

Dynamic Spatial Hearing by Human and Robot Listeners

by

Xuan Zhong

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2015 by the
Graduate Supervisory Committee:

William Yost, Chair
Yi Zhou
Michael Dorman
Stephen Helms Tillery

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

This study consisted of several related projects on dynamic spatial hearing by both human and robot listeners. The first experiment investigated the maximum number of sound sources that human listeners could localize at the same time. Speech stimuli were presented simultaneously from different loudspeakers at multiple time intervals. The maximum of perceived sound sources was close to four. The second experiment asked whether the amplitude modulation of multiple static sound sources could lead to the perception of auditory motion. On the horizontal and vertical planes, four independent noise sound sources with 60° spacing were amplitude modulated with consecutively larger phase delay. At lower modulation rates, motion could be perceived by human listeners in both cases. The third experiment asked whether several sources at static positions could serve as “acoustic landmarks” to improve the localization of other sources. Four continuous speech sound sources were placed on the horizontal plane with 90° spacing and served as the landmarks. The task was to localize a noise that was played for only three seconds when the listener was passively rotated in a chair in the middle of the loudspeaker array. The human listeners were better able to localize the sound sources with landmarks than without. The other experiments were with the aid of an acoustic manikin in an attempt to fuse binaural recording and motion data to localize sound sources. A dummy head with recording devices was mounted on top of a rotating chair and motion data was collected. The fourth experiment showed that an Extended Kalman Filter could be used to localize sound sources in a recursive manner. The fifth experiment demonstrated the use of a fitting method for separating multiple sound sources.

DEDICATION

To my Mother, Dr. Xiuyan Han

ACKNOWLEDGMENTS

I have enjoyed the great support of all my dissertation committee members. William Yost is simply an outstanding mentor. He is among the most knowledgeable in the spatial hearing field and also open-minded. He supported me intellectually and to pursue research ideas in and out of the traditional boundaries of psychoacoustics. He also demonstrated the importance of organization and persistence in the process of translating good ideas into publications. Yi Zhou provided valuable insight into the topics of the dissertation, in particular, the Bayesian model. Michael Dorman was always the first to try the experiments and have been my collaborator in several related projects. Stephen Helms Tillery kindly gave me helpful criticism and inspired me to compare human and robot hearing in more meaningful ways. Without the guidance of any of them, the completion of my project would not have been possible.

I would like to officially thank my friend Liang Sun, an expert in robotics, who taught me the principles and practices of recursive filtering, which formed the foundation of the robot hearing part of my dissertation.

Shuai Wang, Anbar Najam, Kate Helms Tillery, Julie Liss, Yishan Jiao, Ming Tu, Sarah Cook, Ileana Ratiu have all generously provided assistance. Most importantly I am indebted to family support from Kangmin Zhong, Wenjing Yang and Emma Zhong.

The behavioral studies in this research were supported by a grant from the Air Force Office of Scientific Research (AFOSR) awarded to William Yost.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
1 BACKGROUND	1
Introduction	1
The Spatial Hearing Cues	3
The Localization of Self Position	8
The Snapshot Theory of Moving Sources.....	11
The Role of Head Turn	12
2 THE MAXIMAL NUMBER OF PERCEIVED SOUND SOURCES	15
Introduction	15
Methods.....	18
Experiment I: Locating Multiple Sources	19
Experiment II: Locating an Added Source	22
Discussion	24
Concluding Remarks	26
3 AUDITORY MOTION PERCEPTION FROM INDEPENDENT NOISES	27
Introduction	27
Methods	29
Results	33
Discussion	35

CHAPTER	Page
Concluding Remarks	35
4 THE USE OF AUDITORY LANDMARKS IN SPATIAL HEARING	37
Introduction	37
Methods	39
Results	41
Discussion	41
5 A ROADMAP FOR DYNAMIC SPATIAL HEARING IN ROBOTS	44
Introduction	44
The Concept of Auditory Objects	46
The Role of Head Turn in Machine Hearing	48
Active Localization of Auditory Objects	48
6 LOCALIZE A SOUND SOURCE IN SELF MOTION WITH ITD CUES	52
Introduction	52
Problem Statement	54
Data Fusion	55
Case Study	64
Concluding Remarks	66
7 DYNAMIC LOCALIZATION OF MULTIPLE SOUND SOURCES	68
Introduction	68
Methods	69
Experiments and Results	71
Discussion	72

CHAPTER	Page
Concluding Remarks	74
8 SUMMARY AND FUTURE DIRECTIONS	75
REFERENCES	79
APPENDIX	124
A INSTITUTIONAL REVIEW BOARD APPROVAL FORM	124

LIST OF TABLES

Table		Page
4.1.	The Results of Sound Source Localization with and without Landmarks	91
7.1.	The Results of Experiment I, Localization of Three Sound Sources	92
7.2.	The Results of Experiment II, Localization of Three Sound Sources	93

LIST OF FIGURES

Figure	Page
1.1. The Framework of Reference for Sound Source Localization	94
1.2. Interaural Time Difference (ITD) and Interaural Level Difference (ILD) ..	95
1.3. The Cone of Confusion	96
2.1. Test Setup for the Localization Tasks	97
2.2. Individual Results of all Listeners in Experiment I	98
2.3. Averaged Results of all Listeners in Experiment I	99
2.4. Individual Results in Experiment I	100
2.5. Mean and Plus/Minus one Standard Deviation Results in Experiment I...	101
2.6. Mean and Plus/Minus one Standard Deviation Results in Experiment II .	102
3.1. The Loudspeaker Setup on the Horizontal Plane for Experiment I	103
3.2. The Loudspeaker Setup on the Mid-Sagittal Plane for Experiment II	104
3.3. Experiment Results of Experiment I on the Horizontal Plane.....	105
3.4. Experiment Results of Experiment II on the Mid-Sagittal Plane	106
4.1. Loudspeaker Setup for the Experiment	107
4.2. The Localization Error with and without Acoustic Landmarks	108
5.1. Using Probability Density Function to Describe Locations	109
5.2. Sound Source Localization with Head Motion on the Horizontal Plane ..	110
5.3. The Active Binaural Hearing of a Robot	111
6.1. The Earth-Centered Coordinate System	112
6.2. The Plane that Embodies the Microphones and the Target Location.....	113
6.3. ITD Estimation	114

Figure	Page
6.4. Outline of an EKF Recursive Filter	115
6.5. The Rest Setup of KEMAR on a Rotating Chair	116
6.6. The Result of Location Calculation based on EKF algorithm	117
7.1. The Change of ITD over Time	118
7.2. Correlagram over Time Depicting the Changing Pattern of ITD	119
7.3. The Loudspeaker Setup in the Modeling and Experiments	120
7.4. Localization results of Experiment I	121
7.5. Localization results of Experiment II	122
8.1. The Structure of the Dissertation	123

CHAPTER 1

BACKGROUND

Spatial hearing is a tool that the human listeners use to explore the world. To accomplish a decent accuracy in spatial perception, the auditory system must analyze the spatial layout of the sound sources, tag the sources to avoid confusion, and ignore the unimportant ones. Because humans are actively exploring the environment, and the environment itself is always changing, the auditory system also must keep track of the moving sound sources and at the same time compensate for both active and passive motions of self to be able to correctly localize the sources in a world-centered coordinate system. According to the daily experience of the normal hearing listeners, the auditory system is very capable of such tasks. Dynamic spatial hearing is concerned with a listener's ability to analyze auditory space when motion of both the target and the listener are allowed. It is the topic of this dissertation.

It is unfortunate that the overwhelming majority of the current literature in the field of sound source localization is concerned with localization of a static sound source by a static listener. Admittedly, this static view in behavioral studies of sound source localization is important for a full understanding of the underlying neural mechanisms in the auditory system. But evidence showing the importance of dynamic listening is also abundant. Small head movement, for example, proved to be able to increase the accuracy of localization experiments to such an extent that some researchers went so far as put stationary bite bars in the mouth of the listeners to keep them still (Blauert, 1997). It seems the reason why head motion affects the results so much is a more important and interesting question than the study themselves in these cases.

Chapters 1~4 focus on human listeners. It starts with a general introduction of background with a focus on auditory spatial perception in the following sections in Chapter 1. The currently known spatial hearing cues were summarized, and the role of head motion in spatial hearing is reviewed. Chapter 2 discusses the maximal number of sound sources that the human listeners could localize at the same time, which is foundation of spatial layout analysis. Chapter 3 presents a method to stimulate the perception of motion with several independent sources. It serves as a challenge to the prevailing view of auditory motion perception. Chapter 4 discusses whether the existence of static sound sources could serve as “landmarks” for the human listeners to better localize other sources.

The idea behind the psychoacoustic studies on humans in Chapters 1~4 is that there may be a motion tracking mechanism that combines motional and auditory sensations. Because one of the best ways to validate a model is to actually implement the model and test it on a machine, Chapters 5~7 focus on robot listeners. The chapters are outlined in Chapter 5, in which the concept of auditory objects are mathematically defined. In Chapter 6, an old paper (Wallach, 1938) that suggested azimuth-altitude localization was possible based on temporal cues alone is revisited; a recursive filter is constructed based on the work of Wallach (1938) to localize a single sound source. In Chapter 7, the same model is extended to multiple source localization. Both the human behavioral and robotics studies are summarized in a final Chapter 8, in which the future directions of research will also be discussed.

THE SPATIAL HEARING CUES

In human listeners, acoustic signals arriving at the two ears are altered in ways suggestive of the locations of sound sources. The parts of signals that can be used to compute the locations are often called spatial hearing cues, which are discussed comprehensively by Blauert (1997). For simplicity, the current section mainly summarizes the interaural differences caused by the physical characteristics of the listeners, which fall into two categories:

1. **Spatial Sampling of the Sound Fields:** When propagating sound waves are measured by sensors at two different locations in space, there could be natural phase and time differences as well as a little level difference. In the simplest form of a model of binaural hearing, the two ears can be regarded as two acoustic receptors with a constant distance in between, which essentially constitute a basic microphone array and represent a typical two point space sampling of the sound fields. This model accounts for the interaural time difference (ITD) to be discussed in the later sections.
2. **Diffraction of Sound by the human body:** The sound waves could be diffracted by objects with dimensions that are similar to their wavelength. Consequently, mid-frequency sounds can be diffracted by the torso; high frequency sounds can be diffracted by the head in addition to the torso; still higher frequency sounds are also diffracted by pinnae. The diffraction distorts the sound waveforms at the two ears of the listeners in a direction-dependent manner. This effect influences cues

for monaural localization, interaural level difference (ILD) and head-related transfer functions (HRTF).

It is commonly believed that multiple processes are involved in the sound source localization in the human auditory system. One reason for the use of multiple processes might be the limitation of the number of sensors, or the ears. The reference framework in which the issues of sound source localization are often discussed is shown in Figure 1.1. The localization is depicted with two spatial angles as well as the distance. Localizing sound sources in three dimensional spaces with only two acoustic sensors is a very challenging task. In artificial systems for computer audition in which a single strategy such as the phase difference is employed, normally at least 4 sensors are needed for the 3D localization of, e.g., a sniper. According to basic geometry of the 3D space, for a two sensor localization system to be of any success, multiple layered processing should be applied on more than one cue that depend on the direction of the source in different ways. For human listeners, the limitation is also due to physiology. For example, phase changes above 1.5 kHz cannot be detected, so the auditory system uses level difference instead. So the multiple process approach is an ecological solution to cues available for auditory spatial processing.

The remaining part of the current section is a list of cues known to be used by the human listeners. This discussion will not consider cues for distance (range) perception and, as such, will only describe cues for the horizontal (azimuth) and vertical dimensions.

Commonly, the psychoacoustic experiment setups used in sound source localization studies consist of a seat and either a moving loudspeaker or an array of non-moving loudspeakers.

Interaural Time Difference (ITD)

Sound waves travel at a certain velocity. The acoustic signal from a certain side of a human listener would first arrive at one ear and later the other. The interaural time difference, or ITD, constitutes one of the most important spatial hearing cues. The value of ITD depends on the spatial angle of the incoming sound wave (Figure 1.2).

The theoretical model for the computation of ITD was summarized by Woodworth (1962). In the model it was assumed that the incoming sound wave was planar wave, i.e. the location of the listener was at an adequately large distance from the sound source to make the curvature of the wave front negligible. When the sound wave comes from an angle that is not 0° or 180° , the additional distance that the sound wave had to travel to reach the other ear consisted of two parts. The first part was the furthest possible linear distance that the wave traveled so as to reach the side of the head. The second part was a curvature over which the sound wave travels along the surface of head to the other ear. A very careful and detailed measurement by Kuhn (1977) showed that for a certain frequency, a certain ITD value corresponded to only a single location in the quarter field.

There is an inherent limitation to using ITD alone for localization of sound sources. At lower frequencies regions where the wavelength is long, the time delay only corresponds to a small phase difference. As the frequency increases, the wavelength is

increasingly larger, causing an increasingly bigger phase difference for the same ITD value. Above 1.6 kHz the phase difference would be over 180° , and it is hard to tell the leading signal from the following signal. As a result ITD is not useful for high frequency horizontal localization of sound sources.

Interaural Level Difference

When sound waves propagate to reach an object, they are diffracted, the extent of diffraction depending on the dimension of the object as compared to the sound wavelengths. At mid- to high-frequencies, the sound level is attenuated due to the diffraction in the spatial area behind the object along the direction of sound propagation. Due to the head shadow effect as shown in Figure 1.2, there would be an interaural level difference (ILD) for the human listeners, except for sound sources in the mid-sagittal plane.

A careful study by Kuhn (1983) demonstrated the ILDs at different spatial angles on the horizontal plane. If the angle changes from 15° to 60° , the ILD increases in the majority of frequencies. But for 90° , the curve intersects with others, meaning that for a given frequency and a given ILD, multiple sound source locations are possible. This is because the head is largely spherical. For angular positions above 60° and especially close to 90° , the acoustic diffractions from different paths coincide in phase at place of head-shadowed ear. Thus, they are adding together and cause a sound level increase. So a major difference between the spatial hearing cues of ITD and ILD is that the latter change with angle in a less linear manner.

Sound waves with bigger wavelengths as compared to the dimension of the head are harder to diffract, which means ILD are small at low frequencies. So another characteristic of ILD is that it is a more effective cue for mid- to high-frequencies.

Duplex Theory

As stated, ITD and ILD are responsible for two different but complementary frequency ranges. The duplex theory (Rayleigh, 1907) stated that humans could use ITD for low frequency localization below 1.5 kHz and ILD for high frequency localization above 1.2 kHz. In the overlapping region, there is an increase in localization error, the reason being that neither localization cue is highly accurate.

Cone of Confusion

The sound sources at multiple places in the 3D space could generate the same ILD and ITD. Mathematically, the set of spatial locations with the same distance difference to the two ears locate on a hyperbolic surface that is most closely approximated by a cone in the sound field, also called “the cone of confusion” (Figure 1.3). If the cone is sectioned on the horizontal plane, there are often two possible directions. For example, a sound source at 45° left-front often provides the same ILD and ITD as a sound source at 45° left-back. The existence of cone of confusion means that additional cues are needed to differentiate front and back sound sources as well as to localize vertically. Whether it is possible to resolve the confusion with motion sensation is an interesting problem to be discussed in Chapter 6. The cone of confusion may be partially resolved by rotating the head. At the same time, it should be noted that in the everyday life of human listeners, the cone of confusion is rarely experience except for certain types of signals.

Head-Related Transfer Function

In reality humans are able to localize in three dimensional spaces. So some additional physical parameters of the sounds must have been changed before reaching the eardrum. The binaural spectral cues to be discussed later also play a very important part in spatial localization of sound sources. The human torso, head and pinnae are able to alter the sound spectrum in ways that adds unique ripples in the spectrum of the original signals. And the changes account for the ability of human to localize in three dimensional spaces (for a summary, see Brungart and Rabinowitz, 1999).

There are also cases when human listeners were asked to rely solely on monaural signals for localization. The only information available is monaural HRTF in this case. The subjects were trained to see if they get better at sound source localization. The general conclusion is that human listeners tend to use binaural cues whenever they are available, but can be trained to use monaural HRTF cues. However, any monaural spectral cues caused by the pinna, torso, and/or head could be confounded by the spectral dimensions of the sound. Thus, monaural spectral cues are only useful for sound source localization if they can be separated from the spectral characteristics of the sound (Wightman and Kistler, 1997).

A very comprehensive and lengthy discussion about sound source localization in general can be found in the book of Blauert (1997). And most of the current studies only concerned with one or two sound sources. Sound source locations are computed based on the auditory inputs from the two ears. Presumably, human's ability to localize sound sources is not infinite. Surprisingly, very few prior works address the problem of the total

number of sound sources that can be localized simultaneously. The topic would be discussed in Chapter 2.

THE LOCALIZATION OF SELF POSITION

When human listeners move around as relative to stationary sound sources, the sources are usually perceived as not having moved, although the relative motion between the sources and the listeners has to some extent altered the interaural time, level and spectral differences. Presumably a mechanism exists to coordinate and align the auditory, proprioception, visual, and vestibular inputs so as to keep the source being perceived at the same spatial position (Yost et al., 2013).

Visual motion mechanisms may give some hints on how this multi-system coordination happens. When an observer moves around in the environment, such as a room, the perception of the changing visual background is combined with the auditory spatial cues. When a significant degree of coherence between the two sources of information is found, the background is perceived as being nonmoving, although its visual position on the retina has changed (for a comprehensive review, see Dichgans and Brandt, 1978). The visual system is divided to a focal mode that solve problems of “what” the objects are, and an ambient vision that analyze the dimension and relative positions of surrounding visual field, so as to answer the question of “where” (Previc and Ercoline, 2004). Whether the same explanation applies in spatial hearing is a very interesting question, and a partial answer would be attempted in Chapter 4.

A basic form of body movement that is often studied is passive whole body rotation around the vertical axis. The human vestibular system has long been known to be

able to adapt to prolonged rotation at constant angular speed due to the mechanics of the semicircular canals (For a brief summary, see Yong and Oman, 1969). After a few tens of seconds subjects adapt to constant velocity rotation and start to have the illusion of being static. Vision has been known to complement vestibular cues; the involvement of vision and vestibular input depends on motion parameters such as visual frequency (Zacharias and Young, 1981). Early literature showed some vague connection between audition and rotation perception threshold (Dodge, 1923). More recently, Yost et al., (2013) studied the simulated moving sound source localization in whole body rotation. The task for the listeners was to report the directional of simulated sound source motion with and without vision. With eye open, the subjects were able to report the results in terms of world centered locations. With eyes closed, the subjects tend to report in a head-centered coordinate system. This observation provides clear evidence that vision is highly involved in sound source localization, i.e. what a listener see could influence a perceived location of a sound source.

When the human subjects are stopped from angular rotation, there is a period of time in which the perception of a loudspeaker in front of the subjects is biased against the direction of prior motion (Clark and Graybiel, 1949). The amount of angular bias depends on the time after cessation of rotation (Mayne, 1952). Another related area of study is audiovection, the effect of induced illusion of self-motion in the presence of rotating sound field (Riecke et al., 2008; Riecke et al., 2009; Våljamäe et al., 2005).

In evolution, the ability to perceive self-location and motion relative to the world is critical to the survival of a species. The experiments in Chapter 4 are also designed to

test whether several static sources could serve as landmarks for human listeners to better able to localize the other sound sources.

Zhong and Yost (2014) compared the postural balance of blindfolded subjects in cases with and without a nearby sound source. They showed that human subjects were better at maintaining balance with the acoustic cues compared to the cases in silence, as indicated by less body sway in Tandem Romberg test and less angular deviation in Fukuda stepping test. Possibly the subjects improved the performance because they are better able to calculate the self-position as relative to a static sound source. The same topic would be expanded in Chapter 4.

THE SNAPSHOT THEORY OF MOVING SOURCES

The research on localization of moving sound sources is less documented as compared to those on static sources. So far the majority of the effort has been made on simple tasks such as detection and discrimination, as measured by minimal audible movement angle, or MAMA (e.g., Perrott and Saberi, 1990; Strybe and Neale, 1994).

A major current theory of perception of sound source motion is the “snapshot theory”, in which the auditory system captures the location of auditory events in a temporal sequence, and combines the “snapshots” for a trajectory of the moving source (Grantham, 1986). According to this theory, no dedicated motion sensor exists as in the case of vision in some species. Rather, the speed and acceleration of motion are calculated based on locations at individual moments in time. If the theory stands, the minimal auditory movement angle (MAMA) would be consistently same as or bigger than minimal auditory angle (MAA) because of the absence of a dedicated motion

detector. This is true in most cases, especially when the auditory event moves at a comparatively high velocity above $60^\circ/\text{s}$ (Grantham, 1997). In physiology, multiple views exist on how auditory motion is coded. Smaller motion detection may be related to whether the population neural response have the resolution to resolve small change of a source location due to motion, regardless of how many discrete channels or motion detectors are activated. Spatial channels still respond to less-preferred source directions, thus contributing to population code.

Some researchers still argue for a separate mechanism for velocity detection of moving sound sources. For instance, Perrott and Marlborough (1989) compared the case of continuous playing moving noise source to the case of pulsed moving sound source, and found a difference in MAMA threshold. They used the results to argue against the snapshot theory. Later, Carlile and Best (2002) reported that human are sensitive to velocity of auditory events, and that the sensitivity could be enhanced with displacement information of static sound source.

Getzmann and Lewald (2007) did a series of experiment to test the localization of sound sources at the beginning, in the middle and at the end of the trajectory of sound source motion. It was found that in cases at the beginning of motion and at the time that the sound source stopped playing, the perceived sound source location was shifted towards the direction of motion, whereas during the middle of motion the localization is more accurate. These findings disagreed with the previous hypothesis of “auditory motion momentum”, in which sound source motions are assumed to be perceived as if they have inertia.

Overall, the prevalent explanation of auditory motion is the snapshot theory, which agrees with most of the MAA and MAMA comparison data up to date. However, a smaller MAMA compared to MAA does not exclude the existence of a separate velocity or change detection sensory mechanism. In Chapter 3, amplitude modulation was applied on a series of static and independent noise sources to see if mechanisms other than the snapshot theory may exist, because the snapshot theory does not predict perceived motion with static independent noises.

Further, a mechanism that fully combines the changes of spatial hearing cues and motion data is suggested in Chapter 5. Because human listener experiments may not be feasible or ethical, the approach is validated with robots in Chapters 6 and 7.

THE ROLE OF HEAD TURN

Vestibular input provides the human subject with two types of spatial information that are very different from each other. In the first category, the three semi-circular canals provide information about three dimensional angular accelerations, and the utricle within the otolithic maculae provides information on the linear acceleration on the horizontal plane.

In the second category, the macula of saccular is responsive to the vertical direction accelerations. Although the information obtained is still acceleration and the mechanisms are about the same as that of utricle, the sensation of gravity directivity is of a totally different meaning for the purpose of orientation. Due to the ubiquitous presence of gravity, the observer is always aware of the direction of “downside”, hence a sense of absolute frame of reference in a single direction (or degree of freedom) is obtained. For

example, if one is rotated on the horizontal plane at very low speed when blindfold, over time one may not notice any motion if it is smooth enough. However, if one is rotated vertically, no matter how slow the motions is, the subject is always able to tell that over time the position is being changed, e.g. upside down.

Voluntary head motion can be categorized into linear motion (front-back, sideways, and up-down) and rotation (yaw, pitch, and roll). Of the 6 sub-categories, the yaw rotation is researched the most, very probably because head turn and whole body horizontal rotation are both common. Human listeners are much better at localizing sound sources in the horizontal plane. Yaw rotation is addressed the most in this aspect. The other types of motion are either less common or very difficult to be measured due to limitation of anechoic spaces.

Wallach (1938, 1940) was among the first to systematically investigate the role of head turn in sound localization. First the benefit of head turn on localization was confirmed. Wallach also reported that auditory localization with head turns was as effective as that in whole body. His suggestion to combine ITD and motion cues for localization on the vertical planes will be revisited in Chapter 6. And the application of the model will be discussed in Chapters 5~7.

CHAPTER 2

THE MAXIMAL NUMBER OF PERCEIVED SOUND SOURCES

For human listeners, the question of the perception of multiple sound sources is not one of a simple addition of the perception of individual sources. According to current knowledge, no auditory receptors exist in a distributed spatial pattern that resembles that of the spatial layout of the objects being perceived, as in the case of vision and touch (Brugge, 1985). As the number of sound sources increases, the number of information inputs remains at two (the two ears), thus the computation is profoundly more difficult in hearing than in vision or touch. Given that the two input auditory signals are often noisy, and assuming that the total computational power of the human nervous system is finite, the maximal number of sources that the humans could explicitly localize at the same time should not be large. Knowing the total number and locations of surrounding sound sources could potentially enable humans to simultaneously locate a group of predators or preys, or to navigate in darkness. However, how well the human auditory system has evolved to cope with complicated acoustic environments to localize multiple sources is an interesting question that remains largely unanswered.

The existing literature in the field of spatial hearing has mainly focused on the localization or identification of a single sound source, in which case human listeners rely on localization cues such as interaural time and level differences as well as head-related transfer functions cues. It has been observed that, when the total number of sound sources increased to two or three, the individual sources could still be individually localized (Blauert, 1997), the accuracy of sound source localization depending on the degree of

coincidence (Gardner, 1969) and the phase of amplitude modulation (Yost and Brown, 2013). For even more sound sources, Blauert (1997) observed that the summing localization would still occur as in the case of two loudspeakers, the precision of the auditory event depending on the degree of coherence. Lower coherence would lead to an increasingly diffused auditory image, which could fill the entire perceived space. If the coherence goes below 0.2, separate sound sources would be more easily localized. It was argued further that, whatever cues are used for two source localization also applies to the case with more than two sound sources (Blauert, 1997), but how many sound sources was not discussed or studied experimentally. Santala and Pulkki (2011) extensively investigated the influence of sound source distribution on spatial sound perception using loudspeaker configurations restricted to the frontal horizontal plane. They found that, for simultaneous independent noise bursts with 15° spacing, up to 3 individual sound sources could be perceived correctly at the same time. Beyond five loudspeakers, the ending sound sources of an array of loudspeakers were often omitted, and the increasingly fused auditory image were likely to obscure the actual location of each individual sound source.

The study of the maximum number of perceived sound sources is also suggestive of new techniques in spatial audio processing, especially audio rendering for virtual reality. A major application of such techniques is in interactive video games, in which locations of virtual sound sources should be generated and updated as the user explores a virtual space. Updating the locations of even a small number of sound sources is a computationally heavy task (for a brief review, see Verron et al., 2010), and novel fast algorithms have been developed to reduce the computational load. For example, Tsingos et al. (2004) developed software that could simulate 174 moving sound sources in real

time. More recently, Moeck et al. (2007) suggested an algorithm that could process 1815 different sources. If listeners are not aware of the locations of all or at least a large portion of, those virtual sources, the audio spatial rendering techniques may not be necessary even if they are computationally advantageous.

In this study we hypothesized that there is a limit to the human listener's ability to localize multiple sound sources. Beyond this point, human listeners would not be able to localize sources at more than chance levels of performance. Two experiments were designed to investigate this question. In the first experiment, multiple speech signals from different talkers were played over loudspeakers at the same time from up to 12 possible locations. The total number of loudspeakers was randomized between 1 and 8. The loudspeaker locations of each talker were also randomized in a trial. The tasks of the listeners were 1) to report the total number of loudspeakers presenting sounds, and 2) to report the individual locations of all the loudspeakers presenting sounds. The independent variable was the total number of actual loudspeakers presenting sounds. The dependent variables were the reported total sources and the accuracy of localization of all sound sources (loudspeakers). In the second experiment, the speech of a number of human voices was played from a certain number of locations (loudspeakers) three times (three intervals). In the second interval and only in the second interval, an additional voice at a new location was added to the existing sound sources. The task was to report the location of the added sound source, and only of that sound source. The independent variable was the existing number of sound sources in the first and the third intervals. The dependent variable was the accuracy of localizing the added source in the second interval.

METHODS

Instrumentation

The tests were done in a reflection-reduced room in Arizona State University (see Yost and Zhong, 2014 for a full description of this room). The dimension of the room was 15'×12'×9'. The surfaces and possible sound reflectors were covered with 4-in. sound absorbing foam with a noise reduction coefficient of 0.9. The broadband reverberation time (RT60) was less than 100 ms averaged over the frequency bands. Sounds were digitally generated from a Matlab program and played from three 12-channel Digital-to-Analog converters (model: Echo Gina 12) running at 44100 cycles/s per channel. The generated signals were amplified with AudioSource AMP 1200 amplifiers before they were fed into the loudspeakers (Boston Acoustics Soundware 100). Twenty-four loudspeakers, with even angular spacing of 15°, were arranged in the horizontal plane in a ring centered on the listener. The vertical level of the loudspeakers was the same as that of the listener's pinnae. All loudspeakers were 5 feet away from the listener with their main acoustic axis pointed at the listener. In the current studies, only 12 out of the 24 loudspeakers on the horizontal plane were used, as shown in Figure 2.1.

Speech Materials

The audio material used in all experiments was speech that was recorded and processed before all the experiment sessions. The voices of 6 female and 6 male American English talkers were recorded. Each sound clip comprised of the recording of a single word. All the words were names of countries with two syllables. A total of 24 sound clips were recorded for each talker. The words were: Belgium, Britain, Burma,

China, Congo, Cuba, Haiti, Japan, Korea, Libya, Mali, Mexico, Nauru, Norway, Oman, Peru, Russia, Sudan, Syria, Togo, Tonga, Turkey, Yemen and Zambia. After the voice clips were recorded, the levels were normalized. Then the recordings were aligned, so that when multiple clips were played together, they started and ended at the same time. The clips were named so that they can be easily played from the Matlab program by specifying the 1) talker index number and 2) the clip index number.

Subjects

Eight normal hearing listeners voluntarily participated in the studies. All listeners passed an audiometric test showing hearing loss of no more than 20 dB HL across the octave frequencies between 250 Hz and 8 kHz. All listeners finished the experiments in a single visit to the Department of Speech and Hearing Science in Arizona State University. For the protection of the listeners, all procedures used in this study have been reviewed and approved by the Arizona State University Institutional Review Board (IRB).

EXPERIMENT I: LOCATING MULTIPLE SOURCES

Tasks

In this part of the study, a number of loudspeakers played sounds (in the form of speech clips) at the same time. The total number of the loudspeakers was between 1 and 8. The total number was evenly distributed for all trials (12.5% chance for each condition). Once the total number was decided, the individual locations of each loudspeaker were also randomly generated. During this process, a list of all 12 possible locations were generated at the beginning as shown in Figure 2.1. Once a specific location was chosen, it was deleted from the list, and the next specific location would be drawn from the

remaining locations. The process continued until the total number of loudspeaker was reached. The minimum inter-speaker angle was 30°.

Speech recordings were played from the loudspeakers at the same time and repeated 10 times. During this process, the same talker index numbers were assigned to the same loudspeakers, but the clip index numbers were randomized. For a total of 2 loudspeakers, an example trial would be loudspeaker 3 (as shown in Figure 2.1) playing “Britain, China, Haiti, Japan ...” with a certain male voice, while loudspeaker 8 was playing “Korea, Libya, Mali, Mexico ...” with a certain female voice. The tasks of the listener were to identify 1) the total number of loudspeakers that were playing sound, and 2) the individual location of all loudspeakers. The sound level from each loudspeaker was about 65 dB SPL.

Procedures

The listeners were instructed of the purpose of the experiment. Then they practiced for 5~10 minutes. In each trial, a Matlab interface first asked them to type in the total number of sources on the keyboard. After that, depending on their answers, the interface would ask them to type in all the locations. Each run of the tests comprised of 8 trials, with total number of sources 1~8 in randomized sequence. Each listener was tested for 20 runs, i.e. a total of 160 trials. Each experiment took 1~1.5 hours. After the listeners finish 25%, 50% and 75% of the experiment, they were offered a short rest.

Results

A listener’s ability to correctly report 1) the total number of sound sources, and 2) the locations of all sound sources was analyzed.

The individual performance of all listeners in reporting the total number of sound sources is shown in Figure 2.2. For each listener, the data were first sorted according to the test conditions in terms of the actual total number of sound sources. Then the average and standard deviation of the reported number of sources was computed in each condition. Overall, a consistent pattern was found among the listeners. When the actual total numbers of sources was small (between 1 and 3), the listeners tended to report the total numbers correctly. When the total number was 4 and more, however, the reports plateaued around 4.

This tendency could be observed more clearly in Figure 2.3, in which the average and standard deviation measures for all the listeners under all conditions were shown. If the perceived number of sources were correct, they should be close to the dotted diagonal line. The significance of the difference between adjacent conditions in terms of paired t-tests is also shown on the top of the figure. For the cases of 1, 2 and 3 sound sources, the reported total numbers averaged 1.1, 2.2 and 3.0, respectively, which were within one standard deviation compared to the correct responses in all cases. For the case of 4 sources, the perceived total number were 3.5, which is smaller than the actual total number, but was still within one standard deviation. In cases of 5~8 sources, the reported total numbers were smaller than the actual numbers, and were never within one standard deviation of the correct responses. Moreover, the reported total number of sources was mostly no more than 4 in all cases. When the actual total number increased to 6 and more, no more significant differences among the responses were observed.

The correctness of localization was calculated as follows: the listener's response of source locations were compared to actual set of sources; whenever a source was

correctly reported, it was counted as a hit; the sum of all the hits was then divided by the actual total number of sound sources, regardless of whether the total number were correctly reported or not. In cases of overshoot (e.g., five loudspeakers were reported but actually only three were presented), the correctness of localization did not exceed 100%, because the denominator was three, and no more than three sources could be correctly reported. The resulting correctness of localization was a measure of the listener's ability to report all the locations of sound sources accurately, regardless of the total number of sources. The individual and overall performance are shown in Figures 2.4 and 2.5, respectively. Both demonstrate a downward trend in correctness as the total number of sources increased. When the number of existing sources was 1, the localization correctness was 0.94 on average, meaning that they could almost always report the location correctly. In cases of 4 sound sources or more, the listeners were only able to correctly report the location of less than half of all sources.

EXPERIMENT II: LOCATING AN ADDED SOURCE

Tasks

In the second part of the study, a number of loudspeakers played sound twice in each of three intervals. In the first and third intervals, the total number and locations of loudspeakers that were playing sounds remained the same. During the second interval, an additional source would also play a sound. The task of the listener was to indicate the location of the additional sound source. During this process, the same talker index numbers were assigned to the same loudspeakers, but the clip index numbers were randomized. As an example, in the first interval, loudspeakers 4 and 10 played "Britain"

and “China” respectively. In the second interval, the same loudspeakers 4 and 10 played “Haiti” and “Korea”, respectively, while a new talker at a new loudspeaker location, number 2, played “Libya”. Then in the final (3rd) interval, loudspeakers 4 and 10 played “Mali” and “Cuba”. The correct answer was to report number 2, which is the location of the added source in the second interval. In each trial the sound levels of the second and the third intervals fluctuated over +/- 2dB compared to level of the previous interval. The sound level was about 65 dBA.

Procedures

The listeners were first instructed of the purpose and procedure of the experiment. Then they practiced for 5~10 minutes to get familiar with the procedures. In each trial, a simple Matlab interface played the sounds and then asked the listeners to type in the number of loudspeaker of the added sound source. Each run of the tests comprised of 8 trials, with total number of sources 1~8 in the base case. The sequence of trials within a run was randomized. Each listener was tested for 20 runs, i.e. a total of 160 trials. Each experiment took 0.5~1 hour.

Results

In this experiment, only the location of a single added source was reported. So the calculation of the RMS (root-mean-square) error in degrees was similar to any of the existing single source localization experiments (see Yost and Zhong, 2014). A minor difference was that, since a circular loudspeaker setting on the horizontal plane was used, the shorter angular path from the reported location to the actual location was used when calculating the errors; e.g., the angle between loudspeaker #1 and #11 was $2 * 30^\circ = 60^\circ$

since they were both “one position away” from loudspeaker #12; the longer angular path (300°) was ignored. The chance level of rms error was 104.6° based on Monte Carlo simulation.

The data in Figure 2.6 showed an overall upward trend of rms error as the number of existing sound sources increased, meaning that as the number of sources increased, it was more and more difficult to accurately locate an additional source. In cases of four or more existing sound sources, the listener’s performance was near chance level. For more than four sources, the performance continued to deteriorate, which agreed with Fig. 2.3 and 2.5, in which the performance appears to reach a plateau at four sources. If the detection of individual sources beyond four and more sources did not deteriorate, more sources would be correctly reported.

DISCUSSION

In the current study, independent narrow-band time-varying speech signals were played from multiple locations. Speech sounds were chosen as the audio material in all experiments. In using different speech waveforms for the sources, the procedure likely maximized the perceptual differences among the sounds from the various sources. That is, performance would in all likelihood be worse if the sounds were more similar (e.g., tones of different frequencies), although it is not clear if more similar sounds would mean that fewer sources (i.e., fewer than 4) could be correctly identified.

Experiment I agreed with our hypothesis that the maximum number of perceived locations by human listeners is limited. It also showed that the ability of human listeners to localize all sound sources would decrease with growing total number of the sources, as

demonstrated by the decreasing localization correctness. In each second of a 10-second trial, different two-syllable words were played from the same set of loudspeakers. That is, the listener has 10 “looks” at the locations of the loudspeakers. It is probably not likely that more repetitions would have improved performance, but fewer, such 1 or 2, may have made performance worse. The commonly used measure of localization, the rms error was not used. As the total number of actual sources increased, the listeners missed some of the sources, and at the same time they were showing larger error in localization of perceived sources. In data analysis, it was difficult to tell which sound sources were missed, and which were not localized accurately.

Experiment II was designed to make the study more complete. By increasing the number of sources by one in the audio stimuli, and by only asking the listener to report the location of the added source, the rms error was computed, which kept increasing as the total actual number of sources grew, and plateaued at chance level when the total number was 4 and more. This result also agrees with the conclusion of Experiment I that the total perceived source grew very little if there are more than 4 sound sources. We believed that this result reflected the localization error of each sound source among a group of sources. This experiment was close to the one of Langendijk et al. (2001) in their conclusions, but the test cases and methods were different, so the data cannot be directly compared.

The results of the current study have several implications on the development of spatial audio rendering techniques. First, simulating a large number of sound sources is unnecessary. Tsingos et al. (2004) created complex auditory scene with 174 sound sources, and claimed that their audio rendering technique did not affect sound source

localization error. However, the behavioral experiment in their pilot test involved exploring a virtual 2-D space that combined interactive video and audio. As a result, their experiment was only suggestive of how well a single sound source location could be virtualized when interactions were allowed, not how good their algorithm is at creating complex auditory scenes or how realistic the scenes could be.

Second, the current results showed that more efficient spatial audio compression is possible. Currently, when the spatial clustering technique is used (Herder, 1999), the single-source localization error of human listener at different angles were used to decide the lobe width of the clusters. The current experiment demonstrated the increase of localization error of individual sources as the total source number grew. So the clusters could potentially be made a lot bigger without influencing the perception of the auditory scene.

CONCLUDING REMARKS

In summary, the maximal number of sound sources that listeners could perceive and localize is limited to around 4 on the horizontal plane for narrow band independent speech signals. For larger numbers of actual sources, the perceived number of sources remains 4. When 4 sound sources already exist, the possibility for the listener to localize an additional source is at chance level. This study presented behavioral data that supported the hypothesis that the number of perceived sound source is limited (and limited to approximately 4 sources).

CHAPTER 3

AUDITORY MOTION PERCEPTION FROM INDEPENDENT NOISES

Given the large body of literature on static sound source localization (for a review, see Blauert, 1997), auditory motion perception has been addressed in only a limited number of studies (for a review, see Gilkey & Anderson, 2014), and competing theories on the underlying psychophysical mechanism exist. On one hand, Grantham (1986) suggested a “snapshot theory” in which human listeners rely on spatial change – not velocity – of sound sources to sense the motions. In this model, the auditory system captures the location of auditory events in a temporal sequence, especially at the beginning and the end of a trajectory, and combines the “snapshots” to form a complete impression of a moving source.

On the other hand, some researchers argued that a mechanism for velocity detection of moving sound sources may exist. Perrott and Marlborough (1989) compared the sensation of continuously moving noise source to that of “pulsed” moving source, and found a difference in minimal audible movement angle (MAMA) threshold, which, they argued, did not agree with the snapshot theory. Carlile and Best (2002) reported that humans were sensitive to the velocity of auditory events. Getzmann and Lewald (2007) tested the localization of sound sources at the beginning, at the end, and in the middle of a trajectory of sound source motion, and reported biased perception for the first two cases. Grantham and Wightman (1978) studied the “sluggishness” of binaural signals, which also suggested a possible auditory motion mechanism that relied on the range of change of interaural cues at lower level of the auditory nervous system.

In both the experimental studies for and against the snapshot theories, the type of sound sources used were either actual dense loudspeaker arrays or stereophonic panning simulation of moving sources (named “stereophonic balancing algorithm” in Grantham, 1986; for a more complete discussion, see Pulkki and Karjalainen, 2011).

Psychophysically, the two types of sound sources were identical. However, if alternative methods could induce auditory motion perception in human listeners, it could also constitute a case in which the snapshot theory isn’t would not be the only explanation of auditory motion.

Recently, Yost and Brown (2013) reported that amplitude modulation (AM) with a rate less than 25 Hz of two independent sound sources could induce perceived auditory motion when the envelopes of signal were out-of-phase. At any given moment, there were only two incoherent noisy sources each at a different fixed location, leading to the perception of two separate auditory events instead of a single phantom source (Blauert, 1997). So the snapshot theory does not predict the perception of motion. It could be argued that when alternating (out-of-phase) AM was applied, the overall acoustic energy shifted from side to side. But the snapshot theory required a beginning and an end of a trajectory. As long as the two sources could still be individually localized, there were no such beginnings and ends; in fact, there were no trajectories either. Instead, what existed was only a consistent changing pattern of amplitude of independent sources. Consequently, additional investigation is needed to explore the role of AM in auditory motion.

This study asked whether a consistent AM pattern could be applied on an array of independent noise sources to generate perceived auditory motion. In particular, the phase

delays of the envelope of the signals were consecutively larger for successive loudspeakers in the array. The carrier noise signals were amplitude-modulated at different rates to see which range of AM rate lead to perception of auditory motion. The task for the listeners was forced judgment of auditory motion direction. Another independent variable was the spatial configuration (horizontal/vertical) of the loudspeaker array. In the first experiment, six loudspeakers on the horizontal played independent white noise at the same time. The amplitude was modulated, with an increasingly larger envelope phase delay for each sound source. A second experiment asked whether the same principle also applied to sound sources on the vertical plane.

The study had practical value in audio engineering, particularly in spatial sound reproduction. Creating steady sound images in listening spaces with a limited number of loudspeakers has been a challenging task. For example, with less than 6 loudspeakers, it is impossible to create stable sound images at arbitrary locations on the horizontal plane (Xie, 2001). The types of loudspeaker configurations for auditory motion reproduction has been little discussed. And studies on auditory motion in the audio engineering field are even more limited. In the current study, the number of loudspeakers and configuration was a practical setting that could easily be applied in a common listening room. So the findings were also suggestive of new approaches of auditory motion.

METHODS

Instrumentation

The tests were done in a sound treated room in Arizona State University (see Yost and Zhong, 2014 for a full description of the test room). The dimension of the room was

15'×12'×9'. The inner surfaces of the room were covered with 4-in. sound absorbing foam to reduce sound reflections. Audio signals were digitally generated from a Matlab program and three 12-channel Digital-to-Analog converters (model: Echo Gina 12) running at 44100 cycles/s per channel. Then the generated signals were amplified with AudioSource AMP 1200 amplifiers before they were played from an array of 36 loudspeakers (Boston Acoustics Soundware 100). Twenty-four loudspeakers were evenly arranged in the horizontal plane in a ring centered on the listener. The vertical level of the loudspeakers was the same as that of the listener's pinnae. A second layer of 8 loudspeakers (45° spacing) was on the sphere at ~30° elevation, and another layer of 4 loudspeakers (90° spacing) at ~60° elevation. All loudspeakers were 5 feet away from the listener. In the current studies, only 6 loudspeakers (60° spacing) on the horizontal plane were used in experiment I as shown in Figure 3.1; only 4 loudspeakers (~60° spacing) on the vertical plane (mid-sagittal plane) were used in experiment II as shown in Figure 3.2. The audio materials used in all experiments were white noise sound clips that were independently generated for each channel. The sound level at the listening position was about 65 dB SPL. The duration of the signal was three seconds.

Subjects

Eight normal hearing listeners voluntarily participated in the studies. All listeners passed an audiometric test showing hearing loss of no more than 20 dB HL across the octave frequencies between 250 Hz and 8 kHz. All listeners finished the experiments in a single visit to the Psychoacoustics Lab in Arizona State University. All procedures used in this study have been reviewed and approved by the Arizona State University Institutional Review Board (IRB).

Motion Simulation

In the first experiment, four loudspeakers on the horizontal plane played white noise at the same time. The noise was generated independently for each channel. The amplitude was modulated with sinusoidal envelope. The envelope phase delay was 60° between adjacent loudspeakers (1~6 in Figure 2.1). The AM rate was the same for all the loudspeakers. As an independent variable, the AM rate changed from 1, 5, 50, 200, to 500-Hz. The sound lasted for 1.6 seconds. The task for the subjects was to tell the direction of auditory motion, whether it was clockwise or counterclockwise, by pushing buttons 1 or 2 on a keyboard.

In the second experiment, the sound stimuli were similar to those used in the first experiment, only that the direction of loudspeaker array was on the vertical direction. Due to the limited size of the lab, only the upper half of the loudspeakers (4 loudspeakers with $\sim 60^\circ$ spacing) on the mid-sagittal plane was used. The subjects were forced to choose the direction of motion to be either forward or backward by pushing buttons on a keyboard.

Procedures

At the beginning of the experiment, the purpose and task of the experiment was explained to the subjects. The subjects went through a training period, in which the correct answer would be given after their inputs. After that the subjects were tested in experiment I with randomized AM rate conditions (1 Hz, 5 Hz, 50 Hz, 200 Hz and 500 Hz). In total, each condition was tested 20 times. Then the subjects were tested in

experiment II (mid-sagittal plane), again with the same 5 AM rate conditions in a randomized order. The experiments took about 30 minutes.

RESULTS

The mean results of all subjects were shown in Figure 3.3 and Figure 3.4. The proportion of correct responses in judging the direction of simulated auditory motion was calculated simply by dividing the number of correct responses with the total number of responses. The chance level was at 0.5. The five AM rate conditions are shown on the horizontal axis, whereas the proportion of correct responses is shown on the vertical axis.

Figure 3.3 demonstrated the mean results of 8 subjects on the horizontal plane condition. For slower AM rates, i.e. 1 Hz and 5 Hz, the mean values were at least one standard deviation away from the chance level. In particular, when the AM rate was 1 Hz, the mean performance as measured by proportion of correct responses was 0.99 (standard deviation 0.02), meaning that the subjects were able to correctly judge the direction of simulated auditory motion in almost all trials. Another observation was that the inclination of the Subjects to make more errors when the AM rate was increased. The proportion of correct responses dropped consistently with increasingly higher AM rate, until it was close to chance level with 500- Hz AM. With higher AM rates, the standard deviation also tended to be larger in general.

Figure 3.4 demonstrated the mean results of 8 subjects on the mid-sagittal plane. The results were similar to those in Figure 3.3 in that with 1-Hz AM rate, the performance of subjects in judging directions of simulation motion were much better than chance, although the standard deviation was larger than for the horizontal condition.

However, as the AM rate increased to 5 Hz, the performance quickly dropped to within one standard deviation of the chance level. At even higher AM rates, performance remained near chance.

DISCUSSION

The current study expanded the findings of Yost and Brown (2013) with a focus on auditory motion perception. In the prior study, when the two independent noise sources were modulated out of phase with AM rate less than 25 Hz, the subjects reported perceived motion in between the two sources. In the current study, we asked the subjects to report the direction of perceived motion instead of the subjective impression of whether there was motion. For an auditory impression of sound source motion to be realistic, the direction of motion should be correctly reported. We found that on both the horizontal and the mid-sagittal planes, the direction of simulated motion could be reported correctly when the AM rate was slower than 5 Hz. In both experiments, the four loudspeakers used for auditory motion simulation generated independent noise at static locations. The AM depth was 1, meaning that there were silent moments in each AM cycle, but the length of silence was negligible. In this case, the snapshot theory would predict no perceived motion but a few location-fixed sound sources with varying amplitudes. The results of the current study were against this prediction of the snapshot theory.

A simple explanation for the existence of a mechanism for motion detection based on AM of independent sources may be important for survival under certain conditions. Take a predator that stirs in the grass and bushes around the listener's habitat for example.

The predator and its movement could be so quiet as not to be perceived. But as the predator moves and crushes through the grass and bushes, amplitude-changing independent noise sources are present. For the sake of survival, it is better for the listener to perceive the consistently changing AM pattern from the grass and bushes as a sort of movement, rather than several independent static auditory events in different snapshots. In fact, any attempt to apply a snapshot model for auditory motion would have to immediately face with the problem of sound source tagging. How could sound sources in difference snapshots or frames be regarded as the same source, given that the sound is even-changing? In Chapters 5~7, a model based on the concept of “auditory objects” instead of “auditory events” would be introduced and applied on data collected from a rotating dummy head.

In the human auditory system, it is known that neurons at different levels are not arranged in accordance with the spatial location of objects, and the spatial hearing cues such as ILD and ITD must be computed. Computationally, ILD or ITD change detection is a relatively straightforward way to provide a motion detector. As pointed out by Middlebrooks and Green (1991), a mechanism using ILD change to detect sound source velocity might be used in auditory motion perception. On the other hand, Grantham and Wightman (1977) discussed the possible existence of a mechanism for movement detection based on ITD changes. The current study implied the possibility of varying envelope spatial hearing cues being used for detection of motion. As discussed later, varying interaural cues may not be the only explanation for the results in the current study. But they may be the reason that motion direction judgment is more accurate in the horizontal plane as compared to that on the mid-sagittal plane. In the latter case, since

inter-channel cues were not available, the most possible explanation for the perception of motion would be due to notch detectors of high frequency HRTF, which is caused by head and shoulder diffraction of sound, and in many cases, change consistently with altitude in the mid-sagittal plane (Brungart and Rabinowitz, 1999).

For sound sources on the mid-sagittal plane, the varying ITD and ILD cues on the envelope were the same during the whole process of simulated motion, so it was unlikely that they were the only cues being used for motion perception, given that motion perception could be accurate on the mid-sagittal plane in experiment II. The changes of certain spectral components could be one explanation, although the commonly used head-related-transfer-function may not apply directly in the current study. Again, all four sound sources were making noise at the same time. So the available spectral cues for motion detection were at best consistent changing patterns of HRTF in experiment II. Additional data are needed for AM rates between 1 and 5 Hz for the vertical plane experiments and 5 and 50 Hz for the azimuth plane in order to determine more precisely the AM rates at which performance approaches chance. However, the current data suggest an order of magnitude difference in the AM rate that can induce auditory motion between the azimuth and mid-sagittal planes.

CONCLUDING REMARKS

In summary, the changes in the envelopes of a number of location-fixed independent noises could be perceived as auditory motion. On the horizontal plane, motion could be perceived at AM rates of 1~5 Hz. On the mid-sagittal plane, motion could be perceived at 1 Hz. The finding was against the simple form of snapshot theory

on auditory motion, in that it only predicts perception of static sources in the current setting. The results also suggest that varying interaural cues such as ITD and ILD may not be the only mechanism that is responsible for auditory motion perception.

CHAPTER 4

THE USE OF AUDITORY LANDMARKS IN SPATIAL HEARING

How human listeners localize sound sources in self-motion is an interesting research question that was little discussed, although several previous studies addressed related topics. An important but even less documented question is about whether distributed sound sources at fixed locations could serve as landmarks for self-localization. This study asks whether the listeners are better able to localize a target with the listener - rotates when there are multiple static sound “landmarks” are present as compared to the case with no sound landmarks other than the target signal.

Because of the relative nature of the localization in space, the problem can be divided into two parts. The first is about the localization of self in the earth-centered coordinate system, which is primarily decided by the vestibular system (see Lackner, 1983 for a review). Easton et al. (1998) found a connection between the static sound field and spatial orientation in both normal vision and blind human subjects. Zhong and Yost (2013) established a more direct relationship between the listener’s performance of postural stability and the existence of a single far field sound source. The second part is about the localization of the target in local self-centered coordinate system, which is mostly decided by acoustic spatial hearing cues. Spatial perception of sound sources is inaccurate when either of the two elements is inaccurately sensed.

A simple form of motion is the whole body rotation on the horizontal plane. The human vestibular system has long been known to be able to adapt to prolonged rotation at constant angular speed due to the mechanics of the semicircular canals (For a brief summary, see Yong and Oman, 1969). After a few tens of seconds the subjects would

have the illusion of being static. Vision has been known to complement vestibular cues; the involvement of vision and vestibular input depends on motion parameters such as visual frequency (Zacharias and Young, 1981). Early literature showed some vague connection between audition and rotation perception threshold (Dodge, 1923), but limited prior research directly addressed the current research question.

A related but different topic of research is on the so-called audiogyral illusion. When the human subjects are stopped from angular rotation, there is a period of time in which the perception of a loudspeaker in front of the subjects is biased against the direction of prior motion (Clark and Graybiel, 1949). The amount of angular bias depends on the time after cessation of rotation (Mayne, 1952). A similar research topic is audiovection, i.e., the effect of induced illusion of self-motion in the presence of rotating sound field (Riecke et al., 2008; Riecke et al., 2009; Väiljamäe et al., 2005).

The current study investigates the possibility of using static sound sources as an absolute framework of reference and the “earth-centric” spatial hearing instead of “head-centric” perception. In evolution, being able to tell self-location and motion from auditory perception, e.g. in darkness, might be important for the survival of animals. In engineering applications, good understanding of sound source perception can potentially help build visual-auditory combined display in aviation and flight simulations (Bronkhorst et al., 1996) and virtual reality in general (Riecke et al., 2009).

In the current study, the listeners were blindfolded. The target of the sound source localization was a short noise stimulus that was played for one time in the middle of the 90 second passive whole body rotation of the listener. Besides, in one test condition there

were four loudspeakers playing speech signals as acoustic landmarks all through the process, while in the other condition there were no such landmarks. The hypothesis was that in the former case the listeners should be better able to localize the target due to the existence of landmarks.

METHODS

Subjects

Eight subjects voluntarily participated in the experiment. All subjects have normal hearing as indicated by audiometry tests. All subjects reported no problems with vestibular system or postural stability. The procedure was approved by the Arizona State University Institutional Review Board (IRB) for the Protection of Human Subjects.

Test Environment

Testing was done in a 15'×11'×9' sound-treated room. The surfaces and possible sound reflectors were covered with 4-in. acoustic foam with a noise reduction coefficient of 0.9. The overall broadband reverberation time (RT60) was less than 100 ms. 24 Boston Acoustics 110 loudspeakers, with even angular spacing of 15°, were arranged in the horizontal plane on a ring centered on the seat of the listener. Every other one of 24 loudspeakers were chosen to play sounds. The radius of the ring was 1.67 m. The loudspeakers were at the height of the listener's pinna. Sounds were digitally generated and played from the combination of three 12-channel Digital-to-Analog converters (Echo Gina 12) running at 44100 cycles/s per channel.

The subject was seated in a RCS ED900 Rotational Chair System, which was originally designed for standard kinetic vestibular tests. During each test, the subject did

a passive whole body rotation. The rotation was accelerated at a constant acceleration of $1^\circ/\text{s}$ for 68 seconds. That was followed by a deceleration period of $-1^\circ/\text{s}$, also for 68 seconds. The subject put on a belt attached to the chair for safety reasons. The subjects were told not to open their eyes or move their head during the chair rotation. The chair rotations were controlled in a separate neighboring room. The subjects were monitored by a commercial web cam and sound system. At any given time during the tests the subject could call for a stop either by gesturing or verbal communication with the experimenter.

Stimuli

The target sound stimulus was a white noise signal that lasted for three seconds and started at the 65th second of the whole body rotation. The bandwidth was 200 Hz ~ 20 kHz. The signal was about 65 dB SPL at the listening position, with a rise-fall time of 150 ms. In one condition, only the target sound was played. In the other condition, beside the target sound, loudspeaker number s1, 4, 7, and 10 continuously played speech signals from the beginning to the end of the rotation as shown in Figure 4.1. The talker's voices were different for the four loudspeakers, and the content was also different. But the sound clips started and ended at about the same time for each word spoken. The content of the speech was two-syllable names of countries, such as "Japan, Mali, Cuba ...". And the position of the 3-second noise target could be from any of the rest of the loudspeakers.

Procedure

The subjects were told the purpose of the experiment and ran a sample test in each condition. After that, the subject ran each condition for six times in randomized order.

The subjects kept their body as still as position as relative to the rotating chair during the tests. The first reason is to avoid changing the ITD, ILD and spectral cues in an uncontrolled manner. Secondly this is to avoid the Coriolis force cue being used for detection of angular acceleration and self-recalibration (Lackner and DiZio, 2000). The subjects were given a break after every four tests. Additional breaks were offered when the subjects reported to be dizzy.

RESULTS

The results of the experiments were shown in Figure 4.2. The individual data was shown in Table 4.1. The localization error was measured in root mean square (RMS) in degrees. Seven out of the eight perform better in sound source localization with acoustic landmarks than without, with their localization error at least halved. With no acoustic landmarks, the averaged localization error was 98.6° (standard deviation 29.4°), with the acoustic landmarks, the averaged error reduced to 38.9° (standard deviation 22.6°). In the two-tailed t-test, the p-value was less than 0.001. The large range of data and standard deviation may be due to the limited number of repetitions for each subject in each condition. But at least for the current study, six repetitions in each condition is the upper limit that most of the subjects could tolerate.

DISCUSSION

Zhong and Yost (2013) related the listener's performance in postural stability to the existence of a single far field sound source. One of the possible explanations for the subjects being able to perform better postural control is that they are better in locating themselves in an auditory background, similar to the case of a stationary visual field. The

current study extended the topic of spatial awareness due to the ambient sound field to the problem of multiple sound sources. When, continuous speech sources were used as acoustic landmarks, the subjects showed an improvement in sound source localization performance as compared to the case without landmarks.

The localization of multiple sound sources or self-localization relative to multiple sound sources is much more complicated than localizing a single sound source. Unlike visual localization, in which light sources at different positions corresponds to receptors at different physicals locations on the retina, the localization of multiple different sound sources depend on the binaural auditory inputs, which is, a computation-heavy task. Langendijk et al. (2001) demonstrated that the performance of sound localization of a single loudspeaker would monotonically degrade as the number of distractor sounds increase from 0 to 2. Although suggestions have been made on the research direction of dynamic sound source localization, to date there is not a lot studies on this, likely due to the lack of necessary equipment (Wightman and Jenison, 1994).

In the current study, the listeners were seated in a chair that rotated. This case has been often referred to as passive whole body rotation. In some other studies, the motion may be initiated by the listener instead of a chair. This case has been often referred to as self-rotation. In the latter case, cues from proprioceptor of muscle are also available, making the localization error much smaller in some cases. A very meaningful direction of future research is to do the same test for actively moving listeners.

When the human listeners move as relative to stationary sound sources, the relative motion between the sources and the listeners alters the interaural time, level and

spectral differences. However, the sources are usually perceived as not moving, although the auditory spatial cues are changed. Presumably a mechanism exists to coordinate and align the auditory, proprioception and vestibular inputs so as to keep the perceived source at the same spatial position. In vision studies, when an observer moves around in the environment, the perception of the changing visual background is combined with the perception of self-motion. When a significant degree of coherence between the two sources of information is found, the visual field is perceived as not moving, although its visual position on the retina has changed (for a comprehensive review, see Dichgans and Brandt, 1978). The human visual system could be divided to a focal mode that solve problems of “what” the objects are, and an ambient vision that analyze the space and relative positions of surrounding visual field, so as to answer the question of “where” they are (Trevarthen, 1968). Whether the same explanation applies in spatial hearing is an interesting research topic that could be addressed in future experiments. That is, in addition to knowing how multiple sound sources may provide visual landmarks when the listener moves, research is needed on the perception of the motion of multiple sound sources when the listener is stationary. If multiple sound sources are not perceived as moving when they do in fact change location over time, this adds a complication in interpreting the lack of a perceived change in a sound source location with multiple sound sources.

CHAPTER 5

A ROADMAP FOR DYNAMIC SPATIAL HEARING IN ROBOTS

The previous chapters were about behavioral studies in human listeners, in particular, about their performance in dynamic spatial hearing. Modeling the mechanism of dynamic spatial hearing is another related and interesting topic. Very few prior researchers attempted modeling the functions of binaural neural pathways involving vestibular inputs. However, among the very limited number of documented attempts, an important one (Wallach, 1938). Given the progress in spatial hearing study in the decades that follows, the work of Wallach (1938) should be revisited.

Another way to look at the same problem of dynamic spatial hearing is to treat the system as a machine with two types of sensors. The auditory system provides the acoustic sensors. The vestibular system provides the motion sensors. Both types of sensors have errors, but the machine has to determine the location of sources upon the noisy data. Data fusion of multiple sensors constitutes a large body of literature in modern robotics studies.

Perhaps the best way to test an auditory model is to actually construct the model and test it on robots. In the remaining part of this piece of writing, a general model of “auditory objects” would be outlined (Chapter 5). The Wallach model would be revisited and included in an Extended Kalman Filter to compute both the azimuth and elevation angles of a sound source (Chapter 6). Further, an approach to localize multiple sound sources would be presented based on ITD and motion data only (Chapter 7). Echo-locating bats have only two ears. Barn owls, which rely on acoustic cues alone to hunt for their prey, have only two ears. The model that would be presented in the rest of the

dissertation is based on the idea that the possibilities of binaural hearing is far from being exhausted.

Currently, two competing approaches exist for robotic spatial hearing. One is based on multiple-microphone-array ($N \geq 2$), which has the advantage of better accuracy compared to other solutions at the cost of complexity of audio front end (Benesty & Huang, 2008). The other is based on a two-microphone-array and often called a binaural approach (although the term binaural means headphone listening in human spatial hearing research), often on a humanoid robot, for which a simple commercial audio front end can be used, but a complicated algorithm is often required to achieve a decent level of accuracy. The latter approach is often related to human spatial hearing models (Jeffress 1948; Lyon, 1983).

THE CONCEPT OF AUDITORY OBJECTS

Imagine a roommate of yours came back from school. He opened the door with the keys, walked in and closed the door. He changed his shoes, unzipped his jacket and took it off. He then walked to his room, put the key ring on the table, opened the laptop and started typing to respond to an email. During the whole process, the human body of the roommate himself may not be making any sound. However, the sounds of the key turning, the door closing, the zipper unzipped, the footsteps and the keystrokes were due to the physical impact of the same object. As a result, grouping acoustic signals based on locations or trajectories is one possible way to analyze complicated auditory scenes.

The perception of a sound source is often described as an “auditory event” (Gaver, 1993; Blauert, 1997). While this view does explain a large number of subjective auditory

impressions, it has several problems. When applying the concept to dynamic hearing, such as tracking a moving sound source or localizing in self-motion, one is immediately faced with the problem of combining different events in the different temporal and spatial frames. Because the content of sound often changes over time, this may not be easy. Real sounds such as speech are also intermittent in many cases. When, for example, a talker makes short pauses in between the sentences, it is reasonable to say he or she still exists as a sound source, only that he or she is not making a sound at that particular moment. Localizing a sound source in each individual moment also requires a lot of unnecessary computation. These inherent problems suggest that auditory events can be an oversimplified view in dynamic spatial hearing.

In the following chapters, the concept of auditory object is introduced as a basis for the perception of sound source locations. It is defined as a lasting object that could impact its environment and cause vibrations and sounds. An auditory object is described by the first moment of a probability density function, i.e. the mean location and the covariance. The definition is continuous in space but discrete in time (Figure 5.1). Other definitions of the same term of “auditory object” has also been suggested before (Griffiths & Warren, 2004), but is less relevant in the current study. Mathematically, how an auditory object is established, maintained and eliminated will be discussed. The result are algorithms that could be easily ported on robots, as shown in Chapters 6 and 7.

THE ROLE OF HEAD TURN IN MACHINE HEARING

In the current studies, of particular interest is the case of sound source localization with head motion. Figure 5.2 shows a simple robot with a body and a head that are

connected with a neck (not shown). The head could initiate a rotation (a type of control in robotics) when necessary. In the modeling, the acoustic cues are collected from the two ears in human or two microphones in humanoid robots. The motion data are obtained from the vestibules in human or gyroscopes in humanoid robots.

When a quick head motion is initiated by a human or a robot listener, the angular speed can be so fast that the body motion and the target motion can be ignored in many cases. In human spatial hearing literature, the benefit from head motion was discussed by early researchers (Wallach, 1939), but largely head motion was believed to be at best a weak cue in sound source localization (Middlebrooks & Green, 1991). In robotics literature, Kneip and Baumann (2008) built a two microphone array based on LEGO blocks to localize sound sources. More recently, Portello et al. (2011) discussed the use of unscented Kalman filter (UKF) in active binaural hearing. The current study will show that a surprisingly simple model based on data fusion can be used to separate and reliably localize multiple sound sources.

In the current literature concerning human listeners, the terms “self-motion” and “whole body rotation” have different meanings. Self-motion is the motion initiated by the listener him or herself, and whole body rotation is totally passive. In the former case, for motion detection, beside the cues from vestibular inputs, the proprioceptor cues are also available, making the self-localization much easier in many cases. Since when the motion is initiated by the listener, the direction of motion could be much more accurately estimated and hence the spatial hearing cues. In robotics literature, the two terms are somewhat interchangeable, because the motion is mostly detected by accelerometers in both the active and passive cases.

ACTIVE LOCALIZATION OF AUDITORY OBJECTS

Active binaural localization of sound sources is a relatively new topic in the emerging field of machine hearing, with a limited number of studies spreading in the disciplines of psychoacoustics, robotics, signal processing and computer science. In this section, a unified model of active robotic hearing as shown in Figure 5.3 is introduced in the application of multiple non-moving sound source localization.

1. Self-localization: to accurately localize the sound source in the world-centric coordinate system, the robot has to localize itself first in the world-centric coordinate system. The self-localization of robot can employ multiple types of sensors: global positioning system (GPS), computer vision, and etc. The simultaneous localization and mapping (SLAM, see the summary of Bailey & Durrant-Whyte, 2006) based on acoustic cues alone has not been discussed before, although Chapter 4 of the current study implied such a possibility in human audition. Overall this interesting topic is beyond the scope of the current discussion. It is assumed that the robot knows its accurate global location at the beginning of the process.
2. Sound detection: The robot is dormant when no or little sounds are present. When the sound level is above a threshold, the system is activated.
3. Voluntary head turn: When the robot is activated, it does a quick head motion. The direction of motion is arbitrary. For simplicity of discussion, it is assumed that the head motion is on the horizontal plane (along the z axis).
4. Update the number of objects: The cross-correlation at each time frame is calculated and stored. The change of cross-correlation is plotted vs. time as a part

- of a correlogram. Since the motion data is also known, the changing pattern of the ITD can be reduced to a limited number of possibilities, each corresponding to a different sound source. The calculation is based on algorithms described in Chapter 7.
5. Raw localization: At the same time of 3), the raw locations of the each sound source were calculated. The auditory objects are established, labelled with their raw locations.
 6. Update object parameters: An Extended Kalman Filter (EKF) is assigned to each auditory object. The filter parameters depend on the nature of the sound. For relatively quickly changing patterns such as footsteps, the confidence of the object staying at the same place is low so a relatively large gain is given to the sensor error in calculation. For slowly changing or non-moving sources such as a clock, the gain is relatively small. The identification of the type of sounds could be done with other tools in machine hearing (Lyon, 2010).
 7. Control update: The head turn may last for a short moment after all the existing auditory objects are identified. The recursive filtering based on EKF would help enhancing the localization accuracy of each object. The change of acoustic cues due to the motion is calculated in the form of the mean and covariance in the self-centered coordinate system. Without self-motion or head turn, the control equation is still updated, without motion data. The Kalman filter gain is also updated, as discussed in Chapter 6.
 8. Measurement update: The ITD is projected based on the previous value and the motion data. The estimation of ITD is then compared to the next measurement of

ITD. The error between the estimation and the measurement is fed back into the model. It should be noted that a new measurement is accounted for in the current model based on two laws: a) the law of saliency, that the ITD has to be strong enough; and b) the law of continuity, that it is highly unlikely for ITD or the location of an auditory object to make abrupt changes. When either of the cues is not met, the filter would update the control function without updating the measurement function. The location of the auditory object would remain the same but the covariance would grow larger over time. The tracking of moving sources as relative to a static robot is another related topic. For simplicity, it is not discussed extensively here. It should be pointed out, however, that it is easy to track a slowly moving auditory object with multiple head turns and observations as shown in 5).

9. Covariance check: An auditory object that stops making sound is not immediately deleted. Instead, the control equation would repeat itself, adding to the covariance at each repetition, meaning that the model is less confident about the location estimation being true. When the covariance is above a certain threshold, the algorithm would eliminate a certain auditory object.
10. New source detection: The distribution of existing auditory objects occupy a portion of the possible cue (such as ITD) range corresponding to the locations. In the new measurement, if an alien cue, which does not belong to any of the existing set of sources, is observed, a head turn would be initiated to localize its azimuth and elevation, and a new EKF would be assigned to it.

11. Delete the object: The auditory object and the corresponding EKF model are deleted.
12. Update the number of objects: The number of auditory objects decreases by one.

CHAPTER 6

LOCALIZE A SOUND SOURCE IN SELF MOTION WITH ITD CUES

To use an array of two microphones to localize a sound source, inter-channel time difference (ITD) has some benefits over the other localization cues. ITD is well defined analytically and thus requires no table-look up (Gardner, 1998); also, simple and mature algorithms exist for the calculation of ITD based on binaural audio inputs (e.g., Knapp & Carter, 1976). However, if ITD is the only localization cue being used, an inevitable problem is the cone of confusion, which consists of locations in space with the same amount of ITD for the two sensors (Woodworth, 1965). The problem exists in both human and machine hearing.

In human spatial hearing, ITD is believed to be the major spatial hearing cue for low-frequency signals on the horizontal plane, and it may not have a lot to do with localization on the vertical plane, especially on the mid-sagittal plane. (Blauert, 1997). It should be noted, however, the possibility of using head motion to resolve the ambiguities has been discussed in the early days of spatial hearing research (Wallach, 1939). It was suggested that the ITD change with head turn is also a potential spatial hearing cue for disambiguating locations on the cone of confusion, especially on the vertical planes. Unfortunately, not a lot of studies have made this model more complete. In the past decades, a disproportionate amount of effort was spent on studying static sound source localization by static listeners. Head motion is believed to be at best a weak cue in spatial hearing (Middlebrooks & Green, 1991).

In robotics literature, using binaural audio inputs to localize both the azimuth and the elevation of the sound source is a difficult task. So far, the majority of the studies

involve some types of processing of the head-related-transfer-function (HRTF), which is essentially the spatial-angle-related acoustic diffraction pattern at high frequencies (Keyrouz & Diepold, 2006; Keyrouz, 2014). Models that combines ITD and inter-channel level difference (ILD) for localization on the horizontal plane has been suggested (Willert, et al., 2006). Recursive filters were used for the tracking of multiple simultaneous sound sources (Roman & Wang, 2003). Of particular interest is two comparatively recent works. Kneip and Baumann (2008) built a two microphone array based on LEGO blocks and constructed a mathematical model that was essentially very similar to the one described by Wallach (1938); some localization error was observed, and a recursive filter was not applied. On the other hand, Portello et al. (2011) extensively discussed the use of unscented Kalman filter (UKF) in the problem of sound source localization in self-motion, but only applied the model in azimuth and range estimation.

This study demonstrates the possibility of calculating azimuth/vertical angular localization of a static sound source using only ITD and motion data with a recursive filter. A dummy head is mounted on top of a rotating chair to mimic the head and body motion of human beings, as well as to collect audio data. A gyroscope was mounted on top of the dummy head to collect motion data. The functions governing ITD change and motion was described with a mathematical model that is more complete than that of Wallach's and much simpler and efficient than that of Kneip and Baumann's. An Extended Kalman Filter (EKF) was used to estimate the spatial angles of the sound sources with respect to the listener using the model and measured data. The effectiveness and robustness of the developed algorithm are shown by both the numerical and

experimental results, which reveal the quick convergence of the estimated spatial angles toward their real values given noisy measurements. Although the term self-motion has special meaning in human behavioral study, it is used on a dummy head that was rotated passively, because in practical use cases the motion is likely to be started by the robot itself.

PROBLEM STATEMENT

A two microphone array is placed on the horizontal plane of the earth-centered coordinate system. The center of the array is also the center of the coordinates. The distance between the left and the right microphone was defined as $2*b$, where b was the distance between the center and each microphone. As shown in Figure 6.1, a spherical coordinate system was used, in which a point in the field is uniquely defined by (r, θ, φ) , where

r – the distance between the source location and center of head, or direct path;

θ (theta) – the spatial angle between the direct path and z axis

φ (phi) – the spatial angle between the direct path and x axis

The direction in space is uniquely defined by (θ, φ) .

The task is to localize a static point sound source A in space. Without losing generality, the sensory array's resting position is along the y axis, and is allowed to rotate around the z direction (on the horizontal plane). The direction of A is defined in the polar coordinates in the form of (θ_A, φ_A) . The distance is not a concern for the current discussion.

DATA FUSION

When head motion is involved in spatial localization of sound sources, several assumptions can be made to simplify computations.

1. Spatial Continuity: it is highly possible that a sound source making continuous sound moves along a continuous trajectory, or is stationary; an intermittent sound source could still be considered to be at the same place when it is not making sound for a short while;
2. Relative Stillness: the slow motion of sound sources can be ignored during quick head motion;
3. Multiple Observations: the observation of the same set of sound sources at different moments and places could be combined.

Based on those rules, a recursive EKF filter is built to calculate the location of a sound source.

Mathematical Model

This section presents a simple model for the estimation of azimuth and elevation of the sound source location. This would be done in preparation for the derivation of an efficient recursive algorithm that fuses motion and binaural audition data. The method is only based on ITD and the change of ITD over time, which is similar to that of Kneip and Baumann (2008), but is far simpler in mathematical representation. The method is similar to the one suggested by Wallach (1938), and is more complete mathematically.

The two robot ears (microphones) are represented by two dots positioned along the x axis with $2b$ distance between them. To better discuss the spatial relations between

the direct path vector and the dual microphone array, it is favorable to discuss it within the plane of the grey rectangular triangle as shown in Figure 6.2. The length of the longest edge of the grey triangle is r . The side on the y axis equals $r \sin \theta \sin \varphi$ (r projected onto the x - y surface, then projected onto the y axis). If the spatial angle between the direct path vector and the y axis is defined as α , then we have:

$$r \cos \alpha = r \sin \theta \sin \varphi \quad (6.1)$$

or,

$$\cos \alpha = \sin \theta \sin \varphi \quad (6.2)$$

The direct paths from the target sound source A to the two microphones L and R are also shown in Figure 6.2. The distance difference between AL and AR is directly proportional to ITD (represented with D , equals $ITD * c\theta$, where $c\theta$ is the speed of sound). When distance between the microphones $2b$ is significantly less than direct path distance r , the longest two sides of the triangle APL are very close to each other, that is $AP \approx AL$; also, the two angles α and α' as shown in Figure 6.2 are close to each other, that is $\alpha \approx \alpha'$. With these two approximations, we have:

$$D = 2b * \cos \alpha \quad (6.3)$$

Combining (2) and (3):

$$D = 2b \sin \theta \sin \varphi \quad (6.4)$$

This equation is critical for the derivation of a recursive algorithm of sound source localization. The distance difference D can be computed from binaural recording. The absolute value of φ is unknown, but when the robot head is rotated relative to the z axis, the angle φ is changed, and this change can be used for prediction of the change of D , and correction can be made in a recursive filter. If both φ and D are known, θ can be

computed directly. The implementation of such a recursive algorithm is discussed in the next sections.

An option other than the recursive model is a non-recursive analytical method, which has lower accuracy and is potentially less stable, but could more clearly demonstrate how the sound source locations on the cone of confusion is disambiguated. Mathematically, the rate of *ITD change* associated with this head turn (denoted Δ) is derived by partially differential equation (4) as relative to φ :

$$\Delta = \frac{\partial D}{\partial \varphi} = 2b \sin \theta \cos \varphi \quad (6.5)$$

Equations (4,5) yield the following solutions:

$$\begin{cases} \varphi = \tan^{-1} \left(\frac{D}{\Delta} \right) \\ \theta = \sin^{-1} \frac{\sqrt{D^2 + \Delta^2}}{2b} \end{cases} \quad (6.6, 6.7)$$

The equations could give a unique spatial direction in the front-upper quarter-sphere defined by (θ, φ) , where horizontal angle $\varphi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and $\theta \in \left[0, \frac{\pi}{2}\right]$. This explains how static sound sources could be localized during self-motion. The solution was limited to the front-upper quadrant only for mathematical reasons (that the same sine value appears twice in most cases in the range of $[0, 2\pi]$). The solution range of equations 6.6 and 6.7 were only defined in their respective regions, and there was still a front-back confusion. One way to resolve the remaining confusion was to use the sign of Δ , which is decided by head motion direction and differs for front and back. It would also easy to simply do a head rotation along another axis. This method is close to the one suggested by Kneip and Baumann (2008), but is simpler. This method shows that the motion and *ITD* data over time suffice as spatial hearing cues for the computation of spatial angle $(\theta,$

φ). In the current study, the discussions were limited to the upper hemisphere. For lower sound sources under the plane of the ears, again using head rotation along another axis would help with disambiguation of sound sources in the upper-lower hemispheres.

ITD Estimation

The estimation of the time delay between two sensors within an array is commonly based on cross correlation (Knapp & Carter, 1976; Azaria & Hertz, 1984). The two signal paths of a robotic audition system can be modeled as:

$$\begin{cases} x_L(t) = s(t) + n_L(t) \\ x_R(t) = ILD s(t + ITD) + n_R(t) \end{cases} \quad (6.8, 6.9)$$

where,

$x_L(t)$ – signal at the left microphone;

$x_R(t)$ – signal at the right microphone;

$s(t)$ – signal;

$n_L(t)$ – noise of the left microphone;

$n_R(t)$ – noise of the right microphone;

And the noise and the target sound are independent of each other. The estimation of cross correlation is given by:

$$\hat{R}_{x_L x_R}(\tau) = \frac{1}{T-\tau} \int_{\tau}^T x_L(t) x_R(t - \tau) dt \quad (6.10)$$

The signal paths of the commonly used general cross correlation (GCC) methods is shown in Figure 6.3, which is largely the same as equation (10), except that two pre-

filters are added as an attempt to whiten the signals for more accurate ITD outputs. Depending on the nature of the target signal, different approaches to the pre-filter design could be implemented (summarized in Azaria & Hertz, 1984). ITD estimation is not the focus of the current study, and wide band noise sources are used as target. Consequently, the pre-filters are all-pass filters. It should be noted, however, that such an ITD model applies to the ideal case of two omnidirectional microphones mounted on two ends of a bar-shaped stand. When this assumption is violated, such as on a KEMAR dummy head for recording, some deviation, although maybe not a large one, to this rule can be expected (Gardner, 1998).

Recursive Filters

In this part, the use of recursive Bayes filters is applied in the problem of single sound source localization. The introduction of Bayes filters is by no means a systematic description or mathematically complete derivation. Rather, the essential parts of the filtering process are kept to bring the concept of probabilistic state estimation to the context of robotic sound source localization, and to theoretically prepare for the multiple sound source separation in the next sections. A more complete summary of implementation of probabilistic methods in robotics can be found in Thrun et al. (2005). A much more complete and lengthy discussion of general dynamic control theories can be found in Maybeck (1982). Some application notes of Kalman filters in small unmanned robots can be found in Beard and McLain (2012).

The term active binaural hearing listening is used to be consistent with existing robotics literature, in which “binaural” simply refers to the number of acoustic sensors (a

two microphone array instead of an array with three or more microphones). In the literature of spatial hearing in humans, it is more accurate to use the term “binaural” only in the case of simulated spatial hearing over headphones, and to use the term sound source localization with localization cases in rooms or free fields.

In the spatial hearing literature about human sound source localization, some discussions existed for cases in which the senses of balance and audition are combined (for a brief summary, see Zhong and Yost, 2013). However, it is rarely noted that no sensor, whether biological or electro-mechanical, are perfect, and that a human subject has to calculate the true state based only on noisy observations, and recursive filters is probably one of most powerful tools to use. Whereas in robotics, such discussions are abundant, especially in recent years (Thrun, 2000).

In robotics, measurement means data that are collected from sensors, and control means the change of environment or the position of the robot itself (gyroscope / odometer data included). A Bayes filter is built on the idea that an estimation of the true state, such as the location of a sound source, could be calculated based on a collection of past measurement and controls, and that errors of estimation could be utilized to change the gain of different components, so that further error is reduced. Bayes filter refers to a concept that includes a large number of applicable filter designs. Of particular interest is the EKF, which address the problem of nonlinear state estimation with first order Taylor expansion of the measurement and control parameters.

In an EKF model, it is assumed that the localization process is probabilistic in nature. A state is continuous in space but discrete in time. In the case of a Gaussian

distribution, the estimation of a state is described by multivariate probability density functions (PDF) in the form of:

$$p(x) = \det(2\pi P)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T P^{-1} (x - \mu) \right\} \quad (6.11)$$

where,

p – probability;

P – covariance matrix;

x – a specific event of a random variable;

μ – means vector;

The algorithm of EKF is outlined in Figure 6.4. The derivation of the equations are more completely discussed in Thrun (2005) and Welch and Bishop (2001).

For a dynamic model as described in (4), define the state vector as $x = \begin{bmatrix} \theta \\ \phi \end{bmatrix}$, in which θ is still the elevation angle of the sound source from z-axis, and ϕ is the horizontal angle of the sound source in a coordinate system that is rigidly linked to the microphone array. When the microphone array moves a horizontal angle ψ in the earth-centered coordinate system, we have:

$$\varphi = \phi + \psi \quad (6.12)$$

It is assumed that φ is constant. So any control that is applied on ψ (self-rotation) is also equally applied on ϕ (change of angular location of the static sound source, as observed by the moving microphone array), but with an opposite sign. The state equation and the measurement equation of the dynamic model become:

$$\begin{cases} \hat{x}_k^- = \hat{x}_{k-1} + Ts * B * u_{k-1} \\ \hat{z}_k = (2b/c_0) \sin \theta_{k-1} \sin \phi_{k-1} \end{cases} \quad (6.13, 6.14)$$

where,

B – motion status coefficient, $B = [0; -1]$;

Ts – sampling frequency of the gyroscope

u – reading of the gyroscope

b – half of the distance between the microphones

c_0 – sound speed

In plain words, the state equation (13) means: when the microphone array rotates in a certain direction, the locations of a static sound source as seen by the moving array can be expect to change in the opposite direction; the measurement equation (14) means: when the microphone array rotates, ITD would change in a sinusoidal pattern; the new value of ITD can be estimated based on past measurements of self-rotation and ITD; the maximum ITD is decided by the elevation angle θ , while the instantaneous ITD is decided by the horizontal spatial angle of the source as observed by the array ϕ .

Algorithm 1 EKF for Sound Source Localization with Head Turn

```
1: Initialize  $\hat{x}$  and  $P$ .
2: FOR each sample time  $Ts$ :
3:   FOR the  $i$ th element of  $x$ :                               {estimation, or control update}
4:      $\hat{x} = \hat{x} + \left(\frac{Ts}{N}\right) * B * u$                    {state estimation}
5:      $P = P + \left(\frac{Ts}{N}\right) * Q$                            {covariance estimation}
6:   END FOR
7:   IF autocorrelation of binaural inputs yield a valid ITD cue:
8:     FOR the  $i$ th element of  $x$ :                               {correction, or measurement update}
9:        $C_i = \frac{\partial h_i}{\partial x}(\hat{x}, u)$                    {Tylor expansion coefficient update}
10:       $K_i = \frac{PC_i^T}{C_i P_i^- C_i^T + R_i}$                    {Kalman gain computation}
11:       $\hat{x} = \hat{x} + K_i(z_i - h(\hat{x}, u))$                    {state update}
12:       $P = (I - K_i C_i)P$                                    {covariance update}
13:     END FOR
14:   END IF
15: END FOR
```

Combining (13, 14) and the EKF model in Figure 5, the result is shown as Algorithm 1. Like most applicable Bayes filters in robotics, the algorithm is efficient and easy to implement. In most cases, it is also highly robust, although sometimes overly confident. As will be discussed later, the model could potentially find evidence in the context of human binaural hearing modeling. Two important features set the model apart from existing models:

1. Recursive filtering: estimations are continuously made based on control data. The estimations are then corrected with binaural hearing data, as shown in line 3~6 and 8~13 in Algorithm 1;
2. Conditional correction: measurement update is made only when ITD cues clearly exists. When there is no sound, or when the ITD cue is obscure, the correction part is skipped (line 7 and 14). What this means is that, when no salient acoustic cues exist, the model still believes the sound source to remain at the same place; when further control such as rotation is applied, the location of sound source in a self-centered coordinate is projected based on the amount of control (line 4), but the confidence in the position is lowered, as shown by the larger covariance (line 5). The process continues until the next valid new ITD measurement comes in to update the localization result (line 7). This view of localization treats the sound source as an object instead of a series of auditory events. This model could be more practical, even biologically, given that natural sound is often intermittent, or could be silent for a short period of time. A similar approach has been suggested in the case of unmanned aircraft tracking (Beard & McLain, 2012). In the context

of machine audition, it is very meaningful because a large portion of natural sounds such as speech is intermittent.

CASE STUDY

Experiment Setup

The tests were done in a sound treated room in Arizona State University (see Yost and Zhong, 2014 for a full description of the test room). The dimension of the room was $15' \times 12' \times 9'$. The inner surfaces of the room were covered with 4-in. sound absorbing foam to reduce sound reflections. Audio signals were digitally generated from a Matlab program and three 12-channel Digital-to-Analog converters (model: Echo Gina 12) running at 44100 cycles/s per channel. Then the generated signals were amplified with AudioSource AMP 1200 amplifiers before they were played from an array of 36 loudspeakers (Boston Acoustics Soundware 100). Twenty-four loudspeakers were evenly arranged in the horizontal plane in a ring centered on the listener. The vertical level of the loudspeakers was the same as that of the listener's pinnae. A second layer of 8 loudspeakers (45° spacing) was on the sphere at $\sim 30^\circ$ elevation, and another layer of 4 loudspeakers (90° spacing) at $\sim 60^\circ$ elevation. All loudspeakers were 5 feet away from the listening position.

The recording device was the dummy head (Knowles Electronic Manikin for Acoustic Research, KEMAR) with dual microphones, which was temporarily mounted on a rotating chair in the middle of the room (Figure 6.5). Simulating head turn with whole-body passive rotation should not have affected the result profoundly, because ITD was the only spatial hearing cue being measured and body parts other than the head mainly

affect ILD and HRTF. The chair rotated at about $32^\circ/\text{s}$ for about two rounds for safety reasons. A gyroscope was mounted on top of the dummy head to collect motion data. ITD was processed with a GCC model in each time frame that corresponds to the sampling rate of the gyroscope, which was 120 Hz. The computation was done on a PC platform based on a Matlab program.

Results

The result of the experiment is shown in Figure 6.6. The dummy head rotated for a total 12 s (about one circular round). The two components of state vector x converged after about 4 s ($\sim 120^\circ$). The target was at $(60^\circ, 0^\circ)$, which was 30° above the horizontal plane at the resting direction of the dummy head since θ represented the angle from the vertical axis instead of the altitude. The localization result after convergence was at (57, 11.5), with an error that was smaller than previous studies without recursive filters (Kneip & Baumann, 2008). The data clearly showed that the localization on the vertical axis is possible, and the use of EKF in this scenario leads to converging results. A minor problem is that the commercially available AD/DA converters commonly have a sampling rate of no more than 44.1 kHz. For a dummy head the spacing between the two ears is only 19 cm, which mean the maximally possible time delay, as measured in digital audio samples, is only ± 19 samples. The stepping change in the first few data points may be the reason of some minor disturbances at the beginning of the converging period. But after a few seconds the effect was filtered out. Changing to higher speed sound card should help with the resolution of ITD. The model could be easily implement to head motion along other axis to further reduce disambiguation and improve accuracy.

CONCLUDING REMARKS

In this study an EKF based on data fusion of rotation velocity and ITD was presented. An acoustic model after Wallach (1938) was constructed to generate the measurement equation. The motion counteracts the change of location of the observed target, as described in the control equation. The combination of the two equations could be projected to compare to observations and produce localization results that outperform the non-recursive methods (Wallach, 1938; Kneip & Baumann, 2008). And both the angles representing the azimuth and altitude of the sound sources could be estimated.

A very interesting question is whether the same model may exist in animals. In physiology, the auditory system was separated into a few levels, with multiple bottom-up pathways for the processing of ITD, ILD and etc. However, afferent pathways from top to lower levels are also abundant. As the current model suggested, the existence of these pathways may serve to update the gain of recursive filters.

A final observation was that human listeners still differ from robot in many ways. In the comparison of robots, as described in Chapter 4. For humans the accuracy of timing is much lower. The spike rate of auditory neural fiber activity saturate at about 8,000 Hz. So any phase difference above 4 kHz could hardly be detected. For this specific reason, for higher frequency localization, the changing pattern of ILD, instead of ITD, may be used for localization. A major problem for using ILD is that around the maxima ILD, the same value may appear twice due to the shape of the head, this would make it very difficult to reliably localize sound sources with ILD alone. On the mid-

sagittal plane, sometimes ear differences may help, but consistency of changing pattern of high frequency peaks that may change, with elevation, in a systematic manner.

CHAPTER 7

DYNAMIC LOCALIZATION OF MULTIPLE SOUND SOURCES

Localizing multiple sound sources with a two-microphone-array on a dummy head is a very difficult task. This study presents a method that when head motion of the robot is allowed, could localize multiple locations of fixed sources that both temporally and spectrally overlapped.

A large portion of the existing literature on the topic of multiple sound source localization belong to the more general field of computational auditory scene analysis (Wang & Brown, 2006). Commonly the auditory space is divided into a large number of time-frequency (T-F) regions and statistical methods, and features related to the position of the sources in all regions are integrated (Roman & Wang, 2008). Most of the existing approaches use some type of statistics or machine learning algorithms (for a brief review, see Woodruff and Wang, 2012). Also, the majority of the existing work focuses on the azimuth location of sound sources, and not a lot of research deals with altitude estimation.

In this study, a simple fitting method that could separate and localize multiple sources is developed based on the ITD model in the last chapter. Localizing sound sources with ITD cues alone would inevitably lead to problems of cones of confusion. However, when head motion is initialized by a robot and the motion data are known, the possible changes of ITD that are associated with head motion are pre-determined for each sound source. The method is based on data fusion of motion and acoustic sensor inputs, and could estimate both the azimuth and altitude of each source. To emulate the head turn and sound sources, a dummy head (KEMAR) was rotated in the middle of a dome-shaped

loudspeaker array, and auditory and motion data were collected for the calculation of the location of sound sources. This method is non-statistical and has the advantages of mathematical clarity and ease of implementation at the cost of a delay of a few seconds and the requirement of some memory by the robot.

METHODS

The generalized cross correlation can be used to compute the ITD. When multiple sound sources are present, multiple peaks on the cross correlation function can be observed. If the observations were made with static binaural recording, such as a frame of the correlogram vs time chart in Figure 7.1, the number of peaks are not indicative of the number of sound sources because they can be reflections, neither are they indicative of the spatial location of the sound sources because of cones of confusion.

However, in dynamic binaural recording, when more degrees of freedom are given to the robotic head, the peaks on correlogram are subject to change. If the changes of cross correlation are plotted over time as shown in Figure 7.2, even though the chart is noisy due to reflections, certain patterns of changes can be observed.

With the mathematical model of ITD change in the previous chapter, with constant speed head rotation, the possible changing patterns of locations of cross-correlation peaks corresponding to real sound sources are sinusoidal and are governed by the following equation, based on discussion in Chapter 6 and Wallach (1938):

$$ITD(t) = \cos \theta \sin(\omega t + \varphi) \quad (7.1)$$

where, $ITD(t)$ is the location of a peak of the cross-correlation function whose pattern changes over time;

ω is the cycle of the changing pattern of *all* components; it is decided only by the angular speed of head rotation.

θ is the spatial angle from one sound source to the vertical axis, it is the complementary angle of the altitude angle; note that it is independent of t ;

φ is the horizontal angle of one sound source.

So the spatial angular locations of a set of sound sources could be fully described by the θ and φ of their corresponding sine wave components on the cross correlation chart. And the principle applies to multiple peaks when multiple sound sources are present. The changing patterns are superimposed in the case of multiple sound sources. The azimuth of an individual source is decided by the phase delay of individual components in the pattern in Figure 7.2. The altitude of an individual source is decided by the amplitude of individual components in Figure 7.2. Only the (θ, φ) combinations that consistently follows the sinusoid pattern would be identified as sound sources. Some reflections and noise could alter the cross correlation in a few time frames, but as long as the changes is not consistent, they would not qualify as sound sources.

Multiple methods could be used to separate the real signal from the noisy observations. For simplicity, a space of (θ, φ) was constructed. The cross correlation values along the path of the sine wave component corresponding to each (θ, φ) values was summated. A threshold of the summation was set to separate the real sound sources from the noise. The model could be easily expanded to cases in which the rotation of head was not at a constant speed:

$$y = \sum \cos \theta_i \sin(\int \omega(t)dt + \varphi_i) \quad (7.2)$$

EXPERIMENTS AND RESULTS

To simulate robot head turn, the KEMAR dummy head for binaural recording was mounted on top of a rotating chair. To simulate multiple sound sources, a loudspeaker array as shown in Figure 3 was used. Three layers of loudspeakers located on a dome with a radius of 2 m. The first horizontal layer consisted of 24 loudspeakers with 15° spacing. The second layer was at the altitude of $\sim 30^\circ$ and consisted of 8 loudspeakers with 45° spacing. The third layer was at the altitude of $\sim 60^\circ$ with 90° spacing.

The KEMAR was rotated at $30^\circ/s$ in the middle of the loudspeaker array. The signals were amplified and collected by a Tascam US-200 sound card. The digital signals were stored on a PC for further processing. The data processing and visualization were based on Matlab.

In the first experiment, three independent noise sources were played from different loudspeakers (#7, #31 and #34 in Figure 7.3). The noises were played at the same level and at the same time. The relation between cross correlation and time is shown in Figure 7.2. The localization results are shown in Figure 7.4 and Table 7.1. The sound sources were correctly separated, and the localization results were close to the real values, with errors shown as a percentile of the full possible range, i.e., 2π for azimuth and π for altitude.

In the second experiment, the content of the sources were different. The three sound sources were #10 (speech), #25 (music), and #36 (white noise). The signal-to-noise-ratios were -3dB for both the speech and music signals. As shown in Figure 7.5 and Table 7.2, the three sources could still be correctly separated. The noise source was the one that

provided the most energy, so the corresponding peak was also high. The music source contains rich and sometimes harmonic spectral contents, which was probably the reason for a more spreading region on the chart. The speech source was intermittent and contains the least energy, so the corresponding peak was a short and focused one. The error of localization were larger than that in the first experiment.

DISCUSSION

For a robot listener, knowing the total number of sound sources in the surrounding environment can be as important as knowing their individual locations. In application scenarios such as fire evacuation, security and surveillance, detecting the total number of nearby personnel is crucial for a robot to carry out its tasks. However, the topic of using active binaural hearing to separate and localize multiple sound sources has received limited attention in the fields of robotics and psychoacoustics modeling.

The current study is a part of an attempt to build a robot applicable algorithm for auditory space perception. For the challenging tasks of separating three sound sources with binaural inputs, the suggested methods could separate and localize up to three sound sources with a decent accuracy. Even though the three sources spectrally overlapped (three white noise sources in the case of experiment I), the method could still reliably separate the sources based on the changing pattern of autocorrelation. Also from the view point of correlogram over time, since the components representing different sources were categorized in their individual frequency “channels”, it does not matter whether or not the different sound sources are concurrent, intermittent, or alternating.

The proposed method for multiple sound source localization is highly connected to the EKF-based method for single sound source localization. The EKF-based method discussed before could be a more accurate method of localization, but it cannot be used directly with the case of multiple sources. The current approach, instead, can be used as a pre-processing stage and initialization of a number of independent EKF filters. The number of EKF filters equals the number of sound sources. Also, the initial value of the EKF filter's states equals the estimations of the fitting method. This way the time it requires for each of the EKF to converge is minimal.

Combining both the multiple source localization method and the EKF method could lead to a more complete algorithm for auditory object processing. The fast head turn is initiated when a sound or several sounds are heard. Each sound source are tagged with their raw locations and ITD. The EKF method is then used to individually track each of them during slow head and body motion. A more detailed “big picture” is shown in Chapter 5.

Several aspects of the current method can be improved in future studies. In the current method, the maximum of localizable sources is three, which is close to the maximum of number observed in human listeners in Chapter 2. This is an indication that the localization mechanism being proposed for robot hearing may be also true for human listeners. Whether more sources could be localized is an open question, which largely depends on the resolution of the (θ, φ) space. For now, the resolution of altitude is decided by the number of audio samples representing the ITD. For a commercial sound card working at 44.1 kHz, the peak-to-peak ITD is less than 60 samples, which is limiting the resolution of altitude estimation. Audio hardware with higher sampling frequency

could generate better resolution. The resolution of azimuth φ is arbitrarily chosen to be 30° in the current study. This value is only limited by computational load and head turn angular speed.

A further question concerns the discrimination of patterns in the noisy correlogram over time. The current study used a fitting method for simplicity. For more complicated scenarios in the future, the matured approaches in pattern recognition in image processing and the general digital signal processing can be employed.

The proposed method could and should be used together with the existing tools in computational auditory scene analysis (CASA). For example, when two singers are singing and standing close to each other, it is likely that they can be recognized as a single sound source with the current method due to spatial clustering. With CASA the spectral-temporal components can be analyzed individually to better discriminate each singer.

CONCLUDING REMARKS

In the current study, a fitting method based on cross correlation change with head turn was proposed to separate multiple sound sources. For the complicated problems of separating and localizing three sound sources with binaural inputs, the proposed algorithm generated decent results. The method was validated using binaural recording from KEMAR mounted on the top of a rotating chair. Future directions include the improvements in the resolution of localization, the total number of sources, and the combination with recursive filters.

CHAPTER 8

SUMMARY AND FUTURE DIRECTIONS

This dissertation consisted of a series of studies on both human and robot listeners surrounding the central topic of dynamic spatial hearing. Although the studies covered a wide range of topics from behavioral psychology to signal processing, they were connected as shown in Figure 8.1, which is a schematic structure of the whole dissertation. This chapter would be organized according to this structure instead of in a numerical sequence.

Regarding the maximal number of sound sources that could be perceived by the human listeners, there should be a limit because of the limitation of the capability of the brain. The only surprise was that the problem has been little discussed before. In Chapter 2 it was found that the human listeners made increasingly larger localization errors when the total number of sources increased. And the maximal total number of sources that could be simultaneously localized was around four. Future studies should explore the same limit for other types of stimuli, such as noises and pure tones. Also, it would be interesting to see whether whole body rotation could help increase this limit, because the motion information could potentially provide extra spatial hearing cues as suggested by the Wallach (1938) model.

A related study on multiple sound source localization was done in Chapter 7 on robotic hearing. When the binaural recording and motion data were available, the correlogram could be plotted as a function of time, and the locations of multiple simultaneous sounds could be estimated with a decent accuracy, in terms of both azimuth

and altitude, which superseded the performance of previous algorithms. An interesting finding was that the maximal number of sources that could be detected by the suggested algorithm for robots was only a little larger than three, and hence was very close to the results in the human behavioral study in Chapter 2. This may be a partial coincidence, especially considering that the robot was moving and that the limit on robots may be largely due to available commercial hardware. However, the similarity in the phenomenon found in human and robot listeners indicated that in both systems there was an inherent limitation, which was due to sensor errors on one hand and due to computational load on the other.

Regarding sound source localization in whole-body rotation, Chapter 4 found that human listeners were easily confounded without landmarks due to the difference between the world-centered coordinate system and the head-centered system. This result agreed with and complemented the finding of Yost et al. (2013). With acoustic landmarks they were much better able to tell where a target sound source had been compared to cases without landmarks. This study also suggested that the human listeners were very capable of keeping track of sounds.

A related study in Chapter 6 investigated a possible mechanism of sound source tracking in whole-body rotation or self-motion. A tentative localization model of machine hearing was built based on the change of ITD cues and motion data only. The model estimated the next value of ITD based on the current ITD value and the motion of self. When the next sample of ITD came in, the error between the estimation and the new observation was calculated and used to modify the coefficients in the model. The localization accuracy was surprisingly good when the model was implemented using

EKF filters. It was doubtful that the auditory system was built on exactly the same model. Nonetheless the possibility of a similar recursive filtering method in the human auditory system should be seriously considered.

Regarding the current prevailing theory of auditory motion, Chapter 3 suggested that not all auditory motion perception can be explained by the snapshot theory. Motion was clearly perceived and the direction of simulated motion was correctly reported when certain pattern of low-rate AM was applied on four independent noises. As suggested in Chapter 1, four independent static noises could still be localized simultaneously. Consequently, the snapshot theory does not predict motion perception in this case, because in each snapshot there were only four static sources. This study suggested that the snapshot theory was an incomplete view of auditory motion.

A related machine hearing study in Chapter 5 also concerned with alternative explanations of auditory motion, in particular, the tracking of static sources in self-motion. In fact, when applying the concept of snapshot theory in any actual signal processing system, one would immediately have to face with the problem of how to link the different snapshots. The major difficulty with the snapshot theory was that it is an overly simplified view of auditory motion and merely stated somehow the snapshots could be combined, but did not state how the combination might happen. And the difficulty would grow exponentially when multiple sources exist. In Chapter 5, the concept of auditory objects was explicitly defined in mathematics. Based on data fusion of acoustic and motion sensors, how an auditory object was established, maintained and deleted was clearly stated. The success of the implementation of the concept in Chapters 6 and 7

suggested that this model may at least serve as a major addition to the current snapshot theory, if not a replacement.

Overall, this dissertation documented an adventure in the Psychoacoustics Lab of ASU into the largely unknown field of spatial hearing and motion. The new model and theory regarding auditory spatial layout and motion perception was by no means exhaustive. But the author hope it would prove to be a major step towards a more complete and realistic understanding of both human and robot spatial hearing.

REFERENCES

- Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, *14*(3), 257-262.
- Alink, A., Singer, W., & Muckli, L. (2008). Capture of auditory motion by vision is represented by an activation shift from auditory to visual motion cortex. *The Journal of Neuroscience*, *28*(11), 2690-2697.
- Anderson, S. J., & Burr, D. C. (1985). Spatial and temporal selectivity of the human motion detection system. *Vision research*, *25*(8), 1147-1154.
- Andersen, R. A., Snyder, L. H., Batista, A. P., Buneo, C. A., & Cohen, Y. E. (1998). Posterior parietal areas specialized for eye movements (LIP) and reach (PRR) using a common coordinate frame. *Sensory guidance of movement*, 109.
- Arnoult, M. D. (1950). Post-rotatory localization of sound. *The American journal of psychology*, *63*(2), 229-236.
- Azaria, M., & Hertz, D. (1984). Time delay estimation by generalized cross correlation methods. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *32*(2), 280-285.
- Bailey, T., & Durrant-Whyte, H. (2006). Simultaneous localization and mapping (SLAM): Part II. *IEEE Robotics & Automation Magazine*, *13*(3), 108-117.
- Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *JOSA A*, *20*(7), 1391-1397.
- Beard, R. W., & McLain, T. W. (2012). *Small unmanned aircraft: Theory and practice*. Princeton University Press.
- Bekesy, G. V. (1955). Human skin perception of traveling waves similar to those on the cochlea. *The Journal of the Acoustical Society of America*, *27*, 830.
- Benesty, J., Chen, J., & Huang, Y. (2008). *Microphone array signal processing* (Vol. 1). Springer Science & Business Media.
- Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT press.

- Bosun, X. (2001). Signal Mixing for a 5.1-Channel Surround Sound System: Analysis and Experiment. *Journal of the Audio Engineering Society*, 49(4), 263-274.
- Borst, A., & Egelhaaf, M. (1989). Principles of visual motion detection. *Trends in neurosciences*, 12(8), 297-306.
- Bortolami, S. B., Pierobon, A., DiZio, P., & Lackner, J. R. (2006). Localization of the subjective vertical during roll, pitch, and recumbent yaw body tilt. *Experimental brain research*, 173(3), 364-373.
- Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.
- Bronkhorst, A. W., Veltman, J. H., & Van Breda, L. (1996). Application of a three-dimensional auditory display in a flight task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 38(1), 23-33.
- Brugge, J. F. (1985). Patterns of organization in auditory cortex. *The Journal of the Acoustical Society of America*, 78(1), 353-359.
- Brungart, D. S., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America*, 106, 1465.
- Caclin, A., Soto-Faraco, S., Kingstone, A., & Spence, C. (2002). Tactile “capture” of audition. *Perception & Psychophysics*, 64(4), 616-630.
- Carlile, S., & Best, V. (2002). Discrimination of sound source velocity in human listeners. *The Journal of the Acoustical Society of America*, 111, 1026.
- Clark, B., & Graybiel, A. (1949). The effect of angular acceleration on sound localization: the audiogyral illusion. *The Journal of Psychology*, 28(1), 235-244.
- Clark, J. J., & Yuille, A. L. (1990). *Data fusion for sensory information processing systems*. Kluwer Academic Publishers.
- Chen, L., & Vroomen, J. (2013). Intersensory binding across space and time: A tutorial review. *Attention, Perception, & Psychophysics*, 1-22.
- Deas, R., Adamson, R. B., Garland, P., Bance, M. L., & Brown, J. (2011, June). Combining auditory and tactile inputs to create a sense of auditory space. In *Proceedings of Meetings on Acoustics* (Vol. 11, p. 015003).

- Dichgans, J., & Brandt, T. (1978). Visual-vestibular interaction: Effects on self-motion perception and postural control. In *Perception* (pp. 755-804). Springer Berlin Heidelberg.
- DiZio, P., Held, R., Lackner, J. R., Shinn-Cunningham, B., & Durlach, N. (2001). Gravitoinertial force magnitude and direction influence head-centric auditory localization. *Journal of neurophysiology*, 85(6), 2455-2460.
- Dodge, R. (1923). Thresholds of Rotation. *Journal of Experimental Psychology*, 6(2), 107.
- Easton, R. D., Greene, A. J., DiZio, P., & Lackner, J. R. (1998). Auditory cues for orientation and postural control in sighted and congenitally blind people. *Experimental Brain Research*, 118(4), 541-550.
- Fernandes, A. M., & Albuquerque, P. B. (2012). Tactual perception: a review of experimental variables and procedures. *Cognitive processing*, 13(4), 285-301.
- Frost, B. J., & Richardson, B. L. (1976). Tactile localization of sounds: Acuity, tracking moving sources, and selective attention. *The Journal of the Acoustical Society of America*, 59, 907.
- Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological psychology*, 5(1), 1-29.
- Gardner, M. B. (1969). Image Fusion, Broadening and Displacement in Sound Localization. *The Journal of the Acoustical Society of America*, 45(1), 328-328.
- Gardner, W. G. (1998). *3-D audio using loudspeakers*. Springer Science & Business Media.
- Gescheider, G. A. (1970). Some comparisons between touch and hearing. *Man-Machine Systems, IEEE Transactions on*, 11(1), 28-35.
- Getzmann, S., & Lewald, J. (2007). Localization of moving sound. *Perception & psychophysics*, 69(6), 1022-1034.
- Gilkey, R., & Anderson, T. R. (Eds.). (2014). *Binaural and spatial hearing in real and virtual environments*. Psychology Press.

- Grantham, D. W. (1986). Detection and discrimination of simulated motion of auditory targets in the horizontal plane. *The Journal of the Acoustical Society of America*, 79(6), 1939-1949.
- Grantham, D. W. (1997). Auditory motion perception: Snapshots revisited. *Binaural and spatial hearing in real and virtual environments*, 295-313.
- Grantham, D. W., & Wightman, F. L. (1978). Detectability of varying interaural temporal differences. *The Journal of the Acoustical Society of America*, 63(2), 511-523.
- Graybiel, A., & Niven, J. I. (1951). The effect of a change in direction of resultant force on sound localization: the audiogravic illusion. *Journal of Experimental psychology*, 42(4), 227.
- Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5(11), 887-892.
- Herder, J. (1999, September). Visualization of a clustering algorithm of sound sources based on localization errors. *The Journal of Three Dimensional Images, 3D-Forum Society* (Vol. 13, No. 3, pp. 66-70).
- Holmes, N.P., 2007. The law of inverse effectiveness in neurons and behaviour: multisensory integration versus normal variability. *Neuropsychologia*, 45, 3340–3345.
- Howard, I. P., & Howard, A. (1994). Vection: the contributions of absolute and relative visual motion. *Perception*, 23, 745-745.
- Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual and motor skills*, 37(3), 967-979.
- Jacobson, G. P., & Shepard, N. T. (Eds.). (2008). *Balance function assessment and management*. Plural Publishing.
- Jeffress, L. A. (1948). A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1), 35.
- Jeka, J. J. (1997). Light touch contact as a balance aid. *Physical Therapy*, 77(5), 476-487.
- Jones, B. (1975). Spatial perception in the blind. *British Journal of Psychology*, 66(4), 461-472.

- Jones, B., & Kabanoff, B. (1975). Eye movements in auditory space perception. *Perception & Psychophysics*, *17*(3), 241-245.
- Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron*, *48*(2), 373-384.
- Keyrouz, F. (2014). Advanced binaural sound localization in 3-D for humanoid robots. *Instrumentation and Measurement, IEEE Transactions on*, *63*(9), 2098-2107.
- Keyrouz, F., & Diepold, K. (2006, August). An enhanced binaural 3D sound localization algorithm. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on* (pp. 662-665). IEEE.
- Knapp, C., & Carter, G. C. (1976). The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *24*(4), 320-327.
- Kneip, L., & Baumann, C. (2008). Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis. *The Journal of the Acoustical Society of America*, *124*(5), 3108-3119.
- Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712-719.
- Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, *62*, 157.
- Kuhn, G. F., & Guernsey, R. M. (1983). Sound pressure distribution about the human head and torso. *The Journal of the Acoustical Society of America*, *73*, 95.
- Lackner, J. R. (1974). The role of posture in sound localization. *The Quarterly journal of experimental psychology*, *26*(2), 235-251.
- Lackner, J. R. (1983). Influence of posture on the spatial localization of sound. *Journal of the Audio Engineering Society*, *31*(9), 650-661.
- Lackner, J. R., & DiZio, P. A. (2000). Aspects of body self-calibration. *Trends in cognitive sciences*, *4*(7), 279-288.
- Lackner, J. R., & DiZio, P. (2005). Vestibular, proprioceptive, and haptic contributions to spatial orientation. *Annu. Rev. Psychol.*, *56*, 115-147.

- Langendijk, E. H., Kistler, D. J., & Wightman, F. L. (2001). Sound localization in the presence of one or two distracters. *The Journal of the Acoustical Society of America*, 109, 2123.
- Lamb, G. D. (1983). Tactile discrimination of textured surfaces: psychophysical performance measurements in humans. *The Journal of Physiology*, 338(1), 551-565.
- Lewald, J. (1997). Eye-position effects in directional hearing. *Behavioural brain research*, 87(1), 35-48.
- Lewald, J., & Ehrenstein, W. H. (1998). Influence of head-to-trunk position on sound lateralization. *Experimental brain research*, 121(3), 230-238.
- Lewald, J., & Karnath, H. O. (2002). The effect of whole-body tilt on sound lateralization. *European Journal of Neuroscience*, 16(4), 761-766.
- Lyon, R. F. (1983, April). A computational model of binaural localization and separation. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83*. (Vol. 8, pp. 1148-1151). IEEE.
- Lyon, R. F. (2010). Machine hearing: An emerging field [exploratory dsp]. *Signal Processing Magazine, IEEE*, 27(5), 131-139.
- Mastroianni, G. R. (1982). The influence of eye movements and illumination on auditory localization. *Perception & Psychophysics*, 31(6), 581-584.
- Maybeck, P. S. (1982). *Stochastic models, estimation, and control* (Vol. 3). Academic press.
- Mayne, R. (1952). The audiogyral illusion and the mechanism of spatial representation. *The bulletin of mathematical biophysics*, 14(1), 27-34.
- McKinley, R. L., Erickson, M. A., & D'Angelo, W. R. (1994). 3-Dimensional auditory displays: Development, applications, and performance. *Aviation, Space, and Environmental Medicine*.
- Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*. 221(7), 389-391
- Meredith, A. M., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research*, 365(2), 350-354.

- Meredith, M. A., & Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus neurons. *Journal of Neurophysiology*, 75(5), 1843-1857.
- Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of neuroscience*, 7(10), 3215-3229.
- Meyer, G. F., & Wuerger, S. M. (2001). Cross-modal integration of auditory and visual motion signals. *Neuroreport*, 12(11), 2557-2560.
- Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual review of psychology*, 42(1), 135-159.
- Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of neuroscience*, 25(25), 5884-5893.
- Moeck, T., Bonneel, N., Tsingos, N., Drettakis, G., Viaud-Delmon, I., & Alloza, D. (2007, April). Progressive perceptual audio rendering of complex scenes. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*(pp. 189-196). ACM.
- Perrott, D. R., & Marlborough, K. (1989). Minimum audible movement angle: marking the end points of the path traveled by a moving sound source. *The Journal of the Acoustical Society of America*, 85, 1773.
- Perrault, T. J., Vaughan, J. W., Stein, B. E., & Wallace, M. T. (2005). Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli. *Journal of neurophysiology*, 93(5), 2575-2586.
- Perrott, D. R., & Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87, 1728.
- Pettorossi, V. E., Brosch, M., Panichi, R., Botti, F., Grassi, S., & Troiani, D. (2005). Contribution of self-motion perception to acoustic target localization. *Acta otolaryngologica*, 125(5), 524-528.
- Platt, B. B., & Warren, D. H. (1972). Auditory localization: The importance of eye movements and a textured visual environment. *Perception & Psychophysics*, 12(2), 245-248.

- Poirier, C., Collignon, O., Scheiber, C., Renier, L., Vanlierde, A., Tranduy, D., & De Volder, A. G. (2006). Auditory motion perception activates visual motion areas in early blind subjects. *Neuroimage*, *31*(1), 279-285.
- Portello, A., Danes, P., & Argentieri, S. (2011, September). Acoustic models and Kalman filtering strategies for active binaural sound localization. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on* (pp. 137-142). IEEE.
- Previc, F. H., & Ercoline, W. R. (Eds.). (2004). *Spatial disorientation in aviation* (Vol. 203). Aiaa.
- Pulkki, V., & Karjalainen, M. (2001). Localization of amplitude-panned virtual sources I: stereophonic panning. *Journal of the Audio Engineering Society*, *49*(9), 739-752.
- Raykar, V. C., Duraiswami, R., & Yegnanarayana, B. (2005). Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *The Journal of the Acoustical Society of America*, *118*, 364.
- Rayleigh, L. (1907). XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *13*(74), 214-232.
- Riecke, B. E., Feuereissen, D., & Rieser, J. J. (2008, August). Auditory self-motion illusions (circular vection) can be facilitated by vibrations and the potential for actual motion. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization* (pp. 147-154). ACM.
- Riecke, B. E., Våljamäe, A., & Schulte-Pelkum, J. (2009). Moving sounds enhance the visually-induced self-motion illusion (circular vection) in virtual reality. *ACM Transactions on Applied Perception (TAP)*, *6*(2), 7.
- Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society B: biological sciences*, *273*(1598), 2159-2168.
- Roman, N., & Wang, D. (2008). Binaural tracking of multiple moving sources. *Audio, Speech, and Language Processing, IEEE Transactions on*, *16*(4), 728-739.
- Saenz, M., Lewis, L. B., Huth, A. G., Fine, I., & Koch, C. (2008). Visual motion area MT+/V5 responds to auditory motion in human sight-recovery subjects. *The Journal of Neuroscience*, *28*(20), 5141-5148.

- Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia*, *45*(3), 561-571.
- Shelton, B. R., & Searle, C. L. (1980). The influence of vision on the absolute identification of sound-source position. *Perception & Psychophysics*, *28*(6), 589-596.
- Soto-Faraco, S., Kingstone, A., & Spence, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia*, *41*(13), 1847-1862.
- Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., & Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and hearing*, *33*(1), 112-117.
- Spahr, A., Saoji, A., Litvak, L., & Dorman, M. (2011). Spectral cues for understanding speech in quiet and in noise. *Cochlear implants international*, *12*(s1), S66-S69.
- Stanford, T. R., Quessy, S., & Stein, B. E. (2005). Evaluating the operations underlying multisensory integration in the cat superior colliculus. *The Journal of Neuroscience*, *25*(28), 6499-6508.
- Santala, O., & Pulkki, V. (2011). Directional perception of distributed sound sources. *The Journal of the Acoustical Society of America*, *129*(3), 1522-1530.
- Stevens, S. S., & Newman, E. B. (1936). The localization of actual sources of sound. *The American Journal of Psychology*, *48*(2), 297-306.
- Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: inverse effectiveness and the neural processing of speech and object recognition. *Neuroimage*, *44*(3), 1210-1223.
- Strybel, T. Z., & Neale, W. (1994). The effect of burst duration, interstimulus onset interval, and loudspeaker arrangement on auditory apparent motion in the free field. *The Journal of the Acoustical Society of America*, *96*, 3463.
- Thrun, S. (2000). Probabilistic algorithms in robotics. *Ai Magazine*, *21*(4), 93.
- Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. MIT press.

- Thurlow, W. R., & Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". *Perceptual and motor skills*, 36(3c), 1171-1184.
- Thurlow, W. R., Mangels, J. W., & Runge, P. S. (1967). Head movements during sound localization. *The Journal of the Acoustical society of America*, 42, 489.
- Trevarthen, C. B. (1968). Two mechanisms of vision in primates. *Psychologische Forschung*, 31(4), 299-337.
- Tsingos, N., Gallo, E., & Drettakis, G. (2004, August). Perceptual audio rendering of complex virtual environments. In *ACM Transactions on Graphics (TOG)* (Vol. 23, No. 3, pp. 249-258). ACM.
- Väljamäe, A. (2009). Auditorily-induced illusory self-motion: A review. *Brain research reviews*, 61(2), 240-255.
- Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2005, July). Travelling without moving: Auditory scene cues for translational self-motion. In *Proceedings of ICAD'05*.
- Verron, C., Aramaki, M., Kronland-Martinet, R., & Pallone, G. (2010). A 3-D immersive synthesizer for environmental sounds. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6), 1550-1561.
- Wallach, H. (1939). On sound localization. *The Journal of the Acoustical Society of America*, 10(4), 270-274.
- Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4), 339.
- Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press.
- Warren, D. H. (1970). Intermodality interactions in spatial localization. *Cognitive Psychology*, 1(2), 114-133.
- Welch, G., & Bishop, G. (2001). An introduction to the Kalman filter. Proceedings of the Siggraph Course, Los Angeles.
- Wightman, F. L., & Jenison, R. (1995). Auditory spatial layout. *Perception of space and motion*, 365-400.

- Willert, V., Eggert, J., Adamy, J., Stahl, R., & Korner, E. (2006). A probabilistic model for binaural sound localization. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(5), 982-994.
- Wolters, N. C., & Schiano, D. J. (1989). On listening where we look: The fragility of a phenomenon. *Perception & psychophysics*, 45(2), 184-186.
- Woodruff, J., & Wang, D. (2012). Binaural localization of multiple sources in reverberant and noisy environments. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(5), 1503-1512.
- Woodworth, R. S. (1965). *Experimental psychology*. Oxford and IBH Publishing.
- Woodworth, R. S., & Schlosberg, H. (1962). *Experimental psychology*. Holt, Rinehart and Winston.
- Wightman, F. L., & Kistler, D. J. (1997). Monaural sound localization revisited. *The Journal of the Acoustical Society of America*, 101(2), 1050-1063.
- Yost, W. A., & Brown, C. A. (2013). Localizing the sources of two independent noises: Role of time varying amplitude differences. *The Journal of the Acoustical Society of America*, 133(4), 2301-2313.
- Yost, W. A., Loisel, L., Dorman, M., Burns, J., & Brown, C. A. (2013). Sound source localization of filtered noises by listeners with normal hearing: A statistical analysis. *The Journal of the Acoustical Society of America*, 133, 2876.
- Yost, W. A., & Zhong, X. (2014). Sound source localization identification accuracy: Bandwidth dependencies. *The Journal of the Acoustical Society of America*, 136(5), 2737-2746.
- Yost, W. A., Zhong, X., & Najam, A. (2013). Where am I, where is the sound source?. *The Journal of the Acoustical Society of America*, 134(5), 4194-4194.
- Young, L. R., & Oman, C. M. (1969). Model for vestibular adaptation to horizontal rotation. *Aerospace Med*, 40(10), 1076-1080.
- Zacharias, G. L., & Young, L. R. (1981). Influence of combined visual and vestibular cues on human perception and control of horizontal rotation. *Experimental Brain Research*, 41(2), 159-171.

Zahorik, P., Bangayan, P., Sundareswaran, V., Wang, K., & Tam, C. (2006). Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. *The Journal of the Acoustical Society of America*, *120*, 343.

Zhong, X., & Yost, W. A. (2013). Relationship between Postural Stability and Spatial Hearing. *Journal of the American Academy of Audiology*, *24*(9), 782-788.

Table 4.1

The results of dynamic sound sources localization in whole body rotation

Subject Number	RMS error in the Condition without Acoustic Landmark (°)	RMS error in the Condition with Acoustic Landmark (°)
1	57.4	12.2
2	122.5	53.4
3	79.4	24.5
4	90.8	84.9
5	116.8	24.5
6	60.0	21.2
7	143.4	56.1
8	118.1	34.6

Table 7.1

The results of Experiment I, localization of three sound sources (noises)

	Azimuth (°)	Estimation of Azimuth (°)	Error (%)	Altitude (°)	Estimation of Altitude (°)	Error (%)
1	90	90	0	60	53	8%
2	-90	-90	0	30	32	9%
3	90	90	0	0	18	20%

Table 7.2

The results of Experiment II, localization of three sound sources (speech, music, and white noise)

	Azimuth (°)	Estimation of Azimuth (°)	Error (%)	Altitude (°)	Estimation of Altitude (°)	Error (%)
1	-90	-120	8%	60	49	12%
2	0	0	0%	30	49	21%
3	120	90	8%	0	0	0%

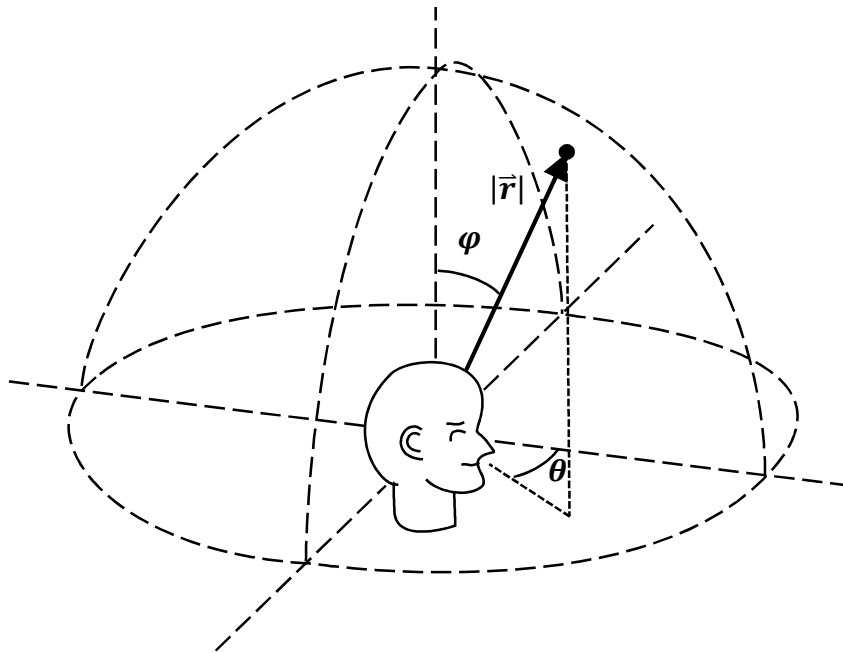


Figure 1.1. The framework of reference for sound source localization. A location in space is defined with its azimuth, altitude, and distance (adapted from Blauert, 1997).

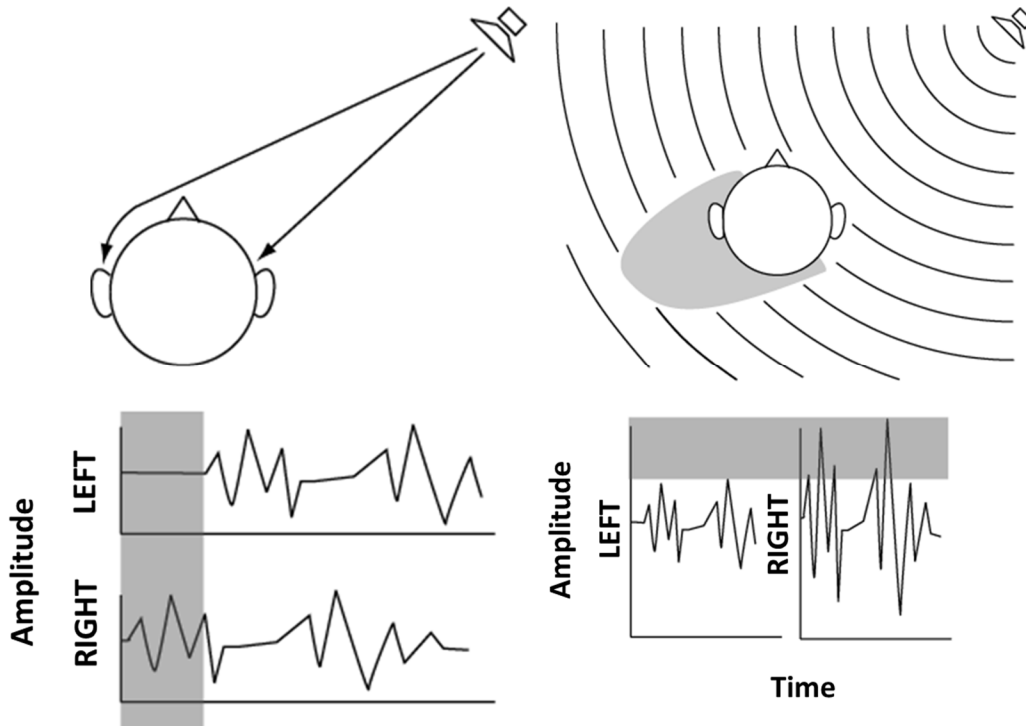


Figure 1.2. Interaural Time Difference (ITD) and Interaural Level Difference (ILD).

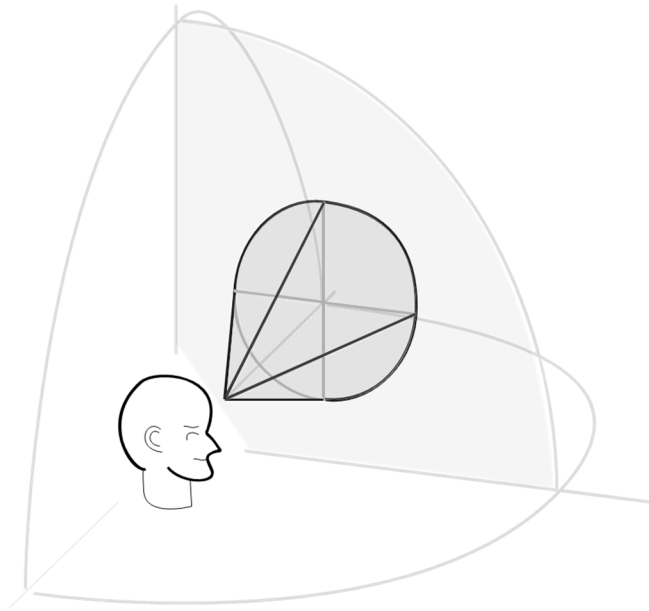


Figure 1.3. The Cone of confusion.

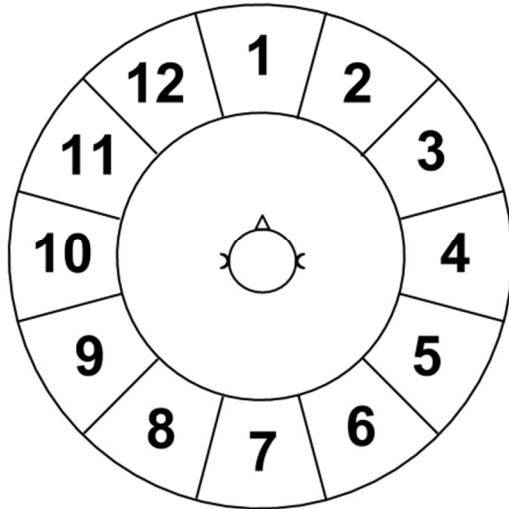


Figure 2.1. Test setup for the localization tasks. Loudspeakers were numbered 1~12 with 30° spacing on the horizontal plane at the level of the ears of the listeners.

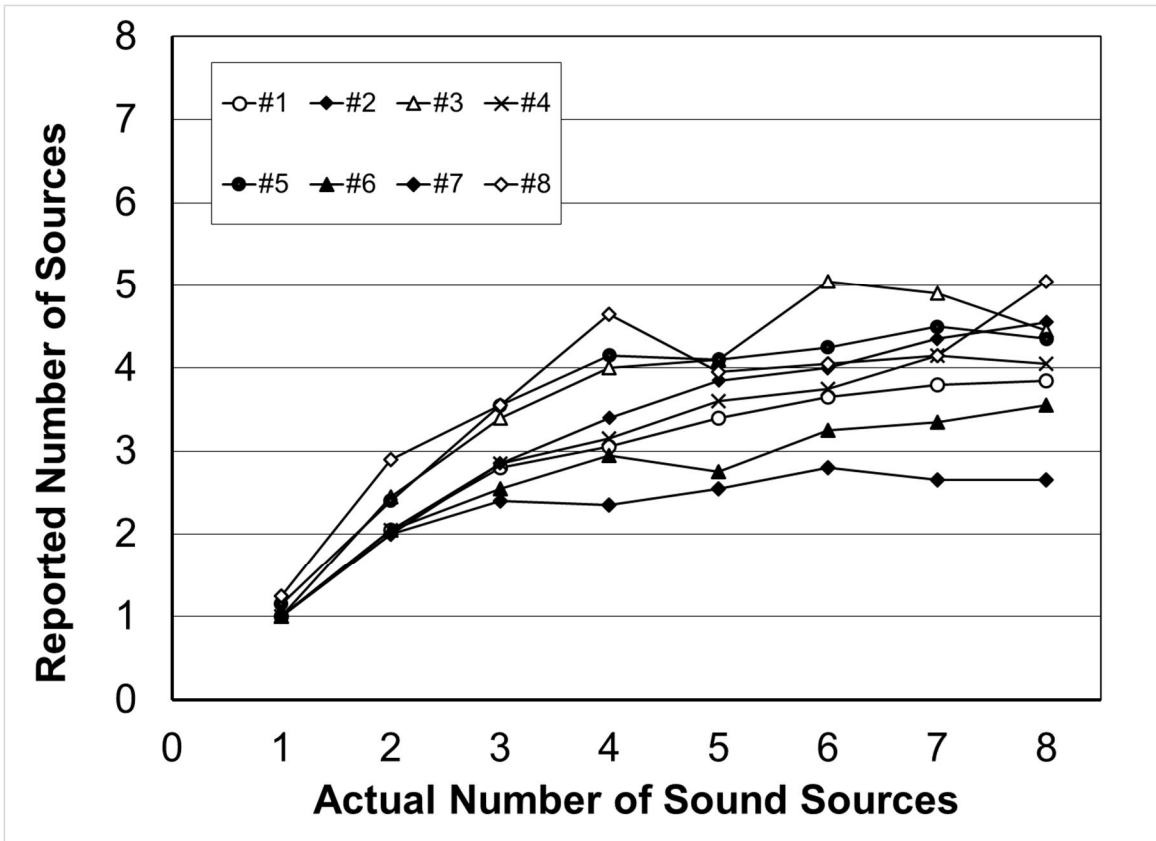


Figure 2.2. Individual results of all listeners (8) in Experiment I showing the relationship between the reported and actual total number of sound sources.

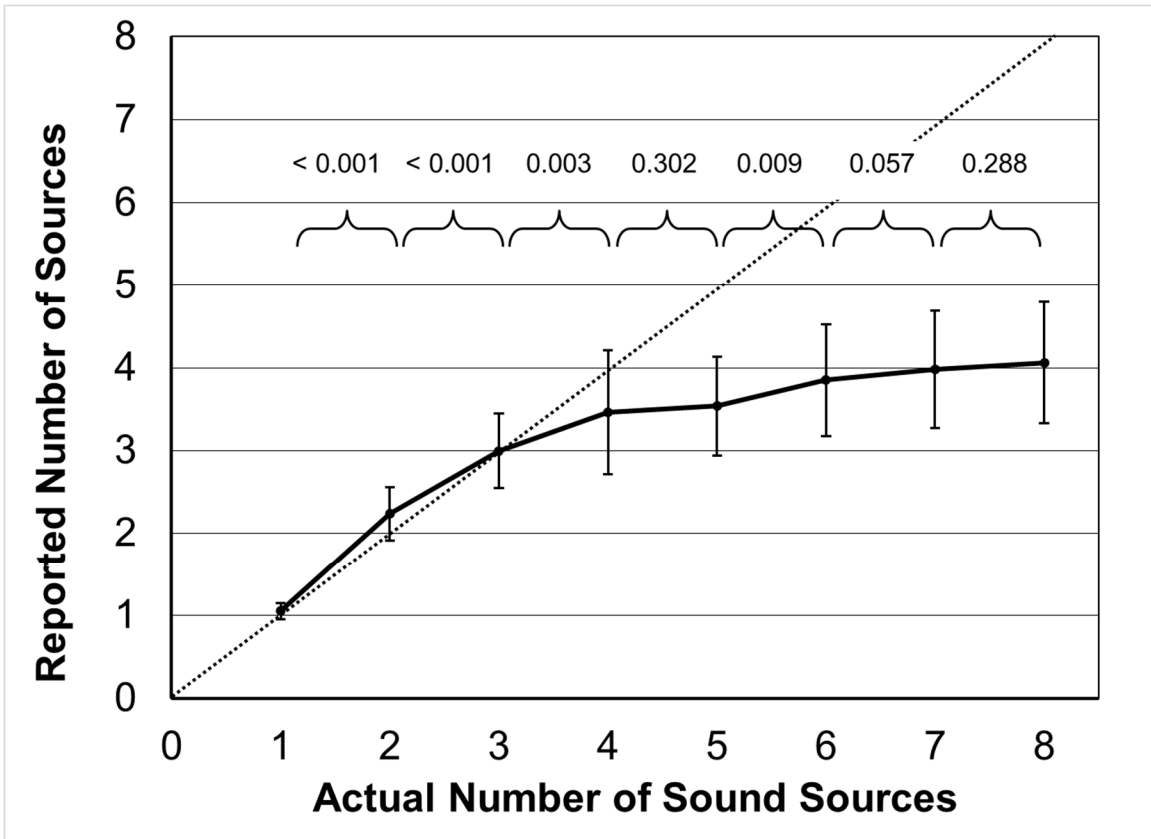


Figure 2.3. Averaged results of all listeners in Experiment I showing the relationship between the reported and actual total number of sound sources. The dotted diagonal line represents correct responding. The significance of the difference between adjacent conditions is shown on top in terms of p-values in paired t-tests.

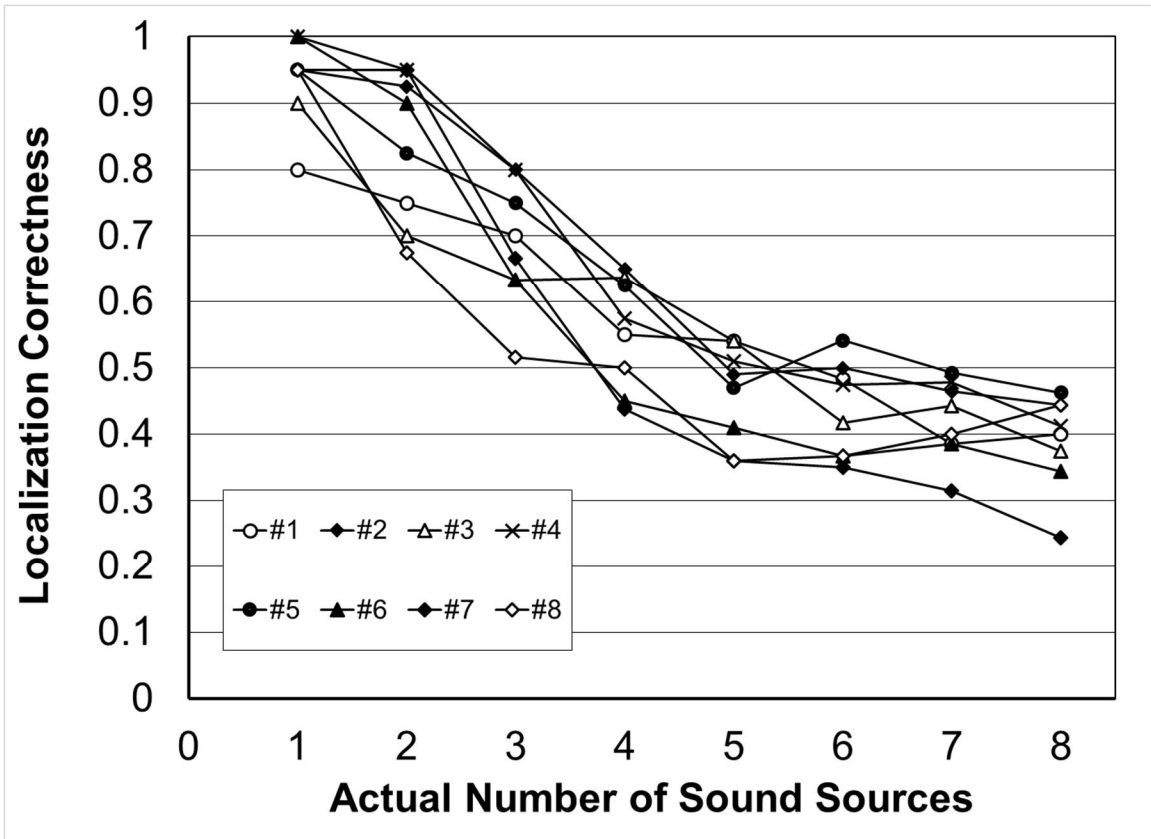


Figure 2.4. Individual results (8 listeners) in Experiment I showing the relationship between localization correctness and the actual number of sound sources.

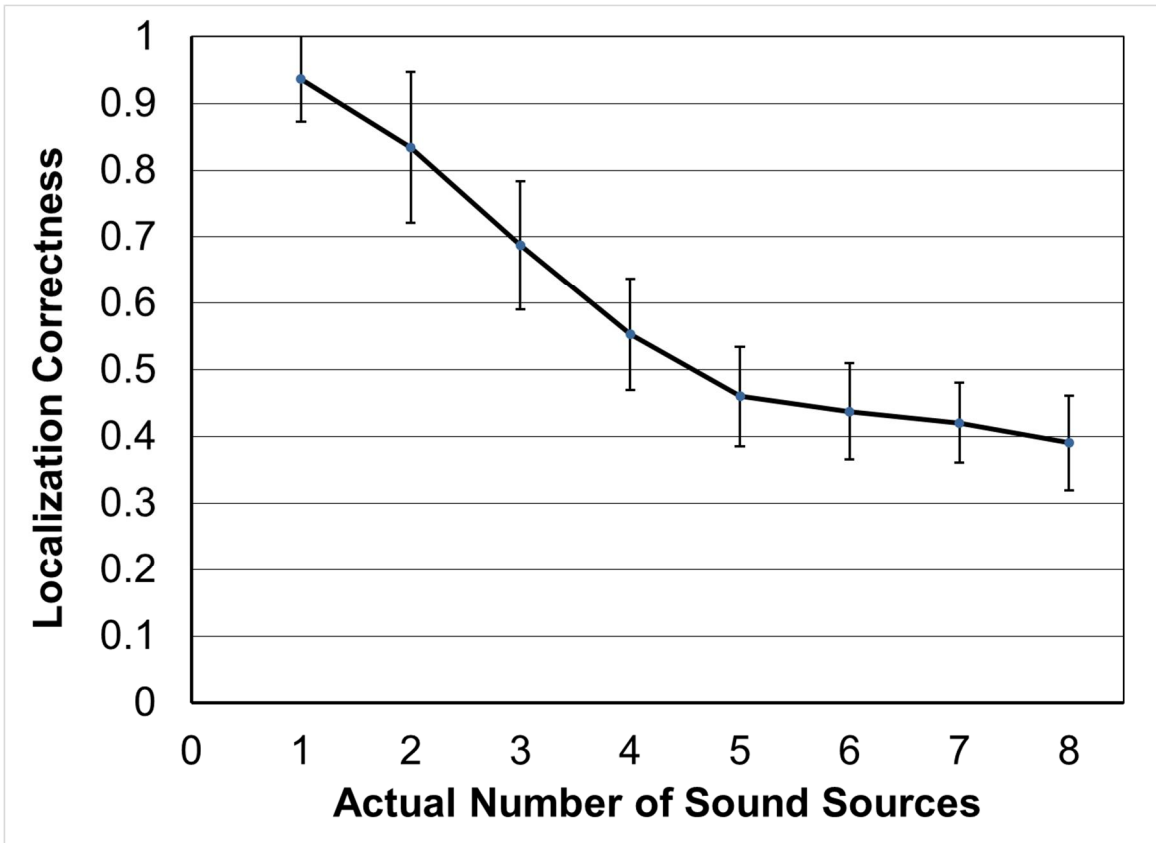


Figure 2.5. Mean and plus/minus one standard deviation (over 8 listeners) results in Experiment I showing the relationship between localization correctness and the actual number of sound sources.

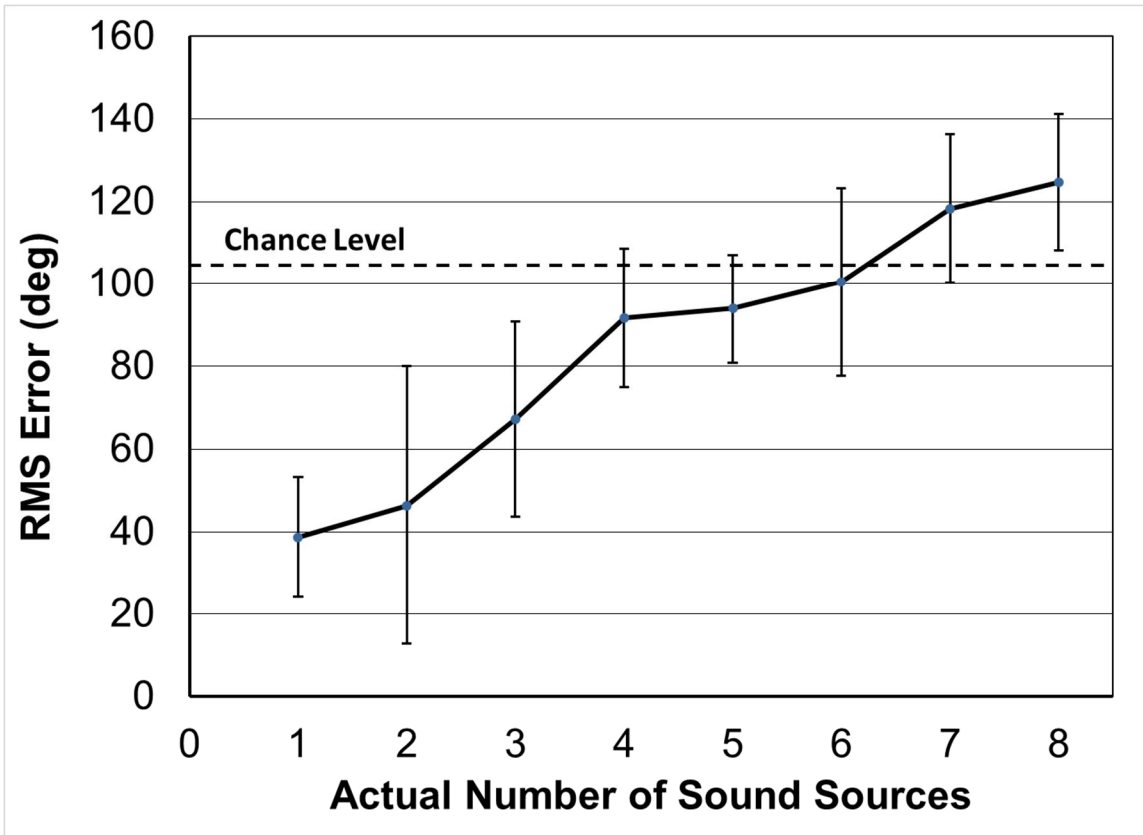


Figure 2.6. Mean and plus/minus one standard deviation (over 6 listeners) in Experiment II: added sound source, showing the relationship between localization rms error and the actual number of sound sources (chance level: 104.8°).

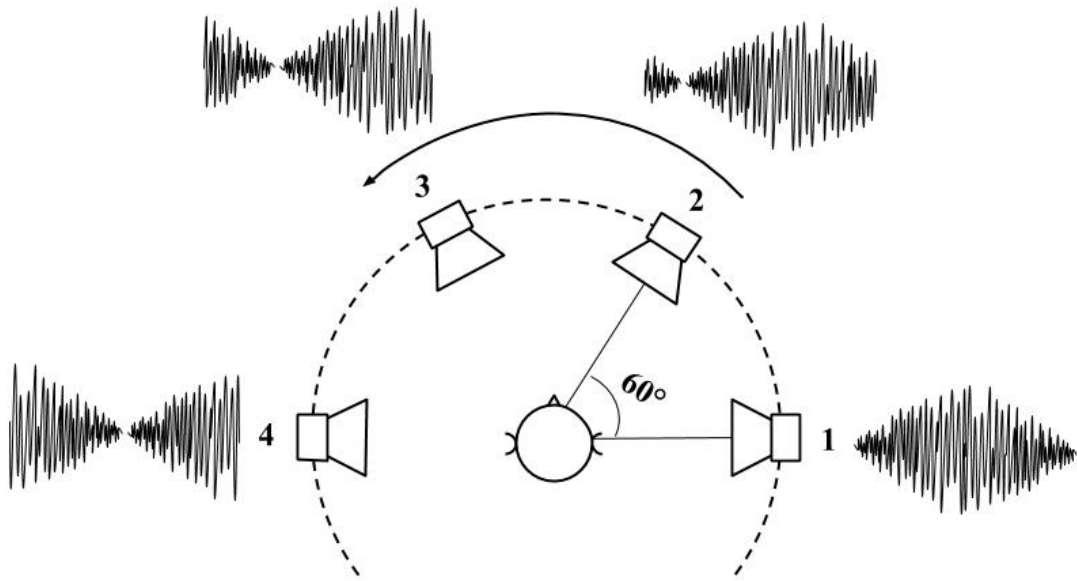


Figure 3.1. The loudspeaker setup on the horizontal plane for experiment I. The signals fed to the loudspeakers were time-delayed and amplitude modulated. The carrier signals were independently generated white noise. The envelope phase delay was 60° in between the adjacent loudspeakers.

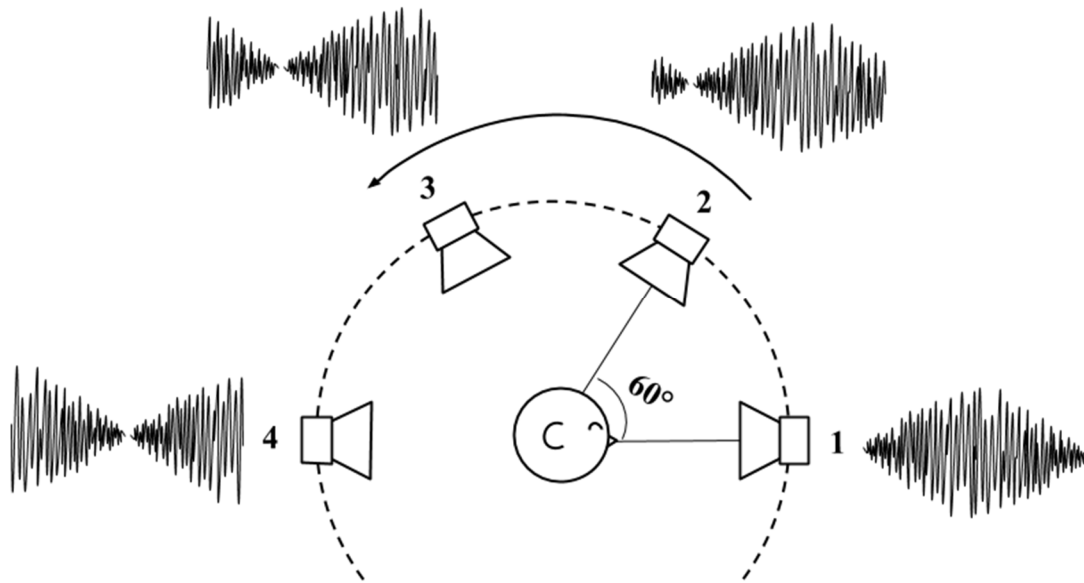


Figure 3.2. The loudspeaker setup on the mid-sagittal plane for experiment II. The signals fed to the loudspeakers were time-delayed and amplitude modulated. The carrier signals were independently generated white noise. The envelope phase delay was 60° in between the adjacent loudspeakers.

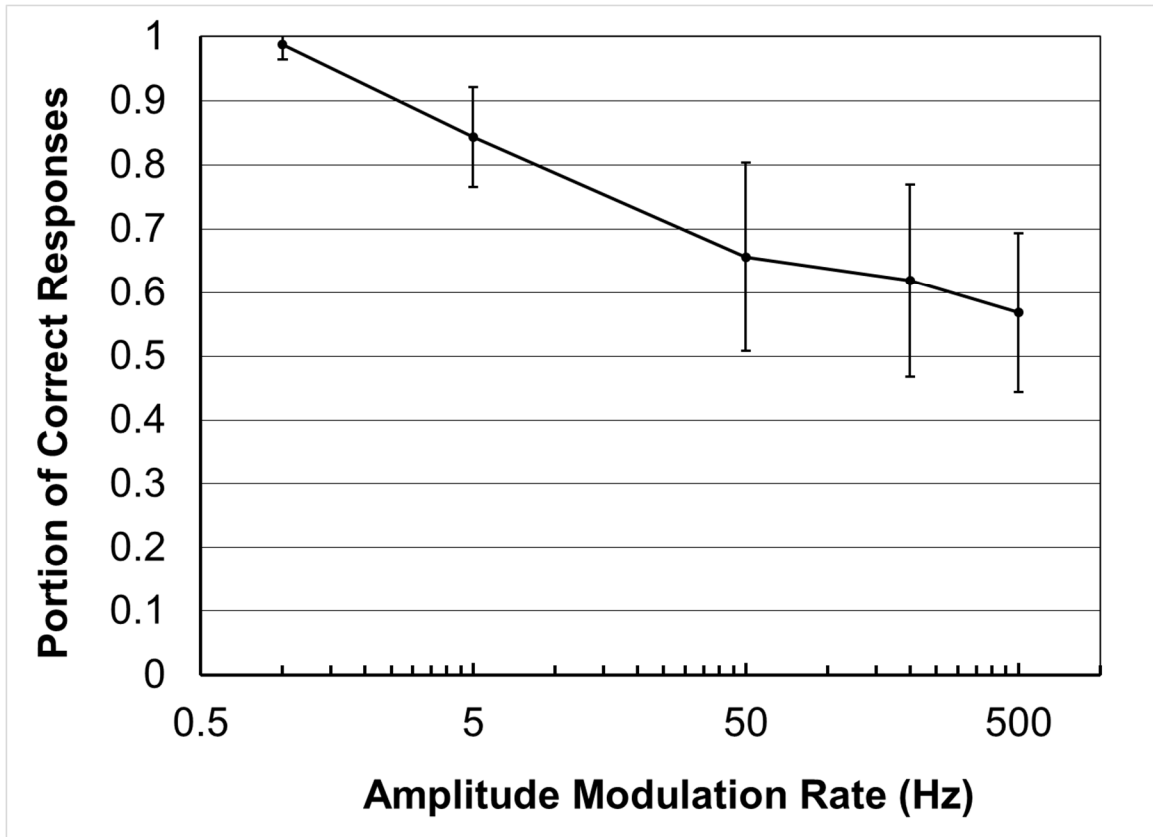


Figure 3.3. Experiment results of experiment I: auditory motion perception on the horizontal plane. The correctness in judging the direction of simulated auditory motion was calculated by dividing the number of correct responses with the total number of responses. The chance level was at 0.5. The mean values were the average of the performance of all 8 subjects. The standard error bars were also shown.

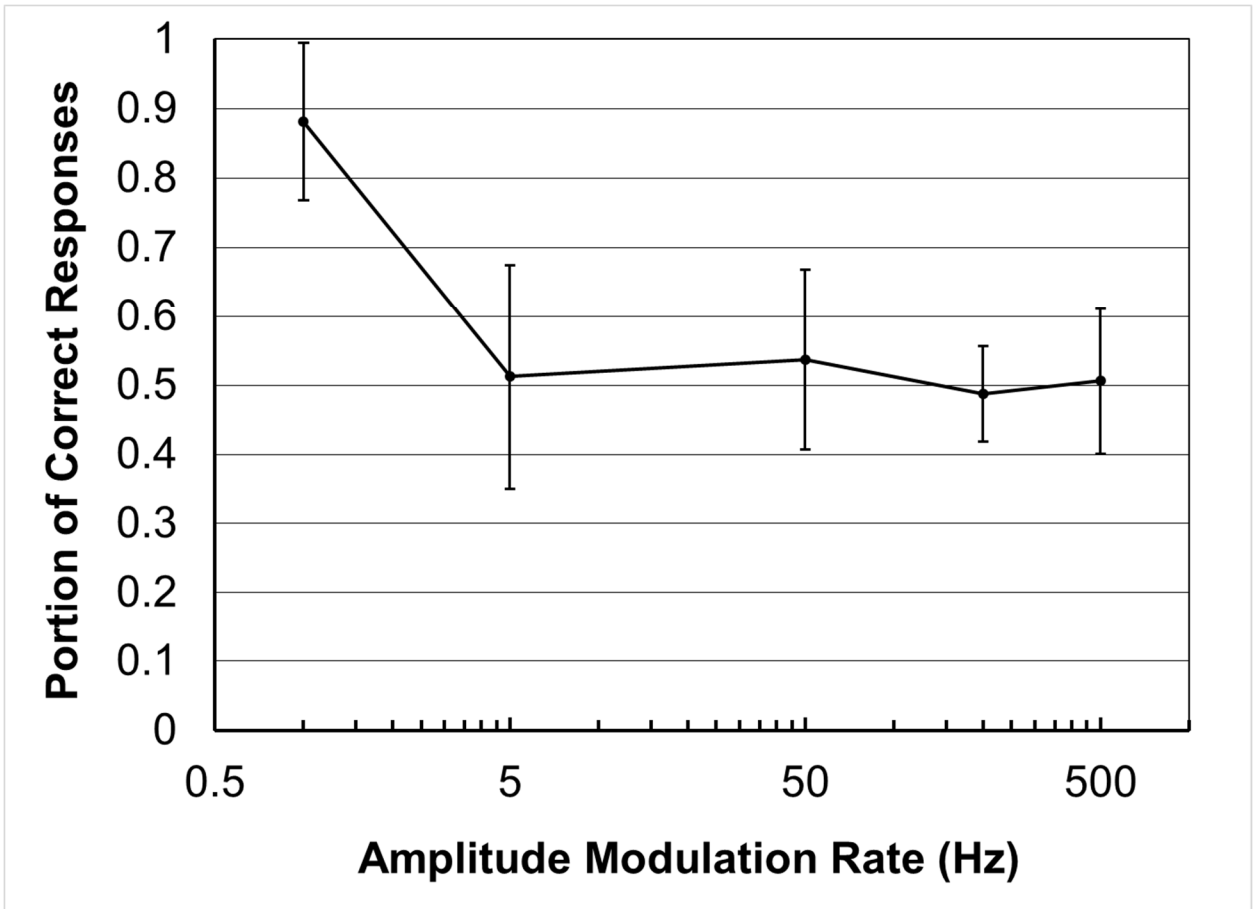


Figure 3.4. Experiment results of experiment II: auditory motion perception on the mid-sagittal plane. The correctness in judging the direction of simulated auditory motion was calculated by dividing the number of correct responses with the total number of responses. The chance level was at 0.5. The mean values were the average of the performance of all 8 subjects. The standard error bars were also shown.

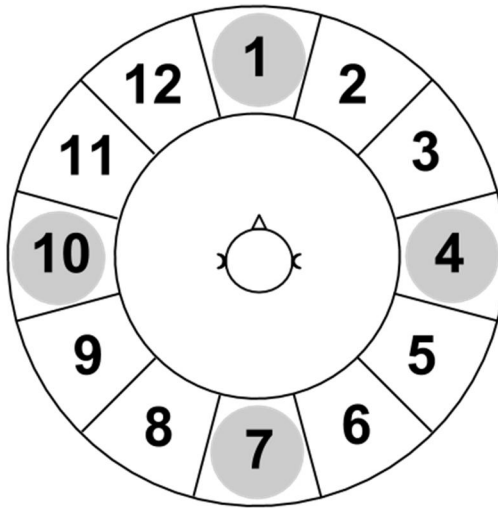


Figure 4.1. Loudspeaker setup for the experiment. Loudspeaker #1, #4, #7, and #10 were playing speech signals as acoustic landmarks. The target white noise stimuli could be from any of the rest of the loudspeakers.

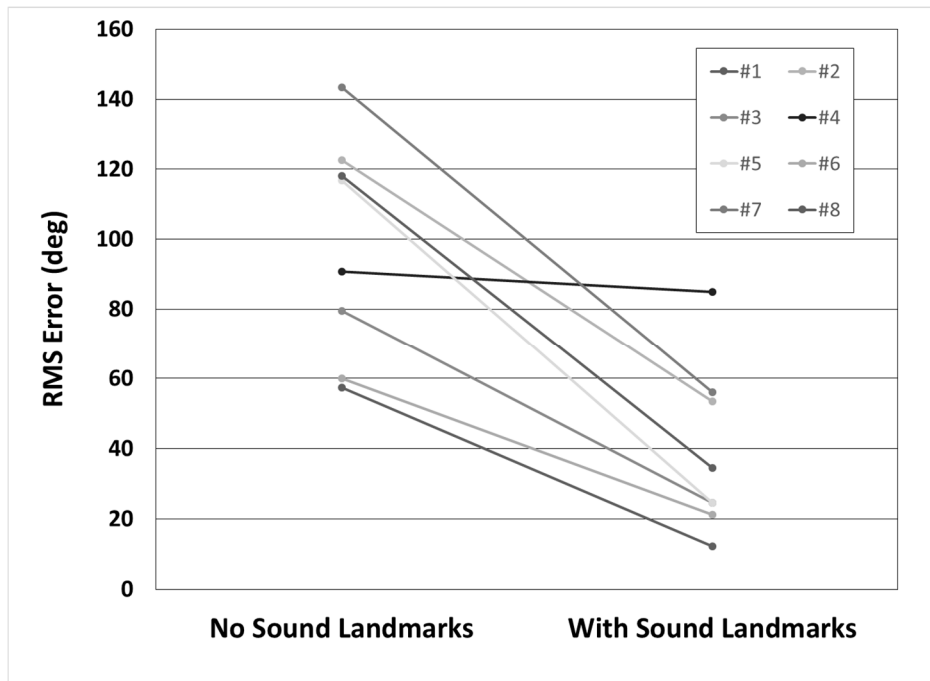


Figure 4.2. The results of localization error with and without acoustic landmarks.

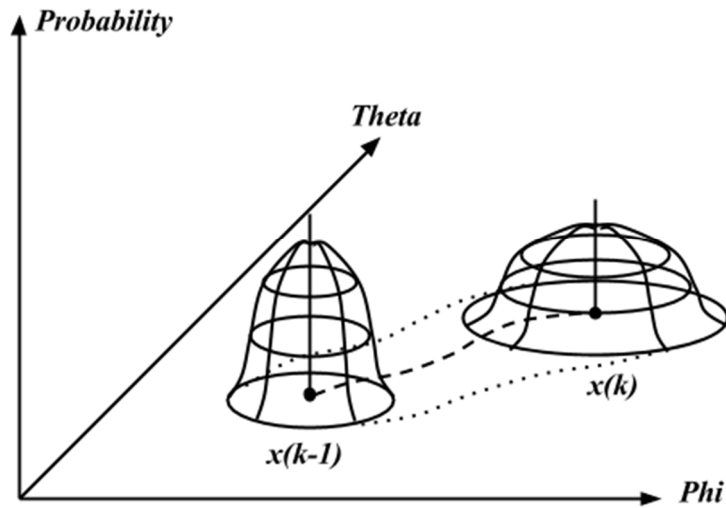


Figure 5.1. Using probability density function to describe the location of sound sources. As time changes from $k-1$ to k , the estimation of the true state also changes its distribution. Here x is location of the source defined by spatial angle Φ and Θ .

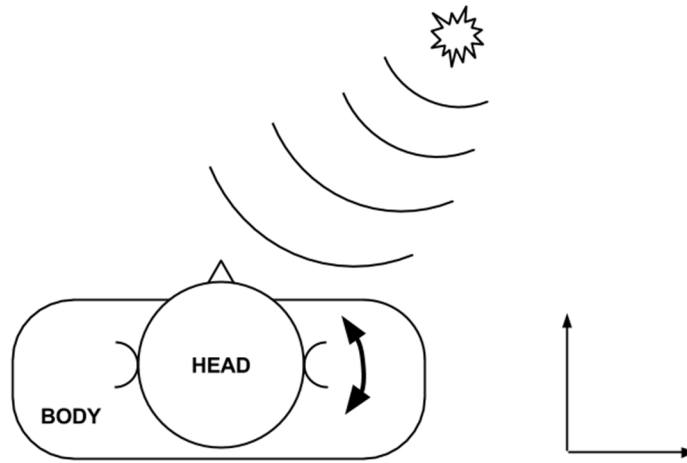


Figure 5.2. Sound source localization with head motion on the horizontal plane.

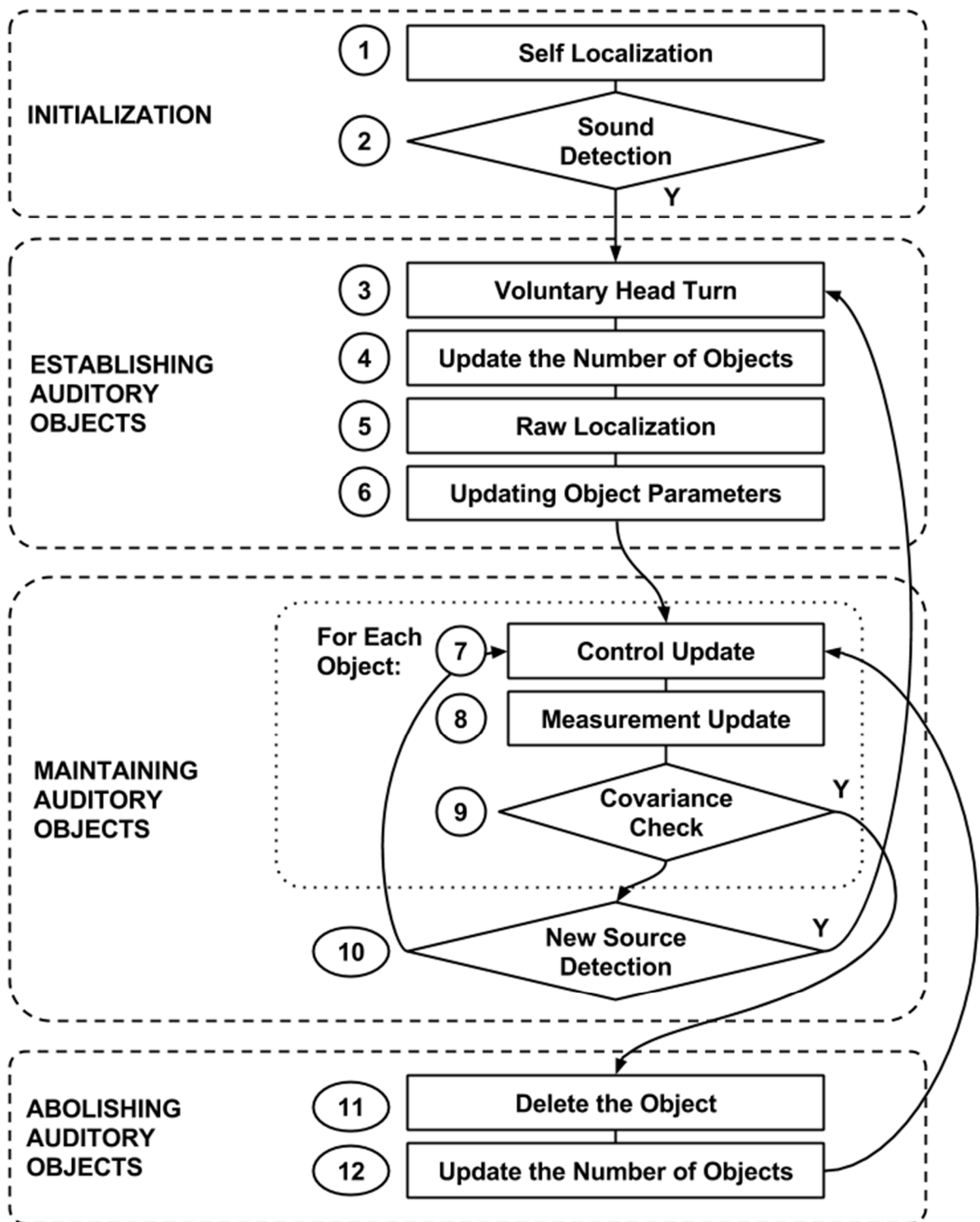


Figure 5.3. The active binaural hearing of a robot.

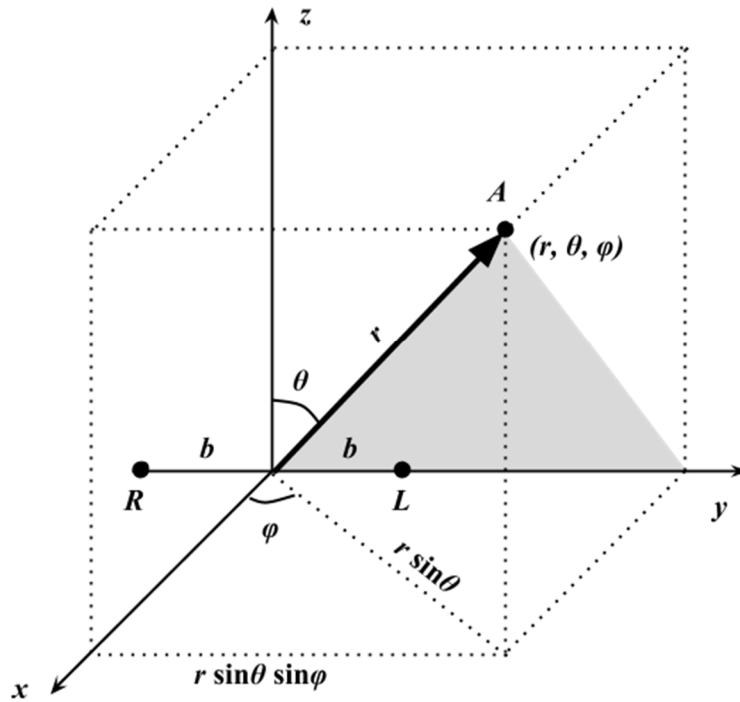


Figure 6.1. The earth-centered coordinate system used for sound source localization. The task is to localize a static point sound source A in space with a rotating two-microphone-array. The array is allowed to rotate around the z direction.

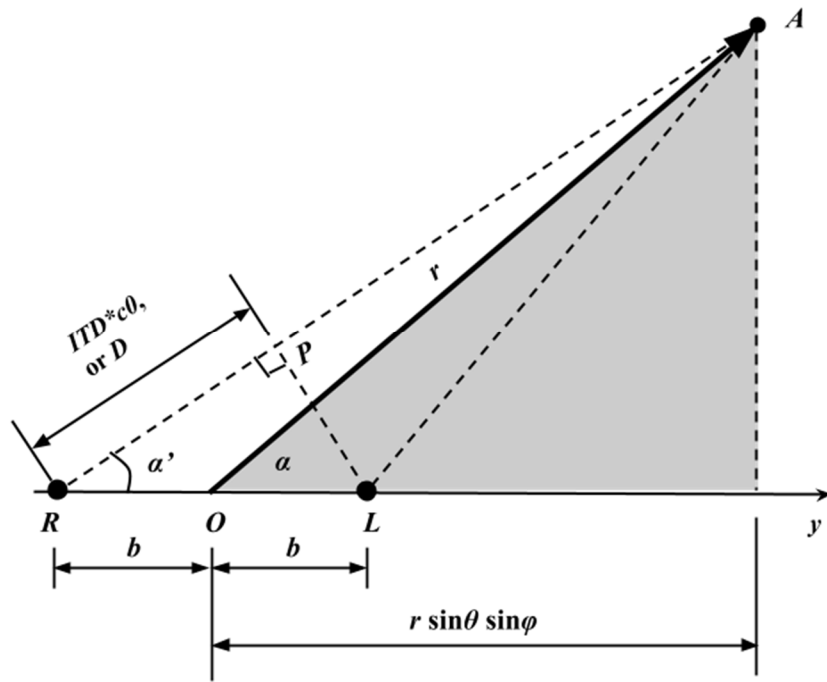


Figure 6.2. The plane in space that embodies the microphones (at R and L) and the target location (A).

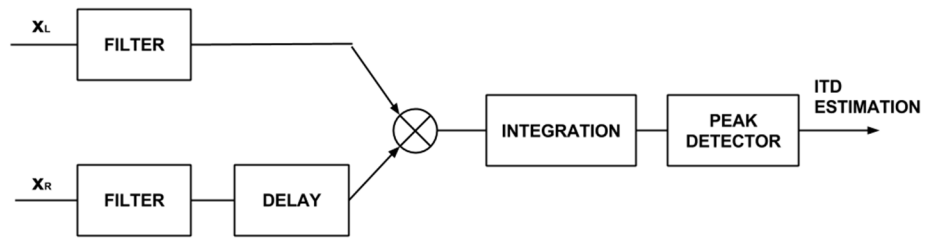


Figure 6.3. ITD estimation (adapted from Knapp & Carter, 1976).

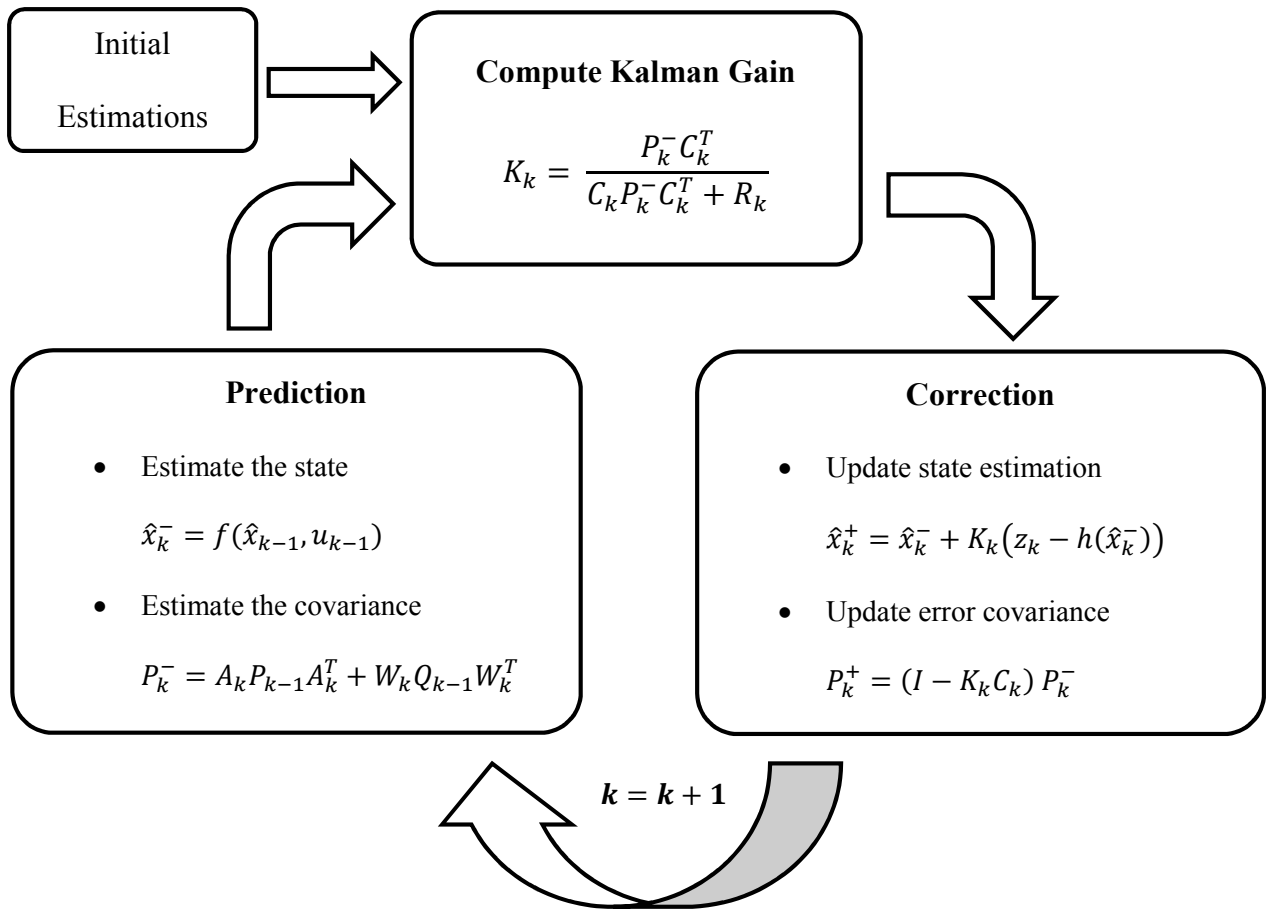


Figure 6.4. Outline of an EKF recursive filter.



Figure 6.5. The test setup of KEMAR on a rotating chair in the middle of the test room.

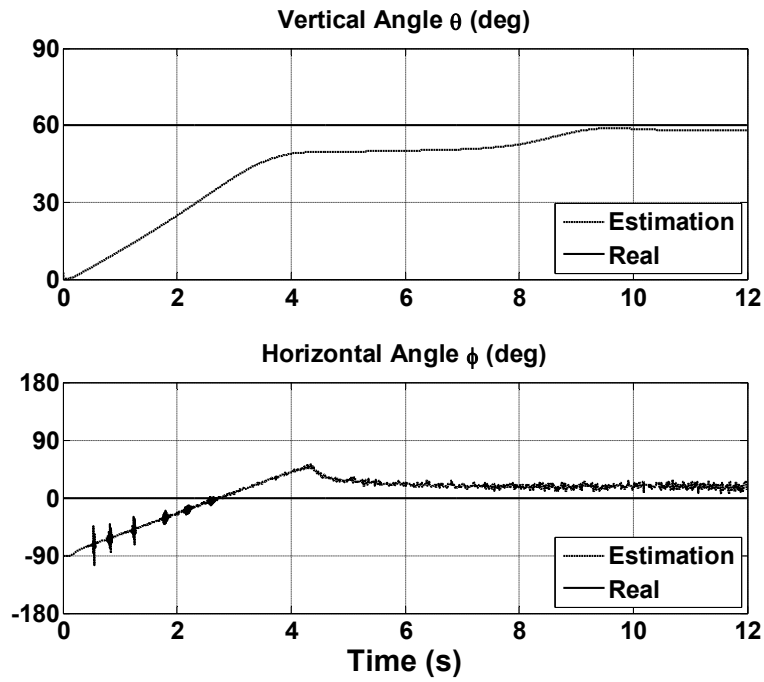


Figure 6.6. The experiment result of location calculation based on EKF algorithm. The target is at (60, 0); the result after convergence is at (60.5, 11.5).

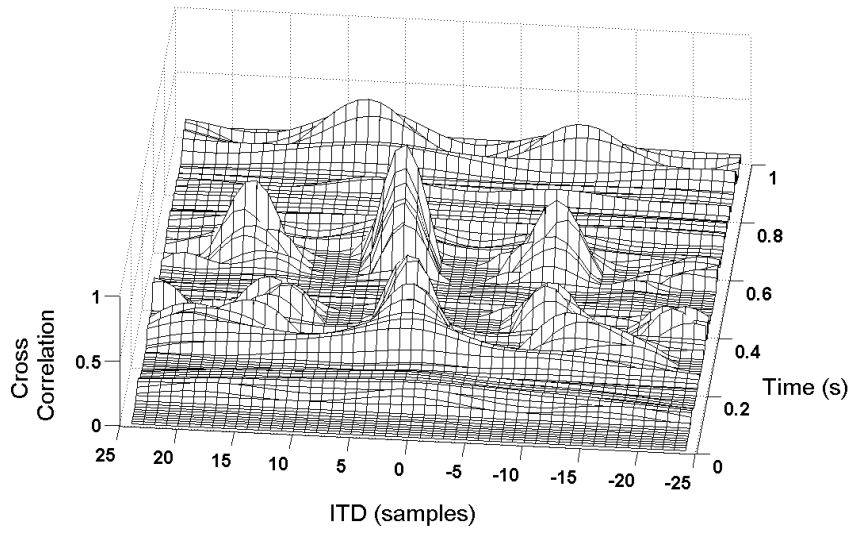


Figure 7.1. The change of ITD over time.

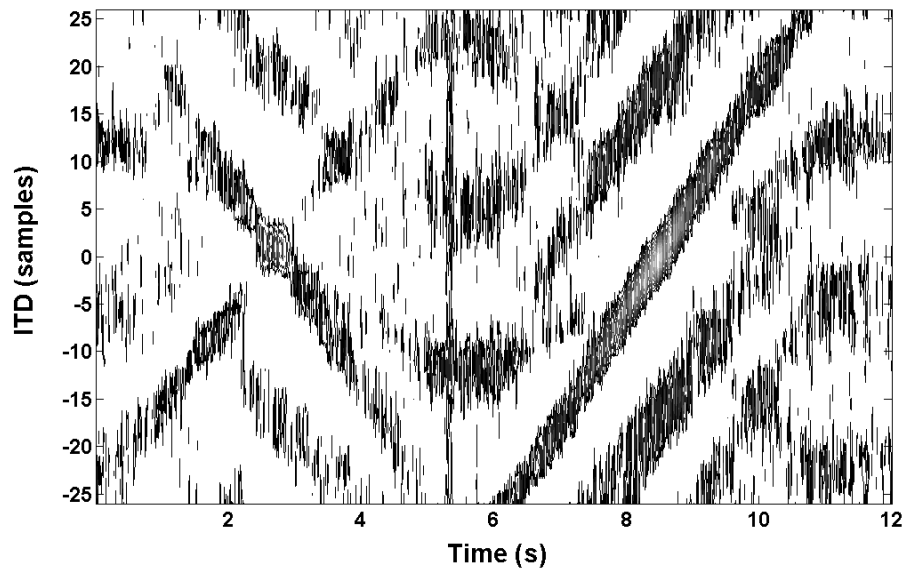


Figure 7.2. Correlogram over time depicting the changing pattern of ITD. The head of robot turned a whole circular round at constant speed. The length of each sample of time for ITD was $22.7 \mu\text{s}$ (sampling rate: 44.1 kHz).

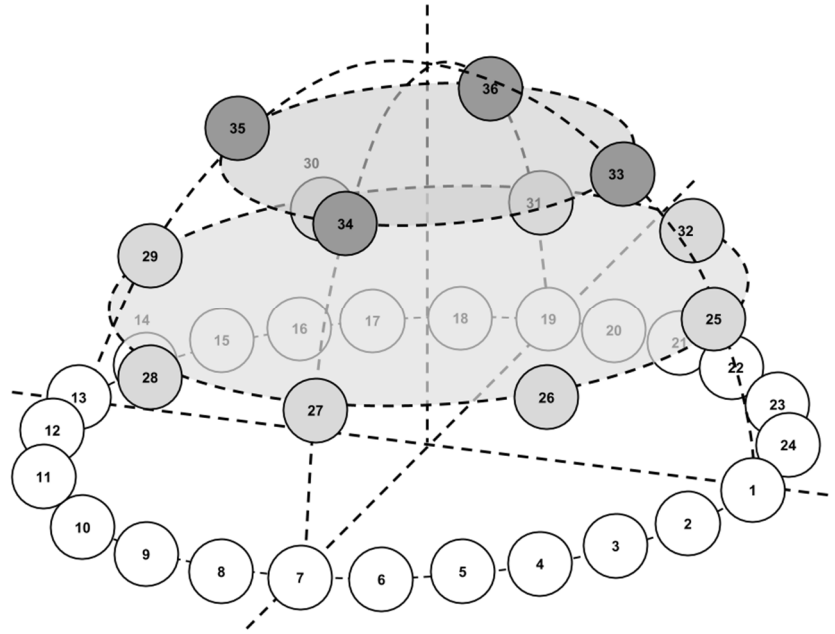


Figure 7.3. The loudspeaker setup in the modeling and experiments.

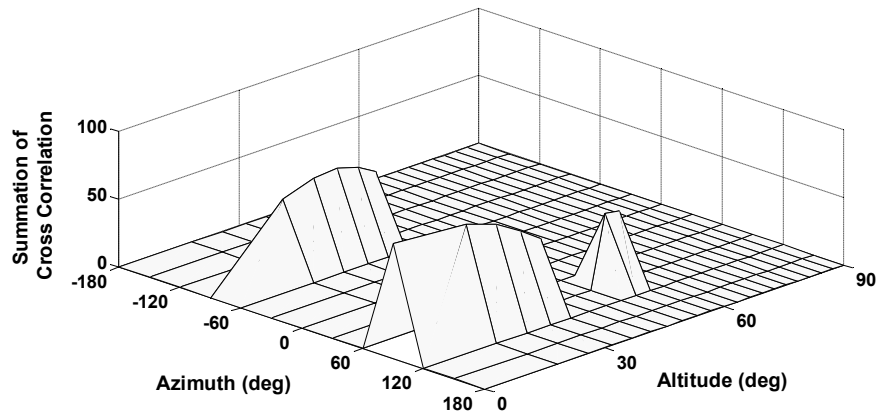


Figure 7.4. Localization results of Experiment I, in which the three independent noise sources were separated and individually localized.

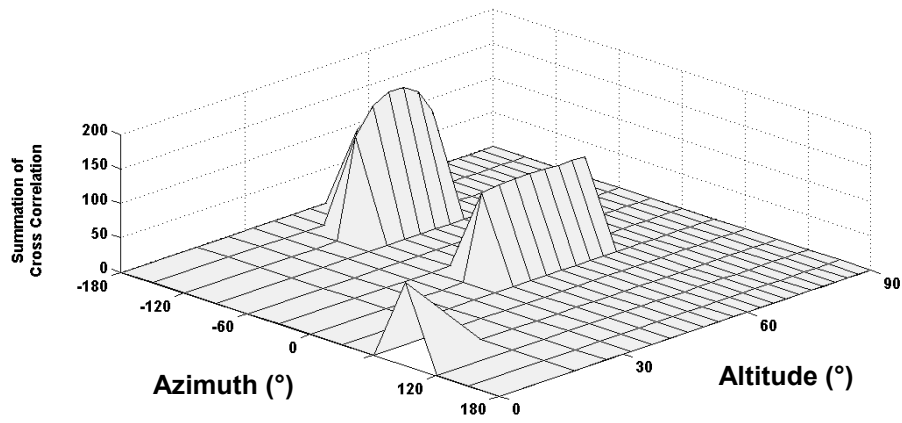


Figure 7.5. Localization results of Experiment II, in which the three sound sources were separated and individually localized. From far to near as relative to the point of observation: white noise, music and speech.

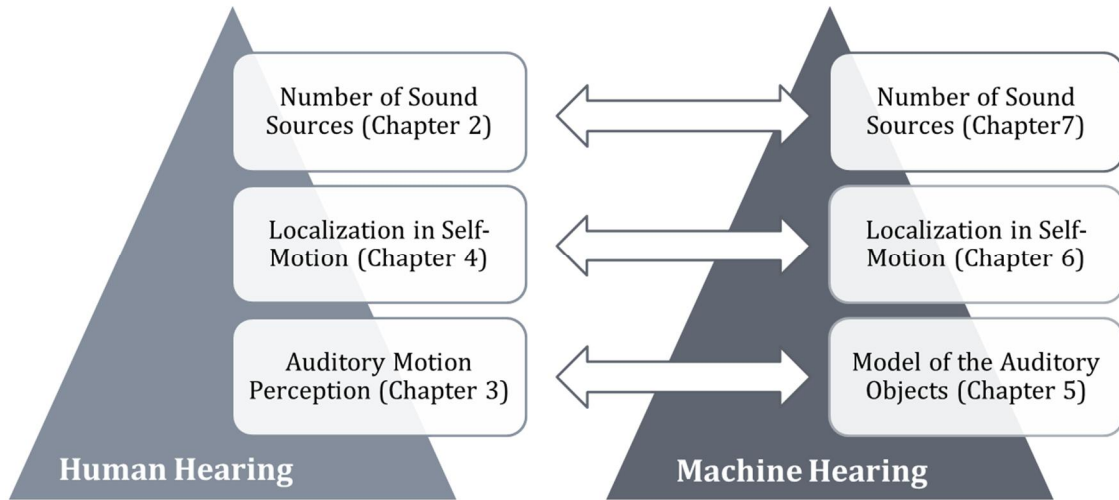



Figure 8.1. The structure of the dissertation.

APPENDIX A
INSTITUTIONAL REVIEW BOARD APPROVAL FORM

Office of Research Integrity and Assurance

To: William Yost
LATTIE F.

From: *for* Carol Johnston, Chair 
Biosci IRB

Date: 11/21/2011

Committee Action: Expedited Approval

Approval Date: 11/21/2011

Review Type: Expedited F4 F7

IRB Protocol #: 1111007097

Study Title: Relationship Between Spatial Hearing and Postural Stability

Expiration Date: 11/20/2012

The above-referenced protocol was approved following expedited review by the Institutional Review Board.

It is the Principal Investigator's responsibility to obtain review and continued approval before the expiration date. You may not continue any research activity beyond the expiration date without approval by the Institutional Review Board.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Biosci IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Biosci IRB. The new procedure is not to be initiated until the IRB approval has been given.

Please retain a copy of this letter with your approved protocol.

126126