand tracking.