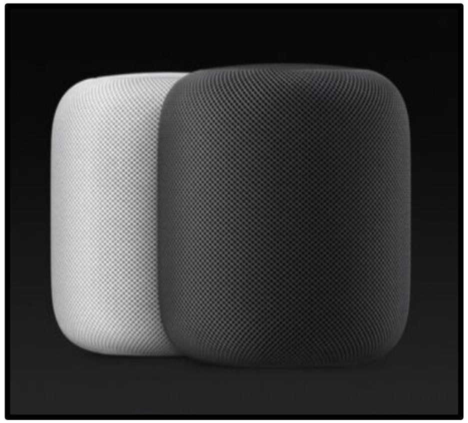
Opportunistic Sensing with Mic Arrays on **Smart Speakers** for **Distal Interaction** and **Exercise Tracking**

Anup Agarwal, **Mohit Jain**, Pratyush Kumar, Shwetak Patel

Smart Speakers

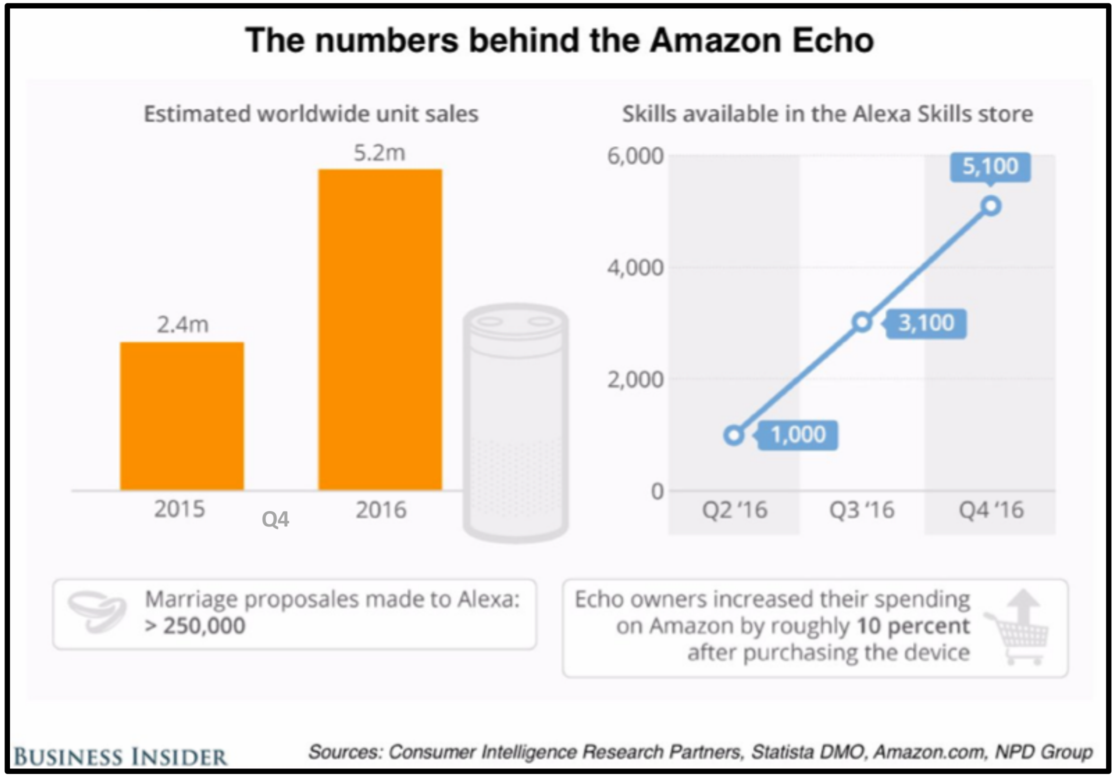
New class of voice-only devices offering hands-free interaction



Smart Speakers

New class of voice-only devices offering hands-free interaction

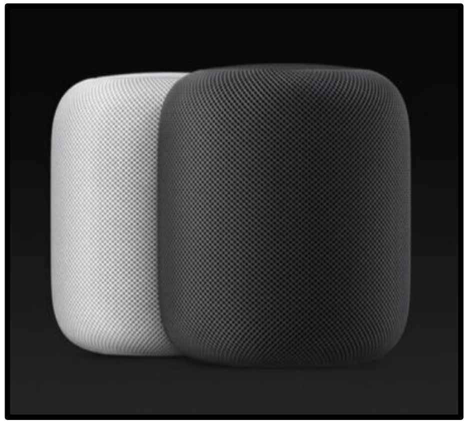
35.6M smart speakers sold in US in 2017, 129% more than 2016



Marriage proposales made to Alexa: **> 250,000**

Echo owners increased their spending on Amazon by roughly **10 percent** after purchasing the device

Sources: Consumer Intelligence Research Partners, Statista DMO, Amazon.com, NPD Group



Smart Speakers

Apple Homepod **6** Sonos One **6** Amazon Echo **7**

To increase the device’s range for recognizing voice commands from across the room using beamforming

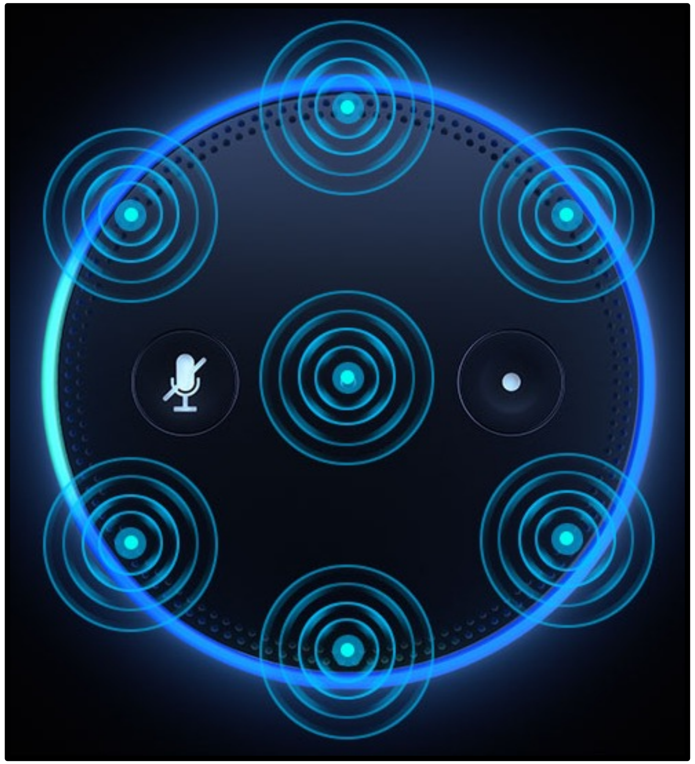
Smart Speakers

Apple Homepod **6** Sonos One **6** Amazon Echo **7**

To increase the device's range for recognizing voice commands from across the room using beamforming

Beamforming: The signals from the each mic are combined in a way that signals coming from a certain direction in space interfere constructively while others interfere destructively.

Delay-and-Sum beamforming



Problem

For certain scenarios voice-only interaction may not be sufficient.

- For instance, when you are busy on phone and want the smart speaker to shut up (without saying it aloud)

Problem

For certain scenarios voice-only interaction may not be sufficient.

- For instance, when you are busy on phone and want the smart speaker to shut up (without saying it aloud)
- Perform a simple hand gesture (similar to stop sign) to shut it up.

Problem

For certain scenarios voice-only interaction may not be sufficient.

- For instance, when you are busy on phone and want the smart speaker to shut up (without saying it aloud)
- Perform a simple hand gesture (similar to stop sign) to shut it up.

No notification

- User needs to explicitly ask a smart speaker to give notifications.

Problem

For certain scenarios voice-only interaction may not be sufficient.

- For instance, when you are busy on phone and want the smart speaker to shut up (without saying it aloud)
- Perform a simple hand gesture (similar to stop sign) to shut it up.

No notification

- User needs to explicitly ask a smart speaker to give notifications.
- The smart speaker detects when a person entered the room, and starts proactive notifications.

Aim

Leverage the mic array in smart speakers for opportunistically sensing gestures

and

classifying and counting exercises

Aim

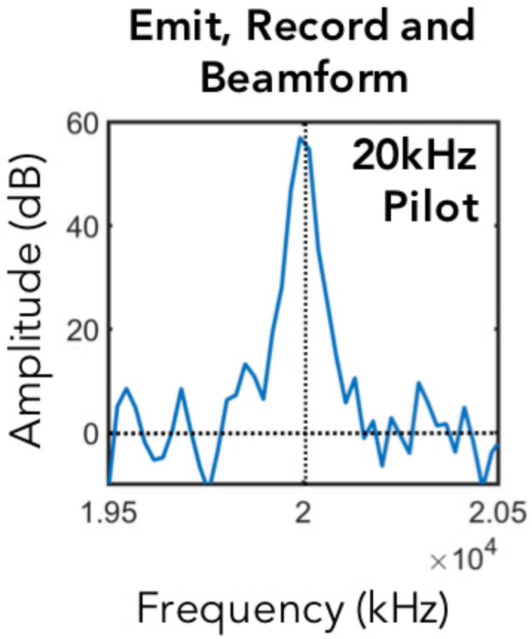
Leverage the mic array in smart speakers for opportunistically sensing gestures

and

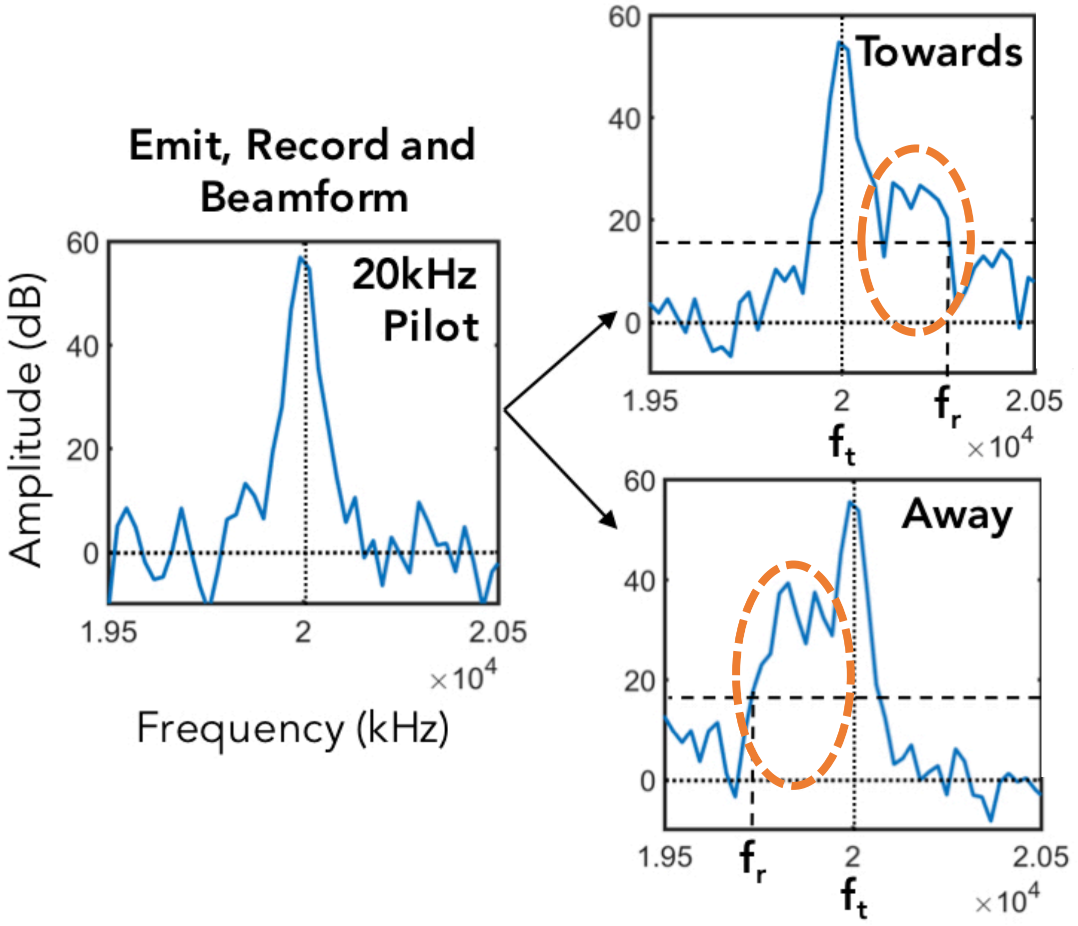
classifying and counting exercises

(without speaking aloud)

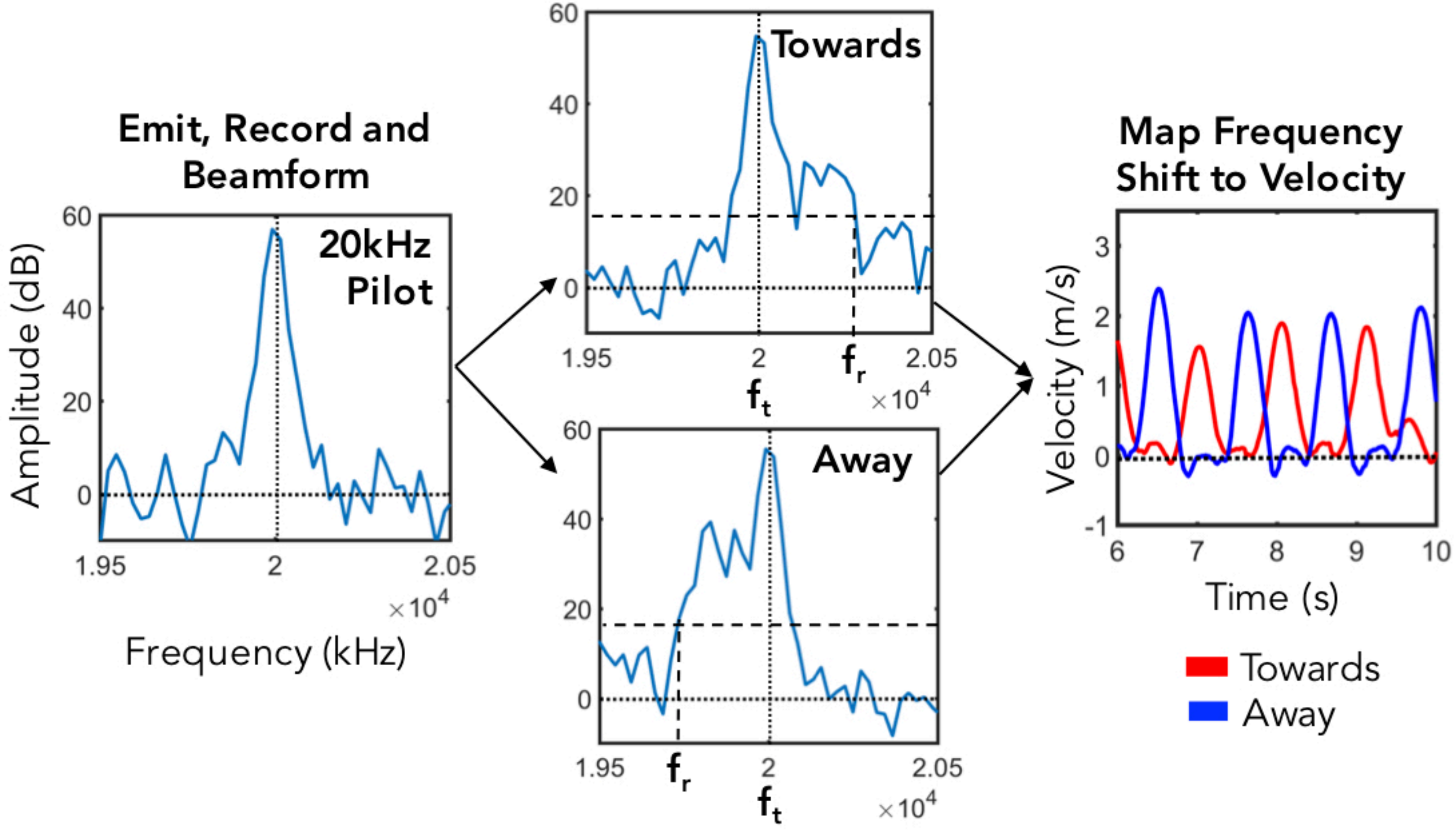
Doppler Shift



Doppler Shift



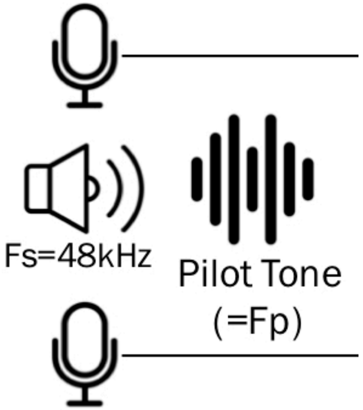
Doppler Shift



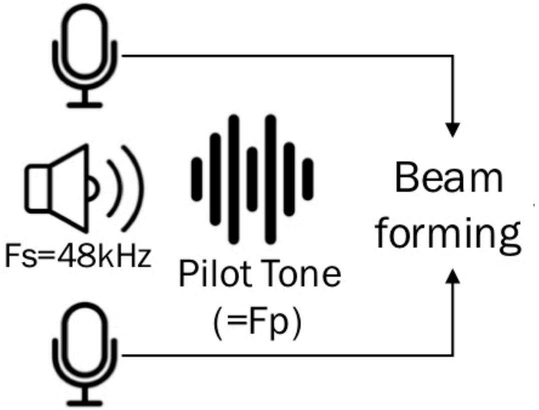
Emit, Record and Beamform

Map Frequency Shift to Velocity

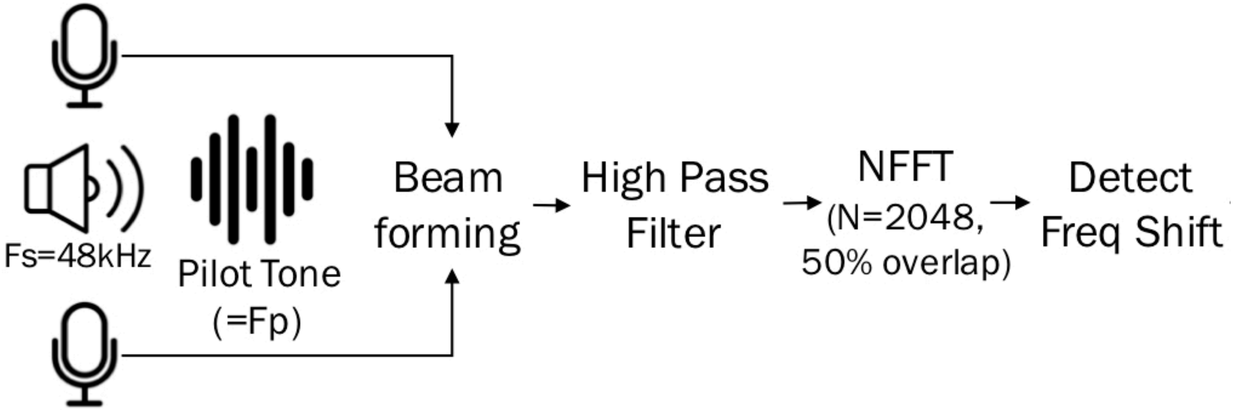
System



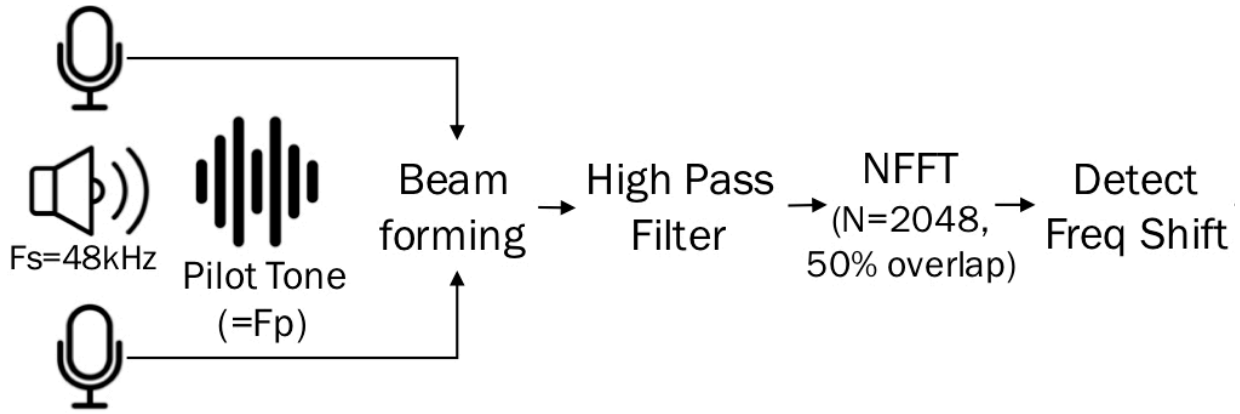
System



System


Fs=48kHz
Pilot Tone
(=Fp)
Beam
forming
High Pass
Filter
NFFT
(N=2048,
50% overlap)
Detect
Freq Shift

System



$$f_r = f_t * (c+v)/(c-v)$$

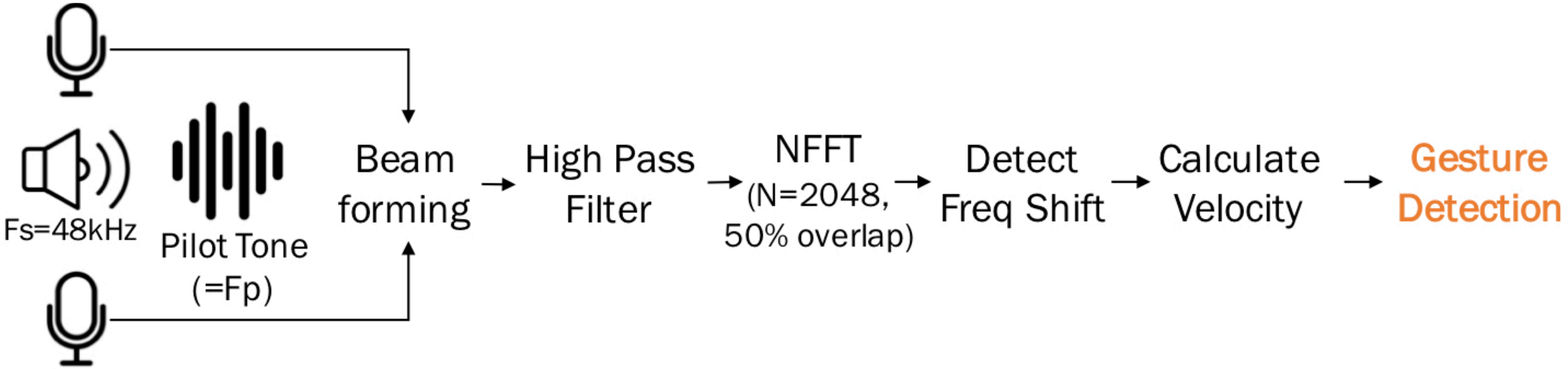
f_r = frequency recorded by mic {farthest from pilot in the interval [f_t-2, f_t+2] kHz above 5dB threshold}

f_t = pilot tone frequency

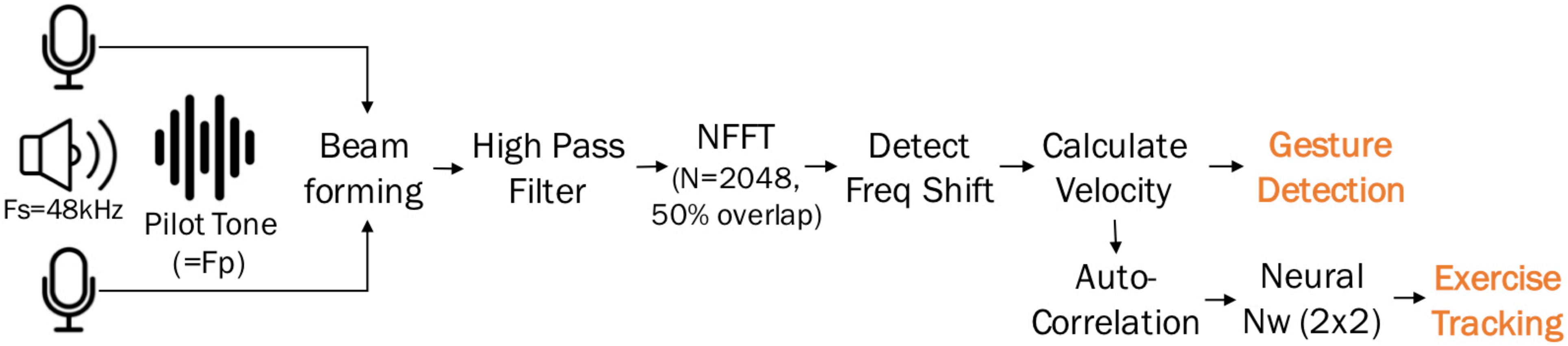
c = speed of sound in air

v = speed of body movement towards the mic

System


Fs=48kHz
Pilot Tone
(=Fp)
Beam
forming
High Pass
Filter
NFFT
(N=2048,
50% overlap)
Detect
Freq Shift
Calculate
Velocity
Gesture
Detection

System



Fs=48kHz

Pilot Tone (=Fp)

Beam forming → High Pass Filter → NFFT (N=2048, 50% overlap) → Detect Freq Shift → Calculate Velocity → **Gesture Detection**

Calculate Velocity → Auto-Correlation → Neural Nw (2x2) → **Exercise Tracking**

veloctiy-time curve
divided into 5s chunks
with 200 ms stride length

Hardware

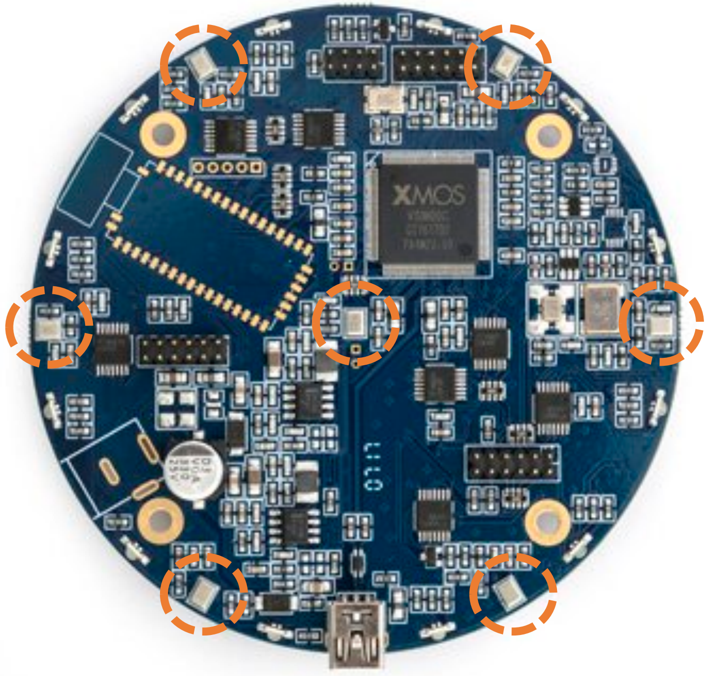
MiniDSP UMA-8 circular USB mic array

7 MEMS microphones

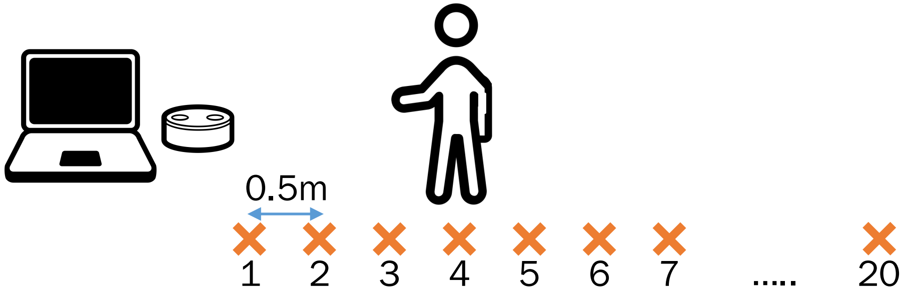
Radius 43 mm

Sampling rate 48kHz (Fs)

Capturing 24 bits per sample



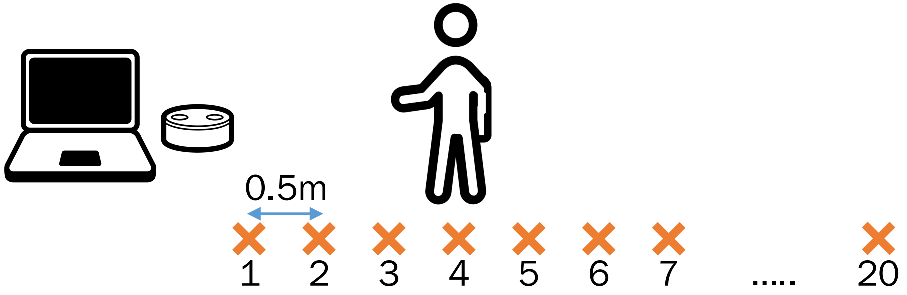
Data Collection: 1



Data Collection: 1

20 markers, 0.5 m away

Forward (pushing hand away from body)
Backward (pulling hand towards the body)



10 times at each marker

Two pilot tones: 20 kHz and 6kHz

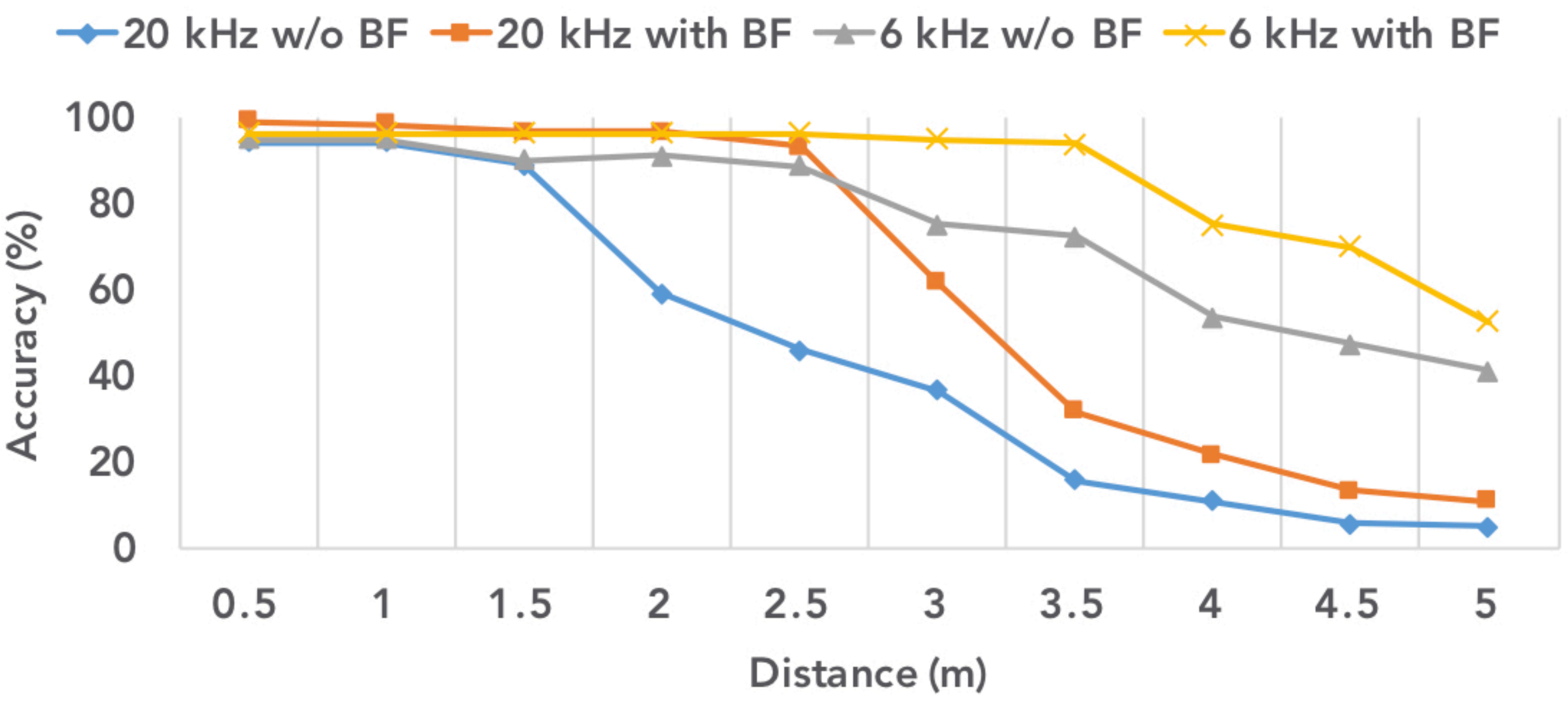
12 participants (10 male, 2 female)

Age = 22.4 ± 4.3 years

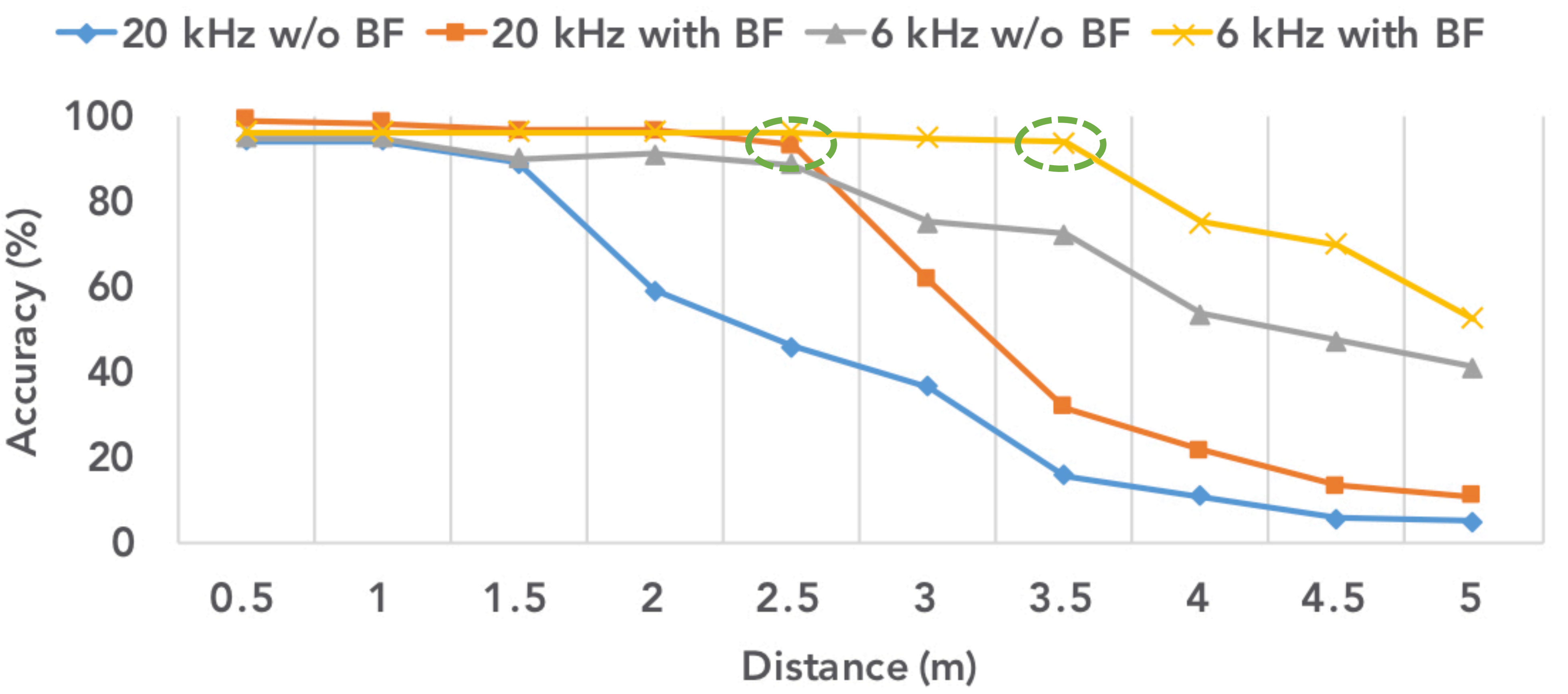
Weight = 73 ± 10.1 kgs

Height = 172.5 ± 8.7 cm

Results: 1


20 kHz w/o BF
20 kHz with BF
6 kHz w/o BF
6 kHz with BF
Accuracy (%)
100
80
60
40
20
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Distance (m)

Results: 1



20 kHz w/o BF
20 kHz with BF
6 kHz w/o BF
6 kHz with BF
100
80
60
40
20
0
Accuracy (%)
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Distance (m)

Data Collection: 2

1. Cross stretch

2. Curls

3,4. Folded shoulder rotation (C and AC)

5. On-spot jog

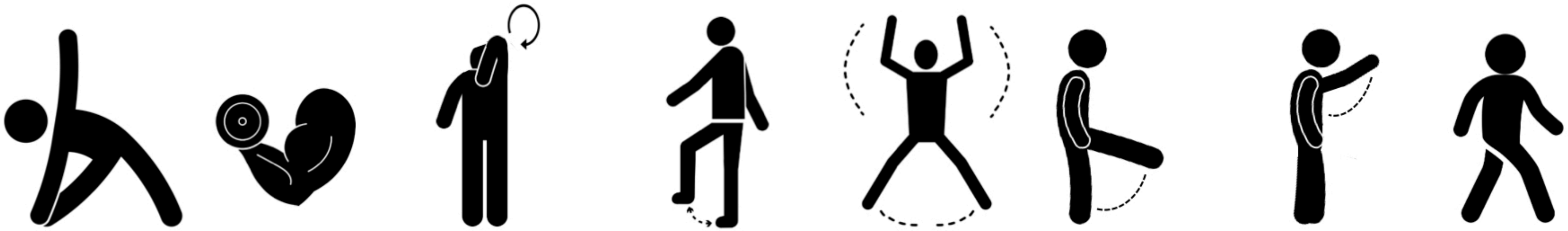
6. Jumping jacks

7. Leg raise

8,9. Shoulder rotation (C and AC)

10. Walk

Data Collection: 2



10 exercises, 20 repetitions each

2.5m from the device

20 kHz pilot tone

17 participants (15 male, 2 female)

Age = 26.4 ± 4.4 years

Weight = 73.6 ± 12.3 kgs

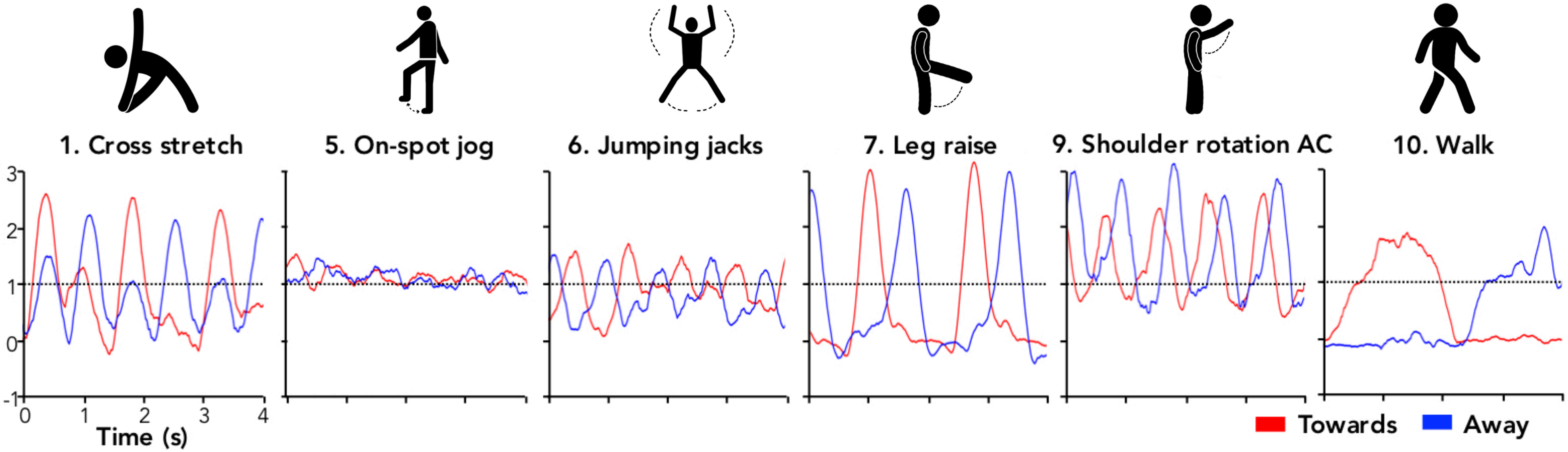
Height = 174 ± 9.6 cms

Average fitness = 3.4 ± 0.8

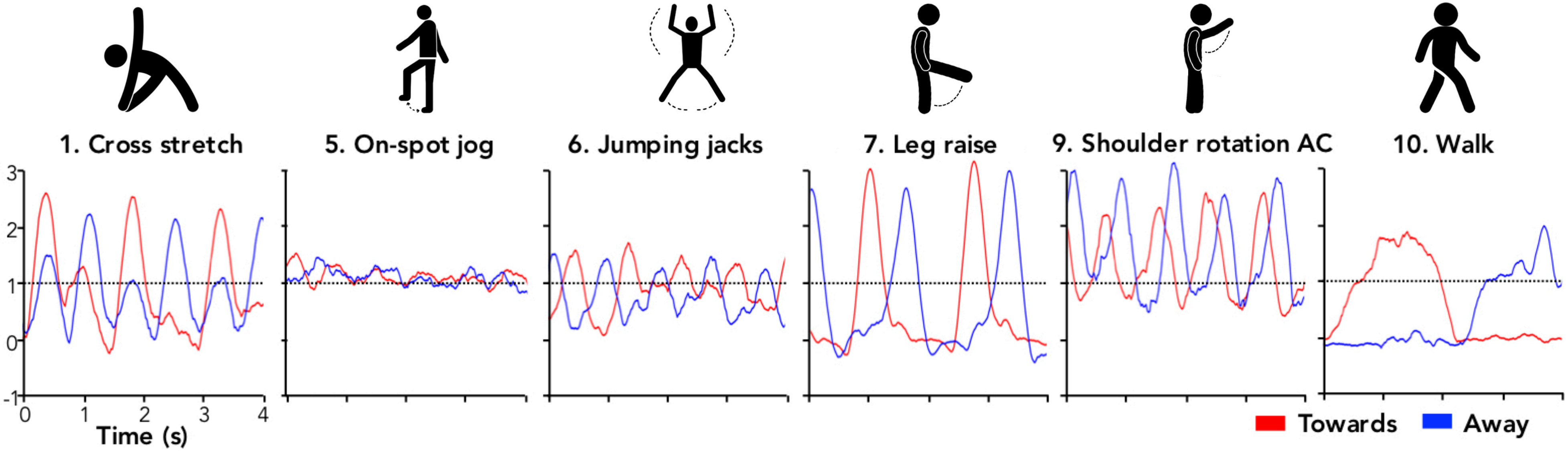
Daily exercise = 6/17

Exercise 2-3 times a week = 4/17

Results: Exercise Identification



Results: Exercise Identification



99.8% on the training set

95.9% on the evaluation set

Results: Confusion Matrix

	Predicted Label									
True Label	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Cross stretch	**0.99**	0	0	0	0	0	0	0	0	0
2. Curls	0	**0.99**	0	0	0	0	0	0.01	0	0
3. Folded shoulder rotation C	0	0	**0.93**	0.05	0	0	0	0	0.01	0
4. Folded shoulder rotation AC	0	0.01	0.04	**0.95**	0	0	0	0	0	0
5. On-spot jog	0	0.02	0.02	0.02	**0.87**	0	0.02	0	0	0.04
6. Jumping jacks	0	0	0	0	0	**0.97**	0	0.01	0	0
7. Leg raise	0	0	0	0	0	0	**0.98**	0.02	0	0
8. Shoulder rotation C	0	0.01	0	0	0	0	0.01	**0.97**	0	0.01
9. Shoulder rotation AC	0.01	0.01	0	0	0.01	0.01	0.01	0.04	**0.91**	0
10. Walk	0.02	0.01	0	0	0	0	0	0.02	0.02	**0.93**

1. Cross stretch

2. Curls

3,4. Folded shoulder rotation (C and AC)

5. On-spot jog

6. Jumping jacks

7. Leg raise

8,9. Shoulder rotation (C and AC)

10. Walk

Results: Confusion Matrix

Predicted Label

True Label

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Cross stretch	0.99	0	0	0	0	0	0	0	0	0
2. Curls	0	0.99	0	0	0	0	0	0.01	0	0
3. Folded shoulder rotation C	0	0	0.93	0.05	0	0	0	0	0.01	0
4. Folded shoulder rotation AC	0	0.01	0.04	0.95	0	0	0	0	0	0
5. On-spot jog	0	0.02	0.02	0.02	0.87	0	0.02	0	0	0.04
6. Jumping jacks	0	0	0	0	0	0.97	0	0.01	0	0
7. Leg raise	0	0	0	0	0	0	0.98	0.02	0	0
8. Shoulder rotation C	0	0.01	0	0	0	0	0.01	0.97	0	0.01
9. Shoulder rotation AC	0.01	0.01	0	0	0.01	0.01	0.01	0.04	0.91	0
10. Walk	0.02	0.01	0	0	0	0	0	0.02	0.02	0.93

1. Cross stretch

2. Curls

3,4. Folded shoulder rotation (C and AC)

5. On-spot jog

6. Jumping jacks

7. Leg raise

8,9. Shoulder rotation (C and AC)

10. Walk

Results: Exercise Counting

1. Cross stretch

2. Curls
X

3,4. Folded shoulder rotation (C and AC)

5. On-spot jog
X

6. Jumping jacks

7. Leg raise

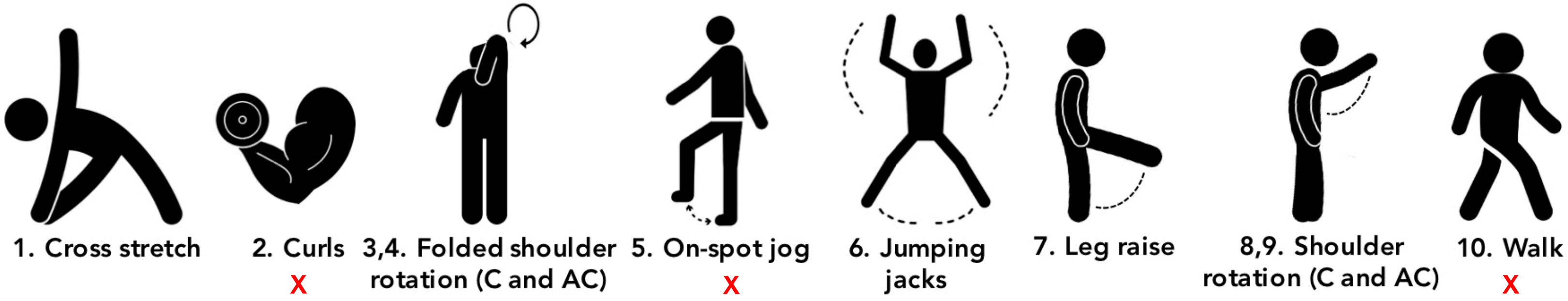
8,9. Shoulder rotation (C and AC)

10. Walk
X

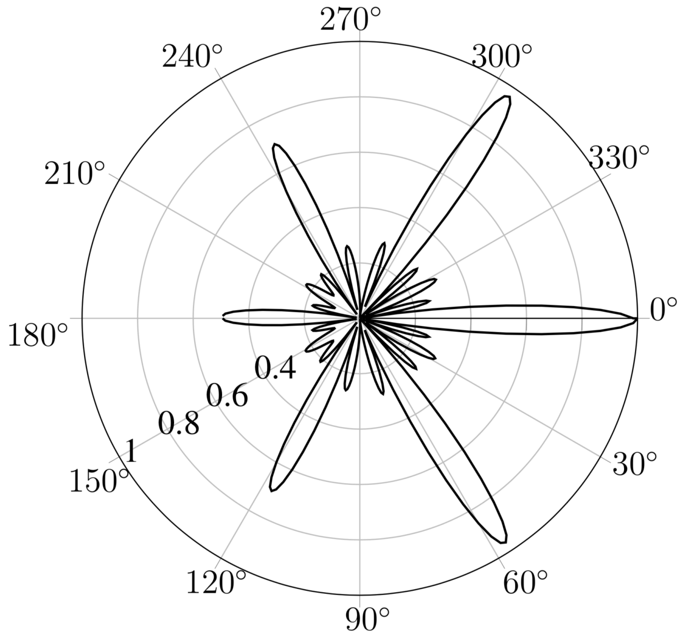
Results: Exercise Counting

	1. Cross Stretch	3. Folded Shoulder Rotation C	4. Folded Shoulder Rotation AC	6. Jumping Jacks	7. Leg Raise	8. Shoulder Rotation C	9. Shoulder Rotation AC
Accuracy (m)	85.7	91.3	94.7	86.7	97.0	95.0	92.2
sd	15.8	16.2	5.1	19.0	4.8	3.7	6.6

91.8% accuracy

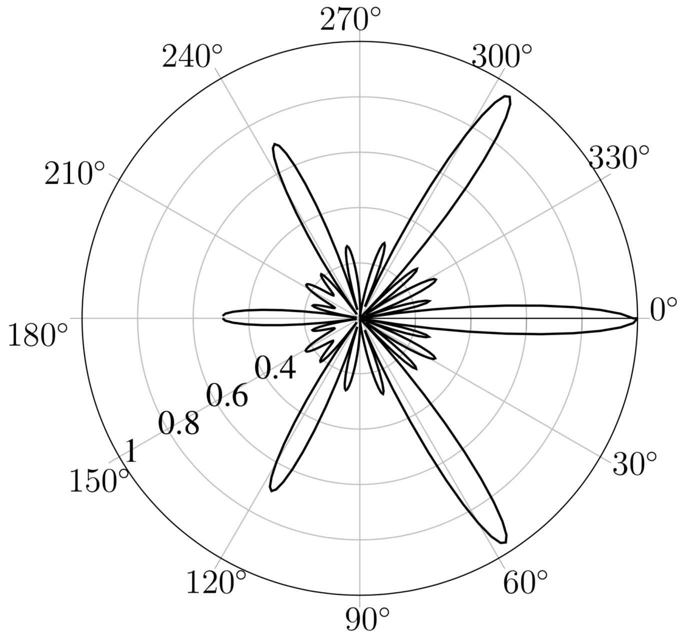


Limitations & Future Directions

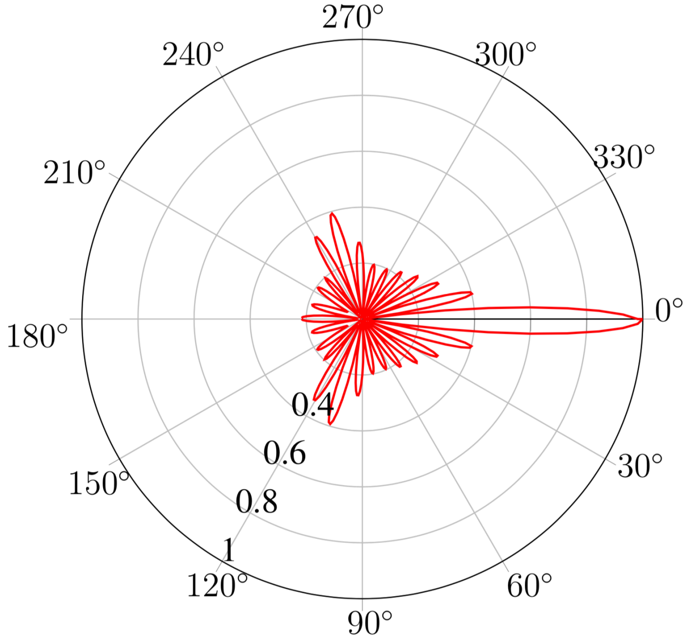


6 mics
43 mm radius

Limitations & Future Directions

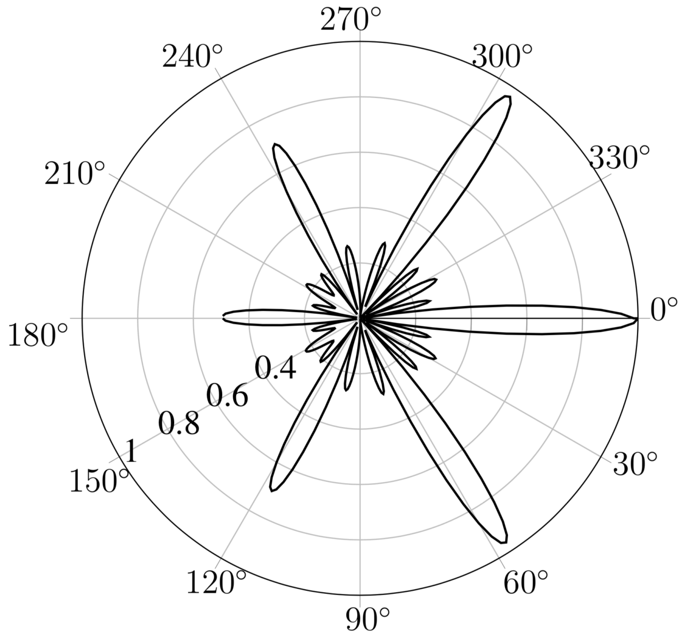


6 mics
43 mm radius

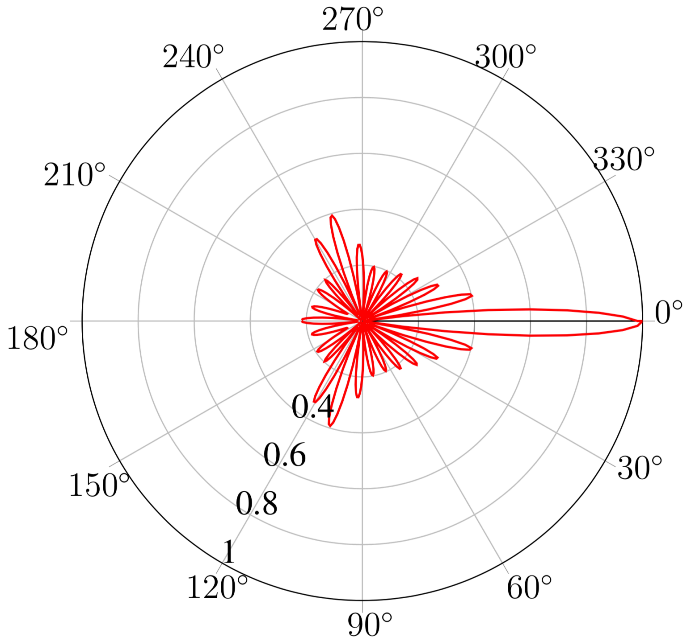


24 mics
43 mm radius

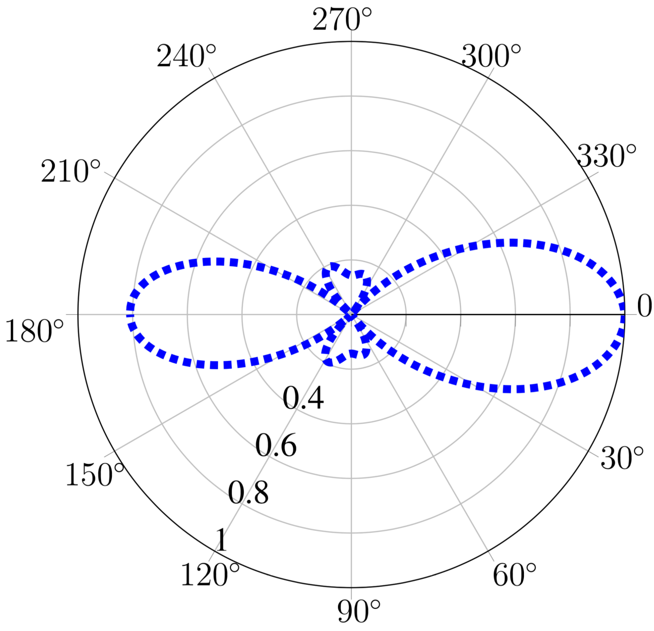
Limitations & Future Directions



6 mics
43 mm radius



24 mics
43 mm radius



6 mics
8.8 mm radius

Conclusion

Accurately tracking hand movement gestures (96.8%) from a distance of 2.5m

Classifying 10 exercises accurately (96%)

Counting 7 exercises accurately (91.8%)

Thank You!

Mohit Jain

IBM Research, India: **mohitjain@in.ibm.com**

University of Washington, Seattle USA: **mohitj@cs.washington.edu**