ACOUSTICAL AWARENESS FOR INTELLIGENT ROBOTIC ACTION

A Dissertation
Presented to
The Academic Faculty

By

Eric Martinson

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Computer Science

Georgia Institute of Technology

December, 2007

ACOUSTICAL AWARENESS FOR INTELLIGENT ROBOTIC ACTION

Approved by:

Dr. Ronald Arkin, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. David Anderson
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Tucker Balch
School of Interactive Computing
Georgia Institute of Technology

Dr. Frank Dellaert
School of Interactive Computing
Georgia Institute of Technology

Dr. Thad Starner
School of Interactive Computing
Georgia Institute of Technology

Date Approved: November 9, 2007

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my wife, Lilia Moshkina, for the countless hours she spent helping me out in many different ways. She was always encouraging, helped out with the robots when it was needed, was a good sounding board for refining ideas, and was an excellent editor for this dissertation. In general, Lilia was invaluable to this work, and this dissertation would likely not have been completed without her.

Second, I would like to thank my advisor, Ron Arkin, along with the rest of my committee: Dave Anderson, Tucker Balch, Frank Dellaert, and Thad Starner. Ron provided me with the freedom to explore this subject and truly identify a topic of my own choosing, while the committee was enthusiastic and helpful all the way to the end.

Finally, I would like to thank the people at the Naval Research Laboratory Center for Artificial Intelligence. In particular, thanks to Alan Schultz and Derek Brock for their early acceptance of this line of research. Their experience and general support was a great asset towards refining the work and getting it published. Thanks also to Ben Fransen, a fellow intern, for being a patient listener on numerous occasions throughout two summers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

With the growth of successes in pattern recognition and signal processing, mobile robot applications today are increasingly equipping their hardware with microphones to improve the set of available sensory information. However, if the robot, and therefore the microphone, ends up in a poor location acoustically, then the data will remain noisy and potentially useless for accomplishing the required task. This is compounded by the fact that there are many bad acoustic locations through which a robot is likely to pass, and so the results from auditory sensors often remain poor for much of the task.

The movement of the robot, though, can also be an important tool for overcoming these problems, a tool that has not been exploited in the traditional signal processing community. Robots are not limited to a single location as are traditionally placed microphones, nor are they powerless over to where they will be moved as with wearable computers. If there is a better location available for performing its task, a robot can navigate to that location under its own power. Furthermore, when deciding where to move, robots can develop complex models of the environment. Using an array of sensors, a mobile robot can build models of sound flow through an area, picking from those models the paths most likely to improve performance of an acoustic application.

In this dissertation, we address the question of how to exploit robotic movement. Using common sensors, we present a collection of tools for gathering information about the auditory scene and incorporating that information into a general framework for acoustical awareness. Thus equipped, robots can make intelligent decisions regarding control strategies to enhance their performance on the underlying acoustic application.

CHAPTER 1

INTRODUCTION

Audition on mobile robots has long been passed over in favor of vision, the argument being that if we could only decipher an image, then vision has all of the data necessary for highly successful navigation. But the proponents of audition have been successfully reversing this trend in recent years by arguing that there is a wealth of information available to the robot outside the narrow confines of a camera's view-space. If nothing else, the omni-directionality of incoming acoustic information can be used to direct more data-rich directional sensors to intriguing or suspicious locations. Beyond that, researchers are also adding microphones to augment human-robot interfaces [Fong et al. 2003], improve security [Huang et al. 1997], localize themselves [Martinson and Dellaert 2003; Hu et al. 2006], and a variety of other applications.

For robots, audition is a relatively young field. Elsewhere, however, it is by no means understudied. Electrical engineering and digital signal processing (DSP) have made great strides over the last 30 years in using static mounted microphones, hand-held microphones, and microphone arrays. It is a great testament to their success in areas such as speech recognition, classification, and source localization that roboticists are now considering equipping their mobile platforms with these sensors. But as researchers have discovered with other sensory modalities, microphones on robots constitute a different problem than other microphone scenarios.

Where traditional microphone mountings have often been subject to environmental interference ranging from ambient noise, to high and low frequency echoes, and overlapping sound sources, robots add their own set of problems to the list. Many of the techniques developed to counter these problems, including filters, *do* work on mobile robots, but they are not as successful when the platform: (1) moves around the environment, changing its proximity to different sources; (2) generates its own noises, wheel and motor, which vary with the executed action; and (3) has limited computational and power resources, but needs to process the data in real time. These inherent problems of mobile robotics, combined with the general problems associated with using microphones, produce daunting obstacles confronting the developers of acoustic applications for these platforms.

Mobile robotics though has unique advantages all its own, which have not been exploited in the traditional signal processing community. The key advantage is that robots can move. They are not limited to a single location as are traditional microphone mountings, nor are they powerless over to where they will be moved as with wearable computers. If there is a better location for performing their task, they can navigate to that location under their own power. Furthermore, we are not limited to a single robot. Robot teams add extra dimensions of control, by allowing fully dynamic microphone arrays that are not limited by a rigid internal structure, nor stuck in randomly distributed locations. The potential that mobility alone adds to acoustical applications is enormous, but we first need to figure out how to best exploit that potential.

In this work, it is our supposition that acoustical awareness is the key to successful development of mobile robotic applications involving sound. Acoustical awareness is defined here as the coupling of action with knowledge about the acoustic environment, where said knowledge could be in the form of maps, rules, measurements, predictions, or anything that indicates how sound flows or will flow through the environment. The underlying premise is that the more acoustical knowledge the robot uses, the better its global performance will be on an acoustic application. The questions of how much knowledge is necessary, and how it is to be integrated into the robotic controller are central to the proposed research.

1.1 TERMINOLOGY

Acoustics is defined as "the science of sound"[1]. Both auditory (listening) and sound generating applications are acoustic, because they work with sound. In the grand picture, the two areas differ by focusing on either the receiver or the source. In either case, the same principles of sound propagation through an environment apply, and an acoustically-aware application would require much of the same information.

Acoustical Awareness is the coupling of action with knowledge about the acoustic environment, specifically anything that indicates how sound flows or will flow in the physical world.

Audition refers to the act of hearing. Like cameras with vision, microphones are the instruments we employ for recording sound, defined as "the mechanical energy transmitted by longitudinal pressure waves in a material medium (like air)."[2] As with computer vision, however, we are not limited to what can be sensed by people. While humans can hear sounds in a frequency range from 20-20000 Hz, microphones can be used to "listen" to much lower or higher frequencies.

---

1 Raichel, D. p.4
2 http://www.webster.com, Accessed 8/29/04

The *Auditory Scene* contains all aspects of that which effect what a listener somewhere in the environment can hear. It includes the sound sources generating the noise, the environment through which the sound travels, and, ultimately, the listener itself.

Noise is an application-specific term referring to unwanted sound. Even if the sound is desired or needed for some other purposes, but is interfering with the intended application, it is called noise.

The *Soundscape* refers to that which can be heard. Although often used interchangeably with the term *Auditory Scene*, the soundscape is a narrower definition, referring specifically to what can be heard at any location in the environment, independent of the listener.

Vocalization refers to the creation of sound by the robot. The emitted sound could be speech, or just noise. It could target either human or robot listeners, or may not target anyone, as does the incidental creation of noise which often accompanies mechanical motion.

1.2 RESEARCH QUESTION

> *How can acoustical awareness be effectively incorporated into a navigational controller?*

In exploring this principal question, we address the following set of three subsidiary questions:

- **What a priori information or sensory data is useful for a mobile robot performing an acoustic application?**

  Acousticians have developed a large body of research on the flow of sound. Potentially everything about the environment, ranging from construction material performance and architectural features, to speaker and microphone models, is useful for a mobile robot. However, not all of it is feasibly acquired, much less usable, given the limited computational, or acoustic processing resources onboard the robot. Therefore, we need to determine which information can be reasonably collected for, or by, a mobile robot, and whether or not that information is usable in a given situational context.

- **How can we combine sensory data from multiple sources to build effective representations of the acoustic environment?**

  Using just a single microphone located on the robotic platform provides a wealth of sensory information available for assisting navigation: sound pressure level, the frequencies present and their loudness, impulse noises vs. continuous streams, classification results, etc. Beyond that a priori knowledge of the environment, other sensory data such as vision, other microphones, or even data from other robotic platforms might be available. Somehow this data needs to be fused together to build effective and coherent representations of the acoustic environment in which the robot resides.

- **How does acoustical awareness change with control over the source vs. the receiver?**

  Vocalization and audition vary only in their control over the sound source. In vocal applications, the robot itself is the source. In auditory applications, the robot is the receiver. In some instances, a robot may need to be both. In many cases, both forms share the same goal, which is to control the sound being heard by the receiver (human or microphone). How these goals are achieved may also be similar across vocalization and audition.

1.3 CONTRIBUTIONS

It is believed that the pursuit of answers to these questions should result in significant contributions to the mobile robotics community, with application to other fields including signal-processing, acoustical engineering, and human-computer interfaces. In particular, this research will provide:

- A conceptual and architectural framework for incorporating acoustical awareness into a navigational controller.
- A novel approach for the storing, retrieval, and fusion of acoustic knowledge for mobile robotic applications.
- Guidelines for applying the resulting framework to acoustic applications. In particular, matching the data available to the task at hand given a particular situational, intentional, and environmental context.

1.4 DISSERTATION OVERVIEW

This dissertation is described with 8 chapters. Chapter 2 covers the types of acoustical applications for robots, and what work has already been done. Chapter 3 discusses the nature of being acoustically aware, identifying what information is needed to understand sound flow in the environment, and how it will be used. Chapter 4 then delves into the robotic question of representations and control for acquiring the necessary information using a mobile platform. Chapters 5-7 describe applications of acoustical awareness to different robotic application domains, including robotic security, stealth robots, and human-robot interaction. Finally, Chapter 8 summarizes the dissertation and outlines the contributions from this dissertation.

CHAPTER 2

RELATED WORK

Acoustic applications on mobile robots, while increasing in number, are still relatively rare. But with the ready availability of microphones, and the increasing processing power of computers and even microcontrollers, the area is primed for an application explosion. One sure sign of this potential is the growing interest in acoustical domains by robotics hobbyists. Voice commands [Williams 2004], noise following [Predko 2003], and sound synthesis[Jones et al. 1999], are all popular applications in the area. Still, the work by hobbyists tends to be overly simple algorithmically. But if enough people become interested in the area, then we will see microphones and speakers on robots become commonplace.

What is making acoustical applications in robotics difficult is the underlying complexity of the acoustical domain. The soundscape is always changing with time, more so than even the visual domain tends to, and the sensors currently available for sampling the soundscape are noisy and only capture a relatively small selection of the soundscape. Additionally, the soundscape itself is not straightforward, and varies significantly from environment to environment, even when the same types of noise sources are present. Altogether, this creates a very hostile perceptual domain for a robot, which is already struggling to successfully handle routine navigational tasks.

In order to overcome these problems in the acoustic domain, much research in mobile computing (including robotics), as well as biology, and digital signal processing, has concentrated on developing task specific solutions that take advantage of the nature of sound flow through the environment in order to improve performance. This chapter presents the current state of the art in robot acoustics. The general organization of the chapter is by broad application categories explored in current research: specifically, sound source localization, natural language interfaces, classification, vocalization, and audio probing.

2.1 Sound Source Localization

Sound source localization applications are primarily concerned with identifying where a sound is coming from, including determining its exact position, or simply the angles to the source(s), or distance estimates. Most sound localization work is not even concerned with what is making the sound, and in fact, a common laboratory assumption is that whatever sound is being heard is the one that the robot is interested in. This category is probably where the greatest amount of acoustic research on mobile robots has occurred to date.

The localization problem can be roughly divided into two areas of interest. The first problem is localizing individual sources in the environment. The second lies in creating maps of sound sources, possibly for a robot to localize itself using a set of detectable noise sources. In either of these areas, an awareness of sound propagation through the environment may potentially assist improving accuracy, reducing false positive responses, and improving the general applicability of the developed algorithms.

2.1.1 Traditional Localization

Localizing sources in the environment may not seem like a difficult task to a person, but that is because the mechanisms for doing so are built into a humans' auditory system. People are equipped with two ears for the purpose of localization, as are most mammals and birds. With two receivers (ears) physically separated from each other on opposite sides of the head, sound arrives at different times, phases, and intensities in each ear. These inter-aural time differences (ITD), phase differences, and intensity differences (IID) can be used to calculate left-to-right angular location of the sound. There are different models of exactly how these features are calculated biologically [Jeffress 1948; Shamma 1989], but these are the physical properties available in the incoming sound stream, and both models make use of them.

If there are three or more receivers, then ITD's alone can estimate the angle of incidence for an arbitrary source location, and this has in fact been used on a number of robots [Yamasaki 1995; Huang et al. 1997; Young and Scanlon 2001]. With only two receivers, however, ITD's can provide only an 180° estimate on the horizontal plane towards the location of the sound. Localizing on front-to-back or elevation is not possible without additional information. The biological solution is found in the shape of the head. Sounds traveling around the back of the head arrive at different times and intensities than when they come from the front. The pinna, or fleshy parts of the ears, also filter or focus sounds depending on which direction they come in from. The shape of the head and the pinna make up the Head Related Transfer Function, or HRTF. The HRTF is a function, different for each individual, which the brain learns in order to pick out localization cues that would otherwise be hidden. Although the use of an HRTF does not guarantee the same resolution at all angles around the head, it does effectively localize sounds from any direction and has been applied in humanoid robotics [Nakadai 2003; Hornstein et al. 2006].

Noise, however, remains a problem with ITD-based solutions. Environmental noise, as well as robot-generated ego noise, can generate misleading measurements or mask the signal of interest. A common feature, for this reason, among many robotics solutions is the use of higher quality microphones with frequency ranges that limit interference, and to mount them high above the robot base, away from motors and other noise-producing equipment. Even then, the noise can cause problems. Furthermore, on some platforms and with some applications, it is not feasible to deploy expensive microphones far from sources of robot ego-noise. Humanoid platforms are a case in particular. Most current humanoid robots are limited to inexpensive microphones mounted within the head, close to the internal machinery of the robot.

A sound source localization solution that has been put forward by a couple of different robotics groups is to physically move the microphones. Barbara Webb [Webb 1998] explored this approach using small robots with 2 microphones. The robot would always turn in the direction of the highest volume, and although the path was not entirely straight, the robot could find the sound source when echoes or obstacles did not interfere. Another more traditional solution using ITD's, was conducted by Nakadai et. al. [Nakadai 2001], which would turn their humanoid robot's head in the direction of the loudest sound as estimated by ITD's. By doing so, the robot could effectively ignore the internal noises when they do not change direction with the rotation, allowing the robot to accurately focus on the noise source.

In addition to robot and environmental noise, another problem confronting real-time implementations of sound localization algorithms is that the accuracy of ITD's is limited in highly echoic environments. If an echo bounces off the nearby floor, or wall, then separating it from the real signal is very difficult, and leads to errors in localization. Unfortunately most indoor environments (especially hard floored ones), unless specially padded, have this echo problem in one place or another. The most common solution in robotics is to add another sensor with a different vulnerability. Cameras are the obvious alternative, as a microphone can assist in orienting a camera towards the object of interest [Huang et al. 1997; Strobel 2001; Blisard et al. 2007], although heat sensors have also been used when applicable [Yamasaki 1995]. Noise remains a problem with the auditory sensing, but its use has been restricted to initializing the application.

Without adding another sensor, another solution that has worked well in echoic environments is to make use of the precedence effect. The precedence effect is a psychoacoustic phenomenon where an acoustic signal arriving first at the ears, suppresses the ability to hear any other signals for the next 40ms, in order to reduce the interference with echoes. Initial work by Jie Huang [Huang 1997] generated robots that were successfully able to locate visually obstructed sound sources in different rooms using ITDs, but still resorted to vision as the primary sensor once the object became visible. More recent work by Martin Heckmann [Heckmann et al. 2006], however, has improved on this early work for binaural microphone arrays to dramatically reduce the need for the visual localization on their Honda robot, Asimo. Their model of the precedence effect combines interaural time, intensity, and envelope differences with an echo suppression mechanism to allow Asimo to localize a human speaker under a large variety of ambient noise and environmental conditions.

2.1.2 CREATING AUDITORY SPATIAL MAPS

Up until this point, we have been assuming that the sound localization algorithm returns angular measurements to the robot. While, theoretically, ITD-based algorithms could also estimate distance from the robot, the closer together the microphones are, the greater the error in the distance measurement. People, too, have a similar problem in estimating distance for similar reasons (the ears are too close together). Nevertheless, if there are multiple sound sources in a room, people are reasonably good at localizing them in space, if not instantaneously, then at least over time. To explain how people and animals can be good at this despite the signal processing limitations, some researchers in biology have proposed the notion of an auditory spatial map that maps multiple sensing modalities in the brain into a single ego-centric representation of stimuli position.

The barn owl has one of the best-studied auditory systems of any species. Researchers of the Owllab at the University of Oregon have, as stated on their webpage, concentrated on "studying the neural mechanisms of auditory localization in barn owls", addressing questions such as how the owl localizes, how it reacts to multiple sound sources [Takahashi and Keller 1994], what the HRTF [Keller et al. 1998] is, etc. What they have discovered is that within the inferior colliculus of the brain, individual neurons become attached to specific locations in the surrounding environment, only firing when a noise is determined to have originated from that location. These neuronal spatial maps, however, are not being constructed from auditory information alone. Hyde and Knudson at Stanford University claim that the Optic Tectum (OT), which directs the eyes toward sensory cues, is critical for the construction of an auditory map [Hyde 2000] . Their studies with barn owls have indicated that without the OT, the owl's brain is no longer capable of updating and aligning the separate maps of space. Hyde and Knudson have also constructed a general framework for the calibration process that applies to both avian and mammalian species [Hyde 2000].

Rucci, Tononi and Edelman [Rucci et al. 1997] at the Neurosciences Institute in San Diego have developed an alternative model specifically for the barn owl, arguing that the existing model does not include enough information. Since the barn owl cannot see without turning its head, they argue that the model should actually be a sensorimotor model. The system takes as input ITDs and visual information to construct the visual and auditory maps. The model developed is neuronal and is trained using "value-dependent learning" and tested in simulation under a variety of inputs.

While this neuronal spatial map of the auditory scene suggests a tantalizing approach for mobile robotics, it should be understood that this is not the only biological model of the auditory scene. In some animals, including people, recent research has failed to reveal any notion of spatial maps [Goldstein 2007]. In some other animals, such as guinea pigs, maps have been found at an early age, but seem to disappear as they get older [Ingham et al. 1998]. This does not invalidate the idea of an auditory spatial map, as the barn owl and the guinea pig have very clearly demonstrated the existence of such maps in some species, but it does suggest that there may be more than one way to track or at least store this information in our brains.

In robotics, however, the idea of localizing all of the sound sources in space relative to the robot has more often been developed as a straightforward extension of the earlier angular sound localization problem, rather than as a mapping problem. To alleviate the problem of microphone proximity in determining distance, the natural solution is to distribute the microphones over a wider area of the environment. Using a microphone array embedded into the walls of a room [Nakadai et al. 2006], some researchers were able to localize multiple simultaneously talking human speech sources, and determine the directivity. This information can then be passed along to a mobile robot for interactive or avoidance purposes. A similar setup in a home environment [Bian et al. 2005] extracted 3-dimensional coordinates of sound sources and estimated the type human activity occurring. Another possibility, when the environment cannot be engineered, is to place the microphones on separate robots [Girod and Estrin 2001]. As the robots separate in space, differences in arrival time become more pronounced, so distance measurements should ideally become more accurate. But for this technique to work the audio streams from the separate microphones need to be synchronized to better than millisecond accuracy. For each millisecond of error in synchronization, roughly 34-cm of error are introduced in the sound source location. Furthermore, the robots need to be localized accurately in space (overhead cameras and/or laser-based obstacle maps), relative to each other, or the time-delay measurements will be incorrectly determined, even with accurate synchronization.

While research into accurate localization of the sound sources remains an active field, there is also work progressing in the opposite direction. If the robot has a map of where the sounds are as it moves through the environment, then the robot could localize itself using the map and the sounds that it can currently perceive. Such work on the surface seems very similar to other landmark-based navigation strategies using databases of vision, laser and/or sonar measurements [Thrun et al. 2001]. In practice, there have been some difficulties in applying the same technique to the auditory domain. Most similar to the idea of landmark-based localization is work by Jwu-Sheng Hu with a Sony AIBO robot [Hu et al. 2006]. Rather than using existing environmental sources, however, the AIBO emits the sound being tracked, a barking sound, and then records the results. Combined with a database of pre-recorded barks from different parts of the environment, along with a history of the robot's movement, they can estimate position and orientation of the legged platform. Also utilizing robot emitted sounds, work by Dellaert et. al. [Dellaert et al. 2003] used distances to other robot noise sources to create maps of where robots had traveled through the environment. Further work by Sebastian Thrun [Thrun 2005] extended a similar methodology to dynamically localizing microphones in an array using a series of easily recognizable impulse noises (finger snaps, clapping, etc.) in the environment.

While all of these probabilistic or landmark-based approaches could potentially be applied to the general mapping problem around passive environmental sources, all of them suffer from a number of problems when applied to a real and naturally occurring auditory scene. The transitory nature of sound sources is one problem, making comparisons between old and new data difficult. Another problem is environmental echoes, which produce large defects in even small maps. To overcome this problem, the robot needs a representation that still has some meaning in the environment despite some changes to the auditory scene. This representational issue is exactly that the problem on which this thesis is focusing. As such, we will discuss this matter in further depth in the next chapter.

2.2 Natural Language Interfaces

Natural Language Interfaces refer to speech interfaces between computers/robots and humans. This could entail the robot speaking to a person, or the person speaking to the robot, or both. People are so used to communicating with each other by speaking that it is only natural that we would want the same interface for communicating with our robots. The argument is still open, however, as to how much of a speech interface is actually necessary. Most would agree though that many areas of robotics that require human-robot interaction would benefit from a real-time speech interface.

The domain of natural language interfaces can be roughly divided into two parts: listening, and speaking. Both parts are necessary for a full interface, but current research is still a long way from a completely integrated solution. As will be discussed in both sections, problems that repeatedly trouble natural language interfaces are the quality of the detected or generated speech, and the effects of masking noise on intelligibility. These are often environmental effects that could be minimized by an acoustically-aware robot utilizing knowledge of sound flow. Chapter 5 will discuss such an application.

2.2.1 The Robot Listens

With the increased use of Hidden Markov Models (HMM), speech recognition rates have improved dramatically over the last decade. Just within the last several years, software has become readily, even freely, available [Lamere et al. 2003] for anybody to add real time speech recognition to their computer application. More and more roboticists are in fact, starting to do just that. The general approach that works best divides all human speech into 40-50 similar sounds, called phonemes. Each word in the

English language can be represented as an ordered set of these phonemes. Using measured transition probabilities between phonemes as input to an HMM, we can identify which phonemes the speaker pronounced, and try to reconstruct the spoken words from those. Russell and Norvig [Russell and Norvig 1995] estimate 80-98% accuracy in the best speech recognition systems, depending on the length of the input, the size of the vocabulary to be recognized, the variety of speakers, and the signal quality.

On robots, however, the word recognition rate tends to be lower than that achieved using just a microphone. Just as with sound localization, the internal noises from the robot and the variety of environments they can be located in can cause problems for robust audition. Sometimes, however, very simple interaction techniques can be used to compensate for the lower recognition rates. One such method, used on the robot HERMES [Bischoff 2000], was to ask the user for confirmation of the spoken command when the recognition results were poor. More complex solutions have combined the spoken command with visual cues, like pointing or other gestures that the human can easily perform [Perzanowski et al. 2000; Kettebekov 2002]. Both of these solutions work best with a limited command set.

Beyond recognizing a limited vocabulary, the next step involves actually understanding the semantic meaning of the speech, so that the computer/robot can respond appropriately. Just because one can recognize words does not imply natural language understanding. Russell and Norvig [Russell and Norvig 1995] state the following regarding natural language understanding :

"We are given a set of ambiguous inputs, and from them we have to work backwards to decide what state of the world could have created the inputs".[3]

This is a very hard problem, and one that we are not close to solving in the general case. However, one technique that has been employed effectively on a number of robots [Roy et al. 2000; 2004]. Often called something like dialog-driven interaction, it implies a script that both the robot and the human are expected to follow. The person says one thing, the robot recognizes that it is the next line in the script, and then "says" its part. Using grammatical formalisms, some variability in the allowable words can be introduced to the script, but the programmer has to anticipate what sentences a human user might respond with, and the human has to stick to the script without changing topics.

Outside of natural language understanding, there is a second problem involving computer audio that is on the DSP side of speech recognition. Stream segregation is the ability to separate two or more different speakers from each other, and from background noises, so as to identify which person is saying what. When multiple people are speaking, the robot/computer has to decide which person to respond to, and remove all of the excess information. Due to the nature of sound propagation, all of that speech information from everyone in the room is normally combined in the stream picked up by each receiver. Unless somehow processed to separate the speakers, that stream is unintelligible as far as computer speech recognition is concerned. This is classically defined as the cocktail party effect [Bregman 1990]. Although the full solution is still

[3] Russell and Norvig, p 654

beyond our abilities, there do exist partial solutions. When the number of talkers is relatively few, independent component analysis (ICA) has demonstrated reasonable performance at separating multiple speech streams and overcoming their masking effects using a binaural microphone array [Takeda et al. 2006]. Traditional ICA, however, is easily confused by reverberant environments and can have difficulties when the people speaking are located too close together. A variant of ICA that works with only a single microphone to separate speech streams [Smaragdis 2001] has partially overcome these difficulties, by focusing on only the speech related cues. This method still suffers in the presence of reverberation, however, as it now includes reverberant effects in the resulting speech stream, making speech recognition difficult. In the case of either method, they remain computationally expensive, and are still limited by the length of audio segmented, the number of speakers or sound sources in the area, and some environmental effects.

An alternative solution to ICA, is to incorporate the geometry of the speaker's location into the algorithm. In the simplest case, where only one speech streams needs to be recognized at a time, a computationally simpler approach mentioned earlier [Nakadai 2003], is for the robot to rotate to face one speaker, and use directional microphones to amplify only that person's speech. Another partial solution based on geometric locations does the same without involving any motor actions by using a microphone array [Claudio and Parisi 2001; Argentieri et al. 2006]. If where the people speaking are located can be identified, then one can mathematically construct and amplify a narrow beam from the streams of multiple microphones. Recent work by Valin successfully applied this approach on a mobile robot, for separating up to three simultaneous speech streams [Valin 2005]. Unfortunately, these geometric-based approaches still have difficulties when the two people speaking are located too closely together, or there is significant reverberation from nearby surfaces. In that case, the speech streams remain too intertwined for a location-based solution to separate, and may require repositioning the microphone array intelligently with respect to the auditory scene to change the relative angles to the speakers, or somehow limit the effects of reverberation.

2.2.2 THE ROBOT SPEAKS

If no speech recognition is required, then robot speech can be a surprisingly easy concept to implement. The simplest systems use pre-recorded sound bites that are activated by a robotic behavior. A human can provide the recordings, and the robot only needs to interject them at the correct point during its human-robot interaction. How the robot chooses the correct sound bite could be in response to a single sensory stimuli like a clap [Jones et al. 1999], or it can even be learned by the robot. Minerva [Schulte et al. 1999], a robotic tour guide, adapted to the environment by monitoring how its voice commands affected the people densities around it.

An alternative approach to pre-recorded audio is text-to-speech (TTS). First, the text is converted to a phonetic code. Next, pitch, intonation, pausing, and rate are added. Finally, the audio is created from the parameterized phonetic code either by grabbing phonemes from a database, or synthesizing speech from basic phonetic units [Venkatagiri 2003]. The resulting speech is clearly understandable, and the newer software available continues to improve in quality. TTS was used on GRACE [Simmons et al. 2003] for the AAAI robot challenge to deliver a talk by the robot at the conference. The greatest advantage of TTS is that the speech can be changed on the fly. Work in affective robotics (i.e. emotions, moods, etc.) has demonstrated the advantages of dynamically adjusting the pitch, pausing, and rate of the synthesized speech (i.e., the robot's prosody) to give a robot personality [Breazeal 2001; Scheutz et al. 2006]. Work in more traditional human computer interfaces [Dusan and Flanagan 2002] changes not just the quality of the speech, but adapts the text input itself sent to the synthesizer by constructing grammars from human speech.

Still, TTS systems are not without their problems. One persistent issue is environmental noise. Although commercial systems are very good in quiet environments, TTS in noisy surroundings can still be hard to understand [Venkatagiri 2003]. The studies on this subject to date, however, have been performed using isolated phrases. With some context behind the phrases, people may be better able to understand the current systems. The problem of context though, is really a full interface problem. Both speech recognition and speech synthesis are necessary to provide the human participant with context. Unfortunately, the two problems are unequally advanced in achieving human-level performance. A study of natural language interfaces at Eurospeech '97 [Bloothooft and Os 1997], suggested that the existing speech recognition capability was holding interfaces back. Some users of the latest systems were at times completely fooled by the accurate TTS, forgetting that they were talking to a machine. But then when speech recognition failed, people fooled by the TTS system took longer than other participants in recovering from the error.

2.3 AUDIO CLASSIFICATION

Passive sound classification is the problem of recognizing sounds and/or indexing sounds in the sound stream. Detection of air ducts, computer fans, machinery, a human voice, music, etc. are all examples of sounds that a robot may encounter in the sound stream and that it may be expected to recognize. To go even further, if we can recognize all of these factors, can we recognize the type of the environment itself? From the robotic perspective, this is a potentially huge field, but can be limited by the sheer processing power available onboard a mobile platform.

2.3.1 ENTITY CLASSIFICATION

The oldest studied problem in audio classification is simply recognizing the start of the interesting signal. The voice/unvoiced/silence problem was the earliest classification problem developed for the telephone industry. The majority of the time in phone conversations is actually silence, or lack of audio content. So if the industry could separate silence from content, they would not have to transmit the entire signal across their lines, therefore increasing the number of calls they could handle simultaneously given a limited bandwidth. To compress the signal even further, they needed to also distinguish between voiced and unvoiced sounds. Because people are so good at recognizing speech, the telephone companies wanted to automatically recognize a minimum signal that could still be understood by the listener, so as to remove all of the excess auditory information from the stream resulting in an extremely compressed signal. The most well-known solution was developed by Atal and Rabiner [Atal and Rabiner 1976]. This approach detects vowel sounds in an audio stream. Later research showed that provided the vowel sounds were present, human listeners could recognize with great accuracy the words in the conversation even when much of the remaining signal was removed. The resulting algorithm, called Linear Predictive Coding [Rabiner and Schafer 1978], compressed the audio stream and reconstructed the stream.

Much of the work in the Voiced/Unvoiced/Silence problem led to useful sets of features that could be extracted from an audio stream and applied to the more general classification problem, like mel-cepstrum coefficients [Quatiri 2002]. Ideally, we would like to classify any type of audio that is put before the computer, but it turned out that successful classification is highly dependent on the mathematical features extracted from the audio sample. Incorrectly chosen features will allow different audio samples in different classes to appear similar [Duda et al. 2001]. For example, water leaks can be detected in pipe networks by listening for certain frequencies (GMIC) [Hetek 2004]. However, the robot needs to know that it is near a pipe network and should be listening for these leaks. The same is true for identifying bird song [Kogan and Margoliash 1998], where the robot has to be told when to start listening for birds. For this reason, classification approaches currently have to be application specific, first choosing a set of classes, and then finding a good feature set and algorithm.

Although classification on robots in general, outside of speech recognition, has remained relatively unexplored thus far, research is already exploding in related mobile computing applications. Current classification applications using cell phones [Philips 2004] and wearable computers [Stager et al. 2003; Lukowicz et al. 2004] may in the long run be applicable to robotics and autonomous control.

2.3.2 AUDITORY SCENE ANALYSIS

One reason for the successes achieved in silence detection involved the telephones themselves. With a telephone, the microphone is placed right next to the person's mouth where the sounds are loudest. As a result, the microphone itself does not pick up much of the background noise. If forced to use a different microphone, however, which picked up more noise, then the classification problem becomes substantially harder. How can an arbitrary noise stream be separated from the "good" stream? What is needed is knowledge of the auditory scene, or contextual information regarding the signal currently being recorded. For instance, if the robot knows it is outside rather than inside, then it can use different algorithms to recognize the start and stop points, and use different filters on the incoming data.

This problem is called auditory scene analysis, and people do this all the time. They listen for different types of events given different auditory scenes in which they are located, and in general, people are relatively good at recognizing the different scenes. Given the task of recognizing the surrounding environment from an audio clip, out of context, people could identify the class of surroundings 70% of the time, and often in 20sec or less [Peltonen et al. 2002]. In comparison, a machine given a much smaller set of classes took almost 2 minutes to identify a much smaller, 5 vs. 25, set of classes. An even smaller set of classes, however, may also be useful in providing some auditory scene analysis. Work by Christophe Couvreur [Couvreur 1998] using HMM's, has attempted to classify environmental audio into a set of 5 categories using transportation-related sounds. It is too small a set of categories for solving the general segmentation problem, but such a set could be useful to a robot which is exposed, as part of its job, primarily to transportation-related auditory scenes. In such a case, transportation sounds could still provide some useful auditory context for identifying the current scene and improving general classification results.

Dividing the world into classes for the machine to recognize is not the only solution to the auditory scene analysis problem. Work from Carnegie Mellon uses a wrist based light and audio sensor to detect places the wearer had been before [Maurer et al. 2006]. Alternatively, another method has been to recognize distinct sounds within the sound stream, and then derive auditory context, or the auditory scene, from those sounds. A good example is, if you hear a car horn, you are probably on the street [Clarkson et al. 1998]. Another, more hardware intensive solution, borrowed a wireless sensor network to assist the traveling microphone/processor. If the sensors know where they are, then they can broadcast this information, providing both context and localization information [Schiele and Antifakos 2002].

2.4 AUDIO PROBING

The field of audio probing includes anything that requires an action by the robot to receive some form of audio feedback. Ultrasonic sensing is one example of audio probing. The robot emits a high-frequency sound click into the environment, and waits for the echo of the sound bouncing off objects in its path. It works on the same principle as bat echolocation, providing the robot with estimated distances to hard surfaces around it. Ultrasonic probing has been used for a variety of purposes including mapping, classification, and, of course, obstacle avoidance. For a more detailed description of how ultrasonic sensors have been used in mobile robot navigation see [Arkin 1998].

Ultrasound, however, is only a very simple form of audio probing. It takes the reflected sound pulse, and translates it into a single unit of distance. More complicated forms of audio probing audio probing can extract different information from the returned sound. For instance, by generating a loud impulse noise at some frequency, computers have already been able to estimate acoustic properties such as reverberation time in indoor environments [O'Keefe 1998]. Perhaps more compelling for robotics is the use of tools to generate noise. By using metal poles to hit surfaces, robots can try to classify the materials used [Krotkov 1995; Femmam et al. 2001], and/or identify other structural characteristics of the object being probed [Amsellem et al. 2006].

More complex forms of audio probing however, such as the classification of material surfaces, tend to have the same problem as classification algorithms. The advantage in audio probing is that there exists a little more context for completing the application. The object being classified is often in a known location, with suspected material properties. However, the breadth of the classification algorithm is still quite limited. In the general case, either more context is needed to narrow the range of possible results, or the classifier needs to be retrained for each application.

2.5 SOUND VOCALIZATION

The final application area is that of sound vocalization. For statically located speakers, the vocalization problem is a highly specialized field. The goal is to provide the best acoustical experience for the listeners, whether they are in a concert hall, or sitting at home in front of the television. Toward that end, researchers have generated a wide number of acoustical parameters and simulation technologies that can assist the creation of excellent acoustics in a variety of spaces. On robots, however, sound vocalization research has remained more limited. Typical robot audio today consists of prerecorded sound bytes and text to speech, played back for the user when and where the program says. Some robots even do this automatically. For example, ActivMedia's AmigoBot plays sounds on startup and shutdown, and Sony's AIBO can beep to indicate current conditions on the robot. Proper sound vocalization, however, should also involve sensing the environment. Humans, for example, will actually adapt their speech, and their singing or music to overcome environmental noise, or to tone down the speech when they are disturbing others. Good vocalization involves analyzing the environmental feedback to determine how the sound output should be modified for the greatest effect.

Sound vocalization research involving robots has largely been limited to two primary areas. The first area is natural language interfaces, which were previously discussed, and which have not traditionally involved adapting the volume, or tone of the generated sound. The second area is in music generation. Waseda University in Japan has been a hotbed for musical robots. Under the guidance of Makoto Kajitani, a series of WAM (Waseda Automated Manipulator) robots that could play pre-programmed routines on the piano were developed through the 70's. A full humanoid, the Wabot-2 [Waseda 2000], could play a keyboard while following a simple score using a camera for input. Several MUBOT's [Katijani 1989], or musician robots, were also developed which could play recorder, violin, or cello. The latest work has been on a Flutist Robot that could not only play the flute, but also perform trills and vibrato on its instrument [Waseda 2000].

The original Waseda robots, however, did not actually adapt to the sounds they were playing. They figured out what note should be played when, and used planning to reach the necessary keys in time. ISAC however, developed at Vanderbilt University [Alford 1999], actually adapted to the sound being played to improve performance on a Theremin. By listening to the output from the instrument, ISAC could achieve perfect pitch using a reactive control system and could play a sequence of notes in time with a human keyboardist. More recently in the same direction, a robot flutist has been added to the band that listens to its pitch and changes its blown air speed [Isoda et al. 2003].

2.6 SUMMARY OF APPLICATION DOMAINS

What this collection of work should demonstrate is that the need for robot acoustics is already large, and only getting larger. Already, robots need to utilize sound for a large variety of applications, many for interacting with people, but also for improving quality in manufacturing and sensing, and even localizing the robot itself. In nearly all of these cases, however, the role of audio is limited by the complexity of the domain itself. As such, designers spend a great deal of effort building filters, and adaptive algorithms to remove the noise. For the remainder of this dissertation, we are going to focus on an alternative to filtering and noise removal. That alternative is acting with respect to the auditory scene. An acoustically-aware robot can utilize knowledge of sound flow to recognize where the noise is coming from and minimize its effects on performance.

CHAPTER 3

ACOUSTICAL AWARENESS

The concept of awareness has many definitions (see below[4]). Typically, though, it implies some knowledge of the surroundings at a conscious level. This could be specific knowledge about a particular object [Dourish and Bellotti 1992], or it could be raw data which generates an appropriate response [Kaelbling and Rosenschein 1991]. More specifically, however, being aware suggests an explicit recognition and understanding of the meaning behind what is sensed. Not only that, but once the recognition and understanding are in place, being aware means being capable of acting upon that sensed information. That is how being aware is different from simply recording the data, analyzing it, and filing it away for other truly aware beings to use.

American Heritage Dictionary[4] Having knowledge or cognizance. Vigilant; watchful.	L. Kaelbling and S. Rosenchein A tight coupling between sensing and action [Kaelbling and Rosenschein 1991]
P. Dourish and V. Bellotti “An understanding of the activities of others, which provides a context for your own activities.” [Dourish and Bellotti 1992]	M. Endsley “The perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” [Endsley 1988]

[4] The American Heritage® Dictionary of the English Language, Fourth Edition Copyright © 2000 by Houghton Mifflin Company. Published by Houghton Mifflin Company.

Awareness means acting with respect to the knowledge of the surroundings. Acoustical awareness, therefore, is coupling action with knowledge about the acoustic environment, specifically anything that indicates how sound flows or will flow in the physical world.

3.1 Types of Awareness

The general definition of awareness, as suggested by the previous definitions, may be too broad for our use. As awareness is defined, it can be applied to any application that perceives the environment, or has prior knowledge of it, and makes control decisions based on that information. But there are at least two different levels of being aware. First, there is simply reacting to a stimulus from the environment. Most of the robotics applications discussed in chapter 2 fall into this category. Second, there is understanding the context and the nature of the perceptual stimuli, and then making an intelligent, informed decision based on that knowledge. In robotics, this has classically been termed as reactive versus deliberative control, and both should have their place in a real acoustically-aware system.

3.1.1 Reactive Acoustical Awareness

Common acoustics applications in robotics today mostly fall under the category of reactive acoustical awareness. Simply stated, all reactive processes can be defined in terms of connections between 4 parts. At any time t, the signal (d_t) from the acoustic perceptual hardware is transformed by some perceptual software processes, referred to here as perceptual schemas (P). Then the perceptual schema results are fed into a motor schema, or behavior (B), which transforms the results of the perceptual schema into commands for the motor controller (M_t). This process, involving 4 parts can be seen in

Figure 3.1. Ideally, the transformation into action would be instantaneous, but in practice, the resulting motor command is generated as a result of the sensor data at some time removed from the original stimulus.

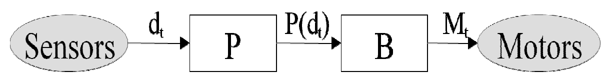


Figure 3.1. Basic reactive acoustically aware system.

Behaviors typically are relatively simple processes. An example behavior (phonotaxis) which uses acoustic perceptual data is moving in the direction of a sound source [Webb 1998]. By inserting one additional component into this architecture, the acoustically-aware behaviors can be easily combined with a multitude of other arbitrary behaviors. Described in [Arkin 1998], a behavioral coordination component (C), accepts the inputs of all behaviors at time *t*, and outputs the commands for the motor controller (Figure 3.2).

The general drawback to purely reactive systems, and this includes reactive acoustically-aware systems, is that there is no memory, so all of the knowledge for

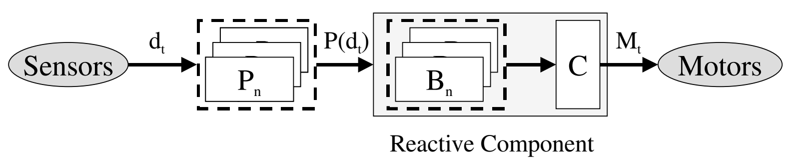


Figure 3.2 Reactive acoustically-aware system with behavioral coordination.

completing the task (both a priori and historical), which cannot be directly perceived by the robot, has to be built into the design of the controller component. This could be algorithmic manipulation of the perceptual data, or it could be facilitated by enhanced behavioral coordination mechanisms [Arkin 1998]. In either case, it places a large onus on the system designer, who must not only know the task extremely well, but also have to predict large numbers of failure situations and design control solutions that do not interfere with each other when operating simultaneously. It is still an open question as to whether or not general awareness can be achieved using sufficiently complex reactive systems, but in today's practice the required complexity for such a task is beyond the scope of a human designer.

3.1.2 KNOWLEDGE BASED ACOUSTICAL AWARENESS

For more complex tasks, some form of internal state needs to be added to the system. Starting with the reactive aware system in Figure 3.3, this can be accomplished by adding a knowledge component to the architecture. The defining characteristics of a knowledge component are: 1) it takes as input data from the perceptual system and the memory of its previous state(s), and 2) outputs the result of some computational processing to the reactive component (behaviors or coordination mechanism). Without the output to the reactive component, the knowledge component is limited to logging and monitoring tasks only. The resulting combined controller forms a hybrid reactive-deliberative robotic controller (Figure 3.3). In this figure, the knowledge component has been further subdivided into two pieces: (1) a knowledge planner (K), responsible for maintaining world models and creating plans for active perception, and (2) a task

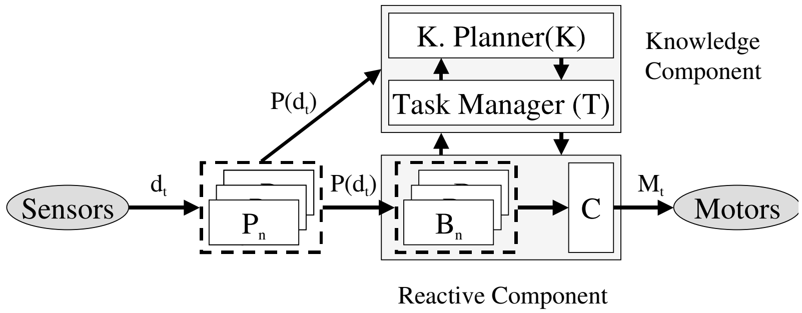


Figure 3.3 Hybrid architecture for supporting acoustical awareness.

manager (T), responsible for selecting, modifying, and following a plan. This subdivision of the deliberative component is equivalent to subdivisions made in earlier hybrid reactive-deliberative robotic architectures such as AuRA [Arkin 1998] or Atlantis[Gat 1991]. In these architectures, perceptual schema data is traditionally passed through the reactive component before reaching the knowledge component, but this is equivalent to the direct link between perception and the knowledge component, which is common in other knowledge acquisition architectures [Johansson 2003] so as to emphasis the importance of data collection.

The drawbacks to a knowledge component within the controller are synchronization, and general processing speed difficulties. As there is no limit on the amount of data stored in the knowledge components memory, there is no guarantee that any processing on that data can be completed fast enough to facilitate real time control. The solution, as proposed by some in the robotic architecture community [Gat 1991], is to run the knowledge acquisition components and motor control components asynchronously. At any given time-step, the motor controller has available to it the perceptual data processed at that time-step, and the most recent result of the knowledge acquisition system. With this design, the knowledge component facilitates the performance of the reactive system, assisting where possible.

In this dissertation, we are focusing on the development of this second type of acoustical awareness. As discussed in chapter 2, there has already been extensive work in developing behaviors that connect acoustical sensors with action, and there is likely to be much more in the future as acoustical applications become more commonplace. What has not received much attention is the augmentation of these behavioral systems with the deliberative form of acoustical awareness. Where behavioral systems can fail because of the difficult nature of acoustic inputs in arbitrary environments, supplemental knowledge of how sound flows through the environment can be used to suggest actions for avoiding failure cases and overcoming local minima inherent to the acoustic domain.

In the remainder of this chapter, we will concentrate primarily on identifying the type of information needed to guide an acoustically-aware robot through this type of deliberative robotic architecture. This corresponds to the first of the three critical sub-questions described at the beginning of this thesis: What types of data and information about the auditory scene are useful for an acoustically-aware robot? After identifying the set of potentially useful information, we will then describe how this information can be utilized in a mathematical framework for estimating sound flow through an environment.

3.2 Knowledge for Acoustical Awareness

Acoustics is an entity-driven perceptual domain. The problem of sound flow through the environment can be thought of in terms of its primary entities and how they interact: (1) Where is the sound coming from? (sources) (2) How will it travel around the environment? (paths) and ultimately, (3) How will it appear to the listener? (receiver) Improving robotic applications requiring sensing or transmitting sound requires at least a basic understanding of all of these parts.

> *"Every building acoustics problem, whether the enhancement of desired sounds, or the control of undesired sounds (noise), can be considered in terms of a system of sound sources, paths, and receivers."*[5]

For the remainder of this section, we will discuss the various aspects of each of these three primary acoustic entities. What are the types of information that an acoustically-aware robot may want to know? In the next section, we will discuss how this information can be combined together to model sound flow through an environment, and discuss why all of this information may or may not be needed for guiding an acoustically-aware robot.

3.2.1 SOUND SOURCE MODEL

The sound source is the most obvious of the three acoustic entities. These are the objects that emit that sound which ultimately arrives at a microphone, or receiver. Furthermore, they are also the entities that are most likely to change over the time during which a robot is executing some task. As such, it is imperative that a robot be able to acquire as much information as possible about the sound sources that may be effecting

---

5

Cavanaugh, W. (1999) , p.3

the environment in which it is situated. For each source in the environment, this includes information about:

- *Position* – coordinates of the sound source's location and orientation in environment.
- *Directivity* – the variation in amplitude of the outgoing signal due to angle of departure. This variation in amplitude may vary with frequency, as some frequencies may be better absorbed or transmitted by the materials from which the sound source is constructed.
- *Sound Function* – the sound produced by the sound source. This includes frequency, volume, and changes in these properties over time.

Of these three types of information, position is the only one that must always be known to some degree of accuracy. If the source position is not known, then the robot cannot predict anything about its effects on the environment. In the absence of further information, simplified models can be substituted for the other two properties. Without directivity information, the source can be approximated as omni-directional, invariant to both frequency and angle. Without specific knowledge of the wave function, a robot can approximate the sound source as a constant volume pink noise source. A pink noise source assumes equal energy frequency bands, and produces sound similar to fan or wind noise.

Given enough time, all three of these pieces of information about a source are determinable to some extent by a mobile robot. Chapter 4 discusses in more detail how this is possible, and Chapters 5 and 6 apply this information to robotic applications.

3.2.2 Environmental Information - Paths

The environment ultimately controls how much of the sound emitted by the sound sources reach the receiver. If there are walls in between the source and the receiver, then the sound will have to either travel around them, or through them. Either method reduces the volume, or otherwise changes the sound arriving at the receiver. How much that sound is changed is dependent upon the following properties of the path model:

- *Geometrical layout*: includes obstacle positions, walls, and all other surfaces in the environment (Figure 3.4). Used for predicting reflections, and regions of acoustic shadow.
- *Material Properties*: what materials are each of the surfaces made of, and what are the acoustic properties of those materials? Used for determining how much sound is transmitted through a wall, absorbed by the wall, or reflected from it.
- *Structural Composition*: what does the support structure of the building look like, such as would be found on an architectural blueprint. Used to determine absorption rates when calculating the strength of transmitted sound.

If this information is not provided a priori to the application, then acquiring detailed information about the environment can be very difficult (although not impossible) for a mobile robot. Luckily, simplistic yet reasonable assumptions are available for all of these. The simplest geometrical layout is an unobstructed outdoor environment. For indoor applications, sound propagation models generated under this assumption are better when the walls are far away, and the environment is relatively uncluttered, but they may still have some value depending upon the application. For better performance, what is needed is a map of the primary architectural features of the

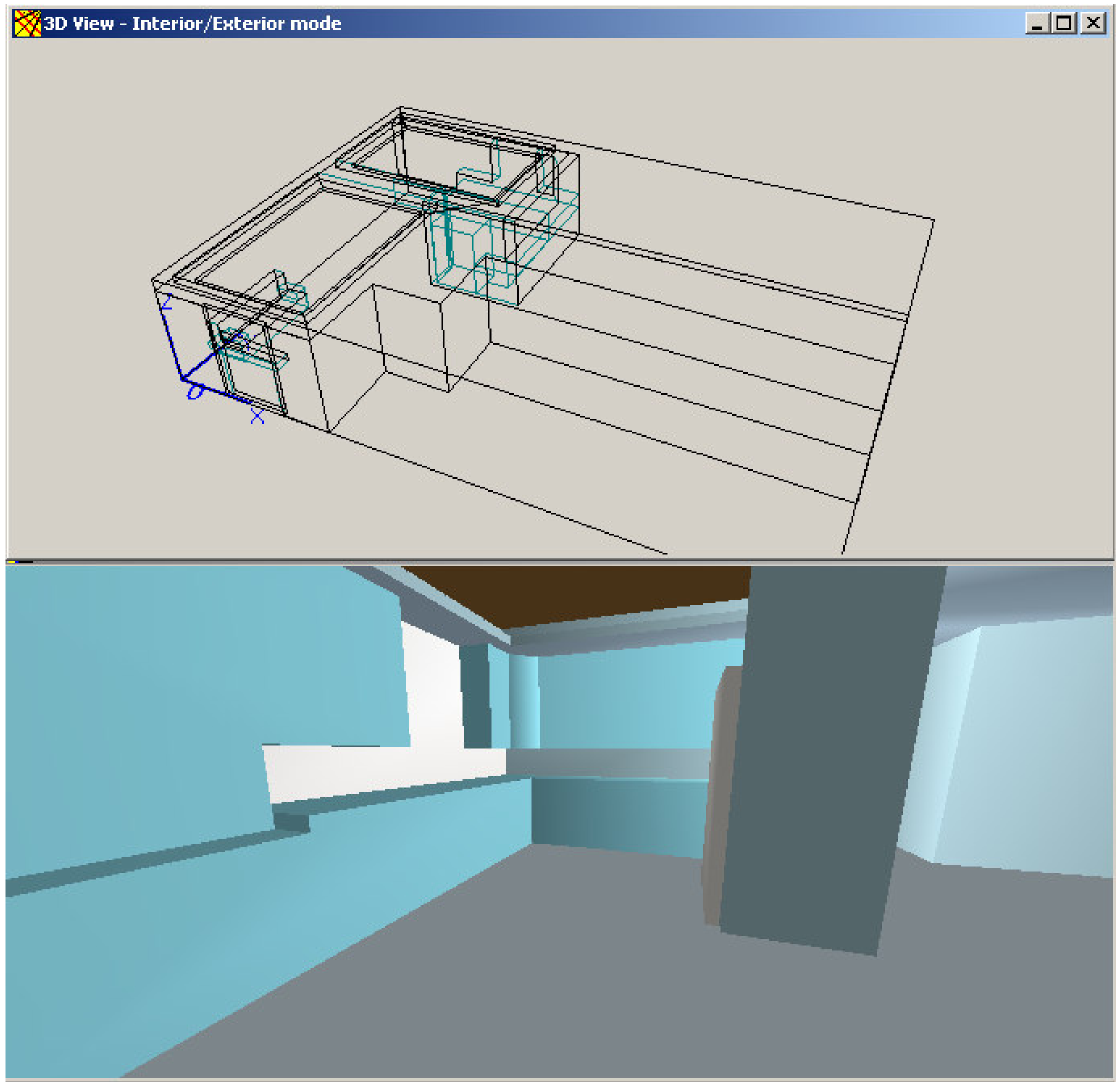


Figure 3.4. 3D model of the Aware Home Laboratory used for estimating reverberation effects and general sound propagation. Small obstacles in the environment, such as plants, dishes, etc. are less important for reverberation models than architectural specifications of surrounding floors, walls, and ceilings. The visualization was created by the software, Odeon 6.5 Combined.

environment, including large pieces of furniture and/or machines. The larger the object the more effect it will have on sound propagation, but also the less likely it will be moved around. For instance, a small houseplant, which could block the path of a robot moving through the environment, does not affect sound flow very much even if it is moved from place to place. Figure 3.4 demonstrates an example 3D geometric layout used with the Odeon acoustical modeling software for sound flow through a kitchen/living room. In Chapter 4, we will discuss the acquisition and use of the geometric layout for an acoustically aware robot. Chapter 6 will then discuss the application of a 2D model to real robotic scenarios.

Without knowledge of the geometrical layout, there is no need for material properties or structural composition of the environment. If the layout is available, however, but material and structural properties are not (a very common occurrence), then a simplifying assumption that can be used is that the walls are thick and solid (i.e. non-transmitting), and that all materials are perfectly reflective. In this dissertation, we will always be using these simplifying assumptions in our sound flow models, since material and structural composition are usually unavailable to the robot. As discussed in chapter 2, however, some material properties are potentially determinable by the robot, as may some structural composition. In future applications of acoustical awareness, additional knowledge, even if pertaining only to individual surfaces or walls, could still be used in conjunction with these simplifying assumptions to improve the accuracy of overall sound flow models.

3.2.3 RECEIVERS

The final acoustic entity needing some description is the receiver. Usually, with a robotic application, this will refer to a microphone, or array of microphones, situated on a mobile robot. Sometimes, however, a robot might be producing noise, or speech, for a human listener to hear, in which case the receiver would refer to the human listener's auditory system. In this chapter of the dissertation, we are interested in building a model of sound flow through the environment, so as to predict what a listener will hear at any location in the environment. In order to estimate what the listener might hear, however, the robot also needs a model of that receiver. While the range of information that can be stored and collected for each receiver will vary significantly between applications, the following are typical attributes for a receiver model that a robot might need:

- *Position* – coordinates of the receivers location in environment
- *Directivity*: the variation in the perceived amplitude of the incoming signal due to angle of incidence.
- *Frequency Response*: a range of frequencies the receiver can detect, and the relative amplitude across frequency bands.

As with sound sources, the position of the receiver is a critical piece of information. Without knowing anything about the position of the receiver, it is impossible to guess what the receiver might hear. Since the receiver is often mobile, however, the position of the receiver is often initially estimated as a set of possible positions, or an area over which the receiver might move about. This allows a broad initial estimate in the form of a map, or guide, that a robot can then use to predict what it will hear at any given reachable location.

The second two attributes of the receiver, directivity and frequency response, are less important to the sound flow estimation process. Without further information, the receiver can be simply modeled as an ideal point sampler. That is, any pressure changes occurring at that location are assumed to be perfectly recorded by the microphone, regardless of the frequency or the angle of incidence to the receiver. In this dissertation, our receiver information is always restricted to position only, as we used omni-directional microphones in all of the experiments, and did not separate out frequencies.

3.3 MATHEMATICAL FRAMEWORK FOR SOUND PROPAGATION

The previous section identified a large amount of knowledge that can be collected from each of the three primary acoustic entities in the soundscape: sources, paths, and receivers. That knowledge, however, does not by itself indicate what a robot would experience as it moves about the environment. Ideally, an acoustically-aware robot should be able to predict what it will hear so that it can make decisions about either avoiding the noise, or moving towards the noise, so as to improve its performance at some acoustic task. How do we bridge the gap between the information available to the robot, and this predictive capability? Thankfully, this problem is already of great interest to another research community, the field of architectural acoustics. When designing a building, engineers often need to consider the ramifications of their design choice on the flow of sound. In the case of concert halls, the aim is to aid sound propagation, so that the sound reaches more people, and does so in a fashion befitting the type of music being played. In more typical mundane buildings, such as offices or homes, the goal is usually the opposite, trying to minimize the effects of ambient noise on people trying to work or live. Whatever the target building of the acoustical engineer, the goal requires the same knowledge, knowledge about the flow of sound from source to receiver, through the environment. This is exactly what our acoustically-aware robot needs.

Of the methods for modeling, or understanding, sound flow through an environment, the theory of sound fields is one of the most commonly utilized. The theory itself is based on the physical principles of sound propagation, as laid out in common acoustic textbooks [Wilson 1994; Cavanaugh 1999; Raichel 2000]. Although this is not the only method for predicting sound flow through an environment, it is particularly suited to mobile robotics as it provides a framework into which many different types of information can be inserted. Furthermore, the resulting framework allows for unattainable knowledge, breaking down gracefully in the presence of unknown quantities. If the robot does not have available to it some knowledge, either a priori or through self-acquisition, then the resulting estimates of sound flow can still guide a robot to or away from sound sources, improving performance over an uninformed robot (see Chapter 5). Other solutions are not quite as appropriate for robotics. In particular, it is possible to solve the wave equation directly using certain assumptions and approximations for unknown quantities. However, solving the equation typically trades speed for accuracy, and degrades quickly with missing information. As computation is still at a premium for a mobile platform, and missing information is common, this solution may not be as applicable in robotics where an estimation of sound flow across large areas may need to be known.

3.3.1 THEORY OF SOUND FIELDS

The theory of sound fields [Svensson 2002] is used to make predictions about the soundscape given the knowledge available. Although called a theory, it is more of methodology that separates different types of acoustic knowledge from each other, so as to make useful estimates about the sound present at any given location in the environment even with only partial information.

To describe the theory, let us first make the following assumptions:

1. Assume there exists a function $S_n(x,y,z,t)$, for each sound source (*n*) in the environment, which can determine the instantaneous pressure generated by that sound source at time (*t*) and location (*x,y,z*) in the environment.
2. Assume that all functions S_n in the environment are independent of each other. In the case of sounds in the audible range at typical volumes, the flow of sound due to each source is generally unaffected by other sources.

Using these assumptions, the theory of superpositioning says that the total pressure at a given time and location ($d_{x,y,z,t}$) can be estimated as the ambient pressure (P_0) plus the sum of the effects of each sound source in the environment.

$$d_{x,y,z,t} = P_0 + \sum_{i=0}^{n} S_i(x,y,z,t) \qquad \text{Equation 3.1}$$

But what is the nature of the sound source function *S*? Although the effects of each sound source in the environment can be separated from each other, the function S cannot be separated from the physical environment itself. If we have a spherical sound source, which generates a single pressure pulse, then that pulse will propagate spherically, decaying in amplitude, until it hits a physical obstacle. At that point, on a smooth surface in a perfectly rigid environment, the entire pulse is reflected from the wall back into the environment at some angle relative to the angle of the incoming pressure wave. If the surface is not smooth however, then some scattering will occur at the impact point, generating uneven reflection from the wall. Moreover, if the surface is not perfectly rigid, then the wall will absorb some of the sound, and some will be transmitted through to the other side.

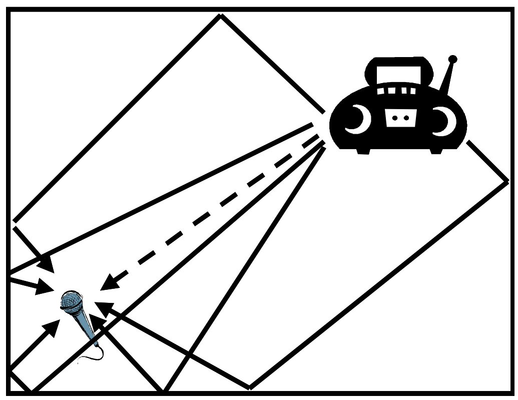


Figure 3.5 Direct (dashed) vs. indirect paths from source to receiver.

Fortunately, however, the nature of sound again promotes the use of superpositioning. Where before the soundscape was split into component sound functions, now the component sound functions may be split into their component parts. In this case, the separable parts of interest are:

- *Direct Sound* – sound waves with an unobstructed path from source to receiver.
- *Diffraction* – the bending of sound around barriers/obstacles.

- *Reflection* – throwing back sound waves from a surface. Includes absorption and scattering effects due to rough, non-rigid surfaces.
- *Structure-Borne Vibration* – sound absorbed by walls/obstacles may generate vibrations that can travel a long way through a solid medium before re-entering the air as sound. A good example is a vibration that travels from the basement to the third floor of a high rise, using the steel skeleton of the building as a conduit.
- *Transmission* – sound waves that continue through the wall/obstacle to emerge as waves on the opposite side.

It is more common in the acoustics literature [Raichel 2000], however, to group a number of these effects together into separable sound fields. A *sound field* can be described loosely as the sound in the region of interest around the source. The *direct field* is then the sound field created only by direct sound. A *reverberant field* includes diffraction, reflection, and surface diffusion effects. Since these fields are assumed to be independent of each other, the sound function (S_n) can then be described as the summation of the effects of each field on the location (x,y,z) plus some transmission effects at given time t.

$$S_n = D + F + T \qquad \text{Equation 3.2}$$

Where:

D = direct field.

R = reverberant field, including reverberations, surface diffusion, and diffraction effects.

T = transmitted and structure-borne sound

Although the exact form of these sound fields is impossible to determine at any given point in time, some reasonable estimates may be generated for a number of these fields using existing tools. As will be discussed in the following sections, the selection of tools, how they are applied, and even what information they need actually varies significantly from field to field. However the estimates for each field are generated, the beauty of the sound fields model is that as long as the effects being modeled are independent of each other (or have relatively small effect on each other), they can still be summed together to estimate the whole. Furthermore, if some effect or sound source is not currently being modeled, for whatever reason, then while the accuracy will decrease, the knowledge about the flow of sound in the environment is still correct, and still potentially useful for effecting robotic movement.

While constructing this model of separable sound fields, we have assumed that all of the fields are independent of one another, it should be noted that in reality the transmission and structure-borne effects are not actually independent of either the direct or reverberant fields. The origin of the sound to be transmitted has to come from somewhere, and may originate in either the direct or reverberant field. Then, any sound that is actually transmitted, i.e., reaching the other side of an obstruction, will continue to reflect around the environment and contribute to the reverberant field. In practice, however, transmitted sound is not included in either field because the calculation methods for estimating either transmitted or structure-borne sound differ substantially from those for calculating the other fields.

In Figure 3.6, we demonstrate the overall relationship between the separable sound fields model and the previously described acoustic entities. The robot will be

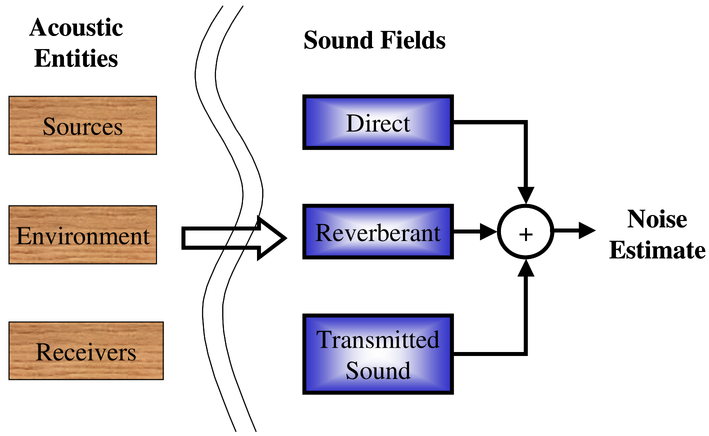


Figure 3.6 Relating acoustic entities to sound fields when building noise estimates. Each of the sound fields makes use of information from all three of the acoustic entities to estimate sound propagation.

gathering as much information as it is able to about each of the three acoustic entities. That information will then be used to construct each of the sound fields that build our final noise estimate.

3.3.2 Direct Field

To describe the direct field, let us start by describing the source itself. Using Fourier's theorem, any waveform can be described as the super-positioning of some number (N) of harmonic waves at different amplitudes, frequencies, and phases.

$$\vec{p}(t) = \vec{A} \cdot \sin(\vec{n}\omega t + \vec{\varphi}) = \vec{C} \cdot e^{j\vec{n}\omega t} \qquad \text{Equation 3.3}$$

Where:

$\vec{A}$ = amplitude vector of length N

ω = fundamental frequency

$\vec{n}$ = [1, 2, 3, 4,…,N]

$\vec{\varphi}$ = phase angle vector

$\vec{C}$ = complex amplitude vector.

Determining the effects of the direct field on the receiver is the simplest sound field to estimate. Assuming an ideal spherical source at a distance (*l*) from an ideal point receiver, and a constant speed of sound (c), then the amplitude of the incident wave can be approximated by a linear drop off with the distance (*l*). If no unobstructed path exists between source and receiver, then *l* approaches infinity.

$$\vec{D}_n(l,t) = \frac{1}{l}\,\vec{p}\left(t - \frac{l}{c}\right) \qquad \text{Equation 3.4}$$

In practice, however, receivers are never ideal. A real microphone is limited in the detectable frequency range. Microphones detect minute changes in pressure (sound waves) only when those changes cause mechanical elements inside the microphone to move, which in turn generates an electric signal. However, the shape and the material properties determine the frequency range to which the microphone responds, and there is no single element design that resonates in response to all frequencies. Even those frequencies that do cause resonance in the sensing element, vary in the size of the resulting amplitude of vibration, in turn affecting the value of the "sensed" signal. This is called the microphones frequency response ($\vec{r}$), and is usually provided by the manufacturer.

The other adjustment that needs to be made for real microphones, is the variation in directivity. An ideal microphone is omni-directional, reacting equally to signals arriving from all directions. A real microphone is not. Even the microphones sold as omni-directional usually have blind spots where the physical cable connects to the microphone. In practice, this can be adjusted for using a scalar adjustment ($Q_{\theta,\phi}$), depending on the incident angle (θ,ϕ). In reality, the directivity value at a given angle should probably be a vector that varies with frequency, but the amount of error introduced by the scalar assumption is relatively low in comparison to other assumption-induced error.

To define the resulting direct field effect, let us first define an element-wise vector product operation (*), where, for all elements in vector $\vec{B} = \{B_1, B_2, B_3, ..., B_N\}$:

$\vec{B} = \vec{C} * \vec{D}$ implies $B_i = C_i \cdot D_i$

Then, then the direct field effects from an omni-directional point source, on a directional microphone can be estimated as the element-wise vector product of the magnitude of the individual frequency components, and the directivity pattern (Q_r) of the microphone at that angle towards the source (θ,ϕ).

$$\vec{D}_n(t) = \frac{Q_r(\theta,\phi)}{l}\left(\vec{r} * \vec{p}\left(t - \frac{l}{c}\right)\right) \qquad \text{Equation 3.5}$$

Where:

l – path length from source to receiver

$Q_r(\theta,\phi)$ – directivity adjustment, for incident angle (θ,ϕ)

$\vec{r}$ – frequency response of the receiver

$\vec{p}(t)$ – sound function generated by the source at time t

c – speed of sound

Receivers are not the only directional entity in the equation. Sources, too, are directional, and are typically modeled by a similar directivity function ($Q_s(\theta,\phi)$). In the source case however, (θ, ϕ), are measured from the front of the source to the receiver instead of vice versa. So if we differentiate between the angle of departure from the source (θ_s, ϕ_s), and the angle of incidence on the receiver, (θ_r, ϕ_r), Equation 3.5 can be re-written to include both source and receiver directivity:

$$\vec{D}_n(t) = \frac{Q_r(\theta_r, \phi_r) \cdot Q_s(\theta_s, \phi_s)}{l} \left(\vec{r} * \vec{p}\left(t - \frac{l}{c} \right) \right) \qquad \text{Equation 3.6}$$

So, to summarize the estimation of the direct field, there are a number of different pieces of information that can be used. A source location is necessary to calculate the path length to each location in the field, and a source function is necessary to scale the response at each possible receiver location. Even the source function, however, does not

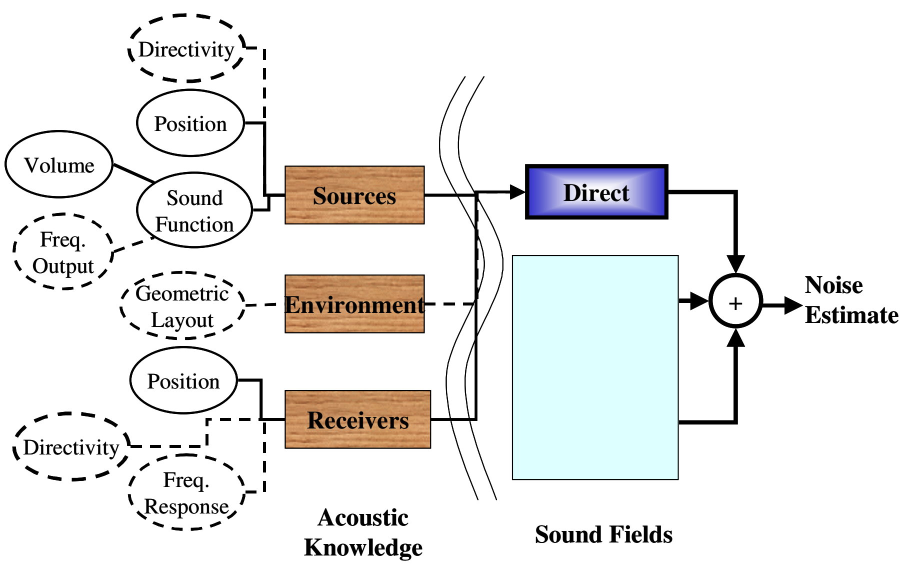


Figure 3.7 The information from each acoustic entity necessary for building a direct field estimate. Only the position and volume of the sound source, and the position of the receiver (solid circles) are absolutely necessary. Other information that can also assist in the calculation (dashed circles) includes directivity of the source and receiver, frequency output of the source, the frequency response of the receiver, and the geometric layout.

need to be complete, as a simple volume is good enough for making an estimate. Everything else is optional. Equation 3-4, which uses just path length and source function is probably the simplest estimate of the direct field to use with new sources. Then, as more information becomes known about the microphone and the directionality of the source, these can also be incorporated into the equation to hopefully improve the results. Figure 3.7 graphically displays the set of information that needs to, or just can, be included in the direct field calculations. Unlike the other fields, recalculating the direct field is a real-time operation even at its most complex that can be done quickly whenever new information becomes available. In Chapter 4, we will discuss the creation of direct field estimates from source position, directivity, and volume using this approach. It is left for future work the incorporation of time varying sound functions and microphone information into the field calculations.

3.3.3 REVERBERANT FIELD

A reverberant sound field can be described as a field created from the reflection, diffusion, and diffraction of sound waves in an environment with physical obstructions to the flow of sound. Without the physical obstructions, it would be only the direct field, but with them, sound waves can take many alternate paths to reach the same location in the environment. When a sound wave hits a hard, smooth surface, it will bounce off at an angle of reflectance (θ_r) equivalent to the angle of incidence (θ_i), see Figure 3.8. If the surface is perfectly rigid, then no energy is lost at the transmission point, and only the phase of the wave is changed. Otherwise, the amplitude of the wave is affected relative

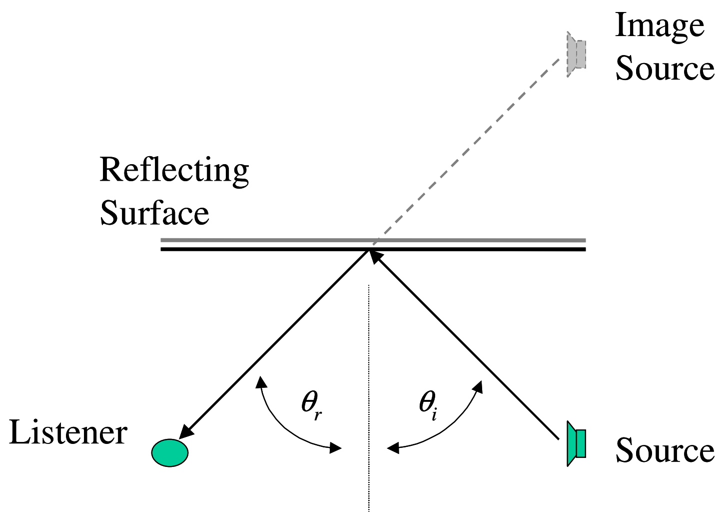


Figure 3.8 Modeling the image source method. In a normal reflection, an incident wave is reflected off of smooth surface at incident angle $\theta_I=\theta_r$. The image source method uses this principle to replace reflections with equivalent sources having the same angle of incidence as the reflection

to the frequency of the incident wave. How many of these reflected waves ultimately reach the receiver is determined by the shape of the environment.

Image Source Method

The simplest model of reverberant environments in fact uses this relation of equal angles to describe the reverberations as the effects of "mirror-image" sources. Figure 3.8 describes the idea. Each reflecting surface can be replaced by an equivalent source ($S^{'}$), which is located at a mirror-image position about the reflecting surface from the original source. The summation of the direct field effects from all of these sources then constitutes an estimate of the reverberant field. This is called the image source method (ISM) [Savioja 1999].

In Equation 3.7 the image source method is used to estimate the combined effects of all 1st order reflections from a single source *n* to a receiver located in the environment. The order in this case, refers to the number of reflections that have occurred in the path from source to receiver, so 1st order means only one reflection between the source and the receiver.

$$\vec{R}_n(t) = \sum_{i=1}^{\#sources} \frac{Q_r(\theta_{r,i}, \phi_{r,i}) \cdot Q_{s,n}(\theta_{s,i}, \phi_{s,i})}{l_i} \left(\vec{E}_i * \vec{r} * \vec{p}_n \left(t - \frac{l_i}{c} \right) \right) \qquad \text{Equation 3.7}$$

Where:

l_i = path length from the real source *n* to the receiver.

$Q_r(\theta_r, \phi_r)$ = directivity adjustment at microphone, for incident angle $(\theta_{r,i}, \phi_{r,i})$

$Q_{s,n}(\theta_s, \phi_s)$ = directivity adjustment due to source *n*, for departure angle $(\theta_{s,i}, \phi_{s,i})$ from the image source.

$\vec{r}$ = frequency response of the receiver

$\vec{p}_n(t)$ = sound function generated by the source at time *t*

c = speed of sound

$\vec{E}_i$ = Environmental adjustment to the amplitude, depending upon the materials reflected off along the path.

The image source method however, is not without problems, largely because of its simplicity. For one, each reflecting surface needs to be flat in order to estimate the reverberant effects as that of a mirror-image source. Second, the surface also needs to be smooth, otherwise diffusion (or scattering) effects occur as a multitude of waves are generated from each intersection point, which image source methods cannot accurately model. Thirdly, image-source methods do not scale well for calculating higher-order reflections. Even with just 1st order reflections, the positions for all mirror image sources have to be calculated first, before directivity and path lengths can be determined. As the order of the reflections increases, however, the number of image sources grows exponentially. Furthermore, the computational effort required for estimating the position, path length, and directivity of each higher-order source also grows along with the order, so even though the final summation of results is relatively easy, the finding of the image-sources themselves is highly inefficient.

Another problem not already mentioned is that the image-source method is very inefficient in estimating the entire field. For each possible position of the receiver an entirely new set of image-sources has to be calculated. Now, if the original source position is known at compile time, the positions of all image-sources for all receiver locations can actually be calculated in advance (which also alleviates the computational inefficiency problem), and then stored for quick access in a database, making the image-source method feasible in real-time. However, if the original source position is not known, then calculating all of the image-sources for all positions in the environment when it does become known will most likely require further compromises in accuracy to speed up the calculation time. Such compromises may include reducing the order of reflections calculated and estimating the field in a more limited area, perhaps for making spot updates to a pre-existing reverberant field model. For this reason, as well as the earlier problems, the image-source method was not implented as part of this dissertation to describe the reverberant field. Instead, it was only included because its simplicity may prove useful to other robot designers. A greater description of the image source method and its effectiveness can be found in [Allen and Berkely 1979; Svensson 2002].

Ray-Tracing

The image source method is loosely based on the concept of wave propagation. Instead of handling individual wave reflections though, it models the reflections as sources. Although it has been revealed to be computationally very expensive it was originally proposed as a faster alternative to a ray-tracing model of wave propagation[A. Krokstad 1968], which was capable of modeling surface scattering. Modern processing speeds however have made ray-tracing feasible and it has been implemented in a number of commercial platforms including Catt-Acoustic, Ease/Ears, and Odeon.

The basic idea behind ray-tracing is that a number of rays (often in the form of cones) are generated at random angles from the source into the room. When a ray hits a surface, it is reflected either specularly (as in image-source methods) or diffusely, depending on the scattering coefficient of the intersecting surface. Greater accuracy could be received by generating a new batch of rays at each surface to reflect on the level of scattering, but doing so quickly escalates computationally as the number of reflections is increased. To then determine the estimated effect on a single receiver then is only a matter of determining which rays intersect with the position of the receiver and summing them together (use Equation 3.6, substituting ray characteristics for sources).

What ray-tracing does not do well are two things. First, if the amount of diffusion at each surface depends on frequency, then without generating new rays at each intersection, the simulation may have to be repeated for each octave band. Second, ray-tracing does not simulate edge diffraction effects. Sound waves will bend around corners, or partial barriers, but the ray-tracing model does not support this phenomenon. However, there are some methods for estimating diffraction around partial barriers [Pierce 1989] that may potentially be used to estimate the effects of diffraction independently of ray-tracing.

Although the later augmentations to overcome diffusion and edge-diffraction effects have not yet been implementetd for this dissertation work, ray-tracing is used for modeling the reverberant field in Chapters 4 and 6. Chapter 4 in particular describes the data collection process by our mobile robot, and the creation of a reverberant field model from that data. A pseudocode description of our implementation is also available in Appendix B.6.

Other Methods for Estimating Reverberation

Besideds the image-source method and ray-tracing, there are also a number of other methods for calculating the reverberant field. Although these approaches were not implemented during dissertation work, these alternatives each have their own advantages/disadvantages, and are included here for completeness:

- Beam-tracing [Funkhauser et al. 2004]

  also uses rays, but creates beams out of adjacent rays. The advantage is that edges can "split" beams to model diffraction effects. Beam-tracing is similar in computational complexity to ray-tracing, but does not easily model surface scattering.

- Radiosity [Korany 2000]

  Another computer graphics inspired method which predetermines the wall reflection values as parts of larger wall elements. It is computationally efficient and easily handles a moving receiver. However, while scattering is simple to predict, specular reflections are more difficult to incorporate. In addition, it does not handle edge diffraction.

- Solving the wave equation

  There are also several numerical solutions to the wave equation for estimating the reverberant field. Volume element methods, such as the finite difference method in the time domain (FDTD) [Botteldooren 1995], discretize the air volume, and calculate the sound propagation as a function of neighboring units. Surface element methods, such as the boundary element method (BEM), discretize all boundaries and estimate their contributions to the sound pressure and particle velocity[Savioja 1999]. Numerical solutions provide very good details about diffraction, diffusion, reflectance, and surface scattering, but are extremely heavy computationally, so are generally practical only in either small or very simple environments.

The general purpose of being able to estimate the reverberant sound field is to integrate some information that was not previously available into the robotic controller. Unlike creating virtual spaces [Svensson 2002], or building concert halls [O'Keefe 1998], where accuracy is critical, there is already so much noise inherent to mobile robotics that obtaining the highest level of accuracy in acoustic simulations of the reverberant field is unnecessary. What the availability of all these tools for estimating the reverberant field means is that the level of accuracy in estimating the field can be tailored to the acoustically-aware application. If diffraction effects are particularly strong in that target environment, then an alternative method to ray-tracing may be used. Hybrid combinations of two or more methods [Svensson 2002] are also easily fitted to this computational approach.

In general, the differences between each of the approaches for modeling reverberant effects are the complexity of the computational model and the specific effects included in the reverberant model. All of the approaches, however, from the image-source method to solving the wave equation, make use of the same information. They need the same information that the direct field required, plus knowledge of the geometric layout. Furthermore, these methods can now incorporate material properties of the surface if they are available. Figure 3.9 summarizes the set of information required for building a reverberant field estimate.

3.3.4 TRANSMISSION

The transmission of sound occurs when not all of the sound is reflected from an obstacle. In that case, some energy from the incident wave (E_I) is reflected (E_R), some is absorbed by the wall (E_A), and some is transmitted through the wall (E_T) and out the other side.

$$E_I - E_R - E_A = E_T \qquad \text{Equation 3.8}$$

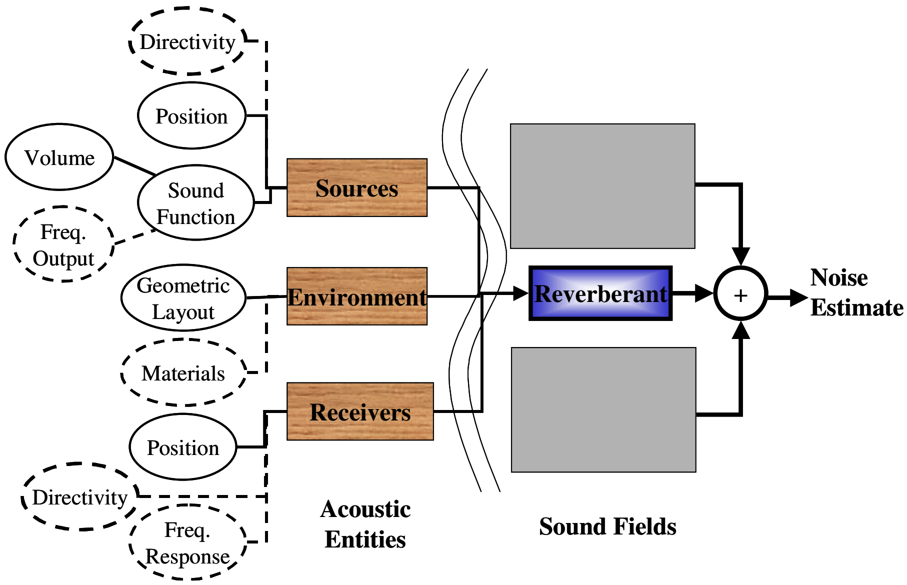


Figure 3.9. The information from each acoustic entity necessary for building a reverberant field estimate. Any non-trivial reverberant field requires a geometric layout in addition to the same information required by the direct field. Also, in the reverberant field equations, we see the use of material properties as optional

Unfortunately, even assuming a flat, relatively thin, panel in air with uniform density, the resulting transmitted energy is not easily estimated. It is not quite as simple as calculating either the direct or reverberant fields, because even if an accurate enough model of the environment (including geometrical layout and material properties) was available, then accurately estimating transmission would still depend on having accurate representations of the driving source functions. Due, in part, to the difficulty of acquiring accurate enough knowledge about the domain, our robots will not be estimating transmitted energy in the experiments described in this dissertation. However, if transmission plays an important role in the environment a robot is being deployed to, then there are still some techniques that can be used to make rough approximates of transmission for use by a mobile robot. These techniques, which will be described in the following sections, are divided into two different sets, each estimating different types of transmitted energy. The first type is direct transmission, which calculates the transmission loss (i.e. the fraction of sound power transmitted to incident sound power) of a wave traveling directly through an obstacle. The second type is structure-borne transmission, which happens when energy absorbed drives vibrations that travel through the structure to reappear as sound in a new location.

Direct Transmission

Direct transmission of sound is the transmission of sound through the obstacle, so that the angle of the waving leaving the obstacle is relative to the angle of incidence. Most typically, it is calculated with respect to walls in an environment, particularly outer walls that are designed to reduce noise interference from outside the structure. Unfortunately, modern building construction tends to make exact models of the structure difficult. Accurate transmission calculations would require knowing, not only what materials were used, but also where the studs, pipes, insulating materials, and especially air gaps between any of the materials are located. This information is just not available. What is available, however, are average transmission loss estimates for walls of similar make and construction. The same thing can usually be found for a variety of other materials that might also affect transmission, including windowpanes, doors, and even floor constructions. These then can be combined together to produce a single estimate of the transmission loss due to a given wall.

If, for the wall seen in Figure 3.10, we have the transmission coefficient (τ_i), and the surface area of each material exposed to the sound source (A_i) then the combined

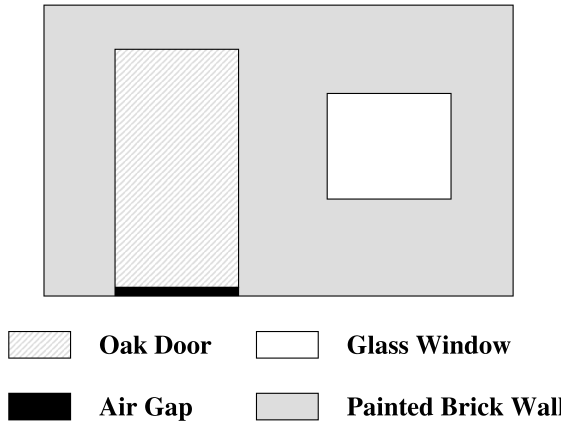


Figure 3.10. An example outer wall of a house used in estimating transmission of sound. This wall has 4 separate materials blocking entrance to outside noise: (1) Oak, (2) Glass, (3) Air, (4) Painted Brick. Even though the oak and brick walls are good at preventing transmission of sound, much noise will be transmitted through the air gap and the glass window.

transmission loss due to the wall can be estimated from the following equations found in [Raichel 2000]:

$$\tau_{combined} = \frac{\sum_{i=1}^{n} A_i \tau_i}{\sum_{i=1}^{n} A_i} \qquad \text{Equation 3.9}$$

$$TL_{combined} = 10\log_{10}\left(\frac{1}{\tau_{combined}}\right)$$

Where:

A_i = Area of surface *i*.

τ_I = transmission coefficient of surface *i*.

TL_i = transmission loss due to the wall

This equation makes the assumption that the source is far enough way that the sound hitting the wall is roughly the same at every location on the wall. While this estimate may not be ideal for sound sources inside a house, it is a good enough approximation for estimating the noise interference due to traffic, industry, or even airports, which are a common source of outside noise for home environments.

If the transmission coefficient is not available for a particular material in the wall then it is also possible to estimate the transmission coefficient given other properties of the material used:

$$\tau = \left(\left[1 + \eta \left(\frac{\omega m \cos\theta}{2\rho c} \right) \left(\frac{\omega^2 B \sin^4\theta}{c^4 m} \right) \right]^2 + \left[\left(\frac{\omega m \cos\theta}{2\rho c} \right) \left(1 - \frac{\omega^2 B \sin^4\theta}{c^4 m} \right) \right]^2 \right)^{-1} \qquad \text{Eq. 3.10}$$

Where:

ρ = the density of air

c = the speed of sound in air

θ = the angle of incidence on the panel

m = panel mass density per unit area

ω = frequency

B = panel bending stiffness

η = a composite loss factor

In practice, estimating transmission loss with any degree of accuracy is very difficult. Even with the ability to estimate transmission coefficients and some idea of the

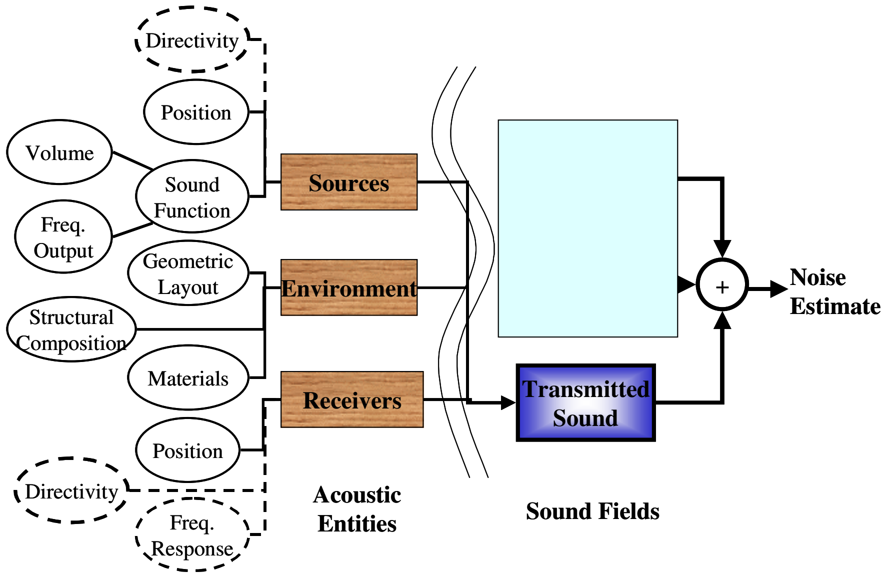


Figure 3.11. The information from each acoustic entity necessary for building a transmitted sound estimate. Estimating transmitted sound requires information about not only the sound source, but also the environment on both sides of the wall, and material properties.

structural composition of the wall, identifying all possible paths for sound to travel through the wall is unlikely, so the transmission loss equations can only provide a good initial estimate of how much sound is lost. To identify how much sound is hitting the wall in the first place, the robot needs to build direct and/or reverberant field estimates for other side of the wall. Figure 3.11 summarizes the large amount of information required to estimate transmitted sound.

Despite the apparent problems with acquiring so much information to build transmitted sound estimates, however, this area of the sound fields model may be of great importance to many applications. The reason is that cars, trains, airplanes, etc all impact everyday environments on a regular basis, transmitting significant quantities of noise through the walls from outdoors. Although a robot cannot predict exactly when a car will pass by, it can use its knowledge of transmission loss to improve its performance by predict where sound levels will be higher or lower due to these external noise sources. This, however, is left for future work.

Structure-Borne Sound

Another type of transmitted noise is structure-borne sound. Structure-borne sound is sound that reaches the receiver in at least part due to the vibration of a solid structure. Direct transmission through the wall is technically structure borne, because the wall must vibrate to generate the noise on the other side, but direct transmission is not usually what is meant by structure-borne sound. A good example of structure borne noise might be the plumbing in the house. As the pipes are all highly interconnected, and pressurized, a vibration at one end could easily generate noise at the other end of the house. Another good example are HVAC (heating, ventilation, and air conditioning) systems in office buildings. Besides the noise generated from the moving air, an HVAC in the basement can generate vibrations that travel through the building's frame to cause noise in seemingly arbitrary locations on other floors.

Of course, for a robot that is working in a small environment, structure-borne noise can often be modeled as a new source seemingly co-located with the wall. Transmission due to unobserved noise sources may also, sometimes, be modeled this way. If the robot needs to know, however, the relation between certain structurally-borne noise and an HVAC system (or any other known sound source) then there are some ways of estimating the structure-borne noise. Most commonly, identifying the transmission due to structure borne sound is done experimentally. If two noise sources are known to be interconnected, then samples are taken by hand at different frequencies to determine the exact relation. A robot with appropriate a priori knowledge, such as the fact that all water noise in the house should be inter-related, could also do this same task.

There has also been some limited work in using Statistical Energy Analysis (SEA) to estimate structure-borne effects. Like direct transmission, however, the models for SEA have to be fairly reliable in order to demonstrate any real accuracy, but there are some available programs that have been successfully demonstrated for analyzing building interiors [Koizumi et al. 2002], so that might be something to check out in the long run.

3.4 CHAPTER SUMMARY

This chapter has focused on identifying the nature of acoustic awareness. More specifically:

- **What does it mean to be acoustically aware?**

  Being acoustically aware means a coupling of action with knowledge about sound flow through the acoustic environment. There are, however, two levels of awareness, reactive vs. deliberative, that could be used for navigating a robot with respect to the auditory scene. Reactive awareness has been demonstrated in many of the applications in Chapter 2, as the robot listens, and then reacts to its current situation. To generalize the application to multiple environments and different types of tasks, we are focusing in this dissertation on the second type of awareness. A deliberative awareness of the auditory scene emphasizes the importance of knowledge of sound flow in the environment for guiding robotic navigation.

- **What types of information can, or should, an acoustically-aware robot acquire?**

  The soundscape can be divided into three primary acoustic entities: sound sources, paths, and receivers. The sound source is the source of noise propagating through the environment. The path entity incorporates knowledge of the environment to identify how the sound travels from the source to the listener. Finally, the receiver describes the listener, microphone or human ear, that has its own limitations on what it can hear, and from what angles.

- **How can the information be combined together to estimate sound flow?**

  The theory of sound fields was identified in this chapter as a convenient framework in which to insert information about the three primary acoustic entities, so as to generate a model of sound flow. The framework itself is not new, and, in fact, has been steadily improved over the past two decades by commercial software companies including Odeon and Catt-Acoustic. Although their exact computational methods are proprietary, their published works [Naylor] suggest they make use of both the theory of sound fields and wave-equation calculations to provide their estimates. What is new to this work, however, is the use of any part of this framework for real-robot applications. The theory of sound fields provides a necessary middle-step along the path to an acoustically-aware robot.

In the following chapters, we will build parts of this generic model of information and sound flow estimates into a real robotic application. Chapter 4 describes how to autonomously gather the necessary information for constructing models of both the direct and reverberant field, and demonstrates how lacking some information can still allow for useful models of sound flow through the environment. Chapter's 5-7 then describe robotic applications, demonstrating how to move the theoretical models from this chapter onto real robots and showing the advantages of being acoustically aware.

CHAPTER 4

ACQUIRING KNOWLEDGE ABOUT THE AUDITORY SCENE

In Chapter 3, we used the theory of sound fields to identify information useful to a mobile robot in understanding the flow of sound through the environment. With knowledge about the receivers, the sources, and the paths through the environment, a model can be created of the auditory scene to guide the robot in improving its performance. From where, however, can a robot reasonably expect to acquire this information? When not available a priori*,* the answer, by necessity, must be that the robot can determine this information using the sensors available to it. In this chapter, we focus on the problem of how, answering the second sub-question posed in Chapter 1. How can we combine data from multiple sensors to build effective representations of the acoustic environment?

The remainder of this chapter is organized as follows. The first section discusses existing work in localizing sound sources with static-mounted arrays and mapping the environment. These domains have been well studied as part of other problems for many years, and have produced well-established algorithms that can be used as a basis for further work. The next two sections build on this groundwork to extend the problem of sound localization to the more general auditory mapping domain. Multiple simultaneously operating sound sources are localized in 2-3 dimensions relative to the moving robot, and their volume and directivity are determined. Then, in the path domain, how to extract the geometric layout is discussed, and reverberant fields are created. Finally, from these sources of information we then build models of sound flow using the equations in Chapter 3 and compare the results to sample-based representations of the environment.

4.1 BUILDING BLOCKS

The development of an acoustically-aware robot has really been made possible by the scientific advances in three fields. The first field, architectural acoustics, contributed the mathematical framework described in Chapter 3 for combining information together into estimates of sound flow through the environment. The second two fields then contributed to the building of tools, and representations, from which the necessary information for the sound fields framework is extracted.

From the field of mobile robotics, the recent successes [Thrun 2002] in localizing the robot with respect to the environment and its past locations have contributed heavily to this work. Without relative position data a robot can be reactively aware to the auditory scene, sampling the environment and making decisions based primarily on local information. By incorporating knowledge of where the robot has been, the robot can now fuse its disparate collection of sampled data together into spatial representations of the environment, predicting where to move and where to avoid. Furthermore, with localization information, a robot can plan paths through the environment using those representations as a navigation guide.

From the field of digital signal processing, the last of the three scientific advances has been the work in localizing sound sources using arrays of microphones. Driven by interests in teleconferencing and military applications, researchers have developed reasonably robust algorithms for estimating angles and, sometimes, distances to the sound sources in the environment. Although the developed solutions for acquiring distance and angle need some adjustments for robotic deployment (Section 4.2.1), the underlying algorithms are essential in acquiring a critical piece of knowledge about the auditory scene, the angle to the sound source.

Together, these three fields together will serve as the building blocks from which we can build additional representations. They will assist in gathering the information necessary for building an acoustically-aware robot.

4.1.1 ROBOT MAPPING AND LOCALIZATION

The problem of mapping the environment has been of great interest to the robotic community for many years, as maps are convenient tools for planning and sharing information with other robots or human observers. One of the simplest methods for building such a map makes use of robot localization, i.e. a robot knowing where it is located with respect to its previous positions, to spatially fuse sensor data into maps. The evidence grid [Elfes 1992], or occupancy grid, is the representation through which data is combined to predict the probability of something occurring in each grid cell in a discrete map of the environment. Traditionally the sensor data being combined together are the robot positions, and the laser or sonar readings to obstacles in the environment, resulting in obstacle maps useful for guiding robotic navigation (Figure 4.1, Bottom). In Section 4.4, we will use such an obstacle map representation for building reverberant field models. The evidence grid representation, however, is not limited to sensing obstacles. In Section 4.2, we will apply the same representation to sound source localization, estimating the likelihood of a sound source appearing within each grid cell.

Figure 4.1. The B21r mobile robot (Top) and the obstacle map it created using the continuous localization algorithm (Bottom).

Unfortunately, reliably determining the robot's relative position to create such fused data maps can be a very difficult problem. The extended Kalman filter is an algorithmic method designed to do just that by fusing data together from a variety of sensors, including GPS, accelerometers, gyroscopes, odometric positions sensors, etc. Its accuracy, however, while substantially better than that of individual sensors, can still vary wildly with the precision, and cost, of the sensors being fused together. Simply placing a map building algorithm on top of this fused localization process may not be the best way to build an accurate map. For this very reason, the problem of localizing the robot has often been combined with mapping the environment. The combined problem is called the simultaneous localization and mapping problem (SLAM). This combined field has been actively researched for over a decade, and many good algorithms have been developed to exploit the advantages of mapping the environment while minimizing the disadvantages and uncertainty of robotic movement. A good source for specific details about many of these algorithms can be found in [Thrun et al. 2005].

After many years of development, robotic localization and mapping tools have recently become available to the wider mobile robot research community. Although such tools still have inherent problems in mapping dynamic or large scale environments [Thrun 2002], they are reasonably robust in structured indoor environments spanning a few rooms. For this dissertation, we made use of two such freely available tools to build two-dimensional maps of the environment and localize the robot using a SICK laser measurement system (LMS). The first such tool is based on the continuous localization algorithm [Schultz and Adams 1998], which extends the evidence grid representation to the simultaneous localization and mapping problem. This localization algorithm was used on an iRobot B21r located at the Naval Research Laboratory, Center for Artificial Intelligence. Both the robot and its evidence grid map can be seen in Figure 4.1.

The second tool utilized for robot localization came bundled with the Player/Stage software for mobile robotic control [Gerkey et al. 2003]. The pmap software developed by Andrew Howard [Howard 2004] uses particle filters to build a discrete, probabilistic representation of the environment. This representation is then used in real-time by a mobile robot to localize itself using an adaptive monte-carlo localization driver equipped with Player/Stage. This combined solution was deployed on an ActivMedia Pioneer 2-dxe robot in the Mobile Robot Lab at the Georgia Institute of Technology. The fully equipped robot and the resulting map can be seen in Figure 4.2.

4.1.2 SOUND SOURCE LOCALIZATION

The problem of sound source localization is one of the fundamental issues in modeling the acoustic environment: from where does the sound originate? Before we can predict the effects on the surrounding environment, even if that environment is assumed anechoic or otherwise acoustically simple, we must know from where the sound is being generated.

The fundamental property of sound that drives most localization algorithms is its finite, and relatively low (compared to light) speed as it propagates across the room. Given an array of microphones in an anechoic environment, the signal received at each microphone should be the same signal, only delayed by different amounts. Therefore, if we knew the delay between the arrival of the signal at each microphone we could also identify the position in environment from which the sound is originating. Assuming a

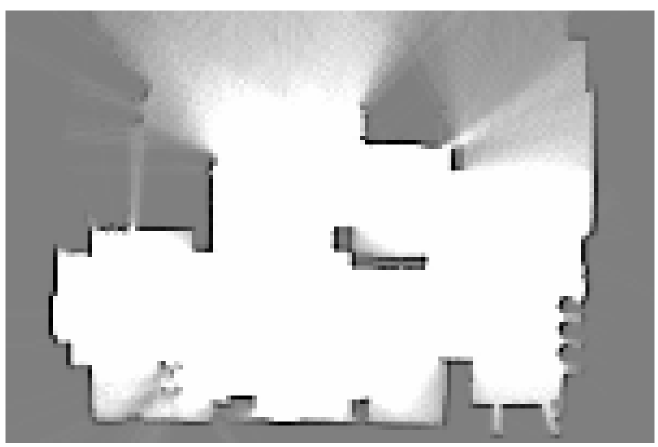


Figure 4.2. The Pioneer2-dxe robot (Top) and the map of the Mobile Robot Lab at Georgia Tech created using the PMAP software (Bottom).

constant speed of sound (c = 343 m/s), the time (T) required to travel the distance ($d_{l,m}$) from a source at position L to a microphone at position m is the distance traveled divided by the speed of sound:

$$T(l,m) = d_{l,m}/c \qquad \text{Equation 4.1}$$

Therefore, the delay between the signal arriving at the microphone located at m and the signal arriving at the microphone located at n is:

$$delay(m,n) = (d_{l,m} - d_{l,n})/c \qquad \text{Equation 4.2}$$

This delay is usually referred to as the time difference on arrival (TDOA), and measuring it can be difficult. If the microphones are in close proximity to each other, as is typical for most on-robot microphone arrays, then the delay between the signal's arrival at each microphone is very small. At 343-m/sec, microphones that are 0.3-m from each other (a likely maximum for a small robot array) experience maximum delays of less than 1-msec when the sound source is in line with both microphones. When the source is not in line, then this delay decays to 0-msec as the angle approaches perpendicular to the center of the array (Figure 4.3). With such small delays, accurate measurement in the time-domain is impossible. Instead, the signals recorded at each microphone are usually converted to the frequency domain using the Fourier transform, where signals can be compared at much finer delay increments than in the time domain.

The next problem in estimating the TDOA is comparing the signals. In any real environment, the signal of interest will be corrupted by some amount of noise, either from other ambient noise sources or reverberation from the environment. Therefore, in

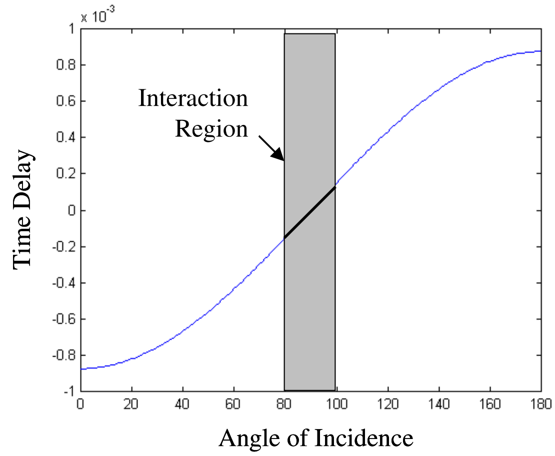


Figure 4.3. Measurable time delay between signals arriving at each microphone vs. the angle of incidence. This graph assumes 0.3-m between microphones in a binaural array, and a 1-m distance to the sound source. The gray region indicates the face-to-face interaction region for a binaural array mounted as ears on a humanoid.

actuality, the microphones are measuring multiple signals arriving from all different directions. Which signal is the robot interested in localizing? If the type of signal to be measured, such as speech, is known, then a filter can often be applied to the incoming signal to significantly reduce the effects of ambient noise sources. For reverberant effects, however, or when the ambient noise sources mask the same frequencies as the source being localized, a filter cannot completely clean the signal. If greater knowledge about reverberant paths or knowledge of ambient noise sources is available, then noise cancellation techniques can go beyond filtering to clean the signal within the frequency bands of interest. Even with both filtering and noise cancellation, the signal will rarely be free of noise, so any method for comparing signals recorded at each microphone needs to take into account the existence of noise.

Adaptations to Robotics

Overcoming these problems in microphone arrays small enough to fit on a robot have generally been restricted to three types of algorithms. The first type of algorithm uses the steered response of a beamformer for identifying the angle to one or more sources [DiBiase et al. 2001]. The advantage of using a beamformer, commonly used for combining acoustical signals in voice capture applications, is that it can be tailored using filters for the specific environment and placement of the microphone array. The basic algorithm uses the general TDOA principle to delay the incoming signal from each microphone some amount, effectively maximizing the energy for a specified angle. Localizing a source is then identifying the angle of incidence with the greatest energy. The drawback to beamforming systems can be their computational complexity. Especially once distance and environmental customization is included in the estimation, the amount of processing can be orders of magnitude greater than other localization methods.

The second type of algorithm uses high-resolution spectral estimation to locate sound sources. The MUSIC algorithm (MUltiple SIgnal Classification) is such an approach that has recently been applied to robotic platforms for speech localization [Argentieri and Danes, 2007]. Adapted from the field of high-resolution spectral analysis, these spectral estimation techniques are designed to handle multi-source localization problems. The difficulty with these approaches, however, is the amount of information required for high-resolution localization. An uninformed system requires many assumptions that decrease its effectiveness, especially under reverberant conditions [DiBiase et al. 2001].

The final approach, and the one most commonly used in robotic systems, is to calculate time-delay estimates using a generalized cross correlation algorithm [Martinson and Dellaert 2003; Blisard et al. 2007; Valin et. al. 2005]. This approach estimates the energy associated with a number of specified angle/distance pairs, and then maximizes the energy to localize a sound source. Although the specific generalized cross correlation approach has also been used with traditional beamforming, the real difference is in the computationally simpler weighting scheme. Instead of customizing the algorithm for a particular environment and/or signal type, an uninformed weighting localizes sources with a wide variety of sound functions. As this final approach is the one used for this entire dissertation, we will now describe this algorithm in greater depth.

Time Delay Estimates Using Generalized Cross Correlation

The solution most commonly employed by TDOA estimation algorithms is generalized cross correlation (GCC). For discrete-time signals, cross-correlation provides an estimate of similarity by bit-wise multiplying two signals together and summing the result. Two identical signals should produce an energy value equal to the sum of square of the signal, where two random signals would be significantly less. To find the TDOA estimate, the cross correlation algorithm is run many times, each time delaying the second signal by some amount. The delay that produces the highest cross-correlation energy is the best estimate of the TDOA between a microphone pair. Cross-correlation in the frequency domain works on a similar principle, maximizing the energy between microphone pairs, only now the cross-correlation is done in the frequency domain. Equation 4.3 demonstrates this equation for one pair of microphones.

$$F_l = \int_{\omega} W(\omega) M_a(\omega) \overline{M_b(\omega)} e^{-j\omega(delay(a,b))} d\omega \qquad \text{Equation 4.3}$$

where (M_a) is the Fourier transform of the signal received by microphone (a), $\overline{M_b}$ is the complex conjugate of the Fourier transform of the signal received by microphone (b), (ω) is the frequency in [rad/s], and (W) is a frequency dependent weighting function. Called the "phase transform" (PHAT) [Mungamuru and Aarabi 2004], this weighting scheme depends on the current magnitude at each frequency to evenly weight all frequencies present in the signal (Equation 4.4):

$$W(\omega) = \frac{1}{|M_a(\omega)| \left|\overline{M_b(\omega)}\right|} \qquad \text{Equation 4.4}$$

The position (l) that corresponds to the highest cross correlation value (F_l) is then the most likely position to contain the sound source.

Besides the PHAT weighting scheme, a weighting scheme that is also commonly employed in sound localization is maximum likelihood (ML) weights. This weighting scheme is most appropriate for tracking sound sources, such as human speech, which are not always present in the environment and cover a broad frequency spectrum. ML weights use knowledge of the noise (i.e. ambient sound not being tracked) affecting each microphone to attach greater weight to frequencies present in the tracked signal that are not present in the ambient sound. Equation 4.5 demonstrates the creation of ML weights from the noise spectra (N_a *and* N_b) corrupting each microphone:

$$W(\omega) = \frac{|M_a(\omega)| |M_a(\omega)|}{|N_a(\omega)|^2 |M_b(\omega)|^2 + |N_b(\omega)|^2 |M_a(\omega)|^2} \qquad \text{Equation 4.5}$$

The drawbacks to ML weights, however, often prevent their regular use. Tracking human speech works well, particularly in conjunction with a speech detection algorithm, because speech sounds are significantly different in frequency from common ambient noise sources like as HVAC systems. Other types of sound sources, unfortunately, are not tracked as well using ML weights. In particular, counter-weighting ambient noise is detrimental to tracking those ambient noise sources, so ML weights should not be used when the sound sources of interest cannot be turned off. Furthermore, sound sources with frequency signatures similar to the ambient noise will be minimized using this weighting scheme. This problem is especially difficult in indoor environments where air vents, fans, and to a lesser extent, plumbing noise, may be of interest to a robot.

Using either PHAT weighting or ML weights, Equation 4.2 still only applies to a single microphone pair. When the microphone array consists of more than two microphones, it can obviously be broken up into pairs of microphones, but how can the results be combined together? The simplest method for extending the GCC algorithm is to identify not the set of delays, but rather all possible locations in the environment from which sound might be originating. Knowing the speed of sound and the distance to the microphone array, we can predict the TDOA associated with a single microphone pair, and calculate the GCC energy for that location in the environment. With multiple microphone pairs (for n microphones in an array, there are $n!/(2!(n-2)!)$ sets of two microphones), we use the same location method, but now identify the delay associated with that location for every microphone pair, calculate the GCC energy for each microphone pair, and sum the resulting energy for that location over all microphone pairs. Equation 4.6 restates the original GCC equation in terms of the location, summing across all microphone pairs *{a,b}* in the array:

$$F(l) = \sum_{a=1}^{N}\sum_{b=1}^{N}\int_{\omega} W(\omega)M_a(\omega)\overline{M_b(\omega)}e^{-j\omega(T(l,a)-T(l,b))}d\omega \qquad \forall\, a \neq b \qquad \text{Equation 4.6}$$

This version of the generalized cross-correlation algorithm is often referred to as a spatial likelihood [Mungamuru and Aarabi 2004] as the resulting GCC energy directly corresponds to the likelihood of a sound source occurring in any given location. Figure 4.4 shows a contour plot of a spatial likelihood created for a 6x6m^2 grid centered about the microphone array. Appendix B.1 provides pseudocode for creating this grid representation of a spatial likelihood.

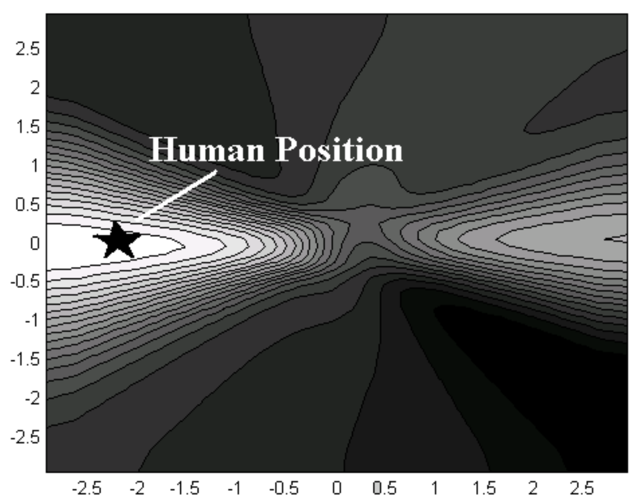


Figure 4.4. A contour plot of a spatial likelihood result for detecting human speech. Light areas are considered more likely. This result demonstrates the common problem of a strong angular performance, but poor distance estimates.

Figure 4.4 also demonstrates the drawbacks to the TDOA approach. In theory, given enough microphones in an array, it should be possible to exactly localize upon the source generating the noise. In practice, however, given the small distances between microphones in an on-robot array, as well as the levels of ambient noise and echoes from the environment, we have observed high amounts of error in the localization from one location. That error tends to be concentrated mostly along the axis stretching from the center of the array out through the sound source location, meaning that the cross correlation results are generally better at estimating angle to the sound source rather than distance.

Identifying Angle and Distance

In general, the problem of accuracy using on-robot microphone arrays is a problem of the domain and not the algorithm. Beamforming suffers from a similar problem in identifying distance to the sound source. With an on-robot array, there is just not enough difference between the signals arriving at each microphone to extract two-dimensional coordinates of the sound source from a single position. To overcome the missing distance problem, a common solution is to widely distribute the sensors about the environment. With a wide distribution of microphones, ideally placing the sound source within the convex hull created by the array, there have been a number of successfully deployed systems that could provide 3D coordinates for sound sources [Girod and Estrin 2001; Nakadai et al. 2006]. Work by Thrun demonstrates how not only the sources can be localized in this situation, but also the microphones themselves [Thrun 2005].

When a widely distributed network of microphones is available, a robot can certainly use this localization system for modeling sound propagation through the environment. A robot of large enough size could even be used to distribute the microphones in the first place if the situation is appropriate [Zhang and Sukhatme 2005], as could a team of robots, each with their own microphones [Martinson and Dellaert 2003; Parker et al. 2003]. However, such a system is not always available. In military or police situations, the robot is likely to be in hostile territory where people or robots cannot move in ahead of time to place microphones. Alternatively, maybe the funds are simply not available for placing, synchronizing, and utilizing large numbers of microphones distributed about the area. When the robot does not have the assistance of a microphone network to localize the sound source, there is another solution to acquiring more accurate coordinates. The robot can move itself to a new location and using its small microphone array with bearing dominated measurements to triangulate on the source from a different angle. In Section 4.2.1, we will demonstrate how this fusion of data can be done in real-time to extract sound the relative position of one or more sound sources in the environment.

4.2 REPRESENTATIONS FOR CHARACTERIZING SOUND SOURCES

The minimal information necessary for estimating sound fields, or the barest of information required by the mathematical framework, is the position of the sound source and the average volume at which it is producing noise. From this information alone, we can estimate the effects of the direct field up to some arbitrary distance about the source. As we acquire more information about the sound source, these effects can be estimated with greater detail and accuracy, ultimately contributing to estimates for both the direct and reverberant fields. If the information about the sound source is not known a priori, however, then what information can a robot collect to build the sound field estimates?

In this section, we will discuss the set of representations, or tools, that a robot can use to acquire the two-dimensional location of one or more sound sources in the room, and then determine their volume, directivity, and sound function. These representations will build on the mathematics presented in the previous chapter. For now, we are focusing on medium to long duration sources that can be expected to remain relatively static over the time the robot is collecting this information. In the long run, however, more information about the sound source function could also be collected by a mobile robot and incorporated into the sound fields framework.

4.2.1 IDENTIFYING SOURCE LOCATION IN 2D

For the purpose of modeling sound sources in the environment, we need sound source localization that includes more than just the angle from the center of the array. Specifically, assuming that every sound source can be represented as a point source, we need to identify a specific centroid coordinate in 2 or 3 dimensions. The reason for this simplifying assumption (point source) is that even though sound is usually generated from a vibrating surface, and the sound-fields framework can work with a more complicated model, sound source localization algorithms typically operate on the assumption that the sound is loudest from a given point. Greater information about the nature and size of the vibrating surface, such as might be needed accurately describing the effects of larger sources, may require more a priori information about the source being localized or more sensory information (e.g. camera).

As mentioned at the end of Section 4.1.2, the way to acquire these 2D coordinates using an on-robot microphone array is through robotic movement. By moving the robot from place to place in the environment, we can use combine the results of an angular source localization algorithm to triangulate upon the two dimensional location of the source centroid.

Depending on which underlying sound localization algorithm is being used, there are different possible methods, concurrently developed, for triangulating on sources in the environment. Using a steered-beamformer, the first method identifies the angle to one or more sources, and then mathematically minimizes the squared error identify sound source locations [Sasaki et al. 2006]. The drawback to this approach, however, is that if the error is high in the original beamforming results, the algorithm may not localize the sound source at all, or have a high false-positive rate. In their paper, the authors developed a specialized 32-element microphone array with an ideal beam pattern for their scenario to counter this effect. Even then, they still could not localize sound sources in the vicinity of walls or other highly reverberant areas of the environment.

As an alternative to using beamforming, a second approach developed for this dissertation [Martinson and Schultz 2006] uses time delay estimates (particularly, spatial likelihoods) to localize sound sources with a moving robot. While the robot moves through the environment, it creates a spatial likelihood for each collected sample, and combines the disparate measurements together into an auditory evidence grid. Unlike the concurrently developed beamforming approach, the auditory evidence grid representation does not require specialized hardware. Furthermore, although better microphone arrays and anechoic environments still demonstrate the best performance, the algorithm degrades gracefully in the presence of noise, allowing for a variety of environmental and hardware configurations. In the worst-case configuration, using the bare minimum microphone array (i.e. binaural) in a reverberant environment, sources can still be localized to some degree of accuracy.

The remainder of this section on localizing sound sources will focus on this second approach using auditory evidence grids. We will first describe the algorithm, including a review of evidence grids in general, and how to adapt them to auditory information. Next, we will investigate the performance of the algorithm under different hardware configurations, robotic movement strategies, and sound source types, with the goal of automating the process. With this experience, we then devise an automated process for extracting the sound source coordinates from the resulting grid and devise, and test, some autonomous robotic movement strategies for accurately estimating the location of sources in the environment.

Some of this work has appeared in other publications, prior to this dissertation. The initial description of how to construct an auditory evidence grid, along with the first phase of testing was originally reported in [Martinson and Schultz 2006]. Then in [Martinson and Schultz 2007] the robotic movement strategies and the second phase of testing for measuring accuracy were described.

Auditory Evidence Grid - Algorithm

The underlying algorithm for auditory evidence grids is the same evidence grid algorithm used in creating obstacle maps of the environment (Section 4.1.1). Only instead of using measurements to obstacles, as returned by lasers or other similar sensors, we are using estimates of distance to sound sources, collected aurally by a microphone array. The name auditory evidence grid, therefore, follows the change in information being mapped, from obstacles and spatial layouts to sound sources and auditory layouts.

As with spatial evidence grids, the auditory evidence grid uses Bayesian updating to estimate the probability of something being located in a set of predetermined locations (i.e. grid cell centers). Since we will be feeding the algorithm spatial likelihood measurements collected by the microphone array, that something being estimated is the probability of a sound source. Initially, it is assumed that every grid cell has a 50% probability of containing a sound source. Then as each new spatial likelihood is created from a sensor measurement, those probabilities for each grid cell are adjusted. For the simplicity of adding measurements together, we use the log odds notation to update the evidence grid. Equation 4.7, from [Thrun 2002], demonstrates this additive process for each new measurement

$$\log\left(\frac{p\left(SS_{x,y} \mid z^{t}, s^{t}\right)}{1-p\left(SS_{x,y} \mid z^{t}, s^{t}\right)}\right) = \log\left(\frac{p\left(SS_{x,y} \mid z_{t}, s_{t}\right)}{1-p\left(SS_{x,y} \mid z_{t}, s_{t}\right)}\right) + \log\left(\frac{p\left(SS_{x,y} \mid z^{t-1}, s^{t-1}\right)}{1-p\left(SS_{x,y} \mid z^{t-1}, s^{t-1}\right)}\right) \qquad \text{Eq. 4.7}$$

In these equations, z_t and s_t are the sensor measurement and robot pose respectively recorded at time t, z^t and s^t are the set of all sensor measurements and robot poses recorded up until time t, and SS_{xy} is a particular grid cell in the evidence grid. Therefore, $p(SS_{x,y}|z^t,s^t)$ is the probability of grid cell SS_{xy} being occupied given all evidence collected up until time t, and $p(SS_{x,y}|z_t,s_t)$ is the *inverse sensor model,* or probability that a single grid cell contains the sound source based on a single measurement.

The *inverse sensor model* used in this work is based on the spatial likelihood measurements described in Section 4.1.2. Every time a sample is collected from the microphone array a spatial likelihood is created by estimating the generalized cross correlation (GCC) energy over a set of pre-determined locations. For creating auditory evidence grids, we generally restricted the set of pre-determined locations to a 3-m radius about the robot, so as to limit the computational requirements of calculating a spatial likelihood. By itself, however, the spatial likelihood needs some additional modification for use with the evidence grid representation. In particular, we need a likelihood varying ideally from 0-100%, but the range of energy values returned is nowhere near that range. The energy from most unlikely to most likely in a single measurement often varies by a factor of 10^4. Furthermore, different environmental conditions, source types, etc. can then shift the entire grid by a similar factor either higher or lower. Therefore, to estimate $p(SS_{x,y}|z_t,s_t)$, we use an inverse sensor model that scales and shifts each spatial likelihood measurement, so that the result lies between two chosen probabilities [P_{low} and P_{high}] where the lowest cross correlation value resulted in a probability of P_{low} and the highest in P_{high}. Equation 4.8 shows how this scaling and shifting process is accomplished:

$$
\begin{aligned}
K_1 &= \left(F_{\min}(t) - F_{\max}(t)P_{low} / P_{high}\right) / \left(1 - P_{low} / P_{high}\right) \\
K_2 &= \left(F_{\max}(t) + K_1\left(P_{high} - 1\right)\right) / P_{high} \\
p\left(SS_l \mid z_t\right) &= \left(F_l(t) - K_1\right) / \left(K_2 - K_1\right)
\end{aligned}
\qquad \text{Equation 4.8}
$$

Where $F_{min}(t)$ and $F_{max}(t)$ are the lowest and highest F_l values calculated for the measurement taken at time (t). To then extract the resulting $p(SS_{x,y}|z_t,s_t)$ from $p(SS_l|z_t)$ the robot pose (s_t) is used to convert from local coordinates (l) to global coordinates (x,y). Figure 4.5 demonstrates an auditory evidence grid resulting from this process, localizing two radios in the environment. Appendix B.2 describes this creation process in more detail, providing pseudocode for the implementation used throughout this dissertation.

The remainder of this section on localizing sound sources in 2D is focused on testing the auditory evidence grid, and then automating the collection of samples and localization of sound sources. For all of these tests using the auditory evidence grid, the spatial likelihood results were typically scaled between [0.2, 0.95], but this could be

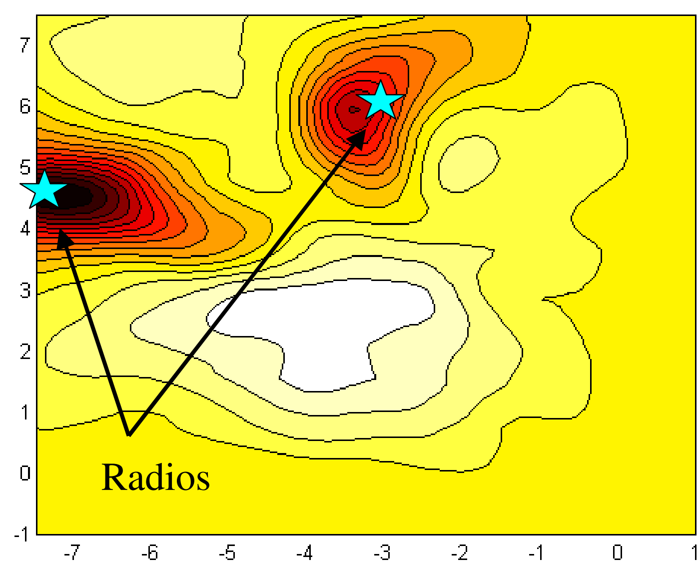


Figure 4.5. Contour plot of an auditory evidence grid localizing two radios. This contour plot combines 190 samples collected from a moving robot with four microphones. The darkest areas indicate the most likely sound source positions. To reduce the noise of a moving robot, a square sliding window (0.6-m in width) was used to produce smoother contours.

varied when tracking different types of sources. These scaling numbers were chosen empirically based on spatial likelihood reliability. As the spatial likelihoods would generally only point at one source at a time, areas not indicated with a high cross correlation result were not necessarily devoid of sources so setting the probability at 0 would unfairly penalize quieter sources. Similarly, spatial likelihoods could also make a mistake in the direction they pointed, and so 100% confidence was inappropriate in scaling the results.

Testing Phase 1 – Investigating the Range of Auditory Evidence Grid Performance

Testing of the auditory evidence grid algorithm was performed in two phases, both taking place at the Navy Center for Research in Artificial Intelligence, located on the Naval Research Laboratory. In this first phase of testing, a human-operator manually tele-operated the robot in a loop in the vicinity of some set of sound sources to test the performance of the algorithm under different operating conditions.

- How many sources can be successfully localized?
- How does robotic movement affect the results?
- How many microphones need to be used?
- Does filtering the data assist in localization?

From the answers to these questions, we will next construct an algorithm for automating this process by extracting sound source coordinates and moving the robot for better sample collection.

This phase of testing, with some modification, was originally reported in [Martinson and Schultz 2006].

Hardware Setup

The robot hardware used in this work was a B21R research robot manufactured by iRobot (seen in Figure 4.6). The robot is equipped with a SICK laser measurement system (LMS) mounted in the robot base, and two onboard computers for processing. Robot pose information is then provided by the continuous localization[Schultz and Adams 1998] algorithm, which uses a spatial evidence grid representation (different from auditory evidence grids) constructed from LMS range data and robot base odometry. The robot also has an additional SICK LMS mounted above the robot base and a full sonar ring that were not used in these experiments.

The equipment used for gathering the acoustic data was an array of (4) Audio-Technica AT831b lavalier microphones mounted at the top of the robot. These microphones were each connected to battery powered preamps mounted inside the robot body and then to an 8-Channel PCMCIA data acquisition board.

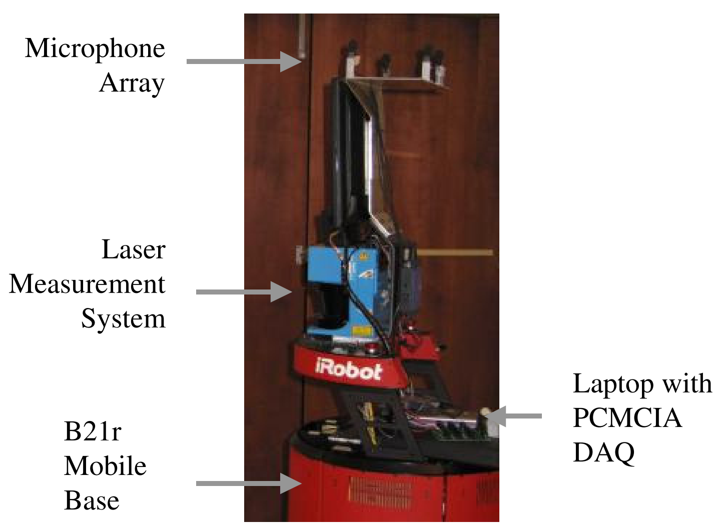


Figure 4.6. Fully equipped B21r mobile robot used for phase 1 testing.

Results

To test the algorithm, the robot was run in 20 trials, varying two parameters: (1) the set of sources active in the environment, and (2) whether or not the robot was moving while gathering data. A total of 10 different configurations of sources were tested, where a source configuration is defined as a unique set of active sources in the environment. For the following trials, 9 sources were mapped by the robot: 2 human speakers (male and female), 1 tape recording of human speech, 2 radios playing different types of music, and 4 air vents in the laboratory. Figure 4.7 shows the relative positioning of each of the sources, along with the grid used for localizing the robot in the 12x12-m^2 laboratory. In

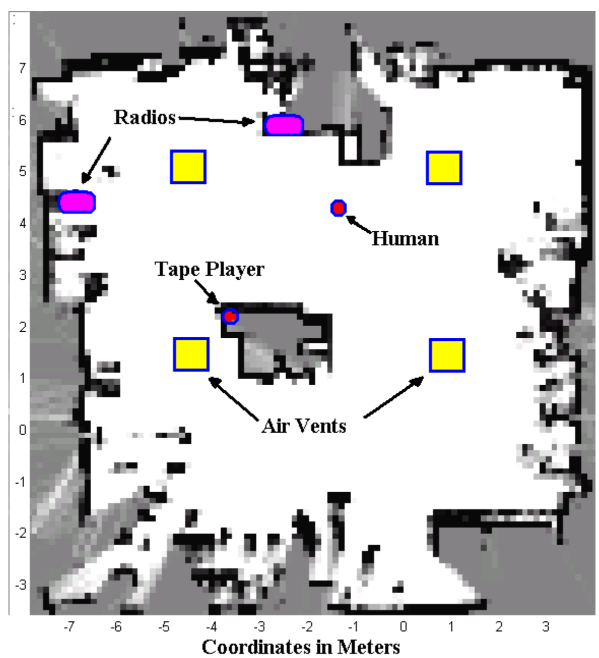


Figure 4.7. Spatial evidence grid used by the robot for localization with source positions shown relative to the obstacle positions in the room. Note that not all sources were active in every test.

general, the robot was not always exploring the entire area, but was instead restricted to a subsection in the vicinity of the sound sources of interest.

Robot movement during these tests was varied according to whether or not it was stationary while sampling audio data. In both cases, the robot was tele-operated in a large circle in the vicinity of the sound sources. In the first case, however, the robot would stop 6-7 times to gather samples of the auditory scene using its microphone array. In the second case, the robot would continue to gather microphone array measurements while it was moving. The reason for the different data collection techniques was to evaluate the effect of increased ego-noise on the robot, as movement increased the volume of wheel and motor noise generated present.

The results of all mapping experiments are shown in Table 4.1, where a successful test is defined as correctly placing a peak in the smoothed contour map within 1-m of the true source location for all active sources. The number of sources listed in this table does not include air vents. As the vents could not be fully disabled, they were on during all trials, but were too quiet to detect except when all other sources were disabled.

Table 4.1. The results of all phase 1 auditory evidence experiments.

# of Sources	# of Source Config.	Successes: Pausing while Collecting	Successes: Moving while Collecting
1	5	5	5 (4)
2	4	4	3
>2	2	1	1

In general, as demonstrated by the table results, the auditory evidence grid algorithm worked very well for mapping one or two sources. In only one test with two sources, did the robot fail to correctly map all of the active sources. There was an additional test using one active source, in which a phantom, or illusory, peak appeared strong enough in the evidence grid to be mistaken for a real source, but the active source was also found in the same evidence grid. Note that in both of these trials, the robot was moving while collecting data instead of stopping, so movement obviously did introduce some additional error, but the algorithm still succeeded in most cases to successfully map 1-2 sources.

Larger numbers of sources were not as successful, but this may have been due to the relative scarcity of samples. The trial that succeeded in localizing 3 sources had all three sources in a relatively small area, while the trial that failed involved a large area and multiple widely spaced sound sources (the air vents). In addition to these being relatively quiet sources, the robot did not sample equally in the vicinity of all sources due to the large area being sampled, thereby limiting the effectiveness of the localization results. A solution to this problem used in the following phase of testing is to investigate individual sources, sampling extensively in their vicinity to improve localization results, and ultimately gather sound source characteristics.

Besides the general sound source distribution problem, this first phase of testing was designed to test to reveal the effects of different design decisions on the quality of the map. What follows here is a discussion of those results:

- **Moving when gathering auditory data**

  As was seen earlier, the evidence grid representation still works when the robot is moving while sampling, but more problems occurred in trials where the robot was moving than when not. There are two reasons for this decreased accuracy in evidence grid. The first reason is that, when moving, the robot pose estimation algorithm introduces more relative error into the representation. As the robot pose estimates are used to align overlapping spatial likelihood measurements, this results in wider, lower peaks in the resulting evidence grid. The second problem when moving comes from the louder robot ego-noise generated by the robots wheels and motors. If the robot is generating more noise when moving, than there will be more noise present in the environment that can partially or totally mask the active sound sources being investigated. Algorithmically, this results in degraded spatial likelihood results, and less certainty on the origin of the loudest sound. The effect of this on the resulting evidence grid is twofold: (1) poorer spatial likelihood accuracy results in more phantom peaks, making it harder to distinguish actual sources; and (2) rougher object contours will be evident in the final map.

- **Number of Microphones**

  Many robots are now being equipped with a binaural microphone array (i.e. two microphones) to mimic human hearing, and there is no reason why spatial likelihoods cannot be computed using only 2 microphones. However, with a binaural setup, the accuracy of calculated spatial likelihoods decreases in both distance and angle. So to test the effect of a binaural setup on auditory evidence grid, we reused the data from the same trials discussed earlier, and only used two microphones streams instead of all four to generate the spatial likelihoods. The resulting effects on the evidence grid from this binaural approach is actually very similar to those seen when moving while gathering audio data: (1) more phantom noise sources, or peaks in the evidence grid are generated, and (2) the object peaks have rougher contours. However, as demonstrated in Figure 4.8, the same sources were generally still evident for both 2 and 4 microphone configurations in most trials.

- **Map Resolution**

To detect sources in real-time the evidence grid and spatial likelihood grid cell size was a minimum of 0.3m. This is a relatively coarse resolution that may have affected the resulting accuracy. To determine exactly how the resulting map was affected, we recreated the trial maps at a higher resolution (0.1m) using the data collected earlier. The result of increasing resolution was that it shifted the center of the peak in the evidence grid towards a more accurate center. However, that center would have otherwise been included in a larger grid cell at a lower resolution, so it was not unexpected. Unfortunately, though, changing resolution did not appear to affect the creation of phantom peaks or rougher contours.

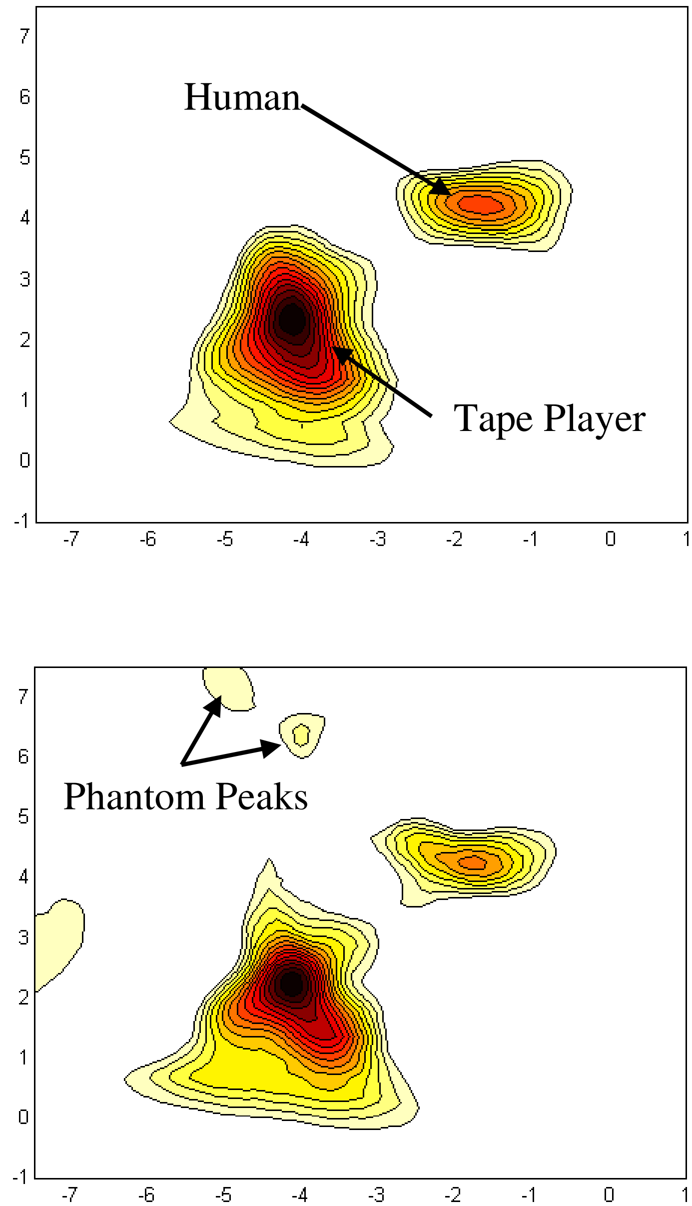


Figure 4.8. Auditory evidence grids localizing two speech sources (a stationary human speaker and a tape player) from 463 data points collected at 6 positions. Both grids are created from the same 463 recorded samples (not all are speech) using either a 2 (Bottom) or 4 (Top) microphone array, and thresholded to display only points more than 50% likely. Note the rougher contours and phantom peaks found in the grid created by only 2 microphones.

- **Filtering the Sample Set**

  If a priori knowledge exists about the set of sound sources being mapped, then another method for removing error from the map is to filter the sample set. One such filter tested was an rms threshold, like that employed in Linear Predictive Coding [Tremain 1982] for detecting speech over the telephone. The resulting maps for speech sources had smoother contours and better defined peaks. There is a tradeoff, however, in that fewer samples were used to create the maps in general, and that some source types (non-speech) were removed by this filter entirely.

Iterative Clustering

After gathering enough data, the resulting auditory evidence grid representation estimates the combined likelihood of a source being located at any position. By itself, however, the auditory evidence grid does not localize sources in space. What we need for use with the mathematical framework discussed in Chapter 3 are two- or three-dimensional coordinates of the source, but, as seen in Figure 4.8, sound sources in the auditory evidence grid are merely peaks of varying heights and contours in the map. An algorithm for extracting coordinates from the grid is required.

By applying a threshold to the auditory evidence grid (Figure 4.8), we can see that sound sources appear as clusters in the map. Using a nearest-neighbor clustering algorithm [Duda et al. 2001] on these evidence grids, we can easily extract the two largest speech sources and estimate their centers. With the large number of samples used to build evidence grids in phase 1 nearest-neighbor clustering algorithm typically produced errors of 0.3-m for clearly evident clusters such as seen in Figure 4.8 (Top). When using only two microphones, however, or moving the robot, phantom peaks often showed up in a thresholded grid. The first question we needed to address when using a nearest-neighboring clustering algorithm, was, therefore, how do we separate these from valid sound source positions?

The second question we needed to address when building the algorithm was discovered while watching the evidence grids form in real-time with more than three sources in the environment. As each new measurement was collected by the robot and added to the grid, all of the sources would appear for a time while the robot was in close proximity, but then, as the robot moved away from the source, one or more sources would be suppressed by new measurements. The reason for these suppressive effects over time is two-fold. First, it has to do with the nature of sound, as the sound volume and resulting GCC measurement will naturally fade both with distance from the source, and, of course, with variations in the source volume. Second, this suppressive effect is then further exacerbated by linearly scaling the GCC data. As mentioned in Section 4.1.2, each spatial likelihood measurement is most strongly associated with a single sound source. By scaling between two set values, however, each measurement that points at one source will effectively suppress the evidence grid in other areas not being pointed at, including areas containing another sources. Therefore, if too few measurements point at a source because it is too quiet or too far away, then the cumulative effect of the suppression may end up being greater than the cumulative positive effect.

Steps in the Iterative Clustering Algorithm

Step 1. Scale and Threshold the map.
Step 2. Cluster points together using nearest-neighbor clustering.
Step 3. Localize the largest source.
Step 4. Identify spatial likelihoods pointing at the largest source.
Step 5. Create a new evidence without largest source.
Step 6. Repeat until all sources are found.

To overcome both the phantom peak problem and the suppressive effect over time, we developed an iterative approach to source localization, extracting one source at a time from the evidence grid. By taking advantage of the angular nature of spatial likelihoods, we can then match newly extracted sources with the spatial likelihoods that point at them. Firstly, this allows us to find more sources that might have been suppressed by those samples. Secondly, a minimum number of associated samples can be used as a convenient baseline for eliminating phantom peaks, which typically are created by one or two very strong echoes.

- **Step 1 - Scale and Threshold the Map**

  To prepare the map for clustering, it is first scaled so that the most likely point is no more than 99% likely, and the least likely point is no less than 1% likely. A threshold of 75% is then applied to the map (or 1 in a log-likelihood grid) to eliminate points unlikely to contain a sound source. Lower thresholds were also tested, but often led to joint clusters when sound sources were located too close together. Figure 4.9 demonstrate an evidence grid before (Top) and after (Bottom) thresholding.

- **Step 2 - Cluster points together using nearest-neighbor clustering**

  A nearest-neighbor clustering algorithm is then used to collect all points together that are within 0.3-m of each other. Appendix B.2.1 provides algorithmic detail on accomplishing this clustering task.

- **Step 3 - Localize the largest source**

  A weighted centroid of the largest cluster is calculated using the likelihood at each grid cell as the weight. Appendix B.2.1 describes how to calculate this weighted centroid as part of the clustering process. If the cluster is larger than 0.5-m^2 in area (determined empirically), then it is identified as a potential sound source and its centroid is used as the source position. If the area is too small, then no sources were successfully detected using this map. Figure 4.9 (Bottom) demonstrates the thresholding and clustering process.

- **Step 4 - Identify the set of samples that point at the largest source**

  While every spatial likelihood measurement does contain information about multiple sources (and echoes), each spatial likelihood measurement at the time of sampling. Appendix B.1.1 provides pseudocode for estimating the most likely angle detected by a single spatial likelihood measurement. So for each measurement we can calculate the most likely angle to the "loudest" source by compressing the log-likelihoods along the angular axis at some increment δ. Let F_θ be the log-likelihood of the

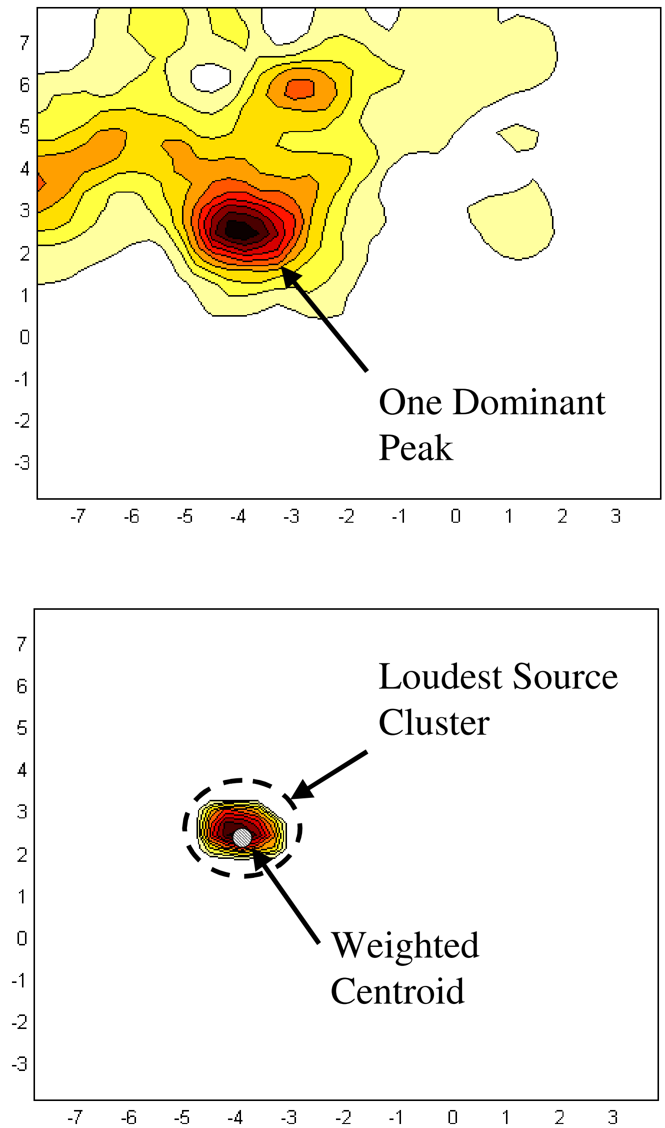


Figure 4.9. Stepping through the iterative clustering process, first round. (Top) Original auditory evidence grid with one dominant peak. (Bottom) Thresholded result with largest cluster circled.

detected source being located along angle θ, and $F_{\phi,r}$ be the log-likelihood of the sound source being located at cylindrical coordinates (ϕ,r). Then the most likely angle towards the detected source is the angle (θ) with the highest log-likelihood:

$$\left|\theta_{best} + \theta_{robot} - \alpha_{source}\right| \leq threshold \qquad \text{Equation 4.9}$$

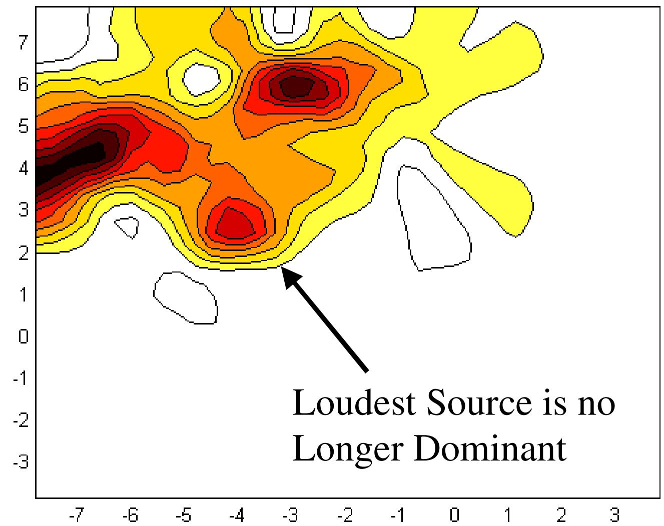
Now, using this notion of most likely source angle, we can determine which spatial likelihood measurements actually point at sources found:

$$F_{\theta} = \sum_{r=0}^{3} \sum_{\phi=\theta-\delta/2}^{\theta+\delta/2} \frac{F_{\phi,r}}{1-F_{\phi,r}} \qquad \text{Equation 4.10}$$

where θ_{best} is the most likely angle as predicted by the spatial likelihood function in local coordinates, θ_{robot} is the orientation of the robot in global coordinates, and α_{source} is the angle from the robot to a detected source in global coordinates. If the difference between the angle to the source location (as predicted by the evidence grid) and the most likely angle (as predicted by a single spatial likelihood measurement) is less than some threshold, then that measurement is estimated to be pointing at the source.

- **Step 5 - Create a new auditory evidence grid without the largest source**

  Using just those spatial likelihood measurements that are not estimated to be pointing at a previously localized source, create a new auditory evidence grid. Figure 4.10 demonstrates a new auditory evidence grid created from this reduced sample set. Appendix B.2 describes how, in



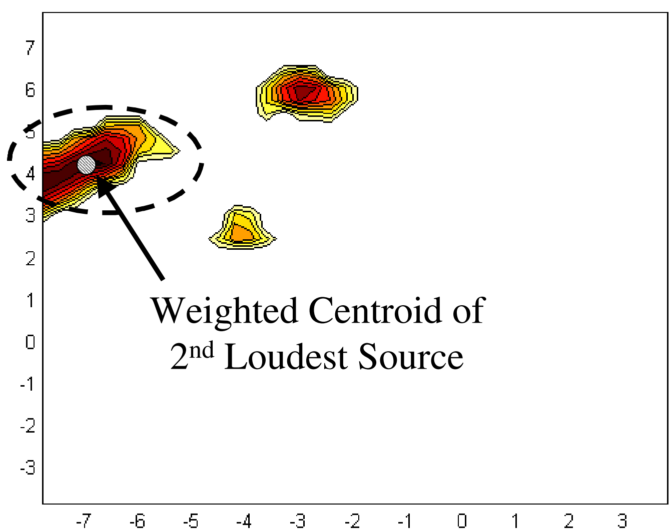


Figure 4.10. Stepping through the iterative clustering process, second round. (Top) Original auditory evidence grid with one dominant peak. (Bottom) Thresholded result with largest cluster circled.

pseudocode, source information can be incorporated into the auditory evidence grid creation algorithm.

- **Step 6 - Repeat steps 1-5 until all sources have been localized**

  Knowing when to stop this iterative process is the difficult part. There are a number of properties that can be used to predict the end of this iterative process:

  1. The largest remaining cluster belongs to a source that has already been detected and removed during this iterative process.
  2. The largest cluster remaining has a very large variance since it was formed from samples dominated by reverberant sound. The variance is defined in Equation 4.11:

  $$V = \sum_i W_i (x_i - \mu_i)^2 \qquad \text{Equation 4.11}$$

  3. where, for all locations (i) included in the cluster, (W_i) is the log odds probability predicted by the auditory evidence grid for that location, (x_i) is the centroid for that cell in the auditory evidence grid, and (μ_i) is the centroid of the detected cluster.
  4. The numbers of samples pointing at the largest remaining cluster is very small, since the peak was formed from samples pointing in arbitrary directions.

  None of these properties are singularly perfect at identifying when no more sources are remaining in the environment to be localized. Together, however, they can usually detect the end of the iterative process.

Therefore, each of the previous steps should be repeated until either no source is localized during step 3, the largest remaining source is already known or has very large variance, or the number of remaining measurements used in step 5 is too small. For this work, a minimum of 20 samples was required for mapping. Figure 4.11 demonstrates a third source detected by iteration (Top), and a final evidence grid on which the iteration stops due to large variance (Bottom).

This iterative approach to localizing stationary sound sources worked well when applied to data collected from the first phase of testing. Except for the test localizing air vents, this iterative clustering approach could extract all of the sound sources from maps creating using either of the sampling strategies. Furthermore, the accuracy appeared high for most extracted sound sources, generally within 0.3-m (note that the source position was only recorded to within 0.3-m accuracy). The one source that was not localized well was a human speaker in the presence of other noise sources. The iterative approach still found a separate cluster for a source within the vicinity of the speaker, but possibly due to movement by the source or the scarcity of samples primarily indicating the human target, at least one localized source was off by 1-m. The next sub-section will focus more on this question of accuracy.

Ultimately, what should be gathered from these results and the results of the next section is that, in the short-term, this iterative approach to localizing stationary sound sources will prove highly useful to an acoustically-aware robot. It allows a robot to localize stationary sound sources in the environment, filling a critical gap in the mathematical framework for sound propagation discussed in Chapter 3. As such, we will

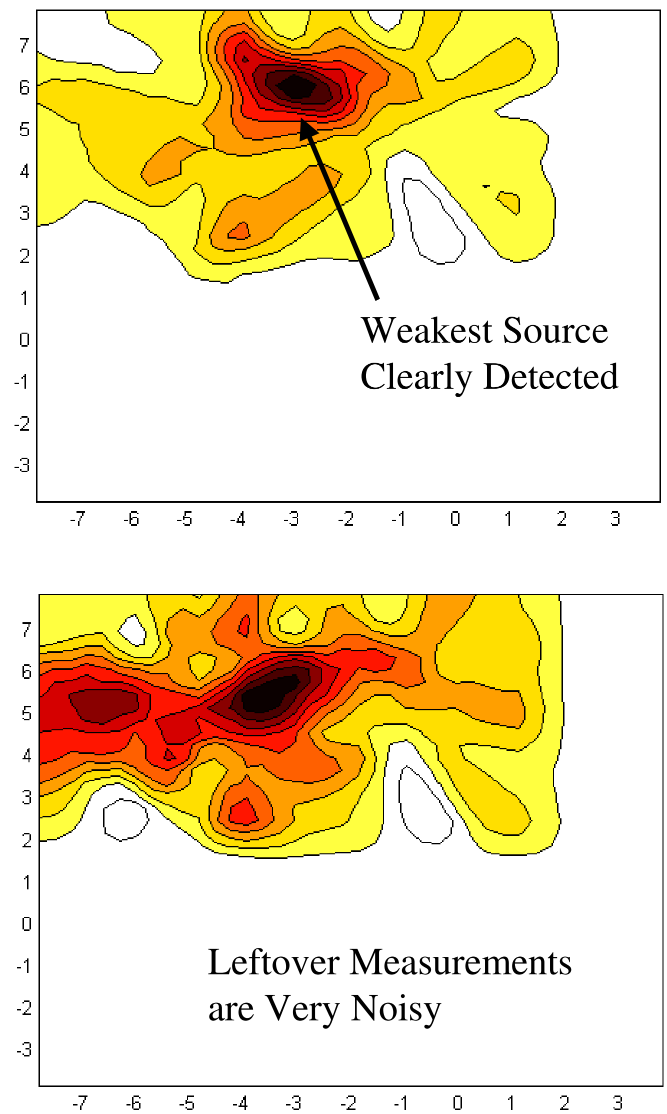


Figure 4.11. Stepping through the iterative clustering process, conclusion. (Top) Third evidence grid, focusing on single source. (Bottom) Final evidence grid on which algorithm stops, because it is very noisy and created from only 8 samples. Notice that the level of noise in the measurements makes separating the sources very difficult.

continue to use auditory evidence grids, and the accompanying iterative clustering approach, for the remainder of this dissertation. However, the results also reveal that there remains work to be done in this area. For one, the current approach has problems in localizing large numbers of sources in the environment. This could be due to a poor sensor model choice, so a better sensor model that incorporates models of sound propagation or environmental effects may actually the need for this iterative heuristic. Alternatively, the use of greedy exploration techniques designed to flush out an evidence grid [Thrun et al. 2005] in combination with short-duration auditory evidence grid creation may serve the same purpose.

A second problem that was highlighted by the tracking of human speakers was localizing sources in the presence of moving sound sources. Although it is explicitly stated that this technique is designed to localize stationary sound sources, moving sound sources are likely to be present in many of the environments. If the source is always moving around, then the source should not be localized, or interfere more than any other type of noise with the localization of other sound sources. If, however, the source remains still for some period of time, and then moves again, like humans or robots in the environment, then that source is likely to have a larger impact on source localization. Not only will it appear as a source in the evidence grid, but it could also mask other sources that a robot is trying to localize. Given that knowledge of moving sound sources is of general interest to a mobile robot anyways, future work in mapping sound sources in the environment needs to incorporate models of moving sound sources into the representation.

Testing Phase 2 – Determining Accuracy

Where the first phase of testing focused on identifying requisite hardware and control strategies, the second phase of testing focused on determining the accuracy of the resulting estimates, when collected by an autonomous robot. While a human may sometimes be available for guiding a robot, the localization of sound sources should not require human assistance. Therefore, given an autonomous robot, what are some different control strategies for acquiring localization information, and how does the accuracy vary between strategies?

The two strategies tested here for autonomous localization of sound sources are a waypoint path and an area-coverage heuristic, each having a different goal in the localization of sound source. The purpose of the waypoint-path is to quickly cover a large area, identifying potential sound source locations in the environment. The purpose of the second algorithm, an area-coverage heuristic, is to spend more time in the vicinity of the source, verify that a source is present, and more accurately identify source properties. Combined, the hope is that the two robotic movement strategies will present a common strategy for localizing unknown sources in an arbitrary environment. The robot follows a waypoint path, effectively patrolling the environment, until something is detected, at which time it changes to an area-coverage heuristic focused on investigating the potential sound source identified during while patrolling the area.

The data used for this phase of testing was originally reported on in [Martinson and Schultz 2007]. Since the original publication, this section has been updated to reflect recent modifications to the algorithm.

Experimental Setup

In this second stage of testing, the same B21r robot used in the first stage was again used in the AI Center Laboratory at the U.S. Naval Research Laboratory (NRL). The layout of obstacles was slightly different from the first stage, however, so as to allow the robot access to multiple sides of the sound source. Figure 4.12 illustrates this modified layout.

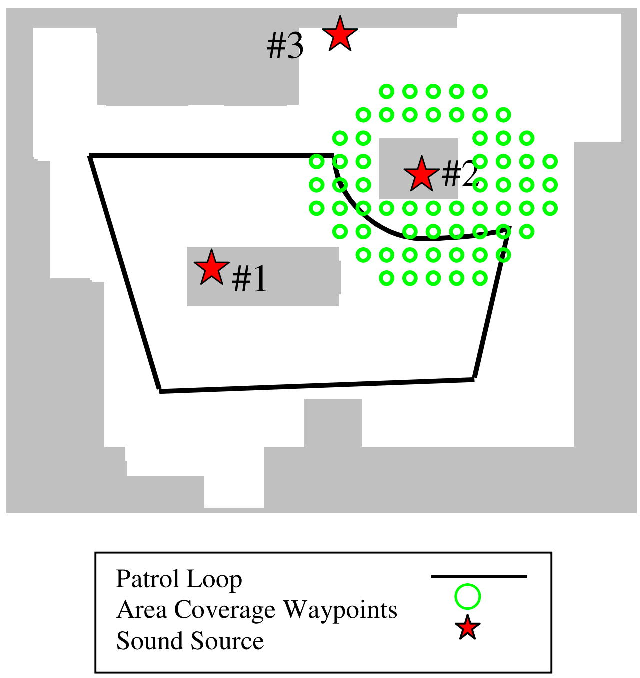


Figure 4.12. An overlay of the NRL environment, showing an example waypoint path, set of area coverage target points, and source locations.

During this second phase of testing, the B21r was used to localize each of three pc-speakers with unknown $\{x,y,\theta\}$ 5 times, for a total of 15 trials. Figure 4.12 illustrates their respective positions in the environment. During any one test, only one speaker was playing and all speakers played the same nature sounds track (rain) at a 65dB volume. For each test, the robot first followed a waypoint path, quickly estimating the location of the sound source in the environment. Then it would dynamically choose where to center and execute its area-coverage heuristic based on the waypoint-path localization results.

Waypoint Path

The first type of autonomous movement is the waypoint-path, described by a set of ordered waypoints in the environment for the robot to visit (see Figure 4.12). The purpose of this phase is to expose the robot to as much of its environment as possible so that it will be able to detect any significant ambient noise sources.

Provided with a waypoint path, the robot uses a path-planner (Trulla, [Hughes et al. 1992]) to guide it from its current position to each waypoint in turn while dynamically avoiding obstacles. Upon arriving within some threshold distance (0.4-m) of the desired waypoint, the robot selects the next waypoint in the specified order as a target, and the cycle repeats. To account for inconsistencies between the real world and the map, a timeout mechanism monitors the robot progress and forces it to move on to the next waypoint after 3 minutes. The task is finished when the robot has successfully visited or tried to visit all specified waypoints. After completing one loop through the environment, the robot then processes its auditory data using the auditory evidence grid and iterative clustering process to search for likely source position candidates.

The expected goal of this type of autonomous movement is the quick localization of possible sound sources. As such, the robot sampled while moving through the environment to reduce time, resulting in an increased localization error. After quickly patrolling the environment (collecting an average of 40 samples per run) the robot was still able to identify an approximate location for the source in every trial (Table 4.2).

Table 4.2. Mean localization error when auditory evidence grids are used with data collected by a moving robot.

	Localization Error (m)
Source 1	1.0
Source 2	0.93
Source 3	1.5
Combined	1.1

Unfortunately, however, this quick patrol strategy produced a relatively high error over each of these trials, even though it was localizing only a single source. This error was due to the very limited number of samples recorded in the vicinity of each sound source. Usually, less than half of the samples were even within the 3-m range over which the spatial likelihoods were calculating, and even then the distance to the source varied significantly. Source 3 was particularly poorly detected for this latter reason, as the robot never came closer than 1.7-m to a source situated on a bookshelf next to the wall.

Given the high error demonstrated during these tests, it is clear that this strategy of using a moving robot to quickly localize the source is not going to produce accurate enough results for estimating sound flow. However, what this patrol strategy did do was correctly localize something in the vicinity of the source. Therefore, to overcome the error introduced by this scarcity of data, a second type of robotic movement is needed.

Either the robot needs to travel slower, and pause and sample, along the waypoint path, or, when the application has enough time, the robot can use the coordinates provided by the quick pass through the environment to thoroughly investigate the source and improve accuracy. This second type of investigatory movement is explored in the next section.

Area-Coverage

The second type of autonomous control is a directed investigation of each source using an area-coverage heuristic. As it is impractical to investigate the entire lab space, this type of control requires an initial target around which to center the area-coverage activities. For this purpose, we used the results of the waypoint-path phase previously described. Therefore, once a coordinate was identified, the robot would move to that area and begin the area-coverage task. The goal of this control strategy was to acquire as many samples pertaining to this sound source as possible, so as to ideally improve on the localization result provided by following a waypoint-path.

Provided with a target set of sound source coordinates to investigate, a set of unobstructed locations is identified within a 3.5-m radius of the target using the obstacle map of the environment. A smaller radius could also be used when identifying unobstructed locations, but the goal of this phase was to identify the location as accurately as possible, so a large area was selected for investigation. These unobstructed locations become waypoints for the robot to visit, effectively performing an area coverage task in the vicinity of the suspected sound source. Unlike the waypoint task, however, visiting these waypoints does not need to be done in any particular order, and so the robot will always travel to the nearest waypoint. The circles in Figure 4-10 show a set of waypoints to be used for investigating a single source. Also unlike the waypoint task, the robot will stop at each target to collect samples. Movement during sampling introduces additional error to the estimate (as discussed during phase 1 of testing), so stopping the robot should improve accuracy at the expense of time. Appendix C.3 describes the area-coverage heuristic in more detail.

After completing the investigation of a single source, the robot now has enough data to refine the position of the source using iterative clustering. The accuracy of the entire process, from waypoint path to investigated source, is given in Table 4.3 for each of the three sound sources:

Table 4.3. Mean localization and orientation error as produced by the discovery process.

	Localization Error (m)	Standard Deviation (m)
Source 1	0.32	0.26
Source 2	0.13	0.20
Source 3	0.19	0.22
Combined	0.21	0.23

Even though the initial investigation coordinates may have had a high error due to the quick nature of the waypoint-path estimation, the resulting estimates after performing an area-coverage task in the vicinity have relatively low error, and are ideal for building estimates of the direct or reverberant fields. Figure 4.13 demonstrates an example evidence grid created from a directed investigation of the source using area

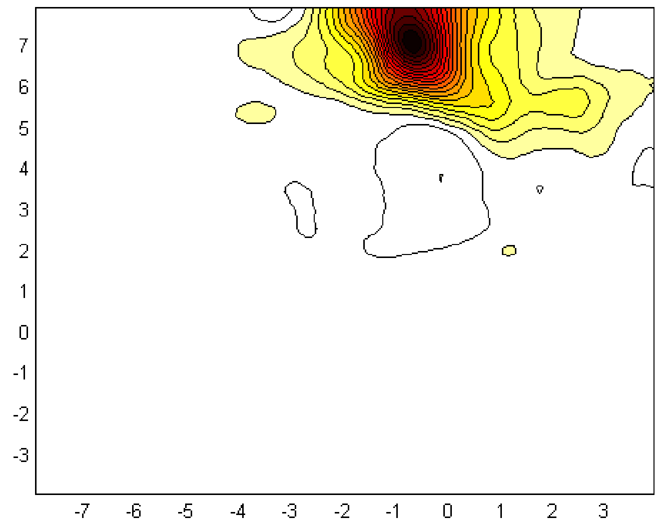


Figure 4.13. Auditory evidence grid created from 137 samples collected during a directed investigation of a source using an area coverage heuristic.

coverage. Notice the smooth, singular nature of the peak, which makes it easy to extract using nearest-neighbor clustering.

4.2.2 Sound Source Volume and Directivity

After determining the position of the sound source, the next step is to determine the volume of (i.e. sound pressure level generated by) the sound source so that we can estimate the effects of the direct field. If there is little time or need for an accurate model of the source, then the simplest source model is an ideal omni-directional source. Using the average sound pressure level of a few samples taken at known distances from the source, a relative volume can be quickly established and the source inserted into our models of the auditory scene. This simplistic model of volume, however, is far from being correct. Most sources, usually due to their physical shape, are not omni-directional, meaning that the direct sound produced by that source varies in volume depending upon the angle to the source. Therefore, without a model of directivity to estimate how the volume changes with direction, averaging a few samples collected at random locations around the source will not produce a good estimate of the source volume.

The challenge in building a model of source directivity and volume is the difference between the ideal method for constructing such a model and the actual nature of the data from which to construct it. In the ideal method for determining source directivity, the sound source would be located in an anechoic chamber where the magnitude of any reflections is negligible, and the sound could be measured at a constant distance from the source. With the robot, however, we are in a real environment where there is a substantial reverberant component to measured sound. Furthermore, due to the presence of obstacles in the environment, the collection of data gathered comes from an arbitrary set of distances and angles to the source. How do we overcome these differences? The solution, originally described in [Martinson and Schultz 2007], is to work backwards from the sound source propagation model discussed in Chapter 3.

Determining Directivity - Algorithm

When the robot records a sample of the auditory scene, it is actually measuring some energy from the direct field of each sound source in the room, plus some amount of reverberant energy and some amount of transmitted energy. Equation 3.2 described this sum in terms of field effects. Equation 4.12 re-writes this equation in terms of pressure:

$$p_s^2 = \sum_i \left(p_{direct_i,s}^2\right) + p_{reverb,s}^2 + p_{trans,s}^2 \qquad \text{Equation 4.12}$$

Where p_s is the rms pressure of the sample (s), $p_{direct\_i,s}$ is the rms pressure due to un-reflected sound from source i, $p_{reverb,s}$ is the rms pressure due to reflected sound waves, and $p_{trans,s}$ is the rms pressure due to transmitted sound. The loudness of the direct sound for one particular source is the quantity we are the most interested in, so we will be limiting the set of samples used to those in close proximity to the source being modeled. As a simplifying assumption, we will, for now, assume that active sources are not located close together, therefore we can ignore the effects of their direct field on samples taken near another source. We will also assume that transmitted sound is negligible, leaving only the direct field component for a single source and the reverberant field:

$$p_s^2 = p_{direct,s}^2 + p_{reverb,s}^2 \qquad \text{Equation 4.13}$$

To remove any more components from the equation is impractical. The direct field component is what we need for estimating directivity, while the reverberant component will almost always be too large to ignore when estimating the volume. As is, however, this equation still has too many unknowns. To identify the separate components, we need additional equations provided by 2 simplifying assumptions. The first such assumption is that the loudness due to reverberant sound will remain constant over the entire room. Since reverberant sound describes the contribution of reflected sound waves, and sound waves will reflect many times all over the room before either decaying to nothing or reaching a receiver, this is a good first approximation often used for estimating the reverberant field [Raichel 2000].

The second simplifying assumption involves the contribution of the direct field. As the direct field describes the volume of un-reflected sound emanating from the source, conservation of energy (p_{rms}^2) tells us that the energy of the direct field should decay linearly with square of the distance. So the farther away the robot is from the source, the greater the energy coming from reverberant sound and the less from direct sound. Equation 4.14 illustrates how to estimate the original volume of the source at a distance d_0 given an actual distance d_s.

$$d_0^{\ 2} p_{direct,d_0}^2 = d_s^{\ 2} p_{direct,s}^2 \qquad \text{Equation 4.14}$$

Therefore, we can also assume that after some distance the contribution due to the direct field is minimal, and then estimate SPL_{reverb} as the mean volume of the sampled data taken beyond d_r meters from the source. In this work, we used two meters as a good approximation, where the volume of the direct field will have dropped 6dB from the volume at 1-m from the source.

Now that we have estimated the contribution of the direct field, the final step is to combine all of the data collected from arbitrary distances and angles into a single model estimating volume for any specified distance and angle. For this purpose, we first use Equation 4.14 to calculate $p_{direct,d0}$ at the specified distance d_0, and then we apply a Gaussian smoothing function centered on the desired angle (ω). The final equation for the model of directivity is:

$$p_{direc}^2(d_0, \omega) = \frac{\sum_s e^{-(\theta_s - \omega)^2 / 2\sigma^2} d_s^2 p_{direct,s}^2}{d_0^2 \sum_s e^{-(\theta_s - \omega)^2 / 2\sigma^2}} \qquad \text{Equation 4.15}$$

Where (d_s, θ_s) is the position of the sample relative to the center of the source, and σ is the standard deviation of the applied Gaussian. Figure 4.14 demonstrates a

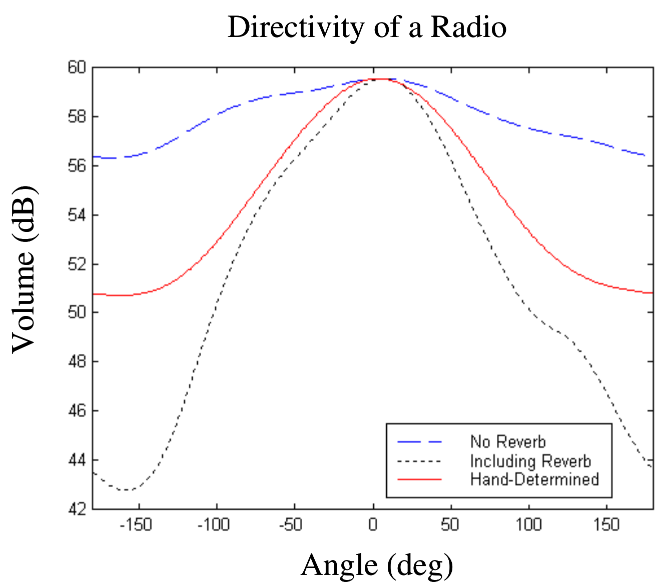


Figure 4.14. Comparison of robot-created directivity models using different reverberation assumptions, with a hand-measured directivity model. The three models assumed a maximum volume of 59.5-dB, as measured by the sound pressure level meter.

directivity model for a pc-speaker at a distance of 1-m. Appendix B.3 presents the pseudocode for creating this directivity model from sampled data.

Given this directivity model, the volume of the source is angle dependant, i.e. the volume is determined directly from the directivity equation using the known angle to the source. Furthermore, this same model can account for changes in volume of the source over time. If the robot returns to the area some time later, it can re-measure some small area (preferably away from the walls), and use that new volume (V_m), measured at a known position (d_m) and angle (ω_m), to add a constant multiplicative offset to the directivity model to reflect the change in the sound source:

$$p^2_{direct}(d_0,\omega) = p^2_{direct,0}(d_0,\omega)\frac{V_m}{p^2_{direct,0}(d_m,\omega_m)} \qquad \text{Equation 4.16}$$

In the same manner, this representation of directivity can also handle differences between measuring equipment. Sensitivity differences between microphones is handled similarly. If two robots measure the same source using different microphone arrays, then their results can be compared by using a similar offset and the difference in estimated volume between the two microphones.

In Figure 4.14 that constant offset is used to compare three different directivity models of the same radio at a 1-m distance. The solid line is created from hand-measured data using a directional sound pressure level meter (Type II accuracy). The dotted line uses robot-measured data with Equation 4-15 and all of the stated assumptions. Finally, the dashed line uses the same robot measured data and equation, but assumes that the reverberant component is negligible (i.e. $p_{reverb} = 0$). What this figure demonstrates is the effect of the reverberant field assumption on our resulting model. Each of the directivity models demonstrate the same cardioid centered at 0°, which should be the result for a radio speaker. However, using the hand-measured model as the ground-truth, assuming a negligible reverberant field underestimated the difference between peak volume and minimum volume. In contrast, assuming a constant reverberant field overestimated the difference in volume.

In practice, the actual difference between these two assumptions and the ground truth is likely to depend heavily on the type of environment. For example, a smaller environment with similar materials may better fit the constant reverberant field assumption. In the future, improving this model should probably incorporate more information about the nature of the reverberant field. As will be demonstrated in Chapter

5, however, the constant reverberant field assumption is good enough for many applications.

Identifying Source Orientation

Given the variation in the resulting directivity models due to environmental effects, experimental testing of the source modeling process was focused on the correct identification of source orientation. If a robot can identify the direction of maximum volume (i.e. the source orientation), then avoiding or maximizing the effects of the sound source is possible, regardless of the noise in the rest of the directivity model. As such, testing this source orientation detection was divided into three stages. In the first stage, we tested the accuracy at which source orientation was estimated by using a mobile to investigate one source with known $\{x,y\}$ and unknown θ. In the second stage, we tested the autonomous localization and modeling process for a single source of unknown $\{x,y,\theta\}$, using the same data from the phase 2 testing in Section 4.2.1. Finally, in the third stage, we tested the ability of the robot to localize, and identify source orientation of multiple simultaneously operating sources of unknown $\{x,y,\theta\}$.

The first two stages use data collected by the B21r for localizing sources, while the third stage uses the Pioneer2-dxe robot in a different environment, so as demonstrate generality. During each of these three stages, we applied a 10^{th} order highpass FIR filter (300-Hz cutoff frequency) to every sample before analyzing the data. Since the ambient noise sources being measured had significant high frequency components, the filter had little effect on the auditory evidence grid creation. What the filter did do, however, is reduce the impact of robot motor noise on determining directivity. Since the robot's motor was in close proximity to the microphone array, it could overpower the weak volumes measured farther away from the source.

Stage 1 - Known {x,y}, Unknown θ

In this first stage of testing, a single source of known centroid position was rotated through 7 different angles in 45° increments. Provided with the ground truth source location, the B21r was used to investigate the source once for each different angle using just the area-coverage algorithm with a 3.5-m range. One angle was not tested due to the source pointing at a solid wall where the robot could not investigate. The sound source used in this stage was a pc-speaker playing nature sounds (rain) measured as being 65 dB at 1-m from the source (including both direct and reverberant sound).

Over 7 trials, the mean error for estimated source orientation was 0.2-rad of ground truth with a maximum error of 0.5-rad. Given that the source itself is a pc-speaker with a wide frontal lobe, this approximation should be adequate to guide the robot away from the loudest areas surrounding the source.

Stage 2 – Unknown {x,y,θ}

In the second stage testing, the B21r was used to localize each of three pc-speakers with unknown {x,y,θ} 5 times, for a total of 15 tests. During any one test, only one speaker was playing. All speakers played the same nature sounds track (rain) at a 65-dB volume. For each test, the robot first moved along the same patrol route, localizing the active source. Then it would dynamically choose where to center its investigation using area-coverage. After sampling the area, the sound source orientation was estimated along with the original location. This same data was used reported earlier for estimating source location coordinates using an auditory evidence grid. The layout of the room and the source positions can be seen in Figure 4.12. Table 4.4 shows the mean error in source orientation, for each source:

Table 4.4. Mean error in identifying the direction of maximum volume, as produced by an area coverage task.

	Mean Orientation Error (rad)	Standard Dev. Orientation Error (rad)
Source 1	0.22	0.23
Source 2	0.18	0.08
Source 3	0.32	0.23
Combined	0.24	0.17

These results demonstrate the reliability of the discovery process in accurately finding and modeling sources. Sources 1 and 2 were located in areas where the robot could completely encircle the source, and therefore gather data from all directions. Source 3, however, was against a wall, so the robot was limited to gathering data in the 180° foreground. Due to this limited area, as well as the proximity to the wall and its echoic effects, the orientation error is highest for this third source.

Stage 3 – Multiple Sources

The final stage of robotic testing demonstrated the ability of the robot to detect multiple simultaneously operating sources and identify their characteristics. Two

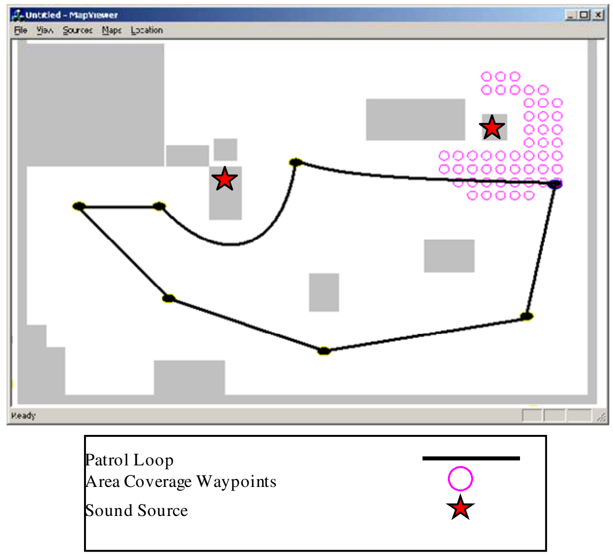


Figure 4.15. Hand coded obstacle map used by the pioneer for navigation in an environment with two sources. The black line shows the waypoint path followed during a patrol, while the circles illustrate a set of targets reached by the robot to complete an area coverage task for a single detected source.

sources, an air filter (0.5-m x 0.3-m x 0.3-m) and a two-speaker radio generating static noise, were placed 5.8-m from each other. Figure 4.15 shows their relative placement. The pioneer2-dxe robot (Figure 4.2) was then used to localize and model each source. Following the initial patrol phase, the robot identified two potential clusters, corresponding to each of the two sources. Both initial clusters were within 1-m of the actual source location. Upon further investigation, the robot improved the localization accuracy for the air filter to within 0.2-m, and to 0.4-m for the radio. The orientation accuracy for each source was 0.64-radians and 0.4-radians respectively.

Table 4.5. Localization and orientation accuracy of the two source discovery process

	Localization Accuracy (m)	Orientaion Accuracy (rad)
Filter	0.2	0.64
Radio	0.4	0.4

4.2.3 USING SOURCE INFORMATION WITH THE SOUND FIELDS FRAMEWORK

So far in this section, we identified three different properties of sound sources that could be used in conjunction with the sound fields framework to estimate the flow of sound through an environment: source location, average volume, and directivity.

The position of the sound source was estimated using an auditory evidence grid in conjunction with an iterative nearest neighbor clustering algorithm. By itself, however, the source location does not contribute much to the sound flow estimation problem. In Figure 4.16 (Top), the shape of the direct field is plotted for some distance around a detected sound source of unknown volume. The displayed map assumes that the receiver can reach any location around the source (i.e. the geometric layout is unknown), and that the source is omni-directional. These estimates of sound flow without knowledge of the volume can certainly be used to guide a robot to or away from a sound source, but they are not very realistic. Without volume, there is little way to compare sources (unless they are assumed equally loud), or estimate how much noise a robot might be exposed to as it moves through the environment.

Volume, however, is not necessarily difficult to estimate for a source that does not vary over time. As discussed in Section 4.2.2, a simple estimate for volume is to collect

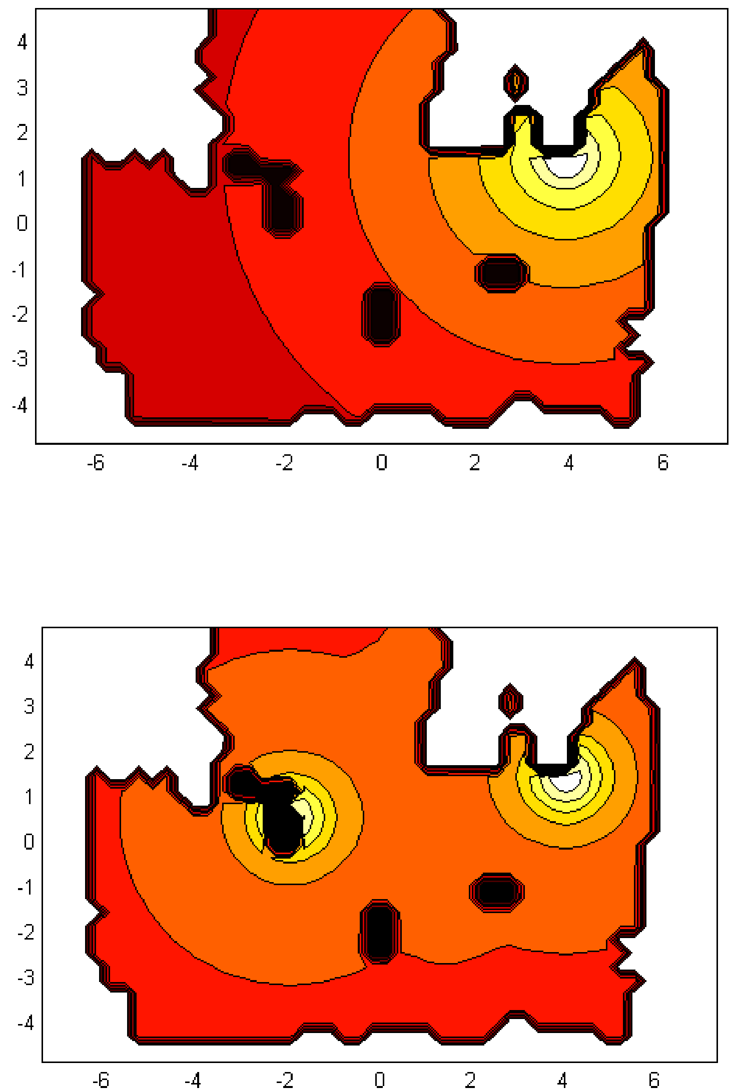


Figure 4.16. Direct field estimates created from a single source of arbitrary volume (Top) and two sources of known volume (Bottom). The obstacles in these field estimates are there for scale purposes, and were not used to predicting occlusions or reverberant effects.

some number of samples in the vicinity of the source and average the result to estimate volume. Using this simple method for estimating volume, and still assuming omni-directional sources, we can estimate the combined volume of two detected sound sources over an arbitrary sized environment (Figure 4.16, Bottom). Still estimating only the direct field, the robot can predict that the lowest sound between the sources is not actually in the middle, but rather closer to the quieter fan source on the left.

To increase the accuracy of the sound fields estimate even further, the last piece of information discussed, which the robot can gather, is the directionality of the sound source. Although we can estimate the volume very simply for each source, the perceived volume should actually depend on the angle from the receiver to the sound source. Given time to investigate a sound source and collect enough samples from a variety of angles, a robot can build a model of directionality, predicting the volume detected at each angle to the source. Figure 4.17-19 demonstrates how knowing the directivity of each of the sound sources can be used to build directional fields for each source, and then combined together into a representation of overall sound flow due to direct sound. Appendix B.5 describes the pseudocode implementation of the algorithm used to construct each of these direct field models.

In general, what this information about sound sources allows a robot to do is make some predictions about the shape of the acoustic environment in the area where it may need to travel. For instance, the robot can use this modeling ability to predict the regions of loudest environmental noise, so as to either avoid or move into areas of loudest sound. These two applications will be demonstrated in Chapters 5 (Improving the Signal-

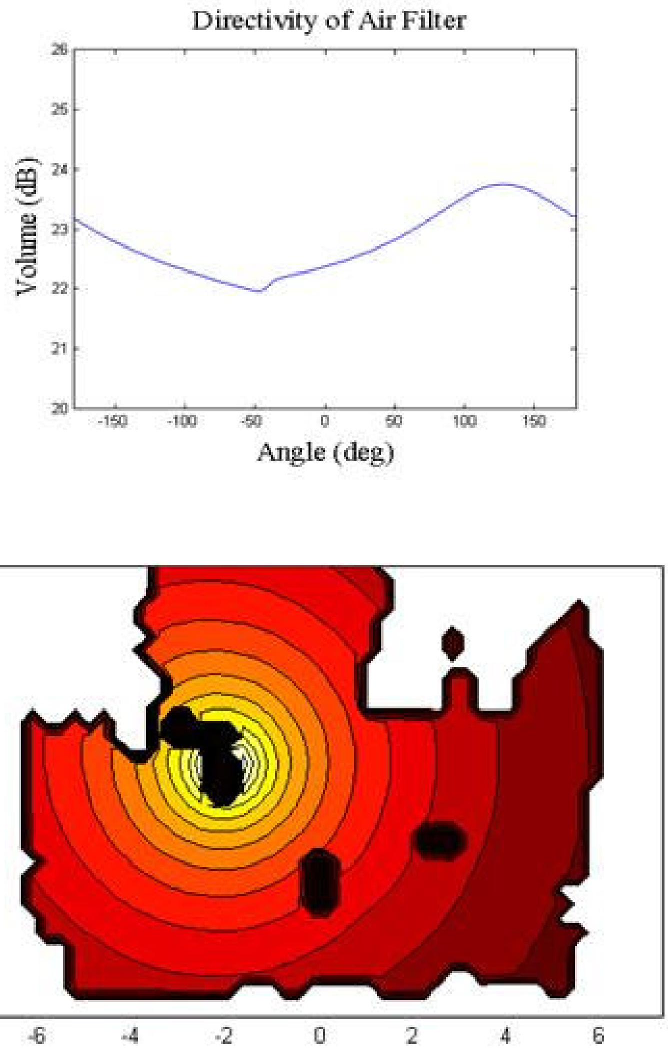


Figure 4.17. Process of creating sound propagation models from sampled area coverage data, part 1. (Top) Directivity results for the mostly omni-directional air filter. (Bottom) Direct field for the air filter, showing the spherical spreading common to mostly omni-directional sources

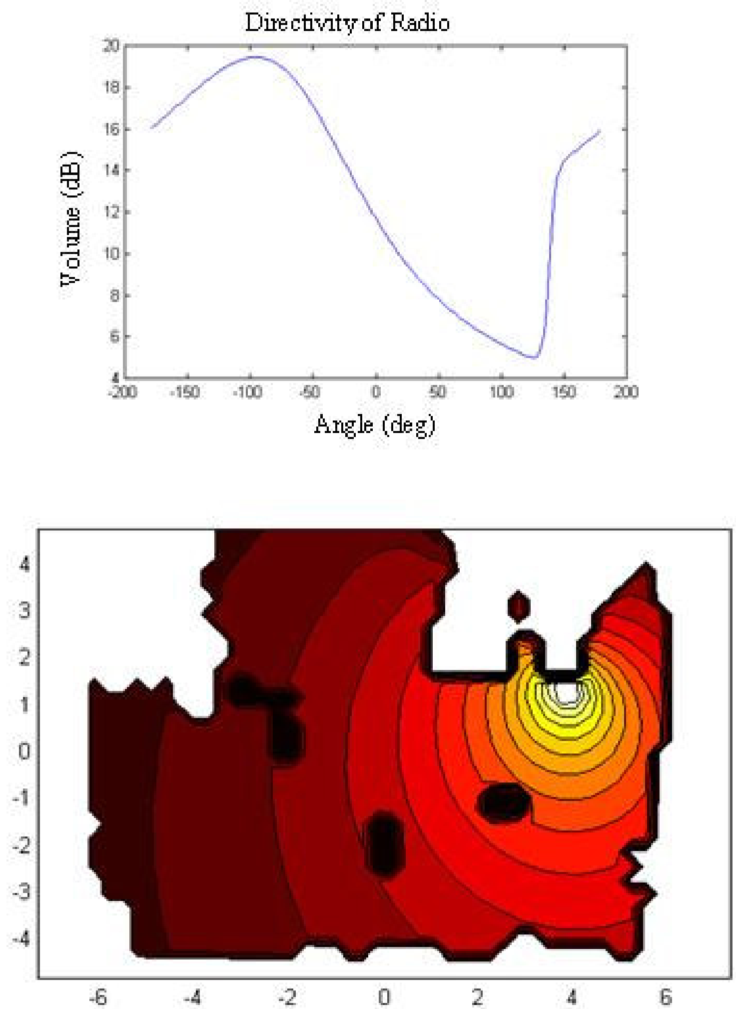


Figure 4.18. Process of creating sound propagation models from sampled area coverage data, part 2. (Top) Directivity results for the radio. (Bottom) Direct field for the radio, showing a distinctly louder region in front of the radio.

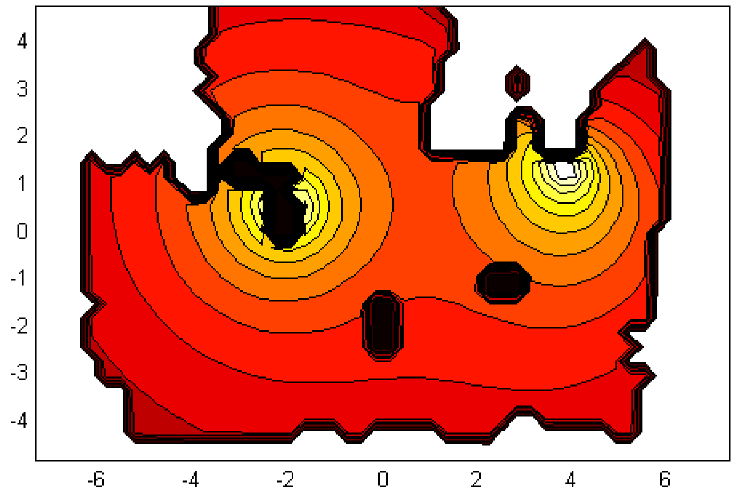


Figure 4.19. Process of creating sound propagation models from sampled area coverage data, part 3 - combined direct field for both sources

To-Noise Radio) and 6 (Stealthy Approach). Another possible use of this information is predicting which sound source will be loudest in a given area. By knowing what source dominates each region of the environment, a robot can detect and track changes to the auditory scene over time (Acoustic Monitoring – Chapter 5). This information that the robot can gather about active sound sources in the environment serves as the basis for many applications. Although receivers and environments are important for refining accuracy, knowledge about the sound sources is critical for virtually all applications reacting to or making use of sound levels in the environment.

4.2.4 IDENTIFYING AND REPRESENTING SOUND FUNCTIONS

The final piece of information about the sound source that we will be using in this dissertation is the sound function. In general, the sound function refers to the sound being generated by a sound source at any given time, including the volume of the sound source. For example, the sound function of a radio playing a single song is the recording of that music, modified by the volume at which it is being played, the properties of the amplifiers/speakers playing the song, and the time at which it is started. Therefore, for any non-repeating source, knowing the sound function exactly would require knowing an infinite stream of data. If the exact output of the sound source is not needed, however, then the sound function can be reduced down to a more compact representation depending upon the application.

In this dissertation, we will use the source sound function for classification. Assuming that our robot has collected some number of samples from the environment, our goal is that the robot should be able to use its representation of the sound function to determine which source was loudest in each of the collected samples. Furthermore, we would like to use the representation of the sound function to separate out any new sources from known sources in the environment, and determine whether or not the sound source is on or off. In short, we would like to use the sound source function representation for classifying samples as belonging to, or most importantly, not belonging to any particular source. The representation that has been used successfully by a number of others for this purpose in audio classification is mel-frequency cepstral coefficients or MFCCs [Slaney 1994; Quatiri 2002].

The MFCC Implementation

MFCCs are a classification feature set based on the mel-scale filter bank, a filter bank designed to group frequencies in a manner similar to human perception. In this filter bank, low frequencies are traditionally grouped equal spaced bands, while higher

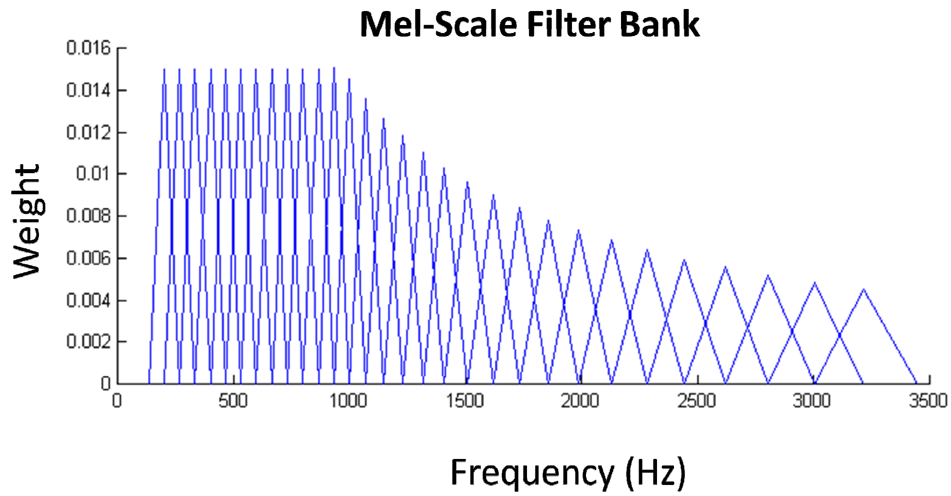


Figure 4.20. Weight vs frequency plot of a mel-scale filter bank.

frequencies are grouped in bands that increase logarithmically in size with the frequency. Figure 4.20 demonstrates the triangular weighting in the frequency band characteristic of a mel-scale filter bank. MFCCs are then calculated by taking the discrete cosine transform of the energy of these mel-scale frequency bands. A good description of how to calculate MFCCs can be found in [Quatiri 2002].

Although the mathematics behind MFCCs is generally the same across all implementations, the specific details of frame size, window size, pre-filtering algorithm, etc. vary from implementation to implementation [Zheng et al. 2001]. In this work, the code for calculating the coefficients was derived from the Auditory Toolbox implementation [Slaney 1994], which can be downloaded from Malcolm Slaney's website[6].

---

[6] http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/, Accessed Aug 25, 2007

- Frame Size – 10-msec
- Window Size – 31-msec (length of 256 at 8192-Hz)
- Pre-Filtering – a hamming window

What these numbers mean is that a single 250-msec sample collected by the robot is run through an iterative algorithm which, for every 10-msec (the frame size):

- Collects the next 31-msec (the window size)
- Applies a Hamming window to the samples
- Calculates the first 8 MFCCs for these samples.

Therefore, for a single 250-msec sample, we get 21 sets of 8 coefficients. We then discard the first coefficient of every set, because it is generally unreliable for classification [Zheng et al. 2001]. Our classification feature set is then the mean and variance of these 7 MFCC's over that sample, for a total of 14 features. Appendix B.4 describes the creation of this classification feature set in more detail.

Typically, to achieve the best classification performance, a longer sample than 250-msec is desired, with 1-sec generally considered long enough to capture the wide range of variation in many typical sources. Our goal, however, was to use MFCC-based classification with the samples already collected by the robot when localizing sound sources in the environment. With a moving platform, even 500-msec samples can introduce a large amount of position error. Therefore, to improve the accuracy of the feature set, each MFCC feature vector is actually calculated over two successive 250-msec samples.

Classification Algorithm

In this dissertation, we make use of nearest-neighbor classification [Duda et al. 2001] to identify sound functions with our MFCC feature set. This classification scheme has successfully been used with MFCCs and other auditory feature vector sets [Ravindran 2006] for distinguishing between a range of sound types including speech, animal noise, and music.

Before a robot can use nearest-neighbor classification to identify sound types, it first needs to build a set of class vectors. A class vector is essentially the mean feature vector for a given sound type that the robot is trying to classify. For instance, if the robot is trying to identify samples that contain noise generated by a particular air filter, it needs to know the average feature vector of samples generated by this air filter or class of air filters. This, in turn, requires that the robot have a set of samples known to predominantly contain noise from the air filter. Luckily for the robot, however, this information is already available.

In the previous sections, we described an area-coverage algorithm used to refine the source localization results and identify directivity and volume of the sound source. This area coverage algorithm involved collecting a large number of samples from many different angles and distances to the sound source. This same collection of sampled data can also be used to build the class vector for the same investigated source. Although some of the samples may be contaminated by other sources in the environment, most of the samples should have been collected within the region most strongly influenced by the target source, and, therefore, should be dominated by that source.

Using the samples collected from the investigation of the sound sources, the robot now has a class vector for every source it investigated. Now, whenever the robot records a new sample, that sample can be classified as belonging to any of the detected sources using the mahalanobis distance [Duda et al. 2001] from the new samples' feature vector to each of the class vectors. The class vector that is closest in distance is therefore closest in sound function to the recorded sample.

Assuming that each of the sound sources has a different sound function, this simple classification strategy works well for distinguishing between known classes. These classes could have been investigated by the robot, or provided *a priori* from hand-sampled data for particularly important sound types like speech. If the task of the robot, however, is to distinguish between known and unknown sounds, then this nearest-neighbor strategy has a problem. When the robot records a sample from a previously unknown source, that sample will still be matched with the nearest-neighbor class vector, which will still correspond to an existing source. To overcome this problem of being matched to existing classes, we inserted a set of 20 random classes into the feature vector set. Now, instead of matching to a known source, there is a good chance that a feature vector belonging to an unknown class will be closer in distance to one of the these random classes.

Classification Results

To test the efficacy of the classification algorithm, we set up a small experiment involving two sources. An air filter with a significant directional component was placed to the right side of the room, generating noise at 50-dB. A second source, a small fountain, was then placed roughly 3-m away at the top end of the room, generating water noise at a measured 54-dB. Figure 4.21 demonstrates the relative placement of these two sources within the environment.

To build a class vector for each of the two sources, a microphone was manually moved about each of the sources to simulate the collection of samples by a moving mobile robot performing an area-coverage algorithm over the surrounding 2-m. Both sources were enabled while the samples were being collected. These samples then served as the basis for a fountain class vector, and a filter class vector. An additional 20 random classes were also generated using the minimum and maximum range of the samples collected by the robot.

After the source functions were approximated with a class vector, additional sampling was performed at each of 10 sample locations distributed about the environment. Roughly the same number of samples were recorded at each location. Figure 4.21 shows the ratio of samples classified as belonging to the fountain vs. the filter at each sample location. Since the fountain was a significantly louder source, more than half of the sample locations detected more samples from the fountain than the filter.

Perhaps the most interesting part of the resulting ratios, however, was when a comparison was made with the predicted direct-field volume of each sound source. Given that the volume, directionality, and position of each source was constant andreadily available, it was a simple matter to estimate the predicted volume of each source at each sample location in the environment. Comparing this ratio to the classification ratio revealed that, at all sample positions where the expected difference in

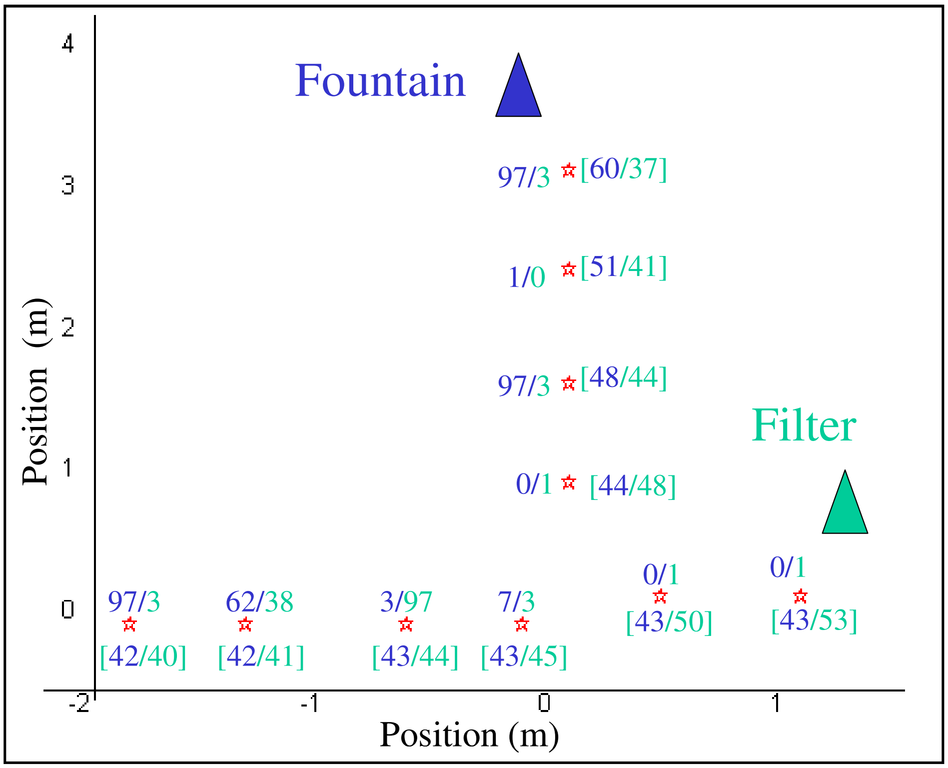


Figure 4.21. Classification results vs. predicted direct field volumes for two sources, a filter (green triangle) and a fountain (blue triangle), at regular intervals (red stars) around the room. On the top and left are the MFCC classification results, with the fountain in blue (nominator) and the filter in green (denominator). On the bottom and right (in brackets) are the predicted volumes of the direct field. At all sample positions where the expected difference in volume is greater than 2-dB, the louder source dominates the classification results. For volume differences less than 2-dB, the classification result is more variable.

volume is greater than 2-dB, the louder source dominated the classification results. For volume differences less than 2-dB, the classification result is more variable.

In general, this test suggests that the predicted volume of the direct-field seems to be related to the classification results. To verify these results, an additional round of testing was performed in the presence of three sources. A third source, a radio playing music in a range of roughly 56-62 dB, was added to the environment at position [-2.5,0]. This time, with the added variability of the music source, there were a lot more sample points with unpredictable classification ratios. Still, however, those sample points that were closest to each of the sources, overwhelmingly favored the nearest source.

These results suggest that MFCCs are a good choice for representing the sound function approximation of constant volume sound sources. When faced with a new sound source, a robot is capable of not only localizing the sound source, but also autonomously building a classification feature vector. Furthermore, using its knowledge of the surrounding acoustic environment, the robot can now also predict where to move to ascertain any changes to the sound function. After determining the loudest locations in the environment for this source, the robot can move to one of those locations, sample, and re-classify the results to potentially determine changes in volume, and/or sound function using MFCCs. This latter functionality will be demonstrated in Chapter 5.

4.2.5 CHARACTERIZING SOUND SOURCES – SUMMARY

In Chapter 3, a general sound source model was presented as being useful to an acoustically-aware robot (Section 3.2.1). For each sound source, the model contains knowledge about the position of the sound source in the room, its directivity, and its sound function. What was not discussed in Chapter 3, however, was where the robot can acquire this information about each sound source. Section 4.2 has now addressed that concern for each part of the model, using a combination of algorithmic data fusion and robotic exploration.

The first part of the model, the location of the sound source, is determined using auditory evidence grids. By collecting samples over a wide variety of locations in the environment and fusing them together in an auditory evidence grid, a robot can identify the presence and general location of new sound sources in the room. Then the robot can improve on its localization accuracy for an individual sound source by investigating the area most likely to contain a source, collecting a large number of samples in the vicinity.

The second part of the model, directivity, is determined using the same investigative sampling technique that served to improve the localization accuracy of the auditory evidence grid. After collecting a large number of samples in the vicinity of the sound source, a robot can use its knowledge of spherical spreading to predict the volume of sound generated by the sound source for each angle. This information, combined with the position of the sound source, can then be integrated into the sound fields framework for estimating sound propagation through the environment. The resulting maps will be very useful in Chapters 5 and 6 for detecting changes to the environment, and avoiding or moving to loud areas of noise.

The third part of the sound source model that the robot can model is the sound function. Using the data collected during investigative sampling of the sound source, a robot can build a representation of the sound function using mel-frequency cepstral coefficients (MFCCs) that allow the robot to classify samples as being generated by a particular sound source. This model of the source sound function will be very useful in Chapter 5, where a robot is trying to identify the presence of new sound sources, or changes in the environment. The model, however, currently lacks temporal information. As will be discussed in Chapter 6, the inclusion of such temporal information into the sound fields framework could be very useful to an acoustically-aware robot, and remains an important area of future work in developing the sound function aspects of the source model.

In summary, it is these tools for filling in information in the sound source model that will form the basis for all of the applied acoustical awareness work in Chapters 5-7. Knowing position and directivity, a robot can estimate how loud different parts of the room will be. Knowing the sound function, a robot can estimate what it will hear at each location. Either set of information allows a robot to make estimates about different parts of the room in which it is not currently located, providing navigational behaviors with the knowledge necessary to improve and acoustically aware application. The remainder of this chapter will now focus on improving those results even further by adding to the pool of knowledge that a robot can gather on its own.

4.3 PATH INFORMATION

Once we have identified characteristics of sound sources in the environment, the next step is to utilize information about the path to enhance the accuracy of the direct field and build reverberant field estimates in place of the simplified, constant-field assumption used in the previous section. As discussed in Section 3.1.2, the attributes of the path entity that we are interested in acquiring for this purpose are the geometric layout of the environment, the material properties of the surfaces, and any structural information for describing transmitted sound. Unfortunately, for the latter two categories there is little that a robot can do to acquire this information if it is not provided a priori.

Limited audio probing work (see Chapter 2) has had some small successes in identifying material properties, as have some computer vision efforts [Ragheb and Hancock 2003], but none of these work in the general case for characterizing an environment. The same is true of knowing the support structure. While DARPA is developing a sensor using RADAR to penetrate concrete and identify floorplans, material compounds, and enemy combatants[Miles 2006], the device is unlikely to be ready for robotic deployment in the near future.

Identifying something of the geometric layout of the environment, however, is well within the capabilities of a mobile robot. Section 4.1.1 described the creation of spatial evidence grids, or obstacle maps, by which a mobile robot can localize itself relative to obstacles in the area. These same spatial evidence grid representations can be used to describe the geometric layout of the surroundings.

4.3.1 Building Reverberant Field Estimates From Spatial Evidence Grids

Figure 4.22 (Top) displays a spatial evidence grid used by the Pioneer2-dxe robot for localization in the Mobile Robot Lab environment. This evidence grid, built by the pmap utility [Howard 2004] using the measurements from a SICK Laser-Measurement-System mounted to the Pioneer robot, consists of a collection of estimates attached to specific locations, or grid cells, in the environment, where each estimate specifies the likelihood of an obstacle occupying this cell. These estimates range in value from 0-

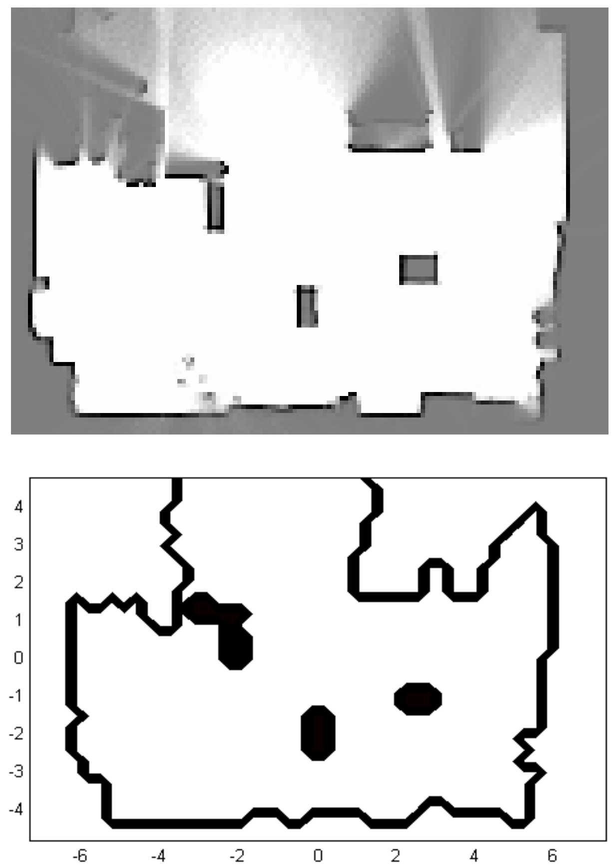


Figure 4.22. Comparison of a spatial evidence grid (Top) collected by the robot for localization purposes to a thresholded evidence grid (Bottom) used for reverberant field estimates

100%, and are not always consistent from neighbor to neighbor. To use this spatial evidence grid as a geometric layout in the mathematical framework, a few assumptions are needed about exactly where the obstacles are in the representation:

- **Which cells are occupied?**

  An estimate of 0-100% cannot be easily used with reverberant field estimation models. A threshold is needed to determine, based on the evidence grid, whether or not any given cell is occupied or not. 75% is a typical threshold for determining occupancy. Figure 4.22 (Bottom) demonstrates the geometrical layout created by applying a threshold.

- **Where is the obstacle?**

  If a grid cell was marked as containing an obstacle, then the obstacle boundary could be located anyplace within the area described by the grid cell. Unfortunately, spatial evidence grid cell sizes often range from 0.01-0.1m^2 in area, providing a large area in which the surface might occur. To simplify the representation for reverberant field estimates, each cell marked as containing an obstacle will be assumed to be completely filled. Therefore, the obstacle boundary is located at the edge of the grid cell.

- **What is the angle of the obstacle surface?**

  Real surfaces in the environment are likely to occur at a variety of angles to the horizontal. By assuming that the obstacle fills the grid cell completely, however, all surfaces will now be restricted to either the vertical or horizontal edges of the grid cell. Given the large size of the average grid cells in spatial evidence grids used for localization, this vertical or horizontal restriction will likely produce minimal error relative to other obstacle positioning errors. However, as smaller area cells become available, this assumption may need to be replaced by some interpolation algorithm across neighboring cell or obstacle surfaces.

After applying these assumptions to the spatial evidence grid and creating a geometric layout for the room, it is relatively simple to apply the ray-tracing algorithm discussed in Chapter 3 for estimating both the direct field and the reverberant field. Works by Savioja [Savioja 1999], and Elorza [Elorza 2005] provide implementation level detail of the ray-tracing algorithm and discussions of accuracy in 3D. As our spatial evidence grids only model obstacles in two-dimensions, our ray-tracing models differ slightly from these described works by limiting all rays to a single height. Furthermore, without knowledge of the specific material composition of the room, we assume that all surfaces are completely reflective (0% energy absorption). Appendix B.6 describes in pseudocode our implementation of ray-tracing.

In Figure 4.23, the ray-tracing algorithm is used to predict direct and reverberant field estimates for a filter generating wind noise at 54-dBA. Figure 4.23 (Top) describes the direct field in which a gradual drop-off from the maximum sound level can be seen as the energy dissipates. "Acoustic shadows" can also be seen in this direct field, where the sound from the direct field is blocked due to the presence of obstacles. Figure 4.23 (Middle) then describes a reverberant field estimate, where sound levels are still highest in the vicinity of the sound source due to the close proximity of the walls below and on the right. Note that the highest volume in the reverberant field is actually 3 dB quieter than the region near the source in the direct field. Figure 4.23 (Bottom) finally combines

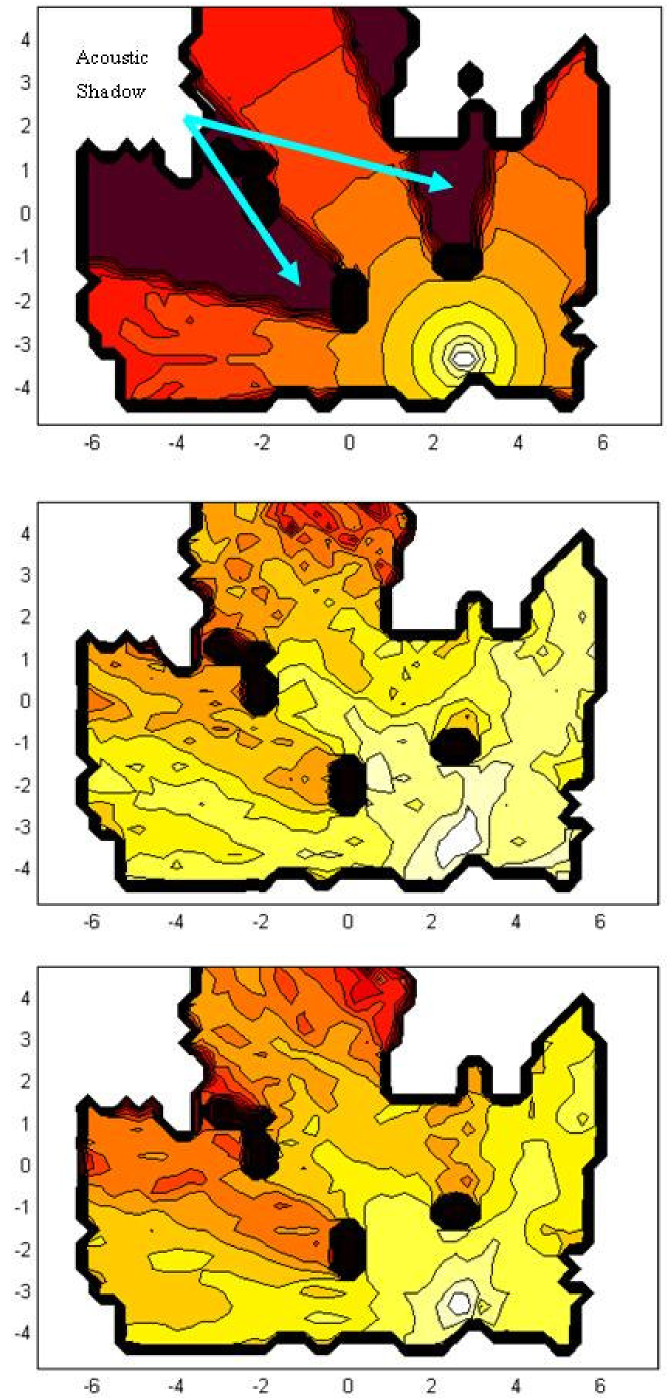


Figure 4.23. Maps of sound propagation created using a 2D robot-created evidence grid of the obstacles in the environment. (Top) Direct field of an ideal point source with known volume and position, note the acoustic shadows due to the 2D assumption. (Middle) Reverberant field created from the same source. (Bottom) Combined field, demonstrating characteristics of both the direct field (strong acoustic shadows) and the reverberant field (strong response near walls).

the two fields, maintaining the key features of each. The gradual drop-off in energy from the direct field dominates the combined field in the region near to the source, but is not as strong as the distance from the source increases and the reverberant field contributes more of the energy.

4.3.2 IMPROVING THE ESTIMATE QUALITY

A great advantage in using spatial evidence grids for finding the geometric layout is that they are created easily using software and hardware found in common mobile robotic applications. As such, spatial evidence grids, unlike more complicated representations of the environment, can be quickly adapted for acoustically-aware applications. Unfortunately, despite the ease with which evidence grids can be acquired and adapted to the problem of sound propagation, there are a number of inherent drawbacks in using them for identifying geometric layout: spatial evidence grids are very noisy, they do not accurately represent surfaces, and they are two-dimensional. What follows here is a discussion of how each of these problems affect the final representation, and how the error might be reduced in future implementations.

<u>Noisy, Incomplete Spatial Evidence Grids</u>

The first problem with the geometric layouts created from spatial evidence grids is that they are very noisy. While a robot travels around the room creating an evidence grid for localization, it does not always gather enough data about the environment to fill in the map completely. This can be seen in the displayed spatial evidence grid (Figure 4.23, Top), where some regions that should be empty are grayed, rather than white. Additionally, the walls have a number of holes in them, and some parts of the room are completely unknown. Our mapping software (*pmap,* with Player/Stage) suffered particularly in this case, as its purpose was localization rather than mapping. Other algorithms for building maps and localizing the robot that have been rigorously initialized can possibly do a better job at creating the resulting map, but will still suffer from the same general problem. If the robot does not enter an area, or something blocks the view of the robots' sensors, then a map cannot be created of the missing area. Without the missing information, however, our sound propagation models will suffer in accuracy.

The solution to this problem requires gathering more data. Putting more sensors onto the robot may have some effect, as the additional data will be gathered from a different viewpoint. More importantly, however, the robot needs to explore the environment thoroughly, actively investigating unknown areas and trying to find better viewpoints for inaccessible regions by incorporating the need for an accurate map into its navigational control strategy [Thrun et al. 2005].

Surface Estimation Errors

The second problem with these particular maps is the accuracy of the representation. As mentioned earlier, if each grid cell covers an area of $0.1m^2$, then the actual surface of the object could be located anywhere within a similar sized area. While our earlier assumptions of completely occupied grid cells and reflections along the cell boundary allow even large-grained spatial evidence grids to be used for constructing an estimate of the reverberant field, knowing exactly where the ray reflects, and at what angle, are important characteristics for accurate sound field estimation.

There are, however, alternative approaches for surface estimation that already exist in the robotic mapping community, although none have yet been tested with sound propagation algorithms. The use of particle filters to estimate the surface directly, rather than the presence of obstacles, should produce a closer match to the real surface. Unfortunately, the resulting surfaces are very noisy, and will produce significant refraction effects in sound propagation. Smoothing the surface may reduce these effects, but may also limit the accuracy gained by using a particle filter approach. Another possibility is to build predictions about the shape of the surface being mapped into the surface estimation algorithm. For instance, if most of the sensor readings will be from a small number of flat surfaces (common with indoor environments), then the algorithm can estimate the number of flat surfaces, and predict which readings belong to which surface[Thrun 2002]. Details about small objects on the surfaces will most likely be lost, but the resulting surface is flatter. If the surface is actually flat, then incorporating such knowledge into the algorithm may produce a better quality map for the purpose of modeling sound propagation.

Two-Dimensional Descriptions of a 3D Environment

The third drawback with these spatial evidence grid representations is that they are 2-dimensional. This is due to the use of just one laser mounted on a flat surface (i.e. the top of the robot), which will produce models at only a single height. Although modeling will still work in 2-dimensions, the ability to accurately predict soundscape features such as echoic locations and acoustic shadows is certain to diminish without 3-dimensional data. This is in part due to missing reflections from above and below, and in part to inaccurate representations of obstacles. Of particular interest are small obstacles, such as small boxes, which should have a minimal impact on sound levels, but are tall enough to be detected by a small robot. These obstacles result can result, when using ray-tracing, in significant acoustic shadows (Figure 4.23, Top) that are not nearly as pronounced in the real environment, due to sound propagating over the top of the obstacle.

Perhaps the most obvious solution to this problem is to use a 3D mapping algorithm on the robotic platform. If the robot is equipped with more sensors for viewing upwards, rather than along a single horizontal plane, the same algorithms described for improving surface estimation accuracy (particle filters, including surface models, etc.) can also be applied to the 3-dimensional modeling problem. There exists extensive work in mapping with upwards pointing lasers [Kaess et al. 2003], cameras, and combinations of both [Biber et al. 2004; Thrun et al. 2005]. At this point, however, none of these models have been tested with sound propagation due to the required hardware and/or software constraints. Furthermore, improved accuracy is not guaranteed by the adaptation of a 3D mapping algorithm. As with 2D evidence grids, they are still subject to the data collection problem and suffer from similar surface estimation errors that may incorrectly predict reverberation effects. In general, the application of robot generated 3D models to sound propagation needs more study.

Given this variety of problems, 3D models of the environment are not a guaranteed way for a robot to improve accuracy. Another possibility for overcoming this problem of small obstacles in 2D is to separate the direct field from the ray-tracing algorithm. The direct field does not actually require ray-tracing for estimation, as the direct field energy can be estimated as decreasing linearly with the square of the distance from the source. The reverberant field can then still be estimated using ray-tracing. Seen in Figure 4.24, the resulting combined field for the filter from this method does not demonstrate as pronounced of an acoustic shadow as the pure ray-tracing solution, which used the geometric layout in calculating both the direct and reverberant fields. An acoustic shadow is still visible, due to reverberant effects, but the difference in volume is much less. Of course, this method is not entirely accurate either, as some obstacles present in the geometric layout may have been tall or massive enough to produce large acoustic shadows, and even the small obstacles likely have some impact. Still, this option allows the designer of a robot controller more freedom for tuning their system to a particular type of environment.

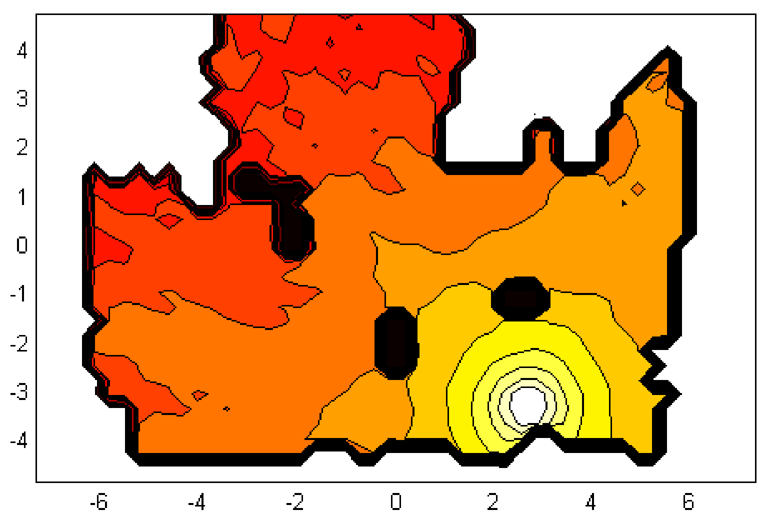


Figure 4.24. Map of the auditory scene combining a simplified direct field model with the reverberant field.

4.3.3 PATH INFORMATION - SUMMARY

At the conclusion of this section on path information, it should be understood that the use of robotic mapping for sound propagation has much work remaining. This work in modeling reverberation from 2-dimensional maps has primarily served to provide insight into the difficulties associated with the mapping the reverberant field, rather than try to build a working system for use with an acoustically-aware robot. It is a significant contribution to the field, because no others have used either environment maps created from robotic data or sound source knowledge generated by an autonomous mobile robot to build estimates of sound propagation through the environment. This dissertation demonstrates the use of both sets of information, and applies them to both the reverberant and direct sound fields. Figure 4.25 demonstrates a combined direct and reverberant field estimate for the two sources localized and investigated at the end of Section 4.2.

Future work in this problem of path estimation will attempt to validate the use of this robot gathered information with sound propagation models. One such validation method is to compare the resulting estimates with measured data at random locations in the environment. Although this is the typical validation method in architectural acoustics, a high level of accuracy does not always transfer to improvements in robotic applications. Therefore, in Chapter 6 we propose work in applying ray-tracing estimates from robotic data to the stealthy approach problem to improve accuracy and generality. It is in the context of this real robot scenario that we intend to demonstrate the usefulness of this information, and resolve some of the remaining issues about accuracy, such as: how accurate does the sound fields estimate need to be to improve performance? And, when is there a need for a better quality robot-generated map?

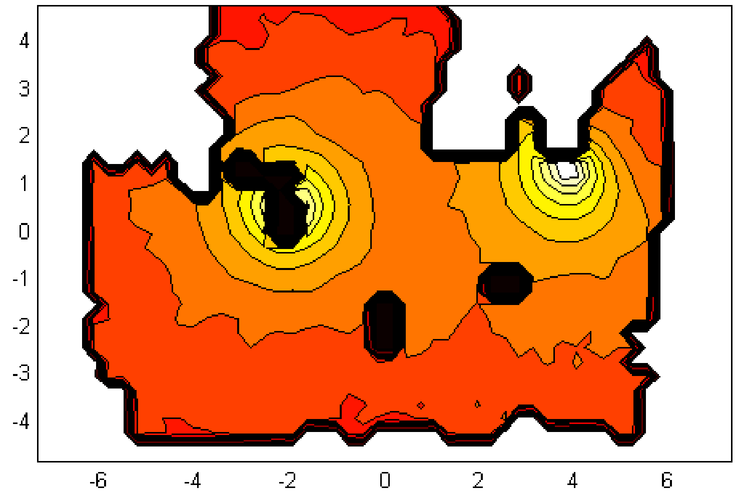


Figure 4.25. Estimated sound levels for the combined direct and reverberant fields, created from purely robot-collected information. The robot collected the obstacle map, localized sources, and identified directivity and volume. For this sound propagation estimate, the geometric layout is only included in the reverberant field estimates.

4.4 Building Maps without Models

The greatest advantage of estimating the auditory scene using models is that all of the information can be described *a priori*. Once a sound source has been discovered, the robot can use it over and over again when building a model of the auditory scene, so long as it detects that it is still operating and has not significantly changed. Identifying such changes may require some work on the robots' part to occasionally check up on the source, but major characteristics of medium duration sources, such as directionality and location, are unlikely to change very often. Such sources are instead more likely to be turned on/off and change volume, things easily determined by a mobile robot. Section 5.1 describes a scenario in which this is accomplished.

The greatest drawback to estimating the auditory scene using models is that the robot may not have, and/or may not be able to acquire, all of the critical information about the environment (or the information is inaccurate). For example, the application may be highly sensitive to the level of reverberant sound. In that case, an adequate model may require 3D models of the environment and material specifications to accurately model the sound propagation. Such information is not always easily determined *a priori* and, although work in both 3D modeling and material characteristics identification [Krotkov 1995] do exist, this info can be very difficult to accurately determine using a robotic platform. As another example, perhaps the robot only has a single microphone for listening to the ambient noise. The robot could try to use a gradient descent strategy to move towards and localize the sound source, but the resulting accuracy is questionable as the robot may become stuck in a local minima. In either of these cases, an alternative to estimating sound flow from models is to measure it directly using the robot.

As originally reported in [Martinson and Arkin 2004], a noise map for robotic use can be created from some arbitrary number of samples through interpolation. Provided with some set of samples collected by one or more microphones, and the relative positions at which each sample was collected, we can use some form of function approximation to estimate the sound field directly from the samples without collecting source localization, directivity, etc. In the original work, K-means interpolation was reported as it was fast and produced a quick approximation over areas that had not necessarily been sampled. Cubic interpolation has since been implemented as well, as it produces smoother contours in densely sampled regions at the expense of increased computational complexity. The creation of a sampled data noise map using any available interpolation method is discussed in more detail in Appendix B.7.

Regardless of the type of function approximation used to estimate the sound field, the advantage of the interpolated noise map is that, while it still requires robot localization, the robot only needs a single microphone to get a rough approximation of the entire sound field. Array information can be incorporated as well, but it is only needed as separate microphone inputs. The drawbacks to this interpolation method, however, are twofold. First, the maps are heavily influenced by robot ego-noise, which may or may not be constant across the sampled region [Martinson and Arkin 2004]. Second, the interpolation method cannot reliably estimate anything beyond the sampled region, limiting its effectiveness for guiding robotic navigation. Given these sizeable drawbacks, we primarily used the interpolation method as an alternative, validating the sound fields estimates in areas that been heavily sampled.

4.4.1 COMPARISON WITH MODELS

Ideally, the sound fields framework described in Chapter 3 would lead to a faithful representation of the auditory scene. In practice, however, it might vary significantly from the ground truth if there is missing information, such as other sources, transmission effects, or 2D vs 3D reverberation models. The interpolation method for constructing noise maps provides a convenient alternative, at least in the sampled region. In Figure 4.26, the source directivity indicates that the fan source in the MRL environment should be loudest to the right, but should still remain fairly loud to the left, at least in some spots. Both representations demonstrate these common features,

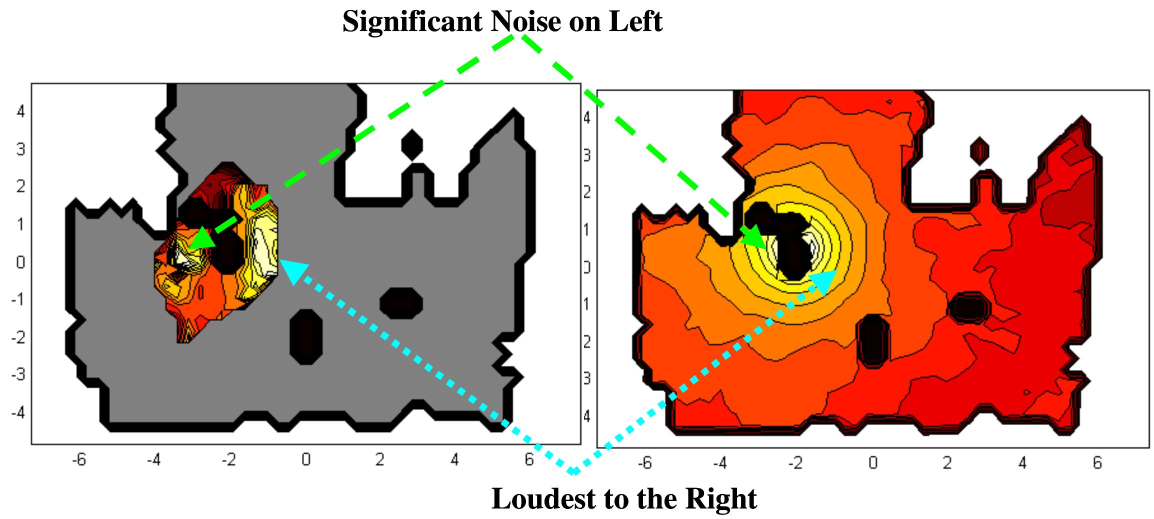


Figure 4.26. Comparison of the interpolated noise map (left) to the sound fields model (right) for the fan source. In both maps, the directivity of the source is indicated as being loudest to the right, but still fairly loud in the rear. Using interpolation, however, generates a noise map that is only valid for the small area that was sampled, while the sound fields map can make predictions for the entire room.

suggesting that the sound fields' estimation is a reasonably good representation of the soundscape. Looking at another example, the radio source (Figure 4.27) from the MRL environment, the different representations indicate common source orientations, or regions of maximum volume. In general, the interpolation method should be reasonably good at predicting local phenomenon such as the direction of maximum volume or other hot spots due to reverberation, so if the two resulting maps do not demonstrate similar characteristics, then it may indicate to a potential problem to a robotic platform.

An example of missing information can be seen in Figure 4.28. The obstacle map of the NRL AI Laboratory is very cluttered, and really needs 3-dimensional data for a faithful reconstruction. In this case, the source, located on a bookshelf next to the wall, is correctly localized and a directivity model is created. But the final model, utilizing the

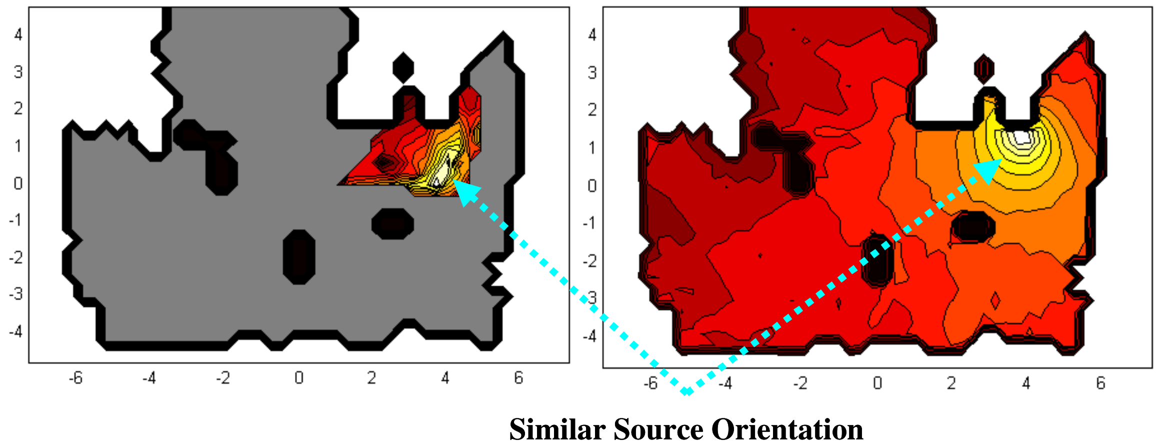


Figure 4.27. Comparison of the interpolated noise map (left) to the sound fields model (right) for the radio source. Both maps indicate the loudest region as directly to the front of the sound source, rather than to a side, or omni-directional.

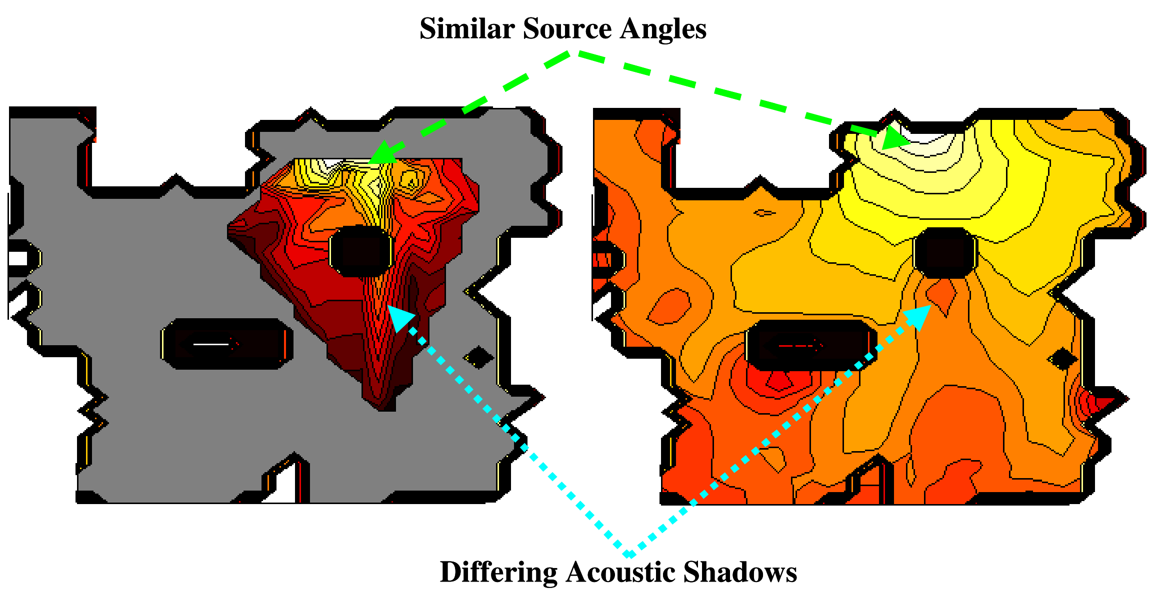


Figure 4.28. Demonstration of the effects of poor reverberation models in the NRL AI Center. The interpolated map shows that the sound can still be quite loud behind the obstacle, while the sound fields map predicts significant decrease in volume.

obstacle map for direct and reverberant fields predicts an acoustic shadow behind the box that does not appear in the interpolated noise map. Although the spike in the interpolated noise map is possibly just an artifact due to the other people in the room, there is still no predicted shadow behind the box. A robot designed to spot such a deviance should probably reinvestigate the area.

Naturally, spotting such discrepancies between these maps is a very hard problem. How significant does the difference need to be before the robot should reinvestigate? Even after identifying a problem, predicting a cause would be even more difficult. For example, the source appearing to point in different directions could indicate a faulty directivity model, a missing source, poor reverberation estimates, etc. Given the difficulty of the comparison problem, this dissertation only used the interpolated noise maps for debugging purposes. If the human designer of the robotic system could see a noise map, then they had more information available for identifying problems, missing information, and faulty sound fields construction. In the long term, however, the combined use of both types of maps could lead to more robust systems. By identifying the discrepancies between maps, a robot can make use of its limited time to better direct its investigatory efforts.

4.5 CHAPTER SUMMARY

Acoustical awareness, as defined in Chapter 1, is the "coupling of action with knowledge about the acoustic environment". In Chapter 3, we then explored this problem in more depth, differentiating between types of awareness, and describing a mathematical framework for making predictions about the soundscape at different locations in the environment. This ability will, over the next 3 chapters, prove to be very useful to a robot performing some acoustic application, as it provides a baseline for constructing plans to maximize the performance of an acoustic application.

By itself, however, the sound fields estimation framework does not make a robot acoustically-aware. In order to make any predictions about the auditory scene, the robot must have some knowledge of the primary acoustic entities: the sound sources, the receivers, and the environment itself. From where, though, does the robot acquire this information? Certainly some of this information could sometimes be made available *a priori*. What happens when the information is not available? Or maybe the information that is available is not enough? Or, what does the robot do when the auditory scene changes? A robot that is completely dependent on a human for its information about the primary acoustic entities is limited in terms of scope and performance. Chapter 4, therefore, focused on resolving this data collection problem that was at the heart of the second sub-question of the thesis, answering how and with what representations can we combine disparate sensory data together for the purpose of enabling acoustical awareness. The emphasis, in particular, was in representations, or tools, that enable the autonomous collection by a mobile robot of the information needed for the mathematical framework presented in Chapter 3:

- **Source Location in 2D**

  Traditional sound source localization algorithms for microphone arrays were extended to incorporate robotic movement for the purpose of localizing one or more simultaneously operating sound sources in two-dimensions.

- **Directivity and Volume Models of Sound Sources**

  Using an area-coverage algorithm, a robot can investigate regions identified as likely to contain a sound source, building models of directivity and overall volume for the source.

- **Sound Function Classification**

  From samples collected during robotic investigation of a sound source, a model of the sound function can be constructed from mel-frequency cepstral coefficients, allowing the robot to classify samples as belonging to a particular source.

- **Geometric Layout**

  Borrowing from existing work in creating spatial evidence grids, the geometric layout of the room can be approximated by a mobile robot for use with ray-tracing algorithms to estimate the direct and/or reverberant fields.

In addition to acquiring information about specific characteristics of entities in the auditory scene, this chapter also identified a method by which a robot can model the combined fields simultaneously:

- **Interpolated Noise Maps**

  The sound fields framework provides a convenient tool for modeling information about known acoustic entities. By creating an interpolated noise map directly from the sample, the robot now has a tool with which it