Audio Location: Accurate Low-Cost Location Sensing

James Scott and Boris Dragovic

Intel Research Cambridge,
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
james.w.scott@intel.com
boris.dragovic@cl.cam.ac.uk

Abstract. Audio location is a technique for performing accurate 3D location sensing using off-the-shelf audio hardware. The use of off-the-shelf hardware allows audio location deployment to be low-cost and simple for users, as compared to other currently available centimetre-scale location sensing systems which have significant custom hardware and installation labour requirements. Another advantage of audio location is the ability to locate users who are not carrying any special location-enabling "tag", by using sounds that the user themselves can make such as finger clicking. Described herein are the various design parameters for audio location systems, the applicability of audio location for novel 3D user interfaces based on human sounds, and a quantitative evaluation of a working prototype.

1 Introduction

Location-aware computing [1] covers many different location granularities, from applications requiring city-scale location accuracy (e.g. online yellow pages), to others which operate on the centimeter or even millimeter scale (e.g. augmented reality). At the heart of location-aware computing is the body of research in location sensing, which has mirrored the wide range of location granularities, from systems such as RightSPOT [2] offering kilometer-scale accuracy to highly accurate systems such as the ultrasonic Bat [3] with its 3 cm accuracy.

However, high accuracy is not the only metric by which to judge location systems. The coverage of a location system is also of great importance. As one would expect, the most accurate location systems also exhibit the lowest coverage; for example, the Bat system is deployed in a portion of a single building, while GPS has worldwide coverage. Recent work on the Place Lab system [4] has focussed on providing high-coverage location information using WiFi-based location, in which the previous work was confined to single buildings [5]. However, such radio beacon–based location is limited in accuracy to tens of metres.

A location system with centimetre-scale accuracy and wide coverage has not yet been achieved. This is due to the high costs inherent in the non-standard hardware required for such systems, and in the installation and maintenance of this hardware. These costs pose too steep a barrier for many potential location-aware application user communities, outweighing the benefits of the applications. The work presented in this paper is motivated by the desire to reduce the deployment costs for potential users of

H.W. Gellersen et al. (Eds.): PERVASIVE 2005, LNCS 3468, pp. 1–18, 2005.

location-aware applications, and thereby enable the wide deployment of high-accuracy location-aware systems.

This paper presents “audio location”, a technique enabling standard audio hardware to be used to locate people and objects with centimetre-scale accuracy. One defining feature of audio location is that it is possible to implement both “tagged” and “untagged” location systems using this technique, i.e. systems where users are required to carry special devices (“tags”) in order to be tracked, as well as systems which have no such requirement and operate by making use of sounds the user themselves produce.

Section 2 will explore the design space for audio location systems, and previous work in this area. Section 3 will present a prototype system using audio location to implement a 3D user interface based on human sounds such as clicking fingers, including experimental results concerning the accuracy of audio location in this context. Section 4 will conclude the paper and outline future work in this area.

2 Audio Location

Audio location is a process whereby the time-of-flight of audio is used to determine the accurate location of people and/or devices. This section addresses the design parameters found in audio location.

2.1 Related Work

Sound source localization has been widely studied by the signal processing community. Two useful introductions to this work can be found at [6] (making use of microphone arrays), and [7] (using sensor networks). However, much of this body of work makes use of custom hardware, and is therefore unsuitable for low-cost and easily deployable location sensing. In contrast, this paper focuses on location sensing using off-the-shelf audio hardware; this area has been looked at by comparatively few research groups.

Girod et al. have developed a system for fine-grained range estimation [8, 9], making use of tags to produce sounds. The tags emit wideband chirps in the audible spectrum together with a RF reference signal to allow the receivers to estimate the time-of-flight accurately. The authors provide an in-depth discussion of the issues of audio signal propagation, chirp sequence correlation and sources of timing errors as well as an extensive set of evaluation results. However, untagged location is not discussed.

The work of Rosca et al., which does consider untagged operation, regards 3D audio-based positioning as a side-effect of a speech interface for use in virtual reality scenarios [10]. However, their discussion is purely theoretical and lacks an evaluation of the difficulties of implementation, in particular the difficulties of extracting a narrow feature of the audio signal for time-of-arrival calculation, and is based on an idealised scenario that may not stand up in real-life use.

Finally, the use of off-the-shelf audio hardware for context-aware applications in a pervasive computing setting, including coarse-grained location sensing, was recently presented by Madhavapeddy, Scott and Sharp [11]; the research presented in this paper was heavily inspired by their work.

2.2 Tagged Versus Untagged

Accurate location systems often require that the objects to be located be "tagged" with small devices which interact with the sensing infrastructure. There are many forms a tag can take, two examples being ultrasonic transmitters hung around the neck of a user or velcroed to a device, and visual barcode tags which can be attached to users' clothes or stuck to devices. Audio location can make use of tagged location sensing, using mobile or wearable devices such as phones to send or receive audio signals.

It is also possible to construct "untagged" systems, in which the user or device is located purely by means of their intrinsic properties or capabilities. One example is the Active Floor [12], which senses a person's body weight using load sensors under the floor. Not requiring tags has a number of advantages: hardware costs may be lower, users of the system do not have to remember to wear/affix tags, and new users do not need to be assigned tags to participate. One disadvantage of untagged audio systems is that the accuracy is likely to be worse than for tagged systems, in which the data exchanged between the mobile object and the stationary infrastructure can be well-defined (e.g. a high-contrast pattern in barcode-like systems such as TRIP [13]) as opposed to relying on what is available (e.g. the colour of the shirt a user is wearing, as used by the Easyliving system [14]).

Audio location may be used for untagged location sensing, since users are capable of generating sounds (e.g. finger clicking). An untagged audio location system will have to cope with two performance-degrading factors, namely the difficulty of detecting a suitable audio feature that can be identified at each microphone, and the lack of synchronisation as the time-of-send of the audio signal is not known. At least one previous research system, using the Dolphin ultrasonic broadband hardware presented by Hazas and Ward [15], has achieved unsynchronised fine-grained location, in which the time-of-send of the signal is determined during the location calculation.

2.3 Infrastructure

The infrastructure used to achieve audio location could involve the use of microphones in the environment and sound generation by the users/devices, or sound generation by speakers in the environment while users carry devices with microphones. This research focuses on the former technique, the most important reason being that this facilitates untagged operation while the latter technique does not.

In order to implement this type of audio location system, a number of microphones must be present in the environment, and they must be linked to one or more devices capable of processing the sound data to determine location. The simplest method of achieving this is to connect a number of low-cost off-the-shelf microphones to a single PC (which may already be in the room), and run the software on that PC. While the hardware cost may be quite low, this infrastructure may require the installation of long wires so that microphones can be optimally placed around the room.

In spaces where multiple computers are already present, e.g. shared offices, it is possible to consider making use of the sound hardware in all computers, where each computer provide only one or a few microphones. Since many PCs are already outfitted with a microphone for multimedia applications, this potentially reduces the cost of an audio location system, perhaps even to zero when deployed in a space that is already densely populated with PCs with microphones. This infrastructure also reduces audio wiring requirements, assuming that the PCs are spread across the space to be instrumented and that the PCs are networked.

2.4 Audio Feature Generation and Detection

In order to determine location, an audio location system must detect sounds as they appear in the data streams from multiple microphones. Furthermore, the system must identify a single feature of the sound which can be localised in each data stream, enabling the collection of a set of times-of-arrival for the same instant from the sound.

For tagged systems, the tag and infrastructure will communicate via radio. This allows the tag to inform the infrastructure of information such as its unique id, the characteristics of the signal it is sending (which may be implicit from the id), and the time-of-send of the signal. The infrastructure can then use this to search for the signal in the sound streams, to associate this signal with the correct tag, and to determine location more easily since the time-of-send is known.

For untagged systems, the infrastructure operates under much harsher conditions. While a tag-generated signal might conform to an easily-detectable format, a user-produced sound will not be so easy to detect. This means that the infrastructure must be able to detect a much broader class of signals, e.g. including sounds such as clapping/clicking of fingers which might be made deliberately for the benefit of the location system, and also sounds such as speech, typing, and others, which may be made by the users during their normal activities. One consequence of this is that the "noise floor" may be much higher, and might include sounds such as music, devices beeping or humming, vehicle noise, and so on. Possible signal detection methods might be based on monitoring each stream for amplitude spikes (e.g. for finger clicking), or on performing continuous cross-correlation between the sound streams, looking for spikes in the correlation coefficient to indicate the arrival of the same sound.

2.5 Location Determination

Timing information from multiple microphones must be gathered to determine location. In order to get some idea of how many microphones are required, one can regard the location problem as a set of equations in up to four variables: the 3D position of the sound source, and the time-of-send of the sound. Naively, a tagged system would require three microphones (since the time-of-send is known, only three variables need to be solved for) and an untagged system would require four. However, this gives no room for error resilience; an erroneous time-of-arrival at one of the microphones would pass undetected, and result in an erroneous location. With one extra microphone, an error situation could be determined, since the times-of-arrival would not "agree". However, it is difficult to determine which was the erroneous microphone without using external data such as the previous known location of the sound source.

To calculate the location, a non-linear system of equations is constructed in the time-of-send of the sound tos, times-of-arrival toa_i at each microphone i, microphone locations $\mathbf{micpos_i}$, the location of the sound source $\mathbf{soundpos}$, and the measurement errors err_i.

$$toa_i = tos + \frac{|\mathbf{micpos_i} - \mathbf{soundpos}|}{speedofsound} + err_i \quad (1)$$

The known quantities are then substituted, and the unknown quantities (sound location, and, in untagged systems, time-of-send) are found using an algorithm such as the Levenberg-Marquardt method [16] to minimize the errors err_i. This approach is similar to that used, for example, in the ultrasonic Bat location system.

2.6 Issues Affecting Location Accuracy

To obtain a precise location, attention must be paid to the placement of the microphones in the room. If all the microphones are co-planar (which may often be the case, e.g. when mounting them against a wall or ceiling), there are always two possible locations for the sound source: one on each side of the plane. This ambiguity can often be resolved by looking at the location of the walls/ceiling/floor of the room, if this is known, as one of the locations may be outside.

Another issue that influences placement of microphones is Dilution Of Precision (DOP). This issue concerns the relationship between the accuracy of a positioning system and the angular relationship between the transmitter and receivers. If all receivers occupy the same narrow angle from the point of view of the transmitter, then small errors in the distance estimates will translate into large errors in the 3D position. To combat this, microphones should be widely distributed in the sensing space so that they have a large angular separation from any position where a sound source may be located.

Location precision is also affected by the speed of sound, which varies according to many factors [17], but most significantly according to temperature and humidity. The change in speed is as much as six percent between cold and warm air, and up to half a percent between dry and humid air. Whether this is regarded as significant or not depends on the location accuracy demanded by applications for which a given system was deployed, and also on the ease of statically predicting these figures based on, for example, the time of day. If it is significant, then computer-readable thermometers (and possibly hygrometers) could be deployed with the microphones.

2.7 Surveying

"Surveying", i.e. the discovery of information about the environment, is important for fine-grained location systems in two ways. Firstly, a survey of the infrastructure (for audio location, the microphones) is required for location determination to function correctly. Secondly, a survey of the characteristics of the environment, i.e. the locations of walls, doors, furniture, fixed electronics, and so on, is needed to enable location-aware applications, e.g. the nearest-printer application needs to know where the printers are. Since surveying can potentially be very time-consuming, and therefore be a discouraging factor for fine-grained location deployment, these topics are described below.

Surveying of the infrastructure can be achieved in a number of ways. Manually measuring the locations with respect to a reference point in the room is the simplest method, but may be time-consuming, subject to human error, and the microphones would have to be firmly fixed since moving them would mean re-surveying is required. Automatic surveying systems are feasible, e.g. using laser range-finders or theodolites. However, such methods require expensive equipment, and again the microphones must be fixed. Finally, “self-surveying” techniques [18] can be used, by which a location system can construct its own survey given enough raw data. For real deployments, this means that the system would not return valid locations when it was activated, until enough data points were gathered for the system to survey itself. This time depends on how much “surplus data” is present in the system, which is proportional to how many extra microphones are installed over the minimum number described previously.

Environmental surveying can also be conducted entirely manually, but again this is a very time-consuming process, and furthermore is likely to be made out-of-date due to objects being moved. Semi-automatic methods such as using the location system to manually indicate the vertices of objects such as rooms and furniture are possible; while quicker than typing in locations manually, this also suffers from falling out-of-date. A final possibility is to use audio location to automatically detect and monitor objects in the room, either because they make sounds during use (e.g. speakers, printers, keyboards, mice), or because they reflect sounds (e.g. walls, large items of furniture) such that the reflections can be detected by the system (the latter method was described by Rob Harle for the ultrasonic Bat system in [19]). The advantages of automatic methods are that they are transparent to installers/users, reduce the deployment overhead, and that they can automatically maintain up-to-date locations. However, mature, reliable and scalable methods for performing environmental surveying have not yet emerged.

2.8 Identification

While the location of sounds is useful in itself, discovering the identity of the sound’s producer and associating it with that location enables a number of additional applications. This can be accomplished in tagged systems by simply having the tag declare its identity over radio, and provide enough information such that the infrastructure can determine which sounds it is making, which may include the time-of-send as well as information on the characteristics of the sound. For untagged systems, determining identity is much harder. One possibility is to use the sounds made to infer the identity of the object making them, e.g. performing voice pattern recognition on speech, using characteristic sounds such as the gait pattern during walking, habitual sounds such as tapping of fingers or distinctive laughter, and so on. However, these methods are not likely to identify users based on many common types of sound, e.g. clapping.

To solve this problem, sensor fusion methods can be used to draw from other sources of information, particularly those which provide accurate identification and inaccurate location (i.e. the complement of untagged audio location’s characteristics). One example of sensor fusion used in this way was shown by Hightower et al. [20], in which coarse-grained RFID tag identification was combined with a very accurate but anonymous laser range-finder. This technique could also be applied to audio location, and in this way both the audio-based identification methods above and other identification methods using technologies such as RFID, Bluetooth, and so on can be combined to generate highly accurate locations for identified entities.

While the above discussion shows that it is possible to design audio location systems which identify users, some may regard that not including this functionality is in fact desirable, for privacy reasons. For example, if audio location were used to control an information kiosk in a shop describing their products, many users may wish to use this facility anonymously, and might refrain from using it if they felt they were leaving themselves open to tracking.

3 3D Interfaces Using Human Sounds

Out of the design space for audio location described above, one application area was chosen to demonstrate some of the novel possibilities of audio location, namely 3D interfaces using human sounds. This is inherently an untagged system, since humans themselves are generating the sounds.

A sound-based 3D interface would enable new types of computer-human interaction, moving away from physical input devices (e.g. keyboards, remote controls, etc), and toward a situation where physical input devices are not required, and the user interface is implicit in the environment. When a user makes a sound at one or more 3D locations in a pre-defined pattern, the environment can perform actions such as controlling appliances (e.g. lights), navigating through data presentation interfaces (e.g. on a wall-mounted display), and so on.

In order to guide the user to make sounds at the correct locations, these locations can be highlighted by marking that place, e.g. using printed paper affixed to surfaces such as walls or desks. Given that such markers cost very little, and that the cost of the audio location hardware is low and is only incurred once at install time, this makes for a very low-cost input method, as compared to the cost of fitting a new device at every location. The audio interface is also easily reconfigurable, in that controls can be added or changed easily. Furthermore, audio interfaces benefit from not requiring physical contact; this is useful in environments where such contact is to be avoided, e.g. in hospitals to avoid the spread of infection.

3.1 Related Work

There are many 3D user interface input methods developed by the research community as well as available commercially. Many of these require users to be equipped with special hardware such as gesture-recognising gloves or ultrasonic Bat tags in order to function. Steerable user interfaces [21] have no per-user device requirement, but rely on expensive cameras and projectors. Tangible user interfaces [22] use cameras to detect movement of physical objects and thereby cause actions on virtual objects, e.g. rotating a map display in an image, but this relies on appropriate physical objects being present and on the user knowing how to manipulate each object to control the environment according to their wishes.

Vision-based gesture recognition systems [23] are perhaps closest to audio location, in that the user does not need any special hardware or devices, and in that it is possible to consider using cheap "webcams" to produce low-cost 3D interfaces (though much of the research presented uses top-of-the-line cameras with significant cost). The advantage of audio location is the very simple interaction method, allowing the user to have a good mental model of when the interface is activated. If the user does not make loud noises, they are sure that audio location will not be activated, whereas a non-expert gesture recognition user may be wary of accidentally making a meaningful gesture.

3.2 3D Audio Interface Primitives

There are various kinds of user interface components that can be achieved using audio location. The first and most obvious one is the "button", in which a user makes a sound at a specific location to indicate an action, e.g. the toggling of a light. A second type of interface could rely on simple gestures, where the user makes a few noises in succession, at slightly different points. An example of this would be clicking one's fingers at a given point, and then again slightly higher or slightly lower, with one potential application being a volume control. The starting point of a gesture could be precisely fixed, or it could be relaxed to a broader area, with the relative location of the two noises being used as the input primitive. While more complex types of interface are possible, e.g. based on making sounds with a changing tempo or amplitude, this may prove counterproductive as the interface becomes less intuitive for users.

It is also possible to consider dynamic audio location interfaces, in which audio location interface components are dynamically created in front of a computer display. These could be used to interface with computers using display types such as projectors, which are difficult to use alongside traditional input devices such as a mouse and keyboard.

To illustrate the potential uses of audio location for user interfaces, four possible application interfaces are shown in Figure 1, parts a to d, which respectively illustrate interfaces suitable for a light switch, a volume control, a web kiosk, and a photo album application displayed on a projector.

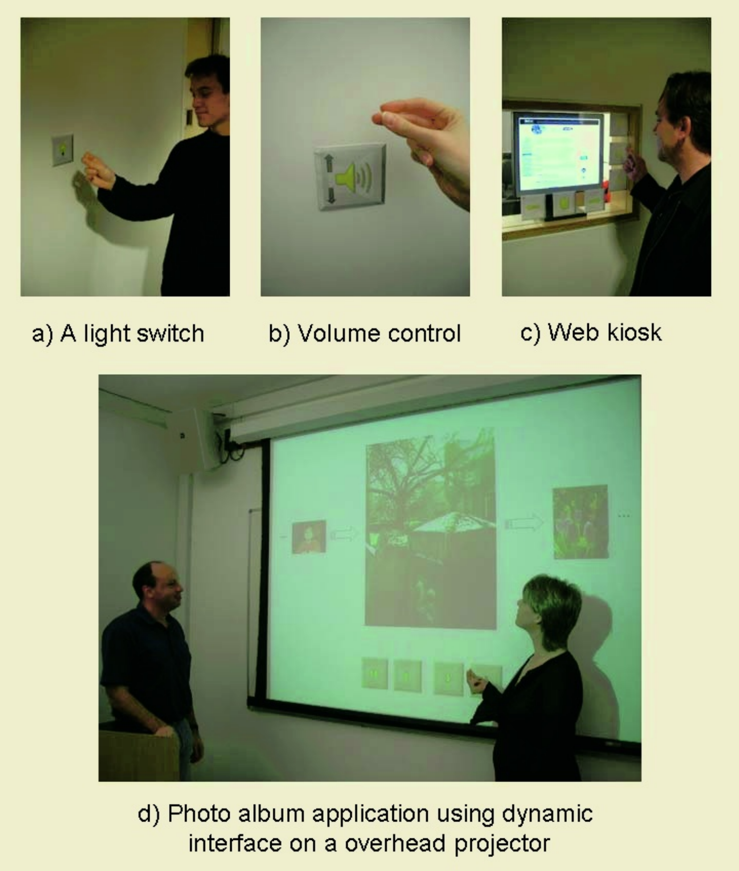
A 3D interface has previously been demonstrated using the ultrasonic Bat system for fine-grained location [24]. One advantage of using audio location is that no tag is required for each user. A disadvantage of audio location is that, by itself, it is not capable of identifying the user, while the Bat system does identify the tag being used at that location (and assumes that it is operated by its owner). However, as discussed in Section 2.8, the identification of users (if required for a given application) can be accomplished using other coarse-grained location methods, and the process of sensor fusion can be used to combine this information with the fine-grained information from audio location.

3.3 Prototype Implementation

An audio location prototype was implemented using a single PC with six low-cost PCI sound cards and six low-cost microphones. The total cost of the sound hardware required was around one hundred british pounds, orders of magnitude less than the custom components required by many location systems described earlier. No temperature sensor was incorporated in the prototype, since this would affect the cost and off-the-shelf nature of the system.

While location can be determined from just four microphones, the use of six microphones allowed the prototype to be robust against occlusion of the path to a microphone by the user, other people, or items of furniture. The provision of redundant data also enables detection and rejection of erroneous sightings.

The software architecture was implemented in Java as an extensible object-oriented framework. Signal detection uses a dynamic amplitude-threshold scheme, whereby each sound stream is monitored for a sound sample with amplitude significantly greater than



a) A light switch

b) Volume control

c) Web kiosk

d) Photo album application using dynamic interface on a overhead projector

Fig. 1. Four examples of 3D interfaces based on audio location

the current background amplitude[1]. Location determination is then accomplished by using the Levenberg-Marquardt algorithm [16] to find the location most closely matching the signals detected, by minimising the sum of squares of the errors in the relative distance estimates, using the Levenberg-Marquardt algorithm as described in Section 2.

The amplitude-threshold signal detection method was chosen since it is good at detecting impulsive sounds such as finger clicking or hand clapping, which users can choose to make when using the 3D interface. However, this algorithm does not detect continuous, low-amplitude, or non-impulsive sounds, including human speech, ambient music, and keyboard strokes. This property is invaluable for the 3D user interface application area, since a system sensitive to such sound sources would generate high levels of false positives when used in a normal home or office environment.

[1] For each sample time t, $BackgroundNoise_{t+1} = BackgroundNoise_t * 0.99 + Sample_t * 0.01$, and $Sample_t$ is marked as a “peak” if $Sample_t > 2 * BackgroundNoise_t$. Parameter values were found by trial and error, and might vary if using different hardware.

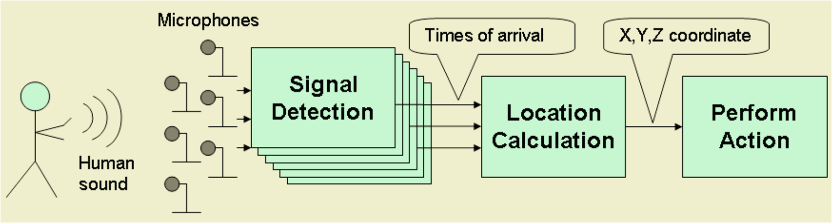


Fig. 2. Audio location–based 3D interface system architecture

One issue that became apparent in early testing was the problem of obtaining an accurate timestamp for a given sound sample, i.e. determining the time-of-arrival of a sound sample at the microphone. This is because there are potentially many delays and buffering points between the microphone and the Java application, including the delay for the hardware to raise an interrupt after its buffer becomes full, the delay for the interrupt to be serviced, and the delay for the Java application to be scheduled. In order to obtain an accurate timestamp, the driver for the chosen sound card[2] was modified such that it took a cpu-clock timestamp for new sound data at interrupt time, and made this available via the Linux `/proc` file system to the Java application. The modification required was modest, and is easily ported to other sound card drivers. Ideally, the sound card itself could maintain a timestamp; this would reduce the potential error in the timestamp from the current low value (the interrupt handling latency) to a negligible amount. It should be noted that Girod et al. [8] encountered similar timing accuracy problems, and suggested a similar solution.

3.4 Experiments, Results and Analysis

Two experiments were conducted in order to determine the performance of the system in one dimension and three dimensions. It is important to note that, in all of these experiments, actual human sounds were used — while recording and playback of such sounds using a speaker was considered, it was decided that this would not demonstrate the accuracy of the system as would occur in a real deployment.

1D Experiment. The 1D experiment investigates the performance of the prototype at the most basic level, namely the accuracy it exhibits in estimating a relative distance between a sound and two microphones.In addition, this experiment reveals how the microphones used perform at various ranges and various angles with respect to the sound.

The experimental setup is shown in Figure 3, and consists of two microphones placed 0.6 m away from, and pointing towards, a 7x6 grid of measurement points, with 0.6 m between each grid point. Both the grid and the microphones were at a height of 1 m above the floor. At each point on this grid, both a “finger click” and a “hand clap”

[2] The sound cards used were the C-Media 8738 model with the CMPCI chipset, chosen because it was the cheapest sound card available, at eight British pounds per card. For the same reason, the microphones used were the Silverline MC220G, also at eight pounds per unit.

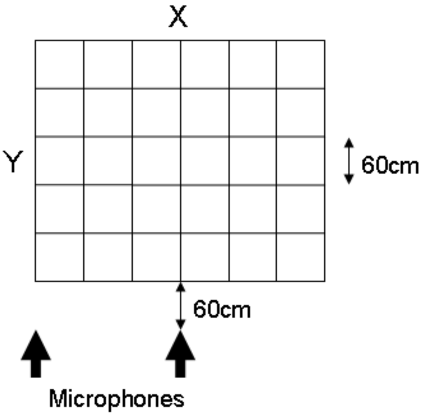


Fig. 3. Experimental setup for 1D experiment

noise was made twenty times, and the time difference of arrival at the two microphones was recorded. These time differences were then multiplied by the speed of sound to obtain estimates of the relative distance between the microphones and the grid point, which were then compared with the actual relative distance, resulting in a 1D distance error for each click/clap.

Figure 4 shows a surface plot of the median error in the relative distance over the 20 iterations, as plotted against each of the grid points used, with “finger clicking” used for sound generation. This illustrates the angular sensitivity of the microphones,

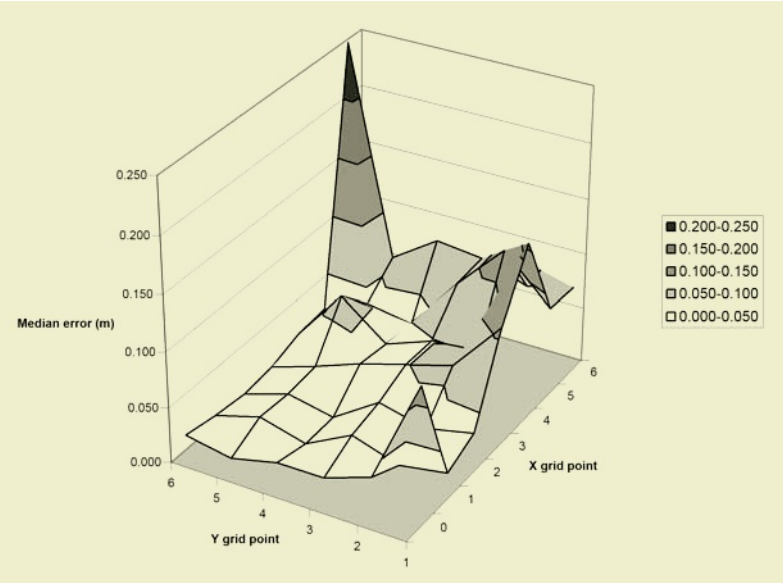


Fig. 4. 1D distance errors at each location: Clicking

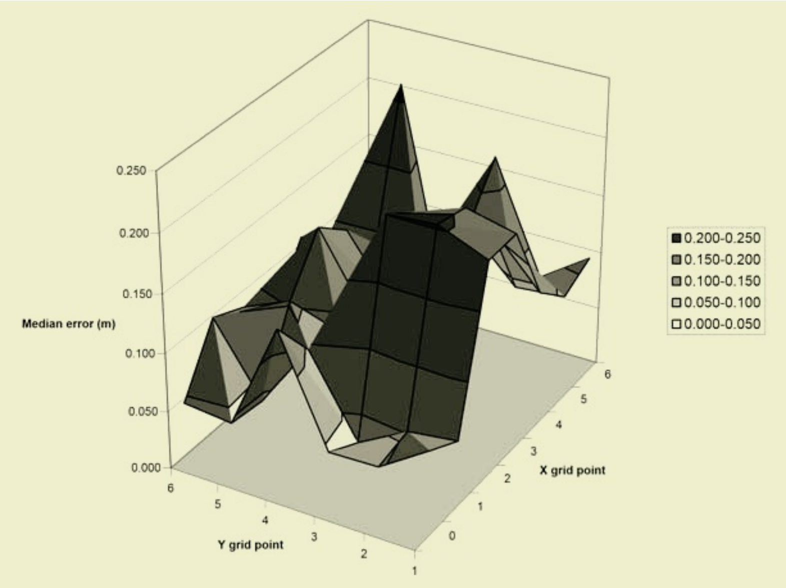


Fig. 5. 1D distance errors at each location: Clapping

as the tests at grid points with high X and low Y do not enjoy accurate relative distance estimates. Extrapolating from this plot, the microphones employed work best over an angle of around 60° either side of their axis; this information affects deployments of the prototype (allowing the installer to choose microphone locations well), and in particular contributed to the design of the 3D experiment described below.

The plot also shows that the microphones allow a location range of up to 4 m with low errors. The precise range of the system will depend on many factors, including the sensitivity of the microphones, the noise floor in the room (which was quiet during the tests presented), and the signal detection method used. In the conditions tested, errors appeared to increase at distances greater than around 4.3 m. These distances, however, are comparable to a typical office or home room. For larger spaces, it is possible to place more microphones around the room.

Figure 5 shows a similar surface plot for experiments using hand clapping for sound generation. It is obvious that the results are significantly worse. This is partly due to the character of the noise being made, which is less "impulsive" and hence harder to determine an accurate timestamp for. The other issue affecting location accuracy is the intrinsic location scale imposed by the use of hand-clapping — human hands measure $\tilde{1}5$ cm across; this imposes a precision limit of this order of magnitude. A human-imposed limit also applies to clicking, albeit at a higher precision of perhaps 5–10 cm. On this basis, the experimental results indicate that finger clicking is more suitable for location sensing; finger clicking was therefore decided upon as the sound generation method for both the 3D experiments and prototype deployments.

3D Experiment. The 3D experiment was conducted over a four-by-four grid with 0.6 m separating the grid points, and at each of three heights at 0.6 m, 0.9 m and 1.2 m.

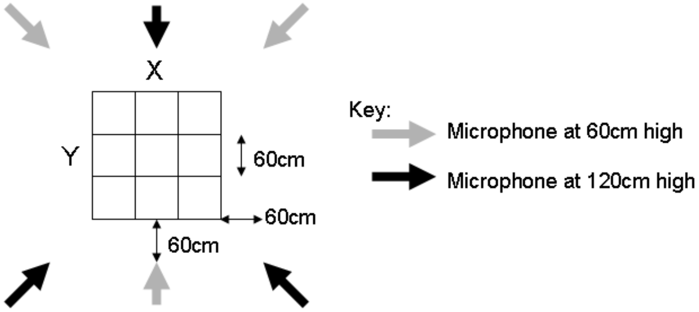


Fig. 6. Experiment setup for 3D experiments

Six microphones were used, three at 0.6 m high and three at 1.2 m high, facing towards this grid, as shown in Figure 6. This layout conforms to the range and angular sensitivity limits of the microphones determined in the previous experiment. At each location, twenty finger clicks were recorded; clapping was not performed.

Figure 7 shows the cumulative frequencies for various 3D distance errors. The rightmost line indicates that the prototype is capable of locating clicks with an absolute 3D accuracy of around 27 cm 90% of the time. In order to determine the various causes of this result, it is useful to examine the three other lines, which show the 3D distance error from the mean reported location, as well as the 2D (XY) errors (both absolute and relative to the mean reported location). Two observations can be made: there are sys-

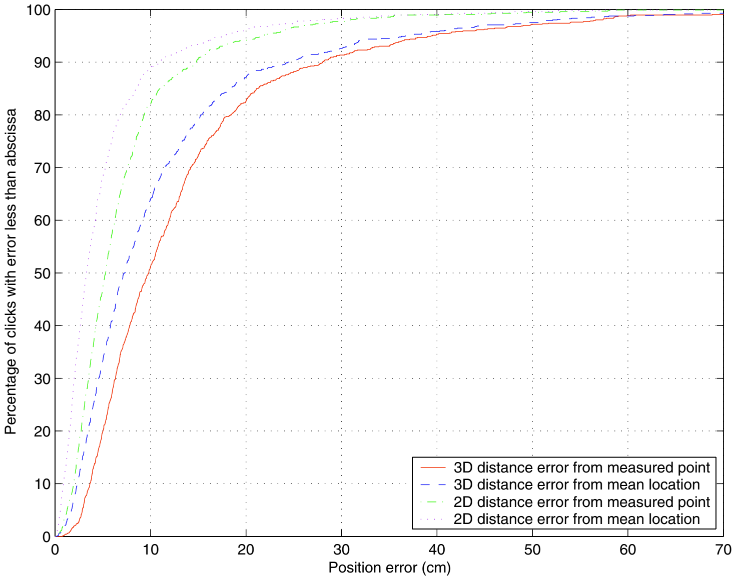


Fig. 7. Cumulative frequency plot of location error measurements

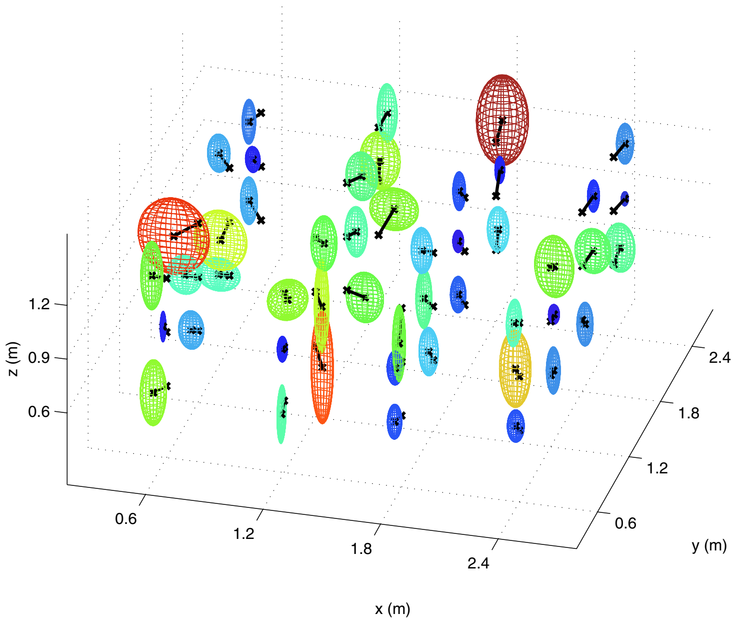


Fig. 8. "Lollipop" diagram showing mean and standard deviation of clicks at each test point

tematic errors of around 5 cm in the experiments, and the Z-dimension error accounts for around half the 3D position error. These are explained below.

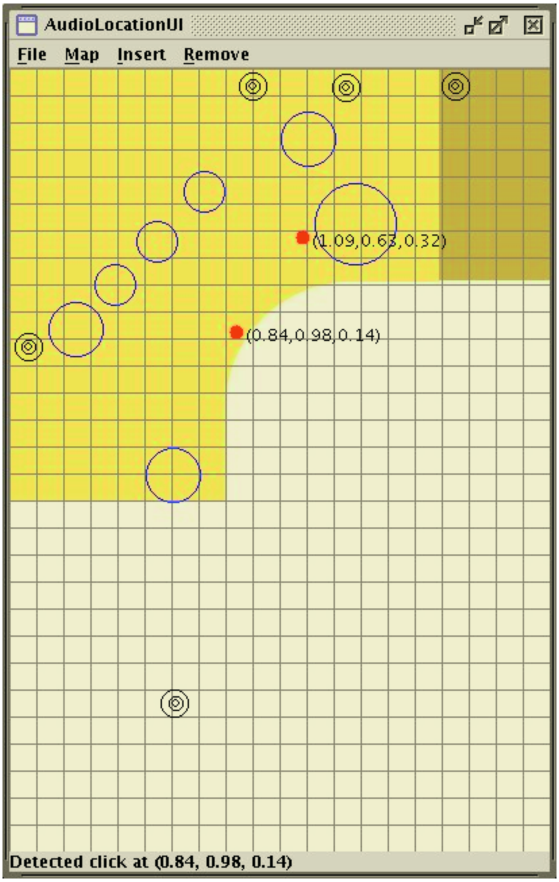
Figure 8 shows a perspective view of the test grid including an ellipsoid for each test point with axes equal to the standard deviations in each dimension. This figure illustrates the higher Z-dimension errors well. These errors are due to the higher Dilution Of Precision (DOP) in the Z dimension, since the microphones are quite closely spaced in this dimension (a spread of 0.6 m as opposed to the 3.0 m range in the X and Y dimensions), thus reducing accuracy. The low spread was used since this may be the situation faced by real-life deployments of audio location systems, where microphones will be placed on objects in the room, and are therefore not freely placed over a wide Z range.

The systematic errors, illustrated by the lines from the centre of the spheres to the test points in Figure 8, may be due to such causes as experimental setup error, lack of temperature compensation, and with the position of the clicking hand as a prime suspect. As discussed above, it is difficult to make a clicking noise with a location accuracy of more than, say, 5-10 cm, because of the size of the hand.

When considering the application area of 3D user interfaces, it is not the absolute position error that matters, but the relative error between the position where a 3D button is defined and the position of the user's click on the button. Some sources of error are therefore tolerated well by a 3D user interface, namely those such as surveying errors which cause the same erroneous offset to be incurred each time. In addition, it is likely that 3D interfaces would be implemented against a plane such as a wall or desk, allowing the physical tokens to be affixed beside the button's location to help users. By organising 3D interfaces appropriately (depending on the layout of the microphones), the effect of DOP can be mitigated, thus providing a location resolution of closer to the lower bound displayed in Figure 7, and indicating that buttons can be as narrow as 20 cm wide while remaining usable.

3.5 Deployments

In order to test 3D user interfaces, a GUI was implemented allowing the audio location system and applications to be quickly deployed. This GUI is illustrated in Figure 9. To create a new configuration, a jpeg file is provided for the background, along with the coordinates of the corners. Microphones are then placed using GUI dialogs, and buttons



Detected click at (0.84, 0.98, 0.14)

Fig. 9. GUI for configuring audio location and defining 3D user interfaces. A plan view of an office room with a corner desk is used for the background. The concentric circles represent microphones, the large unshaded circles represent buttons, and the small shaded circles represent finger clicks recently detected, along with their coordinates. Of the six microphones in the prototype, one is off the side of this plot and therefore not shown

can then be added either by clicking on the map, typing in coordinates, or by indicating a location by finger-clicking at the appropriate point 3 times. Buttons can be spherical, cylindrical, or cuboid, and an arbitrary shell command can be entered for execution when the button is triggered.

This interface has been used to implement a 3D finger-clicking interface controlling an mp3 music player (namely `xmms`) to demonstrate audio location. Over various occasions this has been set up at four different environments, with the setup time from a boxed to a working system being approximately 2 hours. The majority of this time is taken up in setting up the hardware and placing and manually surveying the microphones (automatic survey techniques described in Section 2 have not yet been integrated), with only a short time required to configure the buttons (play, stop, next track, choosing an album, etc) for the mp3 application.

While formal user studies have not yet been undertaken, it was observed that novice users are quick to understand the concept of "clicking" on a point in 3D space, which are marked with paper tags to indicate the virtual buttons. It was also discovered that few spurious location events are reported by the system even in very noisy environments. While such environments result in constant sound events across the various microphones, the location determination algorithm is able to discard the vast majority of these events as not representing the same sound, due to the high residual errors it finds when attempting to fit a 3D location to them. Furthermore, even when a rare false "click" is generated, it is unlikely to perform an incorrect action, since the majority of the 3D space is not marked as part of a button. These observations support the claim made earlier that the amplitude-threshold algorithm works well for 3D user interface applications.

Of the hundreds who have seen this system demonstrated, many (approximately half) were unable to click their fingers well enough to allow for location determination. This was overcome by providing cheap mechanical clickers, which have the disadvantage of introducing a hardware requirement for users. However, this clicker is easily shared, e.g. by leaving it next to the interface, and frequent users of the system could train themselves to click their fingers.

4 Conclusions and Future Work

This paper has described "audio location", a technique for low-cost centimeter-scale location sensing, which makes use of off-the-shelf audio hardware that is already deployed with many PCs, and cheap to add to PCs. The various design parameters for an audio location system were discussed, including the use of tagged (i.e. where the user carries some sort of locatable device) or untagged sensing, with untagged operation being identified as a key advantage over other types of location systems.

A prototype untagged audio location system was built, targetted at the novel application area of detecting human-generated sounds to enable 3D user interfaces. Experiments show that finger clicking can be detected with location errors of under 27 cm in 3D (14 cm in 2D) for 90% of the tests. This system does not suffer from the high cost and difficult installation of other centimeter-scale 3D location systems, with the audio hardware costing around one hundred British pounds, and installation being accomplished in a few man-hours. Audio location has therefore been shown as a viable technology to remove the barriers to entry for high accuracy location systems, opening the door to the wide deployment of location-aware applications relying on high accuracy location information.

There is much research left in the area of audio location. Topics for future work include the use of multiple PCs with fewer microphones per PC, a formal user study for the 3D user interfaces, and development of algorithms based on cross-correlating sound sources rather than amplitude-thresholding, which would for example allow the location of human voices.

Acknowledgements

The authors would like to thank Mark Williamson, Tim Sohn, John Krumm, Richard Sharp, Anil Madhavapeddy, and Mike Hazas for assistance during this research.

References

1. Hazas, M., Scott, J., Krumm, J.: Location-aware computing comes of age. IEEE Computer **37** (2004)
2. Krumm, J., Cermak, G., Horvitz, E.: RightSPOT: A novel sense of location for a smart personal object. In: Proceedings of UbiComp 2003. Volume 2864 of LNCS., Springer-Verlag (2003)
3. Ward, A., Jones, A., Hopper, A.: A new location technique for the Active Office. IEEE Personal Communications **4** (1997)
4. LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., Schilit, B.: Place Lab: Device positioning using radio becons in the wild. In: Proceedings of Pervasive 2005. LNCS, Springer-Verlag (2005)
5. Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based user location and tracking system. In: Proceedings of InfoCom 2000. Volume 2., IEEE (2000)
6. Brandstein, M.S., Adcock, J.E., Silverman, H.F.: A practical time-delay estimator for localizing speech sources with a microphone array. Computer Speech and Language **9** (1995)
7. Chen, J., Yao, K., Hudson, R.: Source localization and beamforming. IEEE Signal Processing Magazine **19** (2002)
8. Girod, L., Estrin, D.: Robust range estimation using acoustic and multimodal sensing. In: Proceedings of the IROS 01. (2001)
9. Girod, L., Bychkovskiy, V., Elson, J., Estrin, D.: Locating tiny sensors in time and space: A case study. In: Proceedings of ICCD 02. (2002)
10. Rosca, J., Sudarsky, S., Balan, R., Comanici, D.: Mobile interaction with remote worlds: The acoustic periscope. In: Proceedings of the AAAI 01. (2001)
11. Madhavapeddy, A., Scott, D., Sharp, R.: Context-aware computing with sound. In: Proceedings of UbiComp 2003. Volume 2864 of LNCS., Springer-Verlag (2003)
12. Addlesee, M.D., Jones, A., Livesey, F., Samaria, F.: The ORL Active Floor. IEEE Personal Communications **4** (1997)
13. López de Ipiña, D., Mendonça, P., Hopper, A.: TRIP: A low-cost vision-based location system for ubiquitous computing. Personal and Ubiquitous Computing **6** (2002)
14. Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., Shafer, S.: Multi-camera multi-person tracking for EasyLiving. In: Proceedings of the Third IEEE International Workshop on Visual Surveillance. (2000)

15. Hazas, M., Ward, A.: A high performance privacy-oriented location system. In: Proceedings of PerCom 2003. (2003)
16. Press, W., Teukolsky, S., Vetterling, W., Flannery, B.: Numerical Recipes in C. Cambridge University Press, Cambridge, UK (1993)
17. Cramer, O.: The variation of the specific heat ratio and the speed of sound in air with temperature, pressure, humidity, and co2 concentration. The Journal of the Acoustical Society of America **93** (1993)
18. Scott, J., Hazas, M.: User-friendly surveying techniques for location-aware systems. In: Proceedings of UbiComp 2003. Volume 2864 of LNCS., Springer-Verlag (2003)
19. Harle, R., Ward, A., Hopper, A.: Single Reflection Spatial Voting: A novel method for discovering reflective surfaces using indoor positioning systems. In: Proceedings of MobiSys 2003, USENIX (2003)
20. Hightower, J., Brumitt, B., Borriello, G.: The location stack: A layered model for location in ubiquitous computing. In: Proceedings of the Fourth IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 2002), IEEE Computer Society Press (2002)
21. Pingali, G., Pinhanez, C., Levas, A., Kjeldsen, R., Podlaseck, M., Chen, H., Sukaviriya, N.: Steerable interfaces for pervasive computing spaces. In: Proceedings of the PerCom 2003, IEEE (2003)
22. Ishii, H., Ullmer, B.: Tangible bits: Towards seamless interfaces between people, bits and atoms. In: Proceedings of CHI 97, ACM (1997)
23. Pavlovic, V., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997)
24. Addlesee, M., Curwen, R., Hodges, S., Newman, J., Steggles, P., Ward, A., Hopper, A.: Implementing a sentient computing system. IEEE Computer **34** (2001)