Mobile Interaction with Remote Worlds: The Acoustic Periscope

Justinian Rosca Sandra Sudarsky Radu Balan Dorin Comaniciu

Siemens Corporate Research, Inc.
755 College Road East
Princeton, NJ 08540 USA
+1 609 734 6500

{rosca, sudarsky, rvbalan,comanici}@scr.siemens.com

ABSTRACT

Strictly speaking, a periscope is an optical device that allows one to view and navigate the external environment. The *acoustic periscop*e is a metaphor for *mobile interaction* that transparently exploits audio/speech to navigate and provide an unobstructed scene in a real or virtual world. We aim at both true mobility – no strings or devices should be attached to human user to be able navigate – and at a smart multi modal. The implementation of our concept highlights un underestimated modality, the acoustic one, for making computers transparent to the actual interaction of the user with a remote world and advancing in the direction of ubiquitous computing. In this paper we describe the basic principles, architecture and implementation of a system for ubiquitous, multimodal and easy visual accesses to the remote world based on the acoustic periscope idea. In order to assemble the required functionality we resort to audio signal processing (in particular array signal processing) for location and orientation estimation, speech recognition and text-to-speech synthesis for natural language interaction, mobile computing, communication in a LAN/Bluetooth network, and streaming of data from or control of a remote telerobotic platform with vision capabilities.

Keywords

Virtual Reality, Multi-Modal Interaction, PDA, Ubiquitous Computing, Mobile Interaction, Smart User Interface.

1. INTRODUCTION

Imagine a mobile robot carrying a tilt-and-pan controllable camera. Our robot is actually a vehicle that would let us remotely explore, for instance, the Rodin Museum of Art in Philadelphia after hours. To appreciate and enjoy sculpture, one has to depart from an apparently rigid and columnar structure, and follow the dynamic flowing lines in Rodin's sculpture. For this, one has to be mobile around the sculpture. Can our robot do this and stream images to a remote display? How would one control it and its camera? Not remote mouse or key movements, please! The latter approach, although possible, is clearly awkward for this goal.

What we would really like is make the robot and its camera smoothly "fly" around the sculpture, and stream the corresponding images to the user display. Moreover, we want the user to actively use her body to search for knowledge in this process. Can the user just naturally and transparently move in her environment with a PDA in her hand, and have the robot follow a similar trajectory in its real (or virtual) environment and stream images onto the PDA? We impose one final constraint: the user should not be tethered in the environment, and the whole process should not necessitate expensive position and orientation sensors mounted on the user's head. Localization and orientation, if possible, should be naturally based on the user's voice and the synthesized speech answers generated by the PDA. The user communicates with her PDA mostly by speech.

Visiting the museum after hours by letting the user herself be a *virtual acoustic-based periscope* in the "other" world be interesting and intriguing. Perhaps more realistic is an industrial application, such as smoothly exploring in 3-D the hidden intricacies of a hardly accessible machinery for diagnosis or repair, or exploring a high-risk or industrial environment. Many other virtual reality or tele-robotics applications are possible by means of the acoustic periscope technique.

This paper describes the architecture and an implementation of our virtual periscope approach in a natural, unobtrusive and inexpensive way. The user only carries with her the PDA, which represents both the virtual window into the other world (museum, real or virtual environment to explore), and the mobile device for dictation and speech commands. The "Rodin museum" experiment is possible in a dedicated room where a *system of microphones* makes it possible to localize sources of sound (human user, PDA).

The structure of the paper is as follows. Section 2 defines more precisely some of the concepts used here and throughout the paper. Section 3 describes the architecture of the system for mobile interaction and telecontrol. Section 4 discusses various implementation issues. Section 5 presents a hardware realization of our system. Finally we summarize this effort and present some challenges for present and future work.

2. SCOPE AND RELATED WORK

The acoustic periscope metaphor and the applications reviewed could be viewed from perspectives that can be cast in several consecrated ways: virtual reality, artificial reality, augmented reality. Below we define these main terms and highlight the nuances exploited in our scenario.

Virtual reality is the process of actively stepping inside (to see, hear, act upon) a computer generated, virtual environment. It usually assumes the use of a head-mounted audio/video display, and position and orientation sensors [1],[2]. This is the general scenario we use, although the applications we mentioned here do not exploit a virtual world – they could as well do that. As a virtual reality does, we also simulate another place to the user by present-

ing a transported visual sensation with *none* of the normally used VR I/O devices (data gloves, head-mounted displays, position and orientation sensors mounted on the user, etc.) In order to simplify requirements on the user, we use a physical space equipped with audio sensors (microphones). The user only carries a standard PDA or tablet computer.

Artificial reality is the process of describing virtual environments such that the user's body and actions combine with the computer generated sensory information to forge a single presence. The human perceives his actions in terms of the body's relationship to the simulated world [3], [4]. This is exactly how we propose to drive the camera in the remote world.

Augmented reality is a technology where the user's display shows a superposition of the real world and computer generated graphics (to augment the presentation of the real world objects) by means of a see-through display [5]. Although this is possible in our scenario, we have not emphasized at all the augmented reality aspects. We replaced the see-through display with a pocket or tablet computer. It is possible to bring a whole new dimension to the problem by incorporating a small camera into the mobile device.

Essential in our endeavor is the accurate tracking of sound sources based on audio signals. Let us review alternatives for spatial tracking solutions presently used in virtual reality systems. A summary of alternatives is presented in [2,6]. We briefly review mechanical, electromagnetic, ultrasonic, acoustic, and optic (vision-based) systems to date. Applications may also exploit non-visual cues of motion from devices that can be physically moved to generate such cues (see the Trike [5] system where self-motion induces in the user a variety of sensory queues: visual, auditory, vestibulary, somatosensory information about limbs, etc.) The main capabilities we are interested in here are location and orientation of the user. Six-degree-of-freedom sensors can provide both position and orientation information in 3-D. Our criteria for comparing the solutions are the accuracy of position and orientation, intrusiveness to user, tethering of user to a physical location, ease/transparency of use, range of use, and cost of deployment.

Mechanical tracking systems rely on a motion-tracking support structure of high precision, e.g. using opto-mechanical shaft encoders (BOOM 3C from Fakespace Labs). The user is generally anchored to the mechanical device. Electromagnetic systems (e.g. Flock products from Ascension Technology) use DC magnetic fields generated by three mutually orthogonal coils from a stationary transmitter that are detected by a similar three-coils receiver. The audio tracking system produced by Logitech uses a three fixedly mounted ultrasonic speakers and three mobile microphones thus detecting all possible 9 distances. Computer vision-based systems use either fixed cameras that track objects with markings (e.g. Northern Digital's Polaris product), or mobile cameras attached to objects that watch how the world moves around (see [2]). Global Positioning System (GPS) based systems receive signals from positioning satellites either directly, or in conjunction with a ground-located additional receiver and transmitter in a precisely known position. Small sized receivers with a small price too make their way into mobile devices (e.g. The Pocket CoPilot from TravRoute). Table 1 summarizes these tracking solutions.

Localization by means of audio signal processing, if possible, would present several advantages. First it would come naturally. Assuming that the user interacts with the PDA by voice, then users's sounds could be used to locate her as well. As mobile phones, PDAs, pocket PCs and the like advance towards the use of speech commands, data needed for localization comes for free. Secondly, the user would not have to wear special expensive helmets or sensors. The approach would be considerably less "intrusive" than others, and also easier to setup and use. The user would not be tethered to some physical location. However audio localization requires an array of microphones and therefore its use is limited to a room/space where the sensor array and a data acquisition system are installed.

It sounds ok so far, but two big questions are not addressed at this point. First, how accurately can we localize sound sources in a room? Second, how can we get orientation information that otherwise would be obtained with sophisticated magnetic or ultrasonic sensors? Next section offers encouraging answers to these questions. Based on that, we advance our architecture for applying the acoustic periscope metaphor.

3. AUDIO BASED LOCALIZATION AND ORIENTATION

Our project aims at location and orientation estimation based entirely on acoustics, more precisely on speech signals assumed to be already used for natural language interaction user-system. This way, location and orientation would come *for free*, being entirely transparent to the other functions of the system. This practically means doing without the additional hardware VR usually necessitates. Below we describe our localization approach, analyze the accuracy of 3-D position estimation, and describe how we can also determine orientation.

3.1 Localization approach

Two microphones would be sufficient to estimate the direction of arrival of a signal in one plane. Assume the following signal model in an anechoic environment:

$$x_1(t) = a_1 s(t-\tau_1) + \upsilon_1(t)$$
$$x_2(t) = a_2 s(t-\tau_2) + \upsilon_2(t)$$

where $s(t)$ is the source signal, $x_1(t)$ and $x_2(t)$ are the two microphone signals recording the attenuated source (by amplitude factors a_1 and a_2) and υ_1, υ_2 are mutually independent noises, and independent with the source signal. Let $\tau = \tau_1 - \tau_2$, and assume it is a multiple of the sampling period $T_s = \frac{1}{f_s}$, where f_s is the sampling frequency. Note that the crosscovariance between $x_1(\cdot)$ and $x_2(\cdot - \delta)$ for some delay δ is:

$$R(\delta) = E[x_1(\cdot)x_2(\cdot - \delta)] = E[s(\cdot)s(\cdot - (\delta - \tau))] \le R(\tau)$$

where $E[]$ denotes the expected value. Therefore, one simple method to estimate direction of arrival is based on the computation of the crosscovariance between the two microphone signals:

$$\hat{\tau} = \arg\max_{\delta}\{E[x_1(\cdot)x_2(\cdot - \delta)]\}$$

In an implementation, expected value would be given by time averaging over a batch of samples, and would be smoothed.

In 3-D, the geometric locus of points that induce a constant delay difference to two microphones (i.e. have constant difference in distances to two microphones) is a hyperbolic surface. To reduce non-determination to a point (a small physical volume around that point if estimation tolerance is introduced) we need to intersect three such surfaces obtained from three pairs of two microphones each. Therefore one has to use four microphones in order to unambiguously estimate the source location in 3-D

3.2 Accuracy of audio location estimation

Let's first discuss the accuracy of audio localization in a plane and then return to the 3-D case.

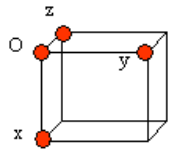
Given the speed of sound propagation *c* and the distance between two microphones *d* the maximum delay inducible in the microphone signals, in samples, is:

$$\tau_{\max} = \frac{df_s}{c}$$

The crosscovariance solution above only deals with integer delays, so that the best angular resolution of the method is:

$$\Delta\alpha = \frac{180}{2\tau_{\max} + 1}$$

For a distance between microphones $d = 3m$ and a sampling frequency $f_s = 16kHz$ we obtain $\Delta\alpha = 0.6\deg$. This corresponds to an error in estimating the source position (in plane) of about *0.7cm*. This implicitly considers that the source moves on a circle centered at the midpoint between microphones. Unfortunately, resolution is nonlinear around the microphones. It is worst if the source moved away from the two microphones, for instance, by sliding away on the median of the two microphones. Nonetheless, more microphone pairs are there to help, and the precision estimation analysis tells us how to place microphones in the environment. In the 3-D case microphones should be placed such that the three pairs to be considered span the three coordinate axis (Ox,Oy,Oz) (see Figure 1).



[3]

Figure 1. Placement of four microphones for acoustic source localization.

A refined computation of resolution in the 3-D case could be estimated as follows. Assume that the audio source to be localized in 3-D is estimated to be placed at $P(x, y, z)$, whose distances to microphones are d_k, $k = 1,...,4$. Also assume that the true source position is $P_0(x_0, y_0, z_0)$, with distances d_k^0, $k = 1,...,4$ to microphones. To estimate the accuracy of localization, we are interested in the size of the geometric locus of points $P(x, y, z)$ where estimated source could be placed. The locus is defined as follows:

$$\left|(d_k - d_j) - (d_k^0 - d_j^0)\right| < c\tau, \forall k \neq j; k, j = 1,...,4$$

We assessed the extent of the geometric volume described by the equation above. We derived the accuracy in position for a room of dimensions 5x4x3m, and microphones placed in three corners of the rooms forming a tetrahedron as in Figure 1.

The above analysis results in a worst-case error in one direction given by the largest distance D to the closest (distance *d*) microphone pair $\arg\min_{i,j}\{d_{ij}\}$. For instance, the largest error along the x-axis corresponding to an error of one sample in delay estimation is given by:

$$\Delta x = 2\sqrt{\frac{2}{\alpha - 1} \cdot D^2 - \frac{2\beta}{\alpha + 1}}, \alpha = \frac{8d^2 f_s^2}{c^2} - 1, \beta = -\frac{d^2}{4}$$

For *c=320 m/s, d = 3m, D=5m, f_s=16 kHz* the above formulae give $\Delta x \approx 0.035m$.

In the worst case the localization error was approximately several centimeters, which implies that the acoustic localization method can be used for our purpose. More complex algorithms based on fractional delays are also possible, but we will not discuss those here.

If the original signal to be "spoken" in the environment is known (e.g. this is the case for the PDA), then the induced delays can be calculated much more precisely by reference to the original signal. This means that localization accuracy is equally increased.

3.3 Determining Orientation

Orientation estimation relies on the estimation in position of both the user's head and the PDA. We assume that the user would talk after each move in her physical space, and that the PDA would respond by emitting a frequency rich signal, (e.g. a speech reply).

The user would normally help the PDA in front of herself, at a distance of about half meter. Assuming that errors made by the localization system are consistent for neighbouring sources, this implies that the two source positions give a reasonable estimate (for our purpose) of the orientation of the user (see Figure 2).

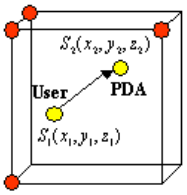


Figure 2. Orientation is obtained from the locations of the two audio sources

4. SYSTEM DESIGN

4.1 System components

The "audio periscope" scenario for the mobile interaction described in the introduction is present in schematic form in Figure 3. The system components are marked (a), (b) and (c) in the Fig-

ure. Data and commands are carried from user's world (a) to the server (b) and further on to the mobile camera in the remote world (c). Data from (c) is routed through (b) to (a).

On the user side (a) (i.e. the user's environment) the system consists of:

- System of microphones connected to local server (b), which contains a real-time data acquisition board. The sensors receive distinct audio signals from both the user and the PDA.
- User's PDA, which can communicate wirelessly with server (b) both to receive streamed images and to send speech and touch commands. It also emits sounds used for its localization on the server side.

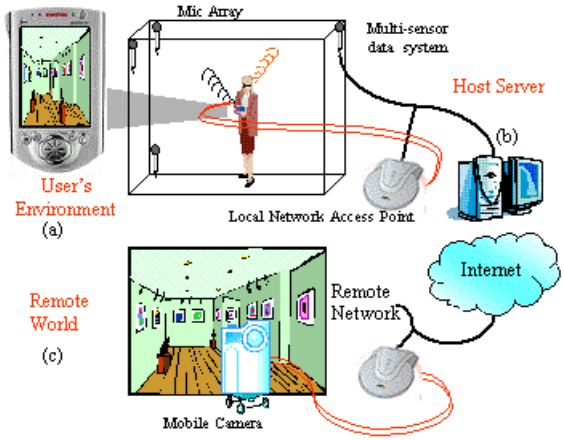


Figure 3. Scenario of operation "audio periscope"

The mobile camera system (c) ensures the desired exploration capability in the remote world.

Software running on the server (b) is responsible for implementing the system functionality and is described in detail next.

4.2 Software architecture

The main system components on the server side are assembled in a multithreaded real-time application controlling the audio acquisition system, the remote video system (or the "camera" in a virtual world) and the video-streaming component (see Figure 2):

- Audio signal processing module is itself multithreaded and is responsible for controlling in real-time the data acquisition board, for processing wav audio data in order to localize sources, perform noise reduction and blind source separation in order to pass clean audio signals to the signal matching and speech recognition components.
- Speech recognition module is responsible for understanding human free speech according to an application dependent command land interaction language. It passes commands further to the camera control system
- Camera control module is responsible for the pan and tilt of the camera and eventually the robot control. To insure a smooth visualization, the camera should execute fast saccades in response to sudden and large movements of the user while providing a smooth pursuit when the user is quasi-stationary [11,12]. An arbiter additionally takes into account commands extracted by speech recognition and implements the overall control that resembles the human system. The fovea subimage occupies laterally about *6 deg* of the camera's *50 deg* field of view, at zero zoom.
- PDA socket server module is responsible for passing commands and voice data from the PDA to the other system components. In noisy conditions it makes sense to interpret the PDA recording (audio) signal for subsequent speech recognition, rather than the signal obtained after processing microphone sensor data.
- Media services control server manages the media encoder and server, for opening session with remote server and streaming data to the PDA. It also arbitrates the various commands extracted from speech or from the PDA.

4.3 Hardware implementation

The system consists of the following components:

- Data acquisition platform and microphones. We are using the M44 Flexible DSP/Data Acquisition board from Signalogic, equipped with a four-channel 96kHz 24-bit Sigma-Delta Analog input output, and four condenser phantom-powered microphones mounted in the upper corners of an office room.
- Host system. This is a Windows 2000 Pentium-based personal computer server configured as a media server as well.
- Wireless local area network. Its hub is 802.11b compliant and insures a 10Mb/s throughput.
- Video system. Interfaces the SONY EVI-D30 camera through a standard RS-232C interface to its host system.
- Mobile computer. An iPAQ 3600 PDA from Compaq, equipped with a WLAN card ensures all the control and communication services required by present scenario.

The data acquisition board has a PCI bus card controller for streaming audio data to the host system.

The video system uses the serial interface to communicate and control the camera's pan, tilt, and zoom.

The software we use to stream live images from camera to the PDA is Windows streaming media player on iPAQ client and Windows Streaming media, Media Services on Win2K server.

5. APPLICATIONS: ROLE OF A NATURAL AND EFFORTLESS USER INTERACTION

Many virtual environment applications try to mimic the real world. Thus it would be ideal if user interaction replicated the

user's natural way of interacting with the real objects. Almost all VR applications involve some kind of navigation through a virtual 3D environment. Navigation in such environments is a difficult problem: users often get disoriented or lost. A number of three degrees of freedom input devices, including 3D mouses [11], spaceballs [12] and joysticks have been designed to facilitate user interaction. However, three degrees of freedom are often not sufficient to define user position and orientation in a 3D scene.

Certainly amongst the most natural ways of navigation is navigation by moving in the physical world without carrying any cumbersome tracking devices. One of the main goals of our metaphor was to create a natural (intuitive), and transparent (effortless) interaction of this type with the remote, virtual world. This is not easy to do with little additional hardware.

Interactive walkthroughs applications [10] are perfect candidates for VR environments. Such applications let the user experience a virtual world by moving through and around virtual objects. In our system, the user location and orientation can be tracked by means of a set of microphones and this information is then used to update the position of the virtual camera. With this type of interaction, the user could walk through the interior of a virtual building to evaluate the architectural design in a natural way, just by walking around a room with only its PDA on his/her hand. Since the user can usually move only on the floor, the orientation information is used to provide the user more degrees of freedom, for example to move up and down staircases. In addition, with a simple speech command, the user can make the walls transparent to further evaluate, for example, the location of pipes and the electrical settings.

Another interesting application where natural user interaction is essential is the use of large wall display systems for business presentations, and immersive, collaborative work. For example Kai Lin et. at.[7] presented the construction of a scalable display where multiple cameras were used to track the user, recognize her gestures and detect the location of some novel input devices. In contrast, our prototype uses audio to track the user position and orientation and also recognize spoken commands. The user can zoom in and out by moving closer and further away from the display, several users can have control over the display without sharing any input devices, and speech recognition can be used to control the speed and other aspects of the presentation.

6. CONCLUSION

.

We exploit an often neglected but very rich modality of our environment: audio signals. This paper proposes the "acoustic periscope" metaphor and our implementation approach from a unique perspective: we work to use presently available hardware and not incur amazingly high costs for making it happen. A low cost increases potential in number of application scenarios and users. This is quite a challenge.

We have tested quite a few components of our system and are working on a system prototype. Below we review some of its highlights:

- Our system presents virtual/remote sensations to the user by means of *none* of the normally used Virtual Reality I/O devices, but a rather much simpler to install and use system of microphones.
- Audio source location estimation, localization and orientation come *for free*, being entirely transparent to the other functions of the system, assuming that user-system speech interaction is a must.
- The acoustic periscope paradigm is aimed for a natural, intuitive, and transparent interaction with the remote, virtual world. Moving into the physical world achieves navigation as in other VR systems but without carrying any cumbersome tracking devices.
- Audio signals from the human user (speech) and PDA (speech generated replies or special signals) are sufficient for determining source location and orientation of the user with sufficient precision (several centimeters for localization) at least for some applications. The acoustic model used in our formal derivations here is anechoic.
- The overall system philosophy and architecture allows a natural integration of virtual reality interaction and speech processing for transcending computers to the ubiquitous stage [10] where the focus is on one's actions and activities rather than the actual mode of interaction.

Present and future work includes assessment of system issues resulting from the integration of the various components, improvement of localization in echoic environments and application of the acoustic periscope scenario to various problems of interest. Last but not least, we hope that our efforts will be one useful step towards integrating present advanced technologies into real problems in an attractive, non-cumbersome way.

Acknowledgements. We would like to thank Heckrodt Kilian, Stuart Goose, Subramanyan Vdaygiri and Arturo Pizano for their good advice regarding various infrastructure and technical issues in this project.

7. REFERENCES

[1] Alen Wexelblat (editor). Virtual reality applications and explorations. Academic Press, 1993.

[2] Blair MacIntyre and Steven Feiner. Future of multimedia user interfaces. In Multimedia Systems, (4): 250-268, 1996.

[3] Micheal Hein. The metaphysics of virtual reality. Oxford University Press, 1993.

[4] M.W. Krueger. Artificial Reality II. Addison-Wesly Publishing Co., Reading, MA, 1991.

[5] T.P. Caudell. Introduction to Augmented Reality. SPIE Proceedings, vol. 2351 : Telemanipulator and Telepresence Technologies, pp.271-281, Boston, MA, 1994.

[6] Robert Allison, et al. First steps with a ridable computer., in the Proceedings of the Virtual Reality 2000 conference, IEEE Computer Society, 18-22 May 2000, pp. 169-175.

[7] Kai Li, Han Chen, et. al., Early Experiences and challenges in Building and using a scalable display wall system, IEEE Computer graphics and applications, vol. 20(4), pp. 671-680.

[8] Rick Lewis and Carlo Séquin. Generation of 3D building models from 2D architectural plans, Computer-aided Design, 30(10), pp. 765-779 (1998). Elsevier Science.

[9] L. Darsa and B. Costa and Amitabh Varshney. Walkthroughs of complex environments using image-based simplification, Computers & Graphics, 22(1), pp. 55-69 (February 1998). Pergamon Press / Elsevier Science.

[10] Mark Weiser. The Computer for the 21st Century. Scientific American, September 1991.

[11] D.W. Murray & all. Driving Saccase to Pursuit using Image Motion. Int.J.Comp.Vis., 16(3) , pp. 204-228, 1995

[12] H.P. Rotstein and E. Rivlin. Optimal Servoing for Active Foveated Vision. IEEE Conf. Comp. Vis. Pat. Rec., San Francisco, pp. 177-182, 1996

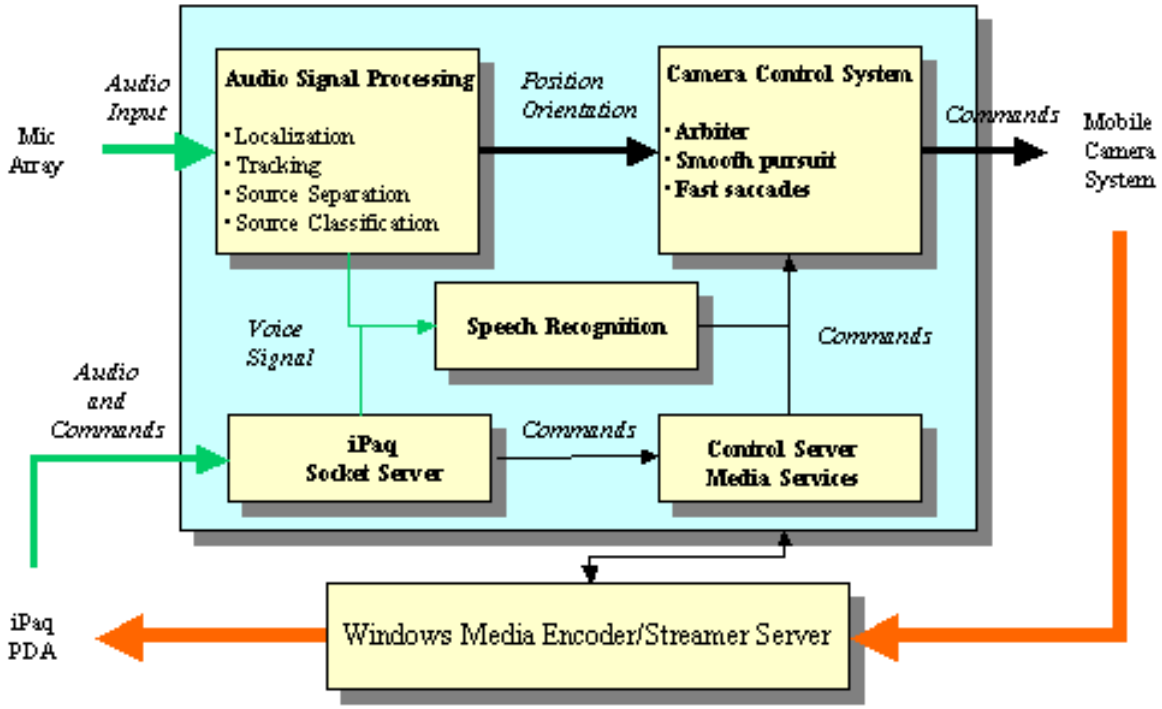


Figure 4. System Architecture

System	Accuracy	Intrusiveness	Range of use	Direct line of sight	Approx. cost	Notes / Environment
Mechanical	high	high	1-10 m	no	high	tethered to a fixed point
Electromagnetic	high	medium	1-6 m	no	high	EM field sensitive
Ultrasonic	high	medium	1-10 m	yes	high	acoustic noise sensitive
Optical	high	medium/high	1-10 m	yes	high	requires special markers
GPS	low	low	world wide	no	low	very general

Table 1. Source localization approaches used in the Virtual Reality literature