LASense: Pushing the Limits of Fine-grained Activity Sensing Using Acoustic Signals

DONG LI, University of Massachusetts Amherst, USA
JIALIN LIU, University of Massachusetts Amherst, USA
SUNGHOON IVAN LEE, University of Massachusetts Amherst, USA
JIE XIONG, University of Massachusetts Amherst, USA

Acoustic signals have been widely adopted in sensing fine-grained human activities, including respiration monitoring, finger tracking, eye blink detection, etc. One major challenge for acoustic sensing is the extremely limited sensing range, which becomes even more severe when sensing fine-grained activities. Different from the prior efforts that adopt multiple microphones and/or advanced deep learning techniques for long sensing range, we propose a system called LASense, which can significantly increase the sensing range for fine-grained human activities using a single pair of speaker and microphone. To achieve this, LASense introduces a virtual transceiver idea that purely leverages delicate signal processing techniques in software. To demonstrate the effectiveness of LASense, we apply the proposed approach to three fine-grained human activities, i.e., respiration, finger tapping and eye blink. For respiration monitoring, we significantly increase the sensing range from the state-of-the-art 2 *m* to 6 *m*. For finer-grained finger tapping and eye blink detection, we increase the state-of-the-art sensing range by 150% and 80%, respectively.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: long-range acoustic sensing, fine-grained activity sensing, contact-free sensing

ACM Reference Format:
Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. LASense: Pushing the Limits of Fine-grained Activity Sensing Using Acoustic Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 21 (March 2022), 27 pages. https://doi.org/10.1145/3517253

1 INTRODUCTION

The great success in utilizing acoustic signals for contact-free human sensing has facilitated numerous applications on widely available acoustic-enabled devices, including smartwatches, smartphones, and smart speakers. Beyond simple audio playing and voice-based control, acoustic sensing extends the capabilities of these devices to support applications such as coarse-grained human tracking [18, 25] and hand gesture recognition [9, 19, 29, 41], as well as fine-grained finger tracking/tapping [24, 32, 40, 43] and respiration/heartbeat monitoring [28, 38, 39, 46]. The extended capabilities not only ease the interactions between the user and device, but also enable sensing rich context information of human targets.

Authors' addresses: Dong Li, University of Massachusetts Amherst, Massachusetts, USA, dli@cs.umass.edu; Jialin Liu, University of Massachusetts Amherst, Massachusetts, USA, jialinliu@umass.edu; Sunghoon Ivan Lee, University of Massachusetts Amherst, Massachusetts, USA, silee@cs.umass.edu; Jie Xiong, University of Massachusetts Amherst, Massachusetts, USA, jxiong@cs.umass.edu.

2474-9567/2022/3-ART21 $15.00
https://doi.org/10.1145/3517253

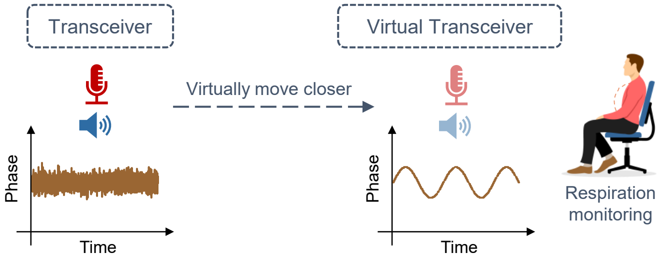


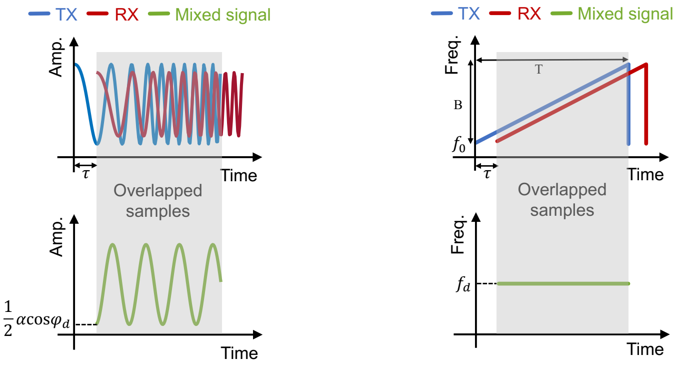
Fig. 1. LASense aims at sensing the fine-grained activities (e.g., respiration monitoring) at distance using a single pair of speaker and microphone, which can be applied to all acoustic-enabled devices.

Although acoustic sensing has extremely fine-grained sensing granularity, the sensing range is very limited, hindering its wide adoption in real life. One representative example is that the sensing range of the prior acoustic-based respiration monitoring systems is approximately 2 *m* [38, 39], which is insufficient for room-scale coverage. The main reason for the small sensing range lies in the essence of contact-free sensing in which the reflection signal adopted for sensing is much weaker than that of the direct path signal. Furthermore, when acoustic signals are used for sensing finer-grained activities (e.g., finger tapping and eye blink), this issue is even more severe mainly due to the much smaller reflection area of the target (e.g., fingers and eyeballs). The state-of-the-art sensing ranges for finger tapping and eye blink detection are merely 40 *cm* [32] and 50 *cm* [20], respectively.

Recent efforts have been devoted to alleviating the above-mentioned sensing range issue [3, 15, 21, 22, 38]. These solutions require a microphone array [3, 15, 22, 38] and/or rely on advanced deep learning techniques [21, 22]. However, the number of microphones in a smartphone is usually two, and this number further decreases to one for most smartwatches. On the other hand, deep learning techniques can be resource-hungry and thus may not be suitable for battery-based wearable devices such as smartwatches. In this paper, we propose a general solution LASense that can be implemented on acoustic-enabled devices to increase the sensing range for fine-grained human activities using a single pair of speaker and microphone. At a high level, LASense extracts the signal variations induced by fine-grained activities and utilizes the variations to sense the corresponding activities. To capture subtle signal variations, LASense transmits inaudible chirp signals using a single speaker and analyzes the signals reflected from the target at the microphone. The key insight of our design to increase the sensing range is that—instead of multiplying the transmitted signal with the received reflection signal, which has been the conventional way to extract signal variations from the chirp signal—we propose to multiply a carefully shifted version of the transmitted signal with the received signal to amplify the signal variations caused by the activity, hence improving the sensing performance. This shifted version of transmitted signal is equivalent to virtually moving the transceiver closer to the target, as shown in Fig. 1.

Though promising, several challenges need to be addressed before the idea can be translated into a functional system. The first challenge is to increase the sensing range using only a single pair of speaker and microphone. Previous studies adopt a microphone array (4-7 microphones) for beamforming to achieve a stronger signal and accordingly a larger sensing range [3, 15, 22, 38]. It is particularly challenging to increase the sensing range using just one microphone-speaker pair without additional information from other microphones.

To address this challenge, through both theoretical analysis and experimental verification, we show that the signal variations induced by fine-grained activities are not only dependent on the signal-to-noise-ratio (SNR) of the received signal but also the number of samples overlapped between the transmitted and the received signals, as shown in Fig. 2. Instead of increasing the SNR of the received signals with beamforming using a microphone array, we propose to increase the number of overlapped samples between the transmitted signal and



(a) Amplitude over time.

(b) Frequency over time.

Fig. 2. The chirp signal representation for the transmitted signal (TX), received signal (RX) and mixed signal.

received signal by shifting the transmitted signal in the time domain through signal processing in software. It is noteworthy that increasing the number of overlapped samples is equivalent to virtually moving the transceiver closer to the target. We demonstrate that this scheme can effectively amplify the signal variations induced by fine-grained activities, improving both sensing accuracy and range.

The second challenge comes in naturally: How should we efficiently move the virtual transceiver closer to the target without knowing the target's physical location? Estimating the physical location of the target when it is relatively close to the transceiver (e.g., 20 *cm* for finger tapping) can be achieved easily. The chirp signal has the intrinsic capability of separating signals reflected from multiple targets into different range bins. The approximate location of the target can be estimated by identifying the target bin (i.e., the range bin where the target locates) with large signal variations. However, when the target is far away from the transceiver (e.g., 80 *cm* for finger tapping), the signal variations become very small, making it difficult to identify the target bin from the other bins due to noise.

To address the second challenge, we propose a solution to detect the target bin even when the target is far away. The solution is based on the following key insight: the signal variations become much larger when the virtual transceiver is moved closer to the bins with targets, and they do not change much when the virtual transceiver is moved closer to the bins without a target. However, both static objects (e.g., furniture) and human target can lead to increased signal variations. We further differentiate the human target from static objects following an interesting observation: the large signal variation for the static object is caused by noise. In the I-Q domain, its signal amplitude variation is relatively large, but its signal phase variation is quite small. On the other hand, even a small human body displacement (e.g., chest displacement induced by respiration) can cause a much larger signal phase variation. To efficiently move the virtual transceiver closer to the target, we employ a divide-and-conquer-based search scheme. The proposed scheme can localize the target bin amongst N bins in $O(\log N)$ iterations.

The third challenge is the severe self-interference when sensing small-sized targets. For example, for finger tapping, the signals can be reflected not just from the finger but also from the hand. Similarly, when we try to detect eye blink, the signals get reflected not just from the eyeballs but also from other parts of the face. Due to the small size, the signal strength of the reflection from the target is much weaker than that of the self-interference. Therefore, it is hard to sense the fine-grained activities from the extracted signal variations, which are the composite of target reflection and self-interference reflection. Even worse, the self-interference reflections are not constant and can change over time due to the unconscious human body movement, which can severely degrade the sensing performance.

To address this challenge, we design a signal processing method to remove the impact of self-interference and amplify the signal variations caused by the fine-grained activities. Specifically, we model the effect of self-interference together with the fine-grained activities in the I-Q domain. Two key observations are obtained: (i) the signal variations induced by the fine-grained activities can form an arc in the I-Q domain, and (ii) the signal variations caused by the self-interference change slower than that caused by the fine-grained activities, indicating that the center of the arc within a short period of time can be viewed as a fixed point. With these two observations, we propose to visualize the signal variations from the arc center rather than from the coordinate origin (0, 0) of the I-Q domain. This presents us two advantages: (i) the impact of self-interference is effectively removed, and (ii) the signal variations obtained from the new origin are larger than that obtained from the I-Q coordinate origin, leading to a better sensing performance and larger sensing range.

To demonstrate the effectiveness of our proposed solution, we implement LASense on two commercial off-the-shelf (COTS) devices, i.e., a smartphone and a smart speaker, for extensive experiments. Furthermore, we apply the proposed system to three fine-grained activity sensing applications: (i) respiration monitoring, (ii) finger tapping sensing, and (iii) eye blink detection. For respiration monitoring, the proposed method significantly increases the sensing range from the state-of-the-art 2 *m* to 6 *m* while still achieving a low median error. For finger tapping sensing, we can track the subtle in-the-air movement of a finger at 100 *cm*, outperforming the state-of-the-art sensing distance of 40 *cm*. For eye blink detection, we can increase the sensing range from 50 *cm* to 90 *cm* while still maintaining a high detection accuracy. Our main contributions are summarized as follows:

- To the best of our knowledge, LASense is the first system which can significantly increase the sensing range for fine-grained activities using only a single pair of speaker and microphone without involving any deep learning techniques.
- We theoretically and experimentally analyze the factors affecting the sensing performance of fine-grained activities, and propose the virtual transceiver idea to increase the sensing range. We believe the proposed method can be extended to improve the sensing performance of other chirp-based signals such as LoRa.
- We propose a sequence of signal processing techniques to search and extract the weak signal variations induced by fine-grained activities from far-away and small-sized targets.
- We implement LASense on two COTS devices (a smartphone and a smart speaker) and systematically evaluate the performance of sensing three fine-grained activities. Extensive experiments show that our system can significantly increase the sensing range, moving acoustic sensing one step further towards practical real-life adoption.

2 RELATED WORK

Recent years have witnessed a widespread interest in contact-free activity sensing that can sense human activities without requiring users to instrument any device. As a novel sensing approach, the research community has explored the opportunities of applying wireless signals in our ambient environment to enable activity sensing, including WiFi [1, 44, 45], RFID [6, 7, 14, 47, 48], visible light [16, 17, 35] and acoustic signals [15, 25, 39]. Compared with systems that employ other types of wireless signals for sensing, the acoustic-based system provides two unique advantages, making it an essential part of the ecosystem for contact-free activity sensing. On one hand, speakers and microphones are widely available in the electronic devices that we interact with on a daily basis, e.g., smartphones, smartwatches, smart speakers, etc. On the other hand, owing to the low propagation speed in the air (340 m/s), acoustic signals have inherent superiority in sensing granularity and precision, which can be employed to sense extremely fine-grained human activities (e.g., heartbeat monitoring [46] and eye blink detection [20]). However, due to the rapid acoustic signal attenuation over distance [15, 22], the sensing range for acoustic-based systems is very limited, which remains an open issue. In this section, we elaborate the similarities and differences between our proposed system and prior studies in increasing the sensing range of acoustic signals.

2.1 Coarse-grained Activity Sensing vs. Fine-grained Activity Sensing

A lot of efforts have been devoted to transforming the commodity acoustic-enabled devices into active sonar systems that can sense human activities. The research studies in acoustic activity sensing can be divided into two categories: (i) coarse-grained activity sensing and (ii) fine-grained activity sensing. The former exploits acoustic signals to sense large-scale human movements, such as human tracking [18, 25], gesture recognition [9, 19, 29, 41] and handwritten recognition [51]. The latter focuses on subtle human movements, including finger tracking/tapping [24, 32, 40, 43], respiration/heartbeat/lung function monitoring [28, 31, 38, 39, 46], eye blink detection [20], lip motion reading [49, 50], etc. Compared with the coarse-grained activity sensing, the issue of limited sensing range is more prominent for fine-grained activity sensing. One representative contrast is that the prior study [18] can easily track human walking at 5 *m* away from the smart speaker while the state-of-the-art study [39] can only achieve a sensing range of 2 *m* for respiration monitoring. The reasons are two-fold: On one hand, the reflection areas for fine-grained activities are usually much smaller than those for coarse-grained activities, resulting in weaker signal strength; On the other hand, the movements for fine-grained activities are usually more subtle than that for coarse-grained activities, which is much harder to be detected when the human target is far away from the device. In this paper, we target at the more challenging case, i.e., extending the range for sensing fine-grained activities using acoustic signals.

2.2 Existing Solutions for Long-range Activity Sensing

In this section, we describe existing solutions for long-range activity sensing using acoustic signals, including signal processing-based solutions and deep learning-based solutions.

2.2.1 *Signal Processing-based Solutions.*

The core idea of signal processing-based solution is to fuse information from different domains, i.e., time domain, frequency domain, and spatial domain, to improve the SNR of the weak reflected signals. One widely-adopted approach to increase the sensing range of acoustic signals is to integrate the information from multiple microphones (spatial domain). For example, Agarwal *et al.* [3] and Xie *et al.* [42] leverage beamforming of the microphone array on smart speakers for distal activity sensing. Besides the spatial-domain information, prior studies also exploit information from time domain and frequency domain to enhance the strength of the received signals. For example, FM-Track [15] designs a multi-dimensional parameter estimation algorithm to fuse information from not just multiple microphones (spatial domain) but also consecutive chirps (time domain), while BreathJunior [38] introduces a novel technique that combines the information from multiple microphones (spatial domain) and orthogonal chirps (frequency domain).

While all these approaches have the potential to address the bottleneck of the limited sensing range for acoustic signals, they can only be applied in some specific application scenarios. First, additional spatial-domain information requires an array of microphones, which is usually unavailable on resource-constrained devices such as smartphones and smartwatches. Furthermore, fusing information from time domain is not suitable for human activities that change very rapidly, e.g., eye blink, since the information from consecutive timestamps cannot be considered to be the same. Different from the above-mentioned approaches, our proposed system provides a general signal processing approach for long-range acoustic sensing without a need of a microphone array. The proposed solution can be applied to a large range of acoustic sensing applications.

2.2.2 *Deep Learning-based Solutions.*

Another solution to increase the sensing range of acoustic signals is the introduction of deep learning networks that can remove the impact of noise and multipath by extensive learning. RTrack [22] leverages the combination of signal processing and recurrent neural network (RNN) to learn the

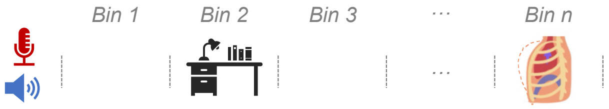


Fig. 3. The chirp signal allows us to separate the reflections from different distances into multiple range bins. The fine-grained activities can be extracted from the signal variations caused by the target movement (e.g., respiration) within a range bin.

signal path corresponding to the target. Another recent work DeepRange [21] generates synthetic training data and then feeds them to a deep neural network to combat noise and multipath interference.

Although deep learning-based solutions [21, 22] could help increase the sensing range, they need to configure the design of the neural network specific to different applications. For example, the network for respiration monitoring cannot be directly adopted to eye blink detection. Furthermore, new data should also be collected to train the network for new applications. In contrast, we provide a general solution that can be applied to a wide range of applications to increase the sensing range without a need of data collection and training. In general, the proposed signal processing method is orthogonal to deep learning-based techniques, and they can be combined to further enhance the capability of long-range activity sensing using acoustic signals.

3 PRELIMINARIES

In this section, we first introduce the basics of chirp signal that is widely used in acoustic sensing. Then, we present the sensing model to detect fine-grained human activities.

3.1 Primer on Chirp Signal

Chirp signal allows us to separate reflections from different distances [2, 37]. The chirp signal is a sine wave (Fig. 2a) whose frequency sweeps linearly over time (Fig. 2b). The transmitted signal can be represented as

$$s^T(t) = \cos\Big(2\pi\big(f_0 t + \frac{B}{2T}t^2\big)\Big), \tag{1}$$

where f_0 is the starting frequency, B is the bandwidth, and T is the sweep time. Consider a simple scenario where there is just a single signal reflected from the target. The reflection signal is a time-delayed version of the transmitted signal that can be represented as

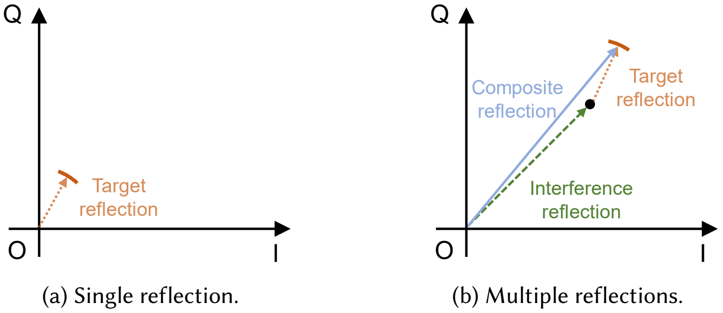
$$s^R(t) = \alpha\cos\Big(2\pi\big(f_0(t-\tau) + \frac{B}{2T}(t-\tau)^2\big)\Big), \tag{2}$$

where α is the signal amplitude attenuation factor, and τ is the Time-of-Flight (ToF) of the signal in the air. To obtain the distance information of the target, we need to mix the received signal with the transmitted signal by multiplying them together, i.e., $s^M(t) = s^T(t) \cdot s^R(t)$. After applying the product-to-sum conversion (i.e., $\cos A \cdot \cos B = \frac{1}{2}\big(\cos(A-B) + \cos(A+B)\big)$) followed by a low-pass filter, the mixed signal can be represented as

$$s^M(t) = \frac{1}{2}\alpha\cos\Big(2\pi\frac{B}{T}\tau t + 2\pi f_0\tau - \frac{\pi B}{T}\tau^2\Big). \tag{3}$$

As shown in Fig. 2, the mixed signal (green curve) is a sine wave with a constant frequency. Therefore, we can simplify the mixed signal in Equation (3) as

$$s^M(t) = \frac{1}{2}\alpha\cos(2\pi f_d t + \varphi_d), \tag{4}$$



(a) Single reflection.

(b) Multiple reflections.

Fig. 4. The illustration of the I-Q trace for single and multiple reflections.

where $f_d = \frac{B}{T}\tau$ is the beat frequency,[1] and $\varphi_d = 2\pi f_0\tau - \frac{\pi B}{T}\tau^2 \approx 2\pi f_0\tau$ is the initial phase. The approximation for the initial phase is based on the fact that $2\pi f_0\tau$ is usually two orders of magnitude larger than $\frac{\pi B}{T}\tau^2$ due to the very small value of τ. In the presence of multipath, the mixed signal in Equation (4) can be rewritten as a superposition of reflections from L paths

$$s(t) = \sum_{l=1}^{L} \frac{1}{2}\alpha_l \cos(2\pi f_{d_l} t + \varphi_{d_l}), \tag{5}$$

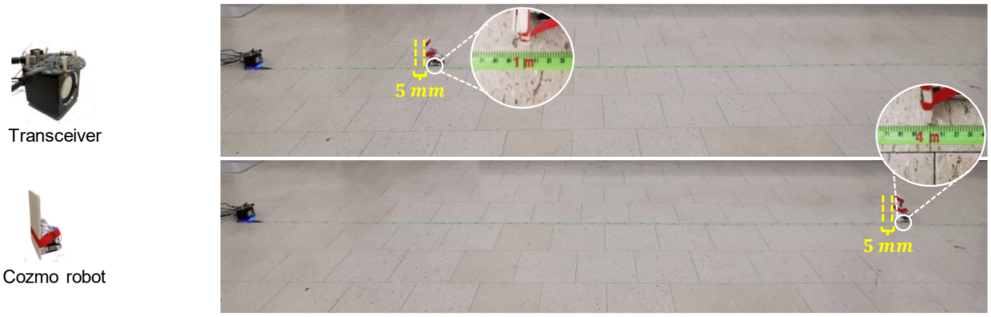
where α_l, f_{d_l}, and φ_{d_l} denote the attenuation factor, the beat frequency and the initial phase for the l^{th} path, respectively. The signals reflected from different distances have different ToFs, resulting in different beat frequencies. By performing Fast Fourier Transform (FFT) on the mixed signal, we can separate the mixed signal into multiple signals reflected from objects at different distances—termed as *bins* hereafter—as shown in Fig. 3.

3.2 Sensing Model for Fine-grained Activities

This section presents the sensing model for fine-grained activities. We first illustrate how the signal varies with the target movement and then describe how to identify the target bin based on the signal variations. Lastly, we explain how to obtain the target activity information from the signal variations.

3.2.1 Understanding Signal Variations. For each chirp, after performing FFT operation on the mixed signal, we can obtain the frequency domain information: one complex value for each frequency bin (i.e., range bin). We term this complex value as a signal sample in this paper. Note that the chirps are sent out one by one in time. For one frequency bin, if we take multiple chirps within a time window for consideration, we can obtain multiple signal samples. Without loss of generality, we consider the case with just one moving target. Among the frequency bins, there is one target bin that corresponds to the target's location and therefore contains the target's information. The signal samples obtained from this frequency bin vary with target movement, which are exactly the signal variation we will utilize for sensing. If we visualize consecutive signal samples in the complex I-Q plane, we can see a signal variation trace as shown in Fig. 4. If there is only one reflection within the target bin, the signal variation trace for a subtle target movement is an arc with the center at the origin, as shown in Fig. 4a. However, in reality, there are usually more than one reflection within the target bin. Take finger tracking as an example, besides the reflection from the finger, there are self-interference reflections from the hand. In this case, the finger-caused signal variation in the I-Q domain is an arc far away from the coordinate origin, as shown in Fig. 4b.

[1] Beat frequency is the frequency difference between the transmitted signal and received signal.



(a) Device.

(b) Experiment setup.

Fig. 5. The setup for benchmark experiments.

3.2.2 Identifying the Target Bin. Before extracting the signal variations induced by the fine-grained activity, we need to first identify the bin where the target locates. Due to target activity, the signal variations of the target bin are large in the I-Q domain. Therefore, prior studies computed the variance of the signal variations for each bin and identified the one with the largest variance as the target bin [28, 39, 46].

3.2.3 Interpreting Signal Variations. Each complex signal sample has both signal amplitude and phase. The amplitude is the distance between the signal sample and the origin of the coordinate, while the phase is the angle with respect to the I-axis. As the signal varies with target movement, the signal variation contains the context information of fine-grained human activities. Both amplitude and phase changes can be utilized for target activity sensing. For respiration monitoring, the frequency of the amplitude/phase changes indicates the respiration rate [31]. For finger tracking/tapping, we can obtain the finger displacement from the phase changes [32]. For eye blink detection, eye blink can be detected when there is a sharp amplitude/phase change [20].

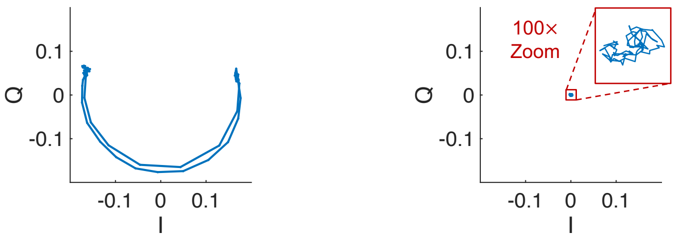
4 BOOSTING FINE-GRAINED ACTIVITY SENSING WITH VIRTUAL TRANSCEIVER

In this section, we first investigate the approaches in the prior studies [28, 38, 39, 46] and explain their limitations when sensing far-away targets via benchmark experiments. Then we mathematically analyze the factors affecting the sensing performance. At last, we introduce the proposed virtual transceiver concept and verify its capability in improving the sensing performance.

4.1 Limitations of Prior Studies

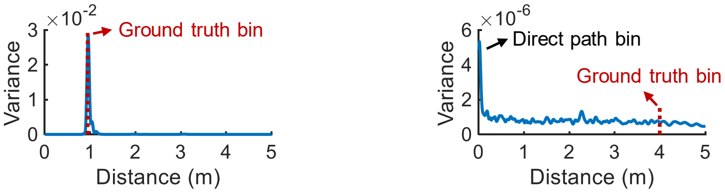
Prior studies pose two limitations for fine-grained activity sensing at distance using a single speaker-microphone pair. The first limitation is that we can hardly identify the bin where the target locates since the extracted signal variations are comparable or even smaller than the noise. Even if the target bin can be identified, there exists the second limitation, i.e., the sensing accuracy is low due to the noisy signal variations.

To demonstrate the limitations, we carry out two benchmark experiments using a single pair of speaker (i.e., ARVICKA speaker [5]) and microphone (i.e., a MiniDSP UMA-8-SP USB microphone board [23]). We employ a Cozmo robot [33] to precisely move the target, which can be controlled to move at a granularity of millimeter as shown in Fig. 5. We first mount a hand-sized cardboard on the Cozmo robot and control the robot to move back and forth on the ground with a displacement of 5 *mm*. The experiments are conducted at two different distances with respect to the transceiver, i.e., 1 *m* and 4 *m*. We employ an optoelectronic motion capture system (i.e., Qualisys [11]) mounted on the ceiling that supports sub-*mm*-level motion tracking accuracy to obtain the ground truths of the robot movement. To intuitively understand the signal difference at two distances, we extract



(a) Target at 1 *m*.
(b) Target at 4 *m*.

Fig. 6. (a) Compared with the signal variations at 1 *m*, (b) those at 4 *m* are much smaller and noisier.



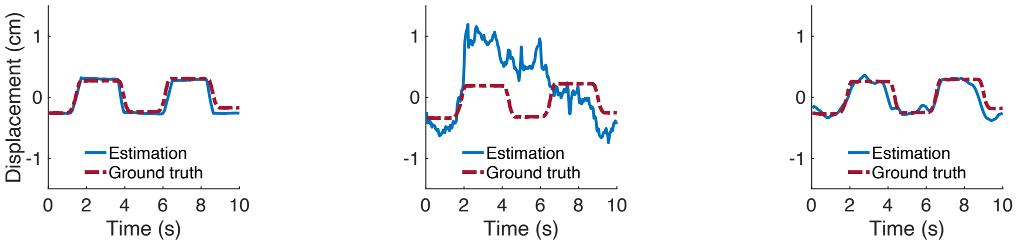
(a) Target at 1 *m*.
(b) Target at 4 *m*.

Fig. 7. (a) The variance of the signal variations for the target bin is much larger than those for other bins when the target is at 1 *m*, while (b) we can hardly identify the target bin when the target is at 4 *m*.

the signal variations induced by the robot movement from the ground truth bins, respectively, as shown in Fig. 6. We can observe that, even with the same amount of robot displacement, the signal variations extracted from the two distances are dramatically different. Specifically, due to the weaker reflections from the further distance, the signal variations at 4 *m* (Fig. 6b) are much smaller and noisier than those at 1 *m* (Fig. 6a). Next we explain how these two observations are related to the limitations of the prior studies.

Limitation 1: Difficulty in target bin identification. To identify the bin where the target locates, prior studies leverage the fact that, due to human activities, there are much larger signal variations at the target bin than other bins. This is true when the target is close to the transceiver, and turns out to be invalid when the target is far away. To demonstrate this, we further compute the variance of the signal variations for each bin, and plot the results in Fig. 7. We can observe from Fig. 7a that the variance of signal variations for the target bin at 1 *m* is much larger than those of other bins. However, we can hardly identify the target bin when the robot moves at 4 *m* in Fig. 7b. The reason is that the signal variations caused by the robot movement at 4 *m* are very small, and the variance of signal variations for the target bin is thus buried in noise. We also observe that, due to noise, the static multipath, e.g., the direct path from the speaker to the microphone (Fig. 7b), can cause relatively large signal variations. Although the impact from the static multipath is negligible when the target is close to the transceiver (Fig. 7a), it can significantly interfere with target bin identification when the target is far away.

Limitation 2: Degradation of the sensing accuracy. The noisy signal variations at a far-away distance would significantly degrade the sensing accuracy for fine-grained activities. To illustrate this, we obtain the phase changes from the signal variations and convert the phase changes to the target displacement. As shown in Fig. 8a and 8b, the robot displacement estimate is accurate at 1 *m* and is very inaccurate when the robot is at 4 *m*.



(a) 1 m without VT.

(b) 4 m without VT.

(c) 4 m with VT.

Fig. 8. (a) The displacement computed from the signal variations are accurate when the target is at 1 m, while (b) inaccurate when the target is at 4 m. (c) Virtual Transceiver (VT) can significantly improve the sensing accuracy for the latter case.

4.2 Factors Affecting the Performance

In this section, we mathematically analyze the factors affecting the sensing performance for fine-grained activities. As presented in Sec. 4.1, the poor sensing performance is due to the small and noisy signal variations. Therefore, we start with the derivation of the signal variations from the mixed signal.

Recall that each signal sample is obtained by performing FFT operation on the mixed signal, and then extracting the FFT value at a certain bin. The FFT operation is a computationally efficient Discrete Fourier Transform (DFT) algorithm for converting a signal from time domain to frequency domain. Therefore, we start the analysis with the definition of the DFT. Suppose that $x[n]$ is a given signal sequence with the length of N, the DFT and the inverse DFT (IDFT) can be defined as [26]

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi k}{N} n} = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi f_k}{f_s} n}, \tag{6}$$

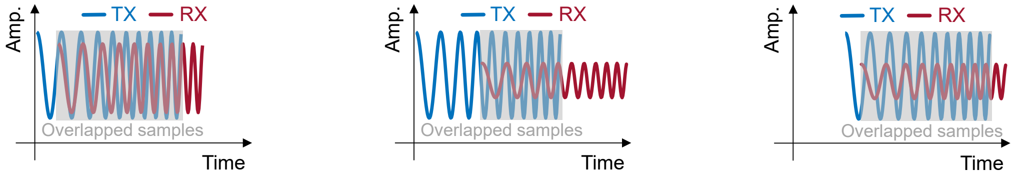
$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j\frac{2\pi k}{N} n} = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j\frac{2\pi f_k}{f_s} n}, \tag{7}$$

where $n = 0, 1, ..., N-1$ is the sample index, $k = 0, 1, ..., N-1$ is the frequency index (i.e., bin index), f_k is the frequency value for the k^{th} frequency bin, and f_s is the sampling rate. The DFT result $X[k]$ from Equation (6) is actually the signal sample extracted from the bin k after performing DFT.

Next, we show how to derive the signal sample $X[k]$ from the mixed signal. Since the mixed signal consists of a sequence of discrete samples, we can rewrite Equation (5) in the form of discrete signal and expand it using the Euler's formula as

$$\begin{aligned} s[n] &= \sum_{l=1}^{L} \frac{1}{2} \alpha_l \cos\left(2\pi f_{d_l} \frac{n}{f_s} + \varphi_{d_l}\right) \\ &= \sum_{l=1}^{L} \frac{1}{4} \alpha_l e^{j(2\pi f_{d_l} \frac{n}{f_s} + \varphi_{d_l})} + \sum_{l=1}^{L} \frac{1}{4} \alpha_l e^{-j(2\pi f_{d_l} \frac{n}{f_s} + \varphi_{d_l})} \\ &= \frac{1}{N} \sum_{k=0}^{N-1} \frac{1}{4} N \alpha_k e^{j\varphi_k} e^{j\frac{2\pi f_k}{f_s} n}. \end{aligned} \tag{8}$$

The derivation from step 2 to step 3 is based on the fact that, each beat frequency (i.e., $\pm f_{d_l}$) corresponds to one of the frequency bins in f_k. For the rest of f_k, we can set α_k to 0. By comparing Equation (8) with Equation (7), we



(a) Close target.

(b) Far target.

(c) Shifted TX.

Fig. 9. (a) Compared with the close target, (b) both the SNR of the received signals and the number of overlapped samples decrease for the far target. (c) Virtual Transceiver can increase the number of overlapped samples by shifting TX in the time domain.

can obtain that the signal sample at frequency bin k extracted from the DFT result of the mixed signal should be $S[k] = \frac{1}{4}N\alpha_k e^{j\varphi_k}$. The amplitude and phase of the signal sample are $\frac{1}{4}N\alpha_k$ and φ_k, respectively. Next, we discuss the factors affecting the amplitude and phase.

Factors affecting the amplitude. There are two factors affecting the amplitude of the signal sample: (i) the number of samples in the mixed signal N, and (ii) the signal amplitude attenuation factor α_k, which is proportional to the SNR of the received signal.

Factors affecting the phase. We leverage the Cramér-Rao Lower Bound (CRLB) to analyze the factors affecting the phase estimate, which indicates a lower bound on the variance of any unbiased estimator. The CRLB for phase estimation for a sine wave is given by [12]

$$var(\varphi_k) \geq \frac{4}{N\eta}, \tag{9}$$

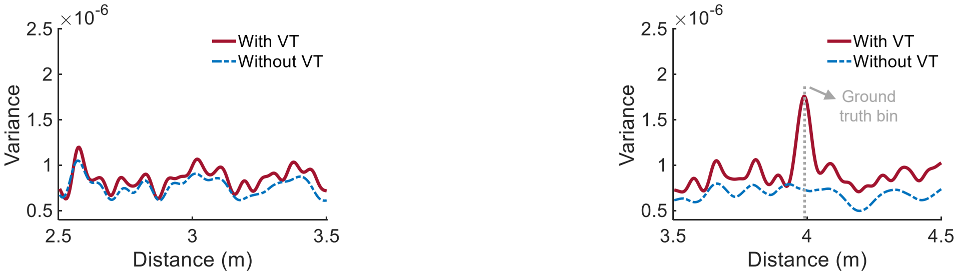
where var is the variance of the estimation, N is the number of samples in the mixed signal, and η is the SNR of the mixed signal, i.e., the SNR of the received signal.

From the above analysis for the amplitude and phase of the signal sample, we can see that the sensing performance is determined by two factors: (i) the SNR of the received signal and (ii) the number of samples in the mixed signal that is equivalent to the number of samples overlapped by the transmitted signal and the received signal. When the distance between the target and transceiver increases, these two factors decrease as demonstrated in Fig. 9a and 9b, which reveals the reason why activity-induced signal variations become smaller and noisier at further distances.

4.3 Improving Sensing Performance with Virtual Transceiver

From the previous section, we know that to improve the sensing performance for fine-grained activities, we need to either boost the SNR of the received signals or increase the number of samples overlapped by the transmitted and received signals. Previous studies choose to boost the SNR of the received signals by constructively combining signals from an array of microphones (> 4) [3, 22, 38]. Instead of boosting the SNR of the received signals which requires multiple microphones, we investigate an approach to increase the number of samples overlapped by the transmitted signal and received signal by introducing an idea of virtual transceiver.

As shown in Fig. 9c, rather than directly multiplying the transmitted signal with the received signal, we propose to multiply a shifted version of the transmitted signal with the received signal to increase the number of overlapped samples, which is equivalent to *virtually* moving the transceiver closer to the target. The increased number of samples can improve the sensing performance in two aspects. On one hand, it can amplify the amplitude of the signal sample and thus increase the sensing range. On the other hand, it can enhance the accuracy of phase estimation, which consequently improves the sensing accuracy of fine-grained activities. Fig. 8c illustrates



(a) Range interval without target.

(b) Range interval with target.

Fig. 10. Virtual Transceiver (VT) can significantly amplify the signal variations at the target bin but only have little impact on those at the other bins.

the accurate displacement estimation after we apply the virtual transceiver scheme, which demonstrates the effectiveness of the proposed method.

5 SEARCHING THE TARGET BIN

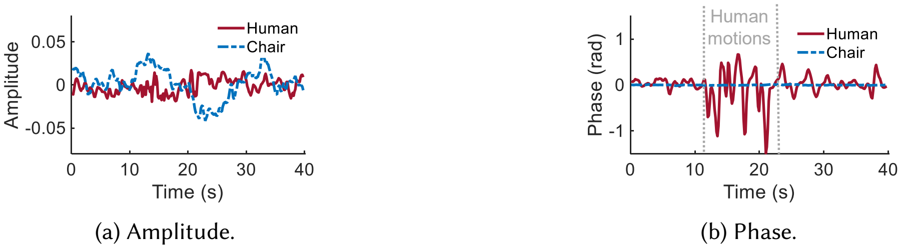
The objective of the proposed virtual transceiver scheme is to virtually move the transceiver closer to the target to increase the number of overlapped samples between the transmitted signal and received signal. However, without the information of the physical location of the target, it is challenging to quickly move the virtual transceiver closer to the target for better performance. In this section, we propose approaches to first identify all the target candidates and then differentiate the human target from the interfering static objects that are also considered as potential targets in the first step. At last, we present our strategy to efficiently move the virtual transceiver.

5.1 Identifying Target Candidates

To determine whether there are possible targets within a bin, we exploit one key observation that, when the virtual transceiver is moved closer, the signal variations at the bins with targets can be significantly amplified while the signal variations at other bins are not changed too much. To demonstrate this, we leverage the benchmark experiment in Sec. 4 where a robot moves 4 m away from the transceiver. We compute the variance of the signal variations for all the bins with and without the proposed virtual transceiver method within two different range intervals, i.e., $[2.5\ m, 3.5\ m]$ and $[3.5\ m, 4.5\ m]$, respectively. As shown in Fig. 10, the signal variations at the target bin are significantly increased when the virtual transceiver moves closer to the target at 4 m. Therefore, we first identify the target candidate bins by picking the bins whose variance is significantly larger than their adjacent bins when we virtually move the transceiver closer. Suppose that $V = [v_1 \cdots v_i \cdots v_n]$ is a vector representing the variances of the signal variation for n bins. We first identify all peaks in V and arrange them in a vector P. Then we adopt an outlier detection algorithm [13] to determine if any peaks in P are target candidates. Specifically, we define a metric called *peak variance ratio*, i.e., pvr, as

$$pvr_j = \frac{p_j - med(V)}{mad(V)}, \tag{10}$$

where p_j is the j^{th} peak in P, med computes the median, and mad compute the median absolute deviation. If there exists any peak in P whose pvr value is larger than a pre-defined threshold, we consider it as a target candidate. We empirically set the threshold as 3.



(a) Amplitude.

(b) Phase.

Fig. 11. The signal amplitude variation for the static object (chair) is relatively large, but its signal phase variation is quite small. The human motions can induce a much larger signal phase variation.

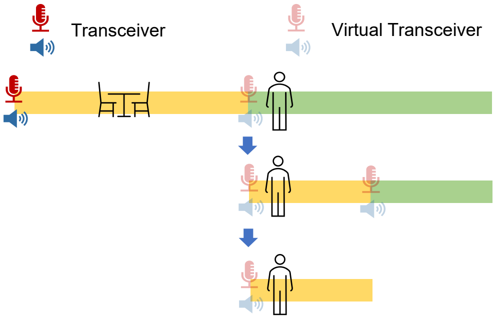
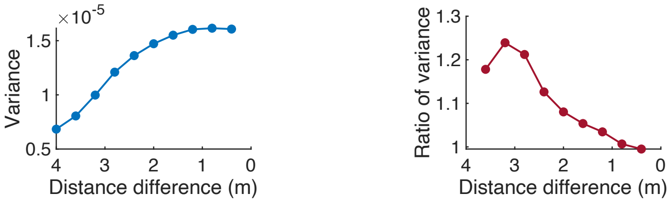


Fig. 12. The illustration of our proposed search algorithm. The yellow and green areas are the search space.

5.2 Differentiating Target from Static Objects

Although the above-mentioned method can identify the true target bins, the bins with the static objects can be mistakenly identified as potential target bins. Through experiments, we found that the bins where static objects are located can also cause relatively large signal variations.

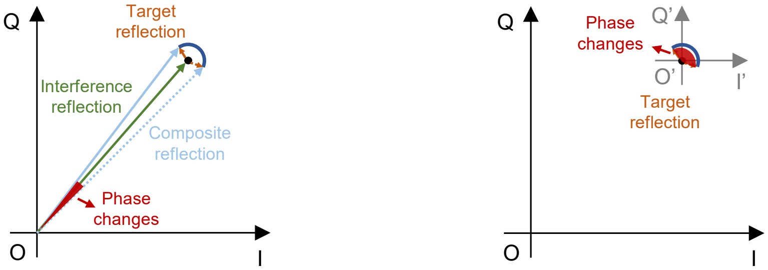
We conduct a benchmark experiment to show the above observation. We extract the signal amplitude changes and phase changes when there is one human target sitting in front of the transceiver and when there is only a static object, i.e., a chair, at the same position, respectively. As shown in Fig. 11, we can obtain an interesting observation that the signal amplitude variation is large (Fig. 11a) but the signal phase variation is quite small (Fig. 11b). Differently, even a small body movement (e.g., chest displacement caused by respiration) can cause a much larger signal phase variation (Fig. 11b). Therefore, we identify the candidate as a static object if its phase variance is smaller than a pre-defined threshold. We empirically set the threshold as 0.2. Furthermore, we only initiate fine-grained activity sensing when no large motions are detected for the target. As shown in Fig. 11b, other motions (e.g., motions from arms) can cause large phase changes, which can impact the fine-grained activity sensing (e.g., respiration monitoring). We detect the presence of large motions if the phase variance of the target is larger than 0.8. Since other motions can interfere the reflected signals from fine-grained activities, we do not sense fine-grained activities during the periods with large human motions. It is worth noting that, since the chirp signal enables the separation of signals reflected from objects at different distances, our system can successfully identify target bins corresponding to multiple human targets.



(a) Variance.

(b) The ratio of variance.

Fig. 13. As the virtual transceiver moves closer to the target, i.e., the distance difference between the target and transceiver decreases, the variance of the signal variations caused by the target movement first rapidly increase and then plateau.



(a) Small phase changes.

(b) Large phase changes.

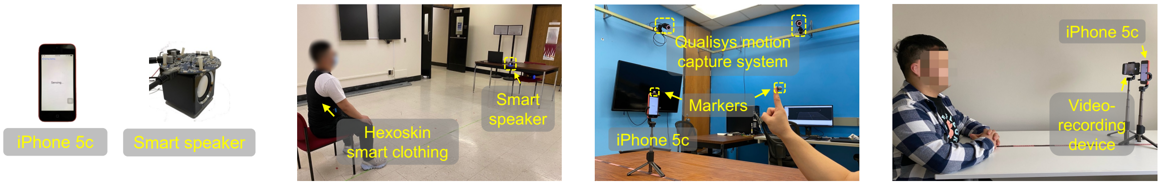
Fig. 14. (a) The phase changes extracted from the signal variations for the small-sized target are very small due to the existence of self-interference reflections. (b) The phase changes can be significantly amplified by translating the coordinate.

5.3 Searching Strategy

One question naturally arises: *how should we virtually move the transceiver closer to the target?* A naïve solution is to exhaustively move the virtual transceiver to all the range bins one by one (i.e., brute-force search), and then identify the target bin. This is, however, computationally expensive and time-consuming.

Instead, we propose a divide-and-conquer-based search scheme that can efficiently move the virtual transceiver, and progressively reduce the search scope. As shown in Fig. 12, in the divide step, we divide all range bins into two parts. For each part, we would first virtually move the transceiver to the starting bin, and compute the variance of the signal variations for all the bins within the part. In the conquer step, we identify the target candidates (Sec. 5.1) and remove the static objects from the candidates (Sec. 5.2). If there exists a candidate in the search scope, we narrow down the search scope and further divide it into smaller parts. The above process terminates when either (i) there are no candidates in the search space or (ii) the variance ratio between two consecutive iterations is smaller than a pre-defined threshold, i.e., 1.05, according to our experiments. The second condition to terminate search is based on the fact that, as the virtual transceiver moves closer to the target, the variance of the signal variations caused by target movement would first rapidly increase and then plateau, as shown in Fig. 13.

Computational complexity. In comparison to an exhaustive search over N bins, the proposed search algorithm can reduce the computational complexity to $O(\log N)$. The number of range bins N is determined by the bin size and the search scope. The bin size is 4.29 mm in our implementation. The search scope is decided by applications, e.g., 5 m for respiration monitoring and 1 m for finger tapping. The target search process only happens once at the beginning of the activity sensing. If there is no activity (motion) for a certain period, we can re-initiate the search process.



(a) Sensing device. (b) Respiration monitoring. (c) Finger tapping sensing. (d) Eye blink detection.

Fig. 15. The experiment setups for three fine-grained human activities using the smartphone and smart speaker.

6 REMOVING SELF-INTERFERENCE

For human sensing, there are always self-interference reflections from uninterested body parts besides the reflections from the target of interest. For example, when we try to detect eye blinks, the signals could be reflected not only from the eyeballs but also from other parts of the face. The self-interference reflections can significantly reduce the amplitude/phase changes caused by the fine-grained activities especially when the target is small. Therefore, before we can extract the amplitude/phase changes, we need to carefully handle self-interference.

To illustrate the issue, we model the effect of the self-interference together with the fine-grained activities in Fig. 14a. Due to the small size, the strength of the reflection signal from the target is much weaker than that of the self-interference. The extracted amplitude/phase changes would be very small since the signal variations are based on the composite of target reflection and self-interference reflection. Even worse, we observe from the experiments that the self-interference reflections can change over time due to unconscious human body movements caused by respiration and heartbeat.

We design a method to remove the reflections from the self-interference and amplify the phase changes caused by the human activity. Specifically, we first leverage the important observation that the signal variations induced by the fine-grained activities can form an arc in the I-Q domain as long as there exists a small target displacement. However, due to the existence of noise, we can hardly obtain an arc when the target is far away. To address the issue, we smooth the noisy signal variations using the Savitzky-Golay filter [30]. Furthermore, we obtain another key observation through experiments that the signal variations caused by self-interference usually change slower than that by the fine-grained activity, indicating that the center of the arc within a short period can be viewed as a fixed point. Therefore, we first extract the signal variations by a moving window and then fit them to find the center of the arc using the Pratt method [27]. By translating the origin of the coordinate O to the center of the arc O', we can remove the impacts from self-interference reflections, as shown in Fig. 14b. The phase changes are significantly amplified based on the translated new coordinate.

7 IMPLEMENTATION

We implement LASense on two COTS devices including a smartphone and a smart speaker, as shown in Fig. 15a. The signals are analyzed using MATLAB on a ThinkPad X1 Extreme laptop.

Smartphone: We adopt an iPhone 5c [4] to evaluate the proposed system for finger tapping sensing and eye blink detection. The software implementation is based on an existing framework, namely LibAS [34], which allows us to develop our sensing algorithm using MATLAB without considering the smartphone-specific details.

Smart speaker: We implement our prototype system for respiration monitoring using a smart speaker prototype, which consists of an ARVICKA speaker [5] and the MiniDSP UMA-8-SP microphone board [23]. We compare the proposed approach using one microphone with the prior study using seven microphones [38].

Acoustic signals: The default acoustic signal adopted in our implementation lies in the inaudible frequency band, which sweeps from 18 kHz to 22 kHz. The duration of the chirp is related to the specific applications, i.e., 150 *ms* for respiration monitoring, 50 *ms* for finger tapping sensing, and 40 *ms* for eye blink detection. The sampling rates for both the smartphone and smart speaker are 48 *kHz*.

Ground truth measurements: As shown in Fig. 15, for respiration monitoring, we employ the Hexoskin smart clothing [10] to collect the ground truths of the respiration rate. For finger tapping sensing, we attach one reflective marker on the finger and employ an optoelectronic motion capture system (i.e., Qualisys [11]) to obtain the ground truths of finger movements. For eye blink detection, we employ a separate smartphone to video-record the ground truths of eye blinking events.

8 EVALUATION

In this section, we first elaborate on the experiment setups for the three fine-grained activity sensing applications. Then we evaluate the performance of our proposed solution for each of the applications under different conditions.

8.1 Experiment Setup

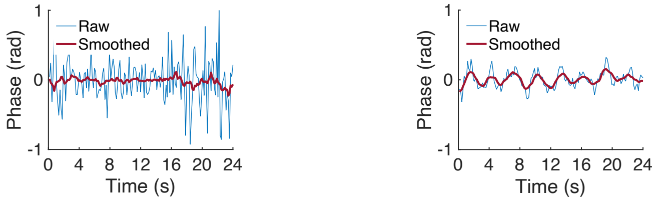
To demonstrate the generality of our proposed solution to increase the sensing range for fine-grained human activities, we evaluate its performance using three applications: (i) respiration monitoring, (ii) finger tapping sensing, and (iii) eye blink detection. The above-mentioned human activities represent three categories of fine-grained activities in terms of the scale of body movements and the size of reflection areas. Respiration monitoring represents human activities with small-scale body movements and large reflection areas. We adopt respiration as the benchmark experiments to investigate the impacts of sweep time, target-transceiver angle, clothing, ambient noise, target diversity, and interference. Finger tapping sensing is chosen to study the effectiveness of our proposed solution on human activities with large-scale movement and small reflection areas. Finally, we chose eye blink to study the effectiveness of our proposed solution on human activities with small-scale movements and small reflection areas.

8.1.1 Human Participants. We recruited 15 healthy volunteers to participate in the study, including 4 undergraduate students, 10 graduate students, and 1 campus staff. The recruited participants were diverse in age (from 18 to 62 years old) and gender (5 females and 10 males). Before conducting the experiments, we asked the participants to read the informed consent document carefully, and went through the details of the experiments with them.

8.1.2 Respiration Monitoring. We chose a hall with a size of 10 $m \times 5$ m to conduct respiration monitoring experiments due to its relatively large space to fully showcase the long-range sensing capability of the proposed system. Each participant was asked to first wear the Hexoskin smart clothing [10] for ground-truth collection. Next, the participant was asked to sit in front of the smart speaker that was placed on the table and naturally breathes. The default distance between the participant and smart speaker was set as 3 *m*. For respiration monitoring, each experiment trial lasted for 60 *s*. We repeated each experiment trial 10 times.

8.1.3 Finger Tapping Sensing. The experiments were conducted in our research laboratory for the convenience of collecting ground truths using the Qualisys motion capture system [11]. One reflective marker was attached to the index finger of each participant for ground-truth collection. Then, the participant was asked to sit in front of the smartphone mounted on a phone holder and freely tap the index finger in the air. The height of the smartphone was set to 0.4 *m* with respect to the table. The default distance between the finger and smartphone was set to 0.9 *m*. For finger tapping, each experiment trial lasted for 10 seconds. We repeated each experiment trial 10 times.

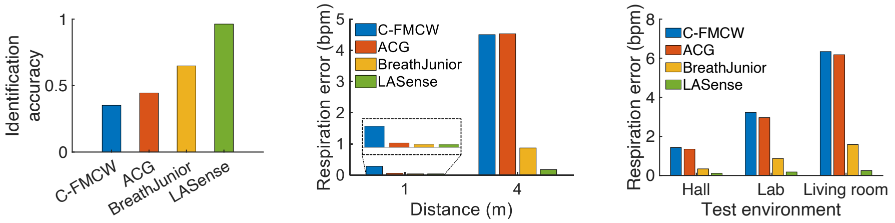
8.1.4 Eye Blink Detection. We conducted the eye blink experiments in our research laboratory. The participant was asked to sit in front of the smartphone mounted on the phone holder along with the video-recording device. The height of the smartphone was set to align with the eyes of the participants. Then the participant was asked



(a) Without LASense.

(b) With LASense.

Fig. 16. The illustration for respiration monitoring at 5 *m* with and without our system.



(a) Human identification.

(b) Long-range sensing.

(c) Robust sensing.

Fig. 17. The performance comparison among various approaches for respiration monitoring.

to blink naturally for data collection. The default distance between the eye and smartphone was set to 0.7 *m*. For eye blink, each experiment trial lasted for 20 seconds. We repeated each experiment trial 10 times.

8.2 Respiration Monitoring

In this section, we report our experiment results for respiration monitoring and discuss critical factors that could impact the performance of our proposed approach. We define two metrics to evaluate the performance: (i) target bin identification accuracy and (ii) sensing accuracy. For target bin identification accuracy, we compute the probability that the target bin can be successfully identified among all tests. For sensing accuracy, we adopt respiration error to measure the accuracy of respiration monitoring, which is defined as the absolute value of the difference between the estimated respiration rate and ground-truth rate.

8.2.1 Illustrative Experiment. To intuitively demonstrate the effectiveness of the proposed system, we extract the phase changes caused by respiration when the participant sits 5 *m* away from the transceiver. Fig. 16a shows the phase changes extracted from the ground truth bin without applying our approach, while Fig. 16b illustrates the phase changes extracted by LASense. We can observe that the respiration pattern cannot be identified without our approach while can be easily identified after the proposed approach is applied.

8.2.2 Performance Comparison. We compare the performance of respiration monitoring between LASense and the state-of-the-arts, including C-FMCW [39], ACG [28], and BreathJunior [38]. BreathJunior employs an array of seven microphones to boost the SNR of the reflected signals. Note that our proposed system LASense only adopts one microphone.

Identifying the human target. Fig. 17a shows the target bin identification accuracy for the four different systems. The detection accuracy for LASense is 96.3%, outperforming the state-of-the-arts. One unexpected observation is that our system with only one microphone outperforms BreathJunior with seven microphones.

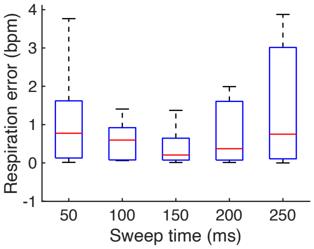
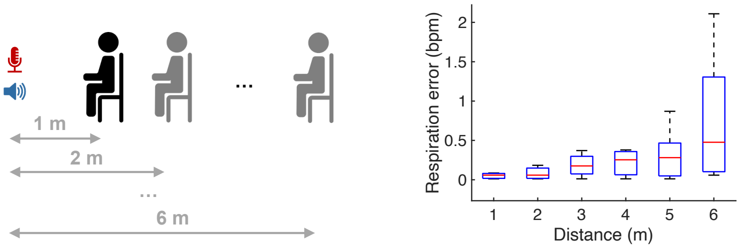


Fig. 18. Impact of sweep time.



(a) Different distances.

(b) Respiration errors.

Fig. 19. Impact of target-transceiver distance.

The main reason is that the microphone array-based beamforming not only improves the SNR of signals from the fine-grained activities, but may also increase the SNR of signals reflected from surrounding static objects. Note that when the human target is close to the transceiver (e.g., 1 *m*), the target reflection dominates. When the target is far away from the transceiver (e.g., 4 *m*), the multipath interference from the surrounding static objects becomes much more severe, degrading the identification accuracy of the target bin.

Monitoring respiration at distance. We compare the respiration sensing performance of different systems in Fig. 17b when the human target is 1 *m* and 4 *m* away from the smart speaker, respectively. We can observe that the performance is comparable for all the four systems when the target is close to the smart speaker (1 *m*). However, when the target is far away from the smart speaker (4 *m*), the performance of other systems significantly degrades, while our proposed system can still achieve a low median error of 0.21 *bpm*. The better performance for our proposed system is benefited from the design of virtual transceiver that can amplify the signal variations caused by respiration and eliminate the impact from other interference.

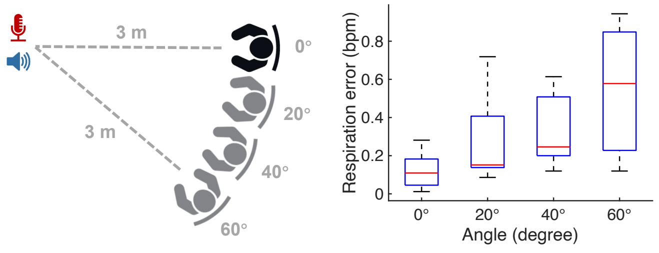
Monitoring respiration in different environments. To evaluate the robustness of LASense in different environments, we asked two of the participants to perform experiments for respiration monitoring in three environments (empty hall, laboratory, and living room) with an increasing amount of static multipath from the surrounding objects. In this experiment, the distance between the target and the transceiver is 4 *m*. As shown in Fig. 17c, benefited from the capability of identifying the target from the surrounding static objects, the proposed system can robustly monitor respiration in all three environments. Specifically, we achieve a low median error of 0.25 *bpm* even in the multipath-rich living room, outperforming the state-of-the-arts.

8.2.3 *Impacting Factors.*

This section evaluates the impact of factors affecting respiration monitoring.

Impact of sweep time. Increasing the sweep time T for the chirp signal can also increase the number of samples overlapped between the transmitted and the received signals. However, a too-long chirp reduces the capability of capturing the subtle signal variations caused by the fine-grained activities. To obtain a suitable sweep time for respiration monitoring, we asked one participant to sit at 3 *m* in front of the smart speaker and varied the sweep time from 50 *ms* to 250 *ms* at a step size of 50 *ms*. As shown in Fig. 18, the errors for respiration monitoring first decrease and then increase as T increases, which is consistent with the above-mentioned analysis. Based on the results, we set T to 150 *ms* for respiration monitoring.

Impact of target-transceiver distance. To evaluate the performance of respiration monitoring at different distances, we asked one participant to sit in front of the smart speaker and varied the distance between the participant and smart speaker from 1 *m* to 6 *m* at a step size of 1 *m*, as shown in Fig. 19a. As we can observe from the estimation errors of respiration in Fig. 19b, although the error increases with distance, the proposed system can still achieve a low median error of 0.47 *bpm* at 6 *m*. The sensing range of the state-of-the-arts is merely 2 *m*,



(a) Different angles.

(b) Respiration errors.

Fig. 20. Impact of target-transceiver angle.

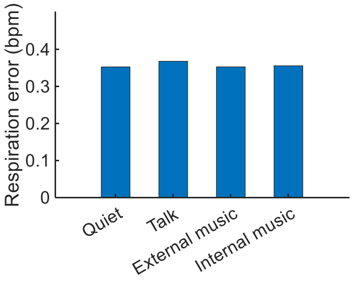


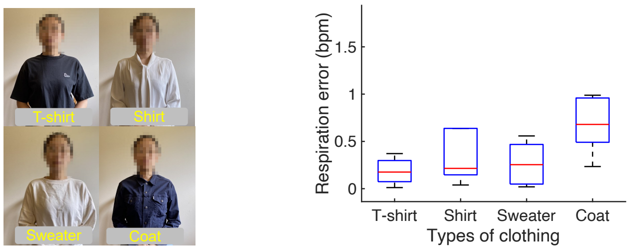
Fig. 21. Impact of ambient noise.

and LASense improves the range by 200%. The extended sensing range attributes the introduction of virtual transceiver that boosts the sensing performance when the respiration signals reflected from the far-away human target are very weak.

Impact of target-transceiver angle. Due to the high radiation directivity of inaudible acoustic signals from commodity speakers [8], the SNR of the received signals decreases as the angle of the target with respect to the transceiver increases. To evaluate the impact of the target-transceiver angle, we asked one participant to sit at 3 *m* in front of the smart speaker and varied the angle between the participant and smart speaker from 0° to 60° at a step size of 20°, as shown in Fig. 20a. We can observe from Fig. 20b, as the angle increases from 0° to 60°, the median respiration monitoring error slightly increases. We can still achieve a low median error of 0.22 *bpm* at 40°, indicating that the impact of the angle on performance is relatively insignificant when the angle is below 40°. The above results demonstrate that our proposed system can support room-scale respiration monitoring even when the user is in different directions with respect to the smart speaker.

Impact of ambient noise. To evaluate the impact of ambient noise, we asked one participant to sit at 4 *m* in front of the smart speaker and introduced three types of noises when monitoring his respiration. The first type of noise is human voice. We asked another participant to sit at 0.5 *m* and 40° with respect to the smart speaker and read an article with the normal speech volume. The second type of noise is external music that is played by an external smartphone. We placed the smartphone at 0.5 *m* and 40° with respect to the smart speaker and played music at its 80% maximum volume. The third type of noise is internal music that is played by the smart speaker itself. The smart speaker played the music together with our chirp signals at its 60% maximum volume. We measured the sound pressure levels by putting the VLIKE sound level meter [36] at the position of the smart speaker. As shown in Fig. 21, the respiration errors for quiet (39.3 *dB*), human voice (55.5 *dB*), external music (69.2 *dB*) and internal music (68 *dB*) are 0.352 *bpm*, 0.360 *bpm*, 0.353 *bpm* and 0.356 *bpm*, respectively. We observe that similar accuracies are achieved for different ambient noises since the frequency band adopted for sensing is much higher than that of noise. Furthermore, we demonstrate that, due to the frequency gap between sensing signals and ambient noises, our sensing applications do not interfere with the common usage of smart speakers such as playing music.

Impact of clothing. Various types of clothes have different attenuation effects on the reflected signal, and thus can impact the performance of contact-free respiration monitoring [31]. To evaluate the impact of clothing, we asked one participant to wear different clothes (T-shirt, shirt, sweater, and coat) sitting at 4 *m* in front of the smart speaker, as shown in Fig. 22a. We can observe from Fig. 22b that the results show small performance variation among different clothes. Even when the participant wears loose clothes like a coat, we can still achieve a relatively high sensing accuracy, indicating the applicability of the proposed system in real-world settings. Furthermore, although the sweater is much thicker than the shirt, it achieves slightly better performance. The



(a) Different clothing.

(b) Respiration errors.

Fig. 22. Impact of clothing.

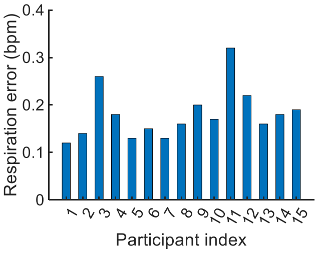
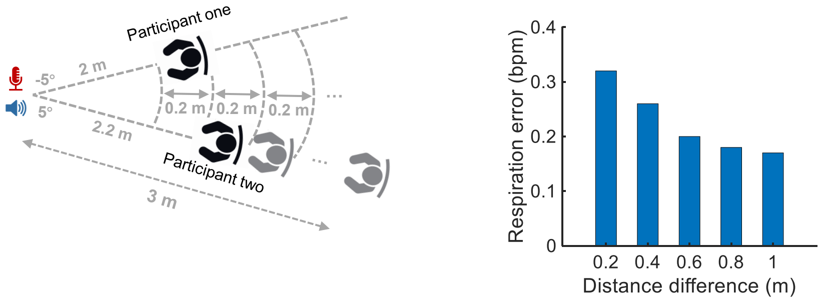


Fig. 23. Impact of user diversity.



(a) Different distance differences.

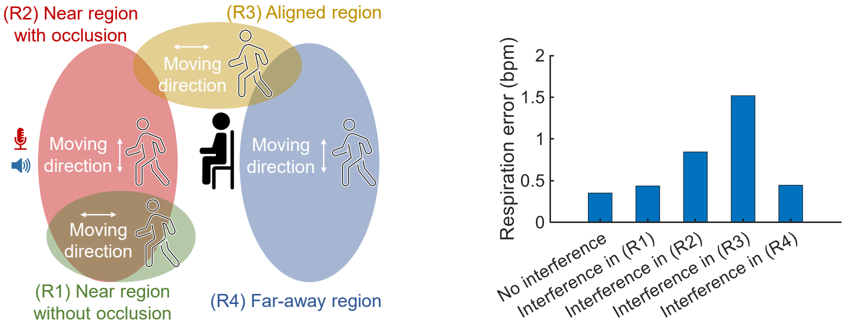
(b) Respiration errors.

Fig. 24. Impact of multiple targets.

reason is that the sensing performance is related to not only the thickness of the clothes but also the looseness of the clothes. Since the proposed system relies on signals reflected from the chest to sense respiration, the close-fitting sweater can lead to larger signal variations and thus better performance.

Impact of user diversity. To evaluate the impact of user diversity, we display the median respiration errors for all fifteen participants in Fig. 23. From the results, we observe that the performance of respiration monitoring is dependent on two factors: body size and sitting posture. Specifically, the higher respiration error for Participant 3 is due to her weaker chest motions and smaller body size. Furthermore, the higher respiration error for Participant 11 is caused by his reclined sitting posture. The reclined sitting posture will result in weaker reflected signals from the chest and thus poorer performance for respiration monitoring.

Impact of multiple human targets. To evaluate the performance of respiration monitoring when there exist multiple human targets, two participants were asked to sit in front of the smart speaker with different distances between them. As shown in Fig. 24a, one participant was asked to sit at 2 m and -5° with respect to the smart speaker. Furthermore, another participant was asked to sit at 5° and vary the distance from 2.2 m to 3 m at a step size of 0.2 m. Fig. 24b shows the average of median respiration errors for two participants under different experiment settings. Although the respiration error increases as the distance difference between two participants decreases, we can still achieve a respiration error of 0.32 *bpm* even when the distance difference is only 0.2 m. The increasing respiration error is caused by the mutual interference of signals reflected from the two participants.



(a) Different dynamic interference.

(b) Respiration errors.

Fig. 25. Impact of dynamic interference.

Impact of dynamic interference. We have already considered the interference from static objects (e.g., furniture and walls) in three different environments in Sec. 8.2.2. Now we evaluate the impact of dynamic environment interference from moving people. We asked one participant to sit at 3 *m* in front of the smart speaker and another participant to walk around in four different regions as shown in Fig. 25a: (R1) near region without occlusion, (R2) near region with occlusion, (R3) aligned region and (R4) far-away region. We compare the experiment results with and without interference in Fig. 25b. We can observe that the interferer has little impact on respiration monitoring when he moves in the near region without occlusion or the far-away region. This is because the design of the chirp signal can separate the reflections from the interferer and the human target into different range bins. And there exists little interference between them when their range bins are far away from each other. In contrast, the impact of the interferer increases substantially when he moves in the aligned region. The reason is that the range bins of the interferer and the human target are so close that it is difficult to separate signals reflected from the interferer and the human target using only one microphone. Furthermore, it is unexpected that we can still monitor the respiration with a slightly high error when the interferer moves in front of the human target, i.e., in the near region with occlusion. By analyzing the experiment results, we find that the interferer only occludes the line-of-sight signals reflected from the human target for a very short period. The computed phase changes during occlusion can be viewed as outliers in the respiration signals and can be filtered out in our signal processing.

8.2.4 Computational Costs. To compute the overall end-to-end execution time, we run algorithms using the experiment data for studying the impact of the transceiver-target distance. The computational cost for LASense mainly consists of two parts: (i) searching the target bin and (ii) extracting the respiration information from the target bin. For the former part, we compare the computational cost of the LASense (i.e., divide-and-conquer-based search) with that of the naïve method (i.e., brute-force search). As the search space is proportional to the distance between the device and target, Fig. 26 plots the median search time at different device-to-target distances for the two search methods. We observe that the computational cost for the naïve method increases significantly as the distance between the transceiver and target increases. In contrast, LASense maintains a low computational cost for all different distances. The median search time for LASense is 0.26 *s*. For the second part, the median execution time for extracting the respiration information is 0.04 *s*. The median overall end-to-end execution time for respiration monitoring is 0.3 *s*, which can support real-time analysis.

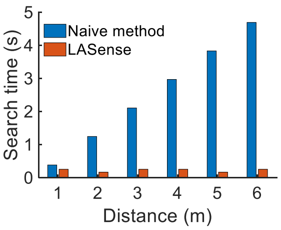
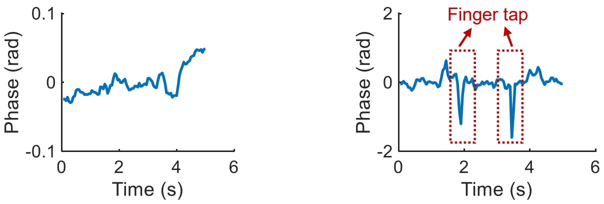


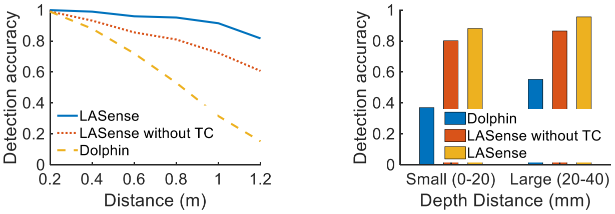
Fig. 26. Search time comparison.



(a) Without LASense.

(b) With LASense.

Fig. 27. The illustration for finger tapping sensing at 1 *m* with and without our system.



(a) Long-range sensing.

(b) Robust sensing.

Fig. 28. The performance comparison among different approaches for finger tapping sensing.

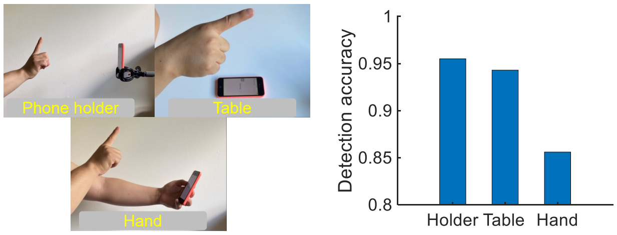
8.3 Finger Tapping Sensing

In this section, we report our experiment results for finger tapping sensing. We evaluate the impact of sweep time and empirically choose 50 *ms*. We adopt the detection accuracy to measure the accuracy of detecting the finger tapping, which is defined as the probability that the finger tapping can be detected among all tests.

8.3.1 Illustrative Experiment. To intuitively demonstrate the effectiveness of our proposed system, we extract the phase changes caused by the finger tapping when the participant performs the finger tapping 1 *m* away from the transceiver. Fig. 27a shows the phase changes extracted from the ground truth bin without applying our approach, while Fig. 27b shows the phase changes extracted by LASense. We can observe that LASense significantly amplifies the phase changes, resulting in clear peaks when the participant performs finger tapping.

8.3.2 Performance Comparison. We compare the performance of LASense with the state-of-the-art, i.e., Dolphin [32]. Note that Dolphin adopts both video and audio information for finger tapping sensing. To achieve a fair comparison, we only compare the proposed system with the audio part of Dolphin and adopt the same experiment setup for both systems. To demonstrate the effectiveness of removing the self-interference, we compare the results for the proposed system with and without translating the coordinate (TC).

Sensing the subtle finger tapping at distance. To demonstrate the performance of our proposed system in increasing the sensing range for finger tapping, we compare the experiment results of finger tapping sensing for the proposed system with Dolphin at different distances. We asked the participants to sit in front of the smartphone to perform finger tapping and varied the distance between the finger and smartphone from 0.2 *m* to 1.2 *m* at a step size of 0.2 *m*. As shown in Fig. 28a, LASense can achieve a median finger tapping sensing accuracy of 91.3% even at 1 *m*, which improves the sensing range by 150% compared with that of the state-of-the-art study 0.4 *m*. Specifically, even without translating the coordinate, LASense can outperform the prior study Dolphin, which is benefited from the amplification of the finger-induced signal variations provided by the virtual



(a) Different placements.

(b) Detection accuracy.

Fig. 29. Impact of phone placement.

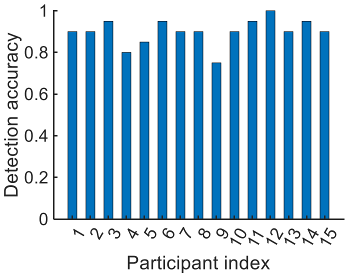


Fig. 30. Impact of user diversity.

transceiver. When the finger is far away from the smartphone, the self-interference reflections from uninterested body parts become stronger, degrading the sensing performance. By introducing the coordinate translation to reduce the impact of self-interference, we can boost the performance of finger tapping sensing especially when the finger is far away from the smartphone. However, we observe that when the finger is further than 1.2 *m* away from the smartphone, the performance degrades significantly, indicating that the impact of self-interference cannot be reduced simply by the coordinate translation.

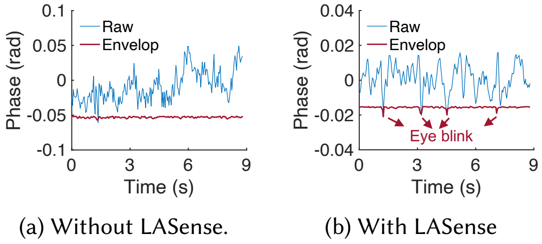
Sensing the finger tapping with different depths. Different tapping depths can affect the sensing performance since the signal variations caused by small finger tapping are much easier to be buried in the noise compared with those caused by large finger tapping. To evaluate how LASense can improve the performance for different tapping depths, we divide the results from the previous experiment into two groups according to the ground-truth measurements from the motion capture system, i.e., small tapping depth ($0 - 20$ *mm*) and large tapping depth ($20 - 40$ *mm*). As shown in Fig. 28b, LASense can still detect the small finger tapping at a high accuracy of 91.5%, and the accuracy further increases to 97.5% for the large tapping depth, demonstrating the robustness of our proposed system.

8.3.3 Impact of Phone Placement. We evaluate the performance of LASense for finger tapping sensing when the smartphone is mounted to a phone holder, placed on a table, and held in hand, as shown in Fig. 29a. The participant places the finger at 0.6 *m* away from the smartphone. As shown in Fig. 29b, the detection accuracies for different placements are 95.5%, 94.3% and 85.6%, respectively. The decreasing accuracy for the case where the smartphone is held in hand is caused by the signal distortion due to the unconscious movements of the hand.

8.3.4 Impact of User Diversity. To evaluate the impact of user diversity, we compute the median finger tapping detection accuracy for all fifteen participants. As we can observe from Fig. 30, the detection accuracies for most participants are larger than 90% except for Participant 4, Participant 5, and Participant 9. Through our careful analysis, we find that Participant 4 and Participant 9 have much smaller finger sizes. Therefore, the signals reflected from their fingers are weaker than those of others, resulting in poor performance. Furthermore, the poor performance for Participant 5 is caused by the shaking hand when she performs finger tapping. Therefore, the reflected signals contain the movement of both hand and finger, degrading the finger tracking performance.

8.4 Eye Blink Detection

In this section, we report our experiment results for eye blink detection. We empirically choose 40 *ms* as the sweep time according to our experiments. We adopt eye blink detection accuracy, which is defined as the probability that the eye blink can be detected among all tests to evaluate the system performance.



(a) Without LASense.

(b) With LASense

Fig. 31. The illustration for eye blink detection at 0.9 *m* with and without our system.

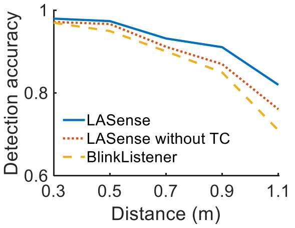
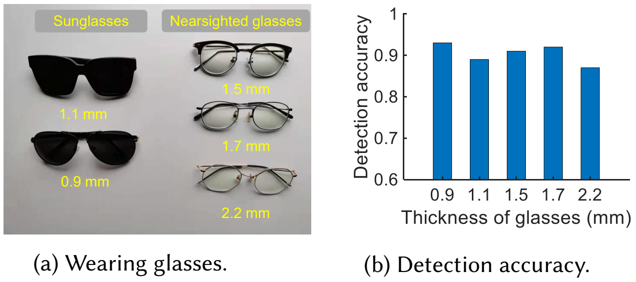


Fig. 32. Eye blink detection comparison.



(a) Wearing glasses.

(b) Detection accuracy.

Fig. 33. Impact of glasses.

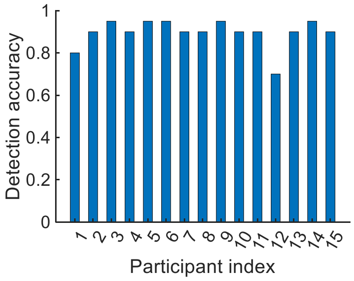


Fig. 34. Impact of user diversity.

8.4.1 Illustrative Experiment. To demonstrate the effectiveness of the proposed system, we extract the phase changes caused by eye blinks when the participant sits in front of the smartphone at 0.9 *m*. Fig. 31a shows the phase changes extracted from the ground truth bin without applying our approach, while Fig. 31b shows the phase changes extracted by LASense. To facilitate eye blink detection, we extract the envelope of the phase changes. We can observe four obvious peaks corresponding to four eye blinks after applying our approach.

8.4.2 Performance Comparison. To evaluate the effectiveness of our virtual transceiver idea, we compare the performance of LASense with the state-of-the-art system, i.e., BlinkListener [20], at different distances. Also, we compare the results of the proposed system with and without translating the coordinate (TC). We asked two participants to sit in front of the smartphone to naturally blink and varied the distance between the eye and smartphone from 0.3 *m* to 1.1 *m* at a step size of 0.2 *m*. As shown in Fig. 32, our proposed system can achieve a median eye blink detection accuracy of 91.1% even at 0.9 *m*, outperforming the sensing range of BlinkListener (0.5 *m*) by 80%. Compared with the other two human activities, i.e., respiration and finger tapping, eye blink obtains the minimum benefit from our proposed approaches. The reason is that the sub-millimeter level movement for eye blink is extremely subtle, and the reflection area is also very small (i.e., only several cm^2). To further improve the performance of eye blink detection, we need to apply beamforming techniques and/or deep learning algorithms.

8.4.3 Impact of Glasses. We evaluate the performance of eye blink detection when the participant wears two types of glasses with different thicknesses as shown in Fig. 33a. The smartphone is fixed on a phone holder which is 0.5 *m* away from the participant. As shown in Fig. 33b, we can detect eye blink at the accuracy of approximately 90% for all different thicknesses of glasses.

8.4.4 Impact of User Diversity. To evaluate the impact of user diversity, we compute the median eye blink detection accuracy for all fifteen participants. As we can observe from Fig. 34, the accuracies of eye blink detection for most participants are higher than 90%. According to BlinkListener, the variation of the detection accuracies is related to the eye size, which explains why we observe a lower accuracy for Participant 1. Furthermore, we find that the detection accuracy for eye blink also depends on head movement. For most participants, their heads can keep stationary when they blink their eyes. However, for Participant 12, her head tends to move forward and backward when her eyes blink, degrading the performance.

9 DISCUSSION

In this section, we discuss the applicability and limitations related to our work.

Applicability. Overall, the proposed system can effectively boost the range of acoustic sensing. However, we also observe dramatically different levels of improvement on different applications. For example, for respiration, the proposed system can increase the sensing range by 200%, whereas we only observed 80% of improvement for eye blink detection. This is because the virtual transceiver method boosts the sensing range by increasing the number of overlapped samples between the transmitted and received signals. The larger the distance is, the more samples can be overlapped. Therefore, the sensing range improvement is proportional to the distance between the target and the sensing device. The proposed system brings more improvement on larger-scale human activities, such as respiration, but less improvement on small-scale activities, such as eye blink detection.

Limitations. While the proposed system works well in single-target scenarios, the performance degrades when there are multiple close-by targets due to mutual interference. Multi-target sensing is a challenge in acoustic sensing, and we believe the transition from a single microphone to a microphone array could be a potential solution to enable multi-target sensing by spatially separating mixed reflections from multiple targets. Another limitation is that for fine-grained activity sensing, such as respiration monitoring, we require the target to be in stationary (e.g., sleeping) or quasi-stationary (e.g., standing or sitting) states for the proposed system to perform well. The proposed system can hardly work when the target performs vigorous movements (e.g., walking or running) because the small-scale respiration-induced chest displacement will be completely overwhelmed by the large-scale body movement. Extracting the subtle movement buried in large body movement is a meaningful yet challenging task that remains as an important future study.

10 CONCLUSION

This paper presents LASense, an acoustic-based system that pushes the range limit of fine-grained activity sensing. We demonstrate a signal processing approach to significantly increase the sensing range with only a single pair of speaker and microphone and showcase its applicability and reliability using three applications. To achieve this, we propose the idea of virtual transceiver to improve the sensing performance and design solutions to address issues associated with long-range acoustic sensing, such as differentiating targets from static objects and removing self-interference. Our demonstration of room-scale respiration monitoring based on LASense provides a new direction to enable smart health using COTS smart speakers. With an increasing number of activity sensing using acoustic signals, the proposed solution can be applied to other activity sensing systems. We believe the proposed methods can also benefit other sensing technologies using chirp-based signals such as LoRa.

ACKNOWLEDGMENTS

This work was partially supported by National Institutes of Health (NIH) under Award Number 5R01MH122371-02.

REFERENCES

[1] Fadel Adib and Dina Katabi. 2013. See through walls with WiFi!. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. 75–86.

[2] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 837–846.

[3] Anup Agarwal, Mohit Jain, Pratyush Kumar, and Shwetak Patel. 2018. Opportunistic sensing with MIC arrays on smart speakers for distal interaction and exercise tracking. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6403–6407.

[4] Apple. 2021. *iPhone 5c*. https://support.apple.com/kb/sp684?locale=en_US

[5] ARVICKA. 2021. *ARVICKA Computer Speaker*. https://www.amazon.com/ARVICKA-Computer-Multimedia-Smartphones-Projectors/dp/B01KC7WGQQ

[6] Bo Chen, Qian Zhang, Run Zhao, Dong Li, and Dong Wang. 2018. SGRS: A sequential gesture recognition system using COTS RFID. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 1–6.

[7] Cao Dian, Dong Wang, Qian Zhang, Run Zhao, and Yinggang Yu. 2020. Towards Domain-independent Complex and Fine-grained Gesture Recognition with RFID. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS (2020), 1–22.

[8] William Evans, Jakob Dyreby, Søren Bech, Slawomir Zielinski, and Francis Rumsey. 2009. Effects of loudspeaker directivity on perceived sound quality-a review of existing studies. In *Audio Engineering Society Convention 126*. Audio Engineering Society.

[9] Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. 2012. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1911–1914.

[10] Hexoskin. 2021. *Hexoskin Smart Garments*. https://www.hexoskin.com/

[11] Qualisys Inc. 2020. *Qualisys motion capture systems*. https://www.qualisys.com/hardware/miqus/

[12] Steven M Kay. 1993. *Fundamentals of statistical signal processing*. Prentice Hall PTR.

[13] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology* 49, 4 (2013), 764–766.

[14] Dong Li, Feng Ding, Qian Zhang, Run Zhao, Jinshi Zhang, and Dong Wang. 2017. TagController: A Universal Wireless and Battery-free Remote Controller using Passive RFID Tags. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 166–175.

[15] 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 18th Conference on Embedded Networked Sensor Systems*. 150–163.

[16] Tianxing Li, Qiang Liu, and Xia Zhou. 2016. Practical human sensing in the light. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. 71–84.

[17] Yichen Li, Tianxing Li, Ruchir A Patel, Xing-Dong Yang, and Xia Zhou. 2018. Self-powered gesture recognition with ambient light. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 595–608.

[18] Jie Lian, Jiadong Lou, Li Chen, and Xu Yuan. 2021. EchoSpot: Spotting Your Locations via Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–21.

[19] Kang Ling, Haipeng Dai, Yuntang Liu, and Alex X Liu. 2018. Ultragesture: Fine-grained gesture sensing and recognition. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[20] Jialin Liu, Dong Li, Lei Wang, and Jie Xiong. 2021. BlinkListener: " Listen" to Your Eye Blink Using Your Smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–27.

[21] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. DeepRange: Acoustic Ranging via Deep Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[22] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-based room scale hand motion tracking. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[23] MiniDSP. 2021. *UMA-8-SP USB mic array*. https://www.minidsp.com/products/usb-audio-interface/uma-8-sp-detail

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

[25] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. Covertband: Activity information leakage using music. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.

[26] Alan V Oppenheim, John R Buck, and Ronald W Schafer. 2001. *Discrete-time signal processing. Vol. 2.* Upper Saddle River, NJ: Prentice Hall.

[27] Vaughan Pratt. 1987. Direct least-squares fitting of algebraic surfaces. *ACM SIGGRAPH computer graphics* 21, 4 (1987), 145–152.

[28] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1574–1582.

[29] Wenjie Ruan, Quan Z Sheng, Lei Yang, Tao Gu, Peipei Xu, and Longfei Shangguan. 2016. AudioGest: enabling fine-grained hand gesture detection by decoding echo signal. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 474–485.

[30] Ronald W Schafer. 2011. What is a Savitzky-Golay filter?[lecture notes]. *IEEE Signal processing magazine* 28, 4 (2011), 111–117.

[31] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. 2020. SpiroSonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[32] Ke Sun, Wei Wang, Alex X Liu, and Haipeng Dai. 2018. Depth aware finger tapping on virtual displays. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 283–295.

[33] Cozmo Team. 2020. *Cozmo Smart Robot.* https://https://www.digitaldreamlabs.com/

[34] Yu-Chih Tung, Duc Bui, and Kang G Shin. 2018. Cross-platform support for rapid development of mobile acoustic sensing applications. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 455–467.

[35] Raghav H Venkatnarayan and Muhammad Shahzad. 2018. Gesture recognition using ambient light. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–28.

[36] VLIKE. 2021. *VLIKE LCD Digital Sound Level Meter.* https://www.amazon.com/VLIKE-Digital-Measurement-Measuring-Function/dp/B01N2RLJ32

[37] Anran Wang and Shyamnath Gollakota. 2019. Millisonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[38] Anran Wang, Jacob E Sunshine, and Shyamnath Gollakota. 2019. Contactless infant monitoring using white noise. In *The 25th Annual International Conference on Mobile Computing and Networking*. 1–16.

[39] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–20.

[40] 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.

[41] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of Acoustic Gesture Recognition. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 566–575.

[42] Yadong Xie, Fan Li, Yue Wu, and Yu Wang. 2021. HearFit: Fitness Monitoring on Smart Speakers via Active Acoustic Sensing. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[43] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*. 15–28.

[44] Youwei Zeng, Dan Wu, Jie Xiong, Jinyi Liu, Zhaopeng Liu, and Daqing Zhang. 2020. MultiSense: Enabling multi-person respiration sensing with commodity wifi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[45] Youwei Zeng, Dan Wu, Jie Xiong, Enze Yi, Ruiyang Gao, and Daqing Zhang. 2019. FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.

[46] Fusang Zhang, Zhi Wang, Beihong Jin, Jie Xiong, and Daqing Zhang. 2020. Your Smart Speaker Can" Hear" Your Heartbeat! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–24.

[47] Jinshi Zhang, Qian Zhang, Dong Li, Run Zhao, and Dong Wang. 2017. RFlow-ID: Unobtrusive Workflow Recognition with COTS RFID. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 333–342.

[48] Qian Zhang, Dong Li, Run Zhao, Dong Wang, Yufeng Deng, and Bo Chen. 2018. RFree-ID: An unobtrusive human identification system irrespective of walking cofactors using cots RFID. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[49] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[50] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–30.

[51] Run Zhao, Dong Wang, Qian Zhang, Xueyi Jin, and Ke Liu. 2021. Smartphone-based Handwritten Signature Verification using Acoustic Signals. *Proceedings of the ACM on Human-Computer Interaction* 5, ISS (2021), 1–26.