Using Sound Source Localization to Monitor and Infer Activities in the Home

Xuehai Bian, Gregory D. Abowd, James M. Rehg
GVU& College of Computing, Georgia Institute of Technology,

801 Atlantic Drive Atlanta, Georgia 30332
{bxh, abowd, rehg}@cc.gatech.edu

Abstract. Recent research in ubiquitous computing has focused both on how to infer human activity from a variety of signals sensed in the environment as well as how to use that information to support interactions. In this paper, we examine the feasibility and usefulness of sound source localization (SSL) in a home environment, which is an implicit location system to support monitoring of a remote space as well as to infer key activities, such as face-to-face conversations. We present a microphone array system that covers a significant portion of the public space in a realistic home setting and discuss monitoring and automated inferring applications that are made possible with this technology in a domestic setting.

1 Introduction

Context aware computing is one of the main themes in ubiquitous computing. In practice, certain types of context, such as location, identity, time and activity are more important than others [1]. Since the early 1990's much research effort has been focused on how to acquire, refine, and use location context information [2].Traditional location-sensing systems rely on either explicit or implicit means of localization. In explicit localization, the user must wear or carry a device or tag which is used to locate them, while implicit localization systems do not require instrumenting the user. Most implicit localization systems use computer vision to track users. We are interested in the use of sound source localization in the home environment, arguing that understanding the location of sound sources can be valuable for context aware computing. Sound events are often associated with human activities in the home, but little effort in the ubiquitous computing community has tried to exploit this.

There are social concerns when sensing video and audio in the home environment. When the actual information retrieved is not the rich signal that a human would see or hear, there is potential for alleviating those concerns. We designed a sound source localization (SSL) system which locates sound events in the environment using microphone arrays. The only information extracted in this case is solely the location of

sound sources. Our system is based on a standard SSL algorithm which uses the time of delay method and PHAse Transform (PHAT) filtering in the frequency domain to locate sound sources [3]. In Section 3 we describe the additions we made to this standard algorithm to make it robustly cover a significant portion of the public space in a realistic home setting. The system runs continuously (24/7) and feeds the detected sound events into a database which is shared with other applications in the home. From our experiments, the accuracy of our system is sufficient to provide rich context information, such as identifying conversations by their pattern of alternating location events.

To demonstrate the feasibility and usefulness of data generated by our SSL system, we built two applications described in Section 4. One is the Sound Event Map, which visualizes the sound events in 3D space allowing a home owner to monitor activities in his house. The other application demonstrates the potentially rich context information that can be distilled from sound events. By using simple heuristics, we are able to distinguish between a two-person conversation and a single person talking, *e.g.*, on the phone or to themselves. We believe this is a promising starting point towards more sophisticated activity recognition based on audio sensing.

2. Related Work: User Location Systems

In more formal environments such as the office, the wearing of an explicit tag or badge can be mandated. However, from our experiences in the home environment, deploying a RFID proximity location system, we observed that one important reason why the system was not extensively used is that some users forget to wear or even lost their tags. User satisfaction and cooperation is the final decisive factor in adoption of ubiquitous location services that require explicit tags or devices to function. For this reason, we have chosen to categorize related work in the location-sensing field by their reliance on explicit tags or devices for user tracking.

2.1 Explicit Systems

Most current location systems in ubiquitous computing are focusing on explicit location methods. Many explicit localization systems, that require users to wear extra devices, have been developed since the 1990s. The Active Badge system, one of the first successful indoor proximity location systems, required users to wear a badge that emitted infrared ID information giving zone level location information [4]. With the improved Active Bat system, users carried a 5cm by 3cm by 2cm Bat that received radio information and emitted an ultrasonic signal to ceiling mounted receivers. This provided location accuracy of 9cm with 95% reliability [5]. The Cricket location system requires user to host the Listener on a laptop or PDA and obtains the location granularity of 4 by 4 feet [6].

We designed our own indoor location service using RFID. Users wear passive RFID tags which are queried by RFID readers at fixed locations to obtain a unique ID [7]. RFID tags are small and passive, and hence, easy to carry and do not require batteries. However, instrumenting an environment with enough readers to obtain decent location information can be expensive.

In order to take advantage of existing radio beacon infrastructure, such as WiFi access points, passive wireless positioning systems use the signal strength of access points received by a wireless network card to determine the location of mobile users with an accuracy of 1-3 meters [8]. This technology has been used on several campuses such as the Active Campus project at UCSD and CMUSKY project at CMU.

For outdoor localization, users can carry GPS receiver and get the global position at the accuracy of 1-5m. Many projects use GPS as the primary outdoors positioning system, including Lancaster's Guide [9], and Place Lab [10]. A good taxonomy of current location systems that mainly focuses on explicit localization systems is described in [2].

Explicit systems generally tend to be more robust than implicit systems, and almost always provide identification information (e.g., unique tag ID) in addition to location. They also allow users to determine for themselves if they wish to be tracked; by choosing whether or not to carry the tag/device. The largest drawback of explicit location systems from the user's perspective is the size and weight of the tag or device they must carry. Many devices require a certain amount of local computation or signaling capability. GPS receivers need a processor to compute their location after receiving satellite signals, while beacons must expend enough energy to be detected. The requirement for computation and/or broadcast power adds to the size and weight of the device. Although different methods are suggested to reduce energy consumption such as human powered computing [11], we feel that the power supply problem will still exist for some time [12]. It seems these user-worn location devices will not be small or light enough to be considered invisible in the near future.

2.2 Implicit Systems

The other category of location system uses implicit characteristics of the users to sense their location, including visual clues, weight, body heat or audio signals. Implicit tracking does not require users to wear tags or carry devices, which pushes the tracking technology into the background. They come closer to meeting Mark Weiser's goal of calm technology [13].

Motion detectors and floor mats open supermarket doors, and motion sensing flood lights and sound activated night-lights ease light pollution while still providing illumination when needed. Although these simple appliances do not track the location of specific users, they implicitly know the location of whoever has activated them for a brief period of time.

With the development of artificial intelligence and increasing computing power, more perception technologies are used to support a natural interaction with the environment. Vision-based tracking and SSL are two important location strategies which have

the ability to passively monitor large spatial areas with only modest amounts of installed hardware. In contrast to motion detectors and contact-based floor sensors they provide greater resolution and discrimination capabilities.

Techniques for tracking people using multiple cameras can be divided into two groups, methods that track parts of the body such as faces [14] and limbs [15], and methods that treat the body holistically as a single moving target [16], often using a "blob" model to describe the targets appearance. See [17] for a recent survey. In a home setting, multiple users can be tracked in real-time using ceiling or wall-mounted cameras. The region corresponding to each user in each of the camera images is described as a blob of pixels, and it can be segmented from the background image using a variety of statistical methods[16, 18, 19]. By triangulating on the blob's centroid in two or more calibrated cameras, the location of the user can be estimated in 3-D. In the EasyLiving project at Microsoft Research, the blob's location can be updated at 1-3Hz in a room environment with two cameras for up to 3 users. Vision requires significant processing power and broadband networking infrastructure in order to get satisfactory real time location updates [19].

Passive sound source localization provides another natural tracking method that uses difference in time of flight from a sound source to a microphone array. With computer audio processing, sound source location can determine the location of sound events in 3-D space. We will discuss SSL in more detail in Section 3.

Because users do not need to explicitly carry tags or devices, these systems allow for implicit interaction, but may not provide identification information. With the help of face recognition, fingerprint, or voiceprint recognition, computer perception based location systems can provide identity information in addition to location.

Although implicit location systems do not allow users to physically opt-out of being tracked (by not carrying a tag/device), we believe that privacy issues can be solved by legal mandate and technical solutions in higher level sensing architectures.

2.3 Implicit Vision Based Tracking versus Sound Source Localization

Vision based tracking and SSL are more accurate than other simple implicit location system, like contact based smart mats or motion detectors. Computer vision systems usually use multiple cameras to circumvent visual obstacles or provide continuous tracking for moving objects over multiple rooms. Vision systems require significant bandwidth and processing power, as a typical color camera with 320x160 resolution at 10 frames per second generates about 1.54Mbyte of data per second.

In comparison, the data throughput of a microphone array is significantly less than a camera system. One microphone generates about 88.2KByte per second for CD quality sound with 16 bit sampling resolution. Because of the relatively lower bandwidth, a microphone array of 16 to 32 sensors can be supported by one Intel PIII desktop computer.

Current vision based location tracking systems suffer from variance in circumstantial light, color, geometric changes of the scene and motion patterns in the view, while sound source localization systems suffer from environmental noise. A sound localiza-

tion system can more easily detect activities that have specific sound features such as a conversation or watching TV, which might be difficult to detect using computer vision alone. However, sound source localization also has obvious disadvantages. Only activities which generate sounds (which may be intermittent) can be detected by the system.

An active research community is addressing the problem of fusing audio and video cues in solving various tasks such as speaker detection [15] and human tracking [20]. For example, the initial localization of a speaker using SSL can be refined through the use of visual tracking [21].

2.4 Sound Source Location as Important Source of Context

One important context from the audio is the capability to detect the sound event's location with some accuracy. We can find a cordless phone when it rings based solely on sound source location. Among the activities which take place in the home, identified by Venkatesh [22], a large portion of them are connected with sound events. Sound events happen when we have conversation, watch TV, listen to the radio, make phone calls, walk over the floor, move chairs around, drop objects onto tables, cook dinner, wash, or eat.

In domestic environments, different activities are often conducted in particular locations. Many researchers in ubiquitous computing are focusing on inferring human activity from a variety of signals sensed in the environment. Noticing the activities are usually connected with sub-areas in the room, a semi-automated method can be used to divide the room into sub-areas which are called activity zones and provide interaction based on status with regard to different zones [23].

Kitchen activities, such as cooking and washing dishes, are most likely to occur in the kitchen around the stove and sink. If the system observes sound events from the kitchen and stove for thirty minutes, followed by sound events surrounding the dinning room table, it can make a good prediction that a meal is occurring. If continuous sound events are detected from where the TV is located, you are probably watching television. Also if you analyze the height of the sound events, footsteps occur at floor level, sound events from the table may indicate that an object was dropped, while conversational noises are likely to be located at a higher position or above chairs.

SSL is good at summarizing activities that generate sound events over a period of time and providing answers to questions like: *when did we have dinner yesterday? Did I cook yesterday?* The update frequency of sound event location is fast enough to recognize some patterns of sound event sequences, like the switching between two persons in a conversation. Sound events can be used to determine the status of the users: *Is the user in a conversation?* It provides substantial information towards high level context such as interruptability determination [24].

Despite the fact that sound location can be an important source of context information in domestic environments, there is little to no research designed to investigate the location of sound events as an important context source to support recognizing

household activities. Most current SSL systems are still in the prototype stage in controlled research lab environments due to the difficulty of deploying a working system in a home.

3 Sound Source Localization in the Home

Sound Source Localization (SSL) systems determine the location of sound sources based on the audio signals received by an array of microphones at different known positions in the environment. All microphones receive time-shifted signals mixed with environmental noise and reverberation. In this section, we first summarize challenges for sound source localization in home environments. Then, we present the improvements to the standard PHAse Transform (PHAT) SSL algorithm we implemented to overcome these challenges. Additionally, we report our strategies to improve the usefulness of the system. We report on the accuracy of our SSL system measured at 6 representative sample locations.

3.1 Challenges to Deploy SSL System in Domestic Environment

Sound source localization research started many decades ago; however, there exists no general commercial SSL system. Based on current SSL research literature [25] and our own experiences [26], the main challenges for deploying SSL systems in domestic environment are:

1. **Background noise** - The background noise in home environments can include traffic noise, noise from household appliances and heating and air conditioners. For single SSL system, noise from the microwave will pose a problem for localizing the person talking at the same time.
2. **Reverberation (echoes) –** Reverberation in the home is difficult to model and can lead to corrupted location predictions when indirect (bounced) sound paths interfere with direct sound paths.
3. **Broadband** - The speech signals and sounds generated from household activities are broadband signals. The failure of narrowband signal-processing algorithms, applied in radar/sonar systems, requires the use of more complicated processing algorithms.
4. **Intermittency & Movement** - The sounds to be detected are usually intermittent and non-stationary. This makes it hard to use adaptive filtering which uses stationery source assumptions.
5. **Multiple Simultaneous Sound Sources** – Current single source sound localization algorithms fail when faced with multiple simultaneous sound sources.

Despite these general challenges, our research system shows that it is feasible and useful to start investigating how sound source location can help to locate human generated sound events which can be used to infer activity both manually and automatically.

3.2 Fundamentals in Passive Sound Localization

SSL systems can be traced back to earlier active radar and sonar localization systems. An active system sends out preset signals to the target and compares it with the echo signal in order to locate the target, similar to how a bat locates its prey using ultrasonic pings. In passive localization, the system only receives signal generated by the targets. If a user wears an explicit tag (such as the Active Bat ultrasonic badge) the receiver can compute the location with high accuracy because of high signal to noise ratio (SNR) in the narrow frequency range. However, for the implicit sound sources in a domestic environment we intend to explore, the signal is often noisy and with broader frequency ranges.

Different effective algorithms with an array of microphones are used in sound source localization. They can be divided into three main categories [25]. Most current sound source location systems are based on computing Time-of-Delay and using PHAT based filtering because they are simple, effective and suitable for real-time localization in most environments. The locating process is divided into two steps: Computing time delay estimation for each pair of microphones and searching for the location of the sound source. Different systems vary in the geometric deployment of sensors, pairing up, filtering and space searching strategies. We will explain the design of our SSL system after a simple introduction to PHAT and correlation based time of delay computations. More details are available in [26].

The incoming signal x received at microphone i can be modeled as

$$x_i(t) = a_i s(t - t_i) + n_i(t) \quad (1)$$

where: $s_i(t - t_i)$ is the signal delay; $n_i(t)$ is the noise; a_i is the attenuation factor for microphone i. For every pair of microphones, we compute the correlation. Usually it is done is frequency domain in order to save time. However, because of noise and reverberation in the environment, some weight functions in frequency domain are applied to enhance the quality of the estimation, such as Phase Transform(PHAT), and Roth Processor etc [3, 27]. The general cross correlation with the PHAT filter is equation 2.

$$\hat{R}_{x_1x_2}(t) = \int_{-\infty}^{\infty} \frac{1}{\left| X_1(f) conj(X_2(f)) \right|} X_1(f) Conj(X_2(f)) e^{j2\pi ft} df \quad (2)$$

Where *Conj* is the complex conjugation function, $X_1(f)$ and $X_2(f)$ are the Fourier transform of $x_1(t)$ and $x_2(t)$. Ideally, the shift of the peak point from the center is the time delay of signal arrival between these two microphones.

3.3 Peak weight-based SSL and other design decisions

We deployed our sound source localization system in an actual home setting and improved the location evaluation function to perform well in a home environment. Figure 1 shows the floor map of the target area.

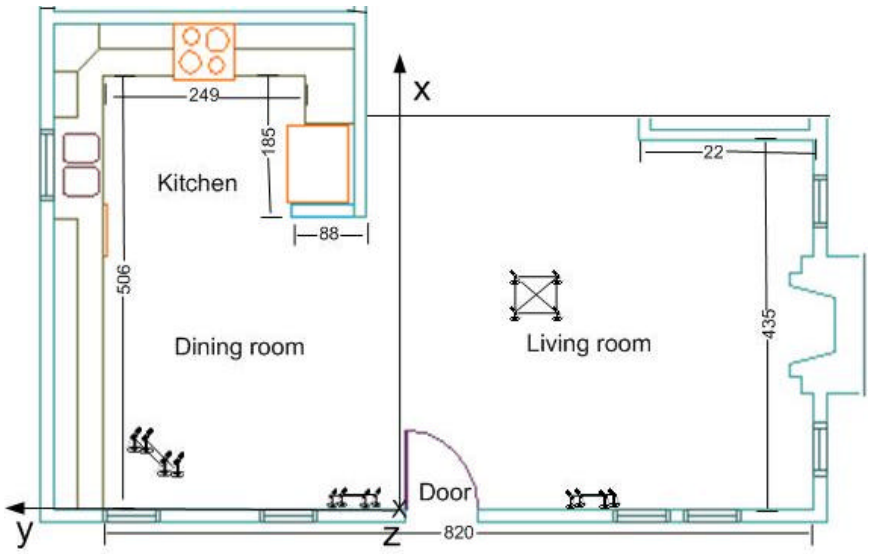

Figure 1. 2D Floor map of covered space in the first floor with four microphone Quads

We improved our system by the following measures:

1) The house is close to a busy street and the noise level is variable throughout the day, so we dynamically update the noise energy threshold in processing before actual localization to ignore street noise.
2) Our target area consists of a living room, dining room and kitchen. To cover the large space 16 microphones were used. However, the microphones were organized into 4 separate Quads (set of 4 microphones in a square pattern). By only computing time of delay between microphones in the same Quad, it effectively limits the peak search range and rules out false delays.
3) To fully utilize the information from each Quad, we correlate sound signals between all six pair-wise combinations of the four microphones.
4) Quads which are closer to the sound source make better location predictions. We calculate the location as determined by each Quad and then use the data from the most reliable Quad. Quads which are far away from the sound source suffer corrupted Time-of-Delay data because of the low signal to noise ratio.

In addition to the above measures, we find it is necessary to reflect the reliability of each Time-of-Delay estimation into the final localization goal function. We use the

ratio of the second peak with the maximum peak in equation (4) to convey the reliability of computed time of delay. Specifically, we define the peak-weight of *ith* pair of microphones to be:

$$W_i = 1 - V_{\text{Second peak}} / V_{\text{Max Peak}} \quad (3)$$

We discard the data items whose peak-weights (W_i) are less than a constant chosen to filter about 60-80% of the measurements. In a home environment with a signal to noise ratio between 5 and 15 db we experimentally determined this constant to be 0.3.

In the second phase of searching for the sound source location, we use deepest gradient methods and final evaluation function E of each potential location is calculated by equation (4). Note that we consider the peak weight (W_i) in the final evaluation function.

$$E = \sum_{i \in PossilbePairs} [W_i (TDOA^{Exp}_i - TDOA_i)^2] \quad (4)$$

$TDOA_i$ and $TDOA^{Exp}_i$ is the measured and the expected time delay from a potential location. During the search for the location of a sound event, we added more initial searching points from the more probable sound source locations, like the kitchen area, dining and living room tables in addition to the previous detected sound source locations. By seeding the initial search points in this manner, we increase the responsiveness of the system to common sound events.

3.4 System design in a home environment and results

Although our system design is not sophisticated enough to work in different environments, it does work well in our target area, a realistic home environment on our campus, where the research initiative is to explore varieties of technologies and applications for futuristic home environments. The dimension of the area is: 820(L)* 506(W)*272(H)cm. The microphone array is deployed in the connected areas including the living room, dining room in the first floor (also see Fig. 1). We are using 16 omni-directional pre-amplified microphones (cost: 10 USD each) that receive audio signals in the 20Hz to 16KHz range. In the living room, one Quad is in the ceiling, two are on the front wall and one is at the corner of the dining room which faces the dining room and kitchen. Figure 2 shows the pictures of arrays (Quads), each of which has 4 microphones. The microphones are small enough to be mounted in the wall or picture frames.

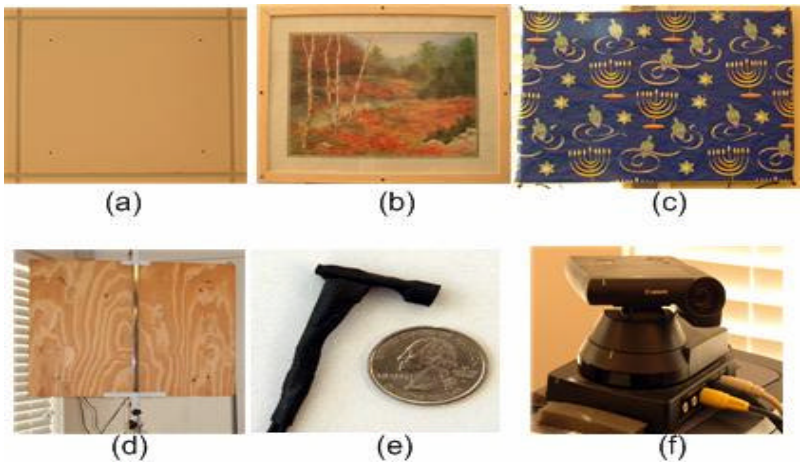

Figure 2. (a)-(d) Microphone Quads, each has 4 microphones. (e) Single microphone with a quarter. (f) PTZ camera driven by detected sound location events.

Currently, one of our demo applications is to drive a pan-tilt-zoom camera to focus on locations where sound events are detected and display it though a large plasma screen. In our home environment the current system can locate sound events from talking, footsteps, putting glasses on the table, chewing food, and clashes of silverware with dishes. Our location updating rate for continuous talking that faces the quads are 1-5 seconds per reading. Generally, it is especially responsive to crisp sound events such as eating, sniffing, or putting a backpack on a table. To test the accuracy of the SSL system, we placed a computer speaker broadcasting a radio news program in 6 representative locations in the space. The SNR of the speaker was set to around 10db. The standard deviation of measured location from sound source is between 6 and 33 cm (Table 1, using coordinate system in Fig. 1), which is enough to generally locate sound events to specific areas in the rooms such as a tabletop, floor, kitchen sink, TV, and to disambiguate between multiple non-moving speakers.

Table 1. The accuracy of sound source in the living room

Location in Room	Source Location (cm)			Num of Measure-ments	Std Error (cm)
	x	y	z		
Left-Front	126	-348	147	1237	31
Left-Behind	286	-325	150	1182	33
Mid-Front	80	-98	148	1002	17
Mid-Behind	299	-76	146	1262	29
Right-Front	147	233	146	1471	6
Right-Behind	310	247	149	1249	14

The standard deviation varies with the source location relative to microphone Quads. Generally, the more Quads which are closer to the sound source, the higher the SNR is and thus the results are more accurate.

4. Applications of SSL in Home Environments

An initial application we developed, mainly to test whether the SSL technology works, drove a pan-tilt-zoom camera to show the area where sound was detected. While this kind of application might be useful for remote monitoring of meetings or for childcare, it was never intended to be the motivating application for our work. With the current sensing capabilities we have in our home environment, we see at least two uses that are now possible, and the potential for other applications that might continue to drive how to improve the sensing. The first application demonstrates a visualization of activity over time, and the second application explores the ability to infer some understanding of human activity.

4.1 Sound Event Map

In a domestic environment, the owner of a home might be interested in viewing what happened in his house yesterday morning, see a summary of activity over longer periods, or remotely access the house of older adults. We developed a sound event map to facilitate this. In the sound localization system, all the sound location data with time stamps are stored into the house database server. The sound event map application connects to the server and retrieves the sound location history. It also pulls the geometric data of the rooms and furniture. Because each sound source location event consists of 16 bytes (X,Y,Z,Timestamp) and events detected are at most a few readings per second, the Sound Event Map application requires very low bandwidth, using orders of magnitude less than a webcam or audio broadcast from the same space.

Our sound event map allows the users to virtually move and turn around in 3D space. It also supports top view, front view and lateral view to better determine what is happening in a particular area. We are assuming the user of the application, mostly the owner of the home, is familiar enough with the area to work with a simplified floor-plan and simplified representations of furniture.

The user can select a timeslot of interest (*e.g.*, 7:30am to 10:00am yesterday morning) or select an area of interest -- the system automatically determines timeslots where activity in the chosen area (*e.g.*, kitchen, dining table) occurred. Events are colored from green to red depending upon their age. Another mode is designed to display a series of sound events in sequence, such as moving from the kitchen to the living room. During the automatic replay, the current event is highlighted with the largest size dot, while the five previous events are rendered with smaller dots.

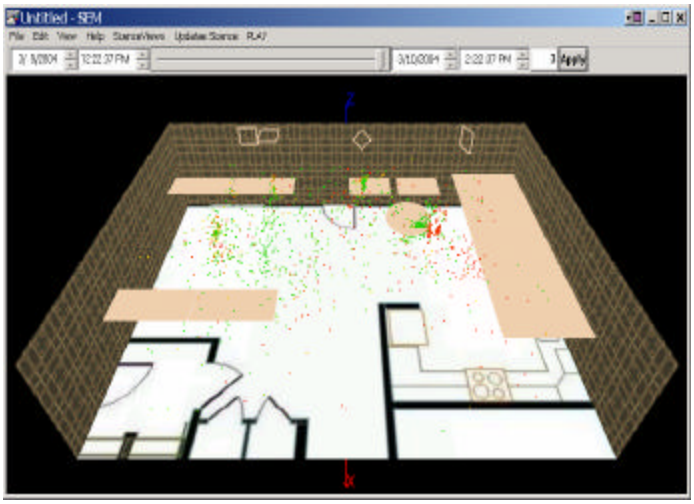

Figure 3 Sound Event Map application, showing sound events (red/green dots) for a day in the home.

4.2 Distinguishing Two-person Conversations from Single-person Talking

The primary benefit of collecting sound location history is the ability to recognize activities in the home. As we mentioned before, the sound location context is usually associated with human activity. Certain household activities are usually linked with specific locations in the home, such as dining, cooking and watching TV.

Currently, users examining the data we collect using the sound event map are able to manually recognize dinning activities by looking for events around the dining room table. Kitchen related activities are also easy to recognize. A little more difficult problem is how to differentiate between single-person talking and two-person conversations. For example, if a smart home would like to visually display a private message to the homeowner, it would be useful to determine if they are alone (talking on the phone) or having a conversation with another person. This problem is difficult to solve using only voice detection algorithms without sound source location.

To demonstrate that the data from our SSL system can differentiate between these two situations, we recorded 10 two-person conversations and 10 people talking over a telephone. These activities are distributed in different places of the covered area. The distances between the two people in conversations ranged from 0.5m to 5m. For people talking over the phone, we have 5 situations where their position is semi-fixed (sitting, but able to sway less than 1m) and 5 situations where the user paced around the area (representing a cordless handset). Three typical cases with dots representing events are shown in Fig.4. Each activity lasted between 2-5 minutes.

Figure 4 Three typical talking cases

To detect two-person conversations, we used a K-means clustering algorithm to separate the data points into two clusters. Then we counted the frequency of the back and forth between these two clusters with their timestamp information.

Table 2. The clustering of 10 two-person conversations and 5 one–person talking at fixed location and 5 single-person talking and moving around.

Cases	total number of readings	Number of Flip-flops	Distance between clusters (cm)	Two-person Conversation?
2-person 1	79	23	191	Yes
2-person 2	29	16	160	Yes
2-person 3	32	12	202	Yes
2-person 4	47	15	118	Yes
2-person 5	59	17	120	Yes
2-person 6	44	16	440	Yes
2-person 7	44	12	258	Yes
2-person 8	51	17	150	Yes
2-person 9	44	12	134	Yes
2-person 10	79	23	191	Yes
1-person mv 1	22	8	201	Yes
1-person mv 2	32	5	118	No
1-person mv 3	43	2	322	No
1-person fix 4	48	22	69	No
1-person fix 5	34	4	101	No
1-person fix 6	31	12	97	No
1-person fix 7	39	18	61	No
1-person fix 8	36	19	64	Yes
1-person mv 9	32	2	340	No
1-person mv10	35	7	283	No

Our proof-of-concept algorithm uses the following two heuristic rules:

1. If (# of flip-flops between two clusters/ # all reading>R1 and distance < D)
   it is a conversation;
2. If (# of flip-flops between two clusters/ # all reading> R2 and distance >= D)
   it is a conversation;
3. else
   it is a single person talking;

Before our experiment, we assigned the parameters R1=0.5; R2=0.25; and D = 100cm. The meaning of these parameters is that when sound clusters are closer than D=100cm, we requite 50% of the sound events to represent the flip-flop between the clusters before the activity is judged to be a conversation. When the sound clusters are farther apart than D=100 cm, then we only require 25% of the sound events to represent the flip-flop before the activity is judged to be a conversation.

When two persons are having a conversation, they will form two clusters of data points in the space and there should be sufficient flip-flopping between these two clusters. However, for a single-person talking around a fixed location, detected location events vary randomly around the true location. We choose D=100 cm to be larger than the maximum distance traveled by a swaying person plus sensing error.

The results in Table 2 show that all 10 two-person conversations were correctly categorized. Four of our five single-person fixed location cases and single-person moving cases were correctly categorized, while one of each was incorrectly judged to be a two-person conversation, giving our proof-of-concept algorithm an accuracy rate of 90% over all 20 cases.

We must point out that we are only using simple heuristic rules to distinguish conversations between two-people and a single-person talking on the phone in the home environment. More sophisticated linear dynamic models should be used to recognize patterns and provide inference for a dynamic number of people. But this work demonstrates the context information inherent within sound source location events; ready to be harvested with more sophisticated inference algorithms.

5. Conclusion and Future Work

In this paper, we summarized two different categories of localization systems and pointed out that implicit localization systems have advantages for deployment in a ubiquitous computing environment. By adapting current SSL technologies, we built a SSL system in a real life home setting and collecting sounds from cooking, dining and conversation. To support a home owner's access to the sound event history, we built the Sound Event Map to allow access to past sound events to determine what has happened in their home. With additional input from other sensors like a camera, speaker identification system and more sophisticated processing, SSL can be used to recognize activities in the home and summarize the habits of users.

By capturing the sound event locations during conversation, we can dynamically cluster the points according to the number of people in the conversation. With more

sophisticated modeling of conversations, we can also find interesting patterns such as who is dominating the conversation.

References

1. Dey, A.K., *Providing Architectural Support for Building Context-Aware Applications*, in *College of Computing*. 2000, Georgia Institute of Technology.
2. Hightower, J. and G. Borriello, *Location Systems for Ubiquitous Computing*. Computer,IEEE Computer Society Press, 2001. **34**(8): p. 57-66.
3. Knapp, C.H.a.C., G. C., *The generalized correlation method for estimation of time delay,"*. IEEE Trans. Acoust., Speech, Signal Process. ASSP-24, 1976. **24**: p. 320-327.
4. Harter, A.a.H., A., *A Distributed Location System for the Active Office*. IEEE Network, 1994. **8**(1): p. 62-70.
5. Harter, A., HOPPER,A., STEGGLES,P., WARD,A. and WEBSTER P. *The anatomy of a context-aware application*. in *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*. 1999. Seattle, Washington, United States: ACM Press.
6. Priyantha, N.B., Chakraborty, A. and Balakrishnan,H. *The Cricket Location-Support System*. in *Proc. 6th Ann. Int'l Conf. Mobile Computing and Networking (Mobicom 00)*. 2000. New York, 2000: ACM Press.
7. Abowd, G.D., Battestini,A. & O'Connell T., *The Location Service: A framework for handling multiple location sensing technologies*, in *GIT-GVU-03-8*. 2002, GVU technical report, Georgia Institute of Technology.
8. Castro, P., Chiu,P., Kremenek,T., and Muntz, R. *A Probabilistic Location Service for Wireless Network Environments*. in *Proceedings of Ubicomp 2001*. 2001. Atlanta: Springer Verlag.
9. Cheverst, K. *Developing a Context-Aware Electronic Tourist Guide: Some Issues and Experiences*. in *Proc. 2000 Conf. Human Factors in Computing Systems (CHI 2000)*. 2000. New York: ACM Press.
10. Schilit, B.N. and G.B. Anthony LaMarca, William G. Griswold, David McDonald, Edward Lazowska, Anand Balachandran, Jason Hong, Vaughn Iverson. *Challenge: ubiquitous location-aware computing and the "place lab" initiative*. in *Proceedings of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots*. 2003.
11. Starner, T., *Human-Powered Wearable Computing*. IBM Systems Journal, 1996. **35**(3-4): p. 618-629.
12. Forman, G.a.Z., J., *The challenges of mobile computing*. IEEE Computer, 1994. **27**(4): p. 38-47.

13. Weiser, M. and J.S. Brown, *The coming age of calm technology*, in *Beyond calculation: the next fifty years*. 1997, Copernicus New York, NY, USA. p. 75-85.
14. DeCarlo, D.a.M., D. *Optical Flow Constraints on Deformable Models with Applications to Face Tracking*. in *IJCV*. 2000.
15. J. M. Rehg, D.D.M., and T. Kanade, Int. J. of Robotics Research, *Ambiguities in Visual Tracking of Articulated Objects Using Two- and Three-Dimensional Models.* Int. J. of Robotics Research, 2003. **22**(6): p. 393-418.
16. Haritaoglu, I., Harwood,D. and Davis,L.S. *W4: Real-Time Surveillance of People and Their Activities*. in *IEEE Trans. On PAMI.* 2000.
17. Gavrila, D.M., *The visual Analysis of Human Movement: A Survey.* Computer vision and Image Understanding, 1999. **75**(1).
18. Wren, C., Azarbayejani,A.,Darrell,T., and Pentland A. *Pfinder: Real-Time Tracking of the Human Body.* in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1997.
19. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., and Shafer, S. *Multi-Camera Multi-Person Tracking for EasyLiving.* in *IEEE Workshop on Visual Surveillance*. 2000. Dublin, Ireland.
20. Checka, N., Wilson,K.,Siracusa,M., and Darrell,T., *Mutlitple Person and Speaker Activty Tracking with a Particle Filter.* ICASSP 2004, 2004.
21. Wang, C., Griebel,S., and Brandstein M. *Robust Automatic Video-Conferencing With Multiple Cameras And Microphones.* in *IEEE International Conference on Multimedia*. 2000.
22. Venkatesh, A., *Computers and Other Interactive Technologies for the Home*, in *Communications of the ACM*. 1996. p. 47-54.
23. Koile, K., Tollmar,K., Demirdjian,D., Shrobe,H, and Darrell,T. *Activity Zones for Context-Aware Computing*. in *Ubicomp 2003*. 2003.
24. Hudson, S., et al. *Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study.* in *CHI Letters (Proceedings of the CHI '03 Conference on Human Factors in Computing Systems)*. 2003.
25. DiBiase, J., Silverman,H., and Brandstein,M.,, *Robust Localization in Reverberant Rooms*, in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S.B.a.D. B.Ward, Editor. 2001, springer.
26. Bian, X., J.M. Rehg, and G.D. Abowd, *Sound Source Localization in Domestic Environment*. 2004, GVU center, Georgia Inst. of Technology.
27. Rui, Y.a.F., D. *New direct approaches to robust sound source localization.* in *Proc. of IEEE ICME 2003.* 2003. Baltimore, MD.