BINAURAL SOUND SOURCE LOCALIZATION IN HUMANOID "BERT" ROBOT

BY

JOHN H. KIM

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Adviser:

Professor Stephen Levinson

ABSTRACT

This thesis explores the models and implementations of binaural sound source location and seeks to implement known and current methods within an existing robotic environment. The two primary methods of sound localization are interaural intensity difference and interaural time difference, also known as ILD and ITD, and they are discussed with their unique advantages and disadvantages and dependency upon each other to produce a meaningful location and output. Within Bert, an iCub humanoid robotic platform, the discussed methods for locating sounds are applied and experimented on to observe their accuracy and usefulness. This specific environment produces the challenge of a being a binaural system, limiting sound input capability to two inputs, along with a relatively noisy environment and limited hardware capabilities. Nonetheless, this development of a sound localization framework in Bert is done with future development in mind, as his pending hardware upgrade alone will allow for a great improvement in accuracy and precision.

ACKNOWLEDGMENTS

I'd like to thank my advisor, Professor Stephen Levinson, for the opportunity to work in his research group. His patience, guidance, and thought-provoking lab meeting discussions introduced many new areas of thought for me to consider. My fellow lab colleagues also made graduate school a pleasure to attend, for they were very kind and helpful whenever I had questions or needed help. Finally, all praise to my God Jesus Christ for the blessing of being able to partake in enjoying His creation through this thesis and graduate school experience.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation

Research and technological improvements seen in the last decade have provided massive improvements in computing power and performance. With these advances, nearly every major technology corporation boasts its own digital assistant, from Google's Google Now to Microsoft's Cortana, each created with the hope to fulfill the role of understanding and conversing in human natural language. Yet, even with these advances in technology the average user will find that voice recognition systems still possess a shockingly low recognition accuracy, especially in higher noise environments. While modern companies are attempting to solve the problem of speech recognition with neural networks and increasingly larger data sets, it remains to be seen if such methods will realistically produce precision levels comparable to that of a human. The goal of the Language Acquisition and Robotics Research Group is to approach language acquisition from a different angle. Rather than feeding a machine an exorbitant amount of data to be processed, our research group hopes to create a model that imitates a human child in his early stages of development.

The average person often has little to no trouble understanding the words spoken to him, even in a high noise environment. This is due to the fact that humans associate words with meaning and experiences, in addition to language largely becoming the means for one's mental calculations and thoughts [1]. As such, if a machine is able to experience the world similarly to how a person does, the hypothesis follows that it will be able to achieve speech recognition, amongst other things, at a similar level of performance to that of a human. Although this thesis does not explore and implement the methods of speech recognition, the acquisition of the closely related skill of sound source localization brings a machine's ability to experience the world a step closer to realization. Accordingly, implementation of a relevant language skill in our robot marks a step toward a broader language performance goal and is the motivation behind this project.

1.2 Background

The concept of locating a sound source may seem completely trivial to an individual. If asked the question of how much effort is exerted in order to approximate the position of a sound in your surroundings, many would be unable to form an answer due to its apparent unimportance. As humans, we do not think about our surroundings so much as feel them. Yet this subconscious, environmental breakdown and calculation that our brains complete in mere moments, with little to no conscious mental exertion, is a technologically staggering feat no modern machine is able to replicate.

The complexity of biological hearing systems is best understood through considering a hypothetical situation. Imagine a Saturday afternoon and you and your friends decide to go to the beach. Having arrived in the parking lot, you hear the sound of other people slamming their car trunks. One is particularly close to you, and you jump to your left as a loud slam sounds from your right. While walking down the road to the main beach area, you hear wheels approaching and slowing down behind you, so you move to the right side of the road, and a car immediately passes you on your left. As you lounge on the beach, you hear seagulls coo directly above you and the waves splashing around on the beach. Amidst all the noise of other people, you recognize your friends' voices as they walk along the shoreline. While you look up at the sky, a volleyball impacts the sand behind you, and you hear your name being called. You look to your left, and your friends wave and shout to you to throw the volleyball back. Without looking, you reach behind yourself, grab the volleyball, throw it back to them, and finally hear the impact of the ball landing in their hands.

The complexities of the aforementioned situations would appear trivial to an average person. In each situation, the specified, actionable sound is isolated, prioritized, and either discarded or acted upon. Amidst such noise and distance, and with startling precision and accuracy, your name and friends' voices are recognized, and each element of sound in your surroundings has its location in your mental map. This feat is even more impressive when thinking about our ears as the actual hardware to accomplish this task. Bregman [2] likens the auditory system to two narrow channels of water, sourcing from the side of a lake with a handkerchief in the center of each channel. As the water and waves from the lake affect these handkerchiefs, simply by observing them one is able to determine if there are any boats in the lake, if any boats are closer to the channels than others, or if wind is blowing. At face value, this problem would seem to be impossible, but this is exactly the task that the biological auditory system performs and excels at. To that end, by understanding the complexity of what the auditory system accomplishes, we are able to better appreciate the processes underneath.

1.3 Applications

While there is a specific goal and purpose behind this thesis, the utility of a computational model that is able to achieve performance similar to that of biological sound localization is manifold. This section touches on just a few uses.

- Virtual Reality - An emerging and potentially explosive market, virtual reality is no longer a figment of the imagination. Devices such as the HTC Vive give a user the ability to immerse himself into an environment with nearly complete visual and auditory sensory experiences. With the expectation that it be as realistic as possible, the program must be capable of producing sound consistent with an object's position within the digital landscape. It would not make sense for a wolf howling into the night in the background and the direct interaction with a character in front of the user to have identical perceived positions. Rather, the system must modulate sound waves in a way that the user will perceive the sound sources to be exactly where they would realistically be. Naturally, the first step to this modulation is to understand how the brain interprets sound waves to determine position and to use that knowledge to know what kind of signal processing may be necessary to achieve such an effect.
- Machine Interaction - Smartphones and digital assistants have developed to the point where they are commodities to the developed world. Companies are constantly looking for new ways to improve these products and a main way to do so is with an improved interaction experience. The human ability to focus on one individual's voice, even in high noise environments, is extremely impressive and would serve as an invaluable asset to many digital assistants. A machine's ability to specifically locate its subject and ignore all other input would certainly prevent many smartphones' incorrect activation when yelling "Okay, Google" into a crowd. Still, there has been much development in this area, as seen in products such as Amazon's Echo [3] with its ability to efficiently isolate and listen to its target through its unique microphone array design. Further development in sound source location would allow machines to improve their performance within non-ideal environments.
- Data Mining - With Big Data becoming the latest trend in computing, the ability to sort through large amounts of data for classification requires a framework. By providing a computational model that is able to interpret data in a different way, many avenues for different research or applications are created. Newly classified, sound localized data could be applied to further

research in biological models of hearing, AI development, security, or many other miscellaneous applications where the ability to process sound is a fundamental requirement.

1.4 Overview

The task of building a framework for sound localization is not a trivial problem and no doubt has its uses in the world. In order to work on a framework most resembling that of a human, this project will be built on an existing machine platform known as Bert. Bert is humanoid robot capable of many sensory motor functions, but in this application mainly Bert's auditory sensors and head movement capabilities will be utilized. Other than for identifying the specific hardware limitations and thresholds necessary for sound localization, there will be no pre-classified training data for Bert. The end goal is for Bert to be able to respond to external stimuli by identifying the stimuli's incident angle and turning his head to meet it as accurately and precisely as possible.

CHAPTER 2

METHODOLOGY

2.1 Mechanisms

Although the exact computational infrastructure used within the human brain for sound localization is not well known, what is widely known is that sound localization is essentially a sound processing problem [4]. This process is first achieved from the sound wave entering through the outer ear into the eardrum. The resulting mechanical vibrations are amplified in the middle ear, and the inner ear basilar membranes are displaced in relation to the input sound wave. This displacement is then evaluated by the human nervous system through the auditory nerves that are fired during this entire progression. With the binaural structure of the auditory system, humans are able to take two separate samples of the same sound waveform and use the two signals' properties and relationship with each other to determine a location.

When comparing the two signals, the two most prominent differences between the signals are their volume, or level, and time differences. Originally proposed by Lord Rayleigh, a British physicist [5], taking advantage of these two characteristics to find location became known as the Duplex Theory. This theory suggests that interaural time differences, or ITDs, used in tandem with interaural level differences, or ILDs, are the main means by which humans are able to locate a source of sound. These methods provide complementary and redundant information to provide a more precise location. Unless the sound emanates directly from the front or the rear of the head, the signals will naturally reach each ear at a slightly different time, with a slightly different amplitude. As a result, the ILD method works best with high frequency sounds since the head's natural shape causes what is known as the shadowing effect. When high frequency waves approach an object, the object will absorb some of that energy, causing a difference in energy arriving at each ear, and higher frequencies are most affected by this attenuation. On the other hand, ITD works best with low frequency signals since the main obstacle of accuracy, the periodic nature of signals, will be irrelevant. Generally speaking, the frequency ranges of best performance for each method exist due to the nature of sound waves and the head as an object. However, it should be noted that while there do exist specific frequency ranges that work best for each

method, that does not mean that frequencies analyzed outside of these ranges for each method cannot be useful.

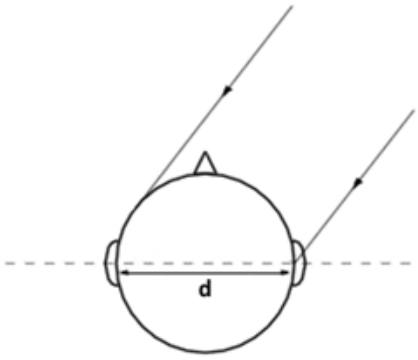

Figure 2.1 Interaural differences acting upon an ideal spherical head from a distant source [6].

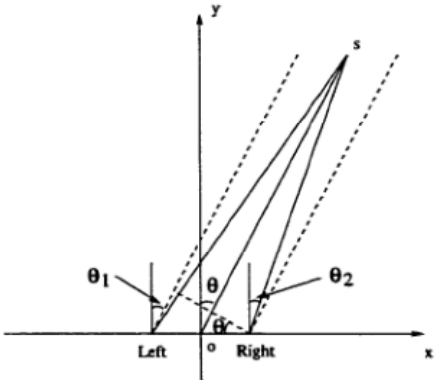

Figure 2.2. Far-field approximation of incident angle in 2D [7].

As seen in Figure 2.1, the head can be approximated as a spherical shape with two microphones on either side of it. The distance between the ears will result in different arrival times, and this difference in time τ can be calculated as

$$\tau = \frac{r_1 - r_2}{C}, \tag{2.1}$$

with C being the speed of sound, and r_1 and r_2 being distances to each ear. When the distance to the head is much larger than the diameter of the head itself, the angle to each ear can be approximated as equal to the angle to the center of the head, as illustrated in Figure 2.2. This simplifies the equation to find the azimuth θ into

$$\theta \cong \sin^{-1}\left(\frac{r_1 - r_2}{C}\right) = \sin^{-1}\left(\frac{C\tau}{d}\right), \tag{2.2}$$

with d being the distance between the two ears, and τ being the time difference between the two signals. Equation 2.2. therefore finds the incident angle to the ears from a distant source, but the ideal maximum frequency must be found in order to prevent non-ideal classification caused by the periodic nature of the waves. As mentioned earlier, signals with wavelengths smaller than that of the size of the head can cause an incorrect time difference when calculating t with the cross-correlation of the two signals. It is within this frequency threshold that the signals entering each ear will have multiple matches. The relationship between frequency f, wavelength λ, and speed of sound C is

$$f = \frac{C}{\lambda}. \tag{2.3}$$

If the speed of sound at room temperature is 340 m/s, and the diameter of Bert's head is 15 cm, then the maximum frequency f_{max} is

$$f_{max} = \frac{340}{0.15} = 2266 \text{ Hz}. \tag{2.4}$$

It is worth noting that Bert's head is smaller than the average adult male's head, with the average head being approximately 17.5 cm in diameter, which would result in a smaller frequency f_{max}.

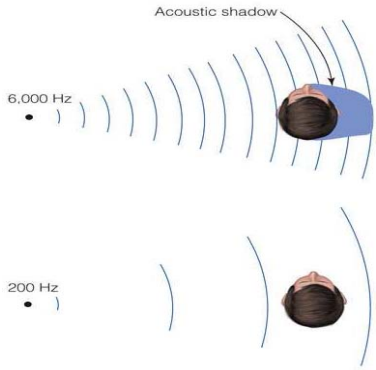

Figure 2.3 The acoustic shadowing effect for higher and lower frequencies [8].

While the upper frequency threshold is established for finding ITDs, it is now fitting to observe ILDs which conveniently are most effective for frequencies higher than those most effective for ITDs. As

seen in Figure 2.3, higher frequencies are more heavily attenuated by the head, while lower frequencies are minimally affected. This phenomenon is noted in diffraction theory, as larger wavelength, lower frequency waves are easily able to bend around the relatively small size of the head. On the other hand, smaller wavelength, higher frequency waves are not easily able to enter the ear due to their tendency to rush past it. Although it is difficult to pinpoint the exact angle of the sound source with only ILDs, it is generally the case that the larger the incident angle upon the head, the larger proportional loss in intensity.

2.2 Development Environment

As mentioned earlier, this project should culminate in a demonstration with Bert. Some background on Bert: Bert is an iCub robot, created by the Italian Institute of Technology [9] as part of the EU project RobotCub for the purpose of studying cognition through implementation of a humanoid robot. Bert possesses 52 motors for movement in the head, arms, hands, waist, and legs, and he is capable of seeing through the two cameras in his eyes, and hearing with an ear on each side of the head. He also has senses of proprioception with accelerometers and gyroscopes. For this thesis, the functionalities utilized will be Bert's hearing and head movement capabilities. But with all of Bert's capabilities, the ultimate end goal would be to combine his hardware for a complete sensorimotor, learning experience.

Development for binaural sound localization will be primarily done with Matlab. Matlab's built-in libraries and sound processing toolkit make it a highly convenient environment for preliminary implementation. Also, the scripting nature of Matlab has the advantage of faster and more efficient debugging. The results from this thesis will largely be from Matlab's, with sound inputs being recorded samples from Bert. By doing this, quantification and analysis of the results to find localization accuracy, precision, and improvement tweaking is achievable. The code that is eventually converted to C++ will be the finalized code from this Matlab step. Nonetheless, the two codes will be very similar as the C++ implementation will be as close to a one-to-one translation as possible. This will allow for smoother future development and improvements. The necessity of C++ code is due to the fact Bert is ultimately interfaced with the C++ written YARP [9], or Yet Another Robotic Platform. As iCubs' already possess their own operating system, YARP is the robot control system necessary to interface with them.

CHAPTER 3

IMPLEMENTATION

3.1 Preprocessing

The first step necessary is to record the sound samples. The sound samples were collected using YARP's audio interface with Bert. Consequently, the sound samples are what Bert would directly hear and operate upon once the code is moved directly onto him. A total of 11 samples were collected at varying frequencies and angles at a sample rate of 48000 Hz. Table 3.1 shows the specifications of the samples. Samples of realistic sounds in the form of speaking and keys jangling were also recorded at each of the angles indicated.

Table 3.1 Frequency for each sound sample recorded at angle theta.

Frequency (Hz)	Theta(°)
500	-90
750	-60
1000	-45
1500	-30
2000	0
2500	30
3000	45
3500	60
4000	90
4500	
5000	

By sweeping a range of frequencies for each angle, the strengths and weaknesses of each localization method are apparent. Also, by having recordings at various angles, the output result can be compared with the expected angle of the sound source.

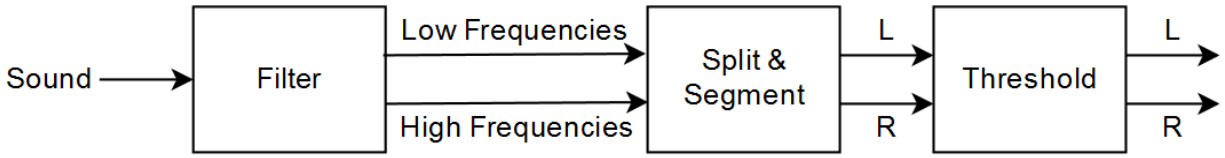

Figure 3.1 Flowchart for preprocessing directives.

After having recorded all the samples, the sound preprocessing commences, as seen in Figure 3.1. The goal of this section is to modify the sound samples into different data structures, so they are easier to work with. As discussed earlier, the methods of ILD and ITD both have their strengths and weaknesses within a frequency range. Knowing this, it becomes ideal to send the best frequency range through each method to acquire the best accuracy. Since f_{max} was calculated to be 2266 Hz, this was selected to be the approximate point at which the sound samples would be separated via filtering.

A Butterworth filter was used in this case to maintain the integrity of the signal in the region of interest. With a larger filter order, the transition boundary cutoff is quickly achieved, so the signal outside the region is quickly diminished. As seen in Figure 3.2, the cutoff frequency for the Butterworth filter is set at approximately the same value as f_{max}, creating two sets of waveforms for each full frequency range waveform. By setting the boundary here, the waveform below f_{max} is used for ITD and above f_{max} is used for ILD.

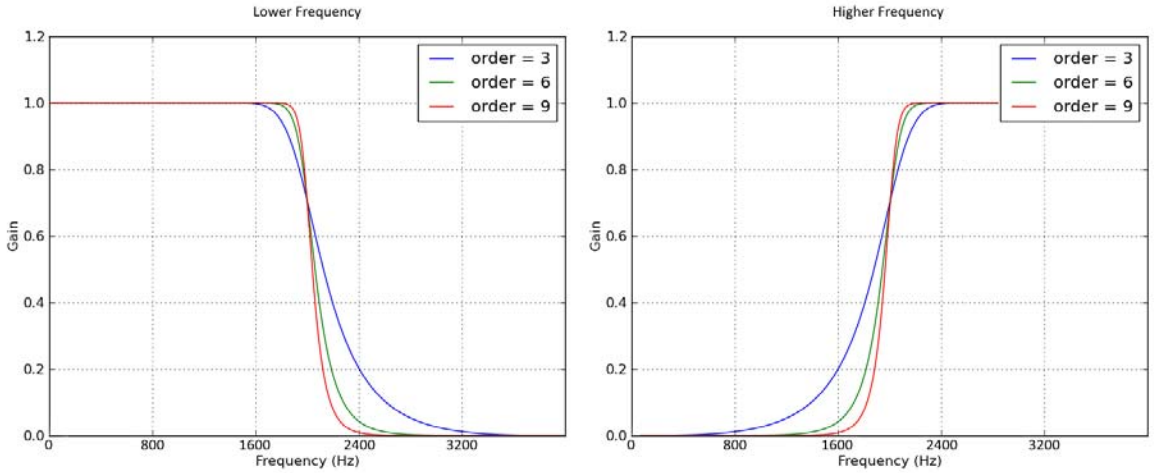

Figure 3.2 Butterworth low-pass and high-pass filters for frequency isolation.

The last step in preprocessing is to split and segment the waveform. Sound samples are normally recorded in stereo, meaning there are separate channels for the left and right ear within one

file. In order to analyze the left and right channels separately, it is convenient to split and save them to different variables. Also, being recorded in stereo means that the channels will have identical time lengths and will be perfectly synchronized. Synchronization is absolutely vital in calculating the interaural differences as the sample numbers being cross correlated must be identical, or the entire process is rendered moot.

Segmenting is the procedure in which the entire sample is divided into equal sized parts. The advantage of this step is to create equally sized blocks of sound samples to perform on, to ensure the same number of samples are correlated each time. Although the number of samples per block is somewhat arbitrary, a short spoken word such as "hey" is around 500 ms long. Due to the later explained buffering method, the size of each block will be around 50 ms long. This translates to a number of

$$\begin{aligned} N &= F_s * t \\ &= 48000 * 0.05 \\ &= 2400 \; samples \; per \; block. \end{aligned} \tag{3.1}$$

As the idea of zero sound is an unrealistic assumption, the microphones are always receiving some sort of vibration and energy, no matter how small. These blocks should ultimately not be considered and removed, for there is no sound source to locate. To remove these blocks, first they must be changed to decibels and averaged. The sound files are recorded as magnitudes, so to convert them to decibels one uses the equation

$$X_{db} = 20 \log_{10} X. \tag{3.2}$$

It was found that blocks of no sound input for higher frequencies had an average value of -21 db, and the lower frequencies had an average value of -10 db. Therefore, these values were set as the threshold of minimum decibels for a sound block to be analyzed. Any block with an average value less than that of the threshold is thrown out. The remaining blocks are then likely to contain sound samples with a potential sound source, so they are passed onto the next section for ILD calculations.

3.2 Interaural Level Difference

Out of the two methods of localization, ILD is considerably simpler to understand and implement. The basic concept is that since human ears are on opposite sides of the head, the volume of a sound source is dependent on its proximity to either ear. The more the source originates from the right side of the head, the more energy will enter the right ear as opposed to the left. The same is true if the source originates from the left. In the case that the source is directly in front, or at 0°, of the head, both ears will receive equal intensity. This relationship between intensity and sound source position can be seen in Figure 3.3, which illustrates the angular sweep of a sound source.

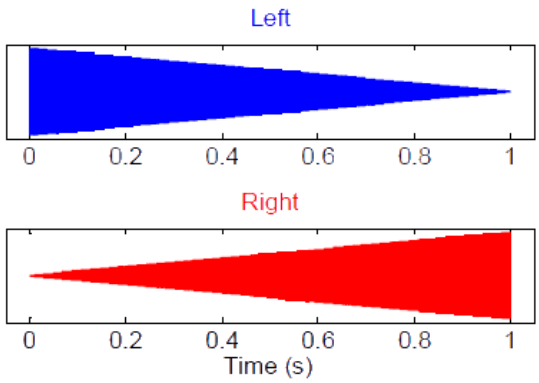

Figure 3.3 Interaural intensity sweeping from left to right ear.

```
1: procedure ILD(leftDb, rightDb)
2:     if leftDb > rightDb then
3:     |  direction is left;
4:     else if leftDb < rightDb then
5:     |  direction is right;
6:     else dbs within 2db then
7:     |_ direction is straight;
8:     end if
9: end procedure
```

Figure 3.4 Matlab ILD implementation pseudocode.

Figure 3.4 steps through the procedure of finding the ILD for each block of sound. In accordance with the theory just discussed, if the left ear receives more power than the right ear, it is more probable that the source is on the left, and vice versa. One thing noticed is that it is highly improbable that the power levels entering each ear will match exactly with one another for the 0°, straight ahead, verdict. After some observation, when a source was placed directly in front of Bert, the difference in decibels

would fluctuate within a range of one to two decibels. As a result, a difference of two decibels was decided upon to be the threshold to serve as the straight ahead verdict. This analysis is done for every block so that each one has an ILD verdict associated with it for future review.

3.3 Interaural Time Difference

Although ILD does provide information on the general location of a sound, ITD is more capable in that it calculates a more exact position. It accomplishes this by obtaining two samples of the same waveform through the separately received signals in either ear and comparing them. This comparison is achieved through the mathematical cross-correlation function. The sampling nature of digital sound input requires the discrete method of cross-correlation of

$$(f \star g)[n] = \sum_{m=-\infty}^{\infty} f^*[m]\, g[m+n], \tag{3.3}$$

where f^* represents the complex conjugate of f. The essence of the cross-correlation measure is to shift the channels on each other and to identify the point at which the two signals produce the largest overlap, or the largest correlation. As mentioned earlier, the synchronization of these signals is paramount to a successful ITD procedure. Due to the high speed of sound, if the difference in time is not measured exactly by the system, a completely incorrect angle will result. For example, if the left and right channels are off by even 50 ms, that would cause a skew of 2400 samples, which translates to a completely off angle of incidence. However, this problem is only worth mentioning in passing and is irrelevant in this experiment as Bert's sound recording is reliable in this area.

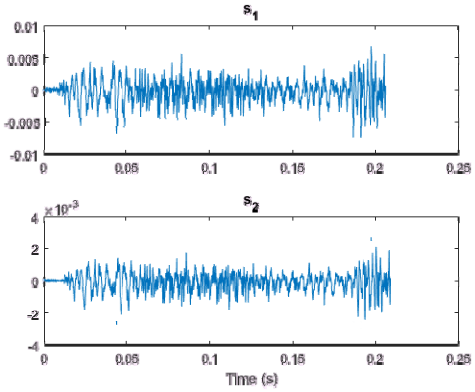

Figure 3.5 Example shifted waveform.

Figure 3.5 can be seen as an example of a binaural input of the same waveform. The signal is largely the same, but s_2 is shifted slightly to the right, meaning it was received at a slightly later time. If the time shift is calculated, the incident angle can be approximated using that value. However, this is the ideal case in which the waveforms do appear identically, and it will not be the case in a realistic scenario. Fortunately, cross-correlation does not require the channels to be identical, as it will still find the point at which the waveforms most resemble each other.

```
1:  procedure ITD(lSound, rSound)
2:      find cross correlation of (lSound, rSound);
3:      find delay of cross correlation max;
4:      if delay < 0 then
5:      |   direction is left;
6:      else if delay > 0 then
7:      |   direction is right;
8:      else delay within 2 then
9:      |_  direction is straight;
10:     end if
11: end procedure
```

Figure 3.6 Matlab ITD implementation pseudocode.

Figure 3.6 outlines the steps that are taken to find the ITD of a given signal. To start, the data that is passed into the ITD section is two 2400 sample size blocks, one each for the left and right channel. Each block is then cross correlated with its counterpart to determine the number of sample shifts necessary. The number that is selected will be at the point where the cross-correlation function reaches its maximum. This infers that shifting the signals in relation to each other by that number of samples will result in a best match. In this case, the left channel is cross correlated with the right channel, or $L \star R$, meaning that the delays are calculated in reference to the left channel at its base. If the delay is negative X, the right sample must be translated X samples earlier to match the left. On the other hand, if the delay is positive X, the right sample must be translated X samples later to match the left. The relationship between this delay and the direction of sound is apparent. A negative delay indicates that the sound had reached the left ear first, and the right ear later, and vice versa for a positive delay. Accordingly, a negative delay means the sound source originates from the left side of the head and a positive delay on the right.

At this point, the directionality associated with delay is explained. However, ITD is also capable of finding an approximate angle of incidence. Equation 2.2 explored the calculation for this and the

delay provides the missing variable needed. *C* and *d* are the constants of the speed of sound and the diameter of Bert's head. By knowing τ, the angle can be calculated. Since τ is the interaural time difference, the delay can calculate this value using the sample rate F_s, resulting in

$$\tau = \frac{delay}{Fs}. \tag{3.4}$$

Consequently, finding the ITD τ through cross-correlation provides the final variable needed to find the incident angle θ using Equation 2.2. After finding the angle for each block, it is stored for later analysis and referencing with ILD.

3.4 Buffering

The advantage of using both ILD and ITD to locate sound is the ability to use the two methods in tandem with one another to throw out inconsistent classifications and refine existing ones. The goal of this section is to implement a way to most effectively use these methods together to produce a meaningful result. This is necessary because a decision-making tree must take place. Due to the continuous input of sound, it would not make sense to act upon a classification for every single block of sound. Section 3.1 discussed the length of a single block of sound 50 ms, and Bert would not realistically be able to turn his head for every impulse of that length. The length of 50 ms was decided as it was one tenth the value of 500 ms, the length of a natural utterance. Therefore, the buffer will provide a direction to be acted upon when the buffer is filled with enough similar classifications.

```
1: procedure BUFFER(ILD, ITD)
2:     if ILD != ITD then
3:         disregard and continue;
4:     else
5:         store ITD in buffer;
6:         if buffer is full then
7:             angle is most frequent ITD in buffer;
8:         end if
9:     end if
10: end procedure
```

Figure 3.7 Matlab buffer implementation pseudocode.

Figure 3.7 provides the flow of the decision-making process. At this point, the ILD and ITD processes have a direction associated with each block of sound and will be fed chronologically into this procedure. To begin, if the ITD and ILD directions do not match for a specific sound block, then that block will be disregarded and not added to the buffer. By ensuring the consistency between separate verdicts, the probability that the final verdict is actually correct is increased. The next step is to begin accumulating ITDs in the buffer. The reason why ILD is not also accumulated is because once it is used to cross-reference the ITD decision, it does not hold any extra information that is not now already included within the ITD direction. The system will continually accumulate ITDs within the buffer until it is filled with ten values. The buffer will continue to accumulate values in a first in, first out fashion, but the most frequently occurring ITD within the buffer will be the direction that is decided and acted upon. In the Matlab environment, this direction will be noted to check consistency with the recorded sample, but in Bert's case he will turn his head temporarily to face that angle. The buffer will also be cleared if there is a long enough time of silence, in which no data will make it through the preprocessing threshold step. Figure 3.8 visually represents the overall flow of the buffering process.

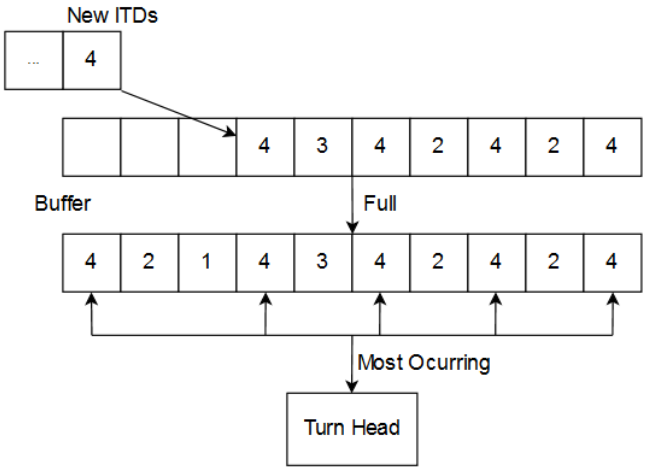

Figure 3.8 ITD buffering flow.

3.5 Limitations

Because ours is a project based on the effectiveness of hardware, there are no doubt many limitations and obstacles that can seriously affect performance. Unfortunately, a number of these factors will negatively impact the results of this project. All of these problems are caused by the

hardware that Bert uses and are unavoidable until pending upgrades are completed. This section discusses these problems and the impact that they will have on the result.

3.5.1 Sampling Rate

Discussed earlier, the sampling rate or F_s is the number of samples the microphone captures per second. The current sampling rate is 48000 Hz, and in normal cases this would not pose any sort of problem for the user. The sampling rate is normally significant in that it must be high enough to avoid any aliasing of the signal. However, in the current application, the sampling rate is necessary in that it defines the minimum increment at which the ITD may classify its angles [10]. A new formula for the angle of incidence can be created with a combination of Equations 2.2 and 3.4, resulting in

$$\theta = \sin^{-1}\left(\frac{C * delay}{d * Fs}\right) . \tag{3.5}$$

Since delay is the number of samples to be shifted to match the cross-correlation of the left and right channels, its absolute minimum increment is the time length of one sample, or 1/48000 seconds. The problem this causes is due to the nature of the trigonometric function $\sin^{-1}$ and the limit it imposes. The argument within $\sin^{-1}$ must maintain a value between -1 and 1. Knowing this, the limits of realistic delay will be added to Equation 3.5 to produce

$$\begin{gathered} -1 < \frac{C * delay}{d * Fs} < 1 \\ \frac{-d * F_s}{C} < delay < \frac{d * F_s}{C}. \end{gathered} \tag{3.6}$$

What the limits translate to is the sound source being farthest left or right of the head. As seen from Equation 3.6, the limits of the delay depend on F_s. The lower the sampling frequency, the more limited the delay is in its discrete values and the less able it is to indicate the smaller angular changes. For example, when solved for this project, the delay must be smaller than 22 samples. This means only 22 samples exist to represent the range of angles from 0° to 90°, which gives a rough average angular granularity of 4°/sample. If the sampling rate was double the amount, then the granularity would be 2°/sample. This is also exacerbated by the fact that inverse sin is not a linear function. In other words, the granularity may start at 1°/sample, but as the delay approaches its limit, it may increase to 5°/sample, causing far more room for error with higher angle of incidence of the sound source. This

problem can potentially be mitigated through a higher sample rate, but it is an interesting obstacle specific to digital sound localization. The biological auditory system faces no problem of this sort, as it does not capture sound in a discrete manner; rather, it is analog and continuous.

3.5.2 Noise

Noise is a classic issue in any sound processing problem, and there are issues with Bert's design that cause major problems. Bert's CPU is located in his head, directly adjacent to his microphones. His CPU is cooled with a loud fan that continuously feeds directly into his sound inputs causing large noise artifacts that heavily interfere with sound clarity. Furthermore, the server unit Bert is connected to also generates large amounts of ambient noise, which are somewhat picked up by the microphones.

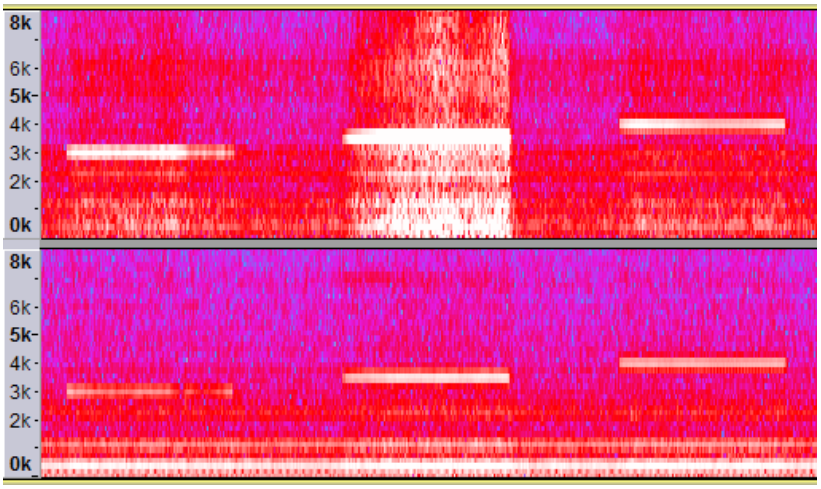

Figure 3.9 Dual channel spectrogram analysis of sound input.

An example of the noise being faced is seen in Figure 3.9. The desired signal being fed here is seen in the 3000 Hz through 4000 Hz band, and these signals are being read properly. However, within the 0 Hz to 1000 Hz range, a large amount of noise is constantly picked up by the right (bottom figure) channel, caused by the white noise that the cooling fan emits. In this specific example, since the desired sound is in the higher frequencies, the noise is not as relevant a problem. The Butterworth filter discussed in Section 3.1 separates the frequency ranges, so the noise will not interfere with the higher values. Still, this remains a major issue for all lower frequency sources like the human voice with natural frequencies well within the 0 Hz to 1000 Hz range. Future upgrades have the potential to completely mitigate this problem, as a relocation of the CPU will render this issue nonexistent.

3.5.3 Gain

Gain can be understood as the sensitivity of the receiving sound unit that affects the volume of the recorded sample. In most cases, gain is an element only useful to musicians or sound mixers to affect the volume of a recording. Human ears operate at a fixed gain level, which may change according to age or external factors, but can be generally assumed to be consistent. This is seen by the fact that given the same sound source at its same circumstance, a person will perceive it at the same volume he has before. Unfortunately, while gain can be usually assumed to be irrelevant, it is not so with Bert.

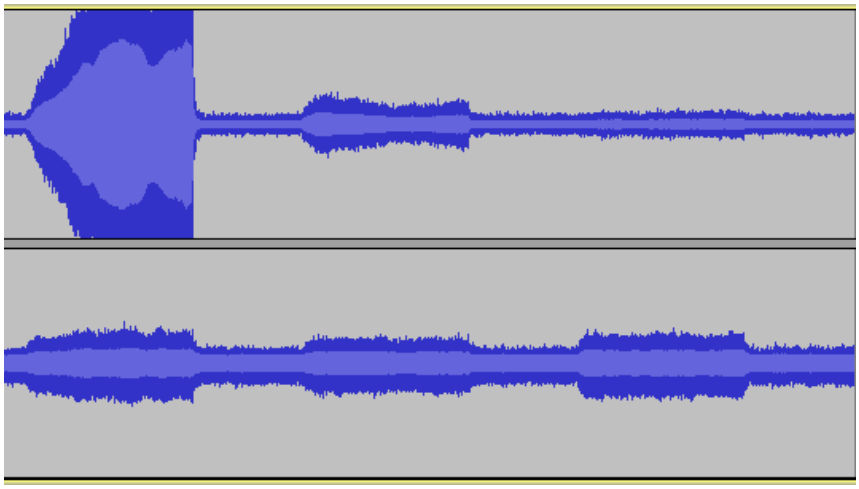

Figure 3.10 Dual channel waveform analysis of sound input.

Bert appears to possess an automatic gain system that changes the recorded volume in an unpredictable manner. Figure 3.10 shows a recording of a sound source of constant volume and distance from Bert. However, the recording clearly shows that the left channel wildly fluctuates in its gain. The first impulse increases in its amplitude, the second is recorded much more weakly, and more so with the third. The right channel appears to be fairly consistent in this test case, but other test cases have shown it to suffer from the same problem as the left. The result of this problem is that the ILD method could be rendered ineffective and useless. ILD depends on the fact that a sound source will be input at a lower intensity at a larger distance. This problem causes that dependency to no longer exist, as the gain can cause a source that is farther from the left ear to actually be recorded at a larger amplitude.

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Performance

The recorded sound samples were sent through the localization system, and each had sets of locations output. Before looking into the results, it is important to note the general frequency range of each sound sample to understand the strengths and weakness of the overall system for each range. Section 3.1.1 explained that there are 11 angles recorded for each sound type. The three sound types were a frequency sweep from 500 to 5000 Hz, a voice utterance of "Hey Bert, look over here," and jangling keys.

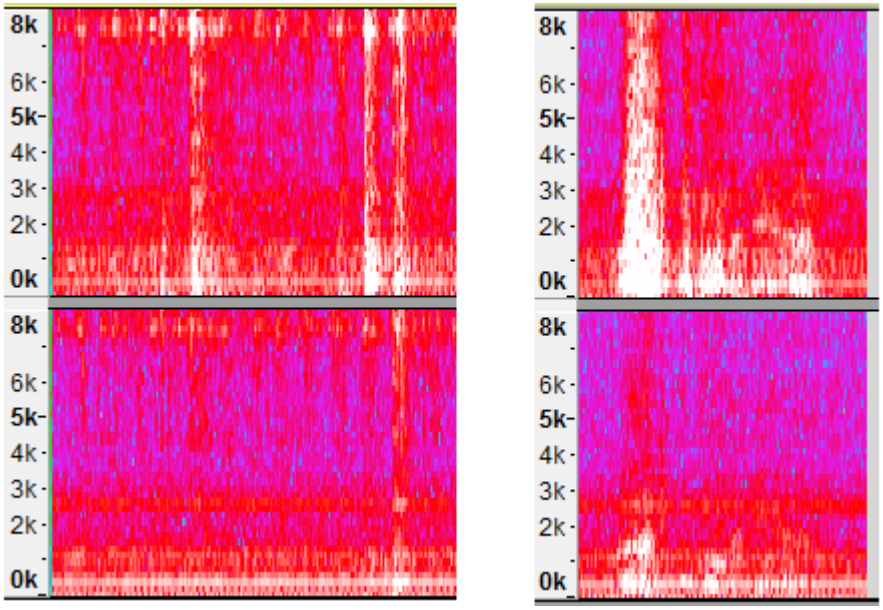

Figure 4.1 Effective frequency ranges for keys and voice recordings.

Figure 4.1 shows that the jangling keys sample populates higher frequency ranges, up to 8 kHz. On the other hand, the human voice generally maintains its frequency under 1 kHz, with clipping causing higher frequencies to be registered. Table 3.1 has already discussed the frequencies that are touched upon in the frequency sweep, so its spectrogram is not shown. The significance of knowing the effective

ranges of each sound sample is to understand the efficacy of the localization methods, as we know that ILD operates better for higher frequencies and ITD for lower.

The experimentation results are obtained and presented as follows: sound samples were first input into the Matlab system. As exampled in Chapter 3, the signals are first filtered, segmented, and then analyzed. For each block of length 2400 samples, it is classified by the ILD and ITD methods. The decisions for both blocks were recorded and cross-referenced to ensure their accuracy and consistency with each other's directional ILD, ITD decision and degree range. Recording each decision permitted analysis of each method's effectiveness individually, but also together, and the identification of the weakest areas in the decision-making process.

In the following subsections, the tables will show the overall success rate of the experimentation. They will show the ratio correct for ILD and ITD correct direction output, ITD accuracy within a certain degree limitation, ILD and ITD consistency, ILD and ITD correct direction output, and ILD and ITD correct angle output. The ratios are calculated by dividing the number of results for that column's header by the total number of classifications made, according to the equations

$$method\ ratio = \frac{specified\ column}{total\ \#\ blocks}, \tag{4.1}$$

$$final\ decision\ ratio = \frac{specified\ column}{total\ \#\ matching}. \tag{4.2}$$

The key difference between Equations 4.1 and 4.2 is that the former is the ratio to the total number of blocks; the latter is the ratio to the total number of ILD and ITD agreeing decisions. Therefore, Equation 4.2 answers the question: Of the total number of matching ILD and ITD decisions, how many of those decisions are actually correct?

4.1.1 Frequency Sweep

The performance for the frequency sweep was of fairly low accuracy, as seen in Table 4.1. Although it managed to perform somewhat effectively for left sources, right sources proved extremely difficult, especially in their ITD output. Because ITD was mostly incorrect for this set, it is seen that the overall output was highly inaccurate due to the disagreement between the two methods. The reason for such a discrepancy is likely to do with the unrealistic nature of the sound being played. Relatively high frequencies were successively played from one direction, possibly causing a large number of incorrect classifications for ITD, and contributing to the success of ILD. Still, it is interesting to note that the ILD directions were fairly accurate in terms of decision-making despite the automatic gain control, seeing as it had an average success rate of over 50%. Also, the seemingly higher performance of the left ear's ITD is unlikely to be actually that; Figures 4.2–4.4 show that the ITD results are heavily skewed to the left.

Table 4.1 Localization results for frequency sweep recording.

Direction	ILD	ITD	ITD (±15°)	ITD (±30°)		Correct (Dir)	Correct (±15°)
Straight	0.35	0.27	0.56	0.82		0.19	0.44
Left30	0.63	0.55	0.37	0.84		0.74	0.52
Left45	0.76	0.78	0.14	0.53		0.91	0.16
Left60	0.94	0.54	0.00	0.00		0.97	0.00
Left90	0.80	0.54	0.00	0.00		0.91	0.00
Right30	0.73	0.02	0.00	0.25		0.29	0.00
Right45	0.90	0.17	0.00	0.05		0.62	0.01
Right60	0.93	0.09	0.00	0.00		0.70	0.00
Right90	0.95	0.18	0.00	0.00		0.82	0.00

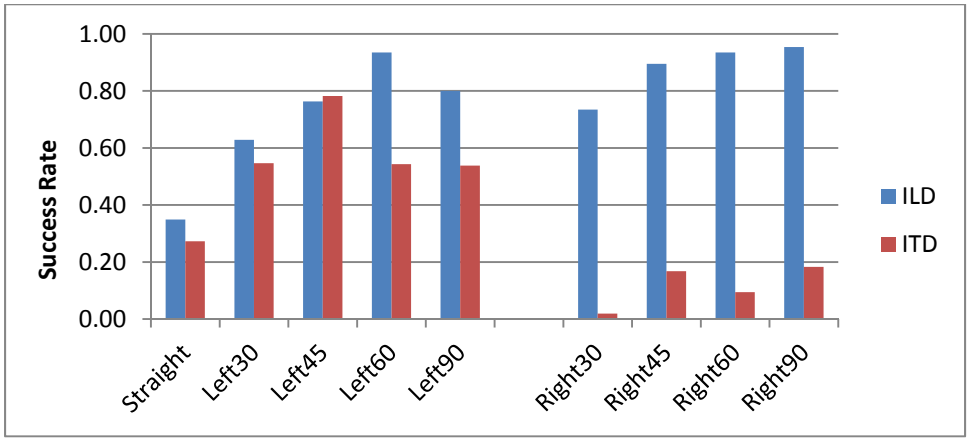

Figure 4.2 Sweep direction classified ILD versus ITD performance.

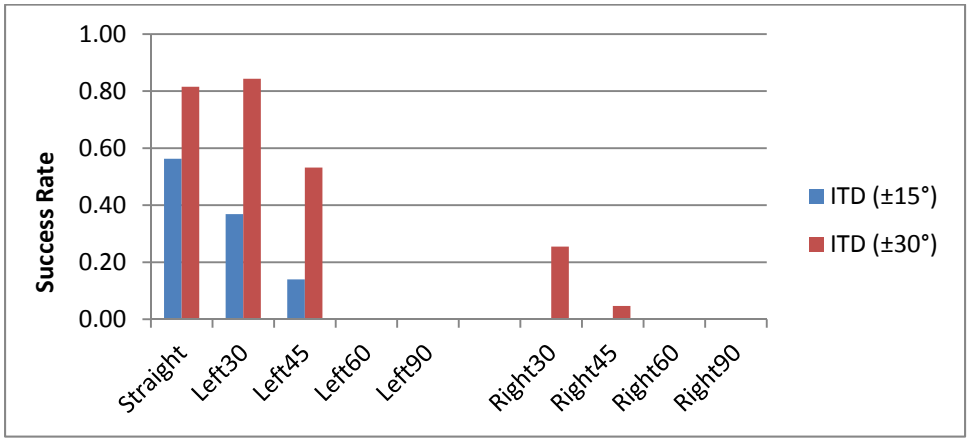

Figure 4.3 Sweep ITD angle classification performance comparison.

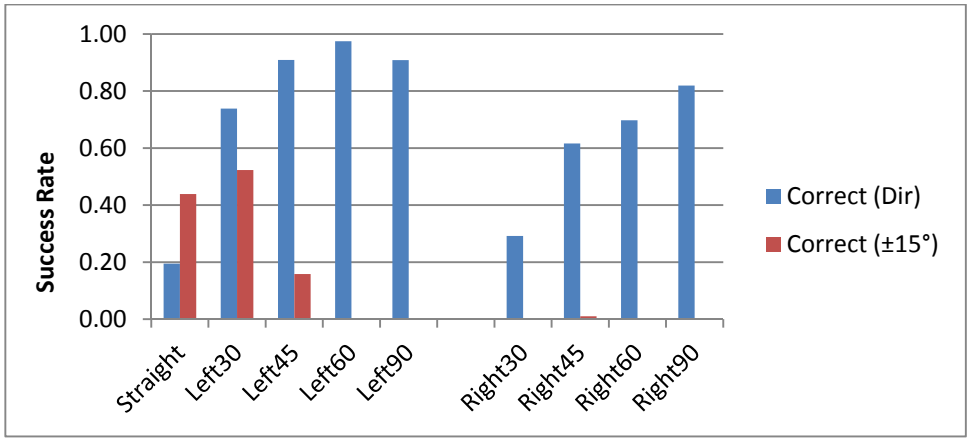

Figure 4.4 Sweep classified direction versus correct angle performance.

4.1.2 Voice

Voice sound samples represented the most realistic experiment in terms of utility, as locating the source of human speech has its obvious uses. This experiment section shows the primary weaknesses of ILD for lower frequency localization and the negative impact of the automatic gain system. Although the combined decision marked by the Correct column in Table 4.2 does show a low accuracy for right angles, Figure 4.7 and Table 4.2's ILD column clearly indicate that the gain was at fault for the erroneous final decisions. In contrast, Figures 4.5 and 4.6 verify ITD's relatively high performance, both in directional accuracy as well as the angle. In this case, it is probable that the automatic gain is the main source of error, seen from Figure 3.10's example of its problematic behavior.

Table 4.2 Localization results for voice recording.

Direction	ILD	ITD	ITD (±15°)	ITD (±30°)		Correct (Dir)	Correct (±15°)
Straight	0.15	0.40	0.75	0.75		0.29	0.86
Left30	0.92	0.58	0.38	0.58		1.00	0.62
Left45	0.79	0.21	0.11	0.16		1.00	0.50
Left60	0.78	0.61	0.13	0.61		0.93	0.21
Left90	0.88	0.73	0.23	0.23		1.00	0.29
Right30	0.07	0.80	0.33	0.60		0.25	0.25
Right45	0.05	0.95	0.26	0.53		0.00	0.00
Right60	0.00	0.85	0.08	0.62		0.00	0.00
Right90	0.06	0.61	0.30	0.33		0.10	0.10

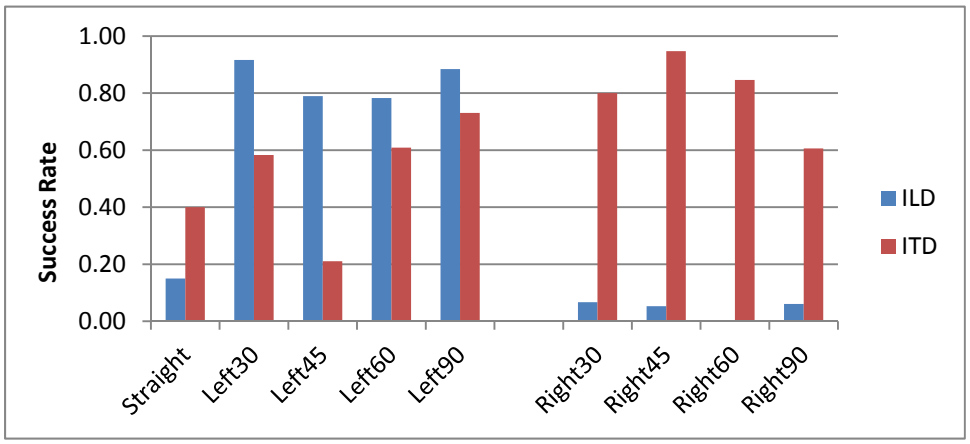

Figure 4.5 Voice direction classified ILD versus ITD performance.

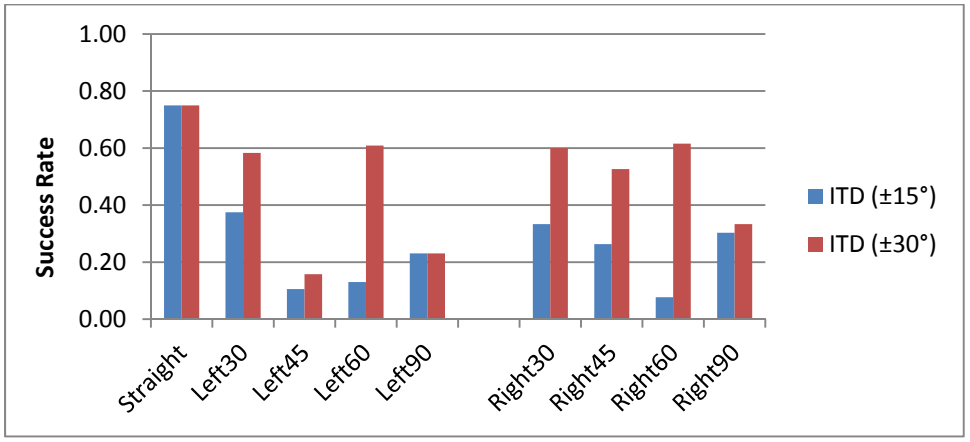

Figure 4.6 Voice ITD angle classification performance comparison.

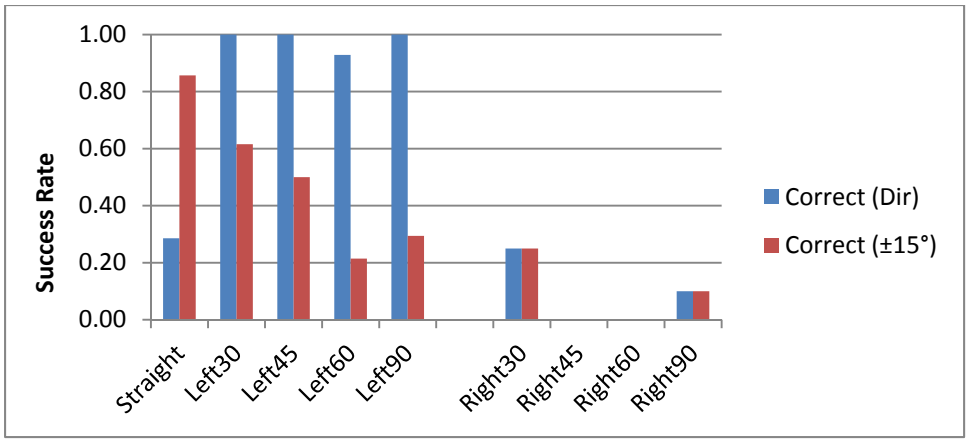

Figure 4.7 Voice classified direction versus correct angle performance.

4.1.3 Keys

The keys sounds results in Table 4.3 were the most heavily inaccurate and skewed. Due to the almost exclusively high frequencies, ITD was nearly useless in its ability to classify angles. In almost every case, it classified the signal as coming from the right at 83°, causing the heavily right-skewed values shown in Figures 4.8 and 4.9. Again, the 1.00 ITD success rates shown in Table 4.3 and Figures 4.8–4.10 are greatly misleading since the ITD's only localizations were directionally right. On the other hand, ILD worked fairly well with an average success rate of over 50%, which is expected because ILD functions well for high frequencies. Still, although direction is found with this nearly ILD exclusive localization set, the angle is completely vague, as ITD is not able to pinpoint the source's azimuth.

Table 4.3 Localization results for keys recording.

Direction	ILD	ITD	ITD (±15°)	ITD (±30°)		Correct (Dir)	Correct (±15°)
Straight	0.43	0.00	0.00	0.00		0.00	0.00
Left30	0.97	0.00	0.00	0.00		0.00	0.00
Left45	0.90	0.02	0.00	0.00		1.00	0.00
Left60	0.94	0.06	0.00	0.00		1.00	0.00
Left90	0.97	0.02	0.02	0.02		1.00	1.00
Right30	0.11	1.00	0.00	0.00		1.00	0.00
Right45	0.40	1.00	0.00	0.00		1.00	0.00
Right60	0.63	0.98	0.00	0.98		1.00	0.00
Right90	0.62	1.00	0.98	1.00		1.00	1.00

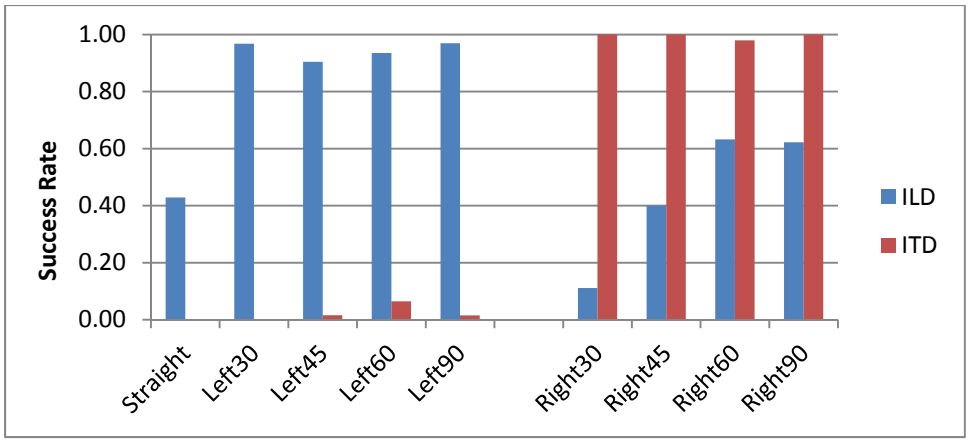

Figure 4.8 Keys direction classified ILD versus ITD performance.

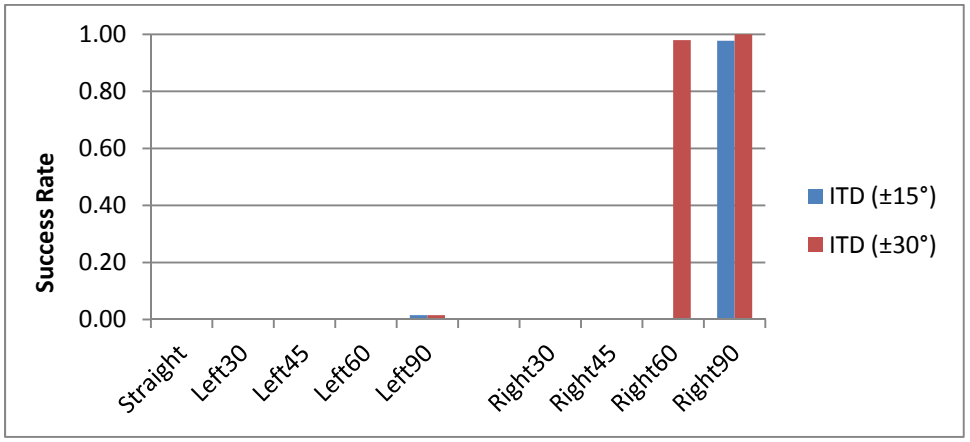

Figure 4.9 Keys ITD angle classification performance comparison.

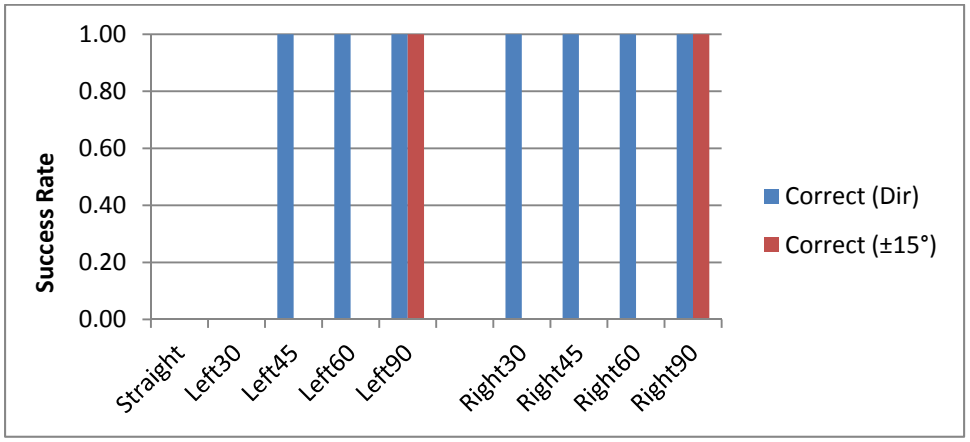

Figure 4.10 Keys classified direction versus correct angle performance.

4.1.4 Comparison

Table 4.4 Average localization results for each sound set.

	ILD	ITD	ITD (±15°)	ITD (±30°)	Correct (Dir)	Correct (±15°)
Frequency Sweep	0.78	0.35	0.12	0.28	0.68	0.13
Voice	0.41	0.64	0.29	0.49	0.51	0.31
Keys	0.66	0.45	0.11	0.22	0.78	0.22

All in all, it appears that each test scenario produced results according to theory and as expected. Visible in Table 4.4, the voice recordings boasted the highest success rate for every ITD category, while the frequency sweep and keys remained relatively close in their values for ILD classifications. One problem all of the test cases faced was a discrepancy between their ILD and ITD outputs, causing there to be far fewer final decisions than if only one method was used. However, if a final decision was made and the individual methods agreed, the advantage of combining both decision-making processes is clear, as the average Correct (Dir) column for every set is over 50%. This cannot be for any of the other columns, which are all dependent solely on one localization method. The gain control also proves to be problematic in its unpredictable behavior, shown in voice and keys sources where results are heavily skewed left, but appears to be subdued in the frequency sweep. Incidentally, the overall success rates were actually much higher than anticipated; initial evaluation of the inadequate hardware did not inspire confidence in the performance of this project.

4.2 Future Work

Although ILD and ITD methods are the most widely used and efficient ways of finding sound, there are other ways to obtain extra information regarding the position of the signal. These extra processes can be added to this project and Bert to improve classification performance. Even if the localization process is primarily done with a multi-aural system, there exist monaural cues which aid in the detection of location. These are known as spectral cues and require the introduction of a head related transfer function, or HRTF [7]. Represented in Figure 4.11, the HRTF is most effective above 4 kHz and is able to show changes in frequency response as a function of azimuth and elevation. This is caused by the ear's pinna and the head affecting the intensities of the frequencies [11]. The disadvantage of this method is that it requires a thorough training process, a process which humans and animals accomplish throughout their lifetime. Spectral cues may not provide the accuracy of ILD and ITD,

but they serve as more data points that our brains use for reference [6]. A number of studies demonstrate that humans are capable of locating sound with just one ear's cues, but with much poorer accuracy. Still, future implementation of the monaural spectral cues can definitely be useful to boost Bert's localization performance.

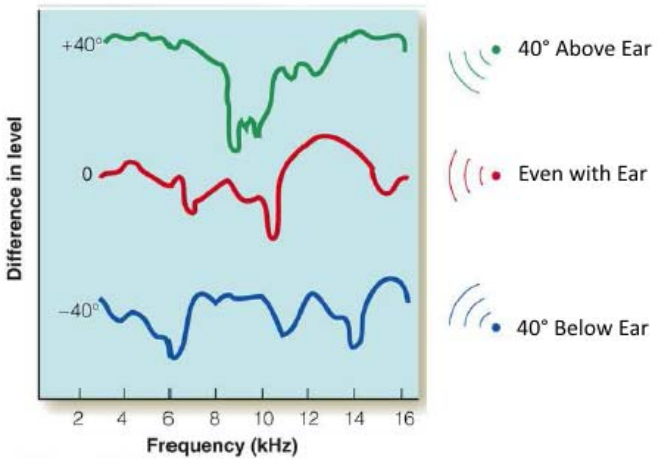

Figure 4.11 Differences in response resulting from differing location sources [8].

Another area for exploration is to locate sound sources within a 3D field instead of a 2D field, like done here. In addition to the azimuth angle calculated in this thesis, the altitude would be considered. The difficulty in this procedure is that it depends heavily on accurate spectral cues for an accurate reading. Most importantly, the head must continually tilt its axis to make use of the ITD and ILD methods in the positive z direction of space. This would require a continual feedback system of head tiling that must also communicate with the spectral cues calculation to hone onto a realistic position. Although these extra features are considerably more difficult as they heavily depend on a robust training set, they are definitely areas of interest for future development.

CHAPTER 5

CONCLUSION

The objective of this thesis was to explore the underlying processes in biological sound source localization and to build a similar framework within an existing environment. The performance level of biological systems is still unmatched, but creating the basis of core functionality within the Bert framework has been largely accomplished. The current hardware setup enormously hinders the potential of localization within Bert, with the low quality of the microphones and his inherent design flaws causing excess noise. However, classification performance was surprisingly acceptable with the voice set's mean 30% success rate for a ±15° restriction and a 50% success rate for ±30° restriction for the ITD method. It was interesting to note that while often ITD and ILD did not match in their classification, largely due to the automatic gain system causing disproportional and unrealistic changes in perceived sound volume, one method did manage to produce acceptable outputs. Furthermore, ILD localization for sources originating on the right side of Bert also proved to be difficult, again caused by the gain system. Nonetheless, there remains potential for higher accuracy with future additions such as spectral cues and head tilting, but it is expected that significant improvement will come with the pending hardware upgrades to remove many, if not all, of the limitations of Bert's system.

REFERENCES

[1] S. Levinson, "Speech and mind," University of Illinois at Urbana Champaign, Tech. Rep., 1998.

[2] A. S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: The MIT Press, 1990.

[3] P. Neil, "Why voice user interfaces with beamforming microphones work better with xCORE-VOICE," Xmos.com, 2016. [Online]. Available: http://www.xmos.com/blog/xmos/post/introducing-xcore-voice-smart-microphone-applications

[4] B. Grothe, M. Pecka and D. McAlpine, "Mechanisms of sound localization in mammals," Ludwid-Maximilians-Universitaet, Tech. Rep., 2010.

[5] T. Letowski and S. Letowski, "Auditory spatial perception: auditory localization," Army Research Laboratory, 2012.

[6] D. Wang and G. Brown, Computational Auditory Scene Analysis, 1st ed. John Wiley & Sons, Inc., 2005.

[7] D. Li, "Computational models for binaural sound source localization and sound understanding," Ph.D dissertation, University of Illinois at Urbana Champaign, 2003.

[8] G. Boynton, "Sound Localization and the auditory scene," University of Washington, Tech. Rep., 2008.

[9] "Icub.Org - An Open Source Cognitive Humanoid Robotic Platform." Icub.org. N.p., 2017. Web.

[10] G. Reid, "Active binaural sound localization techniques, experiments and comparisons," M.S. thesis, York University, 1999.

[11] C. Lenz, "Localization of sound sources," Swiss Federal Institute of Technology Zurich, Tech. Rep., 2009.