

10

# AUDIO-BASED LOCALISATION FOR UBIQUITOUS SENSOR NETWORKS

Benjamin Christopher Dalton

Honours Class 1 MPhys. University of Leeds (2002)

Dip. Art., Middlesex University (1998)

Submitted to the Program in Media Arts and Sciences,  
School of Architecture and Planning,  
in partial fulfillment of the requirements for the degree of  
Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2005


© Massachusetts Institute of Technology 2005. All rights reserved.

Author \_\_\_\_\_

Program in Media Arts and Sciences

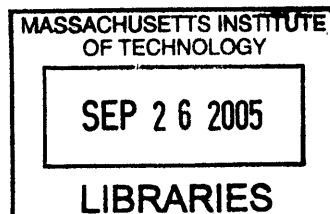
August 9, 2005

Certified by \_\_\_\_\_

  
V. Michael Bove Jr.  
Principal Research Scientist  
MIT Media Laboratory  
Thesis Supervisor

Accepted by \_\_\_\_\_

Andrew Lippman  
Chair, Department Committee on Graduate Students  
Program in Media Arts and Sciences



ROTCH



# AUDIO-BASED LOCALISATION FOR UBIQUITOUS SENSOR NETWORKS

Benjamin Christopher Dalton

Submitted to the Program in Media Arts and Sciences,  
School of Architecture and Planning,  
on August 9, 2005, in partial fulfillment of the  
requirements for the degree of  
Master of Science in Media Arts and Sciences

## *Abstract*

This research presents novel techniques for acoustic-source location for both actively triggered, and passively detected signals using pervasive, distributed networks of devices, and investigates the combination of existing resources available in personal electronics to build a digital sensing 'commons'. By connecting personal resources with those of the people nearby, tasks can be achieved, through distributed placement and statistical improvement, that a single device could not do alone. The utility and benefits of spatio-temporal acoustic sensing are presented, in the context of ubiquitous computing and machine listening history. An active audio self-localisation algorithm is described which is effective in distributed sensor networks even if only coarse temporal synchronisation can be established. Pseudo-noise 'chirps' are emitted and recorded at each of the nodes. Pair-wise distances are calculated by comparing the difference in the audio delays between the peaks measured in each recording. By removing dependence on fine grained temporal synchronisation it is hoped that this technique can be used concurrently across a wide range of devices to better leverage the existing audio sensing resources that surround us.

A passive acoustic source location estimation method is then derived which is suited to the microphone resources of network-connected heterogeneous devices containing asynchronous processors and uncalibrated sensors. Under these constraints position coordinates must be simultaneously determined for pairs of sounds and recorded at each microphone to form a chain of acoustic events. It is shown that an iterative, numerical least-squares estimator can be used. Initial position estimates of the source pair can be first found from the previous estimate in the chain and a closed-form least squares approach, improving the convergence rate of the second step.

Implementations of these methods using the Smart Architectural Surfaces development platform are described and assessed. The viability of the active ranging technique is further demonstrated in a mixed-device *ad-hoc* sensor network case using existing off-the-shelf technology. Finally, drawing on human-centric onset detection as a means of discovering suitable sound features, to be passed between

nodes for comparison, the extension of the source location algorithm beyond the use of pseudo-noise test sounds to enable the location of extraneous noises and acoustic streams is discussed for further study.

Thesis Supervisor: V. Michael Bove Jr.

Title: Principal Research Scientist, MIT Media Laboratory





AUDIO-BASED LOCALISATION  
FOR UBIQUITOUS SENSOR NETWORKS

Benjamin Christopher Dalton

The following people served as readers for this thesis:

Thesis Reader \_\_\_\_\_  
Barry Vercoe  
Professor of Media Arts and Sciences  
MIT Media Laboratory

Thesis Reader \_\_\_\_\_  
Seth Teller  
Associate Professor of Computer Science and Engineering  
Computer Science and Artificial Intelligence Laboratory

# ACKNOWLEDGMENTS

I would like to thank Mike Bove for the research environment he has cultivated, and for all of his support and input. My thesis readers, Barry Vercoe and Seth Teller for their inspiring research and for their comments and conversations through this project.

I enjoy learning most through dialogue and discussion, and I am grateful to those people who have been happy to answer questions and explore ideas. Particularly to Carlos Rocha, James McBride, Jim Barabas and Max VanKleek for their constant lessons and inspiration. To Noah Fields, Amber Frid-Jimenez and Joe Dahmen for creative collaboration. To my research group Arnaud Pilpré, Diane Hirsh, Gauri Nanda and Jacky Mallet. To Amanda Parkes, Burak Arikan and Saoirse Higgins for late night wine, ideas and theories at our house. To everyone else at the lab and MIT who have made it such an interesting and fun place to be, I will thank you in person, rather than trying to list you all here. Thanks to Cameron Marlow for the last minute thesis stylings.

To Chris Csikszentmihályi for his grounded, insightful perspective of technology and technologists, and John Maeda for his constant energy and creativity. And to the other professors who have taught me while I have been here, including Hiroshi Ishii, Glorianna Davenport, Neil Gershenfeld and Judith Donath for their direction and support. To Michèle Oshima and Dora Kelle, and WMBR, and the Collision Collective for keeping MIT full of art.

To my far away friends and family, many thanks for making me who I am. Liz Oliver for everything. Charlotte Burck and Adrian Dalton in their roles as parents and for a life time of inspiration. And to my grandmothers, Jeanne Burck-Rischen and the late Vi Dalton.

# CONTENTS

Contents 8

List of Figures 10

1 Introduction 13

- 1.1 Algorithmic Contributions . . . . . 14
- 1.2 Outline of Work . . . . . 16

2 Spatial Listening 19

- 2.1 Sound . . . . . 19
- 2.2 Sound Sensing . . . . . 26

3 Audio Sensing Everywhere 33

- 3.1 Ubiquitous Computing . . . . . 33

4 Sensor Networks 37

- 4.1 Organising Distributed Sensing . . . . . 37
- 4.2 Smart Architectural Surface . . . . . 42
- 4.3 Heterogeneous Device Network . . . . . 46

5 Collaborative Colocalisation 49

- 5.1 Assumptions and Approaches to Localisation . . . . . 49
- 5.2 Pair-wise Ranging and Relative Localisation . . . . . 52
- 5.3 Colocation Method . . . . . 60

6 Source Localisation 69

- 6.1 Inferring Position from Multiple Recordings . . . . . 70

6.2	Evaluation of Method . . . . .	77
7	Discussion	83
7.1	Colocalisation . . . . .	83
7.2	Source Location . . . . .	85
7.3	Sound Event Selection . . . . .	87
7.4	Enabling Impromptu Sensing . . . . .	89
	Appendices	93
A	Technical Specifications	95
	Bibliography	97

# LIST OF FIGURES

2.1	Examples of spatial mapping in conversation analysis. . . . .	30
2.2	Dialogue occurs in social setting. How can machine listening complement this? . . . . .	31
3.1	Examples of personal and public mapping and memory. . . . .	35
4.1	Desired components of a node sensor. . . . .	42
4.2	Consumer electronics devices can act as sensors. . . . .	43
4.3	A Smart Architectural Surface tile . . . . .	44
4.4	Picture of typical Smart Architectural Tile array . . . . .	45
5.1	Diagram of timing measurements that yeild time of flight on asynchronous networks. . . . .	57
5.2	Raykar <i>et al.</i> , ranging and localisation in general purpose computers. . .	59
5.3	Ranging uncertainty in measurements at a $7.1m$ device pair separation. . . . .	63
5.4	Ranging accuracy over the length of a room . . . . .	64
5.5	Variance in speed of sound with temperature, pressure and humidity. . . . .	64
5.6	The node arrangement derived experimentally using collaborative localisation. . . . .	65
5.7	Peak finding performance despite $8kHz$ sampling. . . . .	66
5.8	Performance tests of a heterogeneous device system. . . . .	67
6.1	Passive and active source localisation training manuals from submarine. . . . .	69
6.2	Plot of hyperbolic function. . . . .	71
6.3	Sound pair source location simulation results. Two, four and six samples of ranging error. . . . .	78

6.4	Sound pair source location simulation results for two step estimation. . .	79
6.5	Sound pair source location results for test using SAS tiles. . . . .	80
7.1	Onset detection for a room recording. . . . .	89
7.2	Correlation between an onset sample and two recordings. . . . .	90





# CHAPTER 1

## INTRODUCTION

Imagine we meet in a cafe and I put my phone on the table along with yours and the phones of everyone in else our group. We now have enough microphones to map the temporal and spatial history of our conversation and the sounds around us.

Far more pervasive than desktop computers with broadband internet connection are the personal electronic devices many of us carry, such as mobile phones, handheld game consoles and music players. Network-connected, digital technologies have become integral parts of our everyday lives for communication, entertainment and organisation of our information, and as essential work-platforms and fashion items. How can we use these collections of devices for public good and personal reward? If individuals choose to allow shared use of whatever embedded resources the devices they carry in their pockets can share, such as their microphones, cameras, speakers, screens, processing power, storage, GPS<sup>†</sup> and communication, then a new family of applications can be built that draw these information-collection and output mechanisms together to form a collaborative whole. The network must have the facilities to adapt to its continuously changing state. This category of distributed systems, although still challenging to implement, carries benefits of flexibility and potential that grows with the number of participants.

We are surrounded by digital devices that can communicate, listen and learn. Although they share many features, we carry multiple processors each with a specific primary use. These devices are often equipped with a microphone and speaker, and many can establish wireless connections to communicate with other devices near by. They need computational power for their primary tasks, but spend most of their time underused. When we meet with someone, the total count of locally available sensors doubles, when we gather in a group the number goes up further. These clusters of devices offer the resources of a distributed audio sensor network. If their owners desired, the impromptu collection of sound sensing nodes could share information and leverage their spatial distribution in order to locate themselves

Global Positioning System - now beginning to be added to some mobile phones.

and the audio sources around them and begin to identify patterns, build models and respond to this information effectively.

The aim of this research is to establish a basis for audio-based spatial localisation as an easily accessible tool in distributed processing sensor networks and to investigate some of the uses of the information such a tool provides. What is desired is a system that works across platforms and processors, functioning equally on our personal devices, on the digital objects in our daily habitats and on distributed *ad-hoc* networks such as the Smart Architectural Surfaces project[5] developed at the Media Laboratory.

A practical approach to self-localisation, that is not reliant on fine grain timing synchronisation, which has been demonstrated in both acoustic[46] and electromagnetic[54] ranging systems, is presented here. In requiring only a microphone, a speaker, a timer and some form of inter-node communication, but no access to low level hardware, it becomes possible to utilise any device with these basic resources as part of an auditory scene analysis sensor network. A microphone array can be built on-the-fly from these pervasive components.

The intention of the current research has been to implement a robust version of this auto-localisation algorithm, demonstrating its viability as a cross-platform ranging mechanism, and investigate extending these techniques to the estimation of external sound event positions. Performance and fault tolerance are discussed. An implementation across nodes of the Smart Architectural Surfaces is presented that establishes the framework for collecting and interpreting long-term spatial sound event patterns from the busy environment in which they are installed.

The thesis of this work is that sound can be used to establish a practical common coordinate system between sensor equipped, but heterogeneous, everyday devices. This information can then be used to enable the impromptu collection to act as a distributed microphone array, spatially mapping unknown acoustic events, which are naturally linked to human activity and scale. This research is discussed in the context of existing audio sensing techniques and the many uses location through sound can play in the context of ubiquitous computing and in response to everyday life.

## 1.1 ALGORITHMIC CONTRIBUTIONS

This work describes how spatial information can be derived from sound measurements taken at dispersed sensors and the techniques to collect and use this information from the microphones of devices. Methods for estimating sound source

locations are developed, motivated by the central role acoustic positioning can play in inferring human activity and context, and by the benefits that pervasive impromptu sensing networks could provide. Two algorithms are presented, one to use sounds intentionally emitted from sensor locations to range between them, and a second to use an array of sensors at known locations to estimate source positions for any sounds occurring around them.

### *Collaborative Colocalisation*

Actively emitted reference sounds from each device in a space are used to measure the ranges between them. It is shown that comparison of the length of the delays between detected sounds at two devices gives the flight time of the sounds across the distance separating the sensors. The delays can be calculated without synchronisation between devices or from recording taken at unknown times. From pair-wise flight times and an assumption of the speed of sound in air, relative device positions can be estimated, using multidimensional scaling to establish a common coordinate system. In tests the ranging is shown to have an uncertainty of a few centimetres dependent on the accuracy with which sounds can be found in a recording at the sampling rate used.

### *Passive Source Location*

A novel approach is presented for extracting spatial information from recordings of sounds arriving at a number of sensor devices. A method is developed for simultaneously estimating the positions of two acoustic sources. This draws on the relative delay-timing approach of the active ranging technique for asynchronous devices described above and the field of research on source location in synchronised microphone arrays. Parametric estimation of a pair of sounds uses a two step process. A closed-form least squares estimate of the newest sound is found based on the position of the previous sound, providing an initial guess that aids iterative numerical least squares estimation for the nonlinear signal model for the two sources. Simulation of this approach shows the estimates to be clustered with a standard deviation of  $\sim 20$  cm for the expected uncertainty in timing measurements. This method is then tested using a series of reference chirps emitted at two locations. The positions of the two sound sources are determined to within half a meter, with the separation of the pair estimated to within a few centimetres.

## 2.2 OUTLINE OF WORK

In Chapter 3, motivation for acoustic based location is grounded in the roles of human vision and audition. Sound and light, because of their linear propagation of energy, are important in sensing the world around us. Further, audio is unique in its strong link to activity, and although it provides sparser detail than vision, with most static scene information lost — merely implied by reverberation and filtering — movement and events are nearly always tied to associated sounds. Sound is also marked by the ease with which it can be generated, in our actions, our communication and entertainment, tying a specific location to each one of these acoustic events. Coupled with the simplicity with which it can be measured, this places sound in a unique position for collecting spatial estimations of audio signatures associated with exact moments of human centric activity or meaning.

In Section 3.2 an overview is given of how machine listening, focused on detecting sounds linked with people, has developed from early voice recognition to music processing and audio scene analysis. These fields can all use expectation, based on assumptions from prior listening or the properties of sound, to improve performance in recognising and extracting meaning from waveforms. The expectation models can be augmented with additional information in the form of acoustic position estimation. The part that spatial estimates can play in aiding segmentation and categorisation of sounds is highlighted. Other uses of spatial acoustic features are discussed including as metadata to mark up long audio recordings, in aiding memory and in understanding otherwise ephemeral sound information. For example human social dynamics and patterns of dialogue can be evoked, and better understood, by examining the traces of spatial history collected from audio.

Chapter 4 highlights the parallels between the information that can be determined from a scene using audio based localisation and the core goals and needs central to ubiquitous computing. The proposed embedding of computing and sensing resources into the fabric and woodwork of our everyday habitats, work and living spaces, with the aim of interpreting situational information of a space, is discussed. A core goal of ubiquitous computing is to estimate the needs of the people in a space and to react in useful ways, and with seeming common sense. The requirement of spatial context — detection of activity and sensing specific to exact locations — matches closely the short range, architecturally contained properties of sound. Some of the ways in which sound can be used in these pervasive sensor and output systems are identified, as a means of determining social context, of complimenting interaction or of recording and retaining information on the use of a space and its acoustic history. Sensors and computing, although not connected or responding as a cohesive whole, have become pervasive in our lives. These resources could be put

to use, to begin to effectively sense people and their actions.

Detecting a sound and estimating source position requires collection and processing of audio captured from spatially separated locations. Chapter 5 examines how the collection of audio can be approached in either a centralised or distributed manner, and how, looking at a number of network approaches, a distributed system carries the advantages of scalability, and tolerance to change and node failure. Key features derived from an architecture that treats sensors as autonomous nodes are identified. The implied constraints of such a system include an emphasis on local information processing, the need to assess changing network topology and neighbourhood resources, and limitations of communication to only salient information. Further constraints are imposed by the use of heterogeneous collections of devices, sharing resources to build an impromptu microphone array. These include a detachment of the algorithm from assumptions of low level hardware properties or communications methods, and designing with code-portability in mind. The Smart Architectural Surface platform is then described, including its role as a testbed for re-configurable hardware and software, and support of distributed acoustic sensing. The typical resources and constraints of current consumer electronics devices such as mobile phones are highlighted, and compare closely to those of the tile testbed.

Given the spatial properties of sound, there exist a number of implemented, and theoretically possible, approaches to the challenge of collaborative collocation of devices in a distributed sensor networks. In Chapter 6 this is discussed, and the motivation for selecting active, pair-wise ranging is given. The use of easily detectable, broadband, deterministic 'chirps', overcoming the challenges of fine grained timing requirements in time of flight measurements, and the steps from range measurements to a complete coordinate system, are outlined. The techniques necessary to determine the network conformation from sounds emitted at each of the devices are given. The selected method for subsequent determination of spatial arrangement from node-to-node ranging is detailed and justified. Evaluation both on the Smart Architectural Surface tiles and using the resources of a collection of other, heterogeneous devices, is given in the final section of this chapter.

Building on the success of colocalisation in asynchronous devices described in the previous chapter, the derivation and testing of a source localisation strategy in the same distributed sensing conditions is then described. The history and theoretical background are presented for source estimation based on time difference of sound arrival, and the use of the given resources given in a distributed microphone array to determine the position of unknown sound emission points. Measurements of sound arrival times at each of the nodes defines a system of equations that describe the intersection of a number of hyperboloid surfaces. Solving these equations, in

the presence of noise, is a nonlinear problem whose solution has seen improved solution methods over the years. A description is given of a currently used technique, in centralised, fixed microphone arrays. The use of this closed-form, spherical model, least squares method, in an asynchronous system, as an estimator, prior to a numerical, iterative error minimisation step, for each pair of sounds from a series recorded by each device is then determined. Experimental evaluation is presented, which demonstrate that coarse position estimation is possible in systems that lack the fine grained synchronisation necessary for typical acoustic tracking.

In the discussion of this research, future directions for this work are given, including possible ways of improving collaborative localisation for real time performance in heterogeneous devices, and the steps necessary to progress the source estimation from proof-of-concept to the robust tracking and mapping of arbitrary audio events. A means of extending this technique to identify arbitrary sounds in real-room conditions is presented, with preliminary practical results. This includes the successful selection of target sounds, based on onset detection, for accurately determining arrival times. Possible improvements to the system towards distributed and real time solutions are suggested. Finally, a description of the novel forms of sensing and collaboration these tools can enable are explored, and how building impromptu sensors from resources we are already carrying can benefit how people interact and use their personal devices is considered.

## CHAPTER 2

# SPATIAL LISTENING

For computing-embedded objects to know where they are, without having to tell them, is a desirable core feature in sensor networks[2], ubiquitous computing[48], robotics and other self-organising device clusters. It will be shown in the following chapters that spatial information can be derived by measuring differences in arrival times of identifiable sounds. Specifically, the relative positions of devices equipped with both a microphone and a speaker can be found, under the constraint that no initial fine-grained time synchronisation or sensor calibration is known. This is the case for a collection of existing sensor equipped devices such as mobile phones. With this inter-node coordinate system, information on the locations of other sounds occurring around the sensors can also be found by selecting audio events as common references, and using the implied constraints from their arrival times to discover their positions. Motivations for choosing audio analysis as an approach to localisation and related work in the field of spatial listening are given below.

### 2.1 SOUND

It is argued here that the role and importance of sound sensing in machine listening is derived not only from its technical facility but from the central part it plays as an indicator of human activity and communication. The strengths and limitations of acoustic positioning are given in this section. An overview of the spatial information a sound carries from its point of creation to a sensor is then described. The end of this section addresses the technical aspects of digital sound detection and production.

### *Motivation for Auditory Sensing*

"Auditory events tell us about the surrounding world. What is characteristic about these events is that they are spatial and temporal; the information specifying the event consists of variations in the sound over both short and long time periods." - Stephen Handel[21]

In designing sensor systems for humans, and in considering the extent of the information that can be gained through machine learning from an environment, much can be discerned from how animals have evolved to detect and use this information. We are sensor rich entities, collecting and processing a large proportion of the available environment changes that occur around us. We develop robust models for interpreting and combining this information. We primarily sense the world beyond our reach through measurements of sounds and light arriving at our ears and eyes respectively. Vision and audition play particularly central roles because the linear propagation of these phenomena enable us to collect information not only locally to our bodies, but from as far as these energy changes can travel. Both sound and light have wave-like properties and travel in a predictable way from their points of creation or reflection, allowing us to determine spatial information from them. Light provides the greatest detail, scattering from the surfaces of our physical surroundings. The useful light we see is mostly reflected from the surfaces of objects, with wavelengths on a scale that interacts with the materials to give an indication of properties in the form of colour. By comparison, sounds are mostly event based — longitudinal pressure waves of compression and rarefaction created by a sudden change or action of possible interest to us. The short temporal signatures, variations in the strength of the pressure waves over time, carry rich information on the corresponding temporal process that caused the sound. These are the results of collisions, breakages, air turbulence and resonance. We recognise and categorise these sounds, inferring a great deal about their creation. Patterns of consecutive sounds also carry meaning, in their timing and changing direction of arrival. We can determine source positions and some information on the surrounding scene from how the sounds streams vary as they move through space and over time. With our eyes closed, static objects and terrain become barriers, detectable only through touch. The dynamic world, marked by sound, is still very much detectable, as those who are visually impaired can attest to. Sound is significant as it is closely linked to activity, and can be generated by us and machines with ease. In our daily lives we leave trails of sounds in almost everything we do. Turning on a light switch, talking to friends, walking down the stairs, listening to music — these are marked by distinct sounds emanating from the point at which they occur.

The structure of the ear is designed to collect and respond to the local sound perturbations that have propagated to it as pressure waves through the air. What



we hear is the brain's interpretation of the measurements of these pressure changes that reach our ears over time. Organisms evolved the ability to detect the constantly varying pressure levels we call sound as it gave them a distinct advantage. Hearing was primarily a method of characterising and locating possible danger as well as a means of collecting background information. Location estimation from distinctive sounds gives an indication of approaching threats and possible prey, while tracking the direction of persistent sound sources over time allows us to update our models of our position within the scene. We maintain a concept of our local acoustic environment by positioning ourselves from measurements within a cloud of sound streams and the corresponding activity they indicate. Our use of sound has enabled us to develop complex forms of voiced communication and musical expression. The range of frequencies we can detect are indicative of human scale events, and sound has been appropriated culturally to carry many messages, subtle meanings and information. Through spoken language, sound has become our main form of communication and through experimentation in the production and ordering of sounds we have developed rich artistic applications in music and beyond.

Sounds can be characterised by variations in amplitude of regularly oscillating and disordered components, that is, the temporal changes in their frequency and noise-like elements. The time between peaks of sinusoidally varying elements of signals, are produced by rapid vibration or oscillation. These frequency components, and how they vary within the signal, indicate properties of the sound source, as do attack and decay rates and noise elements. Our physiology is particularly suited to detection of onsets. These provide clues to new sound elements and are sharp, feature-rich signatures that aid timing estimation. Our ears are wide angle collectors. The sound processed by our brains, or passed to a computer from a microphone, is a constant stream consisting of mixtures of sounds. By taking repeated measurements of the pressure variation over time, the waveforms can be interrogated. The properties of the human auditory system offer clues into effective ways to approach audio event detection and position estimation. Our ears and brains have evolved to make sense of the layered spectral components through acoustic assumptions. As an example two spectral components starting at the same time tend to be perceived as being part of the same sound. Such assumptions may also be beneficial for machine listening analysis, and require some form of probabilistic management when conflicts occur between acoustic common sense assumptions.

As an example of the information and acoustic clues we are constantly processing, imagine emerging from an underground subway onto a busy city street. By analysing the audio arriving at our ears from the many sound sources in the scene we are able to quickly start to make sense of the environment. Cars, people, trees are all generating sound. We move from the enclosed, reverberant space to a wide

open one. As each of the sounds reach us we also detect small differences between the signals at each of our ears, this adds information on the direction of the sounds and our position relative to them. The paths along which the sounds travel also add information that we can recognise and use. Sound travelling away from us can be reflected back to arrive later than the direct path, adding a number of repeated overlapping signals in a reverberation tail. We also interact with the acoustic streams, moving through the space, shifting our body, head and pinnae, and listening to the changes this makes.

### *Estimating Distance and Position with Sound*

Listeners are clearly capable of extracting considerable information on the people and activity around us from the sounds heard. We use the spatial separation of our ears to gain further information from the sound mixtures they collect. Detecting the passage of sound through a space, as it passes by the sensors capable of measuring it, provides a spatial component model of the sound, and the possibility of using that to infer information about its source. We identify and group components within the sound mixture stream at our ears that belong to one source, these can be compared inter aurally for timing, volume and frequency variation. By sampling the waveform originating from a single source as it passes two positions, the differences in the two signals indicate direction of wave propagation, environmental filtering and the effects of our ears, heads and bodies on the sound.

A local pressure change in air will propagate at a speed which depends on the temperature and pressure of the medium. The wave front spreads out spherically from its point of creation or reflection, and assuming a static, constant-density medium, reaches all points on the surface of the sphere at the same time. As the energy of a sound wave spreads out over the surface, the measured signal amplitude decreases further from the source. This volume change is inversely proportional to the square of the distance from the source. From the arrival times and volume of a direct signal at each ear we can therefore infer the azimuthal direction of the source in the plane in which our ears lie. The volume is further varied by occlusion of the head. We use this learned volume change to strengthen our direction and position estimations.

Inter-aural comparison provides a number of indicators of direction which combine to corroborate their agreement — or are assessed in combination to reach a single solution. For nearby sounds, an estimate of position can be triangulated by combining a number of direction estimates from different positions as we move our heads, and our proprioception<sup>†</sup>. This requires a number of sounds to be emit-

Our sense of our own movements.

ted in succession from a single sound source. If we could compare what others in the room hear with the same accuracy that we compare the signals from each of our ears, we could establish even more fine grained spatial sense from the acoustic information. Turning a number of direction estimations from pairs of detectors into a triangulated position estimate of the sound source. Recording sounds across an array of microphones allows this to be achieved. As we will see, the challenge is then how best to create such a network of microphones with known separation distances.

A sense of distance can also be construed from prior modelling of the expected sound and the effects of distance. Higher frequencies are absorbed more readily by the air than lower frequencies, so the degree of filtering can indicate distance travelled. Modelling of multi-path components of a sound as it is reflected back to the ear from surrounding surfaces can be combined with a knowledge of a space to further ascertain a sense of distance. Through microphone detection and processing of the signal, these signifiers of sound direction are possible constituents of machine listening. Repetition of audio sensors allows us to use the relative filtering, loudness or timing differences to infer a great deal about direction and position of sounds, as well as allowing suppression of unwanted noise and background sounds. As with our own hearing, building estimates based on several different estimators of position are inherently more robust to measurement errors. Of the indicators of position that audio offers, timing is well suited to machine listening. Delay measurements can be made with significantly more accuracy than comparative volume level measurements, and are less likely to require calibration between detectors of varying characteristics. Timing measurements are general to any space<sup>†</sup>, there is no need to determine the transfer function — the filtering due to a microphone position and surrounding relative to others — of a particular environment, or relate measurements to proprioception. Modelling an acoustic environment in a machine listening system requires prior assumptions about sound properties and considerable iterative prediction building.

However, as described in Chapter 6, the speed of sound varies with the local properties of the air.

Source location information can be inferred from the delay between spatially sampled measurements. Direct distance measurements can be made if both the emission and arrival times of a sound are known, or the time of flight can be deduced. Once the length of the path along which a sound has travelled can be measured, sounds can be actively produced to act as a ranging mechanism. It will be shown in Chapter 6 that this technique is suitable to a number of independent sound sensors which are able to share information to determine spatial estimations.

The benefits of audio ranging advanced in this chapter show it to be a potentially robust, room scale form of information gathering. However, there are constraints

to audio based localisation. Sound travels through a medium such as air at a finite speed, which varies with local temperature, pressure and air movement. A possible source of error is the fluctuation in the speed of sound with air temperature throughout a room. Multipath reflections and signal occlusion can cause sound path lengths to appear longer than they are. Sounds are muffled or stopped by physical blocks. Compared to radio frequency electromagnetic signals, sound has limited penetration of solid barriers. However, the wavelength range of audio is large, so that scattering from features of a certain size is likely to only affect a small component of a signal. A further source of confusion is multipath through different materials, such as a knock on a table travelling both through the air and through the wood.

In ranging between devices, we will see, generation and emission of reference sounds is a typical approach. Ultrasound, beyond the threshold of human audibility, offers the promise of signals that can pass undetected by people sharing the space with the system. Generating sharp onsets to accurately establish timing requires broad band signals which means that ultrasound chirps contain audible frequencies that lead to a click heard with each emitted pulse. The alternative, sounds in the detectable frequency range, must then address the questions of coexisting in this acoustic channel. That sounds made by technology can be both so effective at catching attention and so irritating when inappropriate further demonstrates the importance we placed on detecting key audio.

### *Activity and Scale*

Sound is a natural choice for human-scale, people-centric sensing. It is closely linked to physical activity, indicating movement, tasks and interaction. Everyday sounds are easily detected within a limited area, having a human scale range of dispersion and architectural containment. The importance of audio, for social communication, entertainment and context make it an information rich channel.

The use of sound in computer-human interaction benefits from its people-centric properties. Since sound waves are intentionally attenuated in our living, social and work environments and drop off in volume over a short range, sensors using sound are naturally bounded to similar near-field attention range as their human users. Sounds measured in a room consist of the local context and human presence, and characteristic dampening and reverberant properties. It may be possible to start to model the local space in addition to recognising distinctive audio signatures of human speech and movement in information gathered from both coarse and fine time-scale analysis.

## *Technical Aspects of Sound Detection*

Throughout this chapter the facility of detection and creation of sounds has emerged as a key feature in its uses. The ease with which sounds can be processed, stored and transmitted technically, and the information that can be gained from relatively simple analysis, together make sound sensing a viable approach using current technology.

The process of digitising an continuously-varying analogue signal into discrete representations of time step and amplitude level carries the inherent considerations of sufficient representation and aliasing effects. Care must be taken to obey the Nyquist-Shannon sampling theorem[50], which states:

"When sampling a signal, the sampling frequency must be greater than twice the bandwidth of the input signal in order to be able to reconstruct the original perfectly from the sampled version",

The incoming microphone signal must be filtered appropriately. Digitisation of an audio signal enables storage, analysis and transmission of the resulting waveform with ease. It is common for signal analysis and digital filters to run in real time, for example, and many hours of audio information, using lossless compression, can be stored in small, pocket-sized, devices. A typical audio sampling rate of 44100 *Hz* yields a spatial resolution on the order of a few centimetres.

The success of the telephone has motivated the development of practical, low cost, low power and physically robust microphones and speakers. The development of a practical foil electret<sup>†</sup> microphone at Bell laboratories in 1962 has lead to small package microphones with a broad frequency range, sufficient characteristics and negligible costs. Technology reached the point, at the beginning of this decade, at which the resources necessary for analysing sound are compact, energy efficient and low cost. Relatively large amounts of audio information can also be stored easily. Many of the consumer electronic devices we use are reaching the point at which advanced signal processing techniques can be implemented, some executing in real time. The inclusion of audio capabilities in a wide spectrum of consumer electronics devices is also driven by benefits that can be offered at a low extra cost. Further to primary sound functions such as telephone, music or meeting recording, examples include voice notes and sound alerts in many devices. It is even conceivable to build large numbers of microphones and speakers with associated in-place processing, into many architectural and household objects. It is this vision of widespread, distributed sensing that will be examined in the next chapter.

Permanent electrical polarisation of foil which acts as unpowered capacitor.

### 3.2 SOUND SENSING

This research is focused on providing methods of inferring distance and location using acoustic sensing. In this section a description is given of the ways in which estimates of positions of sounds, and the events that cause them, can be useful to people. The machine listening tools that have been developed to support daily living are presented, and the ways in which these can be enriched by position information are given. If the location of objects and actions that make sounds can be determined, then this provides a basis for recognising, recreating or interpreting them. Devices can emit and record sounds as a means of self location and begin to map the audio context and history of a space. Tracking and clustering of this data could then allow longer-term patterns to be found, which in turn can be used to perhaps develop a better understanding of situations and daily rhythms. These resources, which use sound as a ruler to measure out space and distance, are valuable to systems that intend to track and draw meaning from the actions of people. Records of transient audio signatures can also be gathered, made more permanent and presented back directly to people in ways that allow them to reflect on and use this information.

As humans, we tend to forget how much we use our auditory system to monitor our surroundings. As described in the previous section, we are able to track and identify sounds, such as the words spoken to us, by using binaural differences, and hence we learn models of sounds occurring around us which augment an expectation of the sound waveform properties. It has been shown that learned assumptions about how the acoustic events in living habitats occur, and the ways in which sensory systems have evolved play a strong part in how sounds heard are interpreted. For example several frequency components arriving at our ears with differing interaural delays are naturally assumed to have come from different sources. The position of sounds are then an important component in listening, and can be used to augment the models of sound spaces implicit in the assumptions on which acoustic sensing technologies are developed.

#### *Machine Listening*

Sound analysis is well suited to human-centred contextual information gathering[13]. The same benefits of sound sensing in living organisms are true for machine listening. It is clear that a broad range of information is available in sound measurements. Acoustic machine analysis is motivated by the success of human audition. This spans low-level waveform signatures indicative of the properties of individual sound sources, analysis of sound clusters and ordering patterns that mark a physical event, and high level semantic or aesthetic meaning, from speech

recognition to mood analysis. Methods of machine analysis reflect this, low level phase and frequency, through onset detection to higher level ordering and composition estimation. Machine listening in everyday life tends to fall into three fields. Speech recognition focuses specifically on features that can be used to estimate the semantic content of spoken words. Music analysis extracts on tonal and temporal patterns, often drawing on assumptions of certain musical styles. Audio scene analysis which is a more general, recent, approach to identifying sound events and modelling sound environment. Inferring where a sound originated, and the journey taken from source to capture can play a part in aiding understanding in each of these cases.

Three concepts central to processing the spatio-temporal acoustic information from a room, 'event', 'context' and 'history', reflect three corresponding scales of both time and space. The terms are given here in relation to this research:

- *Event* - Describes a single action or short series of steps (such as the series of clicks made when open a drink can, for example) which we recognise as a single sound. These are distinct features in time and space.
- *Context* - Encapsulates concepts of situation, place and time. Activity leading up to and following a sound. Audio location cues hint at current place, tasks, emotions, etc.. Determination of context has become a central goal in human-computer interaction.
- *History* - Shifting in context or patterns of events over time. Use of space, and both recent and long term patterns. Inferring normality and abnormality through clustering and modelling from past observations.

### *Sound Recognition and Audio Scene Analysis*

The recognition or transcription of speech by machines is a core challenge in the AI<sup>†</sup> and interface design communities, and was one of the first areas of audio scene analysis to be thoroughly explored. Here an approach must be developed that focuses on a single sound source, excluding the distracting features of many others, in a real room environment, in order to extract the meaning from the words we ourselves find so easy to understand. The techniques popular today are starting to make use of the approaches it is believed support our own listening. It seems we develop an expectation of the sounds arriving that greatly aids comprehension. This is evident in the theory of glimpses too, where it is proposed that we use the redundancy of information in spoken words to recognise the meaning even when the word is masked by noise in parts. We also use expectation in the meaning of the

words to predict what makes sense for someone to say next. Music analysis too, has benefited from expectation built on prior examples, in that case, of musical genre or composition.

The concepts used in voice recognition are being extended to audio scene analysis in general. For example, how can a machine recording a street scene, or the sounds in a park, pick out and identify separate audio events and streams. A number of recent attempts to enable machines to listen, in the sense that we ourselves take for granted, identifying distinct sounds as they arrive, are based on building effective models of the expected sounds, and the types of sounds that the system expects to encounter. This reflects the ways in which our own brains may approach listening. The key to human success at a cocktail party, for example — the ability for someone to track a single conversation in a chaotic acoustic environment — seems to be the redundancy of information in the system we have developed. It is worth noting that this example is tempting and possibly misleading, as we do not extract a completely separated audio stream before comprehending it, but rather extract meaning from the mixture of sounds we hear. The time and amplitude differences between ears, allow us to focus on sounds at a single point in space. Visual information, such as the movement of someone's lips in time with the sounds we hear, further aid our comprehension. This successful approach of robustness from redundancy can be used to improve machine listening. Expectation can be aided by clues of position. Although we can listen to a monophonic recording of a conversation, and pick out the participants and meaning, our accuracy is reduced as the channels of information become limited. Estimates of spatial information of incoming sounds are then not only useful as a separate sense, for mapping or tracking sounds, but also become a valuable component of a robust 'listening' system. This entails collecting partial information on the nature of the sounds recorded to be combined with knowledge of the statistical features of the sounds, video, prior assumptions and models of a scene and a wide variety of other sensing techniques. Spatial analysis stands out as demanding little resource overhead or prior assumptions in microphone array systems.

### *Extracting Acoustic Events*

The goals of Blind Source Separation (BSS) are to extract a single sound from others that overlap it, and from complex filtering effect of passing through an environment. Westner, for example, examines the extraction of sounds in a room to aid an object-based audio approach[61]. In a single recorded stream BSS is possible for mixtures of statistically independent sounds using Independent Component



Analysis. Smaragdis *et al.* provide a recent overview[49] of the current abilities and applications of ICA in machine listening, and its constraints. In real room environments the technique is degraded by non-localised noise, reverberation and non-stationarity of the source. Attempts have been made to more fully model these interactions in order to account for them, but an effective combination to perform robustly in real room separation cases has yet to be demonstrated. It is also unclear where the distinction of room filtering should end, as the human body, for example, also adjusts sound as it is produced or detected. The inspiration for complete separation comes from our own ability to focus on single sounds in noisy environments, but it is clear that we do not completely separate sounds, just extract the meaning we need.

For cases in which several sensors can be used, the signals from each can be compared to aid source separation. Focusing on a sound at one point in space involves, in part, adjusting the delays in a signal to match those caused by the separation of our ears and the angle of sound arrival. This causes sounds from other directions to lose phase coherence and therefore prominence, and is the principle behind beam forming[37]. This often accompanies a pre-formed model of the expected positions in which the sounds may have occurred. Beam forming can be applied concurrently on any number of positions in a room given the resources to track target sounds. Beamforming is used to focus on a sound stream from a single position, by degrading the other sounds, with different delays, that arrive at the separate sensors. From the delay that delivers the strongest target signal, an estimate of where the source is can be obtained. In larger microphone arrays, repeat matches from a stream of sounds further improves position guesses, strengthening the recorded sounds as they combine in phase. Other sounds, which do not arrive at each microphone in phase, and microphone noise, average toward zero and are suppressed. Performance in delay estimation for unknown source location, can quickly degrade if the effects of reverberant environments and non-stationary sources such as people speaking while moving their heads are not taken into account. This is a fine grained source location problem. It will be shown in Chapter 7 that given looser position accuracy constraints, real-time passive localisation systems have been made. Acoustic event positioning has a long history in SONAR<sup>†</sup> and ballistics location (for examples see [8] and [17]). It is now beginning to be used to mimic our own ability to track and concentrate on a single sound source, for implying attention in robot gaze and other affective computing scenarios.

Sound navigation and ranging.

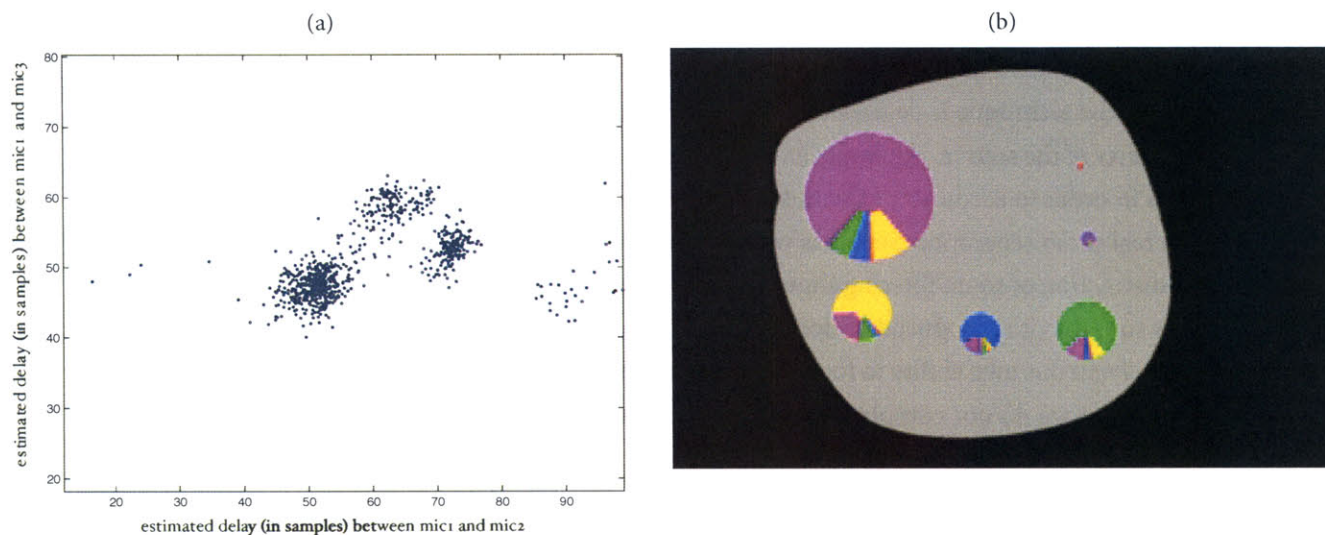


FIGURE 3.1. (a) In the work of Basu *et al.*[3], plotting delay estimations from two microphones yields clusters corresponding to conversation participants. (b) DiMicco[14] presents a method for displaying conversation participation to those involved.

### *Using Spatial Information from Sound Signals*

Localisation of sonic events can play a unique part in many computer tasks involving categorising acoustic scenes. Looking again at ourselves as a successful example, our understanding of the environment we are in is enriched as we track groups of sounds temporally and spatially. Imagine the case of someone approaching you, from behind, through a forest. They will make a large variety of sounds as they move, breaking sticks, rustling leaves, breathing, etc., but each sound occurs along a slowly varying spatial path. We are able to easily identify these as a related stream, and concentrate on them closely. Spatial information also allows us to separate two otherwise similar sounds because of the variation in their arrival times and volumes. If these 'common sense' assumptions of listening can be applied with spatial information collected with sound recordings, scene analysis may become better able to correctly interpret confusing recordings. Dan Ellis[16] describes how his prediction driven approach to auditory scene analysis could be effectively augmented with position information if it could be collected.

Spatial acoustic mapping can provide a means of tagging features in audio recordings to aid ease of browsing and comprehension. For example, interpretation of long term audio recordings. We have the resources to store continuous streams of audio from our lives. This form of memory aid and information collection is potentially very useful. However, effective methods of searching these audio



FIGURE 3.2. Informal, unplanned conversations form many of our work and social interactions. How can spatial machine listening compliment this? (after David Rees)

streams are currently limited, with high speed audio only effective up to a certain speed, and direct visualisation of waveforms or spectrograms only cryptically useful. Approaches to this problem have tended to focus on the extraction and use of semantic information from audio for memory aids and rapid retrieval. Spatial information plays an important part in memory and provides an extra channel of information. For example, Basu[3] used a body-worn, synchronised, three microphone array to estimate azimuth of incoming sounds, based on cross-correlation. The data collected exhibited clustering of the positions of talkers. Attempts to segment and categorise the recording based on situation and place could benefit from partial maps of the arriving sounds. Work on labelling and segmenting personal audio as a memory aid or as a searchable record of events, could use spatial flow as a method of quickly visualising or clustering the sounds recorded. Presenting back information on conversation turn taking, based on acoustic participation, has proven effective in aiding groups who consider their information sharing process to be in trouble. In the research of DiMicco[14], groups were presented with spatial representations of their relative participation in a meeting. It was found that information exchange in groups was encouraged by this spatio-temporal acoustic analysis. The system used close-talk microphones for each participant, and spatial position was manually entered. People read rich social aspects into very simple transcriptions of audio cues.



## CHAPTER 3

# AUDIO SENSING EVERYWHERE

Knowledge of sound location is a valuable component in systems that attempt to track, model and pre-empt people's actions and needs. Examples include the provision of location specific content, environments that respond appropriately to the people moving through them, covert surveillance, health monitoring and 'smart' rooms. Location plays two main roles, the first is self-calibration. In systems intended to contain large numbers of sensors and independent processing objects, manual calibration becomes impractical and undesirable, and yet position within a space is desired information to allow objects to adapt or best fit their situation. This need reoccurs in many systems involving numerous separate subsystems or disparate processors. The use of self-localisation is two fold, not only does this allow objects to calibrate their positions, but it also enables them to track themselves as they are moved. The second vital use of the ability to locate is to identify and respond to people and events in an environment. Not only to focus on one person speaking to gain meaning from those sounds, for example, but also in developing a general impression of awareness and modelling context of a scene.

### 3.1 UBIQUITOUS COMPUTING

The field of ubiquitous computing<sup>†</sup> explores the possibility of pervasive invisible processing, sensing and actuation, built into the fabric of the objects and architecture around us, activating everything. Accompanying this shift from personal fixed computing to universally accessible computing commons, was proposed a corresponding shift from users translating their needs into the language and architecture of the computer, to ubiquitous computing, monitoring the context and actions of people and modelling their intentions to determine how best to intervene and offer assistance. By adding computing to everything, it is argued, it is always at our fingertips. Van Kleek offers a recent overview of the field[56]. Core chal-

The origin of the term and the concept is generally attributed to the late Mark Weiser from the Xerox Palo Alto Research Center (PARC) in the early 1990's.

lenges include embedding resources into the environment, organising information from a scene and inferring situation and intent in useful ways. As we will see in chapter 4, a large body of research has investigated how processing hardware can be shrunk, embedded and organised. An early experiment in ubiquitous computing was Xerox PARC's CoLab Project[58], which used 'inch', 'foot' and 'yard' scale devices throughout an office environment, interconnecting all three over wireless networks. A considerable body of work has been developed into how semiautonomous programs can operate in this environment, acting on behalf of people to monitor needs and provide pertinent information. The Agent-based Intelligent Reactive Environments (AIRE) group part of the CSAIL Oxygen project at MIT, for example, uses the Metaglove programming framework[11] — a system of co-operating agents that monitor external information, the space and the people within it, negotiating confidence and usefulness of the information they can provide.

The tools that spatial audio listening offer, of sensor calibration, tracking and space monitoring, are central to ubiquitous computing environments and 'smart agent' systems. Attempts to cluster this behaviour information, such as use of space over time, can reveal patterns from which predictions can be made. Sound positioning can play a role in fine grained detection of space usage history. Positioning of objects within a space and estimation of human speakers and other sound sources facilitates the determination of social context. Machine learning offers the tools to draw together sensor data. For example, in his research, Gorniak[19] describes how situational modelling and semantic grounding can be used to link position information and speech recognition to disambiguate a phrase directed at a 'intelligent environment' such as "help me with this" through interpretation of context. Research into effective tracking through spaces and buildings suggest that by combining predictions of position determined from a number of sensing modalities, robust estimates of the movement of people can be found[22]. Noisy measurements can be combined in a failure tolerant method using Bayesian statistical descriptions of the confidence in each measurement.

The study of the spatial elements of human dynamics include how people move relative to one another and the space they are in. In Weiser's paper[60] outlining the dreams and initial achievements of ubiquitous computing, he describes a scene from a future world in which someone can look out of their window in the morning and see the activity of the last few hours imprinted on the ground of their neighbourhood. Sound sensing lends itself well to collecting audio footprints of activity which can be displayed as digital patina, mapping routes through a space. It is worth noting the fields of psychogeography and location based content, which study people's interactions with their geography, and how to enable the marking up of these spaces with 'just-in-place' information, respectively. The relationship



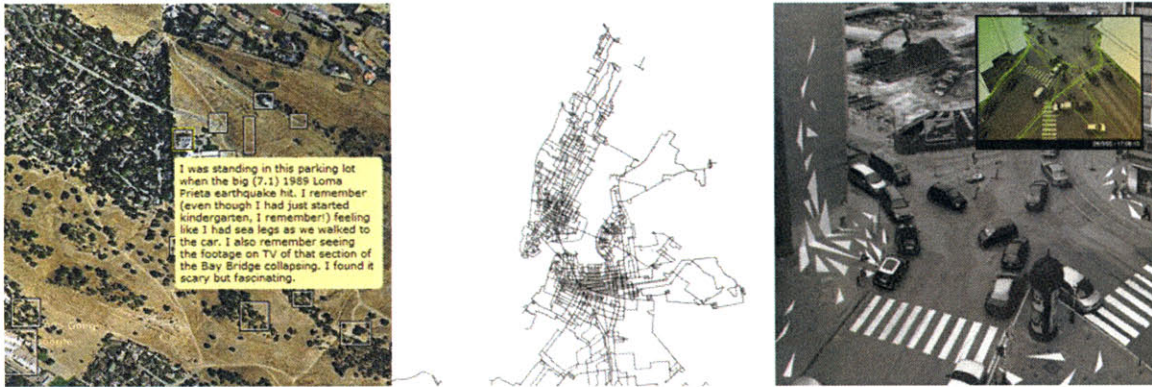


FIGURE 4.1. Examples of personal and public mapping and memory. (a) Image taken from the memorymaps pool on Flickr, in which people mark Google satellite images with personal memories of the space. (b) Partial map of New York based on everywhere one person walked in the city. (c) Street installation made by tracking and then exaggerating pedestrian movement.

between personal mapping and memory is already being explored in a number of projects. Figure 4.1 shows examples of Google maps with childhood memories added in Flickr[25] to create 'memory maps', New York walking map[24] and a street art piece derived from the usage of the space[31].

### *Common Resources on Heterogeneous Devices*

The pocket tab and note pad size devices in PARC's project, envisaged in daily use by Weiser, are now commonplace in the form of laptops and mobile phones. The hardware is still used as single entities, with programs running in isolation on a person's devices. There is some synchronisation offered between an individual's computers. Online services such as Gmail.com and creative tools such as PLW's Open Attaché[62] are shifting data into accessible-anywhere, networked formats. However, the interconnected information sharing networks and the location, situation and context awareness have proven difficult to coordinate and deploy. The resources needed for ubiquitous computing are not yet built into everything in our environment, but they are being carried into our work and living spaces in our personal devices.

As mobile devices become common place, the formation of impromptu distributed networks is possible. The resources to connect are also available in these consumer electronics components. Examples of protocols that support clustered connectivity include Bluetooth, which can form piconet clusters and supports rout-

ing between these, and the resource discovery protocols of Rendezvous ('Apple Bonjour') and UPnP (Universal Plug and Play). Organising the use of these resources remains a challenge to be implemented and adopted. The nodes of these networks should be able to offer up spare resources such as sensors, processing power, output, storage and access to alternative communication channels. Most of these devices also serve in their primary role, telephone, music player, camera for only a small amount of time. This is not fully ubiquitous computing, the devices are still devices, each sold with a primary application, but just as distributed computing can use unused processing time in many machines, running across the PCs of people looking for signs of extra terrestrial life, so too, our existing devices can contribute resources to a pervasive computing ecology. These impromptu resources can compliment dedicated sensor environments that have been built to adapt to new components and resources.

There seems, then, to be a strong motivation for nomadic architectures, subsystems that can function on a number of disparate infrastructures and can offer up their resources to unknown systems. It can be argued that it is unlikely that any single networking protocol will be pervasive. The needs of people using the system, seem to always be changing, and evolving. The current economic and academic models do not fit with the needs of an open connected system. Competition and new ideas always complicate a single, sufficient protocol. Further, an agreed protocol must be fixed, and so may not be beneficial to the growth of new capabilities. Successful systems such as the internet and email, tend to be augmented as their limits are discovered for example RSS[1], BitTorrent[12] and Instant Messenger[45]. So too, in distributed audio sensing systems, heavy reliance on network or device architecture should be avoided to flexibly fit with future systems.



## CHAPTER 4

# SENSOR NETWORKS

Our daily acoustic environments are information rich and linked to activity, with audio localisation playing a core role, or able to enrich many listening tasks. Chapter 2 and 3 described the ways in which sound position information has been used in machine listening and the part it can play in the emerging field of ubiquitous computing. This chapter outlines the motivation for using distributed sensor networks as a means of collecting and using sound, and identifies the suitability of the features of viral networks for organising distributed microphone arrays. The assumptions introduced by distributed systems are discussed, as are the further assumptions implied in the use of generic systems and existing personal devices.

### 4.1 ORGANISING DISTRIBUTED SENSING

Predictions of pervasive computing present an appealing future in which resources are spread through our daily habitats and establish close links to personal or wearable computing. Sensors must be abundant enough to detect situation and connected enough to interpret context and react with 'common sense' to our needs. In order to support effective sensing and interpretation of results, especially apparent in large densely populated collections of sensors, it is clear that an infrastructure that can adapt to a regularly changing ecology of resources is desirable. If the system is to reflect the modular way in which our living environments are built and used, then an approach based on sensors as autonomous nodes, with individual processing and communication seems potent. The system must then self-organise — sharing tasks and identifying new resources on the fly. The current technology in our pockets and living spaces is composed of heterogeneous and largely unconnected devices, but the resources for communication and sensing are already pervasive. The aim of this research is to formulate tools that can build on existing mobile computing towards a ubiquitous sensor environments.

### *Microphone Arrays*

Microphone networks use their spatial distribution to sample propagating sounds at a number of locations and infer information from the differences in these measurements. From timing of sound arrival at each sensor, they can collect information on their own spatial positions or other sound sources. Measurements of sound arriving at a point in space will contain a mixture of all the sounds crossing that place at that time. It is by comparing sound arrival times and other properties at several separate positions that information on source positions and paths can be inferred. The challenge then is to capture these signals at each place and transfer the parts that are of most interest for comparison amongst each of the sensors.

### *Viral Computing*

Organising distributed networks of sensors to coordinate detection and tracking of sound, and to collect together useful information requires a dynamically adapting architecture. Taking the end-to-end principles that have proved so successful in the growth of the internet, Andy Lippman and David Reed describe networks that can operate without central authority, in a scalable and agile way, 'viral networks'[34]. The driving motivation in viral communication is to produce systems which can be incrementally complimented with additional resources, providing new information or connectivity to the ecology, in such a way that increasing the number of participants improves the overall performance of the system. Many existing networks demonstrate the opposite property; as the number of elements increases, the load on centralised components of the system increase. Valuable information collected becomes unwieldy, or the communication or processing requirements scale directly with the number of nodes, and bottle necks are formed. Some of the key properties of a successful distributed network can be explored in several simple examples of current networks. Examples of these standard, centralised systems include many security cameras feeding live video to a single monitoring facility and mobile phone networks. By examining how these centralised systems can benefit from decentralised infrastructure, through developing an understanding of distributed sensor networks, the constraints on tools implemented in such a system can be established.

1. In the case of a camera network, full frame video is streamed from each of the cameras to a central system for analysis. Not only does this demand high network bandwidth, but a single processor, or human, must sort through and identify points of interest on the multiple streams, in real time. The alternative is to process or analyse the video at each of the cameras, passing only key

information or scenes of interest back to the central monitoring facility. This leaf node processing reduces communication bandwidth considerably. It is a key element in scalable systems, allowing for further nodes to be added on the fly, with little increase in central overhead.

2. In current cellular telephone systems, as the number of phones using a single cell tower increases, the overall performance is eventually degraded. In phone usage at a large music festival, for example, many people are placing calls locally, and yet they must all be routed through the central system. In a peer-to-peer alternative, the devices must negotiate routing between nodes in an ever changing system.

From the above examples the desired properties of a distributed sensor network begin to become apparent. Processing at each node, passing only salient information improves performance and limits communication bandwidth whatever the network size. Spreading computational, storage and routing load across the network may reduce bottlenecks and can allow the system to increase in performance as it scales. A decentralised network structure is required that can self organise on-the-fly, adapting to addition and removal of nodes.

### *Related Work: Distributed Sensor Networks*

The limits and benefits of large systems of collocated sensor devices have been explored in a number of recent research projects. Building working prototype systems motivate strategies and programming solutions that are invaluable as the size of processors and devices continues to shrink to the point at which they can be scattered onto or embedded in the surface of objects of every room.

- *Paintable Computing* - Butera imagined grain sized computing that could be mixed with a medium and painted on to any surface[7]. An ecology of resources scattered through dense network is formed. This very constrained system demands distributed programming and code execution across asynchronous processors.
- *Smart Dust and the NEST Project* - Working towards self contained millimetre packages, with the aim of eventually suspending a cloud of nodes in the air. Smart dust embodies many of the challenges of small sensor devices[30]. How to distribute power or whether energy can be collected or stored locally. Programming environment must conserve power and exhibit redundancy to failure. TinyOS, a limited operating system specifically for small nodes in a

sensor network has been developed.

- *Pushpin Computing* - Provides a test bed for the paintable computing project, this system solves power constraint by pushing into wall and power substrate, power harvesting by mounting in to power and ground layers below a surface. Lifton describes the localisation capabilities of the system using ultrasound from a single beacon and colateration[33]. Stackable modules can add resources. Local neighbourhood infrared (IR) communication is used, which allows simulation of much larger node systems through local neighbour infra red communication. Timing is established through external IR beacon.
- *Crickets and Motes* - Small battery powered devices. Based on research into active badge tracking, crickets provide location-support to aid device positioning[44]. Devices maintain privacy, using map and beacon location information. Motes were built as a testbed for smart dust development. These nodes often work in collaboration with higher resource hub nodes, such as iPags. They tend to use infrared communications and ultrasound ranging.
- *Robomote and Eye Society* - These projects explore proprioception-like feedback to sensor reading with dynamic nodes and the effects of varying sensor position. These move to improve sensor readings and scene understanding, or use movement through space to compliment readings and infer path. Robomotes are free to move on a surface, ranging using the mote ultrasound system[51]. Eye Society investigates mounting camera and microphone sensors on ceiling tracks, that are free to adjust positions to update measurements through spatial variation[38].
- *Swarm Computing* - Much more dynamic systems, practically exploring the same constraints as those identified in the air borne smart dust system. Routing, sensing and code execution must be formulated in a constantly changing neighbourhood. These problems of flocking behaviour and dynamic graphing are beginning to be addressed in swarm computing systems such as the helicopter cloud proposed by Owen *et al.*[23], that approaches this problem using devices with greater individual resources and hence autonomy.

Many systems are constantly dynamic, requiring a running update of graph information such as shortest route and local neighbours. Group forming strategies allow for temporary allegiances to be formed.

Auto calibration and self-organisation are clear benefits to an organic network of sensor nodes. These properties become vital in many circumstances, for example as the number of nodes increases beyond practical manual calibration, or in remote deployment, or indeed where including a calibration interface on each node

is unfeasible. Position and synchronisation are core system information, sensitivity level, orientation and knowledge of capabilities of local neighbours may also be useful to determine. Coarse timing is necessary for many network interactions, such as maintaining version history in file systems, and logic flow in distributed programming. Position and fine synchronisation, as we will see, are vital to many sensing, and particularly acoustic analysis tasks. Many devices are characterised by short term timing accuracy, but lack long term synchrony with neighbours. Sensing often relies on gathering information from a change in measurements over time, this question of how to compare the times at which recognisable changes occur will be a common theme in the following chapters.

Strategies must then be developed not only for inferring the resources available in a network — tracking and adapting to their changes over time — but also ways of determining how best to balance the local processing and storage of nodes and the cost and performance of passing information between nodes. The limited resources of memory, computation, power and bandwidth, may also be time varying, furthering the challenge to organising tasks in such a system. New coding approaches are required, that have adaptation and optimisation at their core. Ultimately a system that seeks to discover resources, assess costs and elect tasks is sought.

### *Assumptions Imposed*

An algorithm must fit certain criteria in order to be usable in distributed, dynamic systems such as those described above. By considering the constraints that collaborative, impromptu sensing with device nodes implies, successful tools can be implemented.

A distributed, viral communications system implies a topology and local neighbourhood that is unknown. Microphones should be associated with a local processor and semiautonomous resources, in such a system emphasis is placed on analysis at the nodes as this reduces communication bandwidth and support scalability of the ecology. Communication is minimised, with deterministic calculation of reference signals or digital compression of sounds, limited by the information that must be retained for it to be meaningful. Practically this relates to sampling frequency and then processing the result, compression perhaps even down to just a few key arrival times. With no centralised tracking or maintenance of nodes, resource discovery is desired; sensors, communication links or processing may be added or removed. System stability can only be assumed in certain cases, such as short term stability if mobile devices are placed on a table, or longer term stability if they are embedded into a room. Motivation for distributed algorithms for large computation is clear.

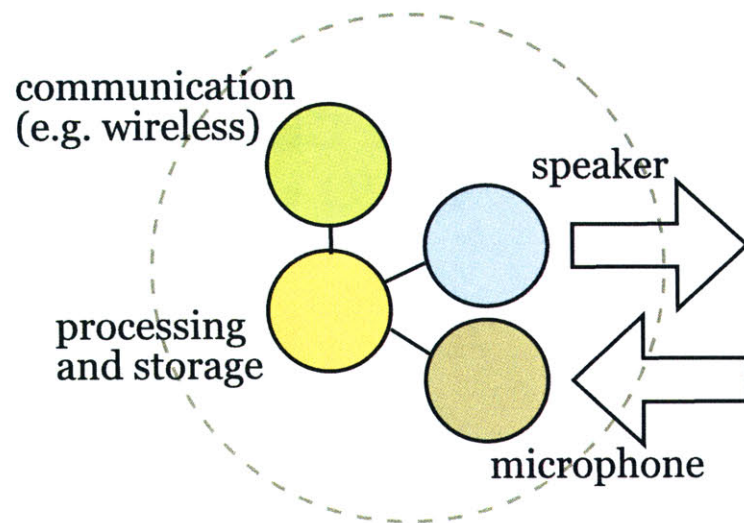


FIGURE 5.1. Placing the sensing and actuation abilities with an autonomous processing node, enables scalable, distributed systems

Where possible, algorithms that can be calculated in clusters and then grown to global solutions are sought. For the centralised tasks that remain, leader election must be accomplished.

Further assumptions are introduced with generic systems of heterogeneous devices. No low level hardware access can be depended upon nor the addition of specialised resources. Microphone, speaker and wireless connection typically included in many electronics goods. Layered communication with hardware resources must be assumed as code may not have direct access or control of settings. Constraints of synchronisation and position and the accuracy with which these can be determined must be considered.

## 5.2 SMART ARCHITECTURAL SURFACE

The Smart Architectural Surface (SAS) is conceived as a distributed system built from unit elements in the form of large wall tiles, which can be affixed to any surface to form an acoustically and visually actuated, sensor rich surface. Drawing together inspiration from smart rooms, distributed sensor networks and responsive architectural building materials research. Devices that cover surfaces can be decorative and non-invasive, but are easily in view and can sense many positions in a room and cover a large real estate.

The goals of the SAS are a physically and logically re-configurable environment



FIGURE 5.2. Many consumer electronics devices carry the resources necessary to establish a distributed sensing network.

that is fault resistant to addition, removal or failure of components. Tiles can operate in isolation, but benefit from connection and collaboration. As a viral network, the performance of the tiles should grow with their numbers.

Two strata of sensor networks are currently pervasive in research, centralised smart rooms and distributed sensor networks built from low cost, limited capability nodes. Benefits of the latter have been discussed in this chapter. They are very cheap, disposable and can be deployed in high density and wide coverage, being small enough to attach to objects and scatter through an environment. They act as test systems for a new scale of tiny computing of the sort proposed in the Smart Dust and Paintable Computing projects. However, limited processing and communication capability currently available heavily constrains the computational tasks that can be achieved locally at each node. On the other hand, augmented and smart room research has pioneered the use of high resolution, specialised sensors and hardware. Following a typical computing infrastructure and centralised approach, information is collected from fixed locations in a room and then transported back to a central processing system. These sensors are networked, synchronised and hand calibrated. They have central failure points, but powerful computation. For example the Large Acoustic Data (LOUD) microphone array[59] is built to pass



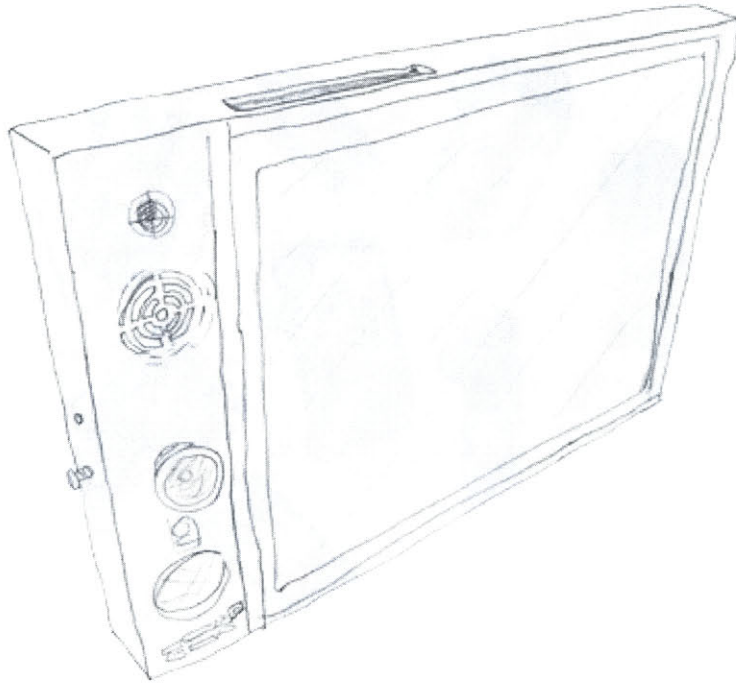


FIGURE 5.3. Each tile has a large front screen face, and is enclosed in a thin plastic casing. Sensors are mounted along one side.

vast quantities of measurement data efficiently back for processing.

The SAS tiles fulfil a third sensor category, and fit between the first two. They are equipped enough to operate in a self-sufficient manner, but are built on a distributed architecture to reap the benefits of re-configurability and rapid installation. The device capabilities are equivalent to currently pervasive consumer hardware, which are already reaching pocket sized devices. So methods developed on the SAS architecture will soon be on Mote size devices. This system addresses how clusters semiautonomous components can share resources for mutual gain and investigates programming of existing consumer electronics for secondary use of the technology. In his thesis on the SAS architecture[43], Pilpré outlines a number of mutually compatible roles the tiles can perform:

- a network of dense sensor nodes (a microphone and a camera array)
- a wall of effecters (a giant screen or an array of speakers)
- a cluster of PDAs
- a testbed for distributed sensing applications.

The tiles can be placed through a room, and quickly rearranged in a practical



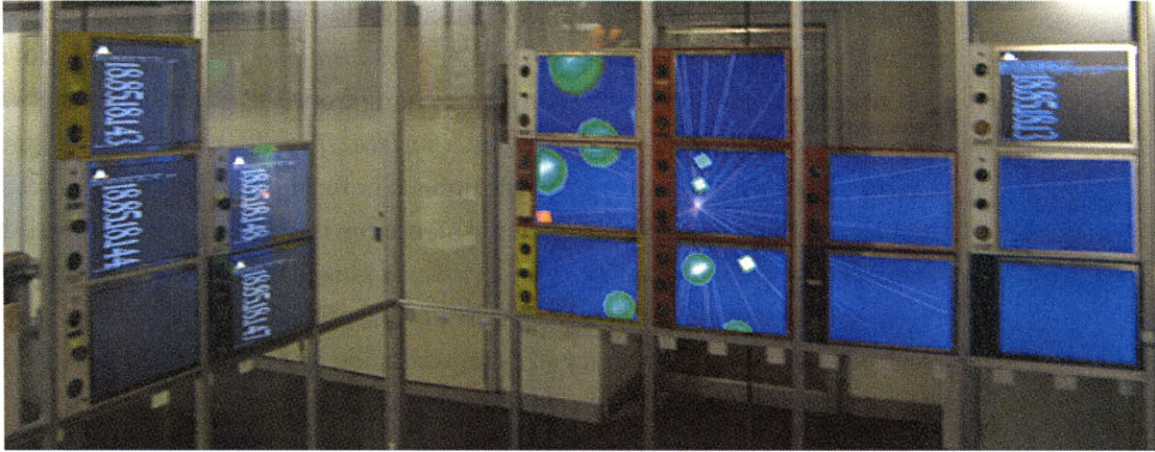


FIGURE 5.4. Example of SAS tiles mounted on a frame along two glass walls. A distributed data display can be seen across the screens of one wall.

way. They are designed to be tested in real room environments, and must demonstrate a tolerance for 'real world' problems. Humans present difficult systems to model and respond to, we multi task and have multiple uses of a space. Acoustically, rooms are very different to test conditions, they contain many sources of background noise and are reverberant spaces. Constructing a system in this busy habitat motivates the research to meet these challenges.

The hardware used is relatively low power, with the screen backlight consuming the largest fraction of the energy. The associated heat benefits result in a solely convection cooled system in a slim wall mounted package. The parts are also relatively low cost, so that large numbers can be constructed and used. The SAS tiles are seen as a testbed for future devices with similar characteristics.

The tiles run the Linux operating system, and offer typical resources of communication, and microphone and speaker performance. Appendix A details the technical specifications of the SAS tile design. The tiles are installed in multipurpose space, used as a conference room, research lab, social area, cinema, dining area, etc., along two glass walls which meet at a corner. There are currently 25 tiles with a range of performance characteristics, due to the non-uniformity across each batch of hardware components. Wireless performance, microphone level, noise level, speaker level are all varied. Figure 5.2 shows a typical layout of tiles in a roughly aligned wall grid. Each tile is capable of an acoustic sensing and a broadcast range of several meters at standard volume levels.

The face of each tile is predominantly a screen, so that by tessellating the rectangular units, a display surface is constructed. A small structural element along the short side each tile supports the sensors in place, exposing them to the room.

The processor, screen power inverter electronics and wiring are enclosed in a case behind the screen. The tiles are  $405 \times 270 \times 40$  mm, and are designed to slot in and out of a support grid. They can also be hung from a structure or on the wall like a framed painting, or stand freely on a surface. The tiles will tend to be static in a space during use, but are easily repositioned or reconfigured for current needs. The tiles are limited in processing performance for intensive applications and by bandwidth for communication. The devices are plugged in to power, but smaller screen, battery operated devices match the power constraints of handheld systems. As we will see, levels of accuracy of synchronisation in these tiles, similar devices and network architectures play a significant part in sound sensing.

### 5.3 HETEROGENEOUS DEVICE NETWORK

The SAS tiles are similar, in resources, to many handheld consumer electronics devices. Given common communication protocols, or methods of routing between them, a single network of sensors can be built by combining the resources of both architectural surfaces and personal devices. These devices are carried into prime locations in a room — where the people are. They are battery powered, but are carefully maintained and charged, and constantly carried because of their proven primary uses.

Mobile phones have been a primary driving force in placing resource rich devices into the pockets of large populations of most wealthier countries on the planet. Low cost of deployment is also spreading mobile phones into regions where wired telephony and computer use have been impractical or unprofitable. As the development of these devices begins to become widely accessible to third-party companies and the public, the resources of the primary telephone function, can become tools for secondary unexpected uses. These include screen, microphone, speaker, camera, GPS, memory, processing, mobile network wireless, keypad, camera flash, FM radio, infra red or bluetooth. The code is moving to open systems, the Symbian programming environment, for example, offers both Java and C++ APIs. Both long and short range communication are well established, GSM<sup>†</sup> and similar standards for communication for example provides consistent connection for Europe and much of America and the world, while the Bluetooth short range communication protocol supports room range discovery and connection between many devices.

Global System for  
Mobile  
TeleCommunications  
protocol

Table 5.1 lists the properties of some of the currently most popular personal devices. We can expect many of these functions to merge in future devices as the hardware becomes cheaper still, and the chips more generic, while new niche use devices are also likely to emerge. The capabilities of PDAs (Personal Digital Assis-

TABLE 5.1. Resources in current consumer electronics devices.

	Mic	Speaker	Communication	Processing	Memory
PDA	yes	often	often bluetooth/IR and wifi	capable	large
Laptop	often	yes	some bluetooth, generally wifi	large	v. large
Camera	often	often	generally wired only for now	small specific	large
Musicplayer	maybe	just in ear	not often	small specific	large
Game	maybe	often	bluetooth or IR, maybe wifi	small	medium
Phome	yes	yes	long range and bluetooth/IR	small	small

tants), for example, are converging with those of mobile phones, and are the closest in resources to the SAS tiles. As with the SAS tiles, continuity in performance between microphones, speakers and other hardware resources of identical and disparate devices can vary significantly. Audio frequency range sensitivity variation, for example, is inherent in the low-cost manufacture process. It is not economically viable or necessary for manufacturers to maintain exact sensitivity correspondence between devices, especially as they tend to be used in isolation.



## CHAPTER 5

# COLLABORATIVE COLOCALISATION

Many sensing tasks are based on the assumption that the position of each of the sensors is known. A group of sensor nodes must also form and self organise based on knowledge of proximity. Determining connection and location is an inherent problem in any sensor network in which the topology is not precisely fixed or pre-calibrated. As described in this chapter, self-calibration steps are further complicated by challenges to ranging in a system with unspecialised or varied hardware nodes, and the constraints these imply.

Justification for using sound for ranging was presented in the first chapter of this document. Audio is easy to produce and sense, with microphones and speakers already prevalent in many devices, and low cost associated with adding it to new technology. Acoustic sensing is predominantly human-scale and activity-centric. It has been shown that the linearity and steady speed of sound can be used to infer direction and position of sources, as can the relation of volume to distance and spatial filtering features. These properties of sound can be used in a variety of ways to establish device topology and a coordinate system between nodes.

In this chapter a colocalisation approach is formulated, in the context of a number of possible techniques, and the corresponding assumptions and properties they imply. Drawing on previous research in these areas, the key challenges of measuring distances between devices and then establishing a common coordinate system across an asynchronous network of sensors are detailed. A practical implementation of this method is described and assessed in the Smart Architectural Surface architecture and a mixed device system.

### 5.1 ASSUMPTIONS AND APPROACHES TO LOCALISATION

A taxonomy of possible approaches to ranging and building connected graphs, and the assumptions and network infrastructure required for each, is presented

here. Given the linear progression of sound at a constant speed through air as a reference by which distances can be determined, a number of possible transmitter and receiver combinations and timing strategies exist in order to establish ranging measurements. A brief account of the scope of approaches is given here accompanied by some infrastructure requirements. A overview of ranging performance including non-acoustic mechanisms is given by [22] and the references therein.

Systems can be classified according to whether the nodes to be located contain the means to emit sound, or the means to detect it, or both. The sensor nodes can be either active or passive; either actively actuating the environment in order to aid ranging, or simply monitoring locally occurring sounds in an attempt to infer position. Active ranging predicated a transmitter sensor pair at each node. Many distributed sensor networks make use of dedicated beacons, which act as reference emitters. Beacons can be fixed, attached to the corners of a room for example, or dynamic, and can be at known or deterministic positions or in unknown locations. Beacons can also be unintentional — a system of sensors can be imagined that establish an estimated conformation by comparing timing measurements from a number of unknown, arbitrary acoustic events. Ranges can be determined from a number of sensors to a reference beacon, and used to establish a common co-ordinate system using lateration [6]. Taylor *et al.* [55] perform Simultaneous Localisation And Tracking (SLAT) with a single moving reference node.

The number of components in a network can also be important. Ranges to four reference node points are needed for triangulation in three dimensional space in order to solve reflection ambiguity. Further nodes provide redundancy that can aid synchronisation or accuracy. Some colatteration techniques that rely on a two-dimensional sensor geometry can succeed with a single beacon, while much larger numbers of beacons can corroborate and strengthen estimates of position over time.

The level of collaboration is another dimension in which systems can vary, and is dependent on resources for communication and considerations such as privacy. Information, such as position of beacons or timing, can be passed in the ranging signal itself, or precalculated, and can permit nodes to localise independently. Alternatively, nodes can communicate with beacons to establish this information, and this can often be openly available or can be broadcast to maintain privacy of node existence. With global communication, it is also possible for sensors to be fixed and nodes to merely act as emitters, being positioned by collaborative sensors which then transmit this information back to the nodes. In sensor node cases, passing measurements between elements in a sensor network is the basis for establishing relative timing, and common coordinates.

Accuracy with which timing or spatial positions of known beacons or sensors

can be determined plays a central role in the estimation algorithms that must be employed. If the timing is accurately known between sensors, or the signals collected can be passed with negligible latency, then timing measurements can be compared to a common reference zero. Equally, knowing or being able to determine the position of transmitters or receivers can greatly simplify localisation equations.

The positioning detail a system seeks to establish spans from fine-grained centimetre scale resolution to coarse room presence detection. The resolution of ranging measurements also plays a part in the infrastructure requirements. Given Boolean values for audible connection to neighbours, for example, a coordinate system can be derived in dense networks which uses the resulting graph and vertex information. Fine grained ranging, on the other hand, can be used to establish relative positions of nodes, subject to rigidity constraints, in even sparse networks.

While this research is focused on the room-range and audible frequencies of human-scale sensing, audio can include infrasonic and seismic frequency pressure waves, and ultrasonic, high frequency signals. Low frequencies can travel great distances and pass around features smaller than the wavelength, while high frequency emitters tend to be highly collimated beams, that are easily reflected and occluded. High frequencies inherently carry greater phase timing information, and broadband signals can encode sharp onset signals.

Systems with certain fixed geometries can build knowledge of their constraints into the algorithm. Assuming nodes are bound to a line, surface, limited volume or fixed geometry, for example, can greatly simplify the calculations. Near and far field cases can also affect the assumptions that can be made. For example, the hyperboloid functions defined by difference of arrival time measurements can be approximated as a cone if the distance from the microphone pair is much greater than the distance between them.

A single node can emit and detect probe signals and measure round trip time to establish echo location. Active echo ranging, which enables a device to map the topology of the surfaces around it, but requires a collimated beam and rotational degrees of freedom to scan through. Mapping requires subsequent comparison with prior maps of the space. Time of flight between two nodes can be established with ping and response measurements. As we will see, this requires synchronisation between nodes or relative timing and self-monitoring. Global or relative coordinate systems can be sought. The former rely on known beacons, or prior models or maps of a space to fix node estimates and the latter can be determined up to global rotation and translation ambiguity.

Several acoustic ranging mechanisms remain unexplored here, such as the use of spectral filtering or reverberation information as a means of estimating distance

by modelling the multipath and frequency dependent effects. Corroboration, by many fixed echo sensors, of moving objects passing through their beams, tied to assumptions of steady or only slowly varying object speeds may also provide a means of positioning.

### *Group Finding*

Prior to fine resolution acoustic ranging, nodes must establish which of the neighbours in a communication group are in direct line of sound and within detection range, establishing a graph of connectivity through coarse audibility tests. Clustering based on wireless communication radius may provide an initial approximation of possible neighbours (such as common connection to a single wireless router). In cases of long range communication infrastructure, such as GPRS<sup>†</sup>, in which wireless proximity provides little guarantee of sound sensitivity range, a common point of interest may provide clues to which nodes share a common acoustic environment. Test signals can be actively generated to determine nodes that are within range of detecting them. As no timing information need be carried in this group forming and graphing step, it has been shown [4] that slowly ramped near-ultrasonic reference tones can clearly establish acoustically collocated nodes without producing audible clicks.

General Packet Radio Service (GPRS) is a mobile data service available to users of GSM (Global System for Mobile Communications) mobile phones.

## 6.2 PAIR-WISE RANGING AND RELATIVE LOCALISATION

Measuring the flight time of sound as a method of mapping and ranging has matured in the fields of location and imaging using echo return time, such as medical ultrasound and marine sonar, and position estimate and tracking of unknown sound sources, such as passive sonar and ballistics location.

The task of ranging is simplified for cases in which signals can be actively emitted and detected at both ends of a distance to be measured. In pair-wise ranging between active nodes, sounds can act as synchronisation and time of flight delay references. Sound signals can be selected with properties that aid detection and passed to the receiver before the sound is emitted. The characteristics that improve arrival detection and timing accuracy are described below. Sensor equipped devices must then establish some common time frame or comparison model in order to calculate time of flight between emission and detection.



## *Signal Detection*

Finding the arrival time of a sound at a sensor relies on accurately identifying when the emitted waveform appears in the recorded signal, against a background of other sounds arriving at a microphone, as well as noise and information loss inherent in both the detection and digitisation processes. The simplest detection method is to use sudden changes in sound level, such as the sharp onset of a loud click generated at the emitter, as a timing reference point. A sharp rise in sound can be easily timed with almost no processing by finding the first sound sample to rise above a threshold level. However, with no further spectral or timing information in the chirp, it can become lost among other loud sounds or background noise. Longer chirps containing specific spectral characteristics can be used to improve detection and reject noise. A matched filter can be run across the recorded waveform to find the closest correlation to the expected signal. This can be hindered by distortion and change in the properties of the waveform as it passes through a space from an emitter to a detector. Filtering of certain frequencies can occur, and reverberation can lead to a number of delayed copies of the signal mixed with the first.

By broadening the transmission frequency band to be much greater than the narrow bandwidth needed for the data improves the detection accuracy and performance. A number of chirps fulfil the criteria for good correlation properties. A signal such as a frequency sweep or white noise are suitably broadband and distinct. If the reference signal used is deterministic, then it can be generated at both the transmitter and the receiver, rather than passed in full over the network. Such signals benefit from processing gain, achieving increased signal detection performance at the cost of greater computational complexity. A sweep frequency chirp, that linearly or logarithmically increases in frequency between two values, for example, can be detected with the corresponding matched filter even if some of the frequency components are missing. The sharpness (kurtosis) of the auto-correlation peak is dependent on the extent of the highest frequency components transmitted. Sarwate[47] outlines the desired properties of reference signals:

"Each signal in the set is easy to distinguish from a time-shifted version of itself and each signal in the set is easy to distinguish from (a possibly time-shifted version of) every other signal in the set."

Maximal length sequences are a family of binary sequences characterised by a sharp peaked auto-correlation function. They are pseudo-random bit sequences, with white noise-like qualities, which can be easily generated using a linear feedback shift register. Taps (which specify the necessary feedback pattern) are selected such that every possible state is produced by the register, hence 'maximal length' or 'm'-sequences. For an  $r$ -stage register these sequences are  $2^r - 1$  in length, which can

be shown is the maximum period. The number of sequences for a given power value varies. These sequences are balanced, containing  $2^{(r-1)}$  ones and  $2^{(r-1)} - 1$  zeros. The bitwise sum of an m-sequence and a shifted version of itself produces a third phase of the same sequence. If each zero is mapped to a  $-1$  and each one to  $+1$  then the resulting autocorrelation function is unity for zero delay and falls off as a triangular peak to a constant value of  $-1/2^2 - 1$  for more than one sample offset. In an environment with a number of such chirps occurring in an unknown order, a further desired property is that each signal be as different from the others as possible, that the cross-correlation between signals be low. The cross-correlation between m-sequences is bounded. As discussed in [47], if it is possible to use only a few sequences, a 'maximal connected' subset of sequences offer improved bounds on cross-correlation properties. In a coarsely synchronised system, a small number of sequences could be looped with little ambiguity, and so a maximal connected set of m-sequences could be used. These signals are used considerably to drive spread spectrum radio communication because of their robustness to noise and jamming and the protection from interception they can offer. Given that the signals are wideband they are less likely to be lost due to the scattering, and filtering in air, of certain frequency bands.

### *Clock accuracy*

To calculate the time of flight of a single acoustic chirp emitted from one device and detected at another, the pair of nodes must maintain a common time frame so that transmission and arrival times can be meaningfully compared. High accuracy clocks are currently an expensive option. Typically processors use quartz clocks which are accurate to about 1 second per day (i.e. 1 part in 86,400) due to crystal oscillation properties and temperature dependence. This is possible to correct, in network connected devices, by adjusting in comparison to a reference clock server. Compensating for drift using a given communication network, is then dependent on the timing accuracy of the connection. Variable latency is present in most standard communication methods. Delays can occur in the MAC<sup>†</sup> layer, network card, network driver and context switching of application process, among others. Through repeat measurements it may be possible to estimate the unknown timing jitter in the communication between devices in order to better synchronise them. Network Timing Protocol[39] has become the internet standard for synchronisation and timing correction. It has proven effective in both small and large networks; using jitter estimation, redundancy and averaging to set and maintain synchronisation with a reference server. Where a processor kernel provides support for precision timing signals, such as a pulse-per-second (PPS) signal, accuracy

Media Access Control (MAC) is the low layer interface between a node's Logical Link Control and the network's physical layer.

can be achieved down to one nanosecond uncertainty. However, in typical usage today, "NTP provides accuracies generally in the range of a millisecond or two in LANs<sup>†</sup> and up to a few tens of milliseconds in global WANs<sup>†</sup>". At the speed of sound, this corresponds to an uncertainty of between half a meter to several meters. NTP is useful for coarse synchronisation, but currently not accurate enough for precise acoustic timing measurements.

Local Area Network  
(LAN)

Wide Area Network  
(WAN)

In some device architectures it may be possible to bypass communication protocol timing overhead and directly use a reference Radio Frequency (RF) packet to recalibrate clock times with each acoustic broadcast. This requires determining the precise emission and arrival times of the electromagnetic pulse. The flight time at the speed of light can be considered instantaneous over the ranges for which audio can be detected. In an acoustic sensor network absolute time may not be needed, only relative synchronisation between every device. In this case Reference Broadcast Synchronisation (RBS) can be used. In RBS, one node emits a reference RF pulse, and all other nodes use the arrival time as a common zeroing point. There may still be delay in the detection and flagging of the arriving pulse, but the time the pulse was transmitted is no longer important, and so any buffering or delay in transmission is not of concern. In a system designed by Estrin and Girod[18], for example, an interrupt triggered through a modified network card is used to tag arriving packets with local time. Comparing the difference in arrival times at each node establishes a common time frame. Reference broadcasts can be emitted synchronously with the sound pulse or energy conservation can be improved by shutting down all resources apart from the low power microphone, and then, once a chirp has been received, using *post-facto* synchronisation.

Reference Broadcast Synchronisation is a popular choice for low resource systems, in which memory for storing sound and processing is limited, but access to low-level hardware timing allows exact arrival and emission times to be measured. Direct interrupts from changes in hardware state can be time coded to improve accuracy. Examples of self-localisation systems using dedicated *ad-hoc* wireless sensor network hardware layers such as the Mica motes or Crickets have demonstrated 1 – 2 *cm* accuracy with this approach, and robust performance in real-room[40], urban[52] and natural[57] environments. These systems use either audible or ultrasound pulses.

Further timing uncertainty occurs in the emission and detection of acoustic chirps. Sound card latency, audio buffering and code execution delay, and to a lesser extent microphone response time and speaker activation time can all contribute. It is clear that precise synchronisation can be achieved, but requires low-level hardware access to time stamp both the acoustic and electromagnetic incoming signals

in order to overcome variation in delays due to buffering in the communications layer. If the device is not dedicated to timing measurements, but instead communication, the raw timing of pulses may be abstracted by several communication layers, each of which may contain inherent delays. So while this solution is convenient in specialised hardware and operating systems, it requires a solution specific to each processor, and so does not migrate easily from one platform to the next.

Further improvements to the challenge of maintaining synchronisation are being made, driven by the need for fine grained timing agreement for many distributed sensing applications. Delay jitter can be modelled and estimated, or included as a nuisance parameter in a system of equations describing the measured results. This global calibration can yield improved synchronisation at the cost of increased complexity in modelling and calculation. It is clear that synchronisation will continue to be improved in terms of reliability, portability between operating systems and efficiency in power and bandwidth consumption, as it is a core resource for many tasks.

### *Asynchronous Time of Flight Ranging*

Can localisation through audio timing measurements be achieved on a network lacking fine grained synchronisation? Chapter 4 predicted a growing prevalence of sensor networks formed from loosely connected, impromptu, heterogeneous devices connected across several communication protocols. In such an infrastructure, where audio and electromagnetic resources may have other primary functions, hardware solutions are impractical, and, further, uniform adoption of a wired synchronisation protocol would be unlikely. The benefits of not having to establish precise synchronisation are clear. The communication mechanism between devices may also not support direct reference broadcast pulses. An example is the GPRS data connection in mobile phone communication, which must pass data through a centralised system, even if both phones are in the same room.

The key to ranging between nodes in an unsynchronised network lies in the symmetry of emission and detection for each device pair. The assumption is made that the distance between the microphone and the speaker in each device can be considered negligible compared to the distance between devices. Taking long recordings of a chirp emitted from one node, and then, some unknown time later, a second returning chirp, provides a direct comparison of the actual emission and arrival times of the two sounds at each of the devices. The time of flight is equal in both directions. The time delay is calculated by subtracting the second arrival time from the first. As Figure 6.1 shows, for two nodes *a* and *b*, the time delay between peaks

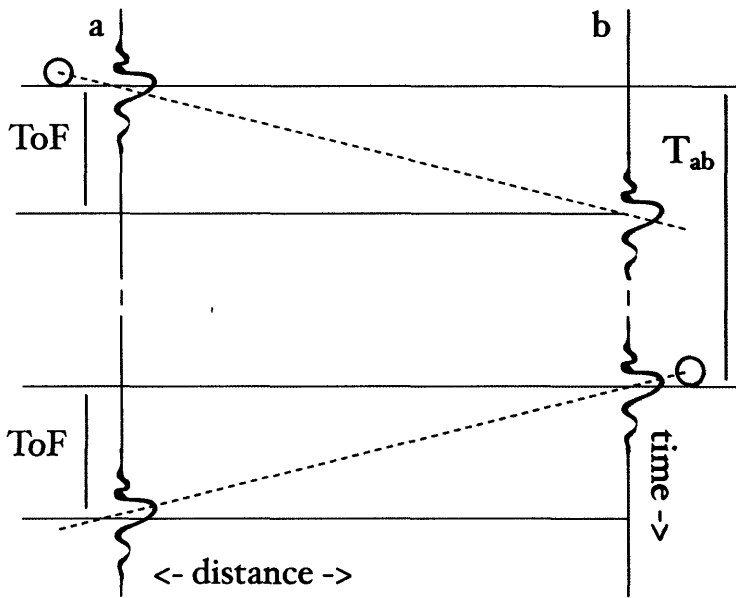


FIGURE 6.1. The sounds recorded at two nodes,  $a$  and  $b$ , are represented as progressing with time vertically downwards, with distance between the times represented horizontally.

measured at the device that chirps first (labelled  $a$ ) is the arbitrary time  $T_{ab}$  between when the two chirps actually occurred, plus the time of flight,  $T_{oF}$ , between  $a$  and  $b$ :

$$A = T_{ab} + T_{oF}.$$

While the delay recorded at  $b$  is  $T_{ab}$  minus the  $T_{oF}$  from  $a$  to  $b$ :

$$B = T_{ab} - T_{oF}.$$

Clearly, calculating the difference between the peak-to-peak delays measured at  $a$  and  $b$  therefore yields twice the time of flight,

$$A - B = T_{ab} - T_{ab} + 2T_{oF}$$

, thus essentially subtracting out the arbitrary time between the two auditory events from the calculation without any knowledge of the absolute synchronisation of the two recordings. Given the signal detection times at  $a$ ,  $t_{A1}$  and  $t_{A2}$ , and at  $b$ ,  $t_{B1}$

and  $t_{B2}$  the time of flight,  $T_{oF}$  can then be found from:

$$T_{oF} = \frac{(t_{A2} - t_{A1}) - (t_{B2} - t_{B1})}{2},$$

at each of the nodes. The separation distance is then given by:

$$\delta_{AB} = T_{oF} \cdot \nu$$

where  $\nu$  is the speed of sound in air.

The algorithm described uses relative timing between the sounds emitted by a pair of devices. This information can be collected in a single recording of both sounds arriving at a microphone, but does not necessitate comparison to an internal clock or local reference measurement. This approach supports the possibility of asynchronous communication. Recordings can be made at a physical location, but then streamed and processed elsewhere or stored for later.

### *Acoustic Ranging Error*

Gaussian error due to uncorrelated noise sources arises from a number of factors in signal detection including external background sounds, air noise on the microphone, thermal noise in the detector and nondeterministic jitter in emission and detection of sounds. The latter is all but removed by taking relative timing measurements between two arriving signals in a recording. The other noise sources act to mask and degrade the chirp signal, but are largely overcome by the processing gain of the maximal-length sequence. In cases of low signal to noise ratio (SNR) the sequences tend to be detected accurately, or not detected at all. False detection can be handled by outlier rejection[42] and through iterative estimates. Quantisation error due to discrete time steps of sampling result in linearly random error of up to one sample length, in the position of the correlation peak. The total error in the range measurement depends on the uncertainty in the position of each of the four peaks detected. Any drift between the clocks at the two nodes during the measurement time can also introduce error, although drift at this time scale is minimal.

Diffraction of sound can lead to excess path lengths on individual range measurements. These will tend to be overcome in comparison to other range measurements between nodes in the determination of a coordinate system. The occlusion of a direct sound path and subsequent detection of a strongly reflected path can lead to unbounded error up to the furthest detectable sound range (errors up to several meters) and consistent over multiple measurements. Comparison with other,

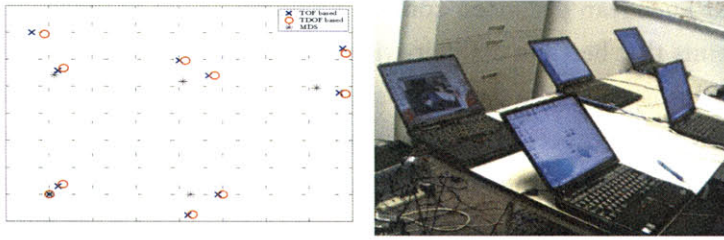


FIGURE 6.2. Raykar *et al.*, use this asynchronous ranging and a closed form estimator, that is then improved by solving for the delay jitters[46].

non-occluded ranges may identify these spurious results in some cases. In [18] the authors suggest a number of alternative sensing mechanisms for detecting these cases. However, occlusion remains a serious complication for acoustic ranging.

The assumption that the microphone and speaker are collocated introduces a consistent offset error up to the actual separation length of the two, depending on the orientation of the device. For the SAS tiles, this distance is 4 *cm*, in handheld devices it is typically about 10 *cm*. The work of Raykar *et al.*[46] on positioning calibration shares goals and constraints close to those described here. Their work establishes separate coordinates of both the microphones and speakers for a collection of heterogeneous devices, which they term General Purpose Computers (GPCs). They introduce the technique used for inferring time of flight used here; closed-form position estimation of sensor-emitter pairs, using it as a close initial guess in an iterative nonlinear least squares minimisation of position error functions, accounting for the synchronisation offset in each measurement. The method is demonstrated in a 5 laptop, two-dimensional case 6.2. The current work does not apply this optimisation step, assuming that the separation of the microphone and speaker at each device, and the uncertainty introduced by this, is small. Sugar and Kloper[54] describe the use of this closed form localisation technique in an unsynchronised wireless radio environment. The comparative accuracy of using electromagnetic and acoustic signals depends on the ranging distance and the timing detection accuracy of the devices. As argued in Chapter 3, the slow speed of sound makes results in accurate measurements on generic, low accuracy equipment, and ideally suited to room range localisation.

### *Solving Geometric Conformation*

For the task of estimating coordinates of points given inter-point distances, Classic Multidimensional Scaling describes a family of algorithms for mapping a set

of estimated Euclidean distances, connecting points, to an optimal conformation of those points that best fits the pair-wise measurements. A conceptually simple example of this is to build a particle system spring model of the nodes, assigning the measured distances between two points to the desired length of a corresponding spring. Relaxing to local minima can result in erroneous results. By tuning drag and elasticity the nodes can relax to an optimal coordinate system given noisy distances between positions in Euclidean space. Simulated annealing, random perturbations applied to the system, can be used to overcome local minima. The scaling is a minimisation problem, and least squares or gradient decent methods can be used. Folding ambiguity in non-rigid systems can result in solutions which are locally correct but globally flawed. Rigidity is defined by the number of edges at each node in the graph. For systems containing large numbers of devices, a distributed approach to MDS is desired, rather than passing all range measurements to a single elected processor. Moore *et al.*[40] propose a method that establishes ranges within local clusters growing this into a complete solution.

### 6.3 COLOCATION METHOD

The approach presented here is intended for devices with sufficient computational and storage resources to record and process several seconds of sound. Full duplex audio is required for simultaneous input and output of sound, allowing a device to record the sounds it emits. Some form of communication protocol is necessary; 802.11\* WiFi, short range Bluetooth, mobile telephone protocols such as GSM (Global System for Mobile Communications), or even infrared are all suitable candidates, as are combinations of these with multi-modal devices to act as intermediaries. It is even conceivable to use an entirely audio based system although the rates possible for transfer of data between devices is inefficient[36].

Initial device discovery and group forming steps are necessary to establish the available resources in an environment and give a count of the nodes, assigning an arbitrary sequential ordering to each of the devices. It is also possible to use other unique information (for example least significant digits of IP address in a local subnet) to establish chirp order.

Unique reference signals for each of the nodes are generated by calculating pseudo-random sequences. These are phase-shift encoded onto a carrier frequency and emitted as noise-like chirps. Maximal-length sequences are used because of high auto-correlation and bounded cross-correlation properties[47]. Using 511 bit length sequences yields 48 unique identifiers, which are assigned sequentially to the nodes.



A synchronisation technique implemented over the available communication protocol is then used to establish initial coarse timing agreement, allowing each device to chirp in order without overlap. In the experiments conducted here Network Timing Protocol (NTP) is used. It is desirable to have as accurate synchronisation as possible, to reduce the overall recording buffer length. However discrepancies in synchronisation can be tolerated with sufficient recording memory. The limiting assumption, however, is that any drift in synchronisation between the clocks in each device, over the length of the recording, must be negligible. The chirp sequences are uniquely identifiable and can therefore occur in any order, further reducing the need for exact synchronisation.

A single recording of all the sequentially arriving chirps is made on each device to remove the significance of any audio circuitry or API buffer latency that can otherwise cause an unknown delay between the arrival and detection of a sound. By correlating the recordings against each of the known pseudo-random sequences, strong peaks are identified at the arrival of each of the sounds. Correlation is performed using a Fast Fourier Transform (FFT) technique. Each device first finds the chirp it emitted, this significantly louder part of the audio is then suppressed before searching for the other chirps, as the volume at such proximity to the speaker otherwise overwhelms the other peaks in the correlation measurements. Following the approach of Girod *et al.*[18], a running threshold, found by correlating the recording with a noise-like sequence first, suppresses false peaks, despite varying background noise levels, and so improves robustness. The most direct path for an arriving sound is found, in most cases, as the first peak and is generally detected, even in reverberant and noisy real-room conditions.

The remaining peaks, having been precomputed from the corresponding maximal-length sequence, are found in the order they were meant to emit. The time delay, in samples, is measured between the chirp emitted at a node and each of the remaining peaks detected in the recording. The number of samples between the two chirps of a node pair, the time between their arrivals, are exchanged between devices over the communication layer (in this case WiFi). From the difference between these two measured delays, the time of flight can be found, and using an estimate of the speed of sound in air, a separation distance can be calculated. Procedure 6.1 describes the generation of reference chirps, recording of sounds emitted from each neighbour, detection of peaks to calculate delays between sounds and finally a multidimensional scaling step to establish relative conformation.

#### PROCEDURE 6.1: COLOCALISATION CODE AT EACH DEVICE

- 1: generate a unique chirp for each node in cluster. chirps 1 to  $n$
- 2: plus one more to act as running threshold. chirp 0

```

3: create duplex audio buffer
4: wait until an agreed time, based on NTP synchronisation
5: start audio stream
6: while there is space in the record array do
7:   store data from buffer in record array
8:   if current time is  $myID * chirpgap$  then
9:     fill buffer with chirp
10:  else
11:    fill buffer with zeros
12:  end if
13: end while
14: stop and close audio stream
15: correlate control chirp with recording to find running threshold
16: correlate own chirp with recording to find own peak
17: mute this section of the recording
18: for each of the remaining chirps do
19:   correlate to find arrival time
20: end for
21: pass own and neighbours' arrival times between neighbours to compute ranges
22: pass pairwise ranges to an elected processor
23: run classic multidimensional scaling to estimate coordinates of each node {up
    to global rotation and translation}

```

A cross-platform audio API (application programming interface) was chosen to maintain as great a separation of the algorithm and code from the specific device hardware interface and requirements of any specific platform. RtAudio is a C++ API supporting distinct features of OSS and ALSA with Jack on Linux, CoreAudio on Apple, and DirectSound on the Windows operating system. By extending this API to other platforms used on new devices, the code should then remain portable. Duplex audio is supported in RtAudio and available on most modern sound cards. The C++ programming language is suitable for high performance real-time audio code, is largely portable between platforms and has few complex requirements which makes it suitable for constrained devices. Both blocking and callback access to the incoming and outgoing audio buffers are offered by RtAudio and can be used by this algorithm. Callback is suited to cases in which the code must run concurrently with other processes (such as updating screen content). NTP is run independently as a background process. The program then uses an absolute timing cue to start recording and chirp time on each device. Enough delay is left between chirps to account for clock error and room reverberation (reverberation time in a large room is  $\simeq 0.73$  s).

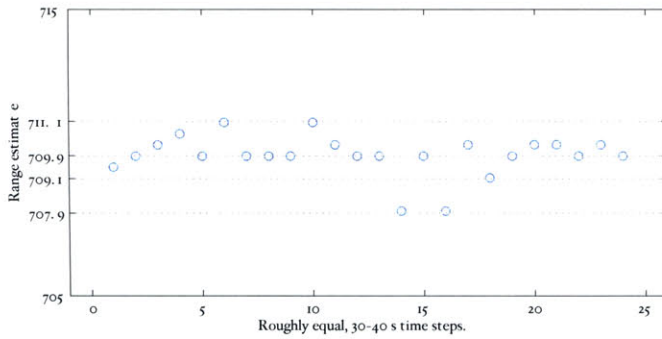


FIGURE 6.3. Ranging uncertainty in measurements at a 7.1m device pair separation.

Ranging accuracy using the SAS tiles was examined using repeated pair-wise measurements. A series of 24 chirp arrival times were taken over  $\sim 20$  minutes in a large meeting room space from two tiles facing each other at a separation of 7.1 m. In three measurements peaks were not found and were automatically rejected. The remaining peaks are plotted, against time in Figure 6.3. The pseudo-random sequences are phase shift encoded onto a 11025 Hz bit rate carrier wave, and are emitted and detected at 44100 samples per second. One sample corresponds to a distance of 8 mm. The variance of these measurements is 3 cm, which corresponds to  $\pm 2$  sample steps. The data seem to display two distinct additive error modes, one is a small, one or two sample size uncertainty, which is thought to arise from error in detecting the peaks. An uncertainty of  $\pm 1$  sample in each peak would cause an error in the calculated time of flight of  $\pm 2$  samples in steps of half a sample. Slight atmospheric changes between the two chirp times, such as air movement, could also contribute to this. The results also oscillate between strata separated by a unity sample step size. Variations in the time of flight, for example due to slight temperature change between each ranging which cause the time of flight length to cross a sampling threshold, are exaggerated by a factor of two resulting in a step of one sample. A separate set of ranging measurements taken over a longer time period demonstrate a slow, long term drift in the values which reflect the variation in room temperature over that length of time due to air conditioning cooling, after the door was closed. Taking a series of measurements at regular steps over several meters demonstrates the inherent linearity of this approach. In Figure 6.4 the results for ranging measurements, incrementally stepped, up to the length of a room are given. In this case, the two devices were moved apart in the plane perpendicular to the direction the microphone and speaker are mounted. The background sound level in the room was measured giving a signal to noise ratio of 16 dB. The ranging has an uncertainty of 3.5 cm. A small offset from zero is due in part to estimation

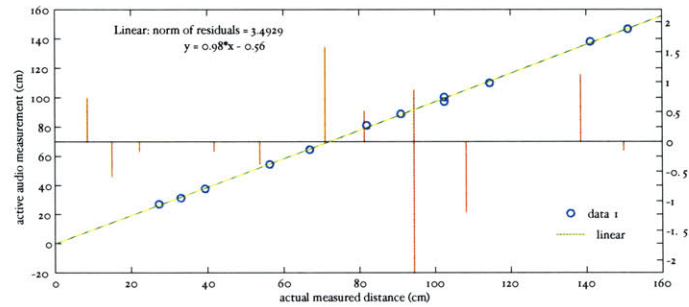


FIGURE 6.4. Ranging accuracy over the length of a room, with an uncertainty of 3.5cm

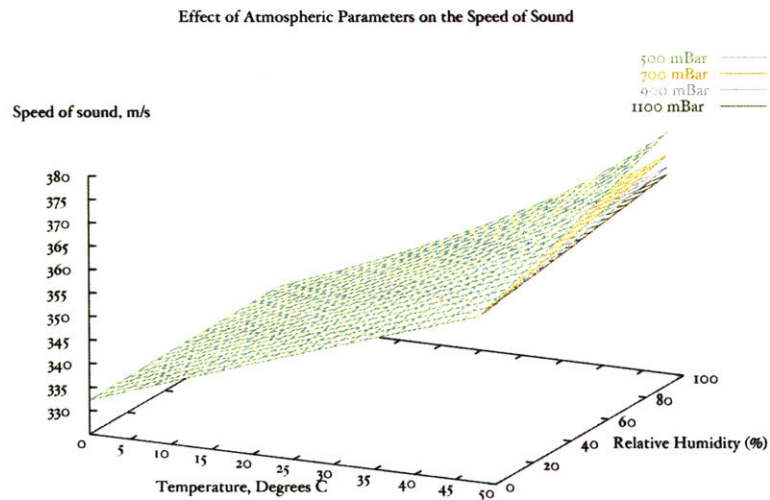


FIGURE 6.5. Variance in speed of sound with temperature, pressure and humidity (after [18]).

error in the speed of sound, which, as Figure 6.5 shows, can vary by 10% indoors due to temperature, humidity and pressure changes. A further offset is introduced by the nonzero distance between the microphone and speaker in each device. The calculated range is shortened from the centre to the edge of an ellipsoid bounded along the major axis by the microphone and speaker, with a minor axes half as long. A minor offset could also be caused by clock drift of one device relative to the other, but drift is negligible on the order of chirp timing. The peak finding performs well despite a stronger reflection signal than the line of sight signal, from a flat surface parallel to the plane in which the tiles lie.



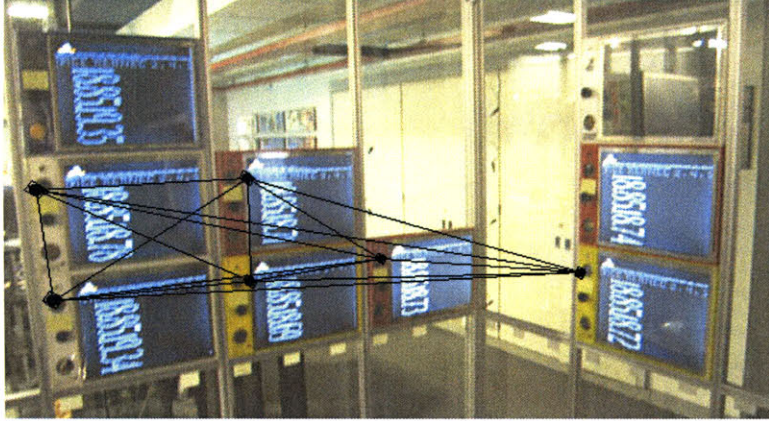


FIGURE 6.6. The node arrangement derived from range measurements and subsequent multidimensional scaling, globally rotated and translated to fit back onto the tiles used

Using these SAS tiles, a 6 device system, occupying two walls, ran the self-localising algorithm to establish the position of each device relative to the others. The range measurements were calculated taking the speed of sound, in air, at sea level to equal  $340.29 \text{ ms}^{-1}$ . All measurements were passed to a single elected device to iteratively calculate likely relative positions. Figure 6.6 shows the experimental set-up overlaid with the sensor layout determined from the range measurements. Neither the grid layout nor planes of position were used as constraints. It was found that even the furthest nodes were detected despite the noise and reverberation levels in this real-room case, leading to a fully connected graph. Rotated and translated onto the actual positions, the average distance between actual and estimated positions was  $3.3 \text{ cm}$ . A multidimensional scaling approach is used to determine a likely three-dimensional conformation from the Euclidean range information between each pair of nodes in the graph. While this performs well for a small number of well connected nodes, as the device count and complexity of the graph increase, the minimisation of errors can settle on local minima and in non-rigid systems, folding ambiguity remains. Moore *et al.* propose a distributed approach to localising nodes from noisy range data[40], by building rigidly connected clusters. The algorithm avoids flip ambiguities and scales linearly with the size of the network.

### *Heterogeneous Devices*

Although the code has not been fully migrated to a number of small device programming platforms, it has been possible to demonstrate its effectiveness across a range of hardware. As proof of concept, and for brevity, the devices are used to

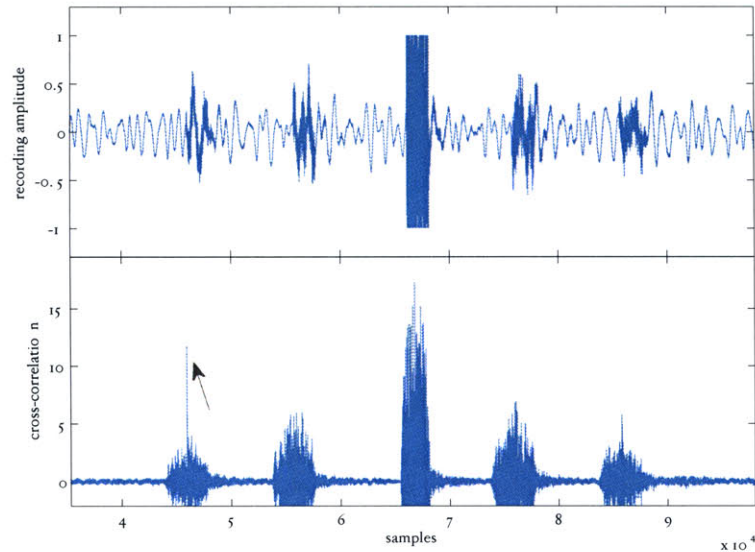


FIGURE 6.7. Peak finding, cross-correlation performance despite  $8\text{kHz}$  recording rate and subsequent linear interpolation upsampling to  $44.1\text{kHz}$ .

record the signals, but the  $44\text{kHz}$  chirp is generated and played from a co-located SAS speaker. A test set of devices consisted of an HP iPAQ running Familiar Linux, a Dell Axim running Windows Mobile and a Nokia Symbian Series phone, plus two SAS tiles. These devices span a range of microphone performance levels, sound circuit jitter and delays, and compression algorithms. Both the Axim and the Nokia provide basic recording programs that sample at a low  $8\text{kHz}$  rate. Figure 6.7 shows the performance of cross-correlation peak detection on the  $8\text{kHz}$  Axim recording, upsampled to  $44100\text{Hz}$ , correctly detecting the first chirp. The phone also uses the AMR, adaptive multi-rate, lossy speech compression file type. These factors reduce the success rate of the peak detection algorithm, and spread the correlation peak, however, peaks can still often be found over a reduced sensing range. Figure 6.8(a) shows the positions of the devices and the calculated estimates. Input compression degrades the detection of chirps, but seems to only contribute a small error to measurements in cases when peaks are found. Figure 6.8(b) shows a comparison of tile only and tile-phone ranging.

The performance of the maximal-length sequences at  $8\text{kHz}$  sampling rate, that the signals can be effectively detected even for heavy compression, demonstrates the broadband nature of the chirps. Sampling at  $8\text{kHz}$  is equivalent to filtering out the high frequency components. This has the effect of spreading the correlation peak, but the centre can still be obtained. The signal must still be emitted at  $44.1\text{kHz}$ , however, to maintain the sharp onset, and timing information.

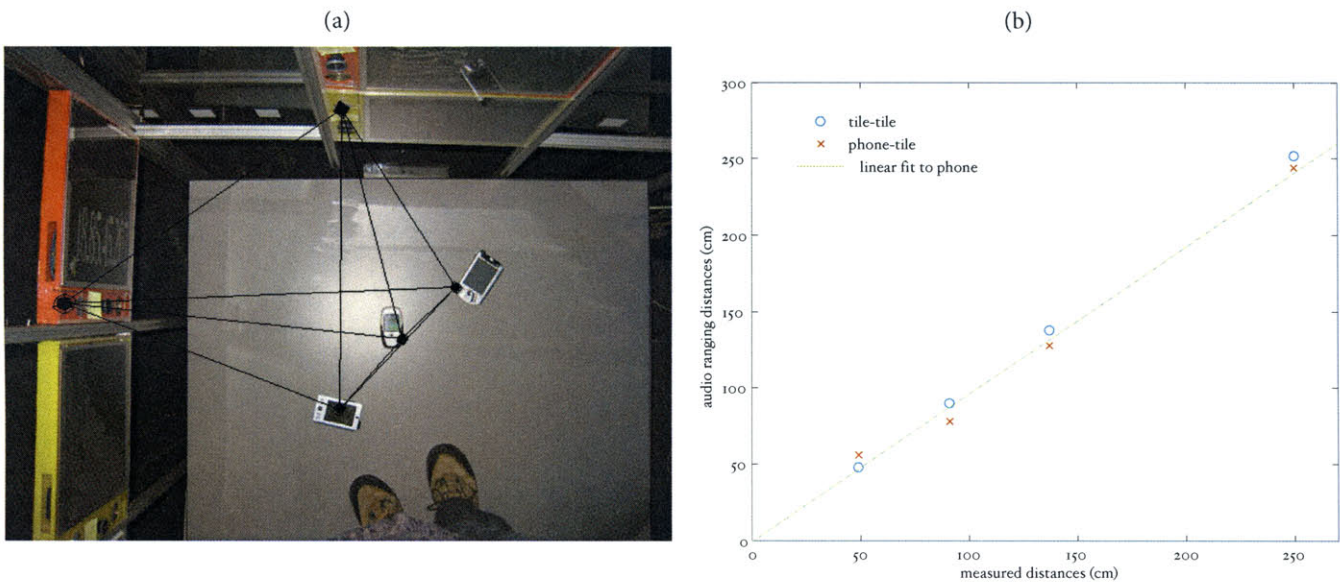


FIGURE 6.8. The three-dimensional coordinates estimated for a variety of devices, each used to record chirps and the ranging characteristics for a mobile phone - device pair. Detection is degraded for heavily compressed recordings, but with interpolation and broadband chirps, timing accuracy remains equivalent.





## CHAPTER 6

# SOURCE LOCALISATION

This research proposes a method of estimating the positions from which sounds have originated, using measurements across a number of asynchronous sensor devices. This necessitates measuring the delay between two consecutive sounds at each microphone and solving for the two source locations simultaneously.

Passive source localisation is the detection, and subsequent position estimation, of unpredicted acoustic events. Whereas collaborative colocation involves ranging between self-organising nodes by emitting chirps and comparing arrival times, source localisation seeks to discover the positions of any sound that can be identified arriving at a number of sensors. By comparing signal arrival times at the microphones of a network of devices at known positions a solution for the originating points of the sounds is sought. This challenge has been explored in a large body of literature, and it is closely related to the calculations of sonar, seismology, ballistics ranging and GPS. Section 2.2 described how source location is beginning to be applied in robotics, human-computer interaction and situated computing among

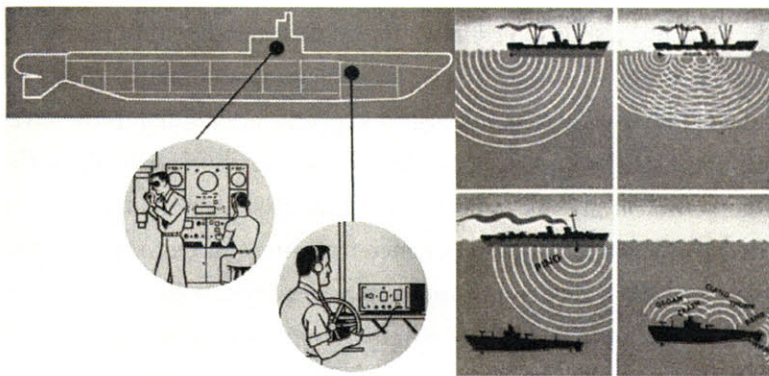


FIGURE 6.1. Passive and active source localisation training manual from WWII US submarine[41].

others. Given here is an overview of currently applied method of solving for source position for systems in which time synchronisation between microphones has been established. Changes to this approach are then proposed to build an algorithm which will work across a network of heterogeneous devices.

Passive localisation has seen considerable use and development in covert naval tracking, and in particular in submarine craft, in which passively detecting others without revealing position is key. Talking with a submarine sonar operator from the second world war, he described listening, in training, to large catalogues of engine, sea life and propellor noises. These enabled him to pick out enemy craft noises from the ocean acoustic background. Having identified a sound, an estimation of its position could be determined from its arrival times at each of a number of microphones. A linear array of hydrophones is often used, dragged on a line behind the submarine, with all the wires running back to the operator. The equal spacing in this linear array makes for a simplified calculation. The use of time difference of arrival was introduced in the context of beam forming in Chapter 2. We will look further at the positioning of unknown sound sources using collections of microphones in this chapter.

## 6.1 INFERRING POSITION FROM MULTIPLE RECORDINGS

An algorithm suited to a distributed sensor network is presented, drawing on the history of approaches to locating a radiative point source using stationary, passive sensors. Phased array source localisation has been investigated considerably in the fields of beamforming and radar[37], although the far-field assumptions, based on considerable distance between points, do not hold for room acoustics. Some sound based application examples include passive underwater acoustics[9], ballistics positioning[17] and wearable computing[3]. Generally the microphone configuration is known and a solution must be found for the range differences calculated from the arrival times of a sound wave at each sensor. Arrival time differences measured at a number of spatially distributed points are related to the array geometry and the originating sound source. For each microphone pair, the difference in path lengths from a single start point to each of the sensors is fixed. This is proportional to the difference in radial distance from the source to two nodes, termed Time Difference of Arrival (TDOA) and results in a hyperbolic constraint for possible position from each difference calculated. A hyperbolic curve is defined in a two dimensional case and a hyperboloid surface in three dimensions. A function can similarly be defined for each microphone pair. For noiseless measurements, the solution then lies at the intersection of these hyperboloids. For a system constrained

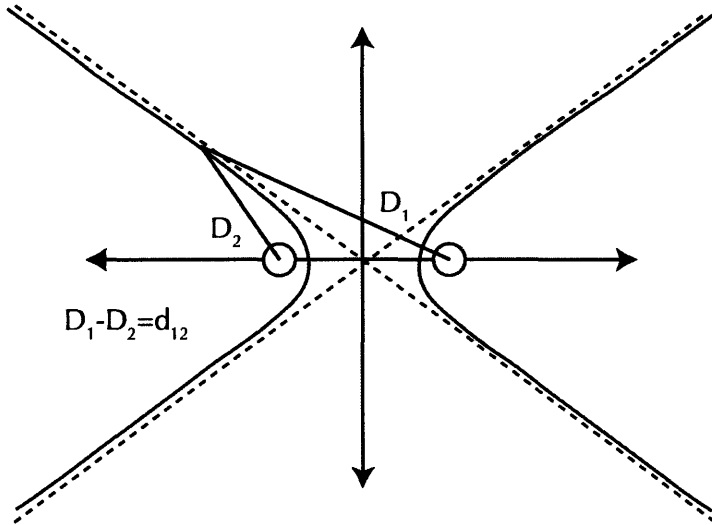


FIGURE 7.2. Plot of the hyperbolic function defined by difference between two range measurements (after Basu[3])

to a surface, three microphones are required, whereas for a system in free space four are needed. The challenge however, is that each of the arrival time measurements is corrupted by noise. The hyperbolic function can vary significantly with small changes, and so a guess must be obtained for the closest point of intersection of each of the curves. This can be solved, and becomes more certain as the number of nodes increases. Measurement uncertainty in practical applications have been addressed in a number of off-line and real-time techniques[27].

Finding a solution for the intersection of hyperboloids corrupted by noise is a nonlinear problem. An additive measurement error model can be used to describe how the data relates to each of the time delays and the uncertainty components. A number of techniques have been proposed for solving this system using maximum likelihood or least-squares approaches. The method employed depends on the constraints implied by the intended application. In particular, how the noise is modelled and whether the system is to run in real-time must be considered. Huang, Benesty and Elko reviewed the recent literature on use of the additive measurement error model, assessing criteria including linear approximation as opposed to direct numerical optimisation, and iterative or closed-form techniques.

### *Maximum Likelihood*

Use of a maximum likelihood estimator (MLE) has been historically popular, performing well in systems with larger numbers of microphones, and reaching a

stable solution (asymptotically consistent). It is constrained, however, by requiring assumptions of the distribution of the errors to be made. Often these uncertainties are assumed to be Gaussian, which has been justified by [20] for continuous-time signals, but does not take into account the non-Gaussian sampling error associated with digitising the signal. Coupled with the difficulty of convergence to the correct minima, without a good initial guess, using typical iterative root finding algorithms such as Newton-Raphson, Gauss-Newton or Least-Mean-Squares linear approximation approaches, has made MLE problematic, especially for real-time use.

### *Least Squares Estimator*

To support real-time or close to real-time processing, closed form algorithms are desired. Methods that include, and make use of, redundancy in sensor information perform better in cases with measurement error. Time difference of arrival measurements can be modelled and expressed in a least squares form. The Least Squares Estimator (LSE) can be used to minimise the squared error between the model and the measured data. With an accurate model, in the absence of noise, the square error function should be zero. The form of the error criteria derived from the source localisation problem can effect the computational complexity of finding a solution.

The hyperbolic error function is formulated from the difference between the observed range differences and the those calculated for the estimated source location. As described above, the range difference between a pair of sensors defines a hyperboloid, with a solution for each point on the surface. Minimising the hyperbolic least squares criterion places a sound source estimate as close to all of the hyperboloids as possible but these functions are unstable in the presence of noise, becoming mathematically intractable as the number of sensors grows large.

In order to overcome these problems a second error function, based on the intersection of spheres centred at each of the microphone locations, can be derived. By substituting in a further unknown variable, the range from each of the microphones to the unknown source position  $R_1$ , a linear set of equations is obtained with a quadratic constraint linking the source coordinates and range estimation. Several techniques for solving for the source co-ordinates using this set of equations have been proposed with progressively improved performance and simplicity. Optimum performance is defined by calculating the Cramér-Rao lower bound on the variance of each estimated coordinate. Substituting in for the intermediate variable,  $R_1$  to give  $N - 2$  linear equations in  $x$ ,  $y$  and  $z$ , or solving for the coordinates first, to give equations in terms of  $R_1$  only first, result in solutions which can be shown to be

mathematically equivalent. This is termed the Spherical Interpolation (SI) method. Redundant information in the source range is not used with this method, and there is a large variance relative to the Cramér-Rao lower bound.

Chan and Ho[10] suggested using a second least squares estimator to make use of the redundancy in the measurements. Their approach takes advantage of the relation of the source coordinates to the range to improve the estimation efficiency using quadratic correction. This is termed the Quadratic-Correction Least-Squares (QCLS) method. Huang, Benesty, *et. al*[27] propose further changes to this approach, that no longer require an assumption on the measurement error covariance matrix to be made. A perturbation approach in the quadratic correction of QCLS, which results in a limit on the error magnitude, is also overcome. This method is termed Linear-Correction Least Squares (LCLS).

### *Derivation of Closed Form Estimator*

Following the derivations of Chan and Ho[10], Hui and So[53], and Huang and Benesty[27] the problem can be stated as follows. Let the positions of the  $N + 1$  microphones, in Cartesian coordinates be denoted by

$$\mathbf{r}_i \triangleq [x_i, y_i, z_i]^T, \quad i = 0, 1, \dots, N,$$

where  $(\cdot)^T$  denotes the transpose of a vector. The first microphone is set to be the origin,  $\mathbf{r}_0 = [0, 0, 0]^T$  and the source coordinates sought are  $\mathbf{r}_s \triangleq [x_s, y_s, z_s]^T$ . The range between the  $i$ -th microphone and the source is given by the Euclidean norm

$$D_i \triangleq \|\mathbf{r}_s - \mathbf{r}_i\| = \sqrt{(x_s - x_i)^2 + (y_s - y_i)^2 + (z_s - z_i)^2}. \quad (7.1)$$

The corresponding range differences, for microphones  $i$  and  $j$  in a pair, to the source are given by

$$d_{ij} \triangleq D_i - D_j, \quad i, j = 0, \dots, N. \quad (7.2)$$

This is related, by the speed of sound  $\nu$ , to the time difference of arrival,  $\tau_{ij}$ , measured in synchronised microphone arrays by

$$d_{ij} = \nu \cdot \tau_{ij}.$$

It is noted[27], that there are  $(N + 1)N/2$  distinct delay measurements,  $\tau_{ij}$ , excluding the  $i = j$  case and repetition arising from  $\tau_{ij} = -\tau_{ji}$ , but any  $N$  linearly independent  $\tau_{ij}$  values determine all the others in the absence of noise. For simplicity, at the cost of improved accuracy in noisy systems, the  $N$  time differences of

arrival with respect to the first sensor,  $\tau_{i0}, i = 1, \dots, N$  are selected. The range differences measured between connected microphones,  $d_{i0}$  are modelled as the actual value and an additive noise term  $\epsilon_i$ , assumed to be zero-mean:

$$\hat{d}_{i0} = d_{i0} + \epsilon_i, \quad i = 1, \dots, N \quad (7.3)$$

where,

$$d_{i0} = \|\mathbf{r}_s - \mathbf{r}_i\| - \|\mathbf{r}_s\|.$$

From the range difference (7.2) and distance definitions (7.1) we have

$$\begin{aligned} \hat{d}_{i0} + \sqrt{(x_s)^2 + (y_s)^2 + (z_s)^2} \\ = \sqrt{(x_s - x_i)^2 + (y_s - y_i)^2 + (z_s - z_i)^2}, \quad i = 1, \dots, N. \end{aligned} \quad (7.4)$$

The solution to this set of hyperbolic functions is nonlinear and sensitive to measurement noise. Instead, a spherical error function can be formulated.

The distances from  $\mathbf{r}_0$  at the origin to the remaining microphones and to the source are denoted by  $R_i$  and  $R_s$  respectively, where

$$R_i \triangleq \|\mathbf{r}_i\| = \sqrt{x_i^2 + y_i^2 + z_i^2}, \quad i = 1, \dots, N \quad (7.5)$$

$$R_s \triangleq \|\mathbf{r}_s\| = \sqrt{x_s^2 + y_s^2 + z_s^2}. \quad (7.6)$$

Squaring both sides of (7.4) and substituting in  $R_s$  as an intermediate variable yields a set of linear, spherical signal model, equations:

$$\mathbf{r}_i^T \mathbf{r}_s + d_{i0} R_s = \frac{1}{2}(R_i^2 - d_{i0}^2), \quad i = 1, \dots, N. \quad (7.7)$$

Writing this spherical least squares error function in vector form,

$$\mathbf{G}\boldsymbol{\theta} = \mathbf{h}, \quad (7.8)$$

where,

$$\mathbf{G} \triangleq [\mathbf{S}|\hat{\mathbf{d}}], \quad \mathbf{S} \triangleq \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_N & y_N & z_N \end{bmatrix}, \quad (7.9)$$

$$\boldsymbol{\theta} \triangleq \begin{bmatrix} x_s \\ y_s \\ z_s \\ R_s \end{bmatrix}, \quad \mathbf{h} \triangleq \frac{1}{2} \begin{bmatrix} R_1^2 - \hat{d}_{10}^2 \\ R_2^2 - \hat{d}_{20}^2 \\ \vdots \\ R_N^2 - \hat{d}_{N0}^2 \end{bmatrix}, \quad (7.10)$$

$$\hat{\mathbf{d}} \triangleq [\hat{d}_{10} \quad \hat{d}_{20} \quad \dots \quad \hat{d}_{N0}]^T, \quad (7.11)$$

and  $[\cdot|\cdot]$  implies adjoining of two matrices. Solving the corresponding least squares criterion is a linear minimisation problem

$$\min_{\boldsymbol{\theta}} (\mathbf{G}\boldsymbol{\theta} - \mathbf{h})^T (\mathbf{G}\boldsymbol{\theta} - \mathbf{h}) \quad (7.12)$$

subject to the quadratic constraint

$$\boldsymbol{\theta}^T \boldsymbol{\Sigma} \boldsymbol{\theta} = 0, \quad (7.13)$$

where  $\boldsymbol{\Sigma} \triangleq \text{diag}(1, 1, 1, -1)$  is a diagonal and orthonormal matrix. The technique of Lagrange multipliers can be used yielding the constrained least squares estimate:

$$\hat{\boldsymbol{\theta}} = (\mathbf{G}^T \mathbf{G} + \lambda \boldsymbol{\Sigma})^{-1} \mathbf{G}^T \mathbf{h}, \quad (7.14)$$

where  $\lambda$  is still to be found.

An unconstrained spherical least squares estimator can be derived by not enforcing the quadratic constraint (7.13) which is equivalent to assuming that the source coordinates and distance to the origin  $x_s, y_s, z_s$ , and  $R_s$  are mutually independent. An estimate of  $\boldsymbol{\theta}$  is then given by

$$\hat{\boldsymbol{\theta}}_1 = \mathbf{G}^\dagger \mathbf{h}, \quad (7.15)$$

where

$$\mathbf{G}^\dagger = (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T$$

is the pseudo-inverse of the matrix  $\mathbf{G}$ . This estimate has been shown [26] to be accurate in some cases, but the source range information redundancy (7.13) can be exploited to improve this estimate.

### *Simultaneous, Two-Sound Solution*

The heterogeneous, impromptu networks envisaged in Chapter 5 are subject to the constraints of unsynchronised member devices. As with the case of collaborative colocation discussed in the previous chapters, this means that arrival times determined at each of the nodes cannot be directly compared as there is no common reference time. Instead the delay lengths between two successive sounds are sought. Comparing these relative timings from each of the nodes yields a system of equations describing the interaction of possible positions for the two sound sources. If the sounds occurred in separate spatial locations then each of the delays will contain timing offsets dependent on the difference in path lengths from each of the two sources to that node. Subtracting the delay measured at one node from that at another, removes the unknown time between the two sounds.

An iterative nonlinear least squares approach can be applied for direct numerical optimisation. No attempt is made, in this case, to formulate a linear approximation or closed form. An interior, trust region approach to nonlinear minimisation is employed, as implemented in the Matlab function *lsqcurvefit*. This provides a proof of concept test as to whether the error on combined measurements, for a sound source pair, can be minimised concurrently, with the additional cost of algorithm complexity and computational time. If an approximate initial guess can be given for the two source locations, then the solution is more likely to converge correctly and quickly. By adapting the closed form linear-correction least square estimation to estimate one source location, given the other, an initial position guess can be derived from the previous source position and the delay between source sounds measured at each of the devices. Using only this method, without the global calibration nonlinear least squares approximation step — chaining a number of source position estimates, each based on the output of the last — would quickly propagate any large errors, as an erroneous output becomes the fixed source position for the next sound source pair.

#### PROCEDURE 7.1: SOURCE LOCATION ESTIMATION FOR A CHAIN OF SOUNDS

- 1: at each sensor, record a series of sounds
- 2: **if** chirp signals are used **then**
- 3:   correlate to find peaks of each
- 4: **else**
- 5:   salient features must be selected from the audio stream (not implemented)
- 6:   *detect onset features*
- 7:   *pass short onset recording to other nodes*
- 8:   *correlate around onset to find precise timing*
- 9: **end if**



- 10: pass chain of sound arrival times to elected processor node
- 11: at elected, collect sensor locations {from colocalisation step or grid position}
- 12: pick centre of sensor distribution for guess of first sound position
- 13: **for** each remaining sound in chain **do**
- 14:   calculate estimate of this second sound in pair from the estimate of the first sound using the closed form least squares approximation
- 15:   compute the numerical least squares minimisation for the positions of both sounds in pair iteratively until a stopping criteria is met.
- 16:   set the new estimate of the second sound as the first sound of the next pair.
- 17: **end for**

## 7.2 EVALUATION OF METHOD

A chain of sounds, emitted alternatively, from two points in front of a typical SAS array geometry was simulated. The microphone positions reflected those of the tiles mounted in a wall grid, with 17 tiles, space over a corner of two walls. The tiles are  $43.2 \times 26.7$  cm in size, resulting in equivalent microphone spacing. The simulated sources were placed in the same positions as the experimental test case below, with one emitter at a height of  $z_1 = 144$  cm, and a distance  $x_1 = 85$  and  $y_1 = 110$  cm from the corner of the room, with the second at  $z_2 = 139$ ,  $x_2 = 5.8$  and  $y_2 = 152$  cm. An error was added to each of the calculated delay observations between the two sounds, consisting of a typical 1 cm Gaussian noise and an error of 2, 3 or 6 samples. Positions of each sound pair were computed using the iterative numerical least squares estimator, from randomly chosen initial source coordinate guesses. Shown in Figure 7.3 are the three distributions obtained for a chain of 100 sounds at three sample uncertainty levels. At each step two randomly chosen starting positions were selected, bounded within a room-sized region around the microphones. The standard deviation in position estimates is  $\sigma_2 = 18.7$  cm,  $\sigma_3 = 21.7$  cm and  $\sigma_5 = 41.7$  cm, where  $\sigma_k$  denotes the mean of the standard deviation in each dimension of the position coordinates, for the simulation case of  $k$  samples of measurement error. The mean pair-wise range estimates lie close to the actual distance  $s = 50.2$  cm, with  $\hat{s}_2 = 53.4 \pm 13.3$  cm,  $\hat{s}_3 = 52.9 \pm 16.1$  cm and  $\hat{s}_6 = 57.5 \pm 27.8$  cm, where  $s_k$  denotes a pair-wise range  $\pm$  one standard deviation, for the simulation case of  $k$  samples of measurement error. The resulting spread of estimations reflects the microphone array geometry. The mean position estimates are offset from the true positions, by  $\simeq 5$  cm in each case.

The closed form least squares spherical interpolation step was then added, estimating a starting point for each new sound from the final estimate of the previous

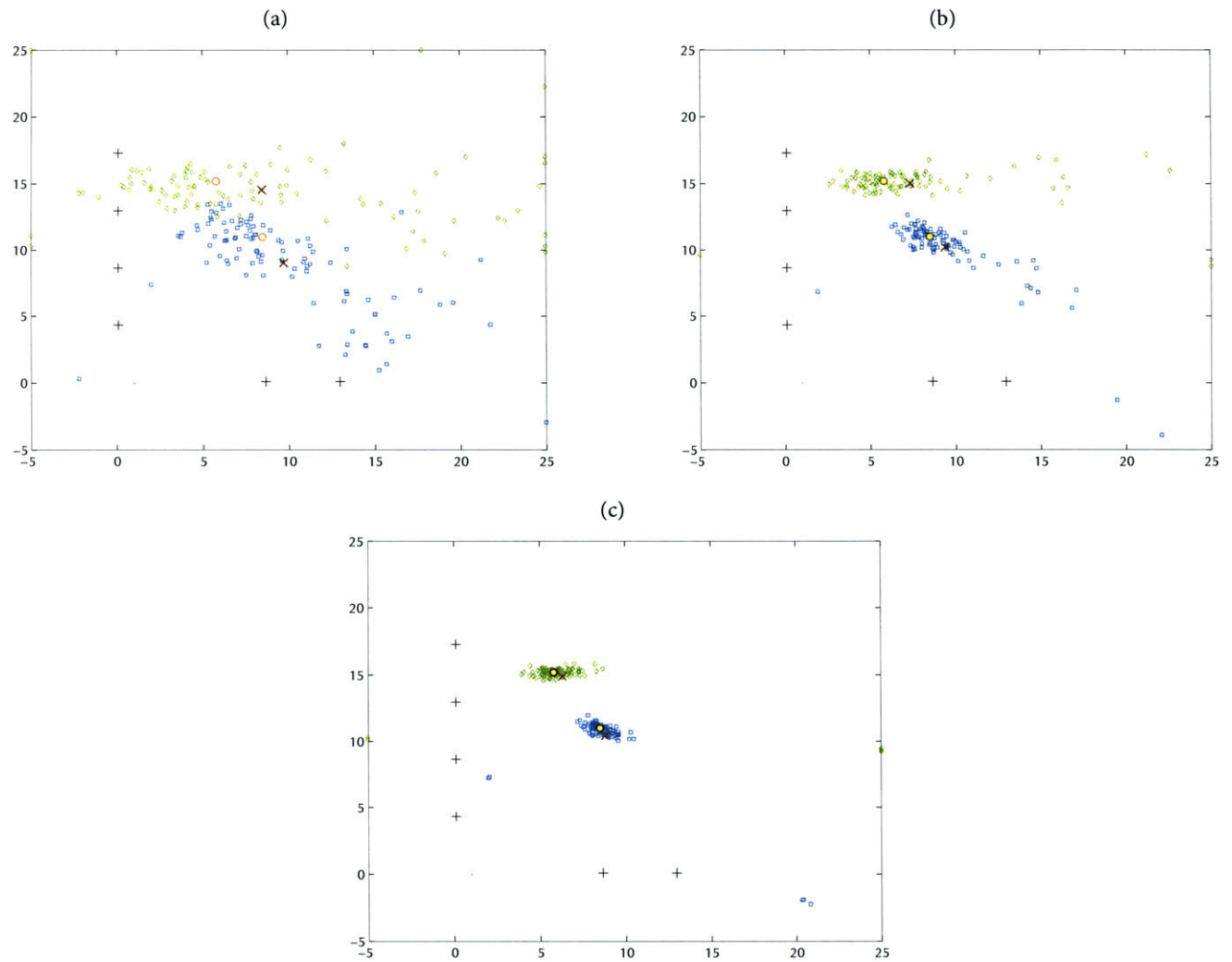


FIGURE 7.3. Clustering of sound pair position estimates for 100 sound chain at 2, 4 and 6 sample ranging error. Key: + microphone, × cluster mean, ○ actual sources, □ and ◇ sound estimates. Both axes are room coordinates  $\times 10^{-2}m$

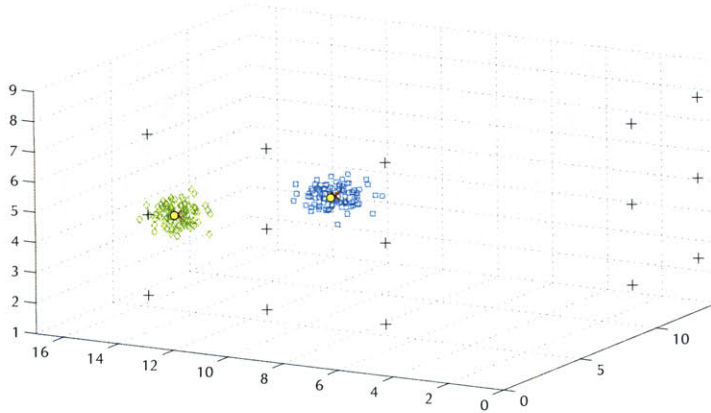


FIGURE 7.4. Two step source estimation, shows an improvement with the initial closed form positioning. Plot is rotated to reveal spread of the data through the space. All coordinates  $\times 10^{-2}m$

one. Although limited in accuracy, this estimation provides a correct region for the new sound and establishes an approximate pair separation. For measurements with, low, 2-sample error on each delay, corresponding to  $1.5\text{ cm}$ , the closed form step provides significant improvement in position accuracy and computation time. For example, along the direction in which the microphone distribution is sparsest, the mean value falls from  $63.4\text{ cm}$  to  $59.5\text{ cm}$ , with a correct value of  $58.0\text{ cm}$ , with the standard deviation falling from  $28.7$  to  $9.8\text{ cm}$ . The mean number of iterations before one of the stopping conditions is reached reduces from 122 to 72 steps. As the measurement error increase, the improvement becomes less significant, but still results in fewer mean iterations, and smaller standard deviation. Figure 7.4 shows a plot of the pair-wise estimates in the room space in front of the wall mounted microphones.

In order to practically test the two step approach to source pair position estimation, two handheld-sized XScale processors were used to alternatively emit a series of reference sound signals at two positions in front of a wall mounted array of Smart Architectural Surface tiles. As with the active, pair-wise ranging algorithm, maximal-length sequence chirps were used as test sounds for their proven accuracy in arrival time detection, and their deterministic nature, allowing them to be calculated separately at both emitters and each of the detectors in the sensor network. The uncertainty in peak detection for the delay between two chirps is on the order of two samples. In the next chapter a discussion on extending this approach from

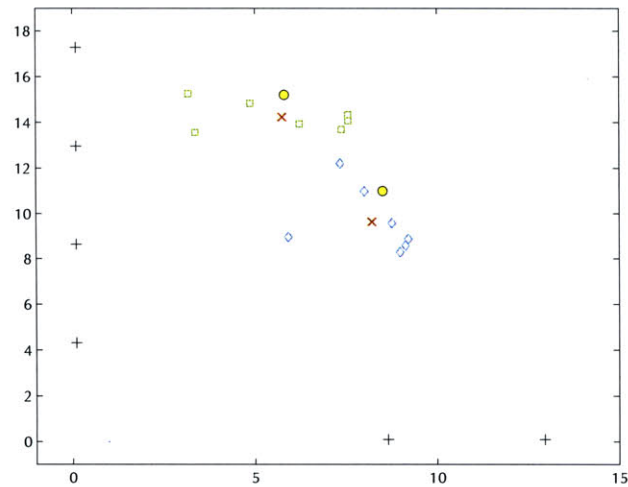


FIGURE 7.5. Sound pair source location results for an 8 noise test using SAS tiles (key as above). All coordinates  $\times 10^{-2}m$

identifying and positioning pseudo noise chirps, to tracking arbitrary sounds. By alternating the sounds between the two test positions a number of spatially separated source pair delay measurements were made. In noiseless conditions, each estimate of the two source positions should be identical to the last. An 8-chirp series of 512 bit sequences was used, each phase shift encoded onto a 11025  $Hz$  carrier rate and emitted from alternating positions, 50.2  $cm$  apart, every sixth of a second. The number of chirps in this test is limited to 8 by the record buffer available on each SAS tile. The signals were transmitted and captured at 44.1  $kHz$ . Of the 16 microphones were used one device exhibited much lower sensitivity than the others, failing to register the arriving signals, and so is not included in the calculation and results. By testing audibility of signals between nodes in an initial group forming step, unresponsive sensors and emitters should be easy to detect automatically. Figure 7.5 shows the plan view of the microphones in the room and the results for the estimated positions. For each pair, the new node coordinates were first estimated from the previously guessed node position using the least squares, closed form, solution. Following this the numerical, nonlinear, least squares step is then applied using the estimates from the first step as initial conditions. The estimates of position converge towards the actual positions in this step, and lie close after falling below the error threshold or reaching a specified maximum number of iterations. The inter-pair distances tend to the exact measurement, the mean measured value is  $53.8 \pm 3.0$   $cm$ , in comparison to an actual value of 50.2  $cm$ . The mean value in the simulated case for equivalent measurement uncertainty is

53.4 *cm*. The absolute position of the pair has a mean of the standard deviation in each dimension of  $\sigma_{exp} = 13.0$  *cm*, with an average Euclidean distance between the estimated positions and the actual coordinates of 35.2 *cm*.



## CHAPTER 7

# DISCUSSION

The thesis of this work, introduced in the first chapter, is that sound can be used to establish a practical common coordinate system between sensor-equipped, but heterogeneous, devices, such as those already in our pockets. The information collected can then be used to enable this impromptu collection to act as a distributed microphone array, spatially mapping unknown acoustic events, which are naturally linked to human activity and scale. The tools of collaborative colocalisation and subsequent source location have been developed and tested, under the constraints of real-room sensing and practical hardware. In this document the case has been presented for the novelty and utility of these algorithms, the distributed, ubiquitous systems that motivate them and the human centred applications that will benefit from robust range estimation between devices and activity triggered sounds.

### 7.1 COLOCALISATION

A simple method for finding pair-wise range measurements between only coarsely synchronised devices, equipped with microphones and speakers, has been described and demonstrated. Correcting for the effect on the speed of sound of environmental conditions, error in the measurements is close to the uncertainty in the sampling rates used. For 44.1  $kHz$  samples in the tests described, this is approximately 3 and 5  $cm$ . Direct face-to-face and lateral sensing test cases are presented. In the lateral ranging case, sensors and emitters are directed perpendicularly to their plane of separation. Pseudo-noise chirps are calculated and correlation peaks found in the recordings at each of the nodes, yielding calculations of ranges. These can be passed to a number of existing algorithms, for example[29], to accurately establish a common co-ordinate system across the sensors in an *ad-hoc* system. It is shown that using a multidimensional scaling approach, the sensor network positions can be retrieved up to global rotation and reflection of the system.

At each node, a peak finding search passes through a single recording of all the reference chirps emitted from each device. In comparison to streaming data to a central processor, only a small amount of information, the length of the delays between peaks, has to be passed over the communication channel. The current results were taken using a correlation algorithm that runs over the entire length of the recording and thus proportionally scales the processing time with the number of nodes. Running an initial peak-finding step on a down-sampled version of the recording to identify areas of interest first may increase the speed of the algorithm. As the number of nodes in a densely populated system grows, the approach can be modified so that each sensor looks for peaks in a constrained time window, such as the time for 6 chirps to have occurred before and after the nodes own chirp. Multidimensional scaling methods that improve performance in networks containing a large number of nodes with a degree of connectivity on the order of 10 neighbours or more, such as Moore *et al.* [40] which avoids settling on false minima in possible conformation, can be used in this case.

The Smart Architectural Surfaces platform has been used in these tests, providing distributed processor and sensor resources, and group-forming and communications protocol. The algorithms have been developed to function, with little modification, over a large range of electronic devices. This is demonstrated in a multi-device, real-room, test case. Despite degradation in peak finding performance due to compression of the audio recordings, accurate range measurements are still obtained with outlier rejection and careful tuning of the threshold levels. Optimisation of the code to use fixed-point only calculations, which improve speed on the limited resources of processors on the SAS, phones and similar devices, is currently work in progress. The authors are also looking to use the temperature and humidity sensors available in the test platform as a possible way to correct for global variations in sound speed, however it has been observed [18] that local variations, such as air outlets between nodes can have a significant effect in sound based ranging approaches. Statistical analysis of repeat measurements and calibration for a known separation between a sensor pair may provide effective scaling in future tests on generic sensor networks.

### *Improvements*

Although conformation can be found, one thing that cannot be determined is the relative orientation of each device. On specialised hardware, two microphones could be used to find this, but on typical consumer electronics, and the SAS tiles orientation cannot be found with acoustic timing. It is possible that some spectral



information indicative of direction could be used, but modelling this would be complex. Similarly, occlusion, may lead to a secondary reflected path being detected, but mapping reflected images of a source back to the original is difficult. The need for orientation information arises, in the tile case for example, in which a number of screens could display elements of a large image or visualised dataset, but the viewpoint correction on each screen must be calculated to accommodate the tile orientation relative to the others. Currently the tiles are assumed to be constrained to be facing perpendicularly to two flat surfaces on which they are mounted to achieve this. Direction and common screen view can be found by comparing video images from the camera on each of the tiles. Extracting prominent features, for example SIFT<sup>†</sup>, present in each of the frames, and comparing the relative shift, a view point and relative rotation and translation between the two tiles can be calculated.

Scale-Invariant Feature  
Tracking[35]

This code currently takes 30 s to determine a single pair-wise range measurement. A large component of this time is spent performing cross-correlation, first with the control chirp and then using the two expected chirps, with the recording. This calculation is currently performed using a floating-point implementation of the FFT<sup>†</sup>, which without corresponding floating point handling on the processor, must be handled inefficiently in code. Converting this to a fixed point implementation would increase speed with little effect on accuracy. This would also aid code portability, by supporting other small devices that do not offer hardware floating point.

Fast Fourier Transform

The ear uses onset detection to identify possible new sounds, and fine grained inter-aural timing comparison for direction estimation. Computing time of the correlation component may be greatly reduced if a rough estimate of when the chirps occur can be found first, before more accurately fine tuning the arrival time based on correlation. This initial step could involve downsampling the recording to reduce the data points the correlation algorithm must parse, or by looking for onset or zero crossing rate patterns for chirp like sounds and subsequently correlating around those times only, rather than the whole recording.

## 8.2 SOURCE LOCATION

Estimating the position of the sources of unknown sounds from the time difference of arrival at a number of microphones is a challenging problem, which involves solving an unstable system of equations in the case of noise corrupted measurements. This is further complicated in a system of unsynchronised devices, in which direct arrival time comparisons are not possible. A technique has been presented that uses the relative timing between two unknown sounds, attempting to solve for

both source positions at once. By stepping through a series of sounds in this way, estimations for previous positions can be used to estimate the next. This approach uses the collaborative localisation presented to establish the relative position of each of the microphones, which is then used in the source localisation algorithm. A two step estimation process is developed, in which a closed-form approximation is first used to estimate the second sound in a pair. These two source positions are then used as initial guesses in a numerical, iterative least squares minimisation. Performance was simulated for a typical sensor layout, demonstrating convergence to the correct location coordinates for a range of sound source positions. The approach was then tested using recordings from a microphone array composed of tiles in the Smart Architectural Surfaces platform. A maximal-length chirp sequence was used to aid accurate detection of the arriving signals. A series of chirps were emitted, alternating between two positions in front of the array. The sound pair separation is estimated with a 3.0 *cm* uncertainty. The absolute position of the pair is less accurate, but successive measurements vary by a standard deviation of 13.0 *cm*, with a mean distance to the actual pair position of 35.2 *cm*. The solution is susceptible to sign error of the reference sounds. This suggests that more than two sounds could be used to strengthen absolute positioning at the cost of increased computation. The use of separation estimates between pairs, only, may also be used to yield a relative coordinate system between sounds. These algorithms for estimating source position from sensor timing measurements are currently centralised in design. They require each sensor node to pass measured recording times to a single node for estimation.

### *Using the Information*

The accuracy of the position estimates is relatively coarse in comparison to that attained in the device localisation case. By selecting the most prominent sounds it is hoped that correlation performance similar to that achieved with the maximal-length sequence chirps can be attained. Sound streams, in which many sound events connected with a single source and location can be found, can iteratively improved in accuracy, by clustering the estimated coordinates of each source. This approach also requires sounds to occur at several different locations in succession in order for the delays measured to contain position information. Such is the case for busy audio scenes, conversations and activity generated sounds along a path through space. A large body of research has investigated ways to combine many position estimates over time into accurate estimates of a number of repeating sound sources. Clustering over short and long time periods reveals a range of information from turn taking by talkers to long-term use of a space. For example, the size of a room and

the permanent obstacles within it are likely to become apparent from the sparseness of sounds originating from those positions.

### 8.3 SOUND EVENT SELECTION

Most beamforming and source localisation approaches use the entire recording from each microphone, comparing correlation between each in small sections to pick out a running estimate of sound source position. This is unsuitable for a distributed system in which streaming each of the recordings to the other nodes in a network, or to an elected central point for processing rapidly becomes unfeasible as the network scales. Instead, to implement source location on a distributed sensor network, a means of picking features to match between recordings must be established. The focus is then shifted from estimates of position for each moment in the recording, to locating key acoustic events, and building longer term patterns of position. It is proposed that such a system, although using less of the signal information is more robust to real room reverberation and better reflects an event focus.

Jehan[28] has presented an effective onset detection algorithm primarily intended for audio segmentation and beat detection in music. The technique is driven by analogy to human experiences of listening, with a number of steps to better match our own acoustic perception of sound onset. This same technique can be used to detect the unknown sound onsets sought, which promise to act as effective markers for timing measurements in room recordings. The onset of new sounds is marked by sharp increases in spectral power. The human centric focus of Jehan's approach compliments the human scale focus of the acoustic sensing and localisation presented in this work. The method, we will see, shifts focus of the analysis to better match the frequency range and temporal response our own auditory systems emphasise. Not only does this favour sounds with significant human meaning but also reflects our evolution to concentrate on initial onset rather than following reverberation, and to pick out onsets that are likely to correspond to new acoustic events.

Taking a waveform recorded from a room, our first step is to compute the Short Term Fourier Transform (STFT) allowing us to monitor frequency information over the length of the signal. Short windows are selected to increase timing information, at the cost of frequency precision. The regularly spaced frequency bins of the STFT are mapped to a nonlinear Bark scale that better matches our hearing, placing less emphasis on low frequency components. Alternatively the Mel scale may also be used. Temporal smoothing is then applied across the spectrogram, by convolving

each critical band with a half-Hanning window [sparkline]. This reflects the temporal masking response of the ear, but has the effect of placing emphasis on initial sounds while smoothing rapidly successive peaks that follow the attack, such as multipath reflections. Frequency masking between the critical bands is matched by smoothing across the frequency bins. Jehan terms the resulting spectrogram “what-you-see-is-what-you-hear”, with just visible in the spectral plot corresponding to just audible for human listeners. The frequency bins are then summed across each window to determine a loudness function. Sharp increases in loudness are found by taking the derivative of this curve. These peaks are smoothed to remove close secondary subpeaks, again reflecting the perceptual fusing of such quickly repeated onset transients, and favouring the arrival of sounds marking new acoustic events. The local maxima which rise above a small threshold, mark each of the candidate sounds. These onsets can be found in the recordings of each device, and locate sound arrival times on a *ms* scale. A small sample can be taken around this point from one recording, and used to correlate against the corresponding onset markers on the other recordings of the same sound to determine exact timing offset. This improves code efficiency, by only applying fine correlation to short sections of the recording, and suits distributed sensing environment, in which only small reference clips are passed between devices, rather than streaming an entire signal.

Shown in Figure 8.1 is a short recording of room noises — tapping, talking and clapping — taken from one of two microphones separated by a distance of 160 *cm*, recorded as two channels to maintain common timing between them. The waveform is translated to a Bark scale auditory spectrogram, below that is the loudness curve, the peaks marking onset transients, and finally the smoothed peaks with the onsets detected marked in as vertical strokes. An implementation of the onset detection algorithm by Lieberman[32] is used here. Figure 8.2 shows the correlation of an onset sample selected automatically in this way with the corresponding onset region in the second channel. The sound was created at a point on the axis of the two microphones. The time delay between the two sounds was measured as 190 samples. This sound flight time corresponds to a distance of 150 *cm*, close to the actual distance of 160 *cm*. Further quantifying the error in this technique remains as future work.

Work must still be done to select, transmit and match these candidate onset samples in a robust way. Each sensor may not detect all of the possible peaks, for example, so a means of matching the patterns of sound onsets to find which ones correspond to which between nodes must be developed.

Zero crossing rate and running entropy calculations can provide indications of highly variant sections of a waveform, indicative of noise-like components that may

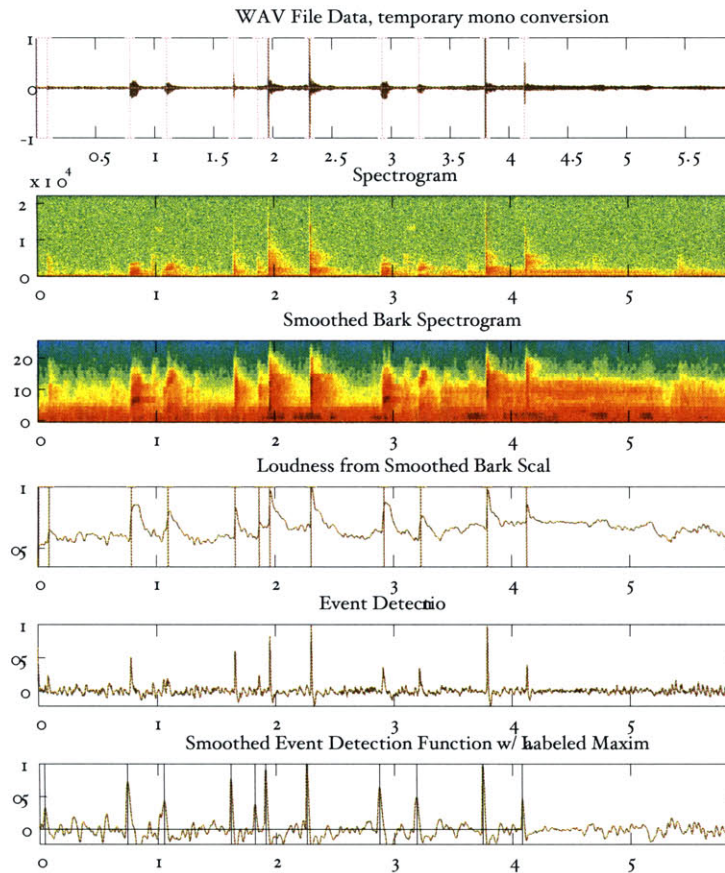


FIGURE 8.1. Onset detection using the approach of Jehan[28] and Lieberman[32] for a short room recording of 4 claps and the word "batfink" spoken twice. Background noise included computer noise and a recording of a dog growling. Onsets are found for each of the claps and two parts of the words, and three background noises.

match the properties of the maximal-length pseudo noise sequences [15]. However, in real room recordings this is often the reverberant part of a signal, leading to poor correlation with other recordings of the same event. Zero crossings, in particular are very efficient to calculate, and so may be able to compliment an onset detection algorithm.

## 8.4 ENABLING IMPROMPTU SENSING

Borrowing the digital resources from the devices of the strangers around us involves challenges of both technical implementation and social negotiation. These are issues that this research raises, addressing concepts of technological and digital commons. Building tools that enable people to share the resources they already

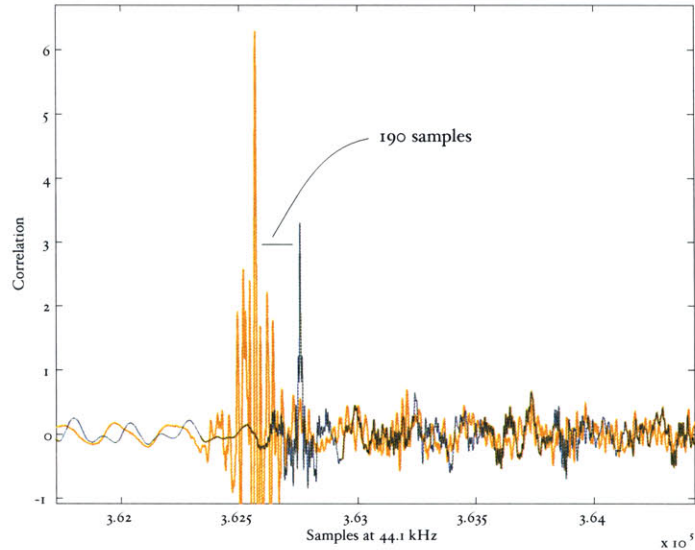


FIGURE 8.2. Correlation between an automatically selected onset sample and two recordings of the sound. Sharp peaks occur at the correct timing points, showing the spatial separation of the two microphones.

carry allows friends and colleagues to achieve tasks that one person's digital devices could not, and challenges strangers to collaborate, when it could be of benefit to them and sharing devices could yield information than any single devices. Two related areas of *ad-hoc* collaboration, spatial mapping and power structures in surveillance, are briefly introduced below, as future avenues of investigation.

### *Context and Connectedness*

This research uses the properties of sound to build tools that can begin to map the history and context of a space. One long term goal is to investigate how technology can compliment conversations. An open question in this research is whether these resources can aid interaction between people. As we saw in Section 3.2, recent research has shown that group discussions can be effected positively through the display of participation information and history as people talk [14] — this work has been developing the resources for the collection of this spatial and temporal acoustic information using impromptu networks of the microphones inside hand-held computers and mobile phones of meeting participants.

## *Surveillance and Sousveillance*

Any work on aiding the tracking of people by machines must address questions of civil rights, ethics and any potentially harmful uses of the tools. Our personal devices and city surveillance infrastructures are already capable of considerable tracking and eavesdropping. We look to those who control them to offer us protection. Throughout our history, the control of information channels has raised important questions on how these systems will be used and misused. Can systems be designed for methods of trust and empowerment in surveillance and security? People may be able to create ways of actively improving their safety and sense of openness without surrendering their civil liberties and without propagating mistrust and fear. The tools presented in this work mandate collaboration of devices and their owners for sensing to occur. Negotiated, informal information collection in our communities may provide a means of empowering people in the use of this information.





# Appendices



## APPENDIX A

# TECHNICAL SPECIFICATIONS

The Smart Architectural Tile processor is a handheld sized, single-board computer, the Applied Data Systems XScale Bitsy with 400 MHz RAM and 64 Mb of onboard memory. Network connection is established through a standard PCMCIA<sup>†</sup> wireless 802.11b card. With no onboard floating point support, any floating point calculations must be handled in code, leading to slow processing in these cases. However, many algorithms can be optimised for fixed point systems.

Personal Computer  
Memory Card  
International  
Association standard

A USB driver and 'smart IO' socket enable connection of resources. Sensors attached to most tiles include a camera, a microphone, an ultrasound sensor, 2.4 GHz electromagnetic signals via the wireless card, temperature, humidity and barometric sensors. Effecters through which a tile can actuate the room include a screen, an ultrasound emitter, 2.4 GHz wireless emission and an acoustic speaker. The microphones are low-cost Panasonic electret capsules, in small 10 mm and 6 mm diameter packages. Tiles use either omnidirectional RF noise-resistant types or a unidirectional model although both exhibit similar wide directional detection. The microphones have a sensitivity of  $-47 \pm 4$  dB with a signal to noise ratio quoted as more than 60 dB. The frequency response is flat over a frequency range of 200 Hz to 20 kHz increasing down to 20 Hz in the omnidirectional case. The microphones have a built-in FET amplifier circuit, and consume under 0.5 mA at a standard operating voltage of 1.5 V. Resistance to radio frequency interference is provided by a  $\sim 10$  pF capacitor. The speaker is a Panasonic EAS3P127A 'micro speaker' which is a slim 4 mm deep, and 36 mm in diameter. It has a broad frequency response extending from 500 Hz to 4 kHz. Audio sampling and emission is handled by an on-board sound card chip, the CirrusLogic CS4202 audio codec, which meets the Audio Codec '97 2.2 component specification. This includes an integrated microphone pre-amplifier circuit, and uses 20-bit digital-to-analogue and 18-bit analogue-to-digital and internal sample rate conversion.

The board runs the 2.4.19 ARM<sup>†</sup> branch Linux kernel, which is held in flash

Acorn RISC Machine  
architecture

GNU Compiler  
Collection

memory, and boots in  $\sim 30$  s. This version of the kernel does not support the Advanced Linux Sound Architecture (ALSA), but does support the older Open Sound System (OSS) which provides a comprehensive sound card driver system. For ease of development and testing, programs compiled for the system, using the GCC<sup>†</sup> cross-compiler, are stored and loaded from an NFS (Network File System) mounted directory. With the addition of slightly larger Flash memory, however the executables could easily be stored locally.

Communication is via a typical TCP/IP connection over an 802.11b network. Interaction is currently implemented via a single router, rather than direct broadcast/multicast protocol, a system that is limited in scalability. Applications are developed with flexibility to use either the former or latter interaction in mind. Streaming raw audio with a sampling rate of 44100 Hz leads to data loss, especially if the network is busy. Streaming mp3 compressed audio is handled successfully in most instances.

# BIBLIOGRAPHY

- [1] <http://en.wikipedia.org/wiki/rss>.
- [2] J. Bachrach and C. Taylor. Localization in sensor networks. In *Handbook of Sensor Networks*. Wiley, 2005.
- [3] S Basu, S Schwartz, and A Pentland. Wearable phased arrays for sound localization and enhancement. *Proceedings of the Fourth International Symposium on Wearable Computing*, 2000.
- [4] G Borriello, A Liu, T Offer, C Palistrant, and R Sharp. Walrus: Wireless acoustic location with room-level resolution using ultrasound. *Proceedings of the 3rd international conference on Mobile and Ubiquitous Multimedia*, 2005.
- [5] V.Michael Jr Bove and Jacky Mallett. Collaborative knowledge building by smart sensors. In *BT Technology Journal*, 2004.
- [6] M. Broxton, J. Lifton, and J. Paradiso. Localizing a sensor network via collaborative processing of global stimuli. *European Workshop on Wireless Sensor Networks*, 2005.
- [7] William Joseph Butera. *Programming a Paintable Computer*. PhD thesis, Media Arts and Sciences, School of Architecture and Planning, February 2002.
- [8] D. Carevic. Tracking target in cluttered environment using multilateral time-delay measurements. *The Journal of the Acoustical Society of America*, 2004.
- [9] Dragana Carevic. Tracking target in cluttered environment using multilateral time-delay measurements. *The Journal of the Acoustical Society of America*, 2003.
- [10] Y.T. Chan and K.C. Ho. A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 42(8):1905--1915, August 1994.
- [11] Michael Coen, Brenton Phillips, Nimrod Warshawsky, Luke Weisman, Stephen Peters, and Peter Finin. Meeting the computational needs of intelligent environments: The metagluе system. In Paddy Nixon, Gerard Lacey,

and Simon Dobson, editors, *1st International Workshop on Managing Interactions in Smart Environments (MANSE'99)*, pages 201--212, Dublin, Ireland, Dec. 1999. Springer-Verlag.

- [12] Bram Cohen. <http://www.bittorrent.com/>.
- [13] P. R. Cook and D. J. Levitin. *Music, Cognition and Computerized Sound: An Introduction to Psychoacoustics*. MIT Press, 1999.
- [14] J.M. DiMicco, A. Pandolfo, and W. Bender. Influencing group participation with a shared display. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 614--623. ACM Press, 2004.
- [15] Daniel P.W. Ellis and Keansub Lee. Features for segmenting and classifying long-duration recordings of ``personal'' audio. *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, 2004.
- [16] D.P.W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996.
- [17] B. G. Ferguson, L. G. Criswick, and K. W. Lo. Locating far-field impulsive sound sources in air by triangulation. *J Acoust Soc Am.*, 2002.
- [18] Lewis Girod and Deborah Estrin. Robust range estimation using acoustic and multimodal sensing, 2001.
- [19] Peter Gorniak and Deb Roy. Probabilistic grounding of situated speech using plan recognition and reference resolution. *Proceedings of the International Conference for Multimodal Interfaces*, 2005.
- [20] W. R. Hahn and S. A. Tretter. Optimum processing for delay-vector estimation in passive signal arrays. *IEEE Trans. Inform. Theory*, 1973.
- [21] S. Handel. *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, 1989.
- [22] J. Hightower and G. Borriello. Location systems for ubiquitous computing. *IEEE Computer*, 34(8):57--66, 2001.
- [23] Owen Holland, John Woods, Renzo De Nardi, and Adrian Clark. Beyond swarm intelligence: the ultraswarm. *IEEE Swarm Intelligence Symposium*, 2005.
- [24] <http://www.fakeisthenewreal.org/nyimade/nyimade.php>. Trace of one person's walking and biking in new york city.

- [25] <http://www.flickr.com/groups/memorymaps/pool/>. Flickr/google memory maps pool.
- [26] Yiteng Huang, Jacob Benesty, and Gary W. Elko. Passive acoustic source localization for video camera steering. *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings, 2*, 2000.
- [27] Yiteng Huang, Jacob Benesty, Gary W. Elko, and Russell M. Mersereau. Real-time passive source localization: A practical linear-correction least-squares approach. *IEEE Transactions on Speech and Audio Processing*, 9(8), November 2001.
- [28] Tristan Jehan. Event-synchronous music analysis/synthesis. *Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04)*, 2004.
- [29] Xiang Ji and Hongyuan Zha. Sensor positioning in wireless ad-hoc sensor networks using multidimensional scaling. *IEEE INFOCOM Conference on Computer Communications*, 2004.
- [30] J. M. Kahn, R. H. Katz, and K. S. J. Pister. Next century challenges: mobile networking for 'smart dust'. In *MobiCom '99: Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 271--278, New York, NY, USA, 1999. ACM Press.
- [31] Thomas Laureyssens. <http://www.pedestrianlevitation.net/process.html> - an artwork in public space, 2005.
- [32] Jeff Lieberman. Intelligent audio segmentation and quantization using dynamic programming - <http://bea.st/text/iq/>.
- [33] Joshua Lifton, Deva Seetharam, Michael Broxton, and Joseph A. Paradiso. Pushpin computing system overview: A platform for distributed, embedded, ubiquitous sensor networks. In *Pervasive '02: Proceedings of the First International Conference on Pervasive Computing*, pages 139--151, London, UK, 2002. Springer-Verlag.
- [34] Andy Lippman and David Reed. Disruptive technologies and opportunities for service providers. In *Viral Communications at IIR Telecoms Transitions*, 2005.
- [35] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the 7th International Conference on Computer Vision*, pages 1150--1157, 1999.

- [36] Anil Madhavapeddy, David Scott, and Richard Sharp. Context-aware computing with sound. *The 5th International Conference on Ubiquitous Computing*, 2003.
- [37] P. Mailloux. *Phased Array Antenna Handbook*. Artech House, 1994.
- [38] Jacky Mallett and Jr. V. Michael Bove. Eye society. *Multimedia and Expo, 2003. ICME'03. Proceedings*, 2003.
- [39] David L. Mills. Internet time synchronization: The network time protocol. *IEEE Transactions on Communications*, 39(10):1482--1493, October 1991.
- [40] David Moore, John Leonard, Daniela Rus, and Seth Teller. Robust distributed network localization with noisy range measurements. *SenSys'04 Baltimore, Maryland, USA*, November 3--5 2004.
- [41] US Navy. *Submarine Sonar Operator's Manual*. Navpers 16167, 1940.
- [42] Edwin Olson, John Leonard, and Seth Teller. Robust range-only beacon localization. *IEEE Autonomous Underwater Vehicles*, 2004.
- [43] Arnaud Pilpre. Self-\* properties of multi sensing entities in smart environments. Master's thesis, Massachusetts Institute of Technology, 2005.
- [44] Nissanka B. Priyantha, Anit Chakraborty, and Hari Balakrishnan. The cricket location-support system. In *Mobile Computing and Networking*, pages 32--43, 2000.
- [45] public. [http://en.wikipedia.org/wiki/instant\\_messaging](http://en.wikipedia.org/wiki/instant_messaging).
- [46] Vikas C. Raykar, Igor Kozintsev, and Rainer Lienhart. Position calibration of microphones and loudspeakers in distributed computing platforms. *IEEE Transactions on Speech and Audio Processing*, 2003.
- [47] Dilip Sarwate and M. B. Pursley. Crosscorrelation properties of pseudorandom and related sequences. 68:593--619, 1980.
- [48] Albrecht Schmidt, Michael Beigl, and Hans-Werner Gellersen. A location model for communicating and processing of context. *Personal and Ubiquitous Computing Journal*, 6(5-6):341--357, 2002.
- [49] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. *Proc. Int. Workshop on ICA and BSS (ICA'99)*, 1999.
- [50] C. E. Shannon. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10--21, 1949.



- [51] GT Sibley, MH Rahimi, and GS Sukhatme. Robomote: A tiny mobile robot platform for large-scale ad-hoc sensor networks. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, 2002.
- [52] Gyula Simon, Miklo?s Maro?ti, A?kos Le?deczi, Gyo?rgy Balogh, Branislav Kusy, Andra?s Na?das, Ga?bor Pap, Ja?nos Sallai, and Ken Frampton. Sensor network-based countersniper system. *Proc. 2nd ACM Conf. Embedded Networked Sensor Systems*, 2004.
- [53] Hing Cheung So and Shun Ping Hui. Constrained location algorithm using tdoa measurements. *IEICE Trans. Fundamentals*, E86-A(12), December 2003.
- [54] Gary L. Sugar and David S. Kloper. System and method for locating wireless devices in an unsynchronized wireless environment. US patent application 20050003828, 2005.
- [55] Christopher Taylor, Ali Rahimi, Jonathan Bachrach, and Howard Shrobe. Simultaneous localization and tracking in an ad hoc sensor network. Master's thesis, Massachusetts Institute of Technology, 2005.
- [56] Max. Van Kleek. Intelligent environments for informal public spaces: the ki/o kiosk platform. Master's thesis, Massachusetts Institute of Technology, 2003.
- [57] Hanbiao Wang, Deborah Estrin, and Lewis Girod. Preprocessing in a tiered sensor network for habitat monitoring. *Journal on Applied Signal Processing*, 2003.
- [58] Roy Want, Bill Schilit, Norman Adams, Rich Gold, Karin Petersen, David Goldberg, John Ellis, and Mark Weiser. An overview of the ParcTab ubiquitous computing experiment. *IEEE Personal Communications*, 2(6):28--43, December 1995.
- [59] E. Weinstein, K. Steele, A. Agarwal, and J. Glass. Loud: A 1020-node modular microphone array and beamformer for intelligent computing spaces. *MIT/LCS Technical Memo*, 2004.
- [60] M. Weiser. Ubiquitous computing. *Computer*, 26(10):71--72, 1993.
- [61] Alex Westner and Jr. V. Michael Bove. Applying blind source separation and deconvolution to real-world acoustic environments. *Proceedings of the 106th Audio Engineering Society (AES) Conference*, 1999.
- [62] Physical Language Workshop. Open attaché <http://plw.media.mit.edu/>, 2005.