MASS 2021

Fine-grained Multi-user Device-Free Gesture Tracking on Today's Smart Speakers

Ningzhi Zhu[1], Huangxun Chen[2], Zhice Yang[1]

[1]ShanghaiTech University

[2]Huawei Theory Lab

The Booming Smart Speaker Market

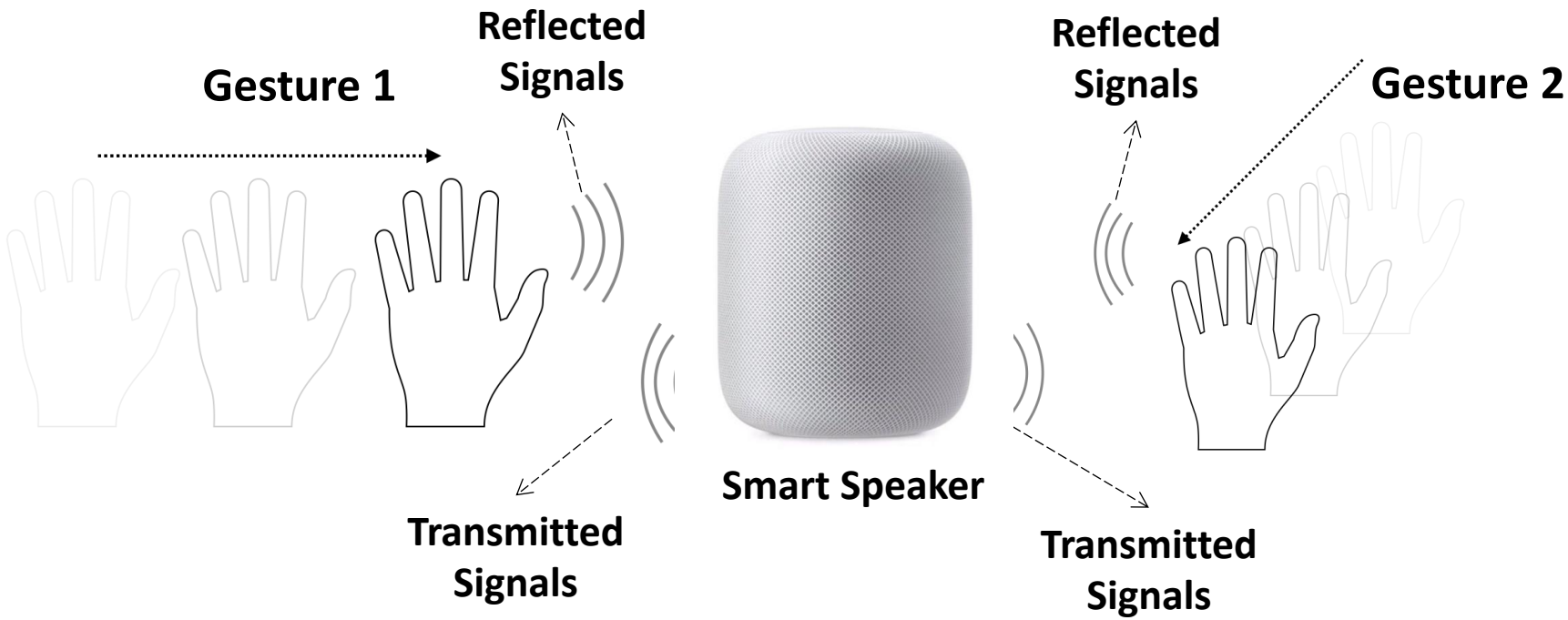
The global market is expected to reach $17.85 billion in 2025

Interaction with Smart Speakers: Voice

Smart Speaker

Mandatory quiet areas and privacy concerns may limit the usage of voice interaction.

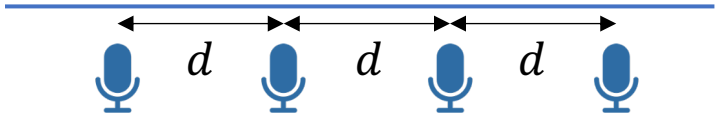
Interaction with Smart Speakers: Gesture



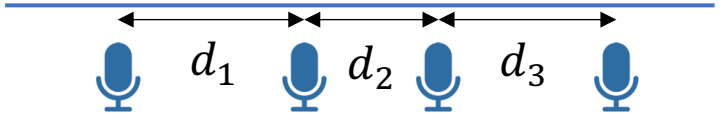
Repurpose the smart speaker as an active sonar to provide gesture-based interaction

Gesture Tracking on Linear Array

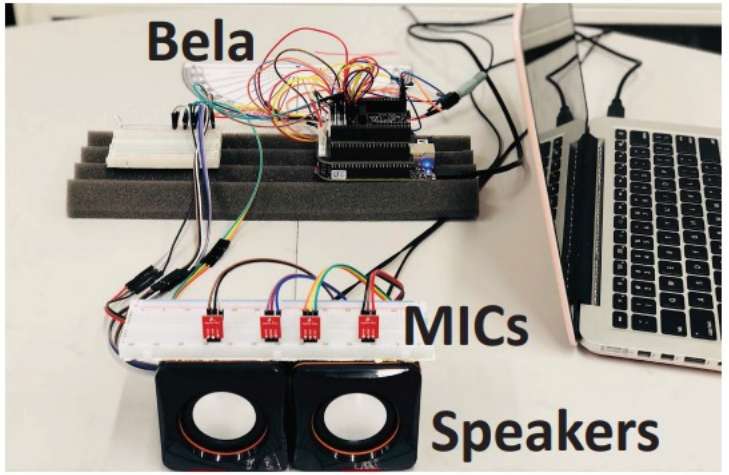
- Prior efforts have realized fine-grained acoustic-based gesture tracking on the uniform/non-uniform linear microphone arrays.



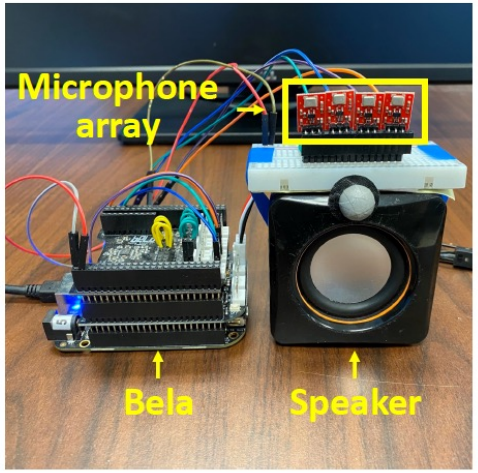
Uniform Linear Array (ULA)



Non-Uniform Linear Array (NULA)

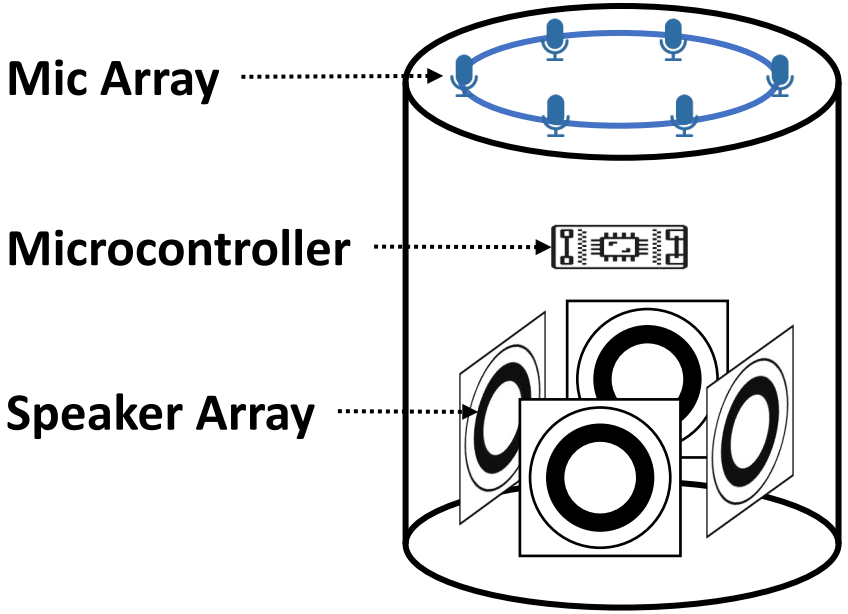


[Mobicom'19]



[Sensys'20]

Gesture Tracking on Uniform Circular Array

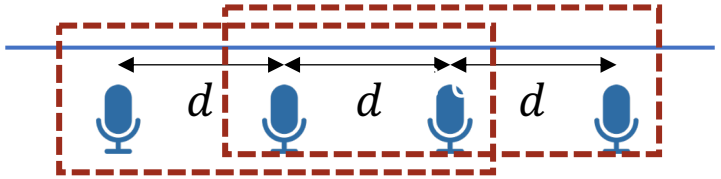


Product	Mic Layout	Mic Spacing
Amazon Echo	6-mic UCA	4.96 cm
Amazon Echo Dot	4-mic UCA	7.00 cm
Apple Homepod	6-mic UCA	7.10 cm
Sonos One	6-mic UCA	5.99 cm

Uniform Circular Array (UCA) is the dominating array layout of most commercial products

Linear Array V.S. Circular Array

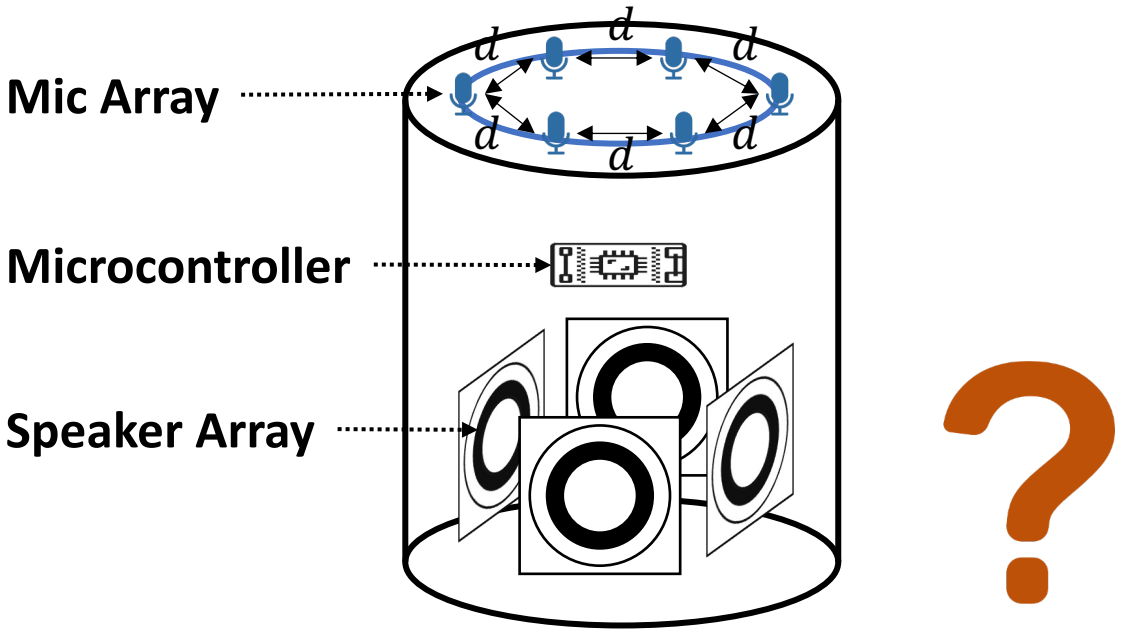
orientational invariant structure



Uniform Linear Array (ULA)

a small separation to reduce aliasing

Non-Uniform Linear Array (NULA)



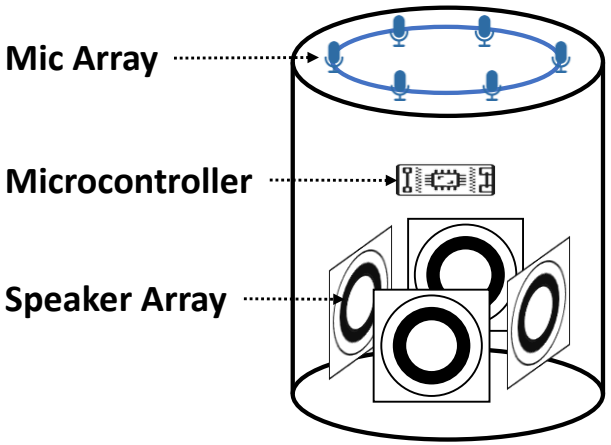
	LA in prior studies	**UCA of commercial smart speakers**
Signal Coherence	orientational invariant structure apply spatial smoothing	lack orientational invariant structure
Spatial Aliasing	ultrasound-wavelength-level mic-spacing non-uniform array	several centimeters mic-spacing uniform geometry

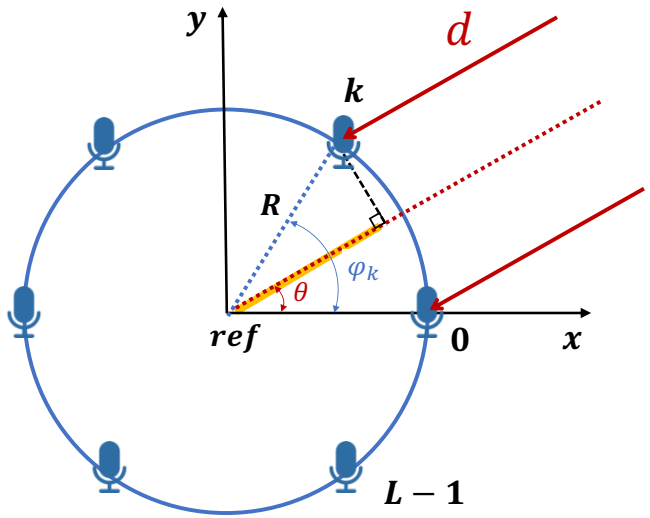
SparseTrack: Key Insights

- **Reflector Sparsity**
  - The number of significant moving reflectors that could have contributed to the overall reflected signal is limited

- SparseTrack treats multi-target gesture tracking from the **sparse recovery perspective**

Casting Gesture Tracking to Sparse Recovery





UCA Signal Modeling

Time domain waveform of candidate reflection positions (d, θ)

$$X = Dic \times C + \mathcal{N}$$

Noise

Measurement Matrix

Sparse vector for true reflection positions

Sparse Recovery Formulation

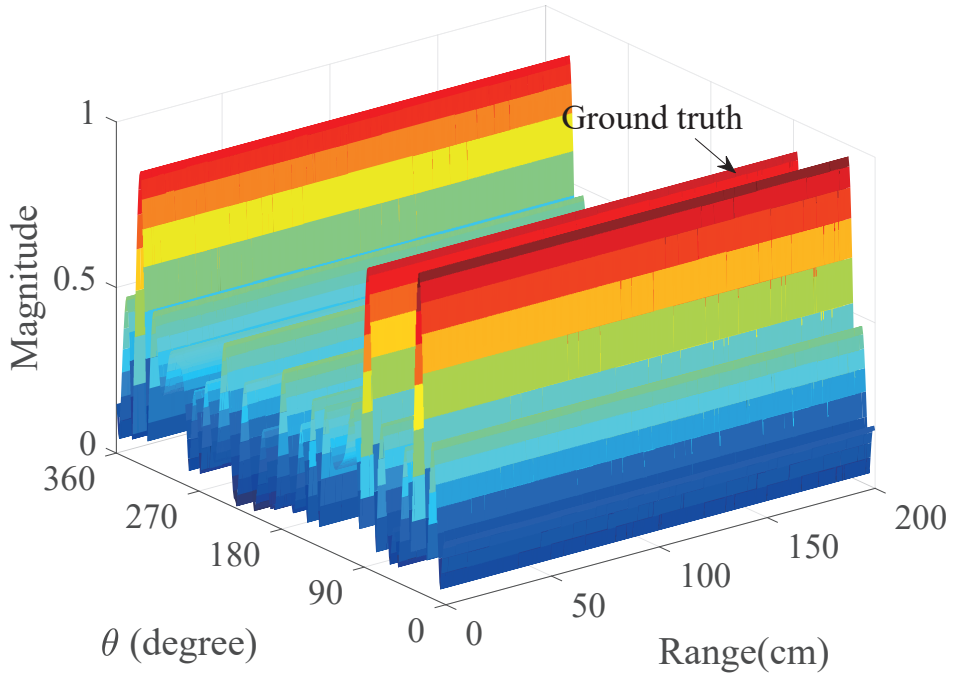
$$min \parallel C \parallel_0 \ \ s.t. \parallel X - Dic \times C \parallel_2 \leq \varepsilon$$

Find the smallest number of scaled and shifted reflection signals that could make up the overall signals received by the mic-array

Challenges & Solutions

Challenge 1: Spatial Ambiguity

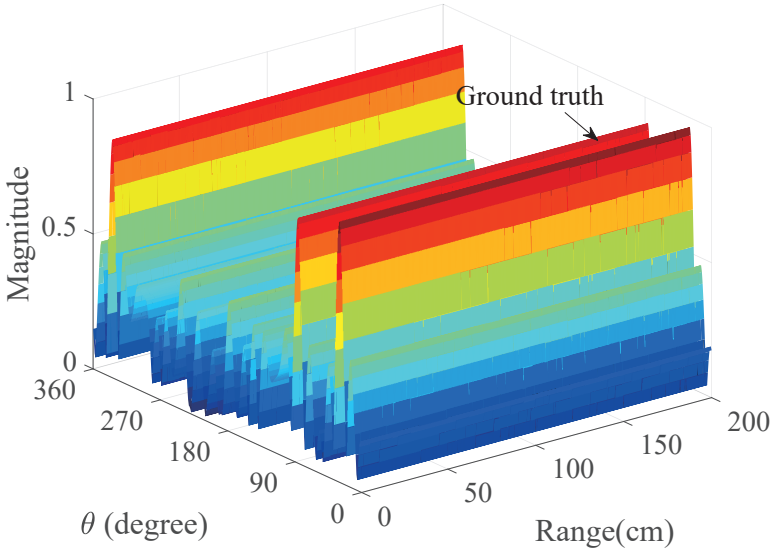
- **Root cause: insufficient spatial sampling rate**
  - Mic-spacing of commercial products: 5-7 cm
  - Wavelength of inaudible ultrasound (17-23 kHz): 1.5-2 cm
- Ambiguity issue exists in both range and direction domain.



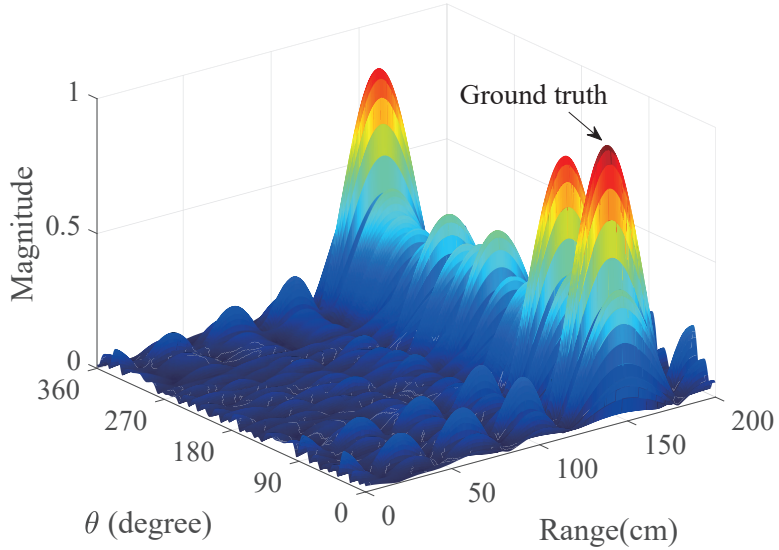
Mic-Spacing: 5 cm
Frequency: 17 kHz

Solution: Synthesizing Wideband Measurements

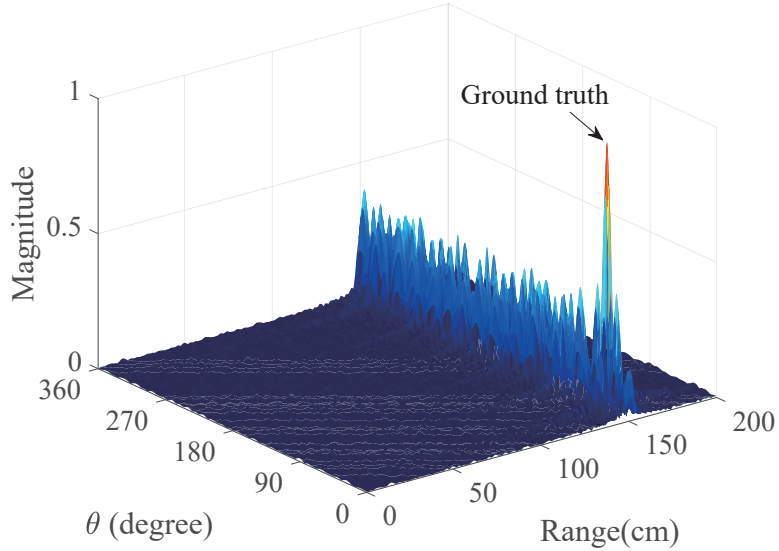
- **Key observation: frequency diversity of the transmitted signal helps!**
  - The measurement from each frequency component experiences different ambiguities, but all measurements include the positions of true reflectors.



Mic-Spacing: 5 cm
Frequency: 17 kHz



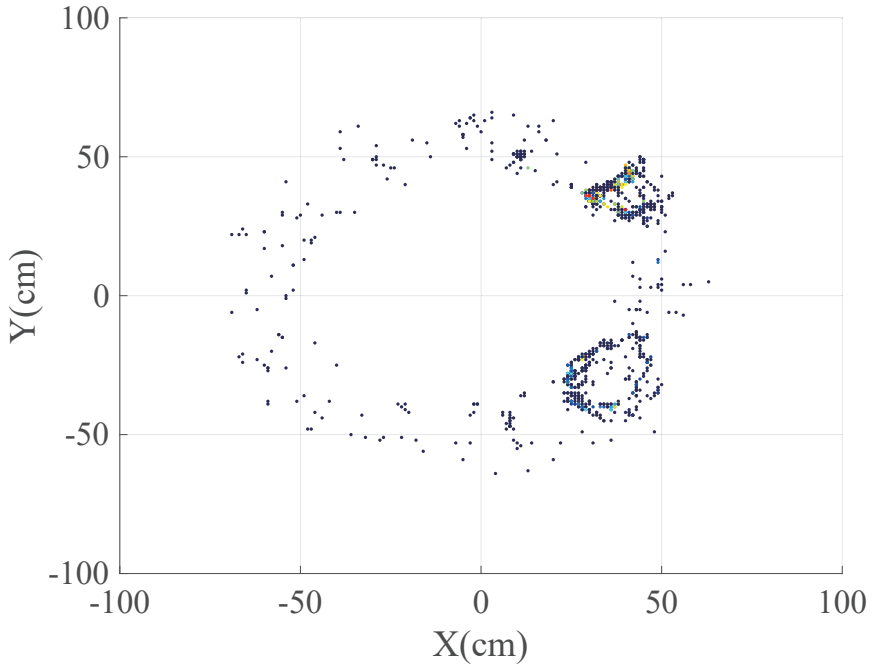
Mic-Spacing: 5 cm
Frequency: 17-17.5 kHz



Mic-Spacing: 5 cm
Frequency: 17-23 kHz

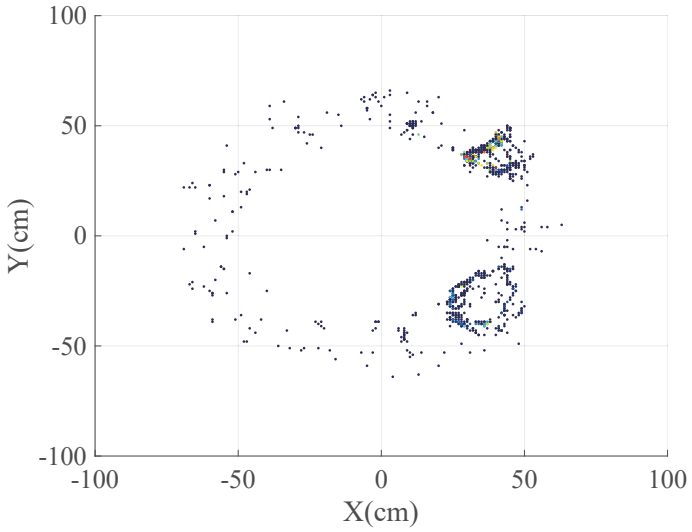
Challenge 2: Noisy Reflection Measurement

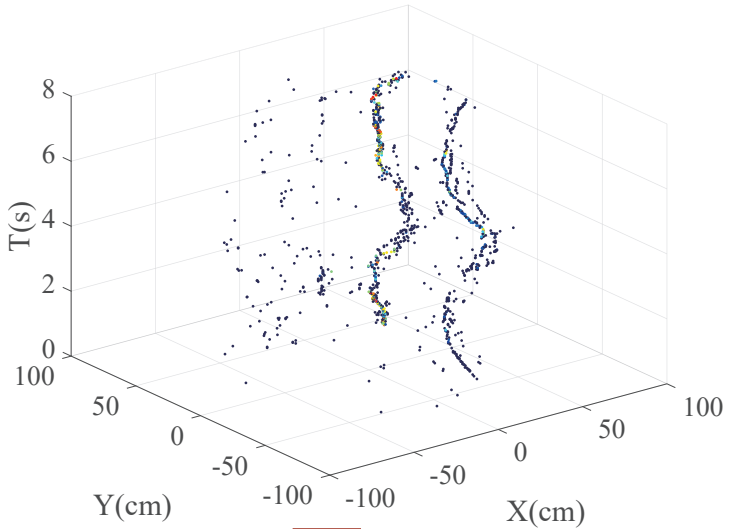
- The output of reflection localization is not noise-free.
- It is non-trivial to extract multiple gesture traces robustly.

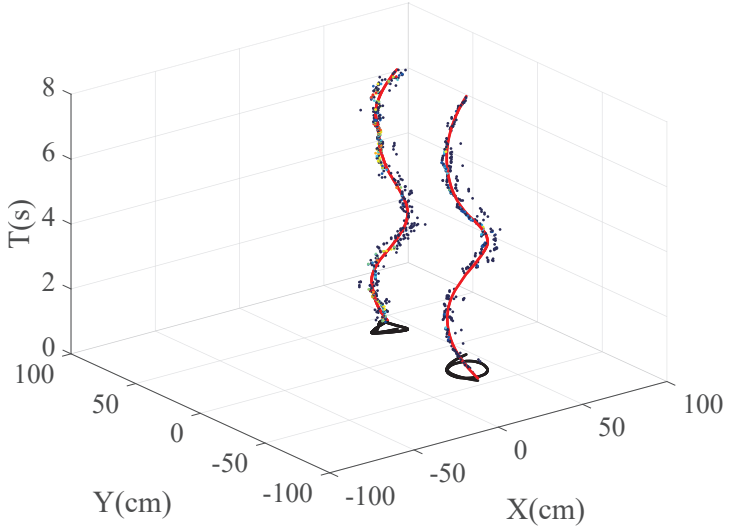


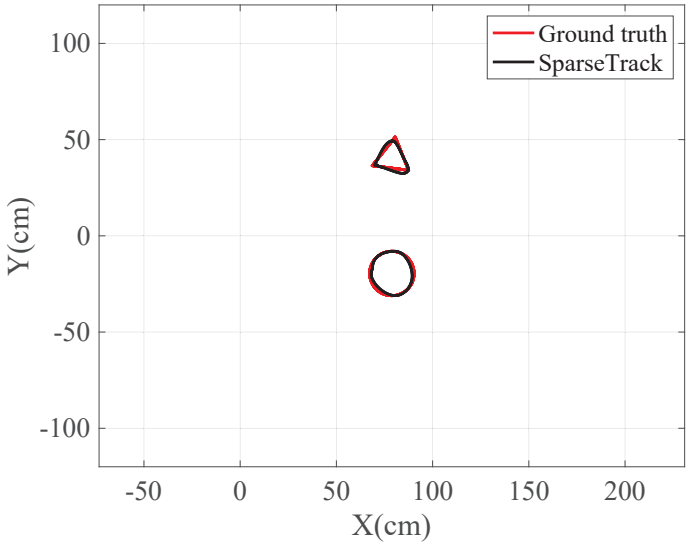
Two users are required to ‘draw a circle’ and ‘draw a triangle’ simultaneously.

Solution: Leverage Time-Domain Information









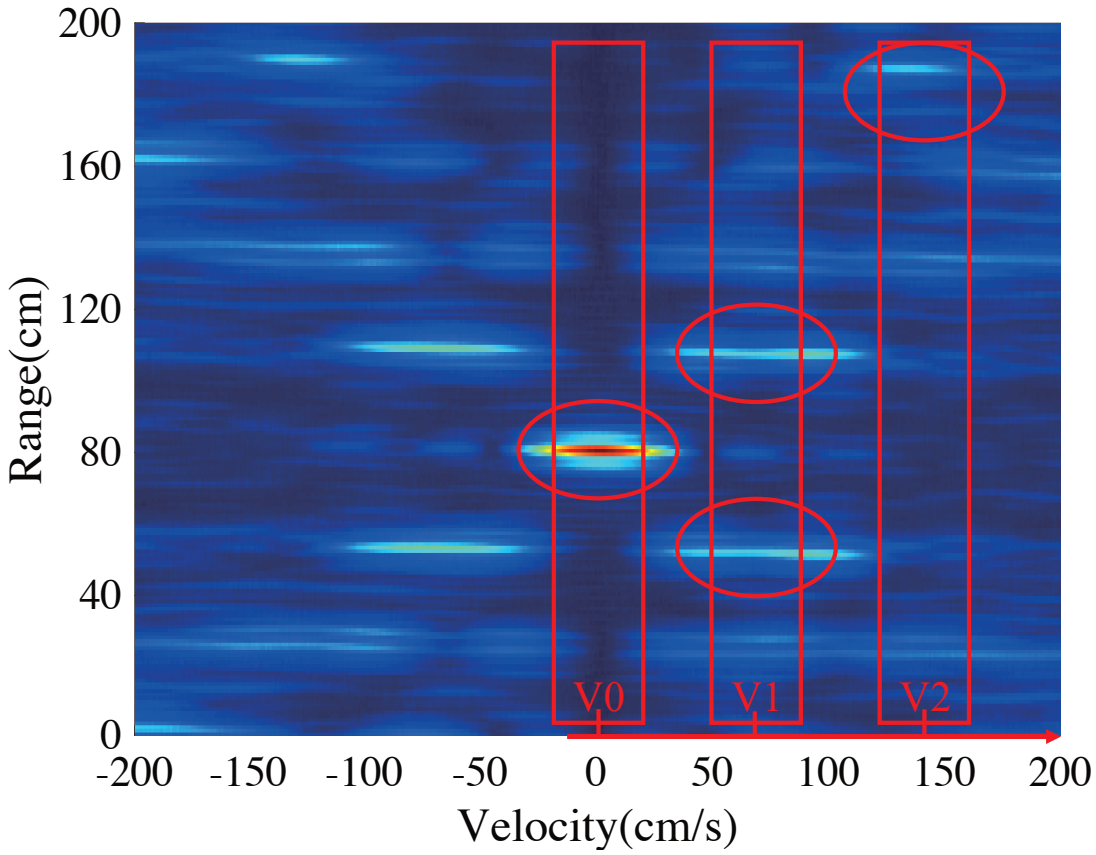
Time Dimension Expansion

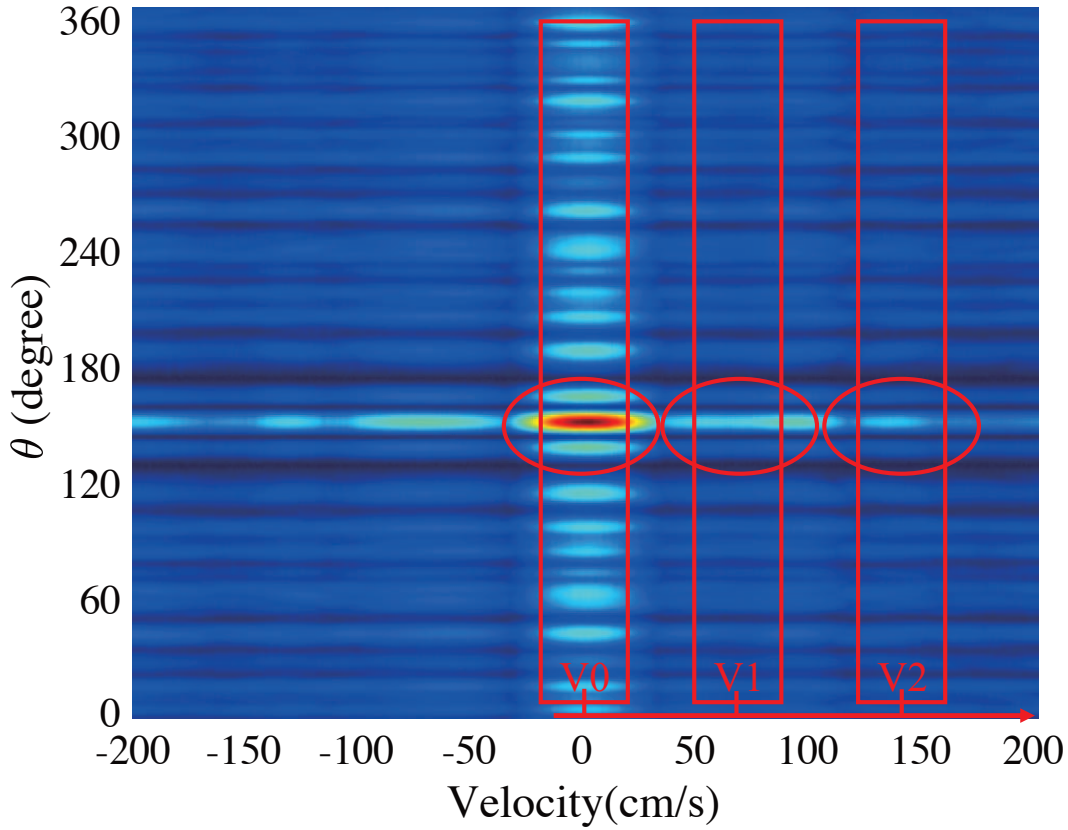
Valid Trace Identification

3D to 2D Projection

Challenge 3: Reflector Dynamics

- Doppler effect downgrades the performance of gesture tracking





Solution: Velocity-aware Dictionary

Time domain waveform of candidate reflection positions (d, θ)

$$X = Dic \times C + \mathcal{N}$$

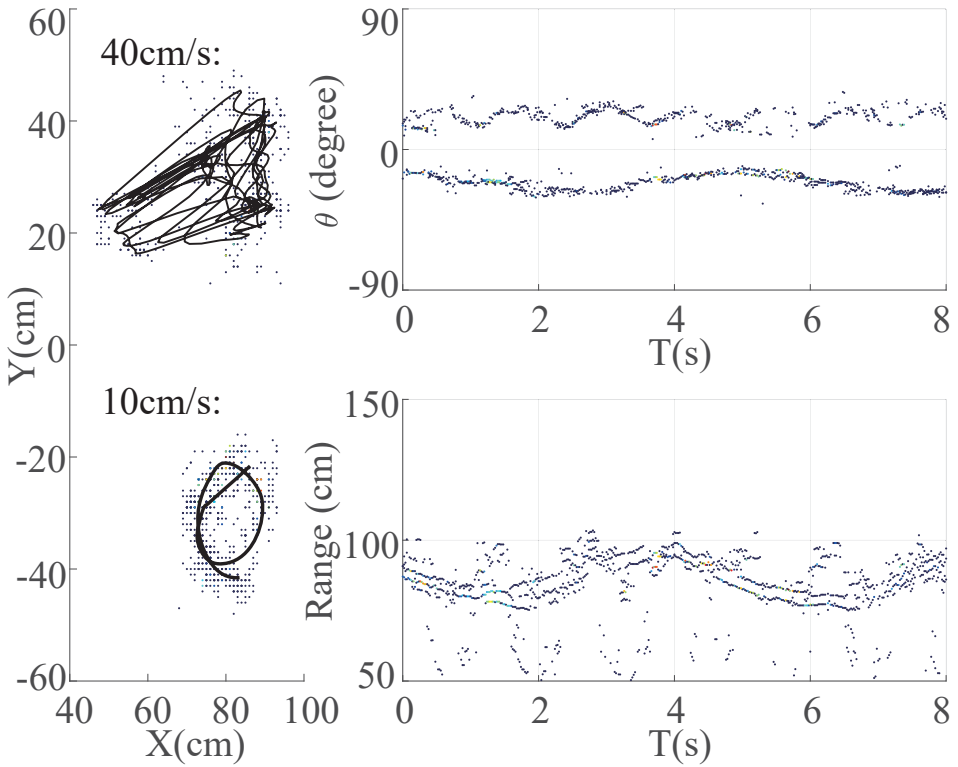
Measurement Matrix

Velocity-aware Dictionary: (d, θ, v)

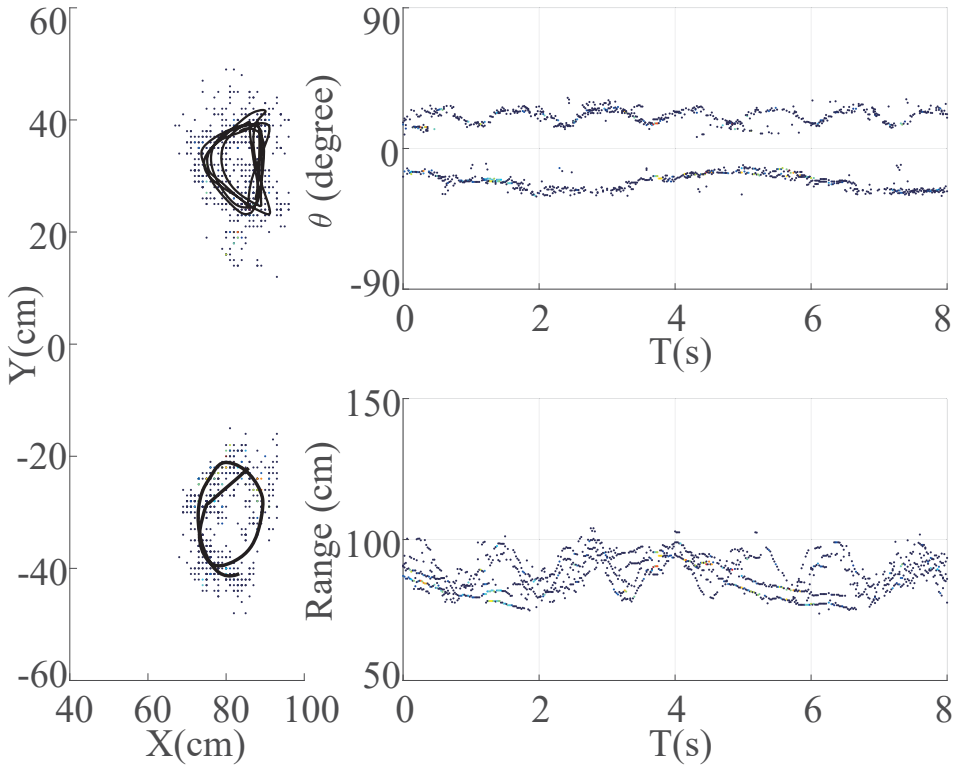
Sparse vector for true reflection positions

Noise

Solution: Velocity-aware Dictionary

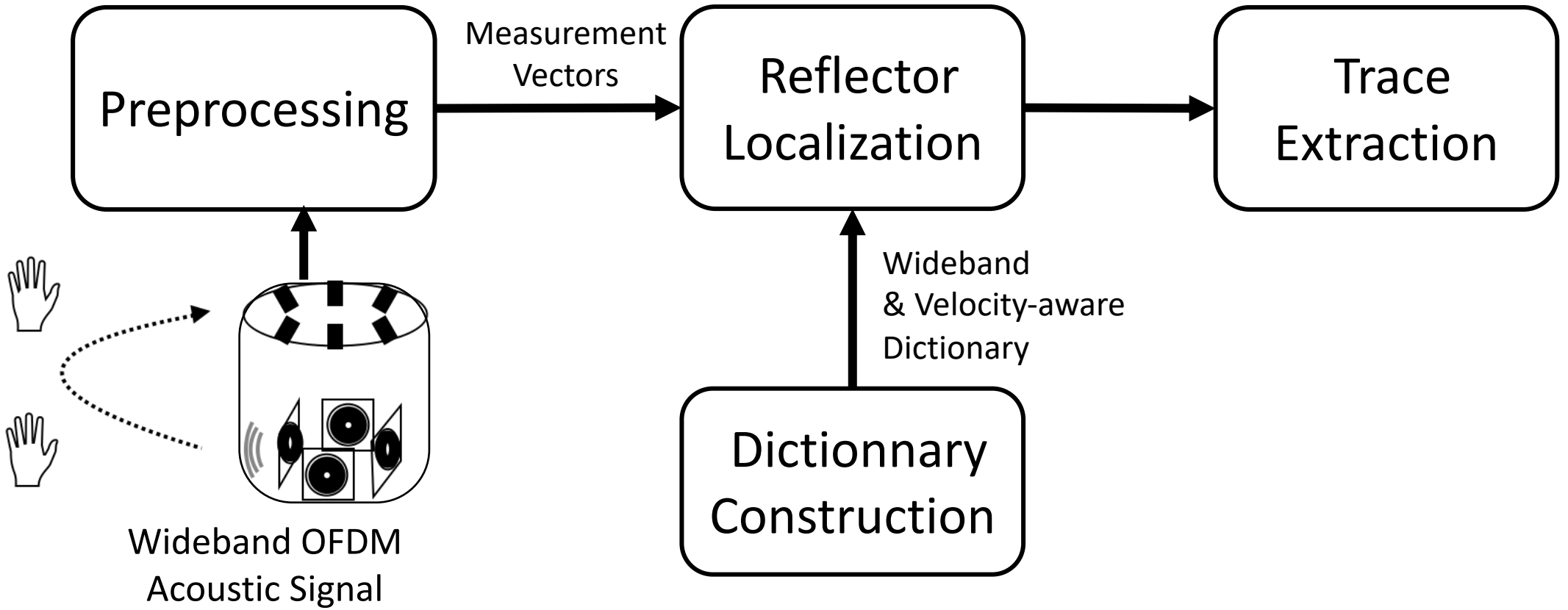


Without Velocity-aware Dictionary



With Velocity-aware Dictionary

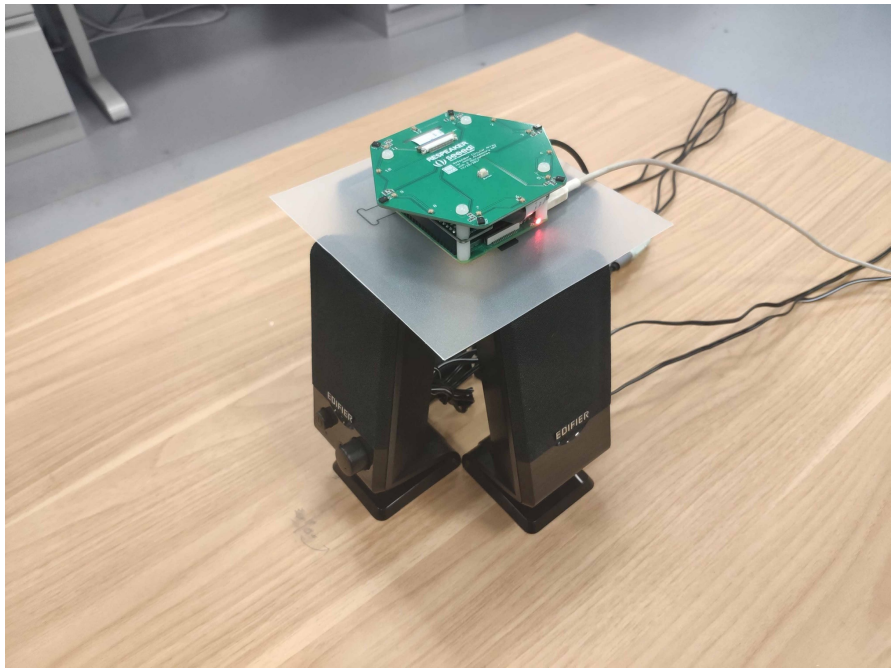
SparseTrack: Putting All Things Together



Implementation & Evaluation

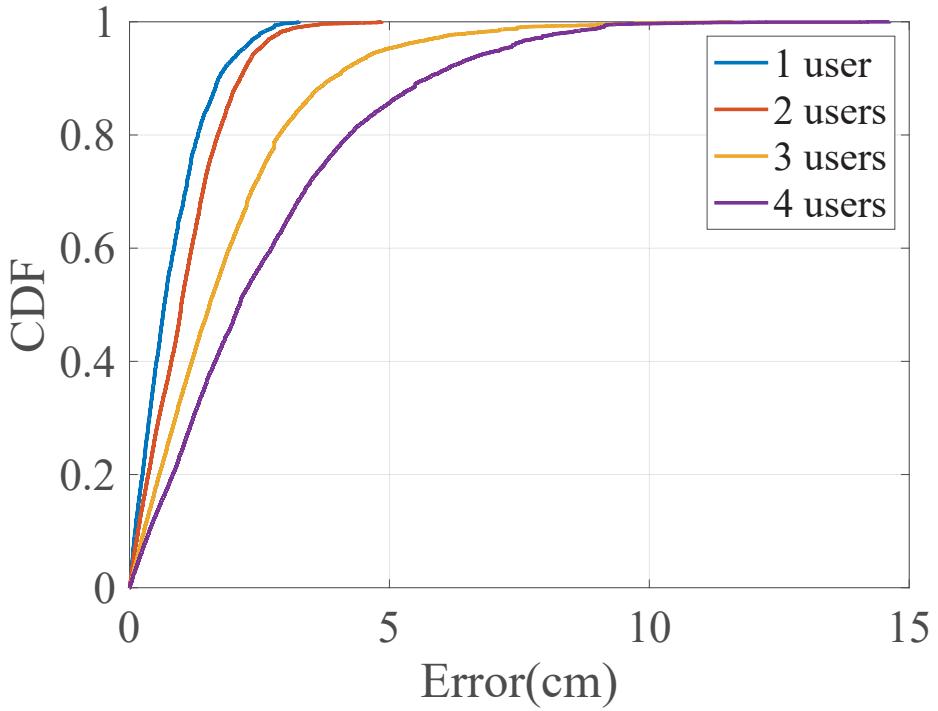
Evaluation Setting

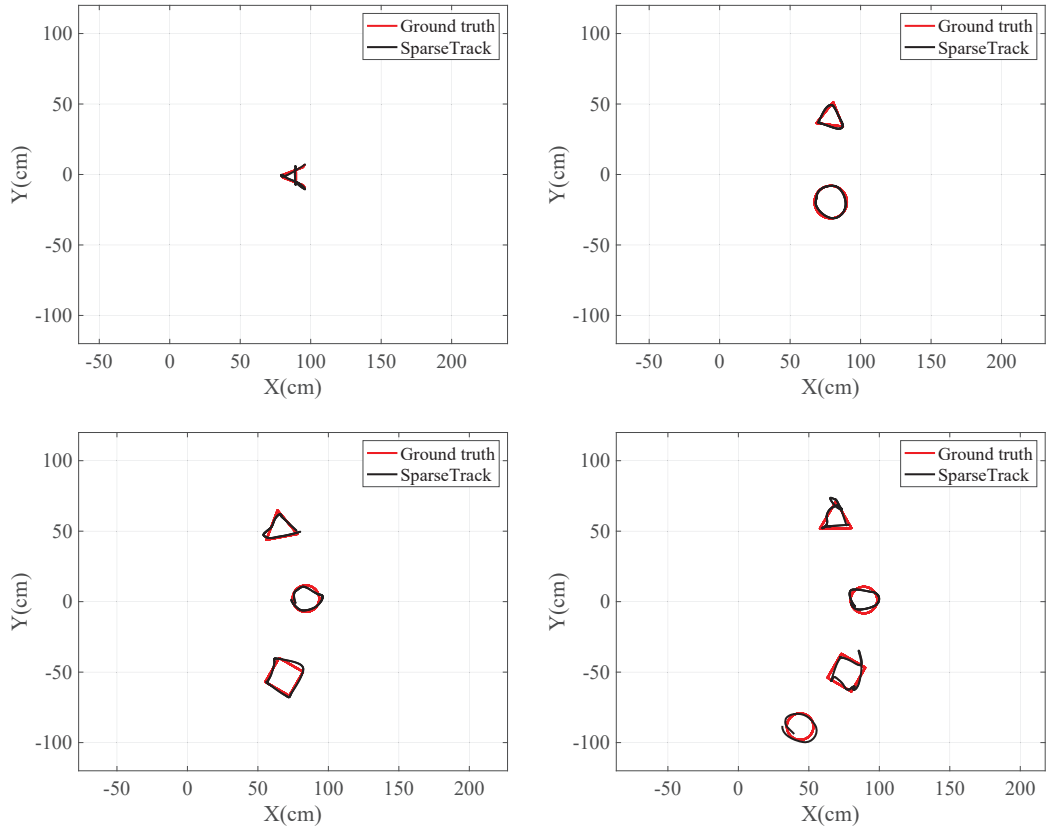
- The prototype has a similar layout as commodity smart speakers:
  - 4 Edifier M1250 speakers + 1 ReSpeaker 6-Mic UCA with 4.7 cm spacing
- 167 collected traces in total under different conditions



Evaluation: Overall Performance

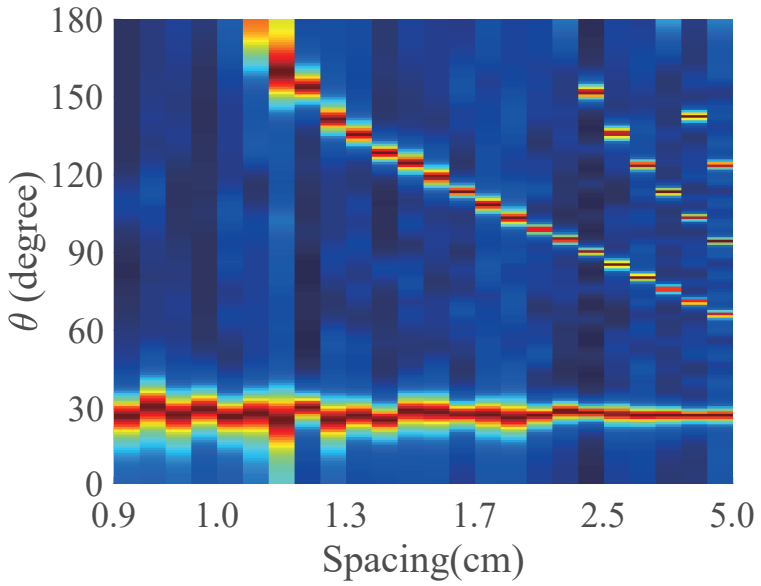
- SparseTrack can simultaneously track 1 to 4 users' gestures
- SparseTrack achieves a mean tracking error of 2.66 cm



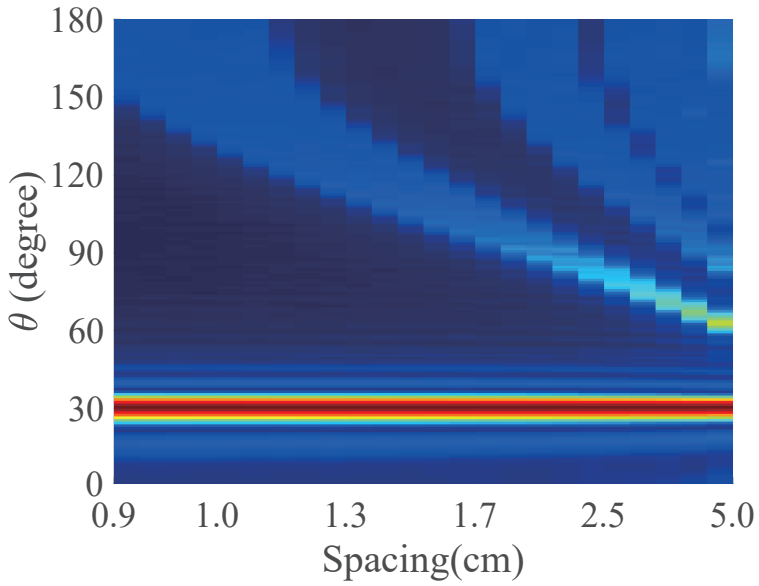


Evaluation: Handling Spatial Ambiguity

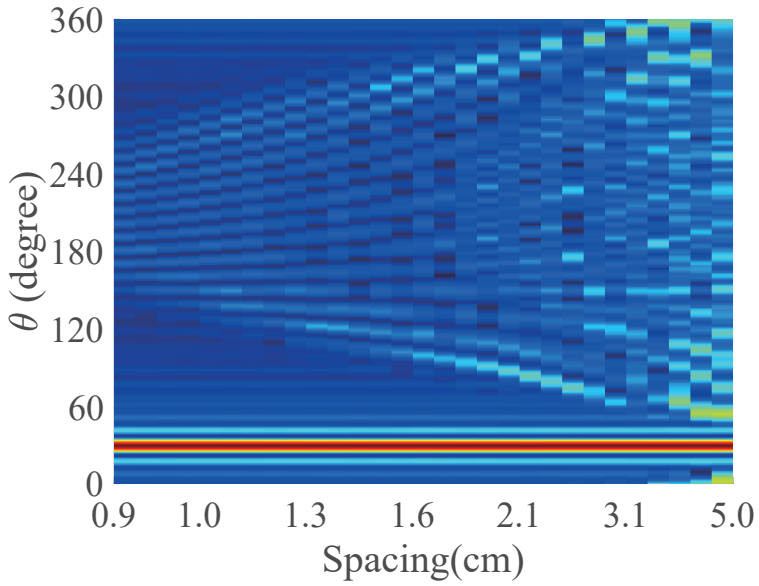
- SparseTrack can handle the spatial ambiguity issue even when the mic spacing is much larger than half of the wavelength.



2D MUSIC with ULA



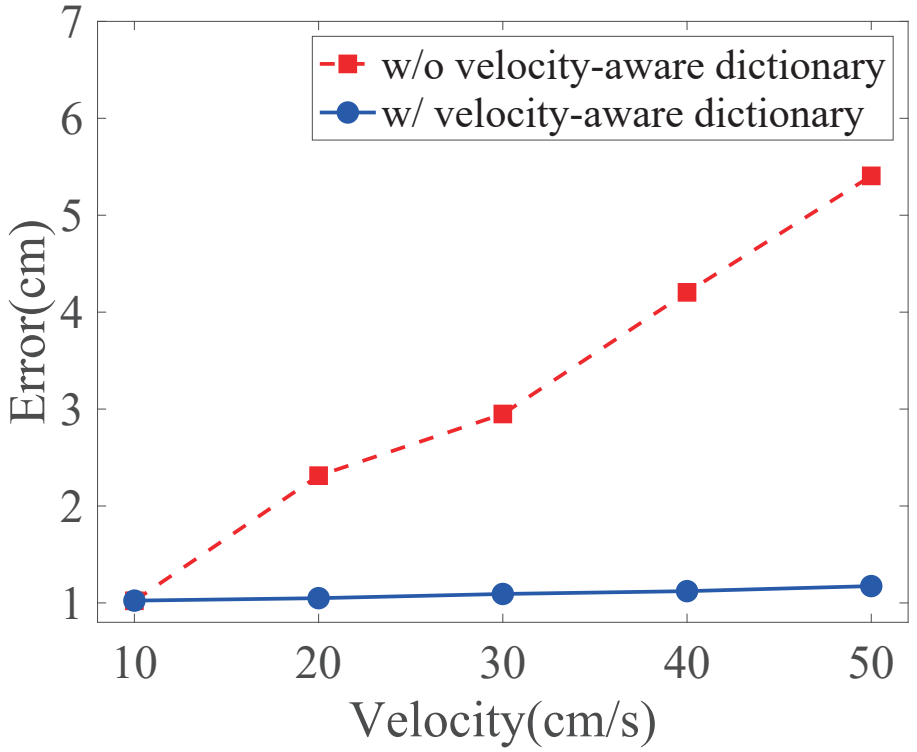
SparseTrack with ULA



SparseTrack with UCA

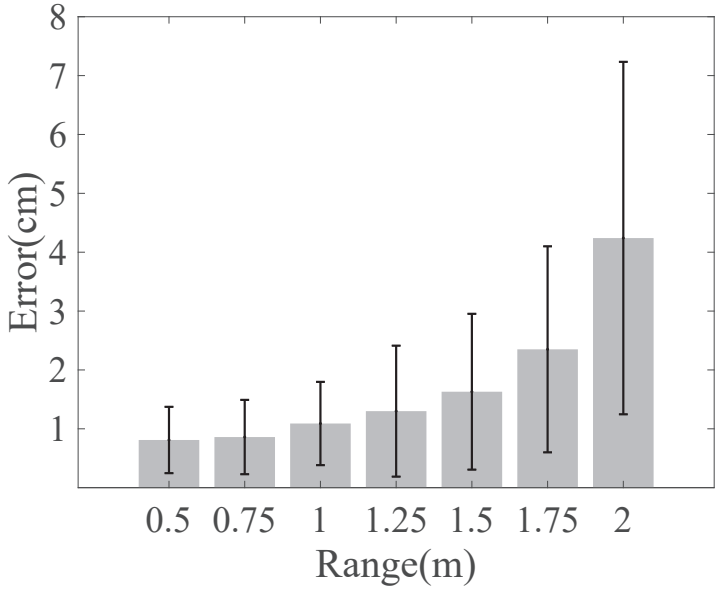
Evaluation: Handling Doppler Effect

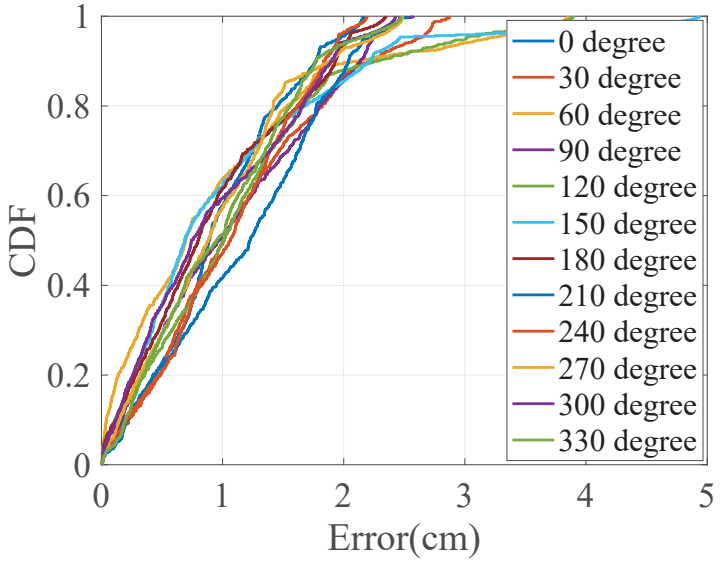
- The velocity-aware dictionary successfully handles the Doppler effect and achieves stable performance.

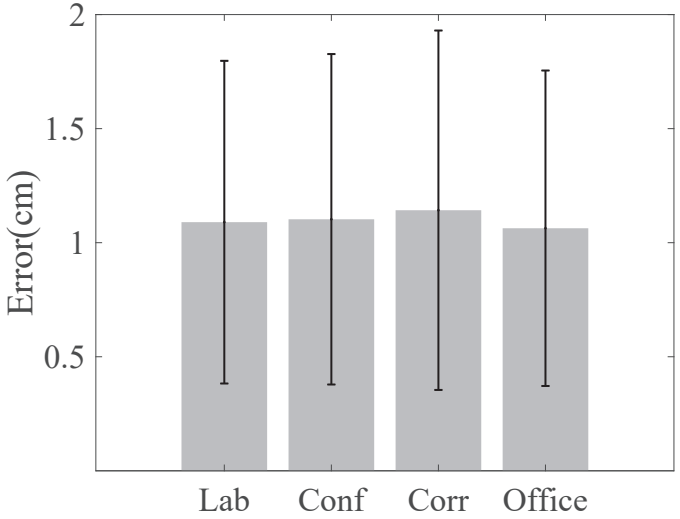


Evaluation: Impact Factors

- Good tracking performance within the range of 1.5 m
- Robust performance under different directions and locations







Conclusions

- SparseTrack achieves fine-grained multi-user device-free gesture tracking on smart speakers with uniform circular geometry.

- SparseTrack takes a step to mature the gesture interaction on commercial products.

Thanks!
Q&A