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

Yin Li, Rohan Reddy, Cheng Zhang, Rajalakshmi Nandakumar
Cornell University
{yl3243,rr784,chengzhang,rajalakshmi.nandakumar}@cornell.edu

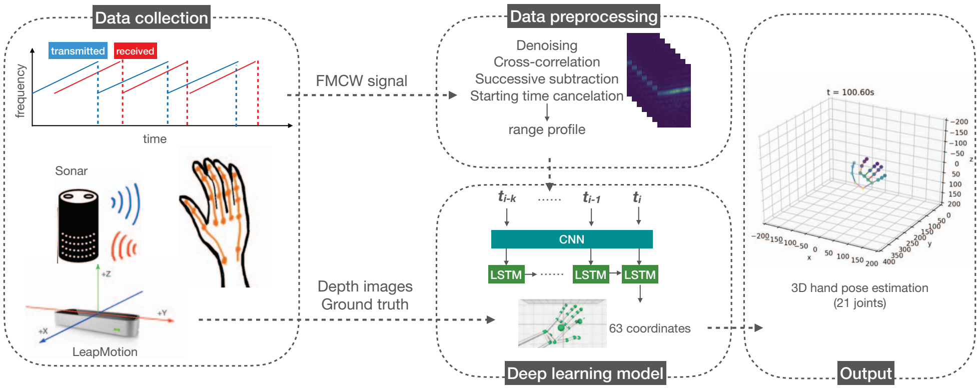

Figure 1: System overview.

ABSTRACT

The surging popularity of home assistants and their voice user interface (VUI) have made them an ideal central control hub for smart home devices. However, current form factors heavily rely on VUI, which poses accessibility and usability issues; some latest ones are equipped with additional cameras and displays, which 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 device into an active sonar system using its existing onboard microphones and speakers. By feeding a high-resolution range profile to the deep learning model, we can localize 21 finger joints in 3D, bringing the granularity for acoustic hand tracking to the next level. 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 for unseen environments and users.

KEYWORDS

acoustic sensing, wireless perception, hand pose estimation

1 INTRODUCTION

Commercial home assistant devices, such as Amazon Echo, primarily employ VUI for interaction. While reliance on a speech interface raises (1) accessibility concerns by precluding users with speech disabilities and (2) usability concerns stemming from misinterpretation of non-native speech or background noise. Some new models feature cameras for motion tracking and touch displays, but they are costly, not widely available, and raise privacy concerns.

In this paper, we propose a beyond-voice method of interaction with these devices. It leverages the existing acoustic sensors of commercial home assistant devices to enable continuous fine-grained hand tracking. In comparison, current acoustic hand tracking systems have insufficient detection granularity, i.e. discrete gestures classification[9, 11], or localize a single nearest point, or up to 2 points per hand [3, 7, 10]. Our system enables fine-grained multi-target tracking of the hand pose by 3D localizing the 21 individual joints of the hand. It's not bounded to predefined gestures, thus significantly improving the expressiveness of the gestures and user experience, allowing for interactions such as i) using continuous gesture commands, like zoom-in or turning a knob to adjust the volume. ii) sign language communication.

As for other modalities, wearables such as EMG wristbands [5] and gloves[6], can implement high-fidelity tracking or haptics. But they are usually cumbersome to wear during daily activities. RGB cameras[4] and depth cameras[8] can track hands in 2.5D or 3D with under-centimeter error. However, vision-based cameras have limitations on non-line-of-sight parts of hand gestures and raise privacy concerns, especially in home-use scenarios. In contrast, our acoustic-sensing-only tracking offers a promising alternative with comparable accuracy and fidelity to penetrate through occlusion, while posing fewer privacy issues associated with wireless perception. Besides, wireless perception via RF signals is popular for human tracking using mmWave[1] or WiFi[2, 12]. But they require custom hardware or huge bandwidth that is computationally expensive. While our system only uses basic acoustics sensors. And sound speed is orders of magnitude slower than RF, which yields a great solution without wide bandwidth.

2 SYSTEM OVERVIEW

Our system enhances the detection granularity of acoustic sensing to enable articulated hand pose tracking by leveraging the existing speaker and microphones on device. The key is to transform the device into an active sonar system, emitting inaudible ultrasound

979-8-3503-6201-5/24/$31.00 ©2024 IEEE
DOI 10.1109/IPSN61024.2024.00045

chirps (Frequency Modulated Continuous Wave, FMCW) from the speaker and recording reflections with a co-located microphone array. Then extracting signal's time-of-flight allows us to 3D localize the 21 finger joints.

Building a continuous hand-tracking system poses several challenges. First, the system needs to work even in unseen environments. Therefore, we design a signal processing pipeline that can eliminate unwanted reflections and then combine multiple microphones to localize the hand in 3D. Nevertheless, the reflections from joints are entangled making it intractable to separate them with rule-based algorithms, especially in the presence of multi-path noise from moving fingers. Hence, we use a Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) model to learn the patterns. In training, we use a Leap Motion depth camera as ground truth and a curriculum learning (CL) and technique to hierarchically pre-train the model. Secondly, ensuring the system works across different ranges and orientations necessitates a large data collection effort. We mitigate this by designing a custom data augmentation method for range profiles, also improving the generalization of standard pose estimation models. Lastly, achieving user-independent performance without adaptive training poses a challenge, since cheap form factors do not come with a ground truth camera. Thereby, we collect a one-time extensive training dataset from multiple users, which can then directly apply to unseen users after deployment. Furthermore, we also demonstrate the user-adaptive results in case cameras are available in some form-factors. Fig. 1 illustrates the full system with sample output.

In summary, this paper presents the following key contributions:

- We develop a novel fine-grained 3D hand tracking system, leveraging the existing acoustic sensors in the home assistant devices. Our continuous tracking of 21 finger joints is unbounded to predefined gestures, enabling versatile downstream applications.
- Our deep learning model can generalize across both environments and users without personalized adaptive training.
- We evaluated our system with a user study with 11 users in three different environments across different days. The system yields an average error of 16.47mm (median 14.57) user-independently.

3 EVALUATION

We deploy Beyond-Voice on a development board with similar hardware settings to Amazon Echo Dot 2. A user study with 11 users provides a total of 64 minutes of data, which consists of carefully selected hand motions that expressively cover the movements of all finger joints. Besides, we pre-trained a model using curriculum learning with 40 minutes of data from two researchers excluded from the user study.

Table. 1 summarizes our evaluation results and generalizability across users, i.e. varies the amount of user's training data, including user-independent, user-adaptive, and user-dependent tests. The user-independent MAE of 16.47mm (median 14.57mm) shows our system's ability for hand tracking without prior training data from new users. If adding two minutes of data from a new user for adaptive training, the MAE can further decrease to 10.36mm (median 9.72mm). Furthermore, subgrouping by the data collection location confirms that it works in unknown environments. Despite different rooms (open-space office, bedroom, small study room), preprocessing steps mitigate environmental noise, resulting in an average leave-one-room-out MAE of 15.73mm.

	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 MAE (mm).

To better understand the usability of Beyond-Voice, we test certain intuitive applications that need continuous and absolute-range hand tracking, such as i) *drawing in the air* ii) *zoom in* iii) *sign language* iv) *grabbing and placing objects at a specific position* v) *playing ping pong*. Their MAEs are 11.70mm, 15.54mm, 13.48mm, 15.54mm, and 19.99mm respectively. And a supplementary video linked here visualizes our results at a 10% subsampled frame rate.

Our long paper further presents benchmark and validation results, considering factors like range, finger/bone, and training data size. We also validate the system's robustness in the presence of different environmental noises, covering audible noise, other motions in the environment, and nearby objects or metals.

REFERENCES

[1] Eiji Hayashi, Jaime Lien, Nicholas Gillian, Leonardo Giusti, Dave Weber, Jin Yamanaka, Lauren Bedal, and Ivan Poupyrev. 2021. Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor. In *Proceedings of the Conference on Human Factors in Computing Systems*. 1–14.

[2] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D human pose construction using wifi. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 1–14.

[3] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2020. FM-track: pushing the limits of contactless multi-target tracking using acoustic signals. In *Proceedings of the Conference on Embedded Networked Sensor Systems*. 150–163.

[4] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. 2022. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2761–2770.

[5] Yang Liu, Chengdong Lin, and Zhenjiang Li. 2021. WR-Hand: Wearable armband can track user's hand. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[6] Meta. 2021. Inside Reality Labs Research: Bringing Touch to the Virtual World. (November 2021). https://about.fb.com/news/2021/11/reality-labs-haptic-gloves-research/

[7] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. Fingerio: Using active sonar for fine-grained finger tracking. In *Proceedings of the Conference on Human Factors in Computing Systems*. 1515–1525.

[8] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. 2019. Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10853–10862.

[9] Penghao Wang, Ruobing Jiang, and Chao Liu. 2022. Amaging: Acoustic Hand Imaging for Self-adaptive Gesture Recognition. In *Proceedings of the IEEE Conference on Computer Communications*. 80–89. https://doi.org/10.1109/INFOCOM48880.2022.9796906

[10] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 82–94.

[11] Zhizheng Yang, Xun Wang, Dongyu Xia, Wei Wang, and Haipeng Dai. 2023. Sequence-Based Device-Free Gesture Recognition Framework for Multi-Channel Acoustic Signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1–5.

[12] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.