Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

9-2017

AudioSense: Sound-based shopper behavior analysis system

Amit SHARMA
Singapore Management University, amit.2015@smu.edu.sg

Youngki LEE
Singapore Management University, YOUNGKILEE@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Computer and Systems Architecture Commons, and the Software Engineering Commons

Citation

SHARMA, Amit and LEE, Youngki. AudioSense: Sound-based shopper behavior analysis system. (2017). *UbiComp '17: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and ACM International Symposium on Wearable Computers, Maui, United States, 2017 September 11-15*. 488-493.
Available at: https://ink.library.smu.edu.sg/sis_research/3839

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

AudioSense: Sound-Based Shopper Behavior Analysis System

Amit Sharma
Singapore Management University
Singapore
amit.2015@phdis.smu.edu.sg

Youngki Lee
Singapore Management University
Singapore
youngkilee@smu.edu.sg

Abstract
This paper presents AudioSense, the system to monitor *user-item* interactions inside a store hence enabling precisely customized promotions. A shopper's smartwatch emits sound every time the shopper picks up or touches an item inside a store. This sound is then localized, in 2D space, by calculating the angles of arrival captured by multiple microphones deployed on the racks. Lastly, the 2D location is mapped to specific items on the rack based on the rack layout information. In our initial experiments conducted with a single rack with 16 compartments, we could localize the shopper's smartwatch with a median estimation error of 15.9 cm in 2-dimensional space.

Author Keywords
Sound Localization, TDoA, AoA, Trajectory Tracking, Audio Sensing

ACM Classification Keywords
H.1.2: [Shopper Behavior]: Models and Principles: User Machine Systems Human Information Processing

Introduction
Shoppers view, touch, and pick various items while they shop. Such interaction with items strongly reflects shoppers' needs and preferences. Accordingly, real-time monitoring of shopper-item interaction will enable useful new ap-

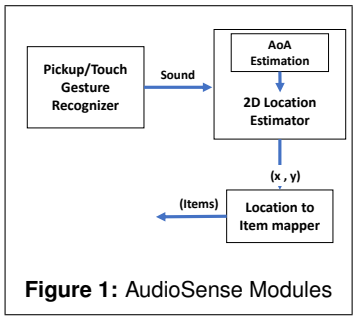

Figure 1: AudioSense Modules

plications. For instance, when a shopper picks up a printer cartridge at a stationery store, an associated 30% discount on a bundle of A4 papers can be sent to the shopper along with its location; the papers and cartridges are located distantly, and shoppers usually do not take the deal despite a high discount rate.

Recently, a few systems were proposed to detect shoppers behaviour and mobility [4, 2, 1]. However, most prior systems have focused on detecting *course-grain mobility*, i.e. which section the shopper is located in; whereas AudioSense detects *fine-grained interaction*, i.e. which item inside the section the shopper is interacting with. For example, IRIS [2] tracks *aisle* level movements of a shopper and estimates the time spent by the shopper in a vicinity of the items and away from the items (non-aisle areas). Although it detects item picking gestures yet it doesn't determine which particular item has the shopper picked. Shopminer [4] uses RFID tags attached to items to understand user-item interactions. The system captures the changes in RF signals between RFID readers and a tag caused by user-item interaction. The accuracy of this system depends on the density of RFID readers, and the readers are often costly. There has been a more closely related work, ThirdEye [3], which focused on detecting shoppers' interaction with items using smart glasses worn by the shopper. The video feed from the glasses is used to identify when and which item the shopper is picking. However, smart glasses are not yet prevailing due to many reasons such as severe privacy concerns, the inconvenience of use, the limited battery power.

To address the challenges in enabling real-time monitoring of shopper-item interaction, we propose AudioSense. AudioSense automatically detects the items that a shopper touches or picks by localising the inaudible sound signals emitted by a smartwatch worn by the shopper. More specifically, Audisense first detects the picking/touching gestures using inertial sensing data on the smartwatch and triggers the watch to emit the sound signals at the moment of interaction. Then, the signal is captured by multiple microphones deployed on the racks. Using each microphone data, we compute Angle of Arrival (AoA) of sound and combines AoA at multiple microphones to estimate 2D location of the sound source relative to the rack. This location is further mapped to the specific items on the rack which completes the shopper-item interaction mapping. The history of these interactions can be used to precisely customize promotions for the shopper. In this paper, we focus on 2-dimensional location estimation of the smartwatch, assuming that the gesture detection technique is already available.

Motivating Scenario

Joey walks in a stationary shop and is looking to buy a book from his favourite writer. He heads for the section where all the books from that author are kept and browse through them. He picks up a book from that section to check out more details about it. The AudioSense system detects this pickup gesture using the smartwatch worn by Joey and localizes the section from where Joey picked up the book. Using the rack layout information, the system identifies which author books are kept in that section and then immediately sends an early bird offer for the upcoming book from the same author to Joey's smartwatch. Joey likes the offer on the new book and hence opt for pre-booking.

System Overview

Figure 1 shows three modules involved in AudioSense. First module is gesture recognizer which utilizes inertial sensor data from shopper's smartwatch and detects an ongoing pickup gesture, if any. If any pickup gesture is recognized, then smartwatch would emit a sound for a moment. Sec-

ond module estimates 2 dimensional location(relative to the rack) of emitted sound. It first computes AoA from the received audio data and then using AoA values, it computes 2D location of smartwatch. Third module is responsible for mapping the estimated location to the specific items of the rack. It uses rack layout information to estimate closest possible item for the given 2D location. Output of this module is set of items that the shopper picked up.

Figure 2 shows detailed workflow & various components involved in AudioSense. It consists of a smartwatch worn by the shopper, multi-microphone system deployed on the racks and lastly a cloud server. The Smartwatch has a gesture recognition application running on it. This application can be either automatically started by the store location or manually before entering the store. Next, each rack inside the store will have 3-4 microphones installed at different positions as shown in the figure. These microphones can be treated as 2 (or 3) independent pairs. All microphones are connected to a computing device, e.g. raspberry pi board, which is further connected to a cloud server. This setup is repeated over all the racks inside the store. Lastly, the cloud server communicates with all the computing devices connected to the racks. Using the store layout information, the cloud server transforms the estimated 2D location to the specific items on the rack.

While inside the store, whenever the user picks up or touches any item on rack N, the gesture recognition application on the watch recognises this gesture (step 1 in figure 2) and immediately emits a specific sound (for a moment) denoted as step 2. All the microphones on the rack N receives this sound and forward it to the computing device to which they are connected. The computing device estimates an AoA value corresponding to each pair of the microphones. Next, the computing device combines these multiple AoA values

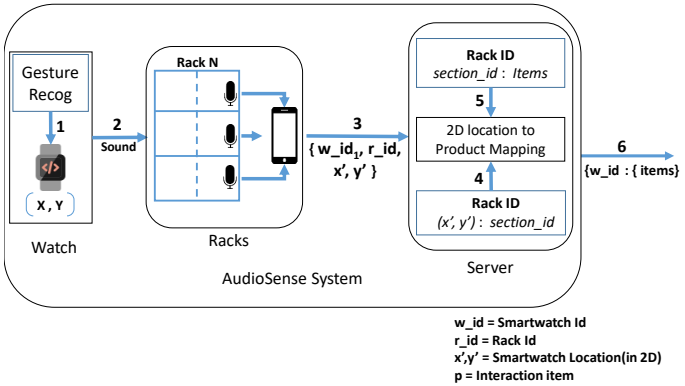

Figure 2: System Architecture

to estimate watch location (x',y') wrt the rack. This is denoted as step 3 in figure 2. This location is further mapped to the specific items (or compartments) on the rack (if rack item layout is known) shown by step 4 & 5. The output of the system is these items that the user picked up (or touched) denoted as step 6. Following sections describe AoA & location estimation procedure in more detail.

Requirements & Challenges
To realize this system functionality, following requirements need to be fulfilled which have certain challenges.

- Real-time Touch/Pickup gesture Recognition
  Output of this step is input for the AudioSense system, so high accuracy and low latency are crucial. These gestures might also vary from user to user, making it difficult to keep the training model accurate as well as generic enough at the same time.

- Robustness from Multi-User & Ambient Noise Interference
  Multiple users might be simultaneously present near a rack thus the sound received by microphones can be a mixture of multiple watch sounds. Also, the presence of dynamic nature ambient noise (e.g. people might be talking) can further complicate the situation and make it hard to separate sounds from each smartwatch.

- Accurate Rack Compartment Estimation
  Goal of the AudioSense system is to understand user-item interaction. Precise identification of these items heavily depends on the how accurately the Smartwatch can be located in 2D and how accurately this location is mapped to the corresponding compartment on the rack.

$$\theta = \cos^{-1}(\frac{v * \Delta t}{k}) \quad (1)$$

$$y = m_1 * x \quad (2)$$

$$y = m_2 * (x - d) \quad (3)$$

We see this work as our first step towards the final system, so in this paper, we address some of the above mentioned challenges. The following sections describe how does the system estimate the 2 dimensional location of watch (wrt the rack) and its mapping to the corresponding compartment of the rack.

2D Watch Location Estimation & Compartment Mapping

We believe 2D watch location estimation and its mapping to corresponding rack compartment are core part of the system because the specific items that a user is interacting with can only be determined by this step. The precision & accuracy of location estimation & its mapping determines how precisely can we determine items of interest to the user. This step makes AudioSense more precise system as compared to other works that track user's overall location inside the store and then send promotions based on the

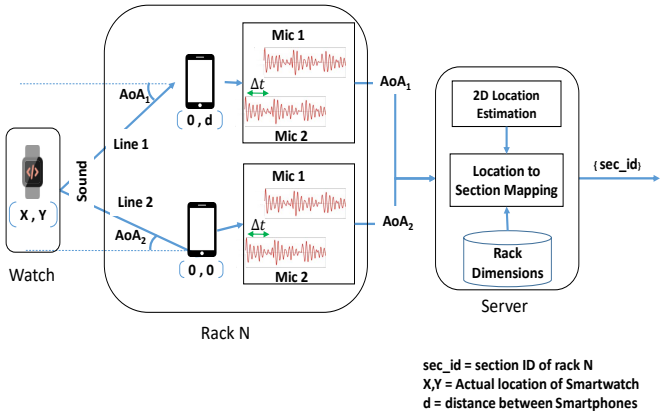

Figure 3: 2D Hand Trajectory Tracking System

these locations. So, in this paper we focus on these two components.

Figure 3 represents architecture of the system used in this paper. The pick-up/touch gesture recognition part is not implemented in the current system and we assume that the input sound to this system is triggered after the pickup gesture is identified by the smartwatch. Two pairs of microphones are required for watch location estimation in 2 dimensional space wrt rack N. So, we have used 2 Smartphones, each with 2 in-built microphones as shown in figure 3. Both Smartphones compute AoA from the received sound and send it to the server. The server then estimates the location of Smartwatch in x,y plane. In contrast to the final system architecture, the location estimation(from AoA) is shifted to server side and instead of 1 computing device per rack, we have used 2 Smartphones. Rest of the workflow remains the same.

Table 1: Experiment Parameters

Parameter Name	Value
Smart phones Used	Samsung Galaxy S7
Smart watch Used	LG Urbane 2
Distance between Micro-phones	14 cm
Distance between 2 Smart-phones	50 cm
Audio Sampling Rate	44100 Hz
Processing Window Size	6500 Samples per channel
Smoothing Window size	1-7
Rack Size	100x100 cm^2

Angle of Arrival Estimation
Smartphones used here have 2 microphones at different positions so sound received will have TDoA (Time Difference of Arrival) represented as Δt in figure 3 and is computed by cross-correlating individual microphone signals. For given audio sampling rate, AoA can be computed as shown in equation 1 where k is distance between the two microphones of a Smartphone and v is velocity of sound.

Optimal size of processing window (number of samples used to estimate each AoA) is crucial for accurate AoA estimation. Depending on the sound type and the ambience, low processing window size might not contain sufficient information about the desired sound to accurately cross-correlate.

But arbitrarily increasing processing window size will increase overhead during cross-correlation consequently increasing estimation latency. Moreover, the system will also have a smoothing window which will average multiple AoA estimates to compute one final AoA value for each location of Smartwatch, thus amplifying the impact of bigger processing window. To address this issue, we used time-domain cross correlation of samples bounded by a maximum possible lag value. This value of lag is computed beforehand for given values of sampling rate and distance between microphones. The cross-correlation module doesn't check for delay beyond this maximum lag value, thus reducing the cross-correlation operation cost. Evaluation section will describe the exact values used for our experiments.

2 Dimensional Location Estimation
Both Smartphones send estimated AoA values to server. For each location of Smartwatch and Smartphone, a straight line can be assumed between them. Using the AoA, server computes this straight line equation for both Smartphones as following.

Location of lower phone is assumed to be at origin (0,0) and that of the upper phone is (0,d) where d is the separation between the two phones. Equations of lines passing through lower & upper phones are shown by equations 2 and 3 respectively.

Where m_1 and m_2 are slopes of the lines which can be computed once corresponding AoA values are known. The intersection of these two lines represents 2D location (x',y') of Smartwatch wrt the origin.

Compartment Mapping
Having estimated 2D location, next step is to map this location to the corresponding compartment(or section) of the rack. For this, we divide the rack into a grid of m x n compartments(physical or logical) and compute coordinates of center of each compartment wrt origin(position of lower phone is origin and we know rack dimensions). Then we compute euclidean distance between (x',y') and each compartment center. The compartment with the minimum euclidean distance is termed as the corresponding compartment for the location (x',y').

Evaluation
This section describes the experimental setup and discusses results for smartwatch location estimation and compartment mapping.

Experimental Setup
Table 1 describes specific parameter values used in our experiments. We divide an area of 100x100 cm^2 into 4x4 grids. Each grid is assigned a number from 1-16 and represents a rack compartment. Next, we generate a random sequence of compartment numbers (1-16) and place the smartwatch at the center of each compartment in that order. We assume that a pickup gesture was detected at the center of each compartment, so the watch emits sound

Seq No	*Accuracy*	*Median Err(cm)*	*Mean Err(cm)*
1	6/10	15.3	17.9
2	6/10	15.9	17.7

Table 2: Exact Accuracy

Seq No	*Accuracy*	*Median Err(cm)*	*Mean Err(cm)*
1	10/10	15.3	17.9
2	10/10	15.9	17.7

Table 3: 1-Hop Accuracy

from each compartment center. This sound is used as input to our system, which outputs an estimated compartment number. So, for each compartment number in the random sequence, AudioSense estimates a compartment number. We present accuracy of these estimations in following section.

Results

We have used two types of accuracy metrics for our results. For each compartment number in random sequence, when AudioSense estimates the exactly same compartment number, then we call it exact-accuracy whereas if estimated compartment is *immediate* neighbour of actual compartment, then we call it 1-hop accuracy. Apart from these, we also report the median and mean errors (in cm) in exact location estimation for each sequence of compartment numbers. Table 2 & 3 show experiment results for *exact accuracy* scenario and *1-hop accuracy* scenario respectively. In *exact accuracy* scenario, AudioSense estimated 6 out of 10 compartments accurately and the remaining 4 compartments were estimated as immediate neighbour of actual compartment. This is evident from *1-hop accuracy* scenario wherein all 10 compartments are estimated either as exact or immediate neighbour. The median error in location estimation is found to be 15-16 cm.

These experiments show that rack compartments of radius around 16cm can be estimated accurately. Hence, in stores where each compartment contains similar items, it is possible to know which specific items a particular user is interested in.

Conclusion

We presented architecture of the AudioSense system to monitor user-item interactions inside a store and presented results from Smartphone based current system implementation. AudioSense could estimate exact location of the smartwatch wrt the rack with a median error of 15-16cm and shows compartment mapping accuracy of 60% and 100% for exact-accuracy and 1-hop accuracy scenarios respectively. In future, we would like to work towards other identified challenges of the system and test system performance in an actual stationary/grocery store.

REFERENCES

1. M. Popa, A. Kemal Koc, L. J. M. Rothkrantz, C. Shan, and P. Wiggers. *Kinect Sensing of Shopping Related Actions*, pages 91–100. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
2. M. Radhakrishnan, S. Eswaran, A. Misra, D. Chander, and K. Dasgupta. Iris: Tapping wearable sensing to capture in-store retail insights on shoppers. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–8, March 2016.
3. S. Rallapalli, A. Ganesan, K. Chintalapudi, V. N. Padmanabhan, and L. Qiu. Enabling physical analytics in retail stores using smart glasses. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, MobiCom '14, pages 115–126, New York, NY, USA, 2014. ACM.
4. L. Shangguan, Z. Zhou, X. Zheng, L. Yang, Y. Liu, and J. Han. Shopminer: Mining customer shopping behavior in physical clothing stores with cots rfid devices. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, SenSys '15, pages 113–125, New York, NY, USA, 2015. ACM.

Acknowledgement

This research is supported by the National Research Foundation, Prime Minister′s Office, Singapore under its IDM Futures Funding Initiative.