OPPORTUNISTIC SENSING WITH MIC ARRAYS ON SMART SPEAKERS FOR DISTAL INTERACTION AND EXERCISE TRACKING

Anup Agarwal [⋆†] *Mohit Jain* [⋆‡] *Pratyush Kumar* [⋆] *Shwetak Patel* [‡]

[⋆] IBM Research, India [†] IIT Guwahati, India [‡] University of Washington, Seattle, USA

ABSTRACT

In 2017, smart speakers (such as Amazon Echo, Google Home, *etc.*) became a commercial success. Most smart speakers have a circular microphone array to provide hands-free, voice-only interaction from a distance. In this work, we exploit this mic array for opportunistically sensing gestures and tracking exercises. To this end, we measure the Doppler shift on a pilot tone caused by a gesturing human body, and use beamforming of the mic array to extend the range of the detection. Data from 12 participants show that gestures can be detected with an accuracy of 96.8% up to a distance of 2.5 meters using an inaudible 20 kHz pilot tone. For exercise tracking, we train a deep neural network to recognize 10 different exercises, and count repetitions by peak-finding heuristics. Data from 17 participants show that exercise classification accuracy is 96% and count accuracy is 91.8%. To conclude, we discuss hardware enhancements to smart speakers to further increase their gesture sensing capabilities.

***Index Terms*—** Microphone array, in-air gesture sensing, smart speakers, exercise detection, exercise counting

1. INTRODUCTION

Voice-enabled speakers, also known as *smart speakers*, have integrated virtual assistants that offer hands-free, voice-only interaction. In 2017, 35.6M smart speakers were sold in the US, 129% more than 2016 [1]. Major corporations, including Apple, Amazon, and Google, are competing in this space with respective offerings. Most smart speakers consist of a circular array of microphones, *e.g.*, Amazon Echo has a 7-mic array, and Apple Homepod and Sonos One have a 6-mic array [2]. The mic array increases a device's range for recognising voice commands from across the room using beamforming, noise reduction, and acoustic echo cancellation [2, 3].

In this work, we propose and design a prototype to exploit the mic array of smart speakers for opportunistic gesture sensing. We use the beamforming capabilities of the mic array to enable gesture-based distal interaction and exercise tracking. Previously, Soundwave [4] demonstrated proximal gesture-based interactions (up to 1 meter) on a laptop's single microphone. They measured the Doppler shift caused by a gesturing hand on an inaudible pilot tone. We build upon that work and use the mic array in smart speakers to achieve gesture-based interactions up to 2.5 meters from the device. The mic array helps extend the range as we beamform to increase the SNR in the direction of the gesturing human. Data from 12 participants shows that gestures can be detected with an accuracy of 96.8% up to a distance of 2.5 meters (using an inaudible 20kHz pilot tone) and with an accuracy of 95.7% up to 3.5 meters (using an audible 6kHz tone).

We observed that the accuracy of the gesture detection is high enough to recognize exercises performed at a distance and to count repetitions. With a deep learning classifier, we are able to differentiate between 10 commonly performed exercises using a pilot tone of 20kHz. Data from 17 participants performing the 10 exercises at a distance of 2.5 meters from the device shows an accuracy of 96% in exercise recognition; the data also shows an accuracy of 91.8% in counting the exercise repetitions using peak finding heuristics. To the best of our knowledge, this work is the first to demonstrate distal interaction with smart speakers enabled by beamforming. We also recommend hardware configurations of the mic-array to further increase the range of the detection.

2. RELATED WORK

We review the prior work in two related areas of interaction: the use of sound waves and exercise detection.

Interaction using Sound Waves: There has been recent interest in utilizing sound waves for gesture-based interaction in commodity devices. Soundwave [4] measures Doppler shift on inaudible tones to recognize gestures in the near-field range of 1 meter from the device. FingerIO [5] uses OFDM-modulated sound to track fingers with an accuracy of 0.8 cm. CovertBand [6] uses similar OFDM-modulated sound for tracking multiple individuals' locations and their activities. Wang *et al.* designed a LLAP [7] scheme exploiting phase changes in sound as it reflects from moving hand/finger and achieved a distance estimation error of 0.7 cm for 1D (1-dimensional) tracking. Strata [8] further reduced the 1D tracking error to 0.3 cm, by estimating the channel impulse response in the time-domain. Most of the prior art explores the possibility of enhancing the granularity and accuracy of the gestures recognized, and are limited to a distance range in close proximity $\sim$1 meter from the device.

Exercise Tracking: Most prior art for tracking exercises [9,

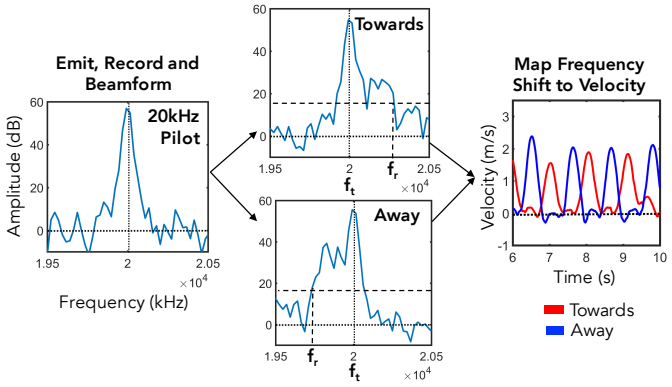

Fig. 1: Signal processing pipeline to transform the audio signal received at 7 channels to a velocity-vs-time plot.

10, 11] use inertial sensors on the body. myHealthAssistant [10] and Muehlbauer *et al.* [11] use an on-body smartphone to recognize the wearer's exercises. RecoFit [9] achieved higher accuracy (compared to [10, 11]) using an arm-worn inertial sensor to track repetitive exercises such as weight training and calisthenics. In this work, we explore gesture detection and exercise tracking from a distance with hardware similar to off-the-shelf smart speakers. To the best of our knowledge, this has not been previously studied.

3. SYSTEM DESIGN

3.1. Hardware and Principle

We use the MiniDSP UMA-8 circular mic array [12], as it best mimics today's off-the-shelf smart speakers while providing access to the raw audio signals. Similar to Amazon Echo devices, it consists of 7 (N) MEMS microphones, with 1 mic at the center and 6 mics uniformly spaced at the circumference of a circle with radius 43 mm (r). A commodity laptop was used to emit a continuous, inaudible pilot tone. N channel audio samples were recorded from the mic array at a sampling rate of 48kHz (Fs), capturing 24 bits per sample. The pilot tone from the laptop was reflected by nearby objects, including the user's body. The reflected signals received at the N audio channels were processed to estimate the speed and direction of any moving entity (Figure 1).

Beamforming: We use the standard Delay-and-Sum beamforming technique [13]. We calculate the time delay-of-arrival of the sound received at the different microphones relative to the sound arriving at the center microphone and then superimpose time-shifted variants of the N signals. This enhances the SNR of the signal received in the direction of the beamforming, which is set to the direction of the user.

Doppler Shift: The apparent frequency shift when there is relative motion between the source and the receiver is well-known as the Doppler effect. This shift is proportional to the relative velocity between the source and the receiver. In our system, both the speakers and the microphones are stationary while the user's body is the moving reflector. The velocity of this moving reflector can be computed as:

$$f_r = f_t * \left(\frac{c+v}{c-v}\right) \quad (1)$$

where f_r is the frequency recorded by the mic, f_t is the pilot frequency emitted by the speakers, c is the speed of sound in air, and v is the speed of the body/hand towards the mics.

Pilot frequency: The pilot tone emitted by the speaker must be inaudible so as to not interfere with the normal operation of the device. We use a 20kHz pilot tone in our experiments. However, the hardware we use is not designed to beamform for such a high frequency. Specifically, the separation between the microphones is 43mm, which is much higher than the separation as per spatial sampling theory of < 8.8mm ($=\lambda/2$) [13]. To compare this with a hypothetically closely spaced mic-array, we also use a 6kHz pilot tone.

3.2. Algorithm and Implementation Details

1. Beamforming and FFT: The $N = 7$ channel audio from the mic array is beamformed with the Delay-and-Sum technique. The resultant signal is passed through a high pass filter to extract frequencies in a 3.5kHz range around the pilot frequency of 20kHz. Then, a $N_{FFT} = 2048$ point FFT was computed to obtain a 1025-point frequency spectrum.

2. Computing f_r: The received frequency f_r (in Equation (1)) is set as the frequency farthest from f_t in the interval $[f_t - 1, f_t + 1]$ kHz with a magnitude above a *threshold value*. This threshold value is set to be 5dB plus the maximum magnitude of the signal outside the search interval, *i.e.*, in $[f_t - 3.5, f_t - 1]$ kHz and $[f_t + 1, f_t + 3.5]$ kHz. The value of 5dB was experimentally calibrated to avoid falsely noting noisy perturbations as Doppler frequency shifts.

3. Computing v: From f_t and the computed f_r, we compute the velocity from Equation (1). Since only one velocity value is calculated for each 1024 sized sample window (2048-point NFFT with 50% overlap), the time resolution for the velocity-vs-time curve was time_res $= (N_{FFT} * \text{overlap})/F_s \approx 21.3$ ms. Further, a frequency resolution of 23.437Hz for a 2048 point FFT with a 48kHz sampling rate and a 20kHz pilot tone yields a speed resolution of 20.62 cm/s.

4. Exercise classification: As exercise is usually repetitive, we leverage the autocorrelation property in the power spectrum to estimate the periodicity of the velocity-vs-time curve, similar to [9]. The v-t curve was divided into chunks of $T = 5s$ with a stride length of $S = 200ms$. For each chunk, we computed auto-correlation features with lags ranging from 0-100 and the first 100 FFT spectrum bins in discrete frequency steps of $\left(\frac{\lfloor (F_s/N_{FFT} * \text{overlap}) \rfloor}{\lceil T/\text{time\_res} \rceil}\right) = 0.199$ Hz from 0-19.74Hz. With these 201 computed signal properties as features, we trained a dense 2-layer neural network to classify the observed pattern into one of 10 exercises We used ReLu (rectifier linear unit) for non-linear activation and the categorical cross-entropy loss function for training the network.

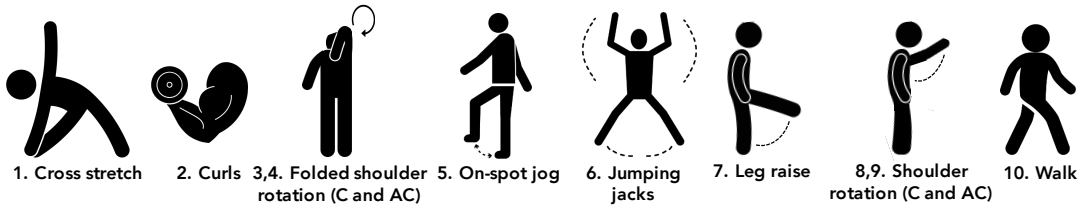

Fig. 2: The ten exercises performed by the participants. (C: Clockwise, AC: Anti-clockwise)

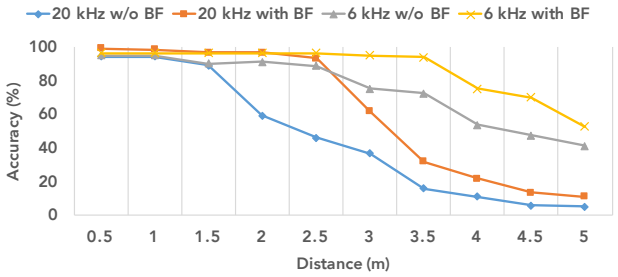

Fig. 3: Distal gesture counting accuracy at different distances from the device under four conditions. (BF: Beamforming)

5. Exercise counting: We count exercise repetitions by identifying repeating peaks in the velocity-vs-time graph. From experiments we see that a single repetition can correspond to multiple peaks with varying ampltitudes and shapes (Figure 4). Thus, we needed to apply both min-peak-prominence and min-peak-distance filters to isolate the peaks. The min-peak-prominence was set experimentally per exercise, while min-peak-distance was set to 0.2s uniformly.

4. EVALUATING GESTURE DETECTION

Setup: The MiniDSP device connected to a laptop was placed at one end of a long aisle in an office space. Twenty markers separated by 0.5 m were placed on the aisle floor. Participants were asked to stand at the first marker with their toes touching the marker. They were instructed to perform the 'forward' (pushing hand away from the body) and 'backward' (pulling hand towards the body) hand gestures at each marker 10 times. The participants were not allowed to take breaks. The experiment was repeated with two pilot tones of 20 kHz and 6 kHz. We also experimented with an audible 6 kHz tone to understand the impact of lower frequency on the range of distal interaction. Twelve participants (10 male and 2 female of age 22.4±4.3 years) were recruited. Their average weight was 73±10.1 kgs and average height was 172.5±8.7 cms. Participants took ∼5 minutes to complete this task.

Results: The signal processing pipeline described in the previous section was employed to compute the velocity-vs-time plot. The gesture count was estimated by peak finding heuristics with the value of min-peak-prominence set at 35 cm/s (experimentally determined). The accuracy was computed as the percentage of the absolute error w.r.t. the ground truth of 10 counts per participant per marker. We consider 4 conditions: for two pilot tones 20kHz and 6kHz we report results with and without beamforming (using audio from the central mic only). The accuracy as a function of the user's distance from the device for the 4 conditions is plotted in Figure 3. With the inaudible pilot tone of 20kHz, using the mic array to beamform increased the detection range with high accuracy (>95%) from 1m to a significant 2.5m. Also using an audible pilot tone of 6kHz increases the corresponding range to 3.5m, suitable for across-the-room uses. At a distance of 3.5m, the 20kHz tone had an accuracy of 96.8±2.2% at 2.5m, while it was 95.7±0.9% for the 6kHz tone.

5. EVALUATING EXERCISE TRACKING

Setup: The MiniDSP device connected with a laptop was placed on a table (height 29 inches). A marker was placed at a distance of 2.5 meters from the device. Participants were instructed to stand over the marker and perform 10 exercises (Figure 2) – *Cross stretch* (right hand touching left toe and vice-versa), *Curls* (both hands together with no weight), *Folded shoulder rotation* (Clockwise and Anti-Clockwise), *On-spot jog*, *Jumping jacks*, *Leg raise* (in the front), *Shoulder rotation* (Clockwise and Anti-Clockwise), and *Walk* – 20 repetitions each, in a random order. For *walk*, participants were asked to walk five steps away and then five steps towards the device, repeated twice for 20 steps. The experiment was run by the participant, without any external intervention. The order of the exercises was written on a white board. Participants were asked to press the 'S' button on the laptop to start the recording, go back to the marker and perform the exercise. Once one set of exercise was completed, participants needed to walk to the laptop and press the 'Space' button to stop recording. One of the researchers stood outside the room observing the session through a glass door. A pilot tone of 20kHz was used. Participants were free to take breaks. Each exercise session lasted for around 30 minutes. Seventeen participants (15 male and 2 female of age 26.4±4.4 years) were recruited. Their average weight was 73.6±12.3 kgs and height was 174 cms±9.6 cms. On a 5-point Likert scale for fitness with 5 being very fit, the average fitness of the participants was 3.4±0.8. Six reported exercising 4-5 times a week, four reported exercising 2-3 times a week, and the remaining did not regularly exercise.

Results - Recognition Accuracy: From the velocity-vs-time graph (examples in Figure 4), the chunks and features for classifying the exercise were computed as described in Section 3. For training the neural network, the dataset was ran-

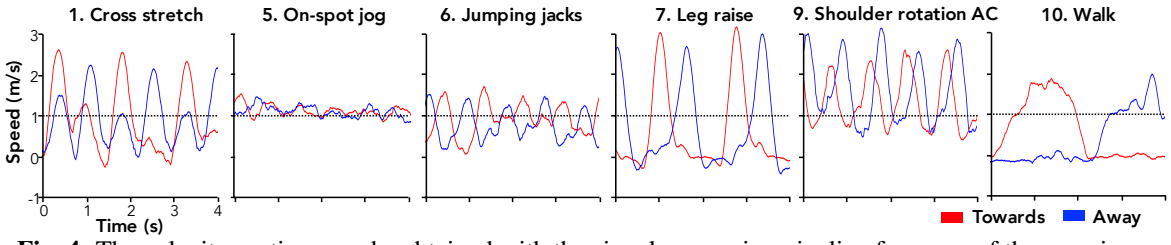

Fig. 4: The velocity-vs-time graphs obtained with the signal processing pipeline for some of the exercises.

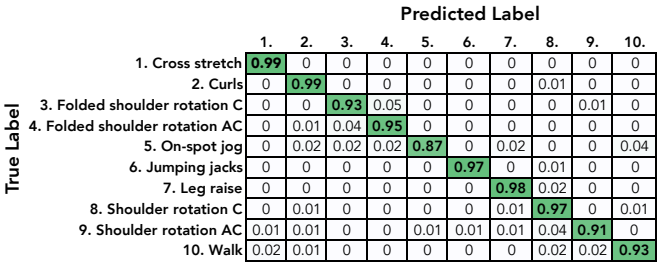

Fig. 5: Normalized confusion matrix for classifying exercises

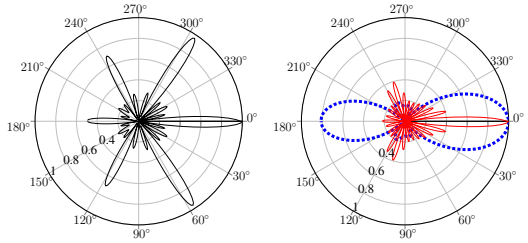

Fig. 6: Beamforming pattern for 20kHz with (left) 43 mm radius, 6 mics, and (right) 43 mm radius, 24 mics in solid red line, and 8.8 mm radius, 6 mics in dashed blue line.

domly shuffled, and 70% of the dataset was used for training, and the remaining 30% was used as the test set. After training, the accuracy the network classified exercises with an accuracy of 99.8% on the training set, and 95.9% on the evaluation set. Figure 5 shows the confusion matrix of the classification for the 10 exercises. On-spot jog had the lowest accuracy (87.3%), as it was confused with walking in 4.4% of the instances. All other exercises had a recognition accuracy above 90%. As an existing benchmark, RecoFit [9] reported a recognition accuracy of 96% recognition accuracy for 13 exercises using data from wearable sensors.

Exercise no.	1	3	4	6	7	8	9
Accuracy (m sd)	85.7 15.8	91.3 16.2	94.7 5	86.7 19	97 4.8	95 3.7	92.2 6.6

Results - Counting Accuracy: Counting was performed with peak detection heuristics with experimentally determined min-peak-prominence values for each exercise. For curls, on-spot jog, and walk, we were unable to determine strong peaks (see example of on-spot jog in Figure 4). For the remaining 7 exercises we report the accuracy in percentage terms w.r.t. the ground truth of 20 repetitions in the table above. Overall counting accuracy was 91.8±12%, with cross stretch 85.7±15.8% and jumping jacks 86.7±18.9% being the lowest. Remaining 5 exercises had an accuracy $> 91\%$.

6. DISCUSSION AND CONCLUSION

With experimental measurements, we have demonstrated that a smart speaker emitting an inaudible pilot tone from a distance of 2.5m can detect gestures with an accuracy of 96.8%, classify exercises with an accuracy of 96%, and count exercise repetitions with an accuracy of 91.8%. The relatively large distance and high accuracy is enabled by opportunistically using the mic arrays present on the smart speakers. However there are two major limitations of our work: (a) a relatively small sample size of 17 participants with 10 different exercises, and (b) applicability of Doppler analysis only for specific angles between user and mic/speaker (fig. 6 left).

Moreover, for truly across-the-room interaction, the supported range of interaction should be about 5m. There are two fundamental ways to enhance the range. Firstly, in *software*, instead of Doppler shift sensing, we can employ methods such as Frequency-Modulated Continuous Wave and Orthogonal Frequency-Division Multiplexing. We plan to report comparative results in future work. Secondly, in *hardware*, the mic array can be optimized, as beamforming for 20kHz with mic separation of 43mm is not effective (Figure 6, left). At the same radius of 43mm, increasing to 24 mics improves the beamforming pattern (Figure 6, right, red solid lines) but at prohibitive hardware and signal processing cost. For the same number of mics, reducing the separation to 12.9mm will roughly improve the performance of the 20kHz tone to that of the 6kHz tone, which according to our experimental results would increase the range from 2.5m to about 3.5m, which is still short of the 5m target. As per spatial sampling theory [13], the separation between mics should not be more than 8.8mm for 20kHz (Figure 6, right, blue dashed lines) for a 5m range. But such a small separation adversely impacts beamforming for the audible range, affecting the core functionality.

Considering trade-offs between number of mics, separation, and beamforming for audible and 20kHz range, we propose a hybrid setup. In this setup, there are two mic arrays: 6 mics in a circular array of 43mm for the audible range for voice interaction, and 6 mics in a circular array of 8.8mm for the inaudible pilot tone for gesture tracking. With these software and hardware modifications, our results show that we can unlock novel across-room interaction modalities and applications with the already popular smart speakers.

7. REFERENCES

[1] Jeff Dunn, "Amazon's echo isn't going to give up its lead anytime soon," http://www.businessinsider.in/Amazons-Echo-isnt-going-to-give-up-its-lead-anytime-soon/articleshow/58602433.cms.

[2] Shannon Liao and Chaim Gartenberg, "Google home max vs. homepod and google home mini vs. amazon echo dot: battle of the smart speakers," https://www.theverge.com/2017/10/5/16425142/google-home-mini-vs-amazon-echo-dot-max-apple-homepod.

[3] Amazon.com, "Amazon alexa 7-mic far-field dev kit," https://developer.amazon.com/alexa-voice-service/dev-kits/amazon-7-mic.

[4] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan, "Soundwave: Using the doppler effect to sense gestures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2012, CHI '12, pp. 1911–1914, ACM.

[5] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota, "Fingerio: Using active sonar for fine-grained finger tracking," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2016, CHI '16, pp. 1515–1525, ACM.

[6] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota, "Covertband: Activity information leakage using music," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 87:1–87:24, Sept. 2017.

[7] Wei Wang, Alex X. Liu, and Ke Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2016, MobiCom '16, pp. 82–94, ACM.

[8] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2017, MobiSys '17, pp. 15–28, ACM.

[9] Dan Morris, Scott Saponas, Andrew Guillory, and Ilya Kelner, "Recofit: Using a wearable sensor to find, recognize, and count repetitive exercises," April 2014, pp. 3225–3234, ACM.

[10] Christian Seeger, Alejandro Buchmann, and Kristof Van Laerhoven, "myhealthassistant: A phone-based body sensor network that captures the wearer's exercises throughout the day," in *Proceedings of the 6th International Conference on Body Area Networks*, ICST, Brussels, Belgium, Belgium, 2011, BodyNets '11, pp. 1–7, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[11] M. Muehlbauer, G. Bahle, and P. Lukowicz, "What can an arm holster worn smart phone do for activity recognition?," in *2011 15th Annual International Symposium on Wearable Computers*, June 2011, pp. 79–82.

[12] MiniDSP, "Uma-8 usb mic array," https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array.

[13] Harry L. Van Trees, *Detection, Estimation, and Modulation Theory: Radar-Sonar Signal Processing and Gaussian Signals in Noise*, Krieger Publishing Co., Inc., Melbourne, FL, USA, 1992.