Using Sound Source Localization in a Home Environment

Xuehai Bian, Gregory D. Abowd, and James M. Rehg

College of Computing & GVU Center, Georgia Institute of Technology,
801 Atlantic Drive Atlanta, Georgia 30332
{bxh, abowd, rehg}@cc.gatech.edu

Abstract. In this paper, we examine the feasibility of sound source localization (SSL) in a home environment, and explore its potential to support inference of communication activity between people. Motivated by recent research in pervasive computing that uses a variety of sensor modes to infer high-level activity, we are interested in exploring how the relatively simple information of SSL might contribute. Our SSL system covers a significant portion of the public space in a realistic home setting by adapting traditional SSL algorithms developed for more highly-controlled lab environments. We describe engineering tradeoffs that result in a localization system with a fairly good 3D resolution. To help make design decisions for deploying a SSL system in a domestic environment, we provide a quantitative assessment of the accuracy and precision of our system. We also demonstrate how such a sensor system can provide a visualization to help humans infer activity in that space. Finally, we show preliminary results for automatic detection of face-to-face conversations.

1 Introduction

Since the early 1990's much research effort has been focused on how to acquire, refine, and use location information [11]. Location-sensing systems rely on either explicit tagging of individuals or objects that facilitate tracking, or they leverage implicit characteristics. Most implicit localization systems use computer vision to track users. We are interested in the use of sound source localization (SSL), techniques that extract the location of prominent sound events, in the home environment. Sound events are often associated with human activities in the home, but few have exploited location of sound as context, particularly in a home environment.

There are appropriate social concerns when sensing video and audio in the home environment. However, 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 SSL system to locate sound events in the environment using microphone arrays. The only information extracted in this case is 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 [15]. In Section 3 we describe the engineering modification we made to this standard algorithm to make it function more robustly in the public space of a

H.W. Gellersen et al. (Eds.): PERVASIVE 2005, LNCS 3468, pp. 19 – 36, 2005.

realistic home setting. The system runs continuously (24/7) and feeds the detected sound events into a database that can be consulted by a variety of applications for the home. For example, in other recent work, we have empirically established the link between face-to-face conversations in the home and availability to external interruptions from distant family members [18]. With this motivation, in this paper we wanted to explore whether temporal and spatial patterns of SSL events might be used to infer face-to-face conversations automatically.

To demonstrate the feasibility and usefulness of a home-based SSL system, we will describe how we engineered a solution based on previously published algorithms, focusing our discussion on modifications that would happen in any home environment. We will demonstrate how good the SSL system is in practice through experimental validation of its accuracy and precision over a large public space in a home, including kitchen, dining and living room areas, but with a controlled sound source. We will try to provide an honest appraisal of the SSL system we built, and attempt to discuss what advantages and disadvantages exist for this technology. Inferring activity, such as conversation, can be done by humans through visualization of the SSL data over relevant intervals of time. More automated forms of conversation detection will not be as robust, but we will show some initial promise in this area that leverages simple spatio-temporal heuristics. We believe this is a promising start, which points to more sophisticated activity recognition based on audio sensing.

2 Motivations: Sound as an Implicit Location Source

An interesting distinction between location-sensing technologies is the reliance on explicit means of marking the people or objects to track. We give a brief overview of location solutions, divided between those that require explicit tagging and those that function based on more implicit means of identifying tracked objects or people.

2.1 Explicit Location Systems

Many explicit localization systems, requiring users to wear extra passive tags or active devices, have been developed since the 1990's. Hightower and Borriello's recent taxonomy of current location systems mainly focuses on explicit localization systems [11]. 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 [9]. 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 [10]. 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 [19]. However, these systems are generally expensive to deploy and maintain.

We designed our own indoor location service using passive RFID. Users wear passive RFID tags which are queried by RFID readers at fixed locations to obtain a unique ID [1]. 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. An alternative method is to tag the environment and place the RFID reader on a person, as suggested by the iGlove used in the SHARP project at Intel Research [14]. Although the iGlove is used to detect hand-object proximity in order to infer activities of daily living (ADLs), this approach also infers locations.

By taking advantage of existing radio beacon infrastructure, such as WiFi access points, 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 [4]. This technique has been explored at length in the mobile and ubiquitous computing communities and 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 a GPS receiver and get the global position at the accuracy of 1-5m. Many projects use GPS as the primary outdoor positioning system, including Lancaster's Guide [6]. Intel Research's Place Lab effort leverages different methods of localization including GPS, WiFi, Bluetooth and GSM cell towers to provide increasingly ubiquitous location services [13].

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. In more formal environments such as the office, the wearing of an explicit tag or badge can be mandated. However, our experience deploying a passive RFID in a home laboratory showed that one important reason why the system was not extensively used is that some users forget to wear or even lost their tags. Another 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.

2.2 Implicit Location Systems

These disadvantages of explicit location tracking techniques motivate others to consider more implicit forms of location sensing. Here, technologies take advantage of natural characteristics of the users to sense their location, including visual cues, weight, body heat or audio signals. Implicit tracking does not require users to wear tags or carry devices, which pushes the tracking technology, for better or worse, into the background.

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. Simple sensors (motion detectors, contact switches, accelerometers) can be spread throughout the fixed infrastructure of a home (walls, cabinets, etc.), and the data from these sensors can be used to infer where human activity takes place [23].

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 that have the ability to monitor large spatial areas passively 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 treat the body holistically as a single moving target [8], often using a "blob" model to describe the targets' appearance. 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 [8, 17, 26]. 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 was 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 [17].

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 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.

2.3 Implicit Vision-Based Tracking Versus Sound Source Localization

Vision-based tracking and SSL are potentially more accurate than other simple implicit location systems, such as 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 1.54 Mbyte of data per second.

In comparison, the data throughput of a microphone array is significantly less than a camera system. One microphone generates about 88.2 KByte of data per second for CD quality sound with 16 bit sampling resolution. Because of the relatively low bandwidth, data collection and processing of an array of 16 to 32 microphones can be easily performed on an Intel PIII-class processor. Because of lower bandwidth requirement of audio, projects like "Listenin" report the ability to monitor remote environments though a wireless IP connection in real time [22].

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 localization 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 [20] and human tracking [5]. For example, the initial localization of a speaker using SSL can be refined through the use of visual tracking [25].

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 [24], many are connected with sound events, such as when we converse, watch TV, listen to the radio, talk on the phone, walk across the floor, move chairs to sit down for dinner, set plates on the dinning table, cook dinner, wash dishes, use utensils and chew during a meal.

In domestic environments, different activities are often conducted in particular locations. Leveraging this activity localization, a semi-automated method can be used to divide the room into activity zones and provide interaction based on status with regard to different zones [16]. Similarly, a previous study suggests availability for interruption from outside the home is strongly correlated to activities within the kitchen [18]. For instance, individuals preparing food at the kitchen counter indicated they would be accessible, but not available when helping a child with homework at the kitchen table. Sensing systems with only room location and presence context do not provide enough information to distinguish these differing states. SSL could provide precise location and sound event data to help a remote family member distinguish the different activities, without revealing the actual content or identities.

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. Thus detecting kitchen activities, such as cooking and washing dishes etc., will have promise in predicting availability for inter-home communication. 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 seated or standing heights.

SSL has potential to summarize 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 in an office environment [12].

Despite the potential for sound location to support relevant activity context, there is little to no research designed to investigate the relationship between sound events and household activity.

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. 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 as well as the design decisions we implemented to overcome these challenges. Finally, we report on the performance of our SSL system.

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 [7] and our own experiences [3], 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 example, 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. **Broadbandness.** 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 sound to be detected is usually intermittent and non-stationary. This makes it hard to apply localization techniques that use stationery source assumptions, such as adaptive filtering localization [2]. Sound generated by a person tends to be fairly directional, since the acoustic radiation in some directions is blocked by the human body.
5. **Multiple Simultaneous Sound Sources.** When faced with multiple simultaneous sound sources, there would be multiple peaks in correlation between microphones. This decreases credibility of computed time of delay and increases location errors.

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 that can be used to infer activity both manually and automatically.

3.2 Fundamentals of Passive Sound Localization

SSL systems can be traced back to earlier active radar and sonar localization systems. An active localization 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 signals

generated by the targets, which are mostly human generated sound signal in our case. 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 the 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: steered-beamformer based locators; high-resolution spectral estimation based locators; and Time-of-Delay based locators [7]. Most current sound source location systems are based on computing Time-of-Delay using PHAT-based filtering, which is simple, effective and suitable for real-time localization in most environments. The Time-of-Delay 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 [3].

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

$$x_i(t) = \alpha_i s(t - \tau_i) + n_i(t) \tag{1}$$

where: $s_i(t-\tau_i)$ is the signal delay; $n_i(t)$ is the noise; α_i is the attenuation factor for microphone *i*. For every pair of microphones, we compute the correlation. This is usually done in the frequency domain in order to save time. However, because of noise and reverberation in the environment, some weight functions in the frequency domain are applied to enhance the quality of the estimation, such as the Phase Transform (PHAT), or Roth Processor [15, 21]. The general cross correlation with the PHAT filter is equation 2, where the first item is the frequency weighting filter.

$$\hat{R}_{x_1 x_2}(t) = \int_{-\infty}^{\infty} \frac{1}{\left| X_1(f) Conj(X_2(f)) \right|} X_1(f) Conj(X_2(f)) e^{j2\pi ft} df \tag{2}$$

Conj is the complex conjugation function, and $X_1(f)$ and $X_2(f)$ are the Fourier transforms of $x_1(t)$ and $x_2(t)$. Ideally the maximum of equation (2) indicates the time offset of arrivals of the two signals if there is no noise. In practice, we compute the location of this peak by finding the time of occurrence of the maximum value of $\hat{R}_{x_1 x_2}(t)$.

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 that environment. Figure 1 shows the floor map of the target area.

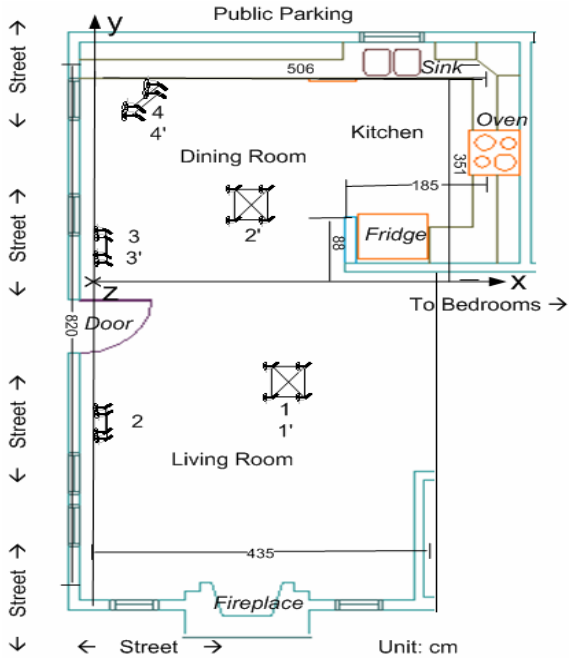

Fig. 1. 2D Floor map of covered space in the first floor with the primary "living room" setting(Quad 1,2,3,4) and "Kitchen" setting(Quad 1',2',3',4'). The only difference between the two settings is the location of the quad labeled 2 and 2'. Each Quad has 4 microphones

Deploying a working SSL system in a home environment introduces several challenges that many researchers do not consider when testing in a controlled laboratory environment. Below is a discussion of those challenges, and how we improved our system in light of those factors in the house where we deploy our system.

1) The house is close to a busy street and the noise level is variable throughout the day, so we dynamically update the noise threshold with equation (3) in processing before actual localization to ignore street noise.

$$Threshold_{t+1} = \alpha * Threshold_t + (1-\alpha) * Energy_t \quad (3)$$

   $Energy_t$ is the current sound energy and α is the inertial factor.
2) Our target area consists of a living room, dining room and kitchen. To cover the large space 16 microphones were used. The microphones were organized into 4 separate Quads (set of 4 microphones in a rectangle 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. Because of the high signal to noise ratio, a Quad selection strategy is needed.

Environmental factors such as noise and reverberation, etc., might corrupt the signal and generate maximum at the time other than true time of delay in equation (2). In addition to the above measures, we also found 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-Delays. Specifically, we define the peak-weight of ***i***th pair of microphones to be:

$$W_i = 1 - \mathrm{V}_{\text{Second peak}} / \mathrm{V}_{\text{Max Peak}} \tag{4}$$

We discard the data items whose peak-weights (W_i) are less than some 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 steepest gradient descent method in the 3D space during the process of minimizing the evaluation function. The final evaluation function *E* of each potential location is calculated by equation (5). Note that we consider the peak weight (W_i) in the final evaluation function.

$$E = \sum_{i \in PossilbePairs} [W_i (TDOA^{Exp}{}_i - TDOA_i)^2] \tag{5}$$

$\mathrm{TDOA_i}$ is the measured time of delay. $TDOA^{Exp}{}_i$ for a potential location is computed according to the distance to the i^{th} pair of microphones divided by speed of sound. 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

Although our system design is not sophisticated enough to work in different environments, it does work well in our target home. The dimensions of the L-shaped area are shown in Figure 1, and the overall area is about 38 square meters. The environmental noise ranges between 60-70db during the day, which is probably more noisy than a typical residential house setting. Our home lab is close to a busy street with much traffic.

Figure 2 shows the pictures of microphones mounted in picture frames and ceiling tiles, each of which has exactly 4 microphones and is called a Quad. The microphone arrays are deployed in the connected areas including the kitchen, living room and dining room on the ground floor (see Figure 1). We are using 16 omni-directional pre-amplified microphones (cost: 10 USD each) that receive audio signals in the 20Hz to 16KHz range. For the initial primary "living room" placement of microphone quads, shown as 1,2,3,4 in Figure 1, our goal was to cover the whole space equally well. For that setting, one Quad is in the ceiling of living room, two are on the front wall and one is at the corner of the dining room that faces the dining room and kitchen.

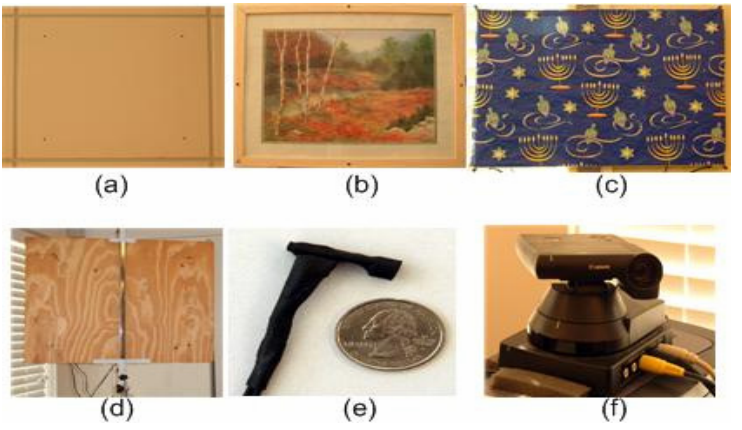

Fig. 2. (a)-(d) Microphone Quads, each has 4 microphones. (e) Single microphone with a quarter. (f) PTZ camera driven by detected sound location events

3.5 System Performance

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. For continuous talking that faces one of the quad arrays, we estimate the update rate as 1-5 seconds per reading. The system is especially responsive to crisp sound events such as clicking, eating, sniffing, or putting a backpack on a table. These crisp sound events generate high SNR signal which will help find correct Time-of-Delays between microphones.

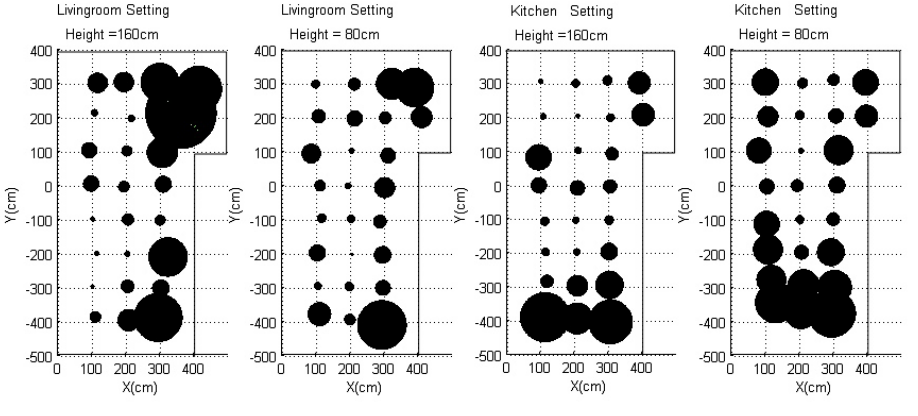

Fig. 3. Visualization of measured sound source location data at 1 meter by 1 meter grid for two different microphone Quad settings at heights of 80 and 160 cm

Table 1. The performance of SSL system for different settings with 25 measurements at each location. ERR is the average deviation with the true sound source location. STD is the standard deviation for 25 measurements

#	Grid Location		Living room 160cm		Living room 80cm		Kitchen 160cm		Kitchen 80cm	
	x	y	ERR	STD	ERR	STD	ERR	STD	ERR	STD
1	100	-400	19	16	27	34	25	73	72	60
2	100	-300	2	5	8	11	23	19	35	43
3	100	-200	14	7	4	24	15	12	30	43
4	100	-100	5	6	17	13	13	13	16	38
5	100	0	5	23	13	16	6	23	7	22
6	100	100	10	23	20	28	20	38	19	37
7	100	200	14	10	8	20	8	9	9	30
8	100	300	18	29	4	12	6	7	2	39
9	200	300	7	29	14	18	3	12	10	15
10	200	200	17	10	15	23	10	6	7	13
11	200	100	4	15	7	8	10	10	7	8
12	200	0	8	16	8	9	14	22	7	18
13	200	-100	9	17	6	11	6	10	4	12
14	200	-200	6	8	7	4	8	10	7	21
15	200	-300	7	19	6	13	12	31	14	46
16	200	-400	12	32	6	16	21	46	32	50
17	300	-400	29	71	14	72	24	65	24	70
18	300	-300	8	25	5	22	7	41	7	50
19	300	-200	45	57	9	25	4	25	8	40
20	300	-100	6	15	15	19	9	13	3	19
21	300	0	11	24	13	30	7	20	20	24
22	300	100	12	45	19	22	14	19	33	43
23	300	200	18	43	6	18	7	12	11	22
24	300	300	3	56	31	46	10	14	9	17
25	400	300	25	67	32	57	10	33	23	38
26	400	200	40	104	11	31	9	34	12	33

To test the accuracy and precision of our deployed SSL system, we created a 1 meter by 1 meter grid on the floor of the kitchen/dining/living room area and systematically placed a controlled sound source at two different heights (80cm, approximating table height, and 160cm, approximating standing height), for a total of 26 data collection points. At each sample point, we collected 25 independent SSL readings from a speaker on a tripod producing a crisp, clicking sound.[1] For the primary "living room" setting, we use microphone quad number 1, 2, 3 and 4 shown in Figure 1. In addition, we have tested the performance of a "kitchen" setting which includes microphone quad number 1', 2', 3' and 4' in Figure 1.

[1] This sound source was selected to allow for quick collection of data at the grid points. Using a more natural sound source, such as a recording of a person talking would not, in our opinion, change the results of the accuracy and precision readings, but would have greatly increased the time required for data collection.

The data collected is summarized in Table 1. Figure 3 shows a top-view visualization of the data collected to give a better at-a-glance demonstration of the accuracy and precision of our SSL system. For each grid point, a circle indicates the accuracy (how far away the center is to the grid point) and precision (radius of circle). We show this visualization at both heights (80cm and 160cm) for each setting of the microphone quads. Among all the measurements in our experiments, the mean of deviations from ground truth location is 13.5cm and 95% of the deviations are less than 33cm. We use the standard error of the 25 measurements to represent the accuracy of the system. Average standard error of all locations is 27.3cm and 95% of them are less than 68cm.

The experimental data verifies that to sense an area with less error and higher resolution, we need to put more microphones around that area. For example, in order to detect sound events better in the kitchen area, we placed a microphone quad in the kitchen ceiling. By doing so, we can enhance the SNRs of the signal and reduce the impact of reverberation. We can also see that between the height of 80cm and 160cm there is no systematically definable performance difference, though there are noticeable differences.

In general, our results demonstrate that the 4 microphone quad arrays give good coverage of this large area, and can provide data that can help determine activity in this space, as we will demonstrate next. This SSL system will help us to disambiguate the placement of relatively stationary or slow moving sound sources.

4 Using SSL to Detect Conversations

An initial application, developed to test how well 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. The advantages of the SSL sensing system we have created is that it covers a fairly large portion of the public living space of a home (kitchen, dining and living rooms) and offers reasonably good accuracy and precision for 3D location without requiring any explicit tagging. The SSL system also only records the location of the single loudest sound source every few seconds, without any other identifying characteristics being archived. Given the justifiable concerns with sensing and privacy in the home, this is a good feature of the sensor. The main disadvantage of this sensor is that location readings from the system are sporadic, with no guarantee of providing data at a fixed data rate. The same sound might not be consistently detected over time because of other environmental noise that cannot be controlled.

Given these advantages and disadvantages, we wanted to explore a use of the SSL technology that would play to its strengths. Using it for any real-time context-aware application would be unwise, given the sporadic nature of the data. However, it would be useful for near-term decision making, as the pattern of sounds in a home should reveal some important characteristics of activity. Previous research on communication support has revealed the potential for using near-term knowledge of home activity to determine whether the household is amenable to an outside interruption, such as an external family member phoning [18]. While the general

capability of determining availability is so subjective as to be impossible to mechanize, this prior result does give some direction for applying our SSL technology. One specific finding in that work suggests that detection of face-to-face conversations, which we define as a conversation being held in the same room of a house, is a good first step.

We address this problem of conversation detection in two ways. First, we look at ways in which the visualization of SSL data over a physical space might inspire a human to make correct inferences about activity, whether looking for conversations or other patterns. Second, we look to implement simple spatial and temporal heuristics that might detect conversational patterns in a time series of SSL data. We present both of these uses of SSL in this section.

4.1 Visualizing SSL Data

In a domestic environment, the owner of a home might be interested in viewing what activities happened in his house yesterday morning, see a summary of activity over longer periods, or remotely access the house of another trusting friend or family member. We developed a sound event map to facilitate this. In the sound localization system described earlier, all the sound location data with timestamps are stored into a database server. The sound event map application connects to the server and retrieves the sound location history. 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, meaning this information could be quickly transmitted outside of the home as needed.

The application, shown in Figure 4, allows a 3D manipulation of the floor plan, with SSL data points distributed throughout the 3D virtual 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 is familiar with the space, so even if there is not very detailed information about furniture, and certainly without knowledge of who might be in the room at any given time, the distribution of sound events can still be meaningful.

The user can select a time interval 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. Within a displayed interval, SSL events are colored from green (least recent) to red (most recent) depending upon their age. Another mode provides a form of replay so that the viewer can see a soundscape unfold over time, under their control. During the automatic replay, the current event is highlighted with the largest size dot, while the five previous events are rendered with smaller dots.

Though we have limited use of this application and cannot, therefore, report on how accurately a human can interpret activity in a familiar space using SSL data alone, our preliminary experiences reveals it is effective for summarizing activity over a reasonable period of time, usually from several minutes to several hours in duration. With this limited visualization, for example, it is possible to detect a moving sound source or alternating sound sources, as you would expect in a conversation. It is also possible to visualize and understand activity around special places, like a dining table. This visualization motivated us to look at more automated ways to detect these activities.

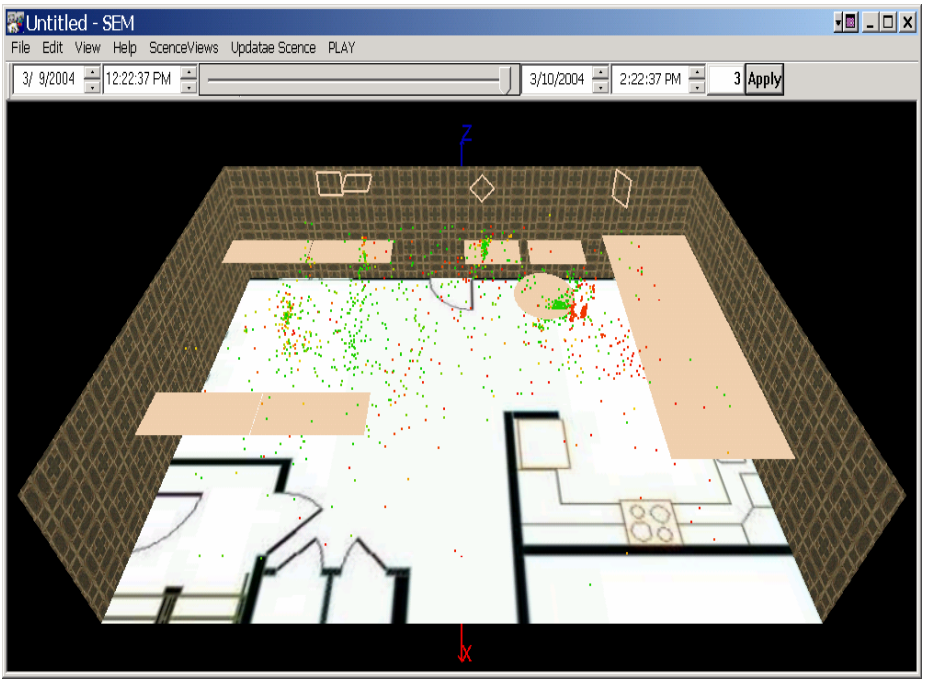

Fig. 4. Sound Event Map application, showing sound events (red/green dots) for a day in the home

4.2 Distinguishing Two-Person Conversations from Single-Person Talking

As previously discussed, 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 recognize dinning activities manually by looking for events around the dining room table. Kitchen related activities are also easy to recognize. While this general direction for activity recognition is an interesting one to pursue, we focus on a simpler kind of activity, specifically how to differentiate between a single person talking and a two-person, face-to-face conversation.

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 were distributed over the kitchen, dining and living room areas covered by the SSL system (using 1, 2, 3, 4 microphone quad configuration in Figure 1). The distances between the two people in conversation ranged from 0.5m to 5m. For a single person talking on the phone, we recorded 5 situations in which the phone conversation is relatively stationary (person sitting, but able to sway less than 1m) and 5 situations in which the user paced around the house (representing a cordless handset). Three typical cases, with corresponding SSL data are visualized in Figure 5. Each black point in the graph represents a detected sound event. To better

visualize the timestamp properties, we use both darkness and size to represent the age of each dot. The more recent events are drawn in larger sizes as well as darker colors. Each activity lasted between 2-5 minutes.

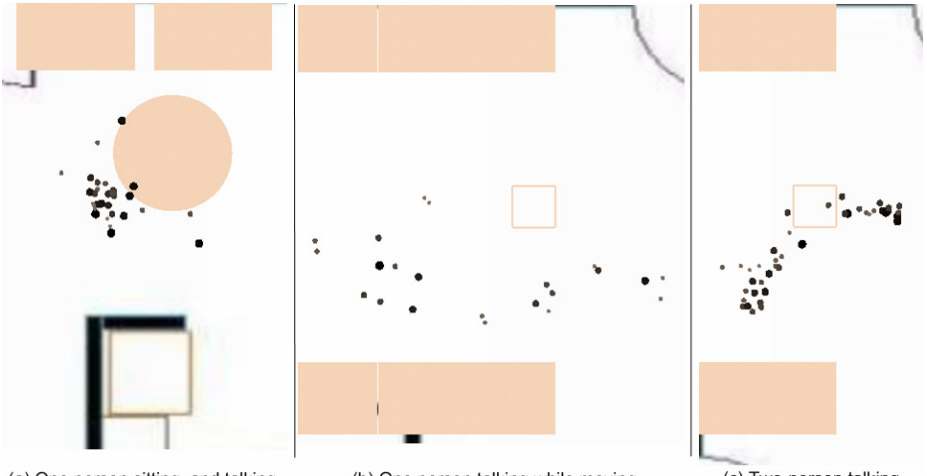

Fig. 5. Three talking scenarios. The darkness and size of dots represent the age of sound events

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

Cases	Total number of readings	Number of flip-flops	Distance between clusters(cm)	Classified as 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	32	2	340	No
1-person mv 2	32	5	118	No
1-person mv 3	43	2	322	No
1-person mv 4	22	8	201	Yes
1-person mv 5	35	7	283	No
1-person fix 6	48	22	69	No
1-person fix 7	34	4	101	No
1-person fix 8	31	12	97	No
1-person fix 9	39	18	61	No
1-person fix 10	36	19	64	Yes

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.

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 require 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-person and a single-person talking on the phone in the home environment. More sophisticated linear dynamic models could 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, highlighting the potential for more sophisticated inference algorithms.

5 Conclusions and Future Work

In this paper we summarized two different categories of localization systems, explicit and implicit, and pointed out that implicit localization systems have advantages for deployment in a ubiquitous computing environment. By adapting current sound source localization (SSL) algorithms, we built a sensor system in a realistic home setting. We demonstrate the accuracy and precision of this 3D localization technology, resulting in a system that is accurate to within 13.5cm with an expected standard error of 27.3cm in average in a realistic home setting.

While this sensor system alone certainly has its limitations, based on the latency and potentially sporadic distribution of data for a noisy environment, the simplicity of the sensor (from the human perspective) and its socially appealing lack of archived

rich data, motivated us to explore what value it might have on its own. We explored solutions that would use time series of SSL data points to help a human infer (through visualization) or be automatically informed of (through pattern recognition) the likelihood of human conversations in the home space. 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 might also find interesting patterns such as who is dominating the conversation. Both as a single sensor modality, and in concert with other sensed data, SSL shows promise for the complex and compelling problem of automated domestic activity recognition.

Acknowledgments

This work was sponsored in part by the National Science Foundation (ITR grants 0121661 and 0205507) and the Aware Home Research Initiative of Broadband Institute at Georgia Tech. The authors thank our colleagues in the Ubicomp Group at Georgia Tech for their helpful input to this paper, particularly Jay Summet and Kris Nagel.

References

1. Abowd, G.D., Battestini, A., and O'Connell, T. The Location Service: A framework for handling multiple location sensing technologies. GVU technical report, Georgia Institute of Technology Technical Report (2002)
2. Benesty, J. and Elko, G.W. Adaptive eigenvalue decomposition algorithm for real-time acoustic source localization system. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP1999).* (2001) 937-940
3. Bian, X., Rehg, J.M., and Abowd, G.D. Sound Source Localization in Domestic Environment. GVU center, Georgia Inst. of Technology Technical Report GIT-GVU-04-06 (2004)
4. Castro, P., Chiu, P., Kremenek, T., and Muntz, R. A Probabilistic Location Service for Wireless Network Environments. In *Proceedings of Proceedings of Ubicomp 2001.* Atlanta Springer Verlag, (2001) 18-24
5. Checka, N., Wilson, K., Siracusa, M., and Darrell, T. Multiple Person and Speaker Activity Tracking with a Particle Filter. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing.* (2004)
6. Cheverst, K. Developing a Context-Aware Electronic Tourist Guide: Some Issues and Experiences. In *Proceedings of Proc. 2000 Conf. Human Factors in Computing Systems.* New York ACM Press, (2000) 17-24
7. DiBiase, J., Silverman, H., and Brandstein, M., *Robust Localization in Reverberant Rooms*, in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D.B. Ward, Editors. 2001, Springer.
8. Haritaoglu, I., Harwood, D., and Davis, L.S. W4: Real-Time Surveillance of People and Their Activities. In *Proceedings of IEEE Trans. On PAMI.* (2000)
9. Harter, A. and Hopper, A., A Distributed Location System for the Active Office. IEEE Network. **8**(1) (1994) 62-70

10. Harter, A., Hopper, A., Steggles, P., Ward, A., and Webster, P. The Anatomy of a Context-Aware Application. In *Proceedings of Int'l Conf. Mobile Computing and Networking (MobiCom 99).* New York ACM Press, (1999) 59-68
11. Hightower, J. and Borriello, G., Location Systems for Ubiquitous Computing. Computer,IEEE Computer Society Press. **34**(8) (2001) 57-66
12. Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. Predicting Human Interruptability with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of CHI Letters (Proceedings of the CHI '03 Conference on Human Factors in Computing Systems).* (2003) 377-384
13. Intel Research Place Lab Project. http://www.placelab.org (Downloaded on Oct. 6, 2004).
14. Intel Research SHARP Project. http://seattleweb.intel-research.net/projects/activity (Downloaded on Oct 6, 2004).
15. Knapp, C.H. and Carter, G.C., The generalized correlation method for estimation of time delay,". IEEE Transaction on Acoustics, Speech and Signal Process. ASSP-24. **24** (1976) 320-327
16. Koile, K., Tollmar, K., Demirdjian, D., Shrobe, H., and Darrell, T. Activity Zones for Context-Aware Computing. In *Proceedings of Ubicomp 2003.* (2003) 90-106
17. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., and Shafer, S. Multi-Camera Multi-Person Tracking for EasyLiving. In *Proceedings of IEEE Workshop on Visual Surveillance.* Dublin, Ireland (2000)
18. Nagel, K.S., Hudson, J.M., and Abowd, G.D. Predictors of Availability in Home Life Context-Mediated Communication. In *Proceedings of Computer Supported Collaborative Work, 2004.* Chicago, IL USA (2004)
19. Priyantha, N.B., Chakraborty, A., and Balakrishnan, H. The Cricket Location-Support System. In *Proceedings of Proc. 6th Ann. Int'l Conf. Mobile Computing and Networking (Mobicom 00).* New York, 2000 ACM Press, (2000) 32-43
20. Rehg, J.M., Morris, D.D., and Kanade, T., Ambiguities in Visual Tracking of Articulated Objects Using Two- and Three-Dimensional Models. Int. J. of Robotics Research. **22**(6) (2003) 393-418
21. Rui, Y. and Florencio, D. New direct approaches to robust sound source localization. In *Proceedings of Proc. of IEEE ICME 2003.* Baltimore, MD (2003)
22. Schmandt, C. and Vallejo, G. "LISTENIN" to domestic environments from remote locations. In *Proceedings of Proceedings of the 9th International Conference on Auditory Display.* Boston University, USA (2003)
23. Tapia, M., Intille, S.S., and Larson, K. Activity Recognition in the Home Using Simple and Ubiquitous Sensors. In *Proceedings of Proceedings of Pervasive 2004: the Second International Conference on Pervasive Computing.* Springer, (2004)
24. Venkatesh, A.: Computers and Other Interactive Technologies for the Home, in *Communications of the ACM*. (1996). p. 47-54
25. Wang, C., Griebel, S., and M., B. Robust Automatic Video-Conferencing With Multiple Cameras And Microphones. In *Proceedings of IEEE International Conference on Multimedia.* (2000)
26. Wren, C., Azarbayejani, A., Darrell, T., and A., P. Pfinder: Real-Time Tracking of the Human Body. In *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence.* (1997)