Direction-Aware, Audio-Based Pedestrian Relative Positioning by Swing Induced Doppler Shift

Liang Wang
State Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing 210023, China
wl@nju.edu.cn

Tao Gu
School of Computer Science
and Information Technology
RMIT University, Australia
tao.gu@rmit.edu.au

Xianping Tao
State Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing 210023, China
txp@nju.edu.cn

Jian Lu
State Key Laboratory for
Novel Software Technology
Nanjing University, Nanjing 210023, China
lj@nju.edu.cn

ABSTRACT

In this paper, we study the problem of pedestrian relative positioning with respect to their walking direction. Existing approaches are mainly based on trajectory information or device proximity detection, and they highly rely on infrastructure or specialized device support. Importantly, most work does not provide relative position information with respect to people's walking direction. To address the above issues, we propose a direction-aware, audio-based solution that only uses daily wearable devices. Based on the fact that pedestrian's arms often swing back and forth during walking, we develop the wrist-body model that formally models the distance change between a user's wrist and his/her walking mate's body when walking together. Based on this model, we design our system by attaching the audio sources to a user's wrists and an audio receiver to the other user's body. We develop key indicators that characterize the received audio signal's Doppler shift induced by arm swing motions and the differences in signal strength. We further propose methods such as cycle segmentation and aggregation to deal with several real-world challenges. The performance of our approach is studied through extensive experiments. Evaluation conducted using real-world data suggests the prototype system achieves 85.9% positioning accuracy, demonstrating its effectiveness.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**;

KEYWORDS

Smartphone and Smartwatch, Relative Positioning, Audio, Doppler Shift

ACM Reference format:
Liang Wang, Tao Gu, Xianping Tao, and Jian Lu. 2017. Direction-Aware, Audio-Based Pedestrian Relative Positioning by Swing Induced Doppler Shift. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Melbourne, VIC, Australia, November 7–10, 2017 (MobiQuitous 2017),* 10 pages.
https://doi.org/10.1145/3144457.3144469

1 INTRODUCTION

The relative positions of pedestrians walking together are affected by many factors. For example, a recent study show that people intentionally walk together tend to form a horizontal formation with respect to their walking direction to facilitate social interactions [10]. When walking in a crowded environment, the pedestrians in a group often form a 'V'- or 'U'-shaped formation to avoid interfering with other pedestrians out of the group [10]. And when walking through a narrow bottleneck area, people tend to walk in a queue due to lack of space [4]. In many cases, people also intentionally maintain their relative positions when walking together, e.g., matching in a square, protecting a VIP, following the leader, etc. As a result, by tracking the relative positions of pedestrians walking together, we can support many applications including but not limited to formation detection [18], social relation and structure analysis [3], crowd dynamics studies [10], etc.

In this paper, we study the problem of relative positioning for pedestrians. More specifically, considering the case that two people are walking together, we aim to determine their relative position as one person walks on the other one's *front*, *back*, *left*, or *right* with respect to their walking direction. This problem is challenging mainly due to the following reasons. **Walking direction awareness**—relative positions such as the *front* and the *back* are defined with respect to the walking direction. So the pedestrians' walking direction must be considered when performing relative positioning. **Infrastructure independency**—people can walk freely in different places which may be out of the coverage of devices deployed in the infrastructure. Hence, a mobile solution is desired.

MobiQuitous 2017, November 7–10, 2017, Melbourne, VIC, Australia
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5368-7/17/11…$15.00
https://doi.org/10.1145/3144457.3144469

Local processing—it is desirable to process the data locally without constant communication and synchronization between devices and servers to achieve high efficiency and scalability.

Existing work on this problem mainly falls into two categories—trajectory- and proximity-based approaches. For trajectory-based approaches, relative positioning can be done by comparing trajectory data for different pedestrians obtained by GPS [24], WiFi [20], RFID [7], or audio [23]. However, this approach is limited because of infrastructure dependency and limited accuracy. For proximity-based approaches, relative positions can be determined by the distance measurement between devices using audio [25] or RF [2] signals. However, this approach is also limited for not providing relative position information with respect to the pedestrians' walking direction. Infrared tags are used in [11] to detect the headings and relative positions of different users which rely on specialized device and global information. In [14, 15], the authors propose to use the Doppler shift of audio pulses for direction finding. However, they cannot provide continuous relative positioning service for relying on specialized user motions such as drawing a cycle. Audio-based direction finding and device tracking is studied in [5, 6, 23] which uses the Doppler shift of fixed audio anchors as sources to track the mobile receivers. However, they assume the audio anchors are stationary so that the local velocity of the mobile receiver represents the velocity of the relative movement between the source and the receiver [5, 6]. Moreover, since the audio sources are stationary, they transmit audio signals continuously without providing any information about the movements of the sources, which is not applicable in our scenario.

In this paper, we propose a novel direction-aware, audio-based relative positioning approach that uses COTS wearable devices. We develop the wrist-body model that formally models the distance change between a user's wrist and his/her walking mate's body when walking together. We design our system by using smartwatch as the audio source and smartphone as the audio receiver. Key indicators are developed to characterize the patterns of frequency and strength differences of the received audio signal for different relative positions. We develop the cycle segmentation and aggregation methods to address real-world challenges such as weakened signal caused by many reasons. Finally, we regard the positioning problem as a classification problem and propose a feature-based positioning approach. We also propose an alternative deployment strategy that attaches the audio sources to the user's ankles to address the issue that arms may not swing during walking. In summary, this paper makes the following contributions.

(1) We propose a novel approach that uses swing induced Doppler shift to perform relative positioning with respect to pedestrians walking direction.
(2) A working prototype system is implemented and a series of studies under different settings are conducted to validate our theory.
(3) Cycle segmentation and aggregation methods are proposed to address the real-world issues.
(4) We collect data in real-world settings and evaluate our system's performance using real data.

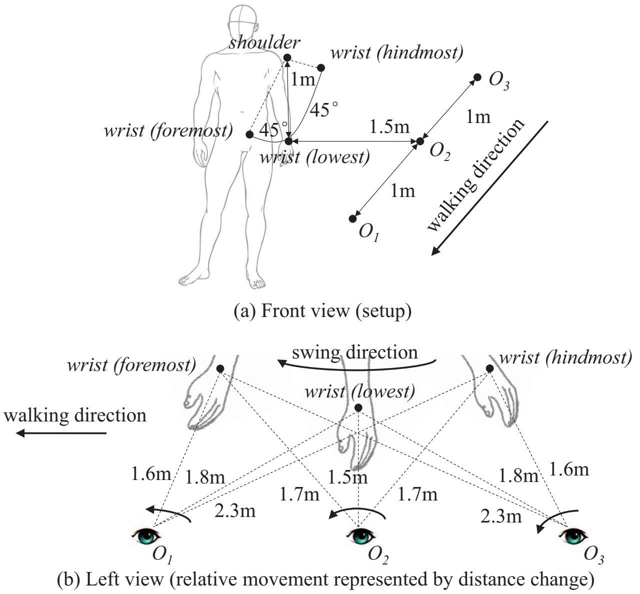
The rest of the paper is organized as follows. We summarize the related work in Sec. 2. Sec. 3 presents a motivating example and the system's design choices. Prototype implementation is introduced in Sec. 4, based on which we conduct preliminary studies in Sec. 5. Sec. 6 introduces the design of the data processing pipeline. The performance of our system is evaluated in Sec. 7. Finally, Sec. 8 concludes the paper.

2 RELATED WORK

Localization for mobile targets has attracted much research interest recently. In [24], the authors explore large-scale human mobility patterns by GPS, cellular, and ad hoc network data. While GPS-based approaches are often considered unavailable for indoor environments and limited in accuracy, RF-based approaches are studied. A large body of literatures have explored WiFi signals for localization and positioning [20]. Other systems like Tagoram [21] explore RFID technologies for tracking. Different from RF-based approaches, recent studies have used smartphone's built-in sensors [19, 22], visible light [7], or audio signals [6] for localization and tracking. Relative positioning is possible if the detailed trajectories of different subjects can be obtained. However, the above approaches often rely on anchor devices, e.g., WiFi APs [20], RFID readers [13], light sources [7], or audio anchor nodes [5, 6, 23] to be present at the environment, limiting their detection coverage. Also, the locations must be sent to a centralized server for analysis [9], which leads to high communication overheads.

Different from localization technologies, proximity detection technologies measure the distance between different objects. Acoustic ranging (AR) has been widely studied in mobile computing [12, 16, 25] which mainly rely on the signal's time-of-arrival (TOA). Other approaches perform proximity detection base on RF signals from cellular networks [8], Bluetooth radio [2], or sensor networks [1]. While proximity detection technologies can indicate whether two people are close to each other, they cannot provide relative position information with respect to the walking direction.

Some recent work has explored the Doppler shift of audio signals for direction finding and tracking [5, 6, 14, 15, 23]. In [14, 15], the authors propose to use the Doppler shift of audio pulses for direction finding. However, they rely on specialized user motions such as drawing a cycle. They do not explore the periodic nature of arm swing motion for signal processing, and cannot provide continuous relative positioning service since they rely on unnatural, specialized motions. Moreover, data communication is required in their approaches to compare the sending and receiving pulse intervals to perform direction finding. Direction finding and device tracking is studied in [5, 6, 23] which use the Doppler shift of fixed audio anchors as sources to track the mobile receivers. However, their approaches are not applicable to our scenario because they assume the audio anchors are stationary so that the local velocity of the mobile receiver represents the velocity of the relative movement between the source and the receiver [5, 6]. Moreover, since the audio sources are stationary in these approaches [5, 6, 23], they transmit audio signals continuously without providing any information about the movements of the sources, which is not applicable in our scenario as discussed above.



(a) Front view (setup)

(b) Left view (relative movement represented by distance change)

Figure 1: Motivating example.

3 MOTIVATING EXAMPLE AND DESIGN CHOICES

Our approach is built upon the fact that people's arms often swing back and forth when walking. As shown in Fig. 1(a), consider three observers with fixed relative positions with respect to the user's body that are in front of (O_1), aligned with (O_2) and behind (O_3) the user with 1m apart from each other, a horizontal distance of 1.5m apart from the body, and the same distance above the ground as the wrist's lowest position. We assume the user's arm is 1m in length and swings between 45 degrees back and front against the body when walking. As shown in Fig. 1(b), when the arm swings from the hindmost point to the foremost point, for O_1 in front of the user, the relative movement between the wrist and O_1 causes their distance to **decrease** from 2.3m to 1.6m. For O_2 aligned with the user, the distance **first decreases** from 1.7m to 1.5m **then increases** to 1.7m. For O_3 behind the user, the distance **increases** from 1.6m to 2.3m. This observation suggests that we can perform direction-aware, relative positioning for the observers from the patterns of relative movement represented by distance change.

Motivated by this example, relative positioning for the observer can be done by finding the patterns of relative movement between the user's wrist and the observer which leads to two design options. For the first option, we can track the detailed relative movement by constantly measuring the distance between the user's wrist and the observer. However, precise distance measurement down to the centimeter level is technically challenging [16] and unnecessary in our case since we only need the direction of relative movement to find the patterns. As a result, we propose the second option which is based on the Doppler shift of the received signal's frequency caused by relative movement between a signal source and the observer which can be written as:

$$f = (1 + \frac{\Delta v}{c}) f_0 \tag{1}$$

where f is the received frequency from the observer, f_0 is the emitted frequency from the source, c is the velocity of the signal wave, and Δv is the velocity of the observer moving relatively to the source, with $\Delta v > 0$ if the observer is moving towards the source.

Following this idea, if we attach a signal source to the user's wrist emitting a signal wave at fixed frequency f_0 when the wrist swings from the hindmost point to the foremost point and a receiver to the observer, the relative movement between the source and the receiver will result in distinctive patterns in the received signal's frequency f caused by the Doppler shift according to the receiver's relative positions. Specifically, we can expect: 1) $f > f_0$ for O_1; 2) $f > f_0$ first and $f < f_0$ later for O_2; and 3) $f < f_0$ for O_3. Note that it is important to restrict the signal transmission period to the period when the wrist swings from the hindmost point to the foremost point. If the signal source is in continuous transmission mode [5, 6, 23], the pattern received at the observer will be $f > f_0$ and $f < f_0$ appear alternatively regardless its position. Relative positioning is then impossible unless the source provides additional information on the period and start time of the swing motion, which leads to intensive data communication and reduces the system's efficiency. In this case, such communication can be avoided if we attach the audio source to the observer and the receiver to the wrist. However, as we will show next, this design option leads to another form of intensive data communication when discriminating *left* from *right*.

So far we can only identify three relative positions, i.e., *front* (O_1), *back* (O_3), and *align* (O_2). It is still difficult to discriminate *left* from *right* since they have the same frequency changing pattern as *align*. We present a simple solution in this work to have two audio sources attached to the users left and right wrists, respectively. Due to the users body blockage, when the observer is on the user's left, the audio signal from user's left wrist will be stronger than that from the right wrist and vise versa. As discussed above, another design option is to attach the audio source to the observer and the receivers to the wrists. However, if we need to compare the signal strength between the left and right wrists to discriminate *left* from *right*, data communication between the receivers is constantly required. Instead, if we follow our original design, we only need to separate the signal from the left and right wrists by transmitting at different frequencies. Computation can then be done locally at the receiver without additional communication.

In summary, by using the patterns described above, it is theoretically possible to discriminate different relative positions using the received audio signal. The advantages of this approach include: 1) walking direction information is naturally contained in the frequency changing patterns; 2) no infrastructure support is needed; 3) once the emission frequency is established between the sender and receiver, the received signal can be processed locally at the receiver without any further communication. While all kinds of signals travel in wave forms can be used, we focus on audio signal instead of RF signal for its slow propagation speed and device simplicity.

4 PROTOTYPE IMPLEMENTATION

We now introduce our prototype implementation following the above system design.

4.1 Hardware

There are two types of hardware used to build our system—the audio source and the receiver. For the audio source, we use the Cross Country Smartwatch[1] equipped with a speaker and multiple onboard sensors including a gyroscope. It has a dual-core 1.5GHz CPU and 1GB RAM. It runs the Android 4.2 OS and can be easily programmed using existing Android SDK. For the audio receiver, we use the Nexus 5 Smartphone[2] powered by a 2.26GHz quad-core CPU, 2GB RAM, and the Android 4.4 OS. The built-in microphone has a maximum sampling rate of 44.1kHz, capable of sensing audio signals up to 22kHz in frequency.

We ask one user to put on two smartwatches, one for each wrist. We then ask the other user to carry the smartphone on the body in a pocket.

4.2 Smartwatch-side System

As introduced above, the smartwatch functions as the audio source that emits a signal on frequency f_0 when the wrist swings from the hindmost point to the foremost point. To determine the periods of wrist swinging forward and backward, we use the smartwatch's gyroscope readings. When worn on the left wrist, the smartwatch is doing clockwise and counterclockwise rotations along its z axis when the wrist is swinging forward and backward, respectively. When worn on the right wrist, the case is opposite. For the smartwatch used in our system, the gyroscope readings are positive when doing clockwise rotation and negative in the case of counterclockwise rotation. As a result, it is straightforward to determine the periods of wrist swinging forward as the gyroscope readings on the z axis is positive for the left wrist (negative for the right wrist). Small vibrations are eliminated by filtering out readings with small rotation values below 5% of the maximum value in history. On detecting the user's wrist is swinging forward, we trigger the smartwatch's speaker to play a previously synthesized audio file which only contains a tone at frequency f_0. We stop signal transmission on detecting the user's wrist starts to swing backward. Experiment result suggests gyroscope is sufficiently accurate and agile to detect swing direction changes, and outperforms the accelerometer.

To discriminate the sources, we set different f_0 for the smartwatches worn on the left and right wrists. Specifically, in our prototype system, we set $f_{0_l} = 17kHz$ for the left wrist, and $f_{0_r} = 16kHz$ for the right wrist. We keep the two frequencies $1kHz$ apart so that the frequency shift cause by the Doppler effect will not affect our judgment on the source of the signal. Also, we choose to use frequencies higher than $16kHz$ so that they are mainly inaudible for people.

4.3 Smartphone-side System

The smartphone-side system implements the data processing pipeline which can perform online relative positioning which we introduce in detail in Sec. 6.

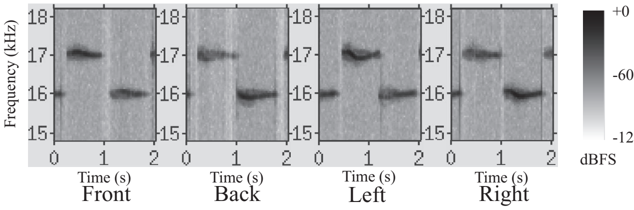


Figure 2: Spectrogram.

5 PRELIMINARY STUDIES

In this section, we conduct a series of preliminary studies using our prototype system to study its effectiveness under different conditions.

5.1 Controlled Indoor Environment

This study is conducted in a controlled lab environment. A subject wears the smartwatches on both wrists and walks in place with his arms swinging naturally. The smartphone, as the receiver, is placed 1.3m above the ground on a shelf, 1.5m apart from the subject. Audio data are collected from the four positions with respect to the subject's facing.

5.1.1 *Observations.*

Fig. 2 plots the signal's spectrogram over time in one swing cycle (left wrist swings from the hindmost point to the foremost point and back to the hindmost point). As shown by the figure, clear patterns matching our theory can be observed. When the smartphone is in front of the subject, the frequency of the received signal is generally higher than the emitted signal (17kHz for the left wrist and 16kHz for the right wrist). Opposite observations can be made when the smartphone is placed in the back. When the smartphone is aligned with the subject (to the left or the right), the received signal frequency first raises above then drops below the emitted signal frequencies.

5.1.2 *Key Indicators.*

From the above observations, we derive the following two key indicators to characterize the signal.

High-to-Low Ratio (HLR). With respect to the emitted frequency f_0 (f_{0_l}=17kHz and f_{0_r}=16kHz for the left and right wrist, respectively), *HLR* is the ratio of aggregated amplitude of frequency components above f_0 over those below f_0 over time during a swing cycle. Specifically, *HLR* is computed by:

$$HLR(t_s, t_e) = \frac{SUM_{high}(t_s, t_e) - SUM_{low}(t_s, t_e)}{SUM_{high}(t_s, t_e) + SUM_{low}(t_s, t_e)} \cdot scale \tag{2}$$

where t_s and t_e are the time when the current swing cycle starts and ends, respectively, $scale$ is a positive real number that scales the *HLR* value within range $[-scale, +scale]$,SUM_{high} and SUM_{low} are two functions that accumulate the amplitudes across frequencies above and below f_0 over time period $[t_s, t_e]$, which are computed as:

$$SUM_{high}(t_s, t_e) = \sum_{t=t_s}^{te} \left[\sum_{f=f_{0_l}}^{f_{0_l}+\Delta f} amp_{f,t} + \sum_{f=f_{0_r}}^{f_{0_r}+\Delta f} amp_{f,t} \right] \tag{3}$$

[1] https://www.youtube.com/watch?v=IpkVNnNl8Ds

[2] https://www.google.com/nexus/5/

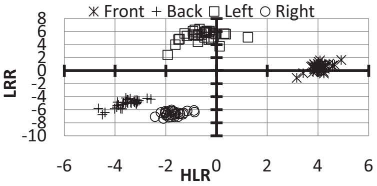


Figure 3: Instances.

	Front	Back	Left	Right
Front	0.0	164.6	93.3	307.6
Back	290.0	0.0	76.4	35.7
Left	187.0	86.6	0.0	466.7
Right	160.6	56.1	137.0	0.0

▨ 352–469
▨ 235–351
▨ 118–234
▨ 1–117

Figure 4: Pair-wise KLD, overall average is 171.8.

		Classified as			
		Front	Back	Left	Right
Groundtruth	Front	34	0	0	0
	Back	0	31	1	1
	Left	1	0	31	1
	Right	0	0	0	31

Figure 5: Confusion matrix, accuracy = 96.9%.

$$SUM_{low}(t_s, t_e) = \sum_{t=t_s}^{te} \left[\sum_{f=f0_l-\Delta f}^{f0_l} amp_{f,t} + \sum_{f=f0_r-\Delta f}^{f0_r} amp_{f,t} \right] \quad (4)$$

where Δf is the single-side frequency range for aggregation, $amp_{f,t}$ is the amplitude of signal's component at frequency f at time t, we take the discrete form in the above equations because we obtain $amp_{f,t}$ using the Fast Fourier Transform (FFT) over the data frame obtained by a short sliding window starting at time t.

Left-to-Right Ratio (LRR). *LRR* is the ratio of the aggregated signal amplitude over time that is from the left audio source over that is from the right audio source. Specifically, *LRR* is computed by:

$$LRR(t_s, t_e) = \frac{SUM_{left}(t_s, t_e) - SUM_{right}(t_s, t_e)}{SUM_{left}(t_s, t_e) + SUM_{right}(t_s, t_e)} \cdot scale \quad (5)$$

where t_s, t_e, and *scale* are the same as defined in Eq. (2), SUM_{left} and SUM_{left} accumulate the amplitudes for the left and right sources which are computed as:

$$SUM_{left}(t_s, t_e) = \sum_{t=t_s}^{te} \sum_{f=f0_l-\Delta f}^{f0_l+\Delta f} amp_{f,t} \quad (6)$$

$$SUM_{right}(t_s, t_e) = \sum_{t=t_s}^{te} \sum_{f=f0_r-\Delta f}^{f0_r+\Delta f} amp_{f,t} \quad (7)$$

all the notations used are the same as defined above.

Given the above two key indicators, we plot the instances (with each one representing a swing cycle) obtained from the simulation data in Fig. 3. We manually determine the time period for each swing cycle, and set $scale = 10$, $\Delta f = 200Hz$, and a sliding window length of 0.1s when computing *HLR* and *LRR*. From this figure, it is clear that we can discriminate the four relative positions from the two indicators. While the *HLR* for the *left* and the *right* positions are close to zero, they are well separated by the *LRR* as expected. The *front* and the *back* positions are separated by both the *HLR* and the *LRR*. The *LRR* for the *front* (*back*) is above (below) zero because of the orientations of speakers on different wrists as explained above.

5.1.3 KL Divergence Used for Analysis. To quantify the distribution of instances for different relative positions in the *HLR-LRR* plane as shown in Fig. 3, we model each position's data as a 2D Normal Distribution and use the pair-wise KL Divergence (KLD) as the indicator. KLD evaluates the dissimilarity between two distributions and has shown to be an effective indicator for discriminative power [17]. Since KLD is asymmetric, computation is done between every pair of positions as shown in Fig. 4.

To understand how KLD is related to discriminative power, we conduct a simple classification test using the above instances. We use a C4.5 decision tree as the classifier and perform a ten-fold cross-validation. Fig. 5 plots the confusion matrix and the overall classification accuracy is 96.9%. By comparing the results shown in Fig. 4 and Fig. 5, it is clear that miss classifications occur between positions that have low KLD. For example, the pair *back-right*, and the pair *back-left*.

In summary, the overall accuracy of 96.9% suggests the proposed approach is effective to discriminate different relative positions from the received audio signal. However, in real-life, there are many factors that may affect the performance of our system. Next, we conduct a series of experiments to study the effect of different factors on the performance of our approach.

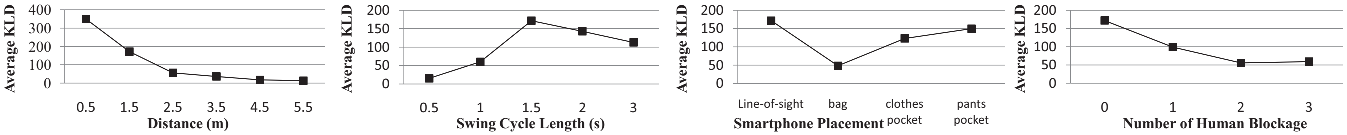
5.2 Under Different Conditions

In this section, we present the results of studies under different conditions. Since KLD has shown to be an effective indicator for discriminative power, we use the average KLD for analysis. The same settings as the simulation study are used when computing the indicators.

5.2.1 Source-receiver Distance. In real-life, two people walking together are often close to each other. However, there is no bound on the distance between two pedestrians. In this experiment, we aim to find out how distance affects the quality of the received audio signal and the performance of the proposed approach. To eliminate the effect of other factors, we repeat the simulation study and only change the distance between the source and the receiver.

Fig. 6(a) plots the average pair-wise KLD computed using data collected with the distance increasing from 0.5m to 5.5m. As shown in the figure, the discriminative power decreases as the distance increases. This is probably caused by the reduced signal strength after traveling for a longer distance. Theoretically, universal reduction in signal strength has no impact on the computation for the indicators. In practice, however, due to the microphone's limited sensibility, the weakened signal can only be identified partially in one cycle. The incomplete and unstable signal received causes *HLR* and *LRR* to vary from cycle to cycle even for the same position, resulting in decreased discriminative power.

5.2.2 Swing Frequency. People may walk in different speeds that result in different arm swing frequencies. This will influence the velocity of the receiver moving relatively to the source and affect the frequency shift of the received audio signal. In this experiment, we study the performance of our approach under different swing frequencies.



(a) Source-receiver distance (b) Swing frequency (c) Smartphone placement (d) Human body blockage

Figure 6: Average KLD affected by different factors.

Fig. 6(b) plots the average pair-wise KLD with different swing frequencies represented by swing cycle length. The data are the most discriminative when swing cycle length is 1.5s. The discriminative power drops when cycle length further increases. This is because the slower the arm swings, the velocity of relative movement between the signal source and the receiver also decreases, which will result in a less significant Doppler shift. However, it is surprising to find that further decrease in swing cycle length also reduces the discriminative power. Further study reveals that because we use a non-overlapping sliding window of 0.1s to segment the audio stream, a shorter cycle increases the cost of imprecise cycle segmentation. For a cycle length of 0.5s, a 0.1s mistake in cycle segmentation will affect 20% of the data, causing large variances in computed indicators and reducing the discriminative power similar to the weakened signal does. A higher time resolution is not a solution because it reduces the frequency resolution which is critical in our system.

This result suggests swing frequency affects the system's performance by two reasons: a) imprecise segmentation; b) less significant Doppler shift. By comparing the results of swing cycle length larger and less than 1.5s, it is clear that the former does more harm to the discriminative power.

5.2.3 Smartphone Placement. In daily life, smartphone may be placement in different positions during walking, e.g., in a pocket or a bag. When placed in a pocket or a bag, the sound is muffled, which may affect the quality of the received audio signal. In this study, we evaluate how smartphone placement affects our system's performance.

Fig. 6(c) plots the average pair-wise KLD when the smartphone is placed in different objects. It suggests the discriminative power drops when the sound is muffled, especially when placed in a bag. Similar to the condition of large source-receiver distance, the signal is weakened when placed in an object, resulting in unstable *HLR* and *LRR* among cycles, which leads to a lower discriminative power.

5.2.4 Human Body Blockage. In real-life, it is possible that multiple pedestrians are walking together and human body may block the signal. Fig. 6(d) plots the average KLD when blocked by different number of human bodies. Comparing to the case of no human blockage (line-of-sight), the discriminative power drops when blocked by human bodies. Similar to the above cases with large source-receiver distance and smartphone placed in objects, human blockage will reduce the received signal strength and make it unstable. Human body blockage is expected and useful to us because we use this feature to discriminate *left* from *right* as introduced above.

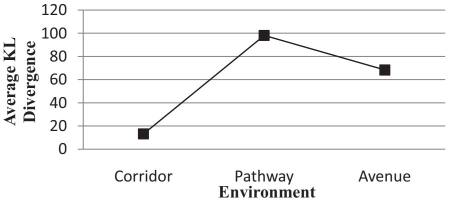


Figure 7: Average KLD varies with environment.

5.2.5 Summary. The above studies suggest that the performance of the proposed approach can be affected by many factors. The main reason behind the decreased discriminative power is the drop of signal strength resulting in incomplete and unstable signal received in each cycle which causes the indicators to vary largely from cycle to cycle. A possible solution is to strengthen and stabilise the signal as shown next.

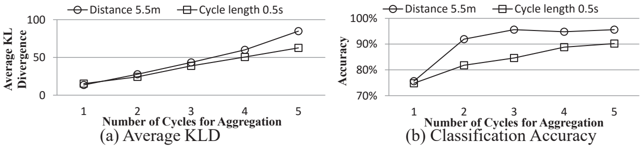
We also test the system's performance under different environmental noise levels. The result suggests that environmental noises do not have an significant impact on the quality of data. All kinds of noise in the audible range, involving talking, tyre and wind noises caused by running vehicles will not affect our signal in the inaudible range.

5.3 Impact of Different Environments

While the above studies are conducted in a controlled lab environment, in this experiment, we evaluate our system's performance in real walking scenarios. Two subjects are involved with one subject (subject *A*) wearing the smartwatches while the other subject (subject *B*) carrying the smartphone in the clothes' pocket. Data are collected in three different environments: a) in a corridor in the lab building; b) on a road on the campus with people and cars occasionally passing by; and c) beside a busy avenue with lots of vehicles.

Fig. 7 shows the average KLD of data collected when walking in different environments. Surprisingly, the data collected in the quiet indoor corridor have the lowest discriminative power while the data collected in outdoor environments still seem to have good quality. Detailed analysis is as follows.

During the experiment, the smartphone is always in subject *B*'s right clothes pocket. Complex effects involving human body blockage and muffling significantly reduce the received audio signal's strength, especially when subject *B* is walking on subject *A*'s right. This combined weakening effect is applicable to all environments, and makes the indicators to vary among cycles as we discussed above. We also find in real-world, *HLR* seems to be more harmed by



(a) Average KLD (b) Classification Accuracy

Figure 8: Effect of cycle aggregation.

the weakened signal than *LRR*. *LRR* which characterizes the blockage effect caused by subject *A*'s own body, is less affected by the weakening effects above. This explains why the data still yield relatively good discriminative power in the outdoor environments.

However, when walking in the narrow (1.7m) corridor in the lab building, the walls on both sides will reflect the emitted audio signal. For example, when subject *B* with the smartphone is walking on the right of subject *A*, the audio signal emitted by subject *A*'s left wrist, which is supposed to be partially blocked by *A*'s body, is reflected by the walls and transmits into the microphone following multiple paths. As a result, the discriminative power of *LRR* also drops in this case.

Another observation from the data is that environmental noises do not affect the quality of our data as significantly as expected. All kinds of noise in the audible range, involving talking, tyre and wind noises caused by running vehicles will not affect our signal mainly in the inaudible range. Currently, the only observed type of noise that may affect our signal is the clothes friction noise from subject *B* when his arms are also swinging when walking. However, the friction noise is similar to a short-term white noise across all frequencies, causing a roughly equal effect on the received signal's high and low frequency components in a very short duration.

5.4 Cycle Aggregation

As discussed above, a major threat to the performance of the proposed approach is the incomplete and unstable signal received caused by many reasons.

The cycle aggregation method is designed to strengthen and stabilise the signal by the periodic nature of arm swinging. To achieve this, we aggregate the current cycle's data with the data obtained in previous cycles. More specifically, given the current cycle's start and end time t_{s_c} and t_{e_c}, *HLR* and *LRR* with cycle aggregation are computed as:

$$HLR(t_{sc}, t_{ec}, n) = \frac{\sum_{i=c-n}^{c} [SUM_{high}(t_{s_i}, t_{e_i}) - SUM_{low}(t_{s_i}, t_{e_i})]}{\sum_{i=c-n}^{c} [SUM_{high}(t_{s_i}, t_{e_i}) + SUM_{low}(t_{s_i}, t_{e_i})]} \cdot scale \tag{8}$$

$$LRR(t_{sc}, t_{ec}, n) = \frac{\sum_{i=c-n}^{c} [SUM_{left}(t_{s_i}, t_{e_i}) - SUM_{right}(t_{s_i}, t_{e_i})]}{\sum_{i=c-n}^{c} [SUM_{left}(t_{s_i}, t_{e_i}) + SUM_{right}(t_{s_i}, t_{e_i})]} \cdot scale \tag{9}$$

where n is the number of cycles for aggregation, t_{s_i} and t_{e_i} are the start and end time of the i-th cycle, other notations are the same as defined in Sec. 5.1.2.

To study the effectiveness of the cycle aggregation method, we evaluate its performance using data from cases with average KLD below 20. More specifically, we choose the traces collected with distance 5.5m in Sec. 5.2.1 and swing cycle length of 0.5s in Sec. 5.2.2. We also evaluate the discrimination accuracy using a C4.5 decision tree model. Fig. 8 plots the average KLD and classification accuracy changed by aggregating the signal with different number of cycles. The result suggests the cycle aggregation method is efficient in increasing the discriminative power.

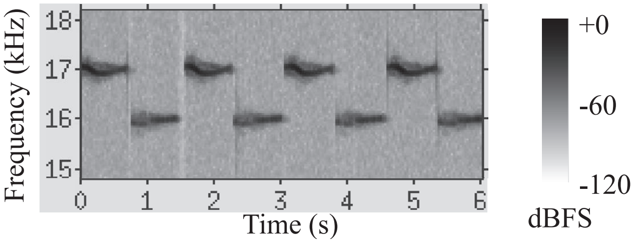


Figure 9: Signal repeats over time.

6 DATA PROCESSING PIPELINE

We have shown the effectiveness of our stage one prototype system in the above studies. However, manual efforts such as cycle segmentation are involved. In this section, we move on to stage two of system design and implementation for a fully automated data processing pipeline for the smartphone. The pipeline mainly involves four steps: sliding window-based segmentation, FFT, cycle segmentation and aggregation, and positioning. While sliding window-based segmentation and FFT are routine procedures for data preprocessing, we mainly focus on the automated cycle segmentation and feature-based positioning methods, in this section.

6.1 Data Preprocessing

On receiving the audio stream from the smartphone's microphone, we first apply a 0.1s non-overlapping sliding window to segment the stream into frames. For each frame, we apply FFT to obtain the amplitudes of the signal's different frequency components. The sampling rate of the smartphone's microphone is 44.1kHz, and there are 4410 samples in each frame. We perform an 8192 point FFT to obtain a fine-grained spectrogram with each frequency bin sized 5.4Hz.

The above data preprocessing steps generate a series of frequency representations of frames, which we use as input for the following steps.

6.2 Automated Cycle Segmentation

After obtaining the series of frames, we perform cycle segmentation to determine the start and end time of each cycle so that indicators like *HLR* and *LRR* can be computed.

Cycle segmentation is done based on the fact that the received audio signal periodically repeats itself over time, with a period equal to the cycle length. Fig. 9 shows the audio signal received over time which includes four cycles. It is clear that the signal is periodic with a period of approximately 1.5s. By definition, a discrete time signal is periodic if for any n, we have $x[n] = x[n+N_0]$, where $N_0 > 0$ is the minimum value that satisfies the equation called

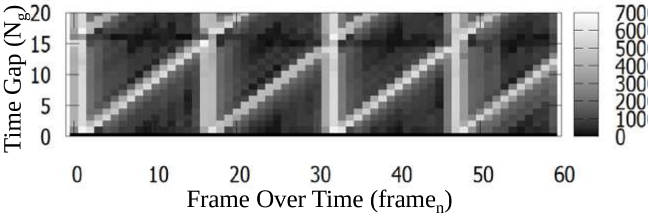


Figure 10: Frame distances over gap and time, the n-th frame is obtained at time $n/10$s.

the period. Following this definition, on receiving the frequency representation of a newly obtained frame, we first compute its Euclidean Distance against the previously obtained frames using the signal's amplitudes around the emitted frequencies. More specifically, given the current frame $frame_n$ and the previous frames $frame_{n-1}, frame_{n-2}, ...$, we compute their distance with a time gap N_g as follows.

$$dist(n, N_g) = \sqrt{dist_l^2(n, N_g) + dist_r^2(n, N_g)} \tag{10}$$

where $dist_l^2(n, N_g)$ and $dist_r^2(n, N_g)$ are the squared distance for the left and right signal sources of $frame_n$ with time gap N_g which are computed as follows.

$$dist_l^2(n, N_g) = \sum_{f=f_{0_l}-\Delta f}^{f_{0_l}+\Delta f} (amp_{f,n} - amp_{f,n-N_g})^2 \tag{11}$$

$$dist_r^2(n, N_g) = \sum_{f=f_{0_r}-\Delta f}^{f_{0_r}+\Delta f} (amp_{f,n} - amp_{f,n-N_g})^2 \tag{12}$$

where $amp_{f,n}$ is the amplitude of $frame_n$'s signal's component at frequency f after FFT, notations such as f_{0_l}, f_{0_r}, and Δf are the same as used in Sec. 5.1.2.

After obtaining the distances for a series of frames (N frames), we search for the minimum time gap N_0 that minimizes the average distances. More specifically, we compute N_0 as follows.

$$N_0 = \arg\min_{N_g} \frac{1}{N} \cdot \sum_{n=n-N}^{n} dist(n, N_g),\ N_g \in [1, max(N_g)] \tag{13}$$

Fig. 10 plots the distance matrix of frames corresponding to the signal shown in Fig. 9. Given the sliding window size of 0.1s, we set the maximum N_g=20 and $N = 20$. From Fig. 10, it is clear that the minimum average distance is obtained around $N_g = 15$ so we have $N_0 = 15$. This means the period of the signal, i.e., the cycle length, is around 1.5s. This result well matches the signal plotted in Fig. 9, suggesting the proposed cycle segmentation method is effective. By comparing Fig. 9 and Fig. 10, we can also notice that the vertical bars with the highest aggregated distance in Fig. 10 highly correspond to the start time for each cycle in Fig. 9. It is caused by a short gap between two adjunct cycles due to the vibration elimination technique presented in Sec. 4.2, which can be clearly seen in Fig. 9, especially near 1.5s. Based on this observation, we find the start time of a cycle t_s by finding the frame with the local maximum of aggregated distance.[3]

So far, we have found the start time t_s and cycle length $l = N_0/10$ (0.1s for each frame) for the current cycle. The period of the current cycle are then determined as $[t_s, t_s + l]$. The cycle segmentation method segments the audio stream into individual cycles which can be used to perform cycle aggregation (Sec. 5.4) and feature extraction for the final positioning method.

6.3 Feature-based Positioning

After the previous processing steps, our final goal is to determine the relative position of the receiver with respect to the walking direction. We model this problem as a classification problem and propose a feature-based positioning approach.

For each cycle, we extract both frequency and time domain features. For frequency domain features, we use the *HLR* and *LRR* introduced in Sec. 5.1.2 to characterize the signal. To characterize how signal changes over time, we compute the ratio of aggregated amplitude in the first half of the cycle against the latter half for each wrist. We further enrich our feature set by computing *HLR* and *LRR* for the left and right wrists independently, and using the normalized signal amplitudes of different frequency components as features. For each swing cycle, we extract the above features and represent it as an instance in the feature space.

Positioning is done using a Support Vector Machine (SVM) classifier with radial basis function kernel. SVM has shown to be promising in previous work to be a light-weight and efficient classifier that scales well to the number of features and training data [17].

Besides the four relative positions considered above, i.e., *front*, *back*, *left*, and *right*, in real-life, it is possible that the system is working with people not walking together. We add the *unknown* position to represent all cases other than the above four relative positions.

7 EVALUATION

In this section, we conduct experiments to evaluate the performance of the proposed approach.

7.1 Data Collection and Methodology

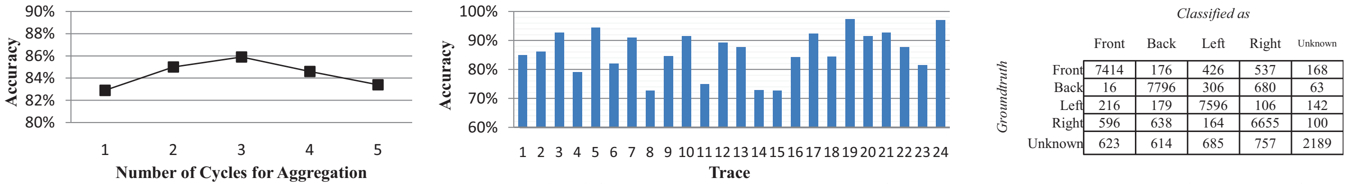
Data are collected by randomly choosing pairs of students in our department and asking them to walk naturally at different places, including both indoor and outdoor environments. No further instructions other than how to correctly wear the devices are given so the students can walk freely with the smartphone placed with their own choices. A total number of 24 traces are collected involving all the directions with each trace lasts for approximately five minutes.

We perform leave-one-trace-out evaluation to evaluate our system's performance. The discrimination accuracy is computed by the time slice accuracy as follows.

$$accuracy = \frac{correctly \quad classified \quad duration}{total \quad trace \quad duration}$$

We evaluate the time slice accuracy using different number of cycles for aggregation. We also evaluate our system's power and

[3] In the above case, the local maximum of aggregated distance always corresponds to the start time of the left wrist's signal. However, in other cases, this may correspond to the start time of the right wrist's signal or both wrists' signal. As a result, we combine the local maximum and the period to determine the start time t_s. And our system is not sensitive to whether t_s is the start time of the left or right wrist's signal. We omit the details in this paper due to page limits.



	Front	Back	Left	Right	Unknown
Front	7414	176	426	537	168
Back	16	7796	306	680	63
Left	216	179	7596	106	142
Right	596	638	164	6655	100
Unknown	623	614	685	757	2189

(a) Overall time slice accuracy

(b) Accuracy breakdown for traces (# cycles aggregated = 3)

(c) Confusion matrix (# cycles aggregated = 3)

Figure 11: System performance using leave-one-trace-out evaluation.

time costs, and evaluate the system's performance when the audio sources are attached to the user's ankles.

7.2 Positioning Accuracy

In this section, we report the system's positioning accuracy. Fig. 11(a) plots the overall time slice accuracy obtained by leave-one-trace-out evaluation under different number of cycles used for aggregation. As shown in the figure, our system achieves an accuracy of over 80%. The strongest discriminative power is obtained when using three cycles for aggregation, which is 85.9% in accuracy. It is interesting to find that the accuracy drops when using more cycles for aggregation. A possible explanation is that the noise contained in each cycle may also be amplified during aggregation which starts to affect the system's performance. This requires us to carefully choose the number of cycles for aggregation to achieve a balance between amplified signal and noise. Noise reduction technologies may be helpful to further increase the system's performance which we leave for our future work.

We further report the detailed performance when three cycles are used for aggregation. Fig. 11(b) shows the breakdown of accuracy for different traces. Nine out of twenty-two traces achieve over 90% accuracy with the highest accuracy of 97.4%. A possible way to further increase accuracy is to dynamically change the number of cycles for aggregation according to the quality of the data in different traces. Fig. 11(c) shows the confusion matrix of the positioning results. Most errors are made with the *unknown* position which does not have a clear pattern by definition. More errors are made between the *right* and *back* positions than other pairs of positions. As shown earlier in Fig. 3, *right* and *back* are close to each other in data distribution. This is possibly caused by different performance or orientations of the smartwatches' speakers. A possible solution is to perform data calibration. We leave the exploration of possible solutions to further increase our system's performance for our future work.

7.3 Power and Time Costs

We evaluate the battery and time performance of our system in this experiment. We use PowerTutor[4] to monitor the power consumption of our program running on smartwatch and smartphone. First, our swing detection and audio transmission programs on smartwatch introduces an additional power consumption of 53mW. The time cost of the smartwatch-side program is omitted because no complex computation is involved. Second, the audio receiving and relative positioning program's power consumption on smartphone is 65mW. It takes our program less than 5 seconds to analyze 10 seconds of data. The results suggest our system does not introduce a high power overhead. Also, the system is fast enough to perform online analysis. We can further reduce the power overhead of our system by turning into sleep when the user is not walking or the relative position does not change frequently, which we leave for our future system implementation work.

7.4 Attaching to Ankles

In some cases, people walks without arm swinging. To address this issue, we can attach the audio sources to user's ankles. We collect five traces of data in this setup and perform leave-one-trace-out evaluation. The result shows our system achieves an overall time slice accuracy of 88.7%, slightly better and still comparative to the wrist data. This result suggests attaching the audio sources to ankles is an effective option when user walks without arms swinging.

8 CONCLUSION

In this paper, we study the problem of direction-aware relative positioning for pedestrians walking together. Built on the fact that people's arms often swing back and forth during walking, we develop the wrist-body model that models the distance change between the user's wrists and his/her walking mate's body when walking together. An audio-based relative positioning approach is proposed based on the theoretical results which discriminates different relative positions with respect to the walking direction using the Doppler shift of the received audio signal. The cycle segmentation and aggregation methods are proposed to tackle the real-world issues. We build a prototype system using COTS wearable devices including smartwatches and smartphones. Experiments conducted under real-world conditions show that our system can perform relative positioning accurately and efficiently. We also show our system can still effectively perform relative positioning if the user walks without arms swinging by attaching the audio sources to the ankles.

ACKNOWLEDGEMENT

We thank all the participants that involve in our study. This work was supported by the National 863 Project No. 2015AA01A203, the NSFC under Grants 61502225, 61373011, 61272456, and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Chakib Baouche, Antonio Freitas, and Michel Misson. 2009. Radio proximity detection in a WSN to localize mobile entities within a confined area. *Journal of*

[4] http://ziyang.eecs.umich.edu/projects/powertutor/

Communications 4, 4 (2009), 232–240.

[2] Trinh Minh Tri Do and Daniel Gatica-Perez. 2013. Human interaction discovery in smartphone proximity networks. *Personal and Ubiquitous Computing* 17, 3 (2013), 413–431.

[3] Margaret Gilbert. 1990. Walking together: A paradigmatic social phenomenon. *MidWest studies in philosophy* 15, 1 (1990), 1–14.

[4] Serge P. Hoogendoorn and W. Daamen. 2005. Pedestrian Behavior at Bottlenecks. *Transportation Science* 39, 2 (2005), 147–159.

[5] Wenchao Huang, Yan Xiong, Xiang-Yang Li, Hao Lin, Xufei Mao, Panlong Yang, and Yunhao Liu. 2014. Shake and walk: Acoustic direction finding and fine-grained indoor localization using smartphones. In *2014 Proceedings IEEE INFOCOM*. IEEE, 370–378.

[6] Wenchao Huang, Yan Xiong, Xiang-Yang Li, Hao Lin, XuFei Mao, Panlong Yang, Yunhao Liu, and Xingfu Wang. 2015. Swadloon: Direction finding and indoor localization using acoustic signal by shaking smartphones. *IEEE Transactions on Mobile Computing* 14, 10 (2015), 2145–2157.

[7] Ye-Sheng Kuo, Pat Pannuto, Ko-Jen Hsiao, and Prabal Dutta. 2014. Luxapose: Indoor positioning with mobile phones and visible light. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 447–458.

[8] Kevin A Li, Timothy Y Sohn, Steven Huang, and William G Griswold. 2008. Peopletones: a system for the detection and notification of buddy proximity on mobile phones. In *Proceedings of the 6th international conference on Mobile systems, applications, and services*. ACM, 160–173.

[9] Liqun Li, Guobin Shen, Chunshui Zhao, Thomas Moscibroda, Jyh-Han Lin, and Feng Zhao. 2014. Experiencing and handling the diversity in data density and environmental locality in an indoor positioning service. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 459–470.

[10] Mehdi Moussaïd, Niriaska Perozo, Simon Garnier, Dirk Helbing, and Guy Theraulaz. 2012. The Walking Behaviour of Pedestrian Social Groups and Its Impact on Crowd Dynamics. *Plos One* 5, 4 (2012), : e10047.

[11] Yoshiyuki Nakamura, Yuko Namimatsu, Nobuo Miyazaki, Yutaka Matsuo, and Takuichi Nishimura. 2007. A method for estimating position and orientation with a topological approach using multiple infrared tags. In *Networked Sensing Systems, 2007. INSS'07. Fourth International Conference on*. IEEE, 187–195.

[12] Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, and Venkata N Padmanabhan. 2012. Centaur: locating devices in an office environment. In *Proceedings of the 18th annual international conference on Mobile computing and networking*. ACM, 281–292.

[13] L.M. Ni, Y. Liu, Y.C. Lau, and A.P. Patil. 2004. LANDMARC: indoor location sensing using active RFID. *Wireless networks* 10, 6 (2004), 701–710.

[14] Yasutaka Nishimura, Naoki Imai, and Kiyohito Yoshihara. 2011. A proposal on direction estimation between devices using acoustic waves. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 25–36.

[15] Yasutaro Nishimura and Kiyohito Yoshihara. 2013. A Device Specification Method Using Doppler Effect of Acoustic Waves. In *2013 IEEE 37th Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, 90–99.

[16] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*. ACM, 1–14.

[17] X. Qi, G. Zhou, Y. Li, and G. Peng. 2012. RadioSense: Exploiting Wireless Communication Patterns for Body Sensor Network Activity Recognition. In *Proceedings of the 33rd IEEE Real-Time Systems Symposium (RTSS)*. 95–104.

[18] Martin Wirz, Daniel Roggen, and Gerhard Tröster. 2009. Decentralized detection of group formations from wearable acceleration sensors. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, Vol. 4. IEEE, 952–959.

[19] Hongwei Xie, Tao Gu, Xianping Tao, Haibo Ye, and Jian Lu. 2015. A Reliability-Augmented Particle Filter for Magnetic Fingerprinting based Indoor Localization on Smartphone. *IEEE Transactions on Mobile Computing (TMC)* (2015), in print.

[20] Jie Xiong, Karthikeyan Sundaresan, and Kyle Jamieson. 2015. ToneTrack: Leveraging frequency-agile radios for time-based indoor wireless localization. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 537–549.

[21] Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. 2014. Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 237–248.

[22] Haibo Ye, Tao Gu, Xiaorui Zhu, Jinwei Xu, Xianping Tao, Jian Lu, and Ning Jin. 2012. Ftrack: Infrastructure-free floor localization via mobile phone sensing. In *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*. IEEE, 2–10.

[23] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 15–29.

[24] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 201–212.

[25] Zengbin Zhang, David Chu, Xiaomeng Chen, and Thomas Moscibroda. 2012. Swordfight: Enabling a new class of phone-to-phone action games on commodity phones. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM, 1–14.