COMPUTATIONAL AUDITORY SALIENCY

A Thesis
Presented to
The Academic Faculty

by

Varinthira Duangudom Delmotte

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2012

COMPUTATIONAL AUDITORY SALIENCY

Approved by:

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

Dr. Pamela Bhatti
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Aaron Lanterman
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Doug Williams
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Bruce Walker
College of Computing; School of Psychology
Georgia Institute of Technology

Date Approved: October 31, 2012

To my family for their love and support

ACKNOWLEDGEMENTS

I want to first express my thanks and gratitude to my advisor, Dr. David Anderson, for his continued patience, guidance, understanding, and support throughout my studies. I also want to thank my committee members, Dr. Aaron Lanterman, Dr. Bruce Walker, Dr. Pamela Bhatti, and Dr. Doug Williams for their valuable feedback and for taking the time to serve on my committees. Dr. Bruce Walker was always available to provide extremely helpful advice and discussions on the experiments.

I want to thank all the friends I have made during my time here at Georgia Tech. They have made this entire graduate experience an enjoyable one. Thank you to all my current and past ESP lab members for making it such a fun place to come to work. I especially want to thank Sourabh for his help, discussions, and friendship. I also want to thank Walter, Faik, and Devangi for their friendship.

Finally, last but not least, I would like to thank my family. Without their continued love, patience, and support, all of this would not be possible. I want to thank my parents and sister for their love and encouragement throughout my life and studies. Thank you to my husband for his love and support. He was always willing to lend a helping hand or listening ear when I needed it. Thank you to my daughter who always puts a smile on my face. She has enriched my life in so many ways with her love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Sounds can change simultaneously along many dimensions. This includes changes in pitch, loudness, timbre, or location. Despite this, humans are still able to successfully integrate these different cues and use them to identify, categorize, and group sounds [4]. Human performance on such tasks continues to far exceed that of computational models, especially under certain conditions, such as noise, speaker change, and speaker variability. Humans, unlike many computational models, are able to easily account for speaker variability along with changes in other auditory percepts. Even in extremely noisy environments, we find ourselves able to relatively easily follow and understand the conversations that we are carrying on with others. One classic example of this is the "cocktail party problem" reported on by Cherry in 1953 [7]. Cherry investigated questions of how humans are able to selectively attend to a particular talker even in very noisy situations where multiple talkers are speaking at the same time.

In this thesis, we are concerned with identifying the sounds that grab a listener's attention. These sounds that draw a person's attention are sounds we consider salient. The focus here will be on investigating the role of saliency in the auditory attentional process. In order to identify these salient sounds, we have developed a model inspired by our understanding of the human auditory system and auditory perception.

By identifying salient sounds we hope to obtain a better understanding of auditory processing, and in particular, the key features contributing to saliency. Additionally, studying the salience of different auditory stimuli can lead to improvements in the performance of current computational models in several different areas, by making use of the information obtained about what stands out perceptually to observers in a

particular scene. Auditory saliency also helps to rapidly sort the information present in a complex auditory scene. Since our resources are finite, not all information can be processed equally. Therefore, we must be able to quickly determine the importance of different objects in a scene. Additionally, an immediate response or decision may be required. In order to respond, the observer needs to know the key elements of the scene. The issue of saliency is closely related to many different areas, including scene analysis.

The thesis provides a comprehensive look at auditory saliency. It explores the advantages and limitations of using auditory saliency models through different experiments and presents a general computational auditory saliency model that can be used for various applications. We begin by first defining specifically what we mean when we say something is salient in audio. From there, we present the model for auditory saliency that was developed based on research into saliency models for the visual domain. Finally, we consider the issues of evaluating auditory saliency as well as investigating the key features contributing to salience. We evaluate our model using several experiments which show that the model matches well with human performance in detecting salient auditory events in a complex scene.

CHAPTER I

INTRODUCTION

The auditory system performs many intricate steps involved in sound processing leading to our ability to hear and perceive various sounds. The sounds that we hear change constantly, and these changes can occur simultaneously along many dimensions. This includes changes in pitch, loudness, timbre, or location. Even with these changes, we are still able to successfully integrate the different cues and use them to correctly identify, categorize, and group sounds [4].

Despite the vast amount of research that has been done, humans continue to outperform machines in many auditory processing tasks, particularly in the presence of noise and other factors. For example, most of the existing computational models are still not able to match human performance when faced with certain issues, such as, speaker change or variability, in addition to changes in the other auditory percepts mentioned above. Humans, on the other hand, are able to deal well with the presence of noise, in such a way that we are able to successfully follow and understand the conversations we engage in with others even in extremely noisy environments. One classic example of this is the "cocktail party problem" reported on by Cherry in 1953 [7]. Cherry investigated the question of how humans are able to selectively attend to a particular talker even in extremely noisy conditions with multiple talkers present. This selective attentional process helps us to determine where our attention should be directed, and it can be either intentional (top-down, task-dependent) or involuntary (bottom-up, saliency-driven process). The bottom-up, saliency driven process does not require attention and occurs very rapidly and effortlessly. The top-down, task dependent process, on the other hand, occurs more slowly and involves

the higher-order processes and attention.

This thesis explores the concept of auditory saliency and presents a computational model for auditory saliency similar in function to the bottom-up, saliency-driven aspect of our attentional process. As it is commonly believed that attention is limited and we can only fully attend to one thing at a time, there must be some mechanisms by which attention can be switched [65]. Auditory saliency can explain how the involuntary part of selective attention occurs. Here, the focus is on identifying the sounds that grab a listener's ear or attention. For example, if we are currently engaged in a task, and we hear an alarm clock ringing, we are likely to stop what we are doing and look up. Thus, a salient sound is one that draws our attention away from where it is currently directed. In this way, salient sounds can drive attention by causing us to change our focus.

Auditory saliency also acts as a filter where important or salient elements of a scene can be passed on for additional processing. It suppresses the less salient stimuli as well, thereby reducing the amount of interference during the higher-order processing of the salient sounds. This can explain how we are able to quickly sort all the information present in any given complex auditory scene. Since our resources are finite, it is impossible for us to process everything equally. Therefore, we must be able to identify what is important in a particular scene. A recent study by Mesgarani and Chang in [49] reinforces the idea of an early filtering process being used by the human auditory system in order to ensure that only relevant information is passed on for higher level processing. Additionally, an immediate response or decision may be required from us. In order to respond, we need to be aware of the most important elements of the scene. The computational auditory saliency model mimics this part of the attentional process and provides us with an auditory saliency map which indicates the areas of a particular auditory scene that stand out perceptually to observers. This information can then be used to improve the performance of current computational

models in many different areas, including noise suppression, audio classification, and speech recognition.

The computational auditory saliency model is analogous to visual saliency models that are used to identify what areas of an image stand out to observers. Much more saliency research has been done in the visual domain, and there are many existing visual saliency models that are used for various applications, such as detection or recognition. Therefore, we begin with a brief overview of visual saliency and some of the existing visual saliency models. We will then follow with a discussion of some of the challenges encountered in going from visual saliency to auditory saliency. Finally, we conclude this chapter with some of the main contributions of this thesis along with an outline of the thesis.

1.1 Visual Saliency Models

The vast majority of all saliency research has been focused on the visual domain. One reason that may explain why visual saliency has been much more researched than auditory saliency is that the primitives for vision are better understood, thus making the concept of saliency more easily definable. Another reason is that the evaluation of visual saliency is more straight-forward as there is a physical correlate, eye-tracking, that can be used. Eye-tracking, which has been shown to correlate with stimulus salience [17], allows us to physically measure or quantify visual saliency. In this section, we provide an overview of some of the motivations behind looking at salience and the current research in this field with regards to visual saliency.

Humans use a saliency-driven attentional process in order to identify objects that stand out in a visual scene. Based on psychological and physiological experiments, it is believed that bottom-up visual saliency is what drives this attentional process allowing us to rapidly scan an entire visual scene and then highlight a small area of the scene as being salient [44, 73, 29]. This highlighted information can then passed

on for higher level visual processing, allowing us to filter out less important aspects of the scene and reducing the computational load to our systems [43]. This filtering process is often referred to as "selective attention" [43].

The goal of visual saliency research is to create a computational model that can determine where an observer's attention will be directed when they are initially confronted by a visual scene [38, 57]. These models can also identify the order in which an observer's gaze will be drawn to different objects in the visual scene. The majority of existing visual saliency models are pre-attentive models that drive attention as opposed to being determined by attention. They were inspired by Treisman and Gelade's "Feature Integration Theory" presented in [73]. Here, Treisman and Gelade proposed that features come before perception, and early visual features are registered automatically in parallel across the visual field [73]. It is then only later, with attention, that the features can be identified as objects.

The feature set of the saliency models include simple visual attributes that neurons in early visual processing are tuned to [37]. Luminance or intensity, along with orientation and color, are three main features typically used in the feature set of all visual saliency models. For example, the model by Itti in [38], extracts these three features, in parallel, along various scales. The different scales are then compared using a center-surround mechanism to obtain the feature maps, where the finer scale is the "center" and the courser scale is the "surround." The feature maps are then combined into a two-dimensional "saliency map" corresponding to the points within an image that are salient. The idea of having a saliency map that corresponds directly to an image may be easier to visualize as opposed to the concept of having a saliency map for audio, which may be another reason that visual saliency has been more widely researched.

The saliency of a particular object will depend on a variety of features, and different features contribute with varying degrees or weights to the overall saliency map [3].

The feature set selected for the saliency models attempt to capture the main features contributing to object salience. Below in Figure 1 is a simple example of the strong pop-out effect of bottom-up visual saliency. One of the bars immediately stands out from the rest of the scene and is considered salient due to its different color and orientation. The pop-out effect of bottom-up saliency demonstrated here works pre-attentively, and it is the salience of the stimulus that drives or directs attention. It occurs very rapidly and requires no effort on the part of the observer to identify the salient stimulus. As illustrated by this example, how different an object is from the rest of the surrounding scene will be an important factor contributing to salience from this bottom-up view.

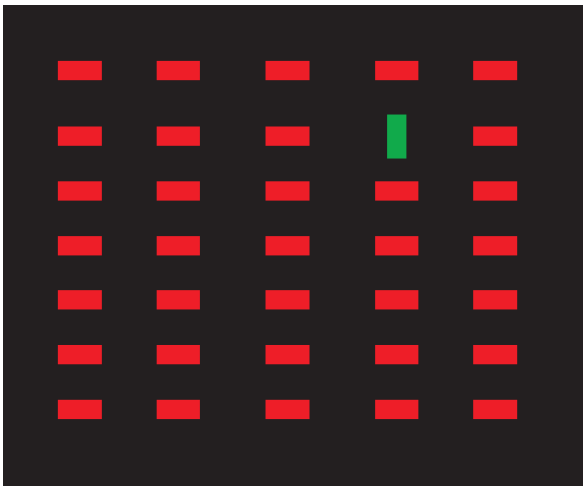


Figure 1: Simple example demonstrating the pop-out effect of bottom-up visual saliency.

By contrast, in Figure 2, we show an example of a conjunctive search problem which results in no pop-out effect. A conjunctive search problem is one where the target cannot be defined by a single unique feature which results in a longer search process as saliency plays a smaller role in guiding our attention. In this case, we have one bar that again differs from the other bars, but it does not stand out to us like the green bar from the previous example. Here, the observer must search through the entire image in order to find the target, which is the one red bar with the different

orientation.

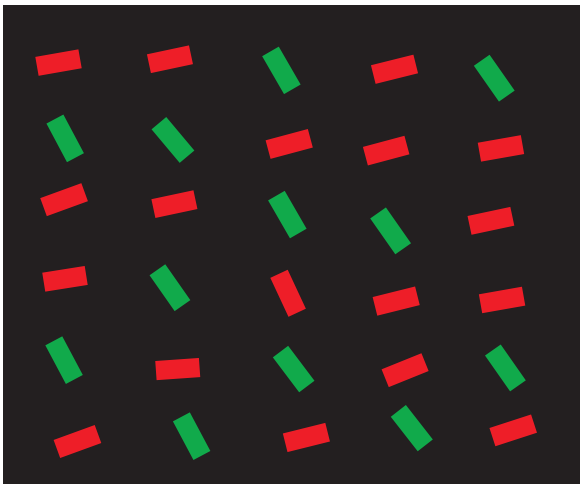


Figure 2: Example demonstrating no pop-out effect when saliency does not guide attention.

The visual saliency models have been evaluated in various experiments by comparison to human observers and have been shown to perform well on certain types of tasks [38, 73]. One method of evaluation that is very commonly used in visual saliency is eye-tracking, and one advantage of this method is that it does not allow observers the time or opportunity to consider their responses. It simply follows where an observer's gaze is naturally drawn, as opposed to requiring an observer to report on what they view as most salient in a scene. Here, the order in which they look at different objects in the scene is also noted and provides information on the degree of saliency. It is important to remember as well that there is no truly objective measure for evaluating saliency, since saliency can vary among different observers and will also be influenced by top-down or task-dependent information.

Since the development of the first computational visual saliency models, the focus of many visual saliency researchers has expanded to include applying the models to various applications. Some of these applications include: video compression, video processing, background subtraction, and 3D visual saliency. There also continues to be a large amount of research focused on further developing the feature set of these

models. For example, many of the recent visual saliency models now incorporate motion and features such as eye and head movement animation [36], along with the three basic features mentioned previously. Adding motion as an additional feature allows the visual saliency models to be applied to not only images, but also to video segments as well, where the models can be used to track salient targets in a video scene. Finally, another area for continuing research in visual saliency is creating visual saliency models that take into account both top-down and bottom-up processing. One top-down implementation that has been introduced is the one by Peters and Itti in [58] where "gist" features are used to capture the top-down information.

The concept of auditory saliency is analogous to that of visual saliency, although creating such a model for the auditory domain is a relatively novel and recent idea. We first started investigating the idea of an auditory saliency model, based on studying visual perception [23, 34]. In particular, different visual saliency concepts and models were examined, such as the one by Itti in [38]. Additionally, observations made from work in other related areas also indicated that an auditory saliency model would be useful [50, 61] for various applications. Studying visual saliency models is a good starting point, as auditory and visual systems are similar in many ways, and the idea of applying a saliency model for other modalities, such as the auditory domain, had not yet been extensively studied. In particular, several different studies have found similarities in the cortical processing for both the visual and auditory systems [70, 18].

A model of auditory saliency is interesting as it can be used to help expand our understanding of auditory perception and attention. The auditory saliency model would be able to detect and rank sounds that are perceptually salient to listeners, much the same way that the visual saliency models can detect the salient points or regions in an image or video. There are also many possible applications that could benefit from applying information provided by our computational auditory saliency model on the saliency of different objects. This includes applications in the areas of

surveillance, noise suppression, auditory scene analysis, and audio classification.

1.2 From Visual Saliency to Auditory Saliency

Several considerations make studying auditory saliency a challenging task. The first issue that must be considered is how auditory saliency is defined. At the time we began researching auditory saliency and applying the concept of saliency to the auditory domain, there were no existing auditory saliency models. One of the key questions asked was what is auditory saliency and, more specifically, what is meant when we say that something is salient in audio.

A second issue to consider is the feature set that should be used by the auditory saliency model. Unlike with vision, where many of the basic features used in early auditory processing have been determined, the primitives in audio are not as well understood. This can make it more difficult to determine the appropriate feature set.

A third issue that makes auditory saliency different than visual saliency is the fact that audio has a temporal component that must be taken into account. In the past, with the visual saliency models, the input image was an instantaneous presentation that did not vary with time. Here, we have an auditory scene that is constantly changing with new sounds being introduced over time. How this time component will be incorporated and whether or not a new sound is now more salient than the previous sounds will need to be considered. Having this time aspect, in some ways, is more natural as when we encounter "scenes" in the real world, we are not just viewing an instantaneous snapshot in time, such as the static images, but instead, we are seeing events occurring continuously over a span of time.

Finally, how to evaluate the performance of the auditory saliency model also needs to be considered. The evaluation of auditory saliency is slightly more complicated than that of visual saliency as the auditory system does not have a physical correlate that can be easily measured, such as eye-tracking, to use for evaluation. Therefore,

it will be necessary to determine a method that can be used to measure auditory saliency. In the next chapter, we will address some of these concerns as well as present the current auditory saliency research.

1.3 Thesis Contributions and Outline

In this section, we state some of the main contributions of this thesis and provide an outline of the thesis. This thesis involves contributions in several different areas, leading to an overall, comprehensive look into computational auditory saliency.

The first contribution of the thesis is presenting a novel computational auditory saliency model. The model presented here uses features that are generated from a cortical model to extract information on spectral and temporal modulations. One novel contribution of our model is it takes into consideration the fact that certain frequencies can be perceived as louder, and as result, be more salient. This is something that can greatly influence what is salient to the model, but it is not accounted for by any of the existing auditory saliency models. Additionally, the model is a general auditory saliency model that matches well with results from human subjects on selecting salient sounds from a scene. One advantage of the model is that it can be applied to a variety of applications. Many of the other currently existing models are tailored towards a specific application and have only been evaluated for that particular task. Therefore, we do not have any knowledge of its overall performance in selecting or ranking the saliency of different sounds from the various sounds present in complex auditory scene. This model was evaluated using experiments with human subjects that involved making comparisons between different stimuli and selecting the one that is most salient.

The question of how to evaluate auditory saliency is an ongoing issue as there are no easily measurable physical correlates. The use of the dual task experimental paradigm to evaluate auditory saliency is a new contribution. Dual task experiments

have long been used in various studies involving attention in the different modalities, but it has never been used to investigate auditory salience. The saliency model presented in this thesis is meant to mimic the function of the bottom-up, saliency-driven part of our attentional process. From this perspective, sounds are considered salient if they can be noticed without our attention. Therefore, we show that dual task experiments are an appropriate and useful tool for evaluating this bottom-up auditory saliency process.

Another contribution of the thesis is the exploration into the dimensions or features contributing to the salience of different sounds. This information can then be used to create a more comprehensive feature set. In particular, we were able to improve our auditory saliency model by adding a pre-processing stage based on the results of this analysis. The new stage resulted in a significant improvement in the correlation values between the model's responses and subjects' responses. Our experiment and the analysis performed using multidimensional scaling (MDS) is the first of its kind, as there is no existing research on identifying the factors the contribute towards making a sound salient.

There are many applications that the auditory saliency model can be used for. In this thesis, we apply the auditory saliency model to the novel application of selecting salient auditory segments from video. Here, the stimuli used are two-channel audio inputs. The current auditory saliency models have all been used and tested primarily on monaural sounds. Thus, another novel contribution of this thesis is the use of the model on two-channel input sounds. While the model was not created for binaural sounds, this experiment shows how it can also be useful for identifying salient stimuli from binaural inputs. We include discussion on how the model can be applied to binaural sounds and show that the model performs as well in selecting salient segments for two-channel input sounds as it did in previous experiments on monaural sounds.

This thesis is divided into seven chapters. In the introduction, we provided some of

the background and motivations for research in saliency. Following the introduction, in Chapter 2, we define the concept of auditory saliency. We also address some of the challenges of creating an auditory saliency model. In Chapter 3, we then present the computational auditory saliency model. This auditory saliency model focuses on the bottom-up processing aspect of our attentional process and uses inhibition of features obtained using auditory spectro-temporal receptive fields to compute a saliency map. This saliency map identifies the sounds that are most salient in a complex auditory scene. We also provide some examples of saliency maps for various auditory stimuli that demonstrate our model is able to replicate some well-known psychoacoustic experimental results. This chapter concludes with an introduction to the auditory saliency tool that we use to easily generate saliency maps for different sounds.

Chapter 4 deals with addressing the issues involved in evaluating or quantifying auditory saliency. Here, we present results from two different types of experiments that were performed. These two experiments are both appropriate for evaluating auditory saliency. The first experiment is a dual task experiment which involves performing two tasks simultaneously. The subjects were asked to attend to one task, while the second task was occurring simultaneously in the background. It is the subjects' performances on each of the tasks when performed separately and simultaneously, which indicate whether or not a stimulus is salient. The second experiment is the saliency scene comparison experiment. In this experiment, we presented subjects with various auditory scene pairs and asked them to indicate which scene from the two scenes contained the most salient element. The results of these two experiments show that our model matches well with subjects' responses regarding the saliency of different stimuli.

In Chapter 5, we investigate what features contribute towards making a sound

perceptually salient. Here, we discuss using the statistical technique of multidimensional scaling (MDS) to analyze subjects' ratings of saliency for different auditory scene pairs. We demonstrate how this technique can be used to identify the underlying dimensions of sound that may affect salience. From the results of this experiment, we examine if there are any additional features that make a sound salient that could be added to the existing feature set. This is the first experiment of its kind to try to identify the dimensions of sound that affect saliency.

Chapter 6 explores applying the auditory saliency model to video scenes with their corresponding soundtracks. The goal is to explore whether or not the most salient auditory segment of a video scene can be used to provide a summary snapshot that is representative of the entire video scene. The stimuli set for this experiment is different from the one used in previous experiments in that the inputs are two-channel, stereo sounds, as opposed to monaural sounds. In addition, the soundtracks include speech, which was not included as part of the stimuli set in previous experiments. We present results comparing the model's responses to subjects' responses in selecting segments containing the most salient sound for when only the audio soundtrack is presented and also for when the video is shown with the corresponding audio.

Finally, in Chapter 7 we conclude with a summary of the main contributions of this thesis. We also briefly discuss the possibilities for further research in the area of auditory saliency.

CHAPTER II

DEFINING AUDITORY SALIENCY

In this chapter, we address some of the issues and challenges involved when applying the concept of saliency to the auditory domain. The first issue that we needed to address was how auditory saliency is defined.

2.1 Auditory Saliency Definition

Saliency defines what is obvious or striking about a particular feature. We found that auditory saliency could be looked at from two different perspectives. The first is a pre-attentive or bottom-up processing perspective, similar to that used by many visual saliency models. This perspective seeks to model the saliency-driven portion of the attentional process, which occurs rapidly and effortlessly. It focuses on the physical attributes of a sound that make it salient, and the main goal of the bottom-up models is to identify sounds that stand out and grab a listener's ear or attention. If we look at saliency from this bottom-up view, we define salient sounds as those sounds that can be noticed without attention. These sounds draw a listener's attention and can cause them to shift their attention from the currently attended task. For example, regardless of the task a subject may be attending to, an alarm sound such as a fire alarm ringing would be considered salient. Although there is also a top-down component that makes an alarm or siren especially salient to us (we have learned they can indicate danger), many of these types of warning sounds have been designed to grab our attention and are inherently salient based on the attributes of the sound itself.

On the other hand, we can also consider saliency from a top-down perspective. In this case, higher level information biases the bottom-up cues, thereby changing

what is considered salient. Sounds stand out due to their meaning despite the fact that they may not be acoustically significant. It involves the higher level processes and takes into consideration our prior knowledge or learning from past experiences. Thus, top-down auditory saliency models need to incorporate some type of learning mechanism for the top-down input or feedback. Here, we consider that humans create different models for various situations based on their past experiences. They then use these previously learned models to understand and interpret complex auditory scenes. Salient sounds would be defined as those sounds that violate these models. From a top-down perspective, sounds that are unexpected for a particular situation would be very salient, while sounds that are typical of or expected will become part of the background and less salient. For example, if you are attending a basketball game, even though the sounds of the crowd cheering or the band playing can be very loud, they do not stand out to us here, since they are all things that we would expect to hear in this environment. Therefore, in this case, these sounds are not salient. On the other hand, if you are in the classroom and you suddenly hear people cheering or a band playing, these same sounds are now much more salient. This is because these sounds are unexpected for the classroom setting. Therefore, saliency from a top-down perspective will depend on the particular context of a scene or task. As illustrated by this example, despite the fact that they are the same sounds, the degree of saliency is dependent on the situation and whether or not the sound is expected for that scene.

Top-down input can also result in certain sounds becoming more salient to only particular individuals based on what they have learned from their past experiences. For example, one person that has an extreme fear of dogs, due to receiving a dog bite as a child, may find the sound of a dog barking much more salient than someone who does not have this same fear. Some other examples of top down feedback making various sounds more salient to different individuals include our name and the sounds of a baby crying. If we hear someone calling our name, it is likely to draw our

attention regardless of any task we may be currently occupied with. Similarly, the sound of a baby crying is particularly salient to mothers, although there may also be aspects of a baby's cry that are salient due to bottom-up cues as well. Overall, as for the siren and baby cry, our perception of sounds will involve a combination of the bottom-up and top-down cues.

The question of how heavily we rely on each of these cues, and how these different cues interact with one another in the attentional process is still largely unanswered. One recent paper demonstrates the importance of attention and both top-down and bottom-up saliency on how we perceive complex auditory scenes [25]. In Figure 3, we provide a block diagram showing how the top-down and bottom-up cues could potentially interact in the control of auditory attention as proposed by our model [22]. Here, the sound input undergoes processing by the peripheral auditory system where different frequencies are resolved by the cochlea. The output of this stage is then transmitted by neurons to the central auditory system for further processing. This processing results in the multi-scale feature maps used by our computational auditory saliency model and discussed in detail in the next chapter. While our auditory saliency model focuses on identifying the sounds that grab our attention based on bottom-up cues, we also show in Figure 3 where top-down cues could bias the bottom-up cues in the evaluation and feature combination stage, thereby influencing where attention will ultimately be directed.

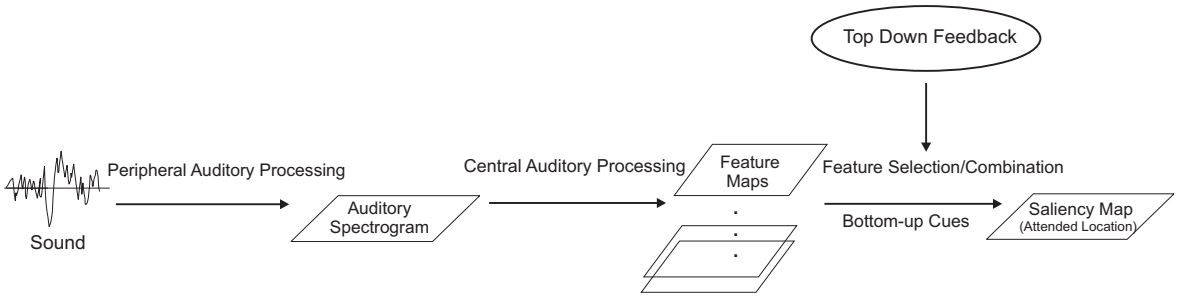


Figure 3: Block diagram of processes involved in auditory attention as proposed by our model [22].

2.2 Feature Set Selection

In creating an auditory saliency model, we needed to determine the main features that make a sound salient. This information is important in ensuring that a proper feature set can be chosen. As we started our research into selecting a feature set for the auditory saliency model, we first considered using a combination of simple features, such as, pitch, zero crossings, intensity, and frequency centroid to form the feature set. Although these simple features do provide information important in auditory processing, we were interested in using features similar to those the brain potentially uses in processing auditory sounds. Since human performance continues to exceed computers in many areas, we wanted to first start by matching human performance by using biologically plausible features. Thus, instead of using a collection of simple features to generate the feature maps for the auditory saliency model, we chose to use a cortical model which works in the spectrotemporal domain and uses various scales of temporal and spectral resolution. The cortical model represents the role of the central auditory system where similar processing is thought to occur on the inputs from the cochlea. The cortical model proposed in [77] was developed based on both psychoacoustical and physiological results. It is based on data obtained from primary auditory cortex neural responses of ferrets to various stimuli [19, 45, 46]. One reason for using this model to generate the feature set is that it extracts many of the cues important in auditory perception, including the spectral and temporal modulation information, both of which play a key role in sound perception and segregation [53, 52, 1]. The cortical features have also been used with success for several other related applications [60, 50, 51, 24].

In the next section, we present models for both the peripheral and central auditory systems, including more details about the cortical model. Exactly how the cortical model is used to generate the feature maps will also be discussed further in the next chapter, where the current computational auditory saliency model is presented. In

addition, in Chapter 5, we present an experiment performed that was specifically aimed at improving the feature set of the model.

2.3 Modeling the Auditory Pathways

Sound is created by the vibration of objects, and the sounds that we perceive are a result of complex processing performed by our auditory system. These vibrations result in changes in pressure creating the waves that carry the sound signal. These pressure waves are then later converted into nerve impulses by the peripheral auditory system.

In this section, we provide a brief overview of the processing performed by the peripheral auditory system. We then present a model used to represent the early auditory processing completed by the peripheral auditory system. We use this model to create two-dimensional auditory spectrograms that are used as an input "image" to the auditory saliency model. This "auditory image" is analogous to the images used as the inputs to visual saliency models and is one way we incorporate the time component of audio. Finally, we finish the section by presenting the cortical model that represents central auditory processing.

2.3.1 Peripheral Auditory System

The peripheral auditory system consists of the outer, middle, and inner ear. Figure 4 shows the structure of the human ear. Sound pressure waves enter the ear and travel down the auditory canal. These waves cause the eardrum to vibrate, and one of the main functions of the middle ear is to transmit these vibrations from the eardrum to the cochlea and inner ear. The middle ear is made up of three small bones: the malleus, incus and stapes. These three bones together are referred to as the ossicles, and they perform impedance matching to ensure the efficient transfer of the sound pressure waves from the air into the cochlea.

The sound waves that enter the cochlea will ultimately be converted into neural

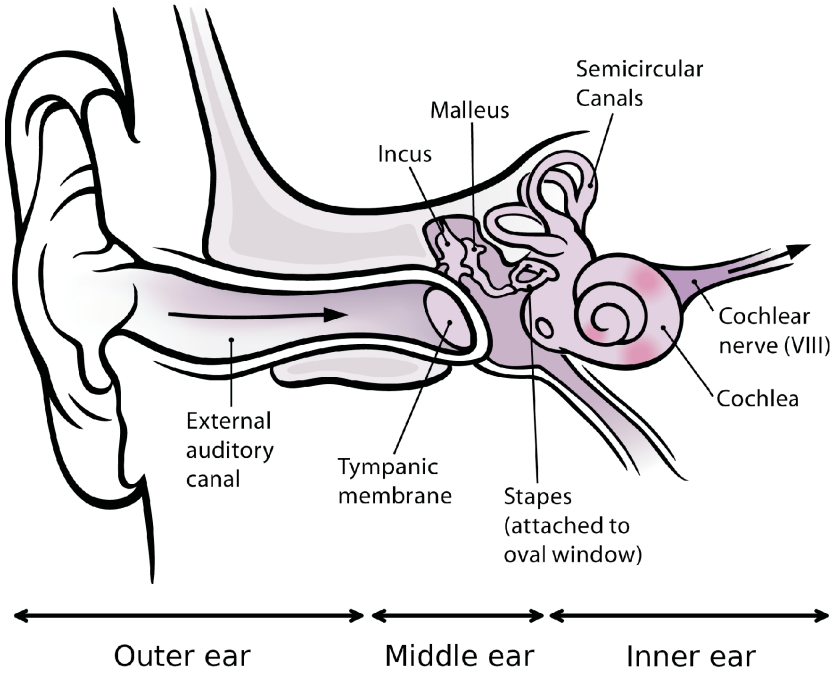


Figure 4: Structure of human ear showing outer, middle, and inner ear. Adapted from [10].

spikes that travel to the brain for further processing. These waves enter through the oval window and cause the fluid in the chambers of the cochlea to vibrate and move. This motion of the fluid leads to pressure changes resulting in the displacement of the basilar membrane. Figure 5 shows a cross-section of the cochlea. As shown here, the cochlea is made up of three fluid-filled chambers (scala vestibuli, scala tympani, scala media) that are divided by two membranes, Reissner's membrane and the basilar membrane.

The basilar membrane plays an important role in the frequency resolution and analysis occurring in the cochlea. It varies in width and flexibility along its length, and this property is what allows it to resolve different frequencies. At the base, it is narrow and stiff, gradually becoming wider and more flexible as you move down the length towards the apex. Thus, different frequencies cause peaks at varying locations on the basilar membrane. For example, high frequencies will result in maximum displacement near the base of the basilar membrane, while lower frequencies will

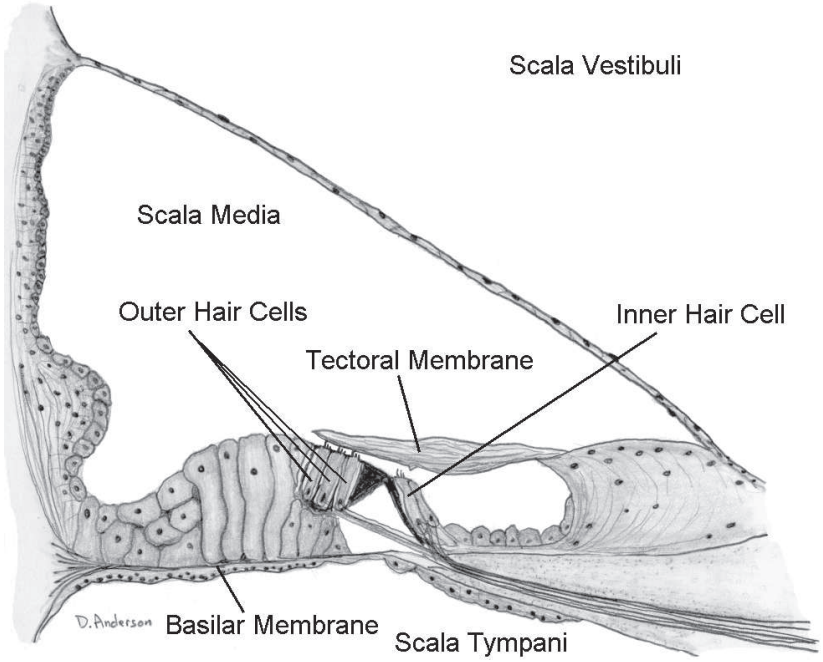


Figure 5: Cross-section of cochlea.

travel further down the length and cause maximum displacement near the apex. As shown in Figure 5, there are hair cells that lie along the top of the basilar membrane. These hair cells, particularly the inner hair cells, are very important in the conversion process of pressure waves into action potentials. This conversion takes place in the organ of Corti, located between the tectorial and basilar membranes.

The conversion process begins when movement of the basilar membrane causes displacement of the stereocilia on top of the hair cells. This movement opens the ion transduction channels, allowing potassium ions to enter the cell, leading to a voltage difference across the membrane of the cell. This voltage difference is what eventually results in the firing of action potentials by the auditory nerve neurons. This information is then transmitted from neuron to neuron until it reaches the cochlear nucleus in the central auditory system for further processing.

In the following section, we discuss modeling the peripheral auditory processing using mathematical models.

2.3.2 Modeling Peripheral Auditory Processing

The models of the auditory system are typically separated into three different stages [79]:

1. Analysis stage - Models how the basilar membrane spatially separates different frequencies along its length

2. Transduction stage - Models how the hair cells convert waves along the basilar membrane into nerve impulses

3. Reduction stage - Models the lateral inhibition thought to be performed in the cochlear nucleus [66].

One such biologically motivated model of early auditory processing is proposed by Yang *et al* in [79]. A block diagram of this model with the three different stages is shown in Figure 6. This model is briefly discussed here, since the auditory spectrograms that are used by the auditory saliency model are generated based on this model. In addition to being physiologically inspired, another advantage of the model presented in [79] is that it has been shown through various experiments to be robust to noise [79, 76]. This is important as our auditory system has been shown to be quite robust to noise compared to its current computational counterparts.

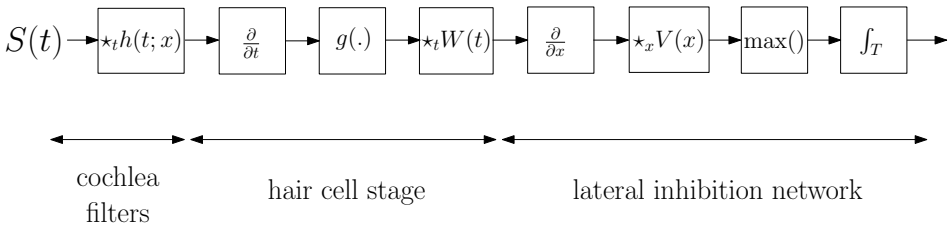


Figure 6: A block diagram of the mathematical model of the peripheral auditory system by [79].

The frequency resolution and analysis performed by the cochlea can be thought of as a bank of bandpass filters, where each point along the basilar membrane has a different transfer function. In this model, the analysis stage is modeled using a parallel

bank of asymmetrical bandpass filters that are uniformly spaced on a logarithmic axis. The output of the analysis stage is given by Equation 1.

$$y_{CF}(t,x) = s(t) * h(t;x) \quad (1)$$

$$y_{HC}(t,x) = g(\partial_t y_{CF}(t,x)) * w(t) \quad (2)$$

$$y_{LIN}(t,x) = max(\partial_x y_{HC}(t,x), 0) \quad (3)$$

$$y_{AS}(t,x) = y_{LIN}(t,x) * [e^{-\frac{t}{\tau}} u(t)] \quad (4)$$

The transduction stage is modeled in three different steps. First, a time derivative is used to model the coupling of the fluid and stereocilia. This is done as it is the flow (velocity) of the fluid, due to the waves, that causes the transduction channels to open. Secondly, the ion channels are modeled using an instantaneous non-linear sigmoidal function, represented by g(.) in Equation 2. In the third step, a low pass filter, w(t), is used to model the ionic leakage in the hair cells.

Lastly, for the reduction stage, a lateral inhibition network is thought to produce a spectral profile of the sound [66]. Equation 3 shows the output for this stage, which is modeled in three steps. The first step is a spatial derivative followed by local smoothing, v(x), to model the lateral inhibition properties between neurons, as this serves to highlight the spatial discontinuities in the auditory nerve patterns [79]. This is then followed by a half-wave rectifier to represent the non-linearity of the neurons and an integrator function to represent the poor response of central auditory neurons to rapid changes.

The final output of this model, y_{AS}, is an auditory representation or spectrum of the signal, which can then be used for further processing and for various applications. In the next section, we present a cortical model that models the role of the higher order processes, such as those that occur in the primary auditory cortex.

2.3.3 Modeling Central Auditory Processing

The auditory representation obtained by the early auditory processing model in the previous section is thought to be similar to the inputs to the cortex, but exactly how the cortex uses this information for further processing is still uncertain [77, 39]. One cortical model, shown in Figure 7, was proposed by Wang and Shamma in [77]. It models spectral shape analysis and the role of central auditory processing. The model was developed based on both psychoacoustical experimental results with human subjects [8] and neurophysiological experimental results in animals [45, 46]. Physiological experiments in the primary auditory cortex of ferrets and cats have confirmed that cortical cells have response properties matching those of this model [5, 30, 39].

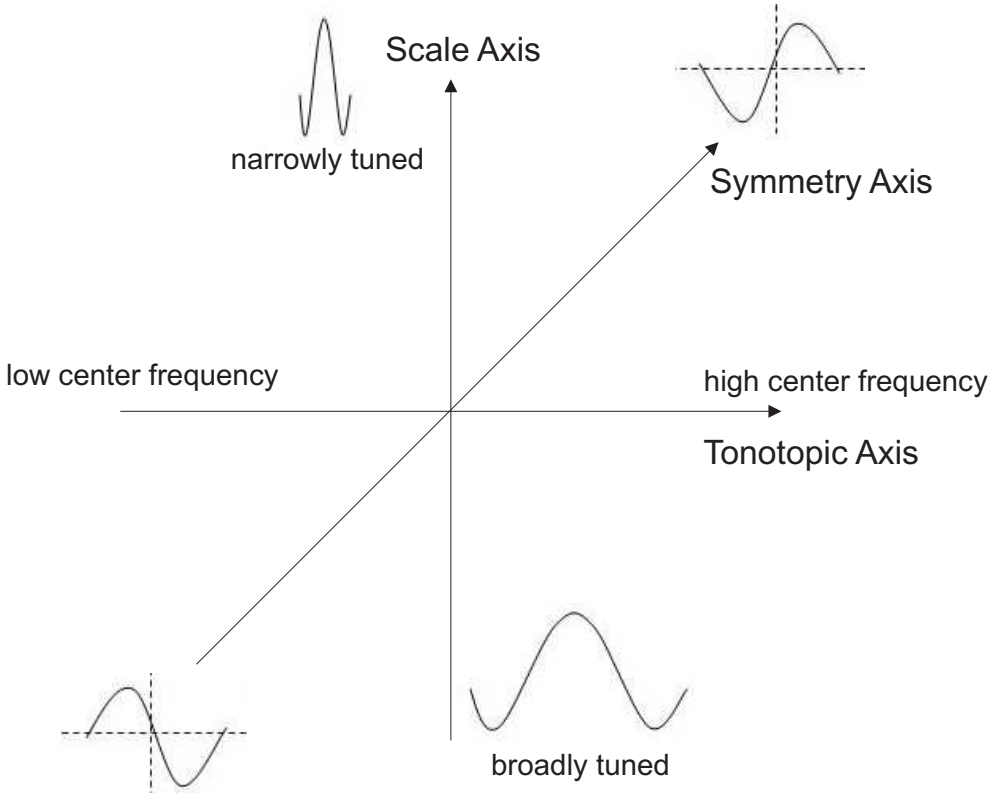


Figure 7: Schematic of the cortical model. Adapted from [77].

As shown in Figure 7, according to the cortical model, the response areas of cortical

neurons in the primary auditory cortex (A1) are organized conceptually along three spatial axes:

- Tonotopic axis - Along the tonotopic axis, neural response areas are tuned to different center frequencies.
- Scale axis - Along the scale axis, the response fields vary in bandwidth from broadly tuned to narrowly tuned.
- Symmetry axis - Along the symmetry axis, the asymmetry of the neural response field varies.

The asymmetry of the response field is based on experimental data showing that at the center of A1, cortical neurons have an excitatory center and symmetric inhibitory side bands, but moving away from the center of A1, the response field becomes more asymmetric with stronger inhibitory side bands in one direction [68].

The output of the cortical model is a multi-resolution representation simulating the response properties of cortical neurons. Performing this cortical analysis on the auditory spectrum can provide us with information on spatial and temporal modulations, both of which are important in auditory perception [2, 60, 9].

This cortical model has been used successfully in many different auditory processing applications. Some of these applications include speech processing, audio classification, and noise suppression [50, 51, 24, 60].

In the next section, we present a summary of some of the current research in auditory saliency.

2.4 Auditory Saliency Models

Using a saliency model for the auditory domain is still a relatively new area that has not been studied extensively the way visual saliency has. There was no existing work in auditory saliency when we first began researching the idea creating a saliency map

for the auditory domain. Since then, particularly within the past year, there has been continually growing interest in the idea of auditory saliency maps and using auditory saliency models for various applications. Recently, several new auditory saliency models have been proposed, although many of these saliency models are application specific [12, 11, 35] and do not provide a "saliency map". They have been tailored for and tested on a specific task, such as speech recognition, or onset and syllable detection, as opposed to being a general auditory saliency model that can be used for a variety of applications. Concurrent to the development of our auditory saliency model [22] were also two other auditory saliency models. There is the model presented by Kayser et al. [42] and the model presented by Kalinli and Narayanan [40]. These three models comprise the first auditory saliency models to be introduced, and all three provide a auditory saliency map of the input.

The three auditory saliency models are all bottom-up processing models, and they follow the same general architecture as the visual saliency models discussed in the previous section. The model presented by Kayser et al. in [42] is very strongly based on the visual saliency models. It mainly differs from the visual saliency models in its interpretation. The features used are generated and processed in very similar ways to that of the visual saliency model. The model presented in this thesis differs from Kayser's model by the features chosen and the processing of the features to form the saliency map. Our model relies on inhibition of feature maps generated from auditory spectro-temporal receptive fields (STRFs) and also takes into consideration the spectral-temporal orientation which represents the ripple selectivity in A1 of the primary auditory cortex. In addition, all the feature maps are generated using the STRFs, and it does not require normalization to account for differences resulting from various maps being generated by different methods. Our model also differs from the model in [42] in that the features are generated from an auditory spectrogram, which is a time frequency representation modeling early auditory processing [80], as opposed

to an "intensity image."

The model presented by Kalinli and Narayanan in [40] also utilizes an auditory spectrogram. They claim the use of the auditory spectrogram as one of the contributions of their model, but an auditory spectrogram has been used as part of our model, which was presented at various workshops and meetings, well before that publication [21, 20]. The model presented by Kalinli and Narayanan also differs from our model and the one proposed by Kayser in that the model's performance was not evaluated using any psychoacoustical experiments involving results from human subjects. Therefore, how it compares to our model and Kayser's model for selecting salient sounds in general is not known. Instead, the model was evaluated based on performance on a specific speech task referred to as prominent syllable detection. The model presented by Kalinli also differs from Kayser in that there are some additional features that have been added to the feature set.

The first three auditory saliency models all rely solely on bottom-up processing cues. In the last year, there has been increasing interest in looking at how top-down input can be incorporated along with the bottom-up processing models. One such model is a top-down, task-dependent model proposed by Kalinli and Narayanan for use on the same task of prominence detection in speech that the bottom-up model was used for [41]. Another recent model using top-down information was proposed by Coensel and Botteldooren in [12, 13]. The model is an auditory saliency model for use on environmental noise, in particular, transportation noise. It is inspired by the three models discussed above, but also incorporates top-down cues using information about which stream is currently being attended to from multiple streams. There is not yet a general auditory saliency model incorporating both top-down and bottom-up processing components into the same model that can be used for a variety of applications. This may be due to the fact that top-down input tends to be very task dependent. Therefore, how the top-down cues will be added and the type of top-down

feedback that will be used varies based on the specific task, making a general model for use on many different tasks difficult.

In the next chapter, we present the bottom-up computational auditory saliency model developed to identify salient sounds. It is inspired by our understanding of the human auditory system and human perception. It is physiologically motivated, and thus, may more typically represent human audio analysis. One key difference in our model from the other auditory saliency models is that it takes into account that fact that different frequencies can be perceived as louder. We show in Chapter 5 how including this information can make a significant difference in what is considered to be salient by the model.

The model presented in this thesis does well in matching humans' selections of salient sounds based on results from several different experiments. We found a strong correlation between subjects' selections and the model's selections of salient scenes in both a saliency scene comparison experiment presented in Chapter 4 and a video summary experiment presented in Chapter 6. In the saliency scene comparison experiment, subjects were asked to choose the most salient scene from a pair of scenes. The correlation values we found between subjects' responses and our model's responses were slightly higher than the correlation values found in [42] for a similar experiment, although it would be better to evaluate both models using the same set of stimuli and subjects to make a direct comparison.

CHAPTER III

COMPUTATIONAL AUDITORY SALIENCY MODEL

In this chapter, we present the computational auditory saliency model which can be used to identify salient sounds that grab our attention. We begin by first providing an overview of the model. This is then be followed by the more detailed discussions explaining how the feature maps are generated and how the saliency map is formed. Several examples of saliency maps for different auditory stimuli are also presented to show that the model is able to replicate some well-known experimental results. Finally, we conclude the chapter with an introduction to the a MATLAB auditory saliency tool that can be used to easily generate saliency maps for various auditory scenes.

3.1 Computational Auditory Saliency Model Overview

The auditory saliency model presented here is a bottom-up processing model, which performs the same function as the bottom-up, saliency-driven portion of our attentional process. This bottom-up attentional process occurs very quickly and helps us to sort through the information present in a complex scene. It can drive attention by drawing our attention to the sounds that stand out from the rest of an auditory scene. It can also be a method by which the auditory system selects important features and sounds to be passed onto the higher-level stages for further processing. Here, sounds that are not salient are suppressed, thus causing less interference for the salient sounds that will passed to higher-order processes in the brain. In this way, selectively attending to a particular stimulus allows the attentional process to allocate the resources necessary to maintain our focus on the stimuli belonging to the stream of interest [1].

The auditory saliency model presented here uses inhibition of multi-rate, multi-scale cortical feature maps obtained using auditory spectro-temporal receptive fields. The feature maps are used to compute an overall two-dimensional, time-frequency saliency map that identifies what is most salient to observers in a complex scene. The saliency map also indicates the degree of salience for different sounds in a complex auditory scene.

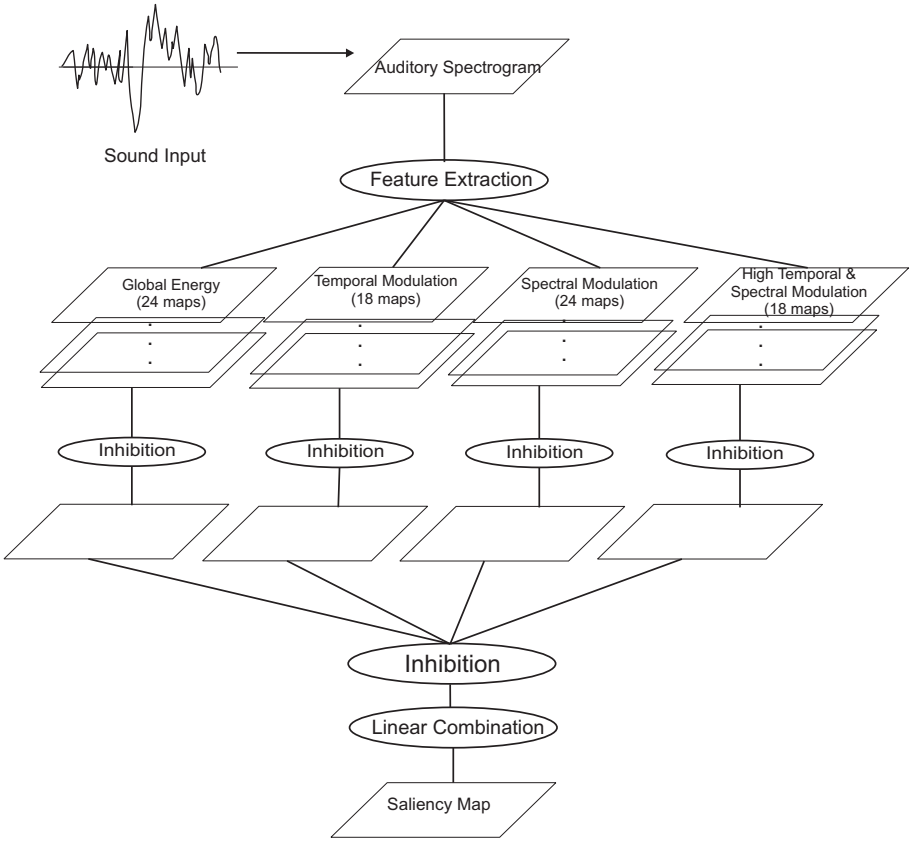


Figure 8: General architecture of the computational auditory saliency model.

The general architecture of the computational auditory saliency model is shown in Figure 8. The sound input is first converted into an auditory spectrogram, which is obtained using a model for early auditory processing that mimics the processing performed by the peripheral auditory system. The feature maps are then generated from the spectrogram and grouped into 4 broad feature classes. Individual feature maps within each class respond to varying rates of change depending on the filter

used. Next, each individual map undergoes inhibition, resulting in the demotion of feature maps with no features that stand out locally. These demoted maps provide minimal contribution to the overall saliency map. The individual feature maps in each of the 4 categories are then combined into a 'global' feature map for each class. Finally, the 4 global feature maps are again subjected to inhibition and summed to form the final saliency map.

The computational model for bottom-up auditory saliency presented here is similar in structure to the visual saliency models, as well as the other existing auditory saliency models [38, 42, 40]. The model can be divided into two main stages:

1. The feature extraction stage where the feature maps are generated

2. The inhibition (or suppression) stage where certain features are promoted or suppressed determining their overall contribution to the final saliency map.

Each of these two stages will be explained in more detail in the next sections.

3.1.1 Generation of Feature Maps

Many different attributes of sound can affect how it is perceived. In selecting the feature set for the auditory saliency model, we wanted to use features similar to those our brain potentially uses in processing sounds. Two cues used by the auditory system that are known to play an important role in auditory perception are spectral and temporal modulations. Some other features that can also influence how a sound will be perceived are its direction, location, frequency, and intensity. In addition, it is generally accepted that both timbre recognition and pitch perception can significantly influence auditory perception [71].

In the previous chapter, we introduced a cortical model, proposed by Wang and Shamma in [77]. It models the responses areas of cortical neurons and extracts spectral and temporal modulation information. We selected a cortical model ([77, 9]) to use in generating the feature maps, since the cortical features are thought to be

similar to those the brain uses in the central auditory system. The cortical features capture many of the attributes of sound that are important in perception. Recent experiments support this as it was found for speech that the cortical representation in the posterior temporal lobe provided not only information on the acoustical properties of the speech, but it was also strongly related to how the speech was perceived [49, 6].

Timbre is one attribute that can be difficult to define, compared to pitch and intensity, but spectral shape has been found to be an important physical correlate for timbre discrimination [59, 69, 31, 67]. Timbre information is well-represented by the cortical model which simulates the spectral shape analysis performed by the central auditory system. One attribute of sound that the cortical features do not provide information on is the location of a sound. While traditional spatial localization cues, such as the inter-aural intensity difference (IID) and the inter-aural timing difference (ITD), are not included in the feature set of our model, in Chapter 6, we demonstrate how the model can be used on binaural sounds as well. The results of that experiment show the model performs well in selecting salient sounds, not only on monaural sounds, but also for two-channel, stereo input sounds.

As shown in the schematic of the model in Figure 8, we start by first converting the sound input into an auditory spectrogram. The auditory spectrogram is a time-frequency representation modeling early auditory processing [80]. More details on the implementation of the early auditory processing model used to generate the auditory spectrogram can be found in the previous chapter. Here, a filter bank of 128 constant-Q bandpass filters is used to model the frequency analysis done by the cochlea. The center frequencies of these filters are uniformly spaced in frequency on a logarithmic axis. Using an auditory spectrogram on the sound input allows us to create an "auditory image" analogous to the static images that are used as the input for visual saliency models.

In the next step, we generate the feature maps using the cortical model in [77]. The

auditory spectrogram is decomposed into its spectral and temporal components using a bank of spectro-temporally selective filters, whose impulse responses are represented by 2-D Gabor functions. By doing this, we obtain features that estimate the spectral and temporal modulations. Each filter has a spectro-temporal receptive field (STRF) centered at a different center frequency (CF). In addition, the receptive fields also take into account the response of cortical neurons to spectral shape, bandwidth, movement, and directionality [77].

The feature maps are obtained by performing a convolution of the auditory spectrogram with each STRF filter. This cortical output is a multi-dimensional representation varying along four different dimensions: time, frequency, rate, and scale. Here, the rate corresponds to the center frequency of the temporal filters, and the scale corresponds to the center frequency of the spatial (frequency) filters.

Receptive fields with varying rates and scales were chosen in order to capture different aspects of the sound. Using a range of rates and scales leads to information on the temporal and spectral modulations, respectively. For example, some of the filters used respond to rapid changes while others respond to slower changes. The filters also differ in how widely or narrowly tuned they are.

Fourteen temporal filters (rates) from ± 0.5 to $\pm$ 32 Hz, and six spectral filters (scales) from 0.25 cycles/octave to 8 cycles/octave were used. For each scale and rate, we generate one feature map, resulting in a total of 84, 2-D (time/frequency) feature maps. The map size will vary depending on the length of the input auditory stimulus. For example, since there are 128 frequency channels, for a 1 second stimulus, the maps are of size 125 x 128.

Figure 9 demonstrates how different aspects of a sound are captured in different feature maps due to the response of the filters. In Figure 9 (b), we show a sample of 6 of the 84 feature maps for the auditory scene which has auditory spectrogram shown in Figure 9 (a). The sample scene consists of an upward and downward chirp that is

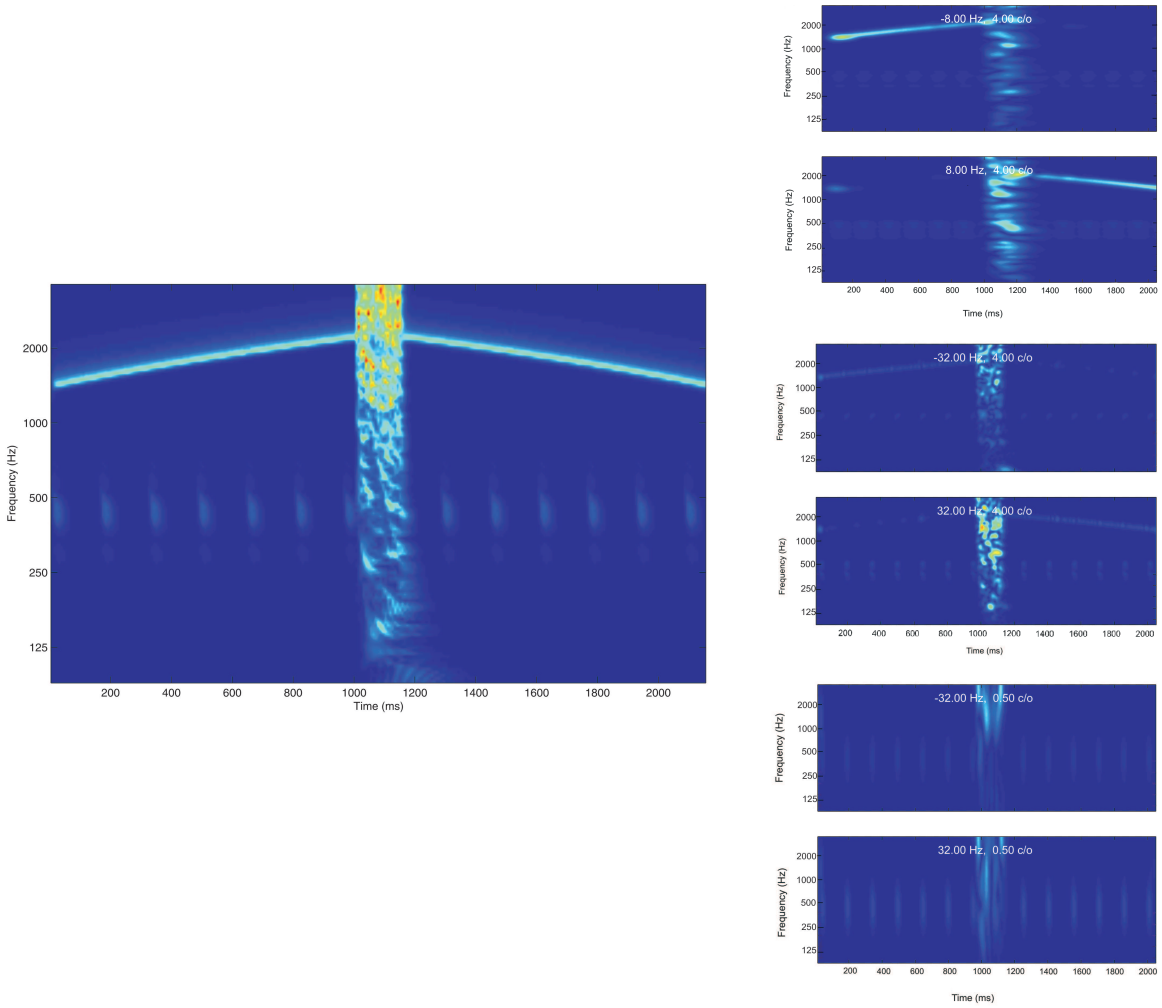


Figure 9: Example showing how feature maps capture different aspects of sound. a) Auditory spectrogram for a sample stimulus b) Six of the 84 feature maps for the sample stimulus.

interrupted by a noise burst. Some clicking noises can also be heard throughout the duration of the scene. Here, we see that the top two feature maps capture the chirp, with one focusing on the upward sweep and one on the downward sweep. Similarly, the middle two feature maps focus on the noise, while the last two feature maps capture the clicks.

Once all the feature maps have been generated, they are divided and grouped into four different feature classes, depending on the rate and scale of the filter used to

obtain the map:

1. A set of feature maps (24 maps) focusing on the overall energy distribution
2. A set of feature maps (18 maps) focusing on the temporal modulations
3. A set of feature maps (24 maps) focusing on the spectral modulations
4. A set of feature maps (18 maps) focusing on areas of both high temporal and spectral modulations.

3.1.2 Formation of Saliency Map

We now have 2-D (time/frequency) feature maps that were generated using the spectro-temporal receptive field filters. The next step is to perform inhibition on each individual feature map. The goal of the inhibition stage is to promote feature maps containing features which stand out locally on the map and inhibit or suppress feature maps without any prominent local peaks. We want to promote maps that have areas of high activity (large global peak) compared to rest of the map, so that these maps make a larger contribution to the overall saliency map. In order to achieve this, each feature map, M_i, is scaled by a factor, D_i. The new scaled feature maps will be denoted M_i^*.

$$M_i^* = D_i \cdot M_i \tag{5}$$

$$D_i = (G_i - \overline{L}_i)^2 \tag{6}$$

where G_i and $\overline{L}_i$ are defined as follows:

$$G_i = \text{Global peak of map i}$$

$$\overline{L}_i = \text{Average of all other local peaks on map i.}$$

Scaling each individual feature map by this factor, D_i, promotes maps with a large global peak compared to the rest of the activity on that map. This is demonstrated

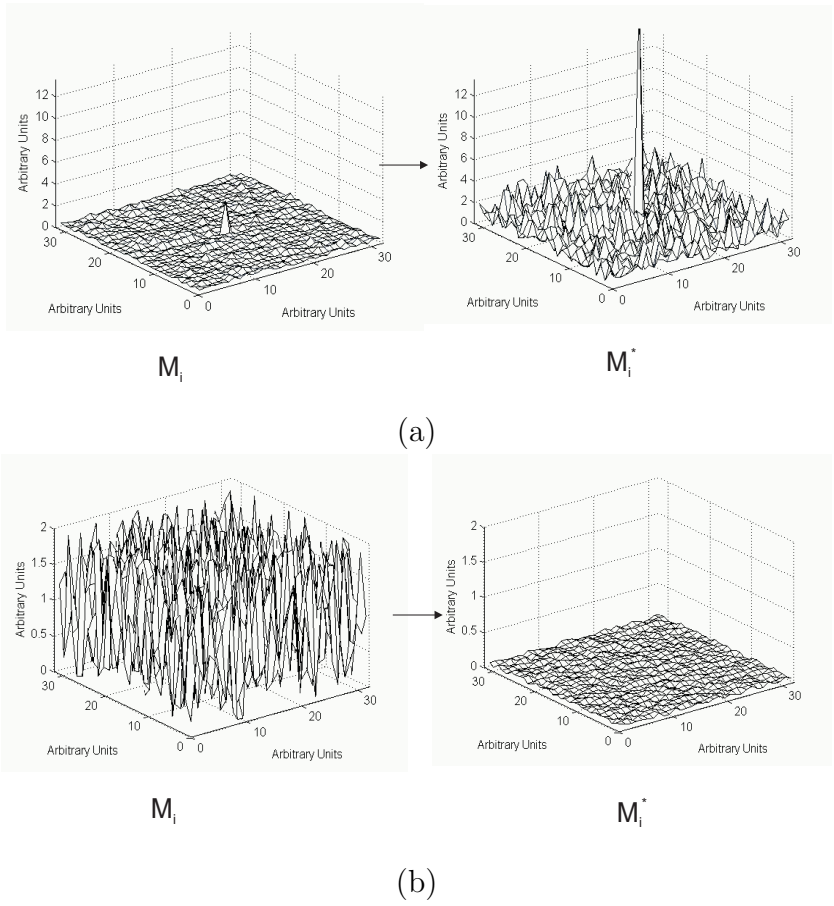


Figure 10: Inhibition of feature maps. a) Feature map with single prominent peak is promoted b) Feature map with many peaks and no prominent peaks is suppressed.

in Figure 10 (a). Conversely, feature maps with high activity everywhere on the map are suppressed (Figure 10 (b)). Using this method, we preserve the general shape of each feature map, while ensuring that feature maps with prominent peaks make larger contributions to the overall saliency map. After this scaling, the feature maps in each respective category are combined to form one global feature map for each category. Each of the four global feature maps are scaled again by D_i before being summed to form the final saliency map.

Several variations of the auditory saliency model were investigated. One variation was to perform the inhibition stage locally rather than globally. We wanted to consider using the local inhibition, since auditory percepts can be more greatly influenced or affected by other auditory events or cues that are occurring closer in time or frequency.

For the local inhibition, the feature extraction stage remains the same as previously outlined. The difference is that once the feature maps have been obtained, each feature map is divided into several non-overlapping 2-dimensional areas. These 2-D areas cover approximately 200 ms in time and 1/3 octave in frequency. In order to retain the peaks for each local area, we take an average of the signal in that area, which is then subtracted from the signal. This is then followed by the previously described method of scaling used to promote the maps with prominent peaks. The differences between the two versions of the model, which we will refer to as Model 1 (global) and Model 2 (local), are summarized in Table 1.

There are advantages and disadvantages to each of the two variations of the auditory saliency model, and depending on the application, one might find one model more useful than the other model. In the saliency scene comparison experiment presented in the next chapter, we find a fairly strong correlation between the responses for both models with the subjects' responses, but there was a stronger correlation to Model 2, where the local inhibition was performed. Although Model 2 performed better on this particular experiment, one advantage of Model 1 is that performing the inhibition globally preserves the shape of each feature map. In Model 2, performing the inhibition locally can cause large parts of the overall saliency map to be lost. In some instances, the saliency map can be comprised of only a few scattered points. In the scene saliency experiment, the loss of this information from the saliency maps for each auditory scene was not important as we were only interested in the most salient point on the map to determine the saliency of that particular scene. In tasks where we would like to use the entire saliency map and where the map is visually important, then Model 1 will perform better.

Table 1: Summary of Model 1 (global) and Model 2 (local) differences.

Model	Description
1	Promotes or inhibits entire feature maps using scaling by D_i
2	Uses local inhibition and then scaling by D_i

3.2 Saliency Map Examples

Now that we have presented the computational auditory saliency model, we will show some examples of different saliency maps. These examples show that the model is able to replicate some known psychoacoustic experimental results.

The auditory system is well-versed in change detection. From an auditory scene analysis perspective, older sounds that remain relatively constant and unchanging will tend to become part of the background. As they fade into the background, these sounds become less salient, while new sounds will stand out from the background and are more salient [4].

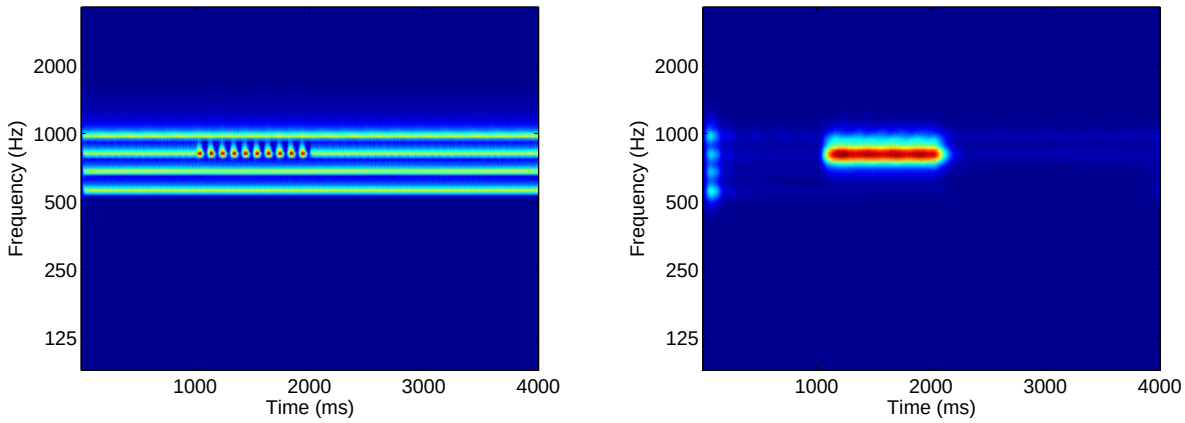
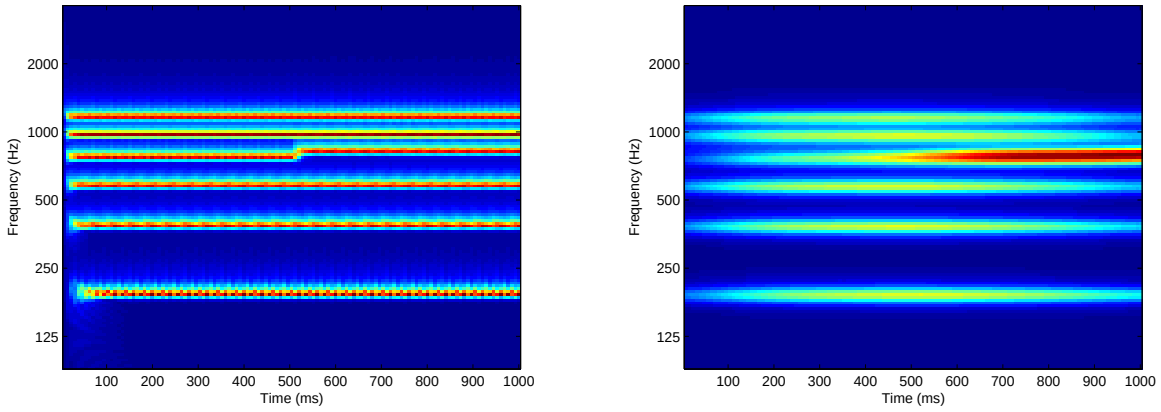


Figure 11: Auditory saliency map for amplitude modulated tone. a) Auditory spectrogram of amplitude modulated tone b) Auditory saliency map of amplitude modulated tone shows the modulated part is salient.

Two change detection examples are presented here where new sounds stand out

perceptually from the other unchanging "background" components. In the first example, there is a tone complex where part of one of the four tones in the complex is amplitude modulated. This particular auditory scene is also used as the peripheral task in the dual task experiment that will be discussed in the next chapter. It is expected that the modulation is salient, since it is known that an amplitude or frequency modulating a tone will make it stand out perceptually from the other unchanging tones, as the modulation helps us to perceptually segregate it from the rest of the complex [53, 48, 54]. In Figure 11, the auditory spectrogram and corresponding saliency map for this example are shown. From the saliency map, we see that the modulated part is, as expected, what is most salient and the rest of the tone complex except for the onsets are suppressed.



(a) (b)

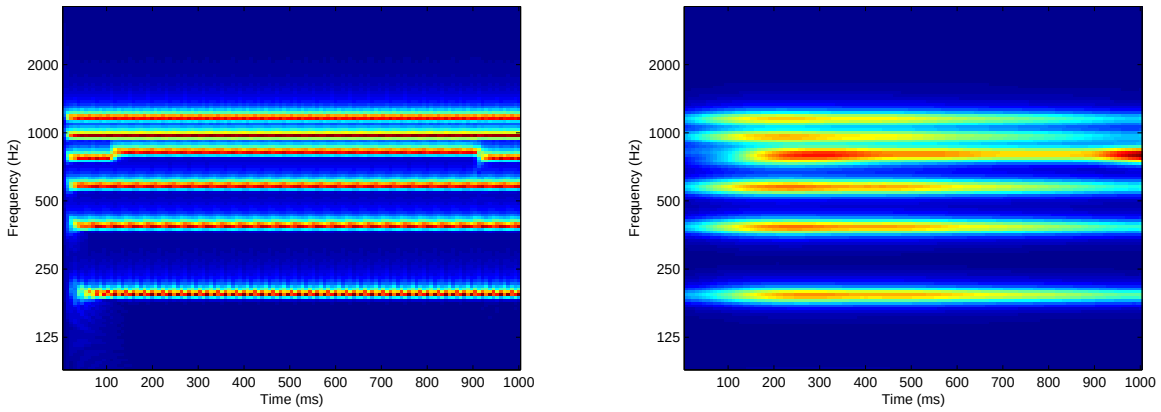
Figure 12: Auditory saliency map for mistuned harmonic. a) Auditory spectrogram of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz for 0.5 s b) Auditory saliency map of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz for 0.5 s.

The second change detection example shows the well-known experimental result of hearing out a mistuned harmonic from the rest of the complex [55, 32, 33]. In this example, the 4th harmonic of a 200 Hz tone is mistuned by 48 Hz for 0.5 s causing the mistuned harmonic to pop-out. The mistuned harmonic is heard separate from the rest of the complex, since it stands out perceptually. The auditory spectrogram

and saliency map for this example are found in Figure 12. The saliency map shows that the mistuned harmonic is salient and stands out from the rest of the complex. One interesting thing to note is that when the mistuned harmonic is in the middle of the signal, the duration of the mistuned harmonic can greatly influence how salient it is to a listener. For example, if we increase the duration of the mistuned harmonic to 0.8 s, such as shown in the auditory spectrogram in Figure 13, the mistuned harmonic will gradually become less salient to listeners. We see in Figure 13 (b), the saliency map for this example indicates that the signal is most salient and the onset and offset of the mistuned harmonic. Specifically, we see that the first 200-300 ms of the mistuned harmonic is salient, and then the degree of saliency gradually decreases. Then for the last 100 ms of the sound, when the harmonic is no longer mistuned, it again becomes very salient. This is what we would expect, since the longer the duration of the mistuned harmonic, the more likely it is to blend into the background. Therefore, the mistuned harmonic gradually becomes less noticeable, and therefore less salient to listeners. It then becomes salient again when we hear the transition going from mistuned to no longer mistuned. Similarly, if we mistune the harmonic for a very short duration (50 ms), then this mistuned harmonic will be extremely salient compared to the rest of the tone complex. The saliency map for this example, seen in Figure 14 (b), verifies this as we see the mistuned harmonic is salient while the rest of the tone complex has been suppressed.

A third example shows two 250 ms, 2 kHz tones in white noise. In Figure 15, the auditory spectrogram and saliency map for this example show that the noise is suppressed while the two tones are emphasized. From this, there are some possible applications of auditory saliency in the area of noise suppression that could be explored.

In one final example, shown in Figure 16 (a) and (b), a gliding tone is interrupted



(a) (b)

Figure 13: Auditory saliency map for mistuned harmonic. a) Auditory spectrogram of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz for 0.8 s b) Auditory saliency map of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz for 0.8 s.

by a noise burst. From the continuity illusion, it is expected the tone will be perceived as continuing through the noise [78, 4]. The saliency map for this example (Figure 16 (b)), reflects the perceptual continuity of the tone through the noise, and this continuity is indicated as being salient.

The four examples presented above are used to demonstrate how the saliency maps from our model is able to predict several known experimental results from auditory scene analysis.

3.3 Auditory saliency map tool

There are many possible applications and uses for auditory saliency maps in computational auditory scene analysis and other areas. In order to easily generate saliency maps for use in the experiments and applications presented in the following chapters, we created an MATLAB GUI interface for our auditory saliency model. In this section, we discuss the MATLAB GUI created for the computational auditory saliency model that can be used to obtain auditory saliency maps for various auditory stimuli. The GUI, shown in Figure 17, takes an input sound file (.wav) and processes it into

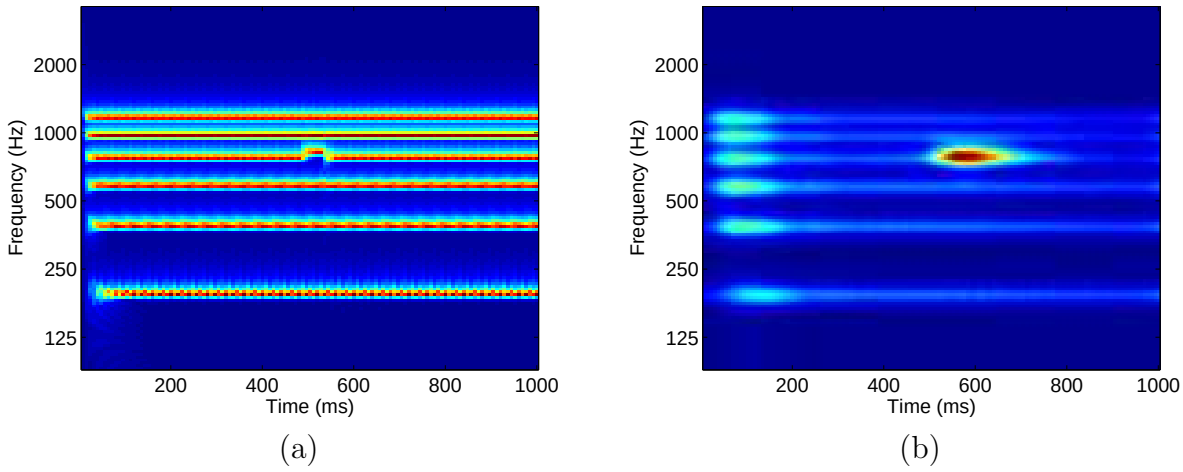


Figure 14: Auditory saliency map for mistuned harmonic. a) Auditory spectrogram of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz for 0.05 s b) Auditory saliency map of 200 Hz tone, with 6 harmonics, where the 4th harmonic is mistuned by 48 Hz for 0.05 s.

the corresponding saliency map.

The drop down box lists all the .wav files available in the present directory. First, you begin by selecting the sound file for which you would like to generate a saliency map and press load. If new sound files have been added to the directory that are not appearing in the drop down list, the refresh button can be used to update the available sound files in that directory. Next, using the AudSpec button will generate a plot of the auditory spectrogram for the sound file. From here, you want to then use the auditory spectrogram to create your feature maps by pressing the Features button. The next step is to perform the suppression, which as discussed earlier in this chapter can be done either locally or globally. Here, you can select which of the two suppression methods you would like to use, and then press the suppress button. As we find in the next chapters, depending on the stimuli, task, or application, one type of suppression may be more suitable than the other. It should be noted that the local suppression often results in many parts of the saliency map being lost or suppressed. Thus, if a visual of the saliency map is important, it is best to use the global suppression option as it provides a more complete picture of the saliency map.

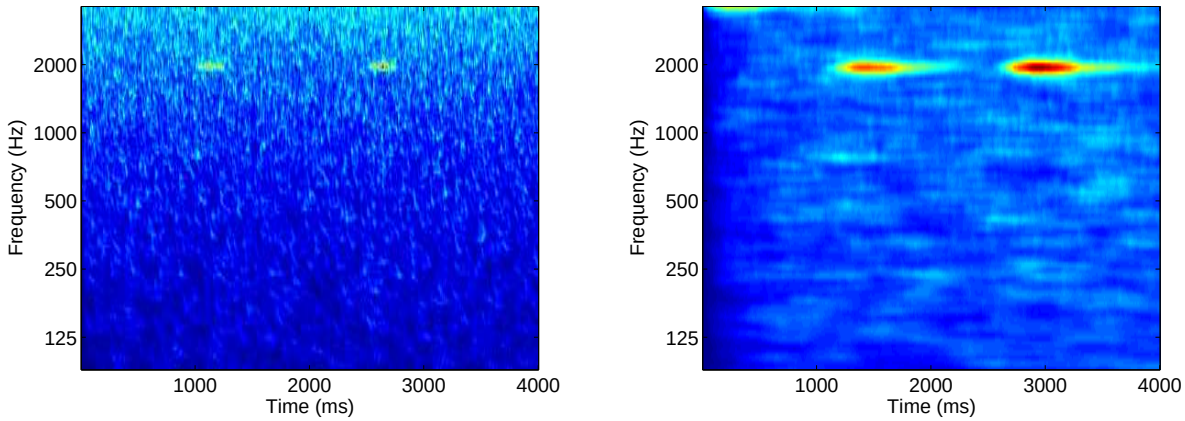


Figure 15: Auditory saliency map for two tones in white noise. a) Auditory spectrogram of two 2 kHz tones in white noise b) Auditory saliency map of two 2 kHz tones in white noise.

Finally, pressing the saliency map button will return a plot of the auditory saliency map. The GUI also displays the value of the most salient point on the map, which may be useful for various applications. For example, we use this value to evaluate our model to human subjects in the scene saliency experiment presented in next chapter.

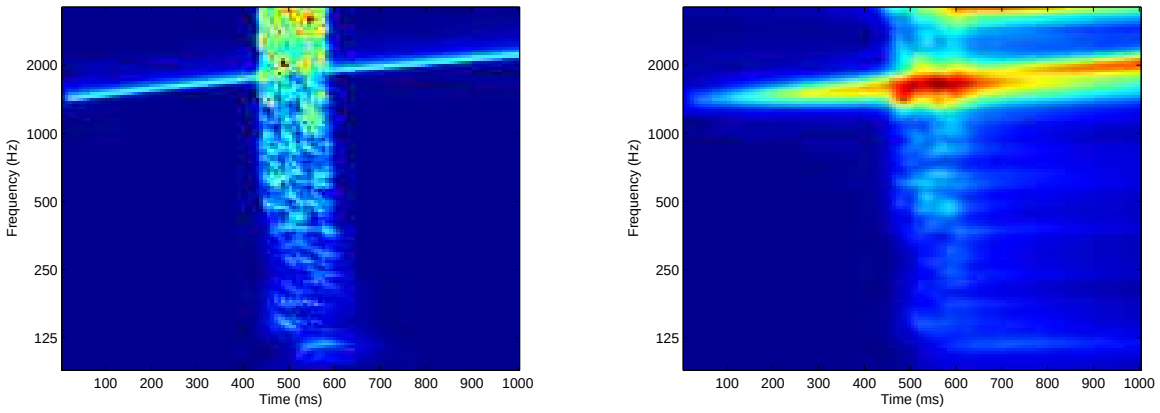


Figure 16: Auditory saliency map for the continuity illusion. a) Auditory spectrogram of gliding tone interrupted by a noise burst b) Auditory saliency map of gliding tone interrupted by a noise burst.

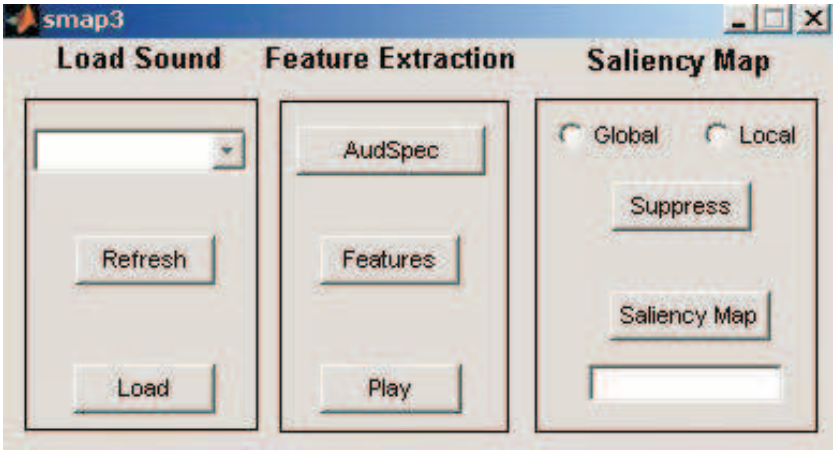


Figure 17: Tool for creating auditory saliency maps.

CHAPTER IV

EVALUATING AUDITORY SALIENCY

There is no easily trackable physical correlate that can be used to evaluate auditory saliency the way eye-tracking is often used to evaluate visual saliency. In this chapter, we focus on the issue of how to evaluate or measure auditory saliency and what types of experiments can be used to do this. We found that we could approach this problem from two different perspectives. In this chapter, we present two types of experiments performed that are appropriate for evaluating auditory saliency and the performance of our computational auditory saliency model. The first experiment is a dual task experiment, and the second experiment is a paired comparison experiment that we refer to as the saliency scene comparison experiment.

4.1 Dual Task Experiments

One way to determine the saliency of a particular stimulus is by looking at it's ability to grab a listener's attention, even when their attention is currently directed elsewhere. Dual task experiments involve dividing a subject's attention by asking them to perform two tasks simultaneously. The subject is asked to attend to a primary task while a secondary task is occurring at the same time in the background. In terms of auditory saliency, it is the subject's performance on this secondary task that will determine the saliency of a particular auditory stimulus. Dual task experiments are a good method for evaluating bottom-up auditory saliency, since salient stimuli are defined as those stimuli that can be noticed without attention.

4.1.1 Subjects

Results were obtained from 18 (9 female, 9 male) undergraduate students at the Georgia Institute of Technology under IRB approval. The students received course credit for participation in the experiment. Normal hearing was determined by self-report. Subjects were informed about the general aim of the experiment, but were naive to the exact purpose of the study.

Subjects listened through headphones (Sennheiser HD 280 pro) to auditory stimuli presented using MATLAB and then entered their responses into a MATLAB GUI interface shown in Figure 18.

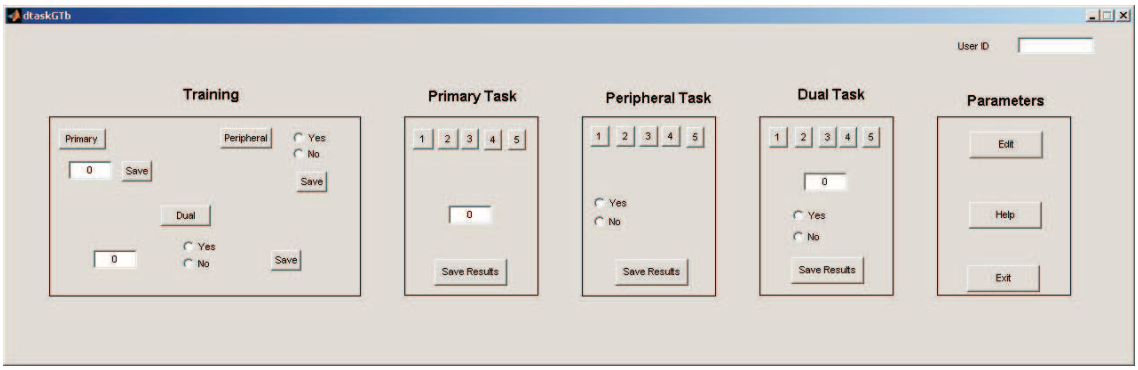


Figure 18: Dual task experiment interface.

4.1.2 Experimental Procedures

As is typical in all dual task experiments, subjects were asked to perform two different tasks, a primary task and a secondary or peripheral task, simultaneously. The primary task is the task which will require the subject's attention, while the performance on the peripheral task will be the one that will indicate whether or not a particular auditory stimulus is salient.

Subjects were given a short training prior to beginning the experiment. In the training, they were given short sample trials of the tasks that they would be asked to perform. In particular, we wanted to accustom the subjects to the primary task, as the task can be extremely difficult to perform when hearing it for the first time.

After the short training, they were then asked to perform each of the primary and peripheral tasks separately before completing both tasks simultaneously in the dual task.

The primary task stimulus for this experiment was a random sequence of twenty-five 100 Hz and 200 Hz tones, each 250 ms long with a 2 ms pause between each tone. Subjects were asked to count the number of low (100 Hz) tones present for each sequence.

For the peripheral task, the stimulus consisted of a target sound and some interferers, and subjects needed to identify whether or not the target was present. The stimulus was a tone complex made up of four tones (not harmonic) centered at Bark filter center frequencies of 570 Hz, 700 Hz, 840 Hz, and 1000 Hz; the target was a tone in the four tone complex that was amplitude modulated to varying degrees. The tone complex for the peripheral task was designed so that there were no harmonic relationships that could generate a pitch percept. Five different modulation depth levels were tested. The entire stimulus was four seconds long but the modulated tone was only modulated for one second and modulation onsets and offsets were gradual over 35 ms.

Several considerations were made in setting up the primary and peripheral tasks. The two tasks were chosen such that they were not spectrally co-located to avoid masking problems. In the peripheral task, the tone complex was chosen such that there were no harmonic relationships. This was done to ensure a missing fundamental would not interfere with the primary task. The spacing of the tones also prevented beat tones and avoided any two tone masking issues.

In dual task experiments, the difficulty level of the primary and secondary tasks chosen is important. Performance on each of the individual tasks and the ability to perform both tasks simultaneously is dependent on the task difficulty and the amount of attention that is required by each of the individual tasks.

For the primary task, the task difficulty is influenced by a number of factors including the number of tones used, the duration of the tones, and the duration of the pause between each of the tones. The level selected for the task difficulty is extremely important as the primary task is the task that demands or require the subject's attention. If the primary task is too easy, then subjects would not need to actively attend to the task in order to achieve high performance. In this case, they would be able to also attend to the peripheral task which would likely result in high performance on both tasks. This would then provide us no information about the salience of the test stimulus.

For this reason, performance on the primary task when performed alone was calibrated so that the performance was between 80-90%. Making the primary task difficult enough when the task is performed alone will ensure that subjects need to actively attend to the primary task when they are asked to perform both tasks simultaneously. If a subject does shift his or her attention away from the primary task, it should be reflected by a large drop in the primary task performance.

In the dual task experiment, the primary and peripheral tasks occur simultaneously. In this experiment, both the primary task and the peripheral tasks were chosen to be auditory tasks. We selected two auditory tasks as we wanted to be consistent with the idea that salient sounds will divert or interrupt our attention from other sounds or stimuli that are present in an auditory scene. Using this dual task paradigm, a sound is defined as salient depending on whether or not a subject's performances on the primary and secondary tasks are affected. If performance remains high on both tasks, this means the sound is salient as this indicates that the sound is noticeable in the peripheral task even when attention is directed elsewhere. Subjects performed 5 trials of each task.

4.1.3 Primary Task and Peripheral Task Trade-offs

As previously mentioned, task difficulty for each of the primary and peripheral tasks is important in dual task experiments as it will affect a subject's ability to perform the two tasks simultaneously.

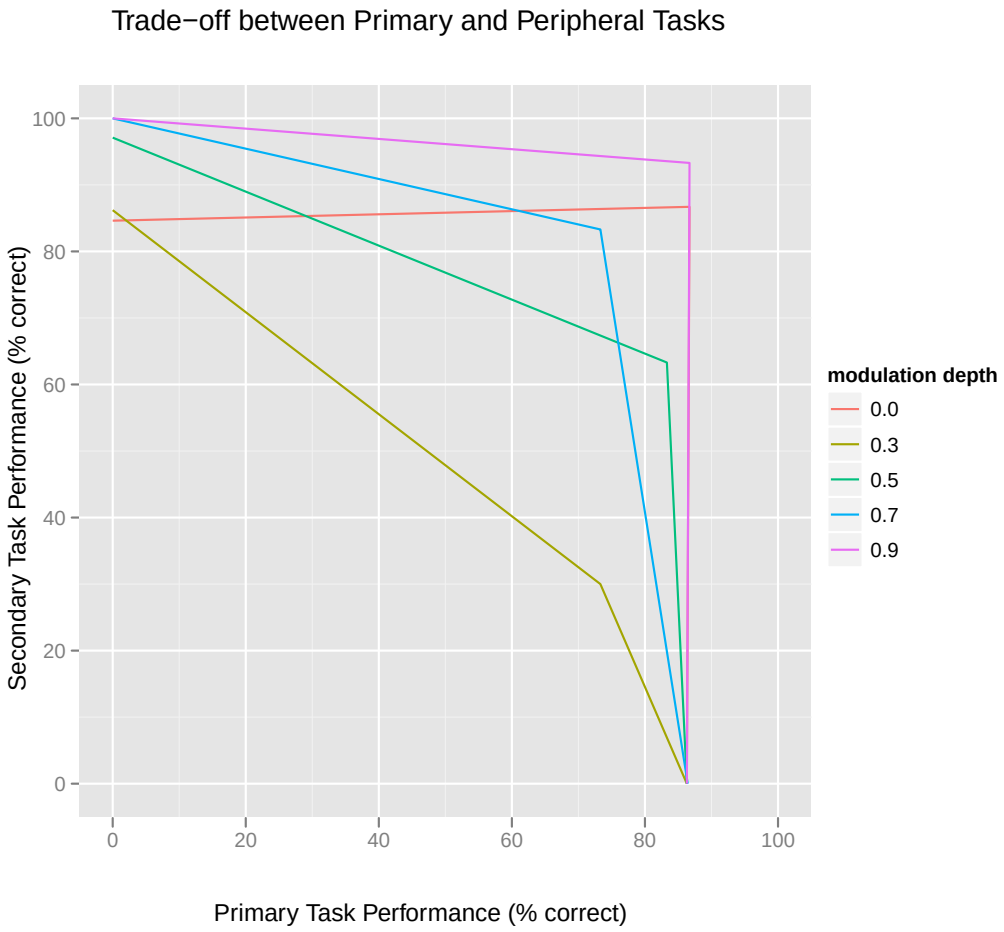


Figure 19: POC curve showing trade-off between primary and peripheral tasks.

Performance operating characteristic (POC) curves can be used to evaluate the trade-off between the two tasks. The POC curve for this experiment is shown in Figure 19. On the far right of the plot, where there are the lines for 0.9 and 0, moving up and down vertically, this is the area of high performance on the primary task. Performance in this region shows that the subjects' attentions were completely focused on the primary task. Moving along the top of the plot, going across horizontally, is the

area of high performance on the peripheral task. Performance in this region, during the dual task, indicates that the peripheral task takes priority over the primary task. The top right corner of the plot is the area indicating high performance on both the primary and peripheral tasks. Salient sounds will have dual task performances in this region. Here, subjects are completely focused on the primary task, but they are still able to achieve high performance on the peripheral task. The fact that the peripheral task can be noticed while the subjects' attentions were focused elsewhere shows the peripheral task sound is salient.

In Figure 19, the 5 curves each represent one of the different modulation depths that are used in the peripheral task. There are three points on each curve, which represent the average performance on each task for all subjects. One point represents performance on the primary task alone, one point is for the peripheral task performance alone, and one point is for the performance on both primary and peripheral tasks when the two tasks are presented simultaneously. The exact values can be found in Tables 2 and 3. These different curves also correspond to the task difficulty. For example, at modulation depth of 0.9, the signal is highly modulated. We would expect this to be salient and easily noticed even while attention is directed elsewhere. As the level of modulation decreases, we expect the modulation to gradually become less salient, therefore, making it more difficult to notice the peripheral task stimulus without attention, leading to a drop in the peripheral task performance when the two tasks are presented simultaneously.

The decrease in saliency that is expected as we decrease the amount of modulation is reflected in the shape of the POC curve. We can see that there is a higher degradation in performance on the peripheral task as the task difficulty increases. For instance, at modulation depth 0.3, there was the largest degradation in performance. At this modulation depth, the difficulty of the peripheral task has increased to the point where subjects were not able to perform both tasks simultaneously. Thus, when

attention is engaged on the primary task, for this modulation depth, the stimulus in the secondary task could not be noticed. From this, we conclude that the stimulus was not salient, since subjects were not able to notice it without attention.

4.1.4 Results and Discussion

The results summarizing overall performance on the peripheral and primary tasks when they were performed alone can be found in Table 2. The "% correct" is an average of the number of correct responses from all subjects. For the primary task, we allowed for a margin of error of one tone. Therefore, a response was counted as a correct response if the subjects were able to come within one tone of the actual tone count.

The results of the dual task are shown in Table 3. For all modulation depths, performance on the primary task remained relatively high. This indicates that the subjects were attending to the primary task as asked. There were some slight drops in performance at two of the modulation depths, but the drops were not large enough to indicate that the subjects were not attending to the primary task. We would expect to see a much larger degradation in performance if the subjects were no longer attending to the task. From the results, as expected, the highly modulated signal (modulation depth 0.9) was extremely salient. From the table, we can see that for this modulation depth, performance on both the primary and peripheral tasks remains high. The high performance on the primary task shows that subjects kept their attention focused on the primary task, and the high performance on the peripheral task indicates that the target in the peripheral task could be noticed while attention was directed elsewhere. Therefore, at this high modulation depth, the modulated target was noticeable without attention and is considered extremely salient.

As the modulation depth decreases, the salience of the modulated signal also decreases. At a modulation depth of 0.7, performance on both the primary and

peripheral tasks drops slightly, showing that while it may still be considered salient, it is less salient than at a modulation depth of 0.9. As we continue to further decrease the amount of modulation, we see that the salience continues to decrease as well. At a modulation depth of 0.3, the signal is not salient at all. Here, peripheral task performance is only 30% compared to 86.2% when the peripheral task was performed alone. Performance on the primary task has also dropped slightly. This large drop in peripheral task performance indicates that the sound is not salient, since subjects were not able to notice it while their attention was directed elsewhere. At modulation depth 0.5, the peripheral task performance was also greatly reduced in the dual task (63.3% compared to 97.1%), although the primary task performance remained high. As expected, the case where there was no modulation was easily observed by subjects.

Table 2: Overall performance on the primary and peripheral tasks when performed alone.

Task	**% Correct**
Primary task	86.3
Peripheral task 0.9	100
Peripheral task 0.7	100
Peripheral task 0.5	97.1
Peripheral task 0.3	86.2
Peripheral task no modulation	84.62

Table 3: Performance on the primary and peripheral tasks when performed simultaneously in the dual task.

Modulation Depth	**Primary task % Correct**	**Peripheral task % Correct**
0.9	86.7	93.3
0.7	73.3	83.3
0.5	83.3	63.3
0.3	73.3	30
0	86.7	86.7

A look at the saliency maps generated by our computational auditory saliency

model shows that the saliency maps for the different dual task scenarios are consistent with the results found in the experiment. In Figure 20, the saliency map for the dual task experiment when modulation depth of 0.9 was used is shown. At this high modulation depth, the experimental results showed that the modulated signal was extremely salient. The saliency map is consistent with these results, and we can see on the map that the modulated part stands out from the rest of the tone complex in the peripheral task. In addition, the saliency map also indicates that the modulated part is more salient than the primary task as well, which reinforces what was observed in the experiment, as the modulated tone was easily noticeable while attention was focused on the primary task.

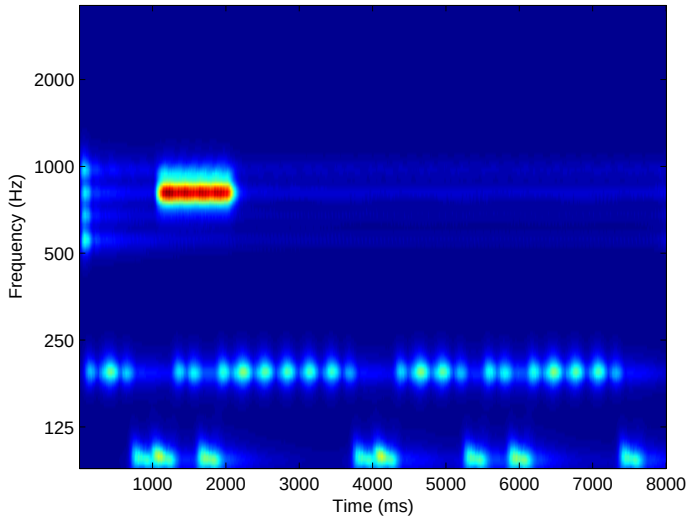


Figure 20: Auditory saliency map for dual task at modulation depth 0.9.

At a modulation depth of 0.7 (Figure 21), the modulation in the peripheral task still remains somewhat salient, but it is less salient than at a modulation depth of 0.9. In addition, the saliency map here shows that the the primary task on this saliency map is more salient than on the saliency map for modulation depth 0.9. Again, this is consistent with the results obtained from the experiment, where performance here dropped slightly compared to modulation depth 0.9, indicating it is less salient.

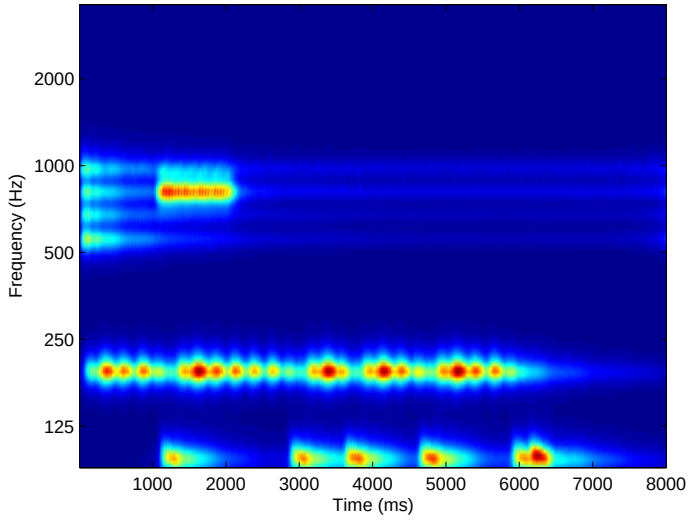


Figure 21: Auditory saliency map for dual task at modulation depth 0.7.

Finally, we can look at the saliency map for modulation depth of 0.3. In the dual task experiment, it was found that at this modulation depth, the modulated signal was not considered salient. The saliency map reflects this finding, and in Figure 22, the peripheral task has been almost completely suppressed leaving nothing, including the modulated tone, salient in the peripheral task. On the saliency map here, it is actually parts of the primary task that are most salient.

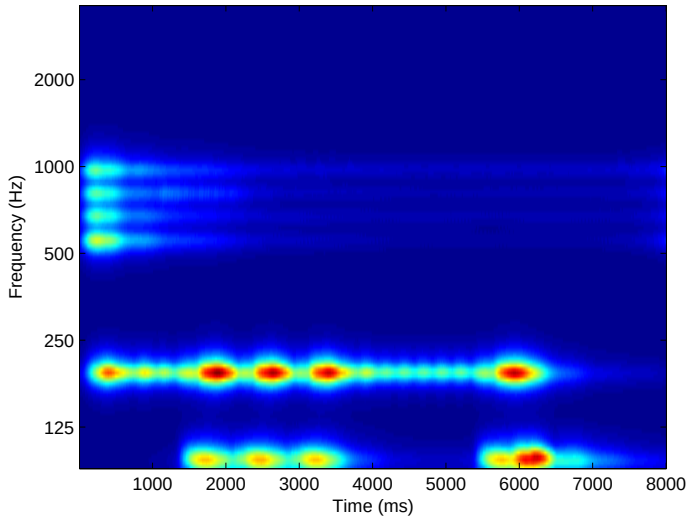


Figure 22: Auditory saliency map for dual task at modulation depth 0.3.

Dual task experiments are a very useful technique for determining the salience of a particular stimulus. The advantage of the dual task setup is that it provides strong evidence for whether or not a particular stimulus is salient. It is also a very focused experiment, which allows the experimenter to determine the degree of salience as well as where the boundary between salient and not salient lies. With dual task experiments, the experimenter is able to make one small change at a time to see the effect of that particular change on the salience of a particular stimulus. Using this method, we can determine the degree of change needed for a target to become salient. For example, in the experiment we performed, we saw the amount of modulation that we needed to use on the tone to make the modulated tone salient.

Some disadvantages of using dual task experiments are that it would be difficult and time-consuming to perform the experiment for a large range of stimuli. For example, in this one experiment, we were only testing the effects of amplitude modulation of a tone in a tone complex on salience, so running the whole experiment provides us with information only on the salience of this one particular type of stimulus. The experiment presented in the next section will allow for comparison of the saliency for a larger variety of different sounds.

4.2 Saliency Scene Comparison Experiment

The primary goal for the auditory saliency model is for it to be able to select the sounds that are perceptually salient to observers in an auditory scene. After this goal is achieved, we can then consider tailoring the model to the specific applications that it would be used for. In this way, the model could potentially exceed human performance on a specific task. In order to evaluate the model performance with regards to the primary objective, we need to see if the model can successfully match human performance in identifying salient sounds from various auditory scenes. Thus, a second way we can evaluate auditory saliency is to look at which sounds humans

consider salient when presented with a variety of different sounds. We can then directly compare our model's selections of salient sounds to that of human subjects.

In this section, we discuss the saliency scene comparison experiment, which uses pairwise presentation of different auditory scenes in order to determine which sounds human subjects find salient. In this experiment, subjects were also asked to give subjective ratings on the saliency for different auditory scene pairs. The same scene pairs were then presented to our computational auditory saliency model for comparison. Using this technique, we can evaluate the model's performance by determining how well it does in selecting the scenes that subjects found salient.

4.2.1 Subjects

Results were obtained from 14 (10 female, 4 male) university students of the Georgia Institute of Technology with IRB approval. Normal hearing was determined by self-report. Subjects were informed about the general aim of the experiment, but were naive to the exact purpose of the study.

Subjects listened through headphones (Sennheiser HD 280 pro) to auditory stimuli presented using MATLAB and then entered their responses into a MATLAB GUI. The response interface for this experiment can be seen in Figure 23. The sound card on the PC used is AD1981A AC'97 SoundMAX Codec (Full-duplex with variable sampling rates from 7040 Hz to 48 kHz with 1 Hz resolution). This same equipment and set-up is used for all subsequent experiments presented in this thesis. The experiment took each subject approximately 50 minutes to complete.

4.2.2 Experimental Procedures

Each subject was presented with a total of 162 scene pairs created from 50 unique target sounds. The auditory scenes consisted of a target (one second) sound on a background (4 seconds). We wanted to test our model on a wide range of stimuli, so the target sounds were randomly selected from a library containing a range of different

types of sounds, including animal sounds, sirens, noise, and music. The background, which was the same for all of the scenes, consisted of white noise combined with a randomly selected sample of the target sounds.

Subjects were presented with each of the different auditory scene pairs. Each scene pair consisted of listening to two 4 second scenes that were separated by a 1 second pause between each of the two scenes. Subjects were then asked to indicate on a MATLAB GUI interface whether the first or the second scene they heard had the most salient element or if the saliency of the two scenes was equal. After selecting which of the two scenes was more salient to them, subjects were also asked to give a rating from 1 to 5 indicating how much more salient they found the scene they selected compared to the other scene. Subjects were told that a rating of 5 meant the scene was much more salient than the other scene, making the decision of which scene was more salient an easy one. A rating of 1, on the other hand, meant that the two scenes were very close in saliency, making the decision of which scene was more salient much more difficult. As the experiment was quite lengthy and repetitive, we divided the experiment up into 6 different tests or segments. The subjects were required to complete each test in its entirety from start to finish once they started it, but they were allowed to take a short break before starting the next segment if they chose to. Each individual test consisted of 27 scene pairs and took approximately 8 minutes to complete.

Two different types of catch trials were also included in the experiment:

- Type 1 - Identical scene pairs where both scene 1 and scene 2 of the scene pair were made up of the same scene
- Type 2 - Duplicate scene pairs where scenes pairs presented earlier in the experiment are presented again later

For type 1 catch trials, subjects should be able to correctly indicate that the two

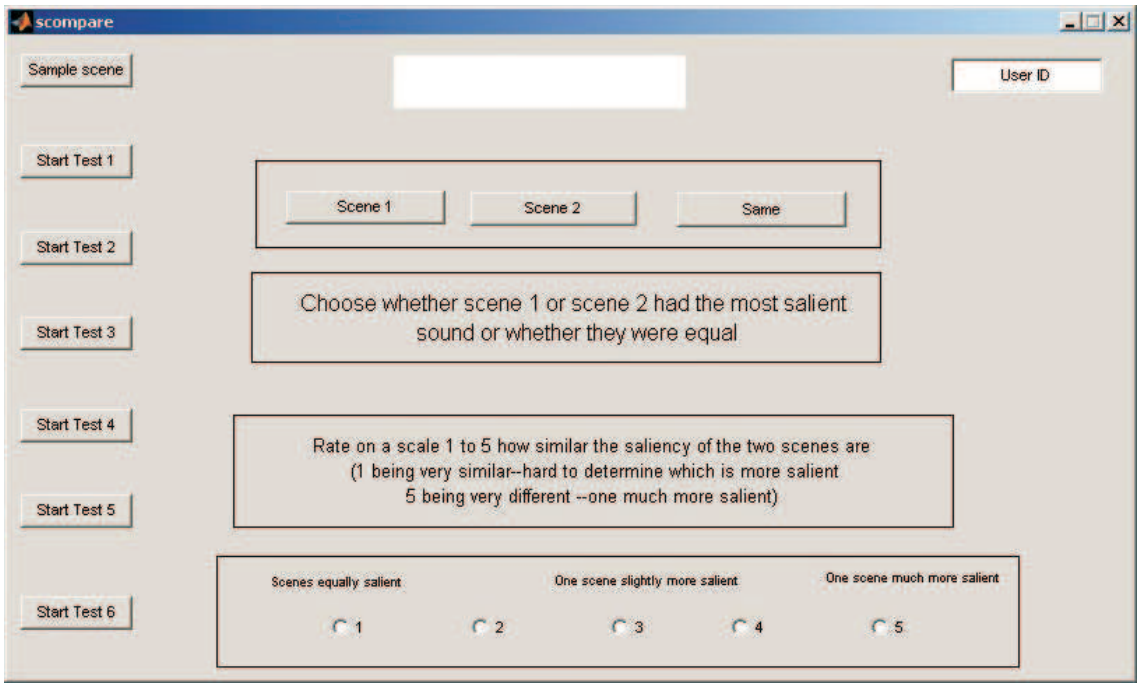


Figure 23: Saliency scene comparison experiment interface.

scenes are of equal saliency. These are used to determine if a subject is properly attending to the task and not just entering random responses. Type 2 catch trials were included to try to get an indication of how consistent a subject was.

4.2.3 Results and Discussion

Results for the correlation of the model's selections and subjects' selections are shown in Table 4. Here, scenes pairs where subjects indicated that the two scenes were equally salient were excluded from the results. This was done as the objective here is to evaluate how the model performs in selecting scenes that are salient to humans. Thus, we are mainly interested by the scene pairs where the subjects considered one scene more salient than the other scene. Additionally, we are particularly interested in the scenes where there is subject agreement on the saliency. These are scene pairs where the majority of the subjects agree on which of the two scenes is more salient. The scenes with subject agreement are not affected by excluding the scene pairs where subjects found both scenes equally salient from the analysis. Removing the scenes

that subjects found equally salient also allows us to make a more direct comparison to our model, since the model always selects one scene as being more salient than the other, unless the two scenes are exactly the same.

Table 4: Correlation between subject and model responses for saliency scene comparison experiment.

	Correlation to	
Subject	**Model 1**	**Model 2**
1	0.7297	0.8725
2	0.4499	0.5297
3	0.4137	0.5309
4	0.7724	0.7856
5	0.0399	0.0131
6	0.0573	0.1079
7	0.3917	0.4680
8	0.4040	0.4147
9	0.5950	0.7014
10	0.5961	0.6895
11	0.4198	0.4198
12	0.6188	0.6753
13	0.6131	0.6300
14	0.5411	0.5519
Average	0.4745	0.5272
Std Dev	0.2166	0.2392

We found a significant correlation between scenes that subjects selected as being salient and scenes that our computational model chose as being salient. All 162 scene pairs were presented to both Model 1 and Model 2. The differences between these two model variations were discussed in the previous chapter and the differences focus mainly on the area on which the inhibition is performed, globally or locally. For Model 1, where the inhibition is performed globally, the average correlation from all subjects was 0.4745 (p=0.08) with a standard deviation of 0.2166 and for Model 2, where the inhibition was performed locally, the average correlation was 0.5272 (p=0.05) with a standard deviation of 0.2392. These results can be found in Table 4.

It is important to remember that saliency can also vary greatly depending on the

top-down input or feedback. Therefore, what is salient to one observer may be very different from what is considered salient by another observer. The computational model we presented in the previous chapter is a bottom-up processing model, and as such, it does not take into account the top-down input. This could be one reason why two of the subjects' responses (5, 6) had almost no correlation with the model's responses. Additionally, subject 6 was also one of the three subjects who had poor performance on the catch trials compared to the other subjects (see Table 7), indicating that they may not have been properly attending to the given task. It is also interesting to note that subjects 5 and 6 along with subject 1 were the only three subjects that performed the experiment in the evening. Several studies have shown that time of day can affect subject performance given the type of task being performed and also given a subjects' own circadian type [27, 28, 47, 56]. Since the task does require a subject's attention, performing the experiment later in the day when they are more likely to be tired or lacking in concentration or patience may affect the results. Later, we present results of an outlier analysis to see if these two subjects should be removed.

As the computational auditory saliency model is a strictly bottom-up processing model, we were particularly interested in comparing the model's performance for scenes where there is some agreed salience among the observers. To determine scenes where there was agreed salience, we looked at the scene pairs where the majority of the subjects were in agreement that one of the two scenes was more salient than the other.

By looking at the scenes where there is subject agreement on saliency, we can eliminate some of the individual variation that may result from top down input which can cause certain types of stimuli to be salient only to particular observers. Out of the 162 scene pairs, there were a total of 118 scene pairs where there was agreed salience. The results for these scenes are shown in Table 5. For these 118 pairs, the results

show a strong correlation between the subject and model responses. The correlation was 0.7652 for Model 1 and 0.8494 for Model 2.

Table 5: Correlation between subject and model responses for scene pairs where more than 50% of subjects agreed on salience.

Model Number	Model Description	Correlation
1	Scaling by D_i on entire feature maps	0.7652
2	Local inhibition	0.8494

For completeness, in Table 6 we show the average correlations of the subjects to the model for all scene pairs, which includes the scene pairs where subjects indicated the two scenes were equally salient. From this table, we can see that there is very little difference in the correlation values, regardless of whether or not the equally salient scene pairs are included or removed. We also tried applying a threshold to our model where saliency differences below the threshold were considered equally salient. For the threshold, we found the median number of scenes that subjects chose as equally salient. We then set the threshold, such that, the model also responded with the same number of scenes being equally salient. There was little to no change in the correlation values when this threshold was applied.

Table 6: Average correlation of the model to the subjects when all scenes were included in the analysis.

Description	Correlation
Model 1 all scenes	0.44
Model 2 all scenes	0.49

Subject accuracy on the catch trials (both types 1 and 2) presented in the experiment is shown in Table 7. All subjects had performance on the catch trials greater than 50% and only three subjects (6, 7, and 12) had performance lower than 70% correct on these trials. Poor performance on catch trials can be indicative of several

things. For example, it can be a sign that a subject was not focusing on the given task, was randomly selecting answers, or did not understand what they were asked to do in the experiment. All subjects correctly identified the catch trials where the same scene was presented twice as being equally salient. We did not exclude any subjects based on their performance on the catch trials as all subjects had performance greater than 50%.

Table 7: Subject performance for both type 1 and 2 catch trials.

Subject	**% Correct**
1	88.9
2	72.2
3	94.4
4	88.9
5	77.8
6	64.7
7	66.7
8	83.3
9	94.4
10	82.3
11	94.4
12	64.7
13	83.3
14	83.3

In this experiment, the scene that the model considers as the more salient of the two scenes in the scene pair is given by evaluating the saliency difference values of the most salient points from the saliency map for each scene. We wanted to see if the saliency maps generated from the model were representative of the auditory saliency of a scene and also if the maps corresponded to the degree of saliency of that scene. From this, we expect that the scene pairs where there is subject agreement on which scene was more salient of the two scenes presented would result in higher saliency difference values from the model, as opposed to the scene pairs where subjects could not agree on which scene was more salient. We would expect to see lower saliency

difference values from our model for the scene pairs where there was no subject agreement, as this may indicate that the two scenes were close in saliency with no clear choice for which scene is more salient.

Looking at the saliency difference values from the model for these scene pairs, we did find a statistically significant difference in the average saliency differences for scenes where there was subject agreement as opposed to the average saliency differences for scenes where there was no agreement. For this analysis, we excluded the same scene catch trials as the saliency difference for these is zero. For the saliency difference values, we took the absolute value of the saliency differences, since the sign is only an indication of which scene between the first or second one presented was more salient. Table 8 shows the results of this analysis. We found that the average saliency difference for scene pairs where there was subject agreement on saliency was greater than the average difference for scene pairs with no subject agreement. For scene pairs with agreed salience, the average saliency difference was 0.0159 with a standard deviation of 0.04, while the average saliency difference for scene pairs where there was no subject agreement was 0.00517 with a standard deviation of 0.009.

Table 8: Saliency difference values for various scene pairs.

Description	**Average Saliency Difference**
Scene pairs with agreement on saliency	0.0159
Scene pairs with no agreement	0.00517

To determine if this difference is statistically significant, we performed a z-test, where the null hypothesis is that there is no difference between the two means. We find that the test statistic, z, is 2.63. The critical z value for rejection of the null hypothesis is $z < -1.96$ and $z > 1.96$ for alpha = 0.05. Since z=2.63 is greater than 1.96, we can reject the null hypothesis. Therefore, we can conclude that there is a statistically significant difference (p=0.008) between the average difference values for

scenes with subject agreement from the scenes with no agreement. This shows that our model and the saliency maps generated do encompass the features that contribute to saliency and are representative of what humans perceive as being salient. The maps are also representative of the degree of saliency of the scene pairs. The scene pairs with a smaller saliency difference value from the saliency map were closer in saliency, and corresponded to the scenes where subjects were not in agreement as to which of the two scenes was more salient. On the other hand, the scene pairs that had larger saliency differences, resulted in subject agreement on which scene was more salient.

We found that there was a strong correlation between subjects' responses and our model. The next question that needs to be considered is how closely we can expect the model to match a subject's responses. It is not reasonable to expect the bottom-up saliency model to be able to predict with 100% accuracy what one person determines is salient, since saliency is a subjective measure that will vary from one individual to another. One measure we used to determine how closely we can expect the model to match subjects' responses was to look at the correlation values for each individual subject with the rest of the subjects. By looking at how correlated one subject is to another subject, we obtain a measure for how high of a correlation value we can expect to achieve between the correlations of the model to the subjects. Figure 24 is a correlation matrix showing the correlation of each of the subjects to one another. The shades of green used show the strength of the correlation. The correlation values for subjects that are highly correlated to one another are shown in darker green. Subjects with lower correlation values are shown in lighter shades of green. The last two columns on the right of the figure, labeled MG and ML, for the global and local models, respectively, show the correlations of each subject to the model. This same information can also be found in Table 4. Finally, the last row in Figure 24 shows the average correlation for each subject with the rest of the subjects. Overall, the average correlation of the subjects to one another is 0.4402 with a standard deviation

of 0.1270. We can see that the correlation of the model to the subjects is actually slightly higher than the average correlation of the subjects to each other.

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	MG	ML
s1	-	0.50	0.57	0.81	0.13	0.22	0.42	0.73	0.76	0.45	0.75	0.66	0.65	0.60	0.73	0.87
s2	0.50	-	0.31	0.53	0.26	0.19	0.41	0.44	0.64	0.47	0.53	0.47	0.63	0.45	0.45	0.53
s3	0.57	0.31	-	0.48	0.08	0.13	0.21	0.48	0.56	0.46	0.39	0.37	0.71	0.46	0.42	0.52
s4	0.81	0.53	0.48	-	0.12	0.16	0.46	0.64	0.68	0.59	0.64	0.62	0.75	0.56	0.77	0.79
s5	0.13	0.26	0.08	0.12	-	0.14	0.12	0.05	0.23	0.22	0.18	0.05	0.34	0.24	0.04	0.02
s6	0.22	0.19	0.13	0.16	0.14	-	0.17	0.21	0.17	0.19	0.20	0.18	0.16	0.26	0.06	0.11
s7	0.42	0.41	0.21	0.46	0.12	0.17	-	0.57	0.40	0.36	0.50	0.42	0.43	0.36	0.41	0.42
s8	0.73	0.44	0.48	0.64	0.05	0.21	0.57	-	0.63	0.28	0.56	0.55	0.54	0.51	0.60	0.70
s9	0.76	0.64	0.56	0.68	0.23	0.17	0.40	0.63	-	0.49	0.62	0.59	0.67	0.57	0.60	0.69
s10	0.45	0.47	0.46	0.59	0.22	0.19	0.36	0.28	0.49	-	0.51	0.58	0.81	0.49	0.42	0.42
s11	0.75	0.53	0.39	0.64	0.18	0.20	0.50	0.56	0.62	0.51	-	0.67	0.59	0.47	0.62	0.68
s12	0.66	0.47	0.37	0.62	0.05	0.18	0.42	0.55	0.59	0.58	0.67	-	0.57	0.45	0.61	0.63
s13	0.65	0.63	0.71	0.75	0.34	0.16	0.43	0.54	0.67	0.81	0.59	0.57	-	0.62	0.54	0.56
s14	0.60	0.45	0.46	0.56	0.24	0.26	0.36	0.51	0.57	0.49	0.47	0.45	0.62	-	0.40	0.47
avg	0.56	0.45	0.40	0.54	0.17	0.18	0.37	0.48	0.54	0.45	0.51	0.48	0.57	0.46	0.48	0.53

Figure 24: Correlation matrix showing the correlation values of each subject to the other subjects as well as to the model.

From this, we hypothesize that our model's performance on this task is the same as that of the human subjects and that the model is not distinguishable from the subjects. Using hypothesis testing, our null hypothesis is that there is no difference between the average correlation of the model with the subjects and the average correlation of the subjects to each other. The alternative hypothesis is that we are able to distinguish the model from the subjects, as there is a statistically significant difference between the correlation of the model to the subjects and the subjects to each other.

Using a t-test, we find the test statistic, t, is 0.51 ($p<0.61$) for the global model and 1.20 ($p<0.24$) for the local model. Looking at the t distribution tables, the criteria for rejection of the null hypothesis is $t < -2.06$ and $t > 2.06$ for alpha = 0.05. Since t is within the acceptable range in both cases, we cannot reject the null hypothesis.

Therefore, we can say there is no difference between the average correlation of the model to the subjects and the average correlation of the subjects to each other.

Next, we investigate whether or not there are any outliers in the data that should be removed.

4.2.3.1 Outlier Analysis

While no subjects were excluded based on the catch trial performances, we did note that two of the subjects (5,6) had extremely low correlation values compared to the rest of the subjects that participated in the experiment. Therefore, we performed outlier analysis to determine if these two subjects should be excluded from the results. Here, we used Tukey's fence method to identify possible outliers [74]. Figure 25 shows the box plots for the data.

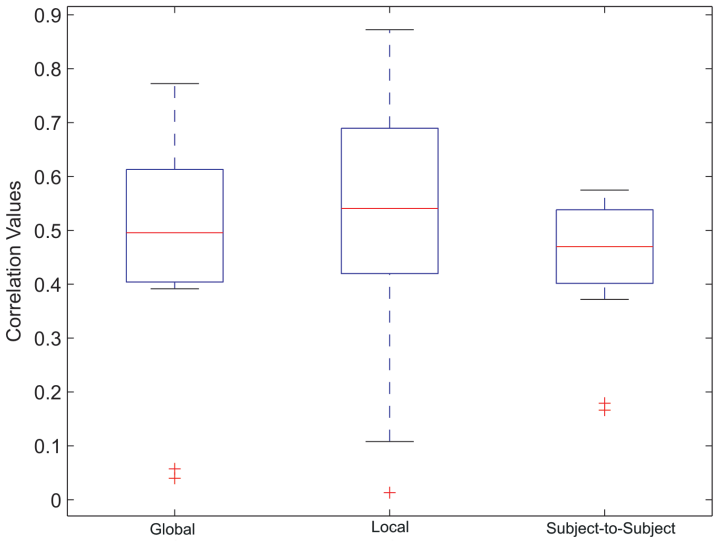


Figure 25: Box plot showing outliers in the saliency scene comparison experiment.

Here, the first box is for the correlation values of the subjects to the global model, the second box is the correlation of subjects to the local model, and the last box is the correlation of subjects to each other. The line in each box represents the median and the top and bottom edges of the boxes represent the 25th (Q1) and 75th (Q3)

percentiles, respectively. The outliers, shown as a red +, are the points that are more than 1.5 times the interquartile range (IQR) below Q1 and above Q3, where IQR = Q3 - Q1. The box plots indicate that these two subjects are indeed "possible" outliers in the data.

Table 9 presents the results of the correlation values between the model's and subjects' responses without the two subjects that were identified as possible outliers. We see an increase in the average correlation of model and subjects' responses from 0.47 to 0.55 for the global model and 0.53 to 0.60 for the local model. Looking at the correlation values for the scenes where the majority of subjects agreed on which scene was more salient, we see a very strong correlation to both the global and local model. For the local model, the correlation is now 0.92. Comparing these correlation values to the average correlation of subjects to each other, we see that the average correlation values to model are slightly higher than the average subject-to-subject correlation which is 0.54, but a t-test found no statistically significant difference, so again we can say subjects were as correlated to the model as they were to each other.

Table 9: Correlation between subject and model responses with outliers (subjects 5,6) removed.

Model	**Correlation**
Global inhibition	0.5454
Local inhibition	0.6049
Agreed salience - Global	0.8262
Agreed salience - Local	0.9227
Subject-to subject	0.5380

Finally, we perform a power analysis to look at how much power there is in our experiment. The results of the power analysis will also serve as a guideline for the number of subjects that should be used for subsequent experiments that will be discussed in the next chapters.

4.2.3.2 Power Analysis

A power analysis provides information on the probability of finding an effect that is present. In other words, it is the probability that we will be able to detect a statistically significant difference when such a difference exists. Therefore, if an experiment does not have enough power, there is a higher chance to miss finding an effect even though one is present.

The information from this power analysis can be used in planning future experiments. Many factors can affect the power. The three main variables that will affect the power are: sample size, effect size, and significance level. The significance level, also known as alpha, represents the probability of a Type I error. A Type I error is when an effect is falsely detected. A commonly used and acceptable value for alpha is 0.05, which is what we have selected to use for this power analysis. This indicates that there is 5% chance of rejecting the null hypothesis when we should not. Of the different variables that much be specified in performing a power analysis, the effect size, which measures the strength of the relationship we are looking at, may be the most difficult to estimate. Cohen's conventions are used as a general guideline in many psychological studies [14, 15]. For correlation coefficients, Cohen considers 0.1 to be a weak effect, 0.3 medium correlation, and 0.5 or greater a strong correlation or effect. Based on these guidelines, in our experiment, we found a strong effect, especially for the scene pairs where there was subject agreement. For these pairs, the correlation was over 0.8 for the local model. In a similar experiment in [42], they also found a moderate to strong effect. Based on this information, for the following studies, we will look for an effect size of around 0.5.

Depending on the type of experiment, a generally accepted value for the power is 0.8 or greater. A power of 0.8 would mean that there is an 80% chance of correctly rejecting the null hypothesis, or a 20% chance of accepting the null hypothesis when we should not.

In Figure 26, we show a plot of power versus effect size for a sample size of 14, which is the number of subjects used for this experiment. The two curves on the plot represent significance levels of 0.05 and 0.1. Looking at the scene pairs where there was subject agreement on saliency, for alpha of 0.05, we found that for 14 subjects and the effect size found, we have a power of 0.9871 for the local model and a power of 0.920219 for the global model. Based on these power numbers, the relatively low number of subjects appears to have been sufficient, due to the fact that such a strong effect was found.

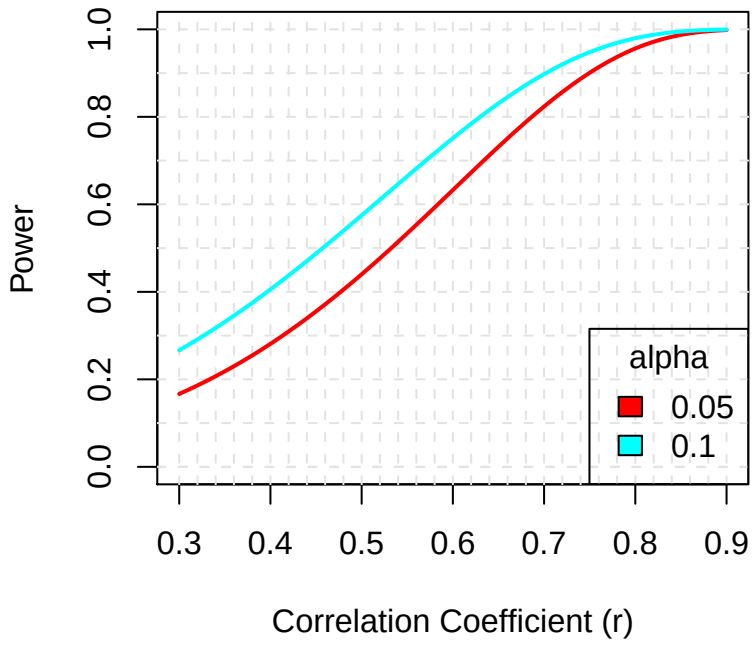


Figure 26: Plot of power vs effect size for sample size, n=14.

The next plot, Figure 27, is a plot of the number of subjects versus effect size for power levels 0.8, 0.85, and 0.9 and significance level of 0.05. As we can see, if we are looking for an effect size of 0.5 and power of 0.9, we would need at least 37 subjects. For power of 0.85, we would need at least 32 subjects and for power of 0.8, we would

need at least 29 subjects. The information from this power analysis is helpful for determining the number of subjects that should be used for other experiments. From this analysis, we chose to have at least 32 subjects for any following experiments.

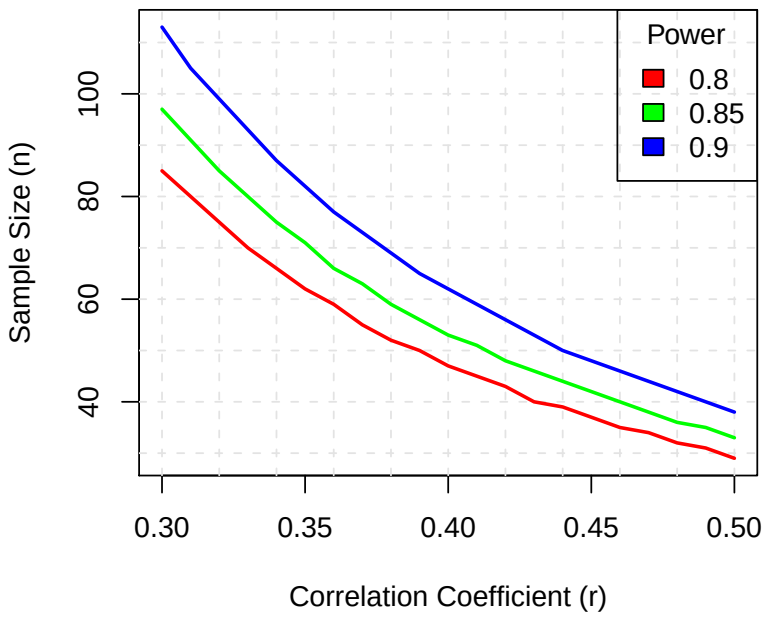


Figure 27: Plot of sample size vs effect size for various power levels.

The results found in this experiment (Tables 4, 5, and 24) shows that our model performs well in predicting which scenes humans consider salient. In the next chapter, we will use a similar experimental setup to further explore what features may contribute to saliency. This information can then be used to determine if there are any additional features that could possibly be added to the current feature set.

CHAPTER V

MULTI-DIMENSIONAL SCALING

Multi-dimensional scaling (MDS) is a statistical technique used to investigate the similarities or dissimilarities of a set of objects. It is also often used as a tool for creating a visual representation or perceptual mapping of the relationships between different variables within a set of data. The different objects are represented by points in typically a two or three dimensional space. For example, two points closer together in the multidimensional space would represent similar objects while two points that are further apart would represent objects that are less similar. Here, we use the novel approach of applying the MDS technique to explore auditory salience. The goal of using the MDS technique to analyze saliency ratings is to determine the features that contribute to the salience of a stimulus. We also use this analysis to determine if there are any additional features contributing to saliency for humans that could be added to the existing feature set of the computational auditory saliency model presented in Chapter 3.

5.1 Multi-dimensional scaling and auditory saliency

Multi-dimensional scaling is now widely used in many different disciplines, but it was first introduced in psychometrics as a way to understand and interpret subjects' ratings of similarities for a set of objects [72, 64]. As multi-dimensional scaling relates to the similarities and dissimilarities of a set of variables, we determined that it could be a useful tool for analyzing the subjects' ratings of saliency obtained in our experiment.

There are two main types of multi-dimensional scaling: metric MDS and non-metric MDS. These two types differ mainly based on the type of data that will be

used or collected during the experiment. Metric MDS deals with quantitative (interval or ratio) data while non-metric MDS involves qualitative (ordinal) data.

There are no known experiments using multi-dimensional scaling to investigate the salience of different stimuli. Here, we use it to determine what some of the dimensions (features) contributing to the saliency of a particular stimulus are. This is a new and novel contribution, since there have not been any other studies or experiments performed that focus specifically on saliency and exploring the underlying factors that make a sound salient. The goal of performing this MDS analysis on our experimental data will be to use this information about the factors contributing to saliency to determine if there are any significant features not accounted for by the model. Additionally, the analysis may reveal new features that could be added to our existing feature set.

As we are interested in examining where the model fails (model and subject selections of salient scenes do not match), we have selected as the stimuli for this experiment the scenes in which the model was not able to match well with subjects' responses regarding which scene was more salient in the saliency scene comparison experiment.

The setup for the experiment used for the MDS analysis is almost identical to the saliency scene comparison experiment described in the previous chapter. One difference is that, in this new experiment, the subjects were presented with all possible scene pairings as MDS analysis works best when subjects provide their ratings for all possible pairwise combinations.

5.1.1 Subjects

Using the results from the power analysis performed for the scene saliency comparison experiment, we wanted to have at least 32 subjects complete the experiment.

Results were obtained from 34 (14 female, 20 male) undergraduate students at the

Georgia Institute of Technology under IRB approval. The students received course credit for participation in the experiment. Normal hearing was determined by self-report. Subjects were informed about the general aim of the experiment, but were naive to the exact purpose of the study.

Subjects listened through headphones (Sennheiser HD 280 pro) to auditory stimuli presented using MATLAB and then entered their responses into a MATLAB GUI. The GUI interface used is the same as the one used in the saliency scene comparison experiment presented in Chapter 4 and is shown in Figure 23. The experiment took each subject approximately one hour to complete.

5.1.2 Experimental Procedures

Each subject was presented with a total of 198 scene pairs created from 20 unique target sounds. As before, the auditory scenes consisted of a target (one second) sound on a background (4 seconds). The scenes used were selected from the scene pairs where the model failed in the saliency scene experiment presented in the previous chapter. These failed scenes were the scenes where there was a majority agreement amongst the subjects as to which scene was the more salient, but the model did not agree with the subjects' selection.

Subjects were presented with each of the different auditory scene pairs. As in the previous experiment, each scene pair was made up of two 4 second auditory scenes with a 1 second pause between the two scenes. Again, subjects were asked to indicate which of the two scenes they heard had the most salient element. They were then also asked to rate from 1 to 5 how much more salient they found that scene compared to the other one.

Two different types of catch trials were included to check that subjects were paying attention to the given task. The first type was where the same scene was presented for both the first and second scene of the scene pair, and the second type was where

scene pairs presented earlier in the experiment were presented again later were also used to get an indication of how consistent a subject was.

5.1.3 Results and Discussion

Subject accuracy on the catch trials is presented in Figure 28. Two subjects (12, 26) had very low performance on the catch trials (less than 40%). Additionally, three of the subjects (15, 26, 30) were not able to correctly identify all the catch trials where the same scene was presented twice as being equally salient.

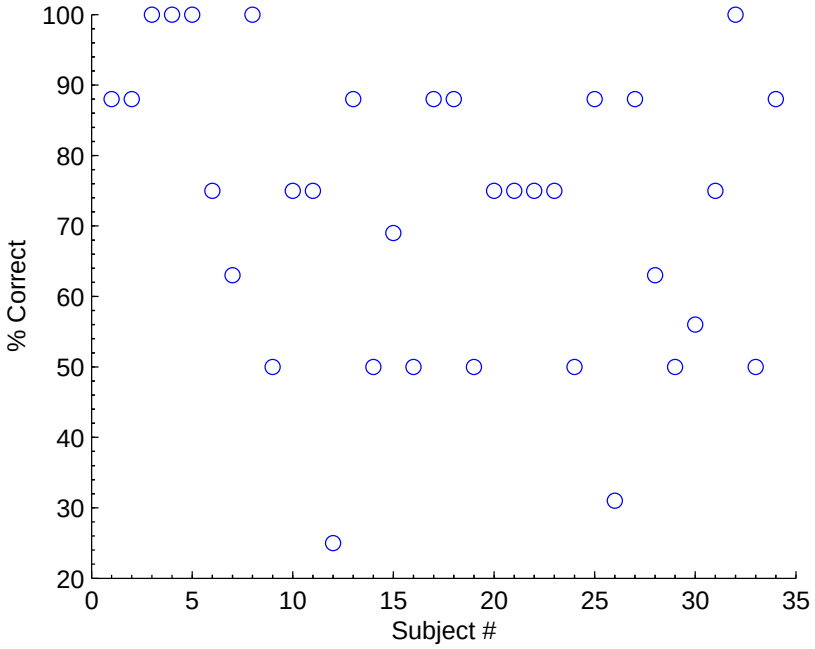


Figure 28: Subject performance on the catch trials for the MDS experiment.

Looking at the correlation values between the scenes our model selected as salient and the scenes that subjects selected as salient, we found that the correlation values were much weaker than the correlation values found in the saliency scene comparison experiment discussed in the previous chapter. This result is what we would expect, since in this experiment, we are using the auditory scenes where the model performed poorly (model's selection did not match subjects' selections in cases where there was clear agreement amongst subjects) in the previous experiment. The correlation

results for each subject with both Model 1 (global) and Model 2 (local) can be found in Table 10.

Table 10: Correlation between subject and model responses for MDS experiment.

Subject	Model 1	Model 2
1	0.2251	0.3822
2	0.2724	0.3560
3	-0.1942	-0.1049
4	0.0562	0.1682
5	0.1949	0.3457
6	0.0790	0.2317
7	0.0436	0.2354
8	0.1860	0.2826
9	0.1774	0.3612
10	0.3285	0.4129
11	0.3045	0.3726
12	0.1607	0.2296
13	0.2286	0.3609
14	0.1155	0.2250
15	0.1885	0.3317
16	0.3865	0.3999
17	0.2803	0.4061
18	-0.1182	0.0585
19	0.1109	0.2345
20	-0.0181	-0.0181
21	0.3505	0.5007
22	-0.1450	-0.0604
23	0.2099	0.4401
24	0.2472	0.2407
25	0.1670	0.3241
26	0.1083	0.1943
27	0.0591	0.2093
28	0.1302	0.2434
29	-0.0070	0.0256
30	0.1791	0.2092
31	0.2326	0.2635
32	0.1548	0.2890
33	-0.1753	-0.0242
34	-0.1178	-0.0022
Average	0.1294	0.2390
Std Dev	0.1512	0.1553

The average correlation for the the model's responses to subjects' responses was much weaker for the global model. The correlation here was 0.1294 (p=0.46) compared to 0.2390 (p=0.17) for the local model. If we remove subjects 12 and 26, due to their poor performance on the catch trials, we do not see any significant difference in the average correlation values between each of our models and all the subjects. The average correlation, without these two subjects, is 0.1291 (p=0.47) for Model 1 and 0.2406 (p=0.17) for Model 2.

As we did in the saliency scene comparison experiment in section 4.2.3, we can eliminate some of the subjective or top-down component of saliency by examining only the scenes pairs where more than half of the subjects agreed upon which scene

was more salient. These scene pairs are of particular interest for comparison to our bottom-up saliency model, since these are the scenes where there was subject agreement on the saliency, therefore eliminating some of the individual variation that may cause some stimuli to be salient to only certain individuals. The correlation of the model to the subjects for these scenes can be found in Table 11. As expected, for these scene pairs, the correlation is higher, and we see a correlation between the model's responses and subjects' responses of 0.3590 for Model 1 and 0.5823 for Model 2.

Table 11: Correlation between subject and model responses for scene pairs where more than 50% of subjects agreed on salience for MDS experiment.

Model	Correlation
1 Scaling by D_i on entire feature maps	0.3590
2 Local inhibition	0.5823

While the correlation values for this experiment where lower than in the previous scene saliency experiment, if we look at the correlation of the subjects to each other, we see that those values are lower as well. Figure 29 is a correlation matrix showing the correlation of each subject to the other subjects. The strength of the correlation is represented by the colors, ranging from dark green (high correlation) to white (no correlation). Subjects that are more highly correlated are shown in darker green, and subjects that are less correlated are shown in light green. The red is used to indicate a negative correlation.

The average correlation values for each subject to all of the other subjects can also be found at the bottom of Figure 29. These values can provide a measure for how closely we can expect the model to match a subject's response. Overall, the average correlation of subjects to each other is 0.2522 with a standard deviation of 0.1047. Here, we see that the average correlation of subjects to one another is about the same as the correlation of subjects to Model 2 (local) but is much higher than the average

correlation of subjects to Model 1 (global).

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24	#25	#26	#27	#28	#29	#30	#31	#32	MG	ML
#1	-	.35	.11	-.01	.47	.41	.20	.32	.53	.46	.33	.33	.35	.60	.46	.42	.18	.43	.02	.47	-.01	.38	.28	.28	.28	.38	.35	.30	.33	.46	.21	.30	.23	.38
#2	.35	-	.12	-.08	.49	.46	.31	.31	.39	.32	.49	.21	.30	.51	.51	.46	.12	.44	.08	.42	.11	.50	.31	.44	.29	.39	.18	.24	.34	.59	.12	.20	.27	.36
#3	.11	.12	-	.11	-.09	.14	.19	-.03	.06	-.08	-.11	.10	.12	.02	-.06	.01	.12	.05	-.12	.02	.08	.10	.25	-.19	-.16	.17	.09	.22	.06	.05	.37	.50	-.19	-.10
#4	-.01	-.08	.11	-	-.19	.06	.08	-.17	.14	-.10	-.24	.14	.18	-.15	.01	-.18	.14	.12	-.03	.06	.05	.02	-.19	-.29	.12	.13	.05	-.19	-.13	-.17	.19	.22	.06	.17
#5	.47	.49	-.09	-.19	-	.49	.20	.57	.33	.66	.66	.44	.42	.63	.51	.48	.30	.35	.21	.39	.03	.54	.41	.63	.16	.31	.14	.36	.50	.64	.08	.25	.19	.35
#6	.41	.46	.14	.06	.49	-	.34	.30	.61	.36	.55	.42	.30	.57	.43	.30	.32	.54	.06	.44	.09	.51	.25	.43	.38	.43	.27	.25	.36	.51	.31	.43	.08	.23
#7	.20	.31	.19	.08	.20	.34	-	.15	.33	.10	.21	.24	.16	.37	.13	.28	.27	.29	-.06	.31	.06	.39	.18	.19	.29	.31	.04	.11	.25	.34	.30	.25	.04	.24
#8	.32	.31	-.03	-.17	.57	.30	.15	-	.23	.46	.54	.42	.18	.45	.35	.43	.19	.12	.14	.37	-.01	.39	.19	.56	.10	.15	-.06	.28	.36	.38	.04	.19	.19	.28
#9	.53	.39	.06	.14	.33	.61	.33	.23	-	.18	.30	.42	.28	.45	.41	.23	.25	.44	-.07	.59	-.02	.46	.33	.31	.45	.44	.29	.36	.42	.44	.35	.42	.18	.36
#10	.46	.32	-.08	-.10	.66	.36	.10	.46	.18	-	.62	.39	.37	.52	.44	.50	.20	.19	.24	.34	-.08	.44	.34	.50	.16	.13	.05	.25	.39	.55	-.01	.11	.33	.41
#11	.33	.49	-.11	-.24	.66	.55	.21	.54	.30	.62	-	.38	.35	.64	.51	.61	.22	.24	.29	.39	.03	.55	.47	.68	.22	.21	-.07	.31	.43	.67	.00	.18	.30	.37
#12	.33	.21	.10	.14	.44	.42	.24	.42	.42	.39	.38	-	.21	.40	.37	.27	.30	.22	.10	.40	-.05	.50	.11	.30	.23	.25	.32	.18	.19	.37	.24	.40	.23	.36
#13	.35	.30	.12	.18	.42	.30	.16	.18	.28	.37	.35	.21	-	.40	.37	.36	.18	.19	.07	.21	.09	.35	.16	.18	.11	.28	-.05	.09	.15	.38	.20	.33	.12	.22
#14	.60	.51	.02	-.15	.63	.57	.37	.45	.45	.52	.64	.40	.40	-	.57	.49	.17	.43	.09	.43	.05	.58	.40	.53	.32	.35	.27	.35	.41	.62	.16	.24	.19	.33
#15	.46	.51	-.06	.01	.51	.43	.13	.35	.41	.44	.51	.37	.37	.57	-	.34	.10	.44	-.06	.53	-.01	.43	.26	.42	.19	.27	.31	.36	.29	.37	-.02	.14	.39	.40
#16	.42	.46	.01	-.18	.48	.30	.28	.43	.23	.50	.61	.27	.36	.49	.34	-	.28	.08	.08	.49	-.06	.52	.34	.52	.15	.28	.00	.32	.37	.59	.06	.21	.28	.41
#17	.18	.12	.12	.14	.30	.32	.27	.19	.25	.20	.22	.30	.18	.17	.10	.28	-	.11	.15	.23	.04	.41	.03	.18	.04	.14	.07	.05	.22	.24	.32	.42	-.12	.06
#18	.43	.44	.05	.12	.35	.54	.29	.12	.44	.19	.24	.22	.19	.43	.44	.08	.11	-	-.03	.39	.13	.37	.22	.22	.49	.52	.31	.26	.32	.35	.24	.23	.11	.23
#19	.02	.08	-.12	-.03	.21	.06	-.06	.14	-.07	.24	.29	.10	.07	.09	-.06	.08	.15	-.03	-	-.01	.01	.02	.13	.24	.04	-.05	-.17	-.07	.13	.10	.07	.04	-.02	-.02
#20	.47	.42	.02	.06	.39	.44	.31	.37	.59	.34	.39	.40	.21	.43	.53	.49	.23	.39	-.01	-	.05	.48	.29	.34	.39	.33	.28	.52	.24	.32	.27	.15	.35	.50
#21	-.01	.11	.08	.05	.03	.09	.06	-.01	-.02	-.08	.03	-.05	.09	.05	-.01	-.06	.04	.13	.01	.05	-	.09	.14	-.02	.30	.03	.20	.09	.00	.08	.15	.06	.14	-.06
#22	.38	.50	.10	.02	.54	.51	.39	.39	.46	.44	.55	.50	.35	.58	.43	.52	.41	.37	.02	.48	.09	-	.22	.54	.34	.44	.15	.24	.37	.54	.32	.36	.21	.44
#23	.28	.31	-.25	-.19	.41	.25	.18	.19	.33	.34	.47	.11	.16	.40	.26	.34	.03	.22	.13	.29	.14	.22	-	.42	.35	.17	.06	.45	.36	.43	-.09	-.08	.25	.24
#24	.28	.44	-.19	-.29	.63	.43	.19	.56	.31	.50	.68	.30	.18	.53	.42	.52	.18	.22	.24	.34	-.02	.54	.42	-	.24	.22	-.06	.34	.51	.54	.02	.12	.17	.32
#25	.28	.29	-.16	.12	.16	.38	.29	.10	.45	.16	.22	.23	.11	.32	.19	.15	.04	.49	.04	.39	.30	.34	.35	.24	-	.43	.36	.34	.28	.33	.16	.05	.06	.21
#26	.38	.39	.17	.13	.31	.43	.31	.15	.44	.13	.21	.25	.28	.35	.27	.28	.14	.52	-.05	.33	.03	.44	.17	.22	.43	-	.15	.26	.28	.37	.31	.32	.13	.24
#27	.35	.18	.09	.05	.14	.27	.04	-.06	.29	.05	-.07	.32	-.05	.27	.31	.00	.07	.31	-.17	.28	.20	.15	.06	-.06	.36	.15	-	.20	.01	.15	.12	.06	-.01	.03
#28	.30	.24	-.22	-.19	.36	.25	.11	.28	.36	.25	.31	.18	.09	.35	.36	.32	.05	.26	-.07	.52	.09	.24	.45	.34	.34	.26	.20	-	.25	.28	-.02	.03	.18	.21
#29	.33	.34	.06	-.13	.50	.36	.25	.36	.42	.39	.43	.19	.15	.41	.29	.37	.22	.32	.13	.24	.00	.37	.36	.51	.28	.28	.01	.25	-	.49	.13	.29	.23	.26
#30	.46	.59	.05	-.17	.64	.51	.34	.38	.44	.55	.67	.37	.38	.62	.37	.59	.24	.35	.10	.32	.08	.54	.43	.54	.33	.37	.15	.28	.49	-	.15	.30	.15	.29
#31	.21	.12	.37	.19	.08	.31	.30	.04	.35	-.01	.00	.24	.20	.16	-.02	.06	.32	.24	.07	.27	.15	.32	-.09	.02	.16	.31	.12	-.02	.13	.15	-	.43	-.18	-.02
#32	.30	.20	.50	.22	.25	.43	.25	.19	.42	.11	.18	.40	.33	.24	.14	.21	.42	.23	.04	.15	.06	.36	-.08	.12	.05	.32	.06	.03	.29	.30	.43	-	-.12	-.00
avg	.32	.32	.04	-.01	.37	.37	.22	.25	.33	.29	.34	.28	.23	.38	.30	.30	.19	.28	.05	.33	.05	.37	.22	.30	.24	.27	.13	.21	.28	.37	.17	.23	.13	.24

Figure 29: Correlation matrix showing the correlation of each subject to the other subjects for the MDS experiment.

The next step will be to perform the MDS analysis on the subjects' ratings of saliency. The statistical program Statistics 19 by IBM was used to perform this analysis. A visual representation of the data resulting from the MDS analysis is shown in Figure 30 and Figure 31 for two and three dimensions, respectively.

5.1.3.1 Determining data dimensionality

An important step in MDS analysis is to determine the number of dimensions. In order to determine the dimensionality of the data, we first looked at the Kruskal's stress values that were calculated. These stress values serve as an indication of how well the model fits our data. We plotted the stress values by dimension on a scree plot

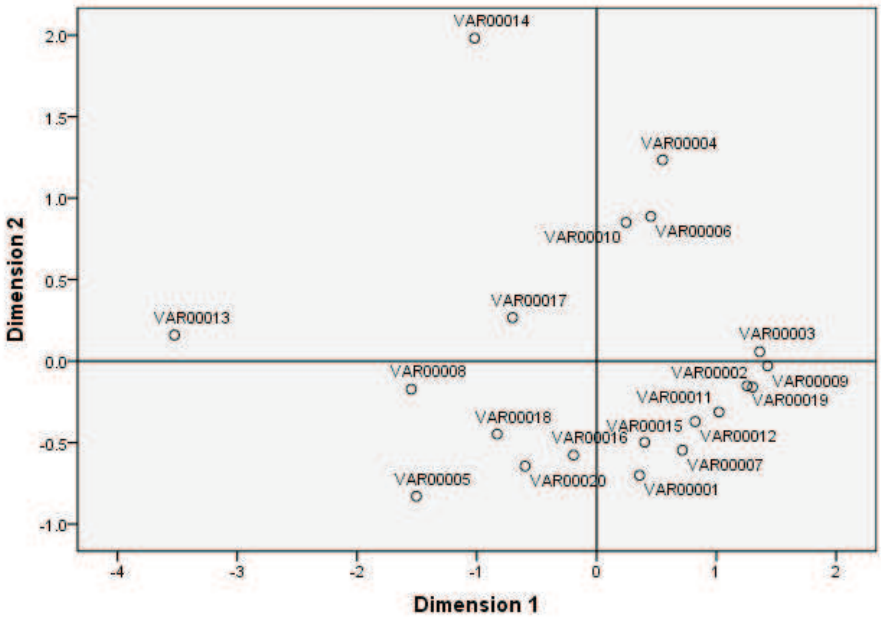


Figure 30: Two-dimensional visual representation of data on subject's ratings of saliency.

shown in Figure 32. In general, high stress values can indicate that the number of dimensions used to represent the data is not sufficient. When using multi-dimensional scaling in order to create a visual representation, it is common to use at most three dimensions, it can be very difficult to visualize the data in higher dimensions. Depending on the stress values though, it may be such that it is not possible to represent well the particular data set using only two or three dimensions. As the number of dimensions is increased, the stress values will either stay fairly constant or decrease. In general, the point at which the rate of decline in stress slows down (ideally seen as an elbow in the scree plot) is the point which may be considered as the true dimensionality of the data. In this case, the rate of decline in stress slows down at four dimensions. However, the analysis using four dimensions may not be reliable as we have only 20 sounds and 190 total scene pairs. Thus, while adding another dimension does appear to increase the fit of our model, the increase is small enough that it

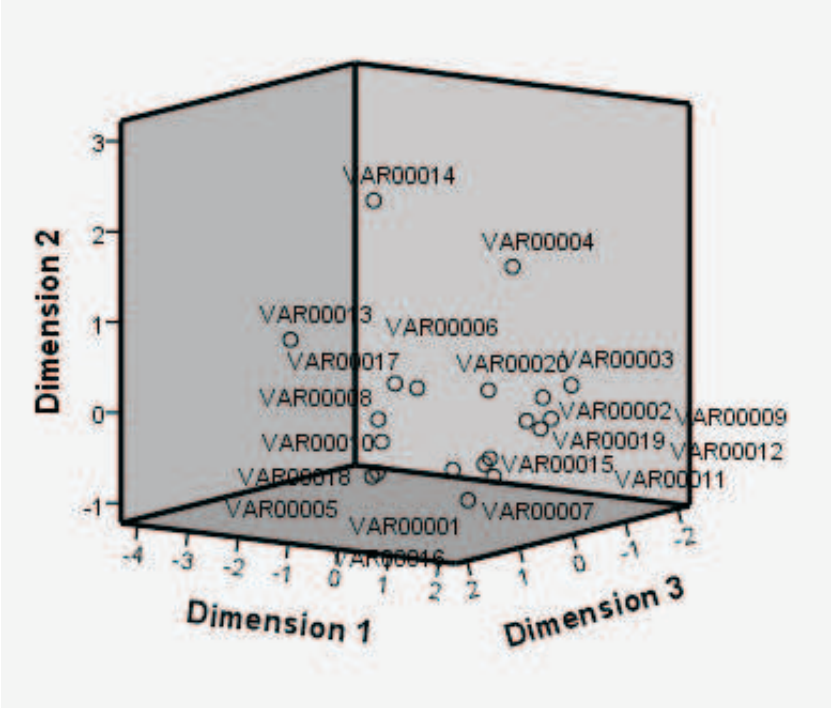


Figure 31: Three-dimensional visual representation of data on subject's ratings of saliency.

may make sense to remain with only three dimensions, as adding another dimension can also make it more difficult to interpret the results. We present the results of the analysis using both three and four dimensions. Additionally, we must also consider that the stress values obtained may not necessarily be the optimal values as they are calculated based on the final iteration only. The data itself is fit by minimizing the S-STRESS loss function. S-STRESS values are calculated as shown in Equation 7 below.

$$\mathcal{SS}1 = \left[\frac{\sum_{(i,j)} (\delta_{ij}^2 - d_{ij}^2)^2}{\sum_{(i,j)} (d_{ij}^2)^2} \right]^{1/2} \tag{7}$$

5.1.3.2 Hierarchical clustering

Now that we have a visual representation of the data set, the next step is to determine what the different dimensions may represent. In order to help determine the

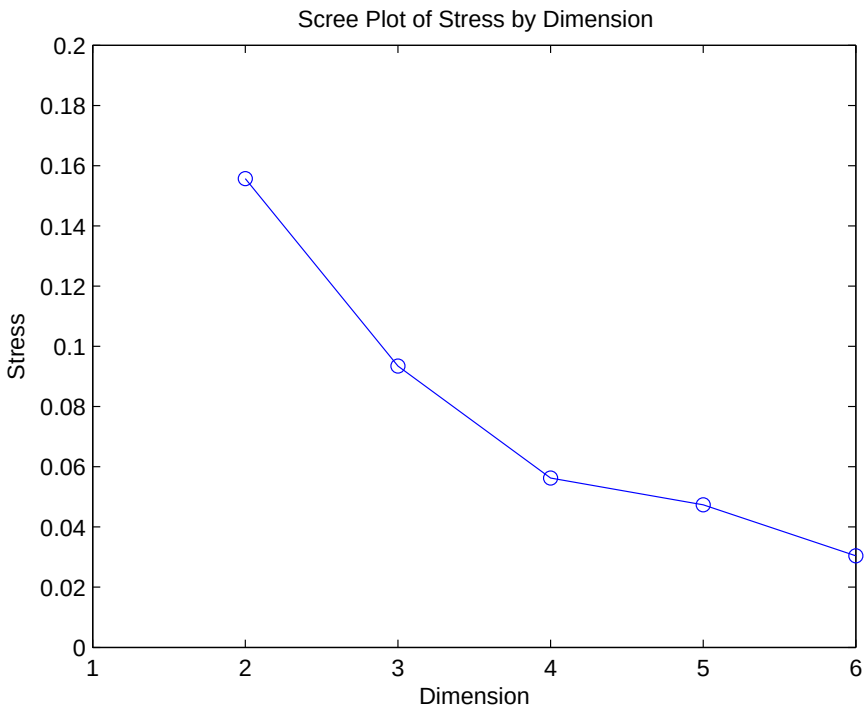


Figure 32: Scree plot showing stress by dimensions.

distances between each of the points, and which sounds should be grouped together, we considered using several different techniques, including k-means clustering, nearest neighbor, and hierarchical clustering. As we are not specifically interested in just clustering the data for this experiment, we chose to use the hierarchical clustering method to create an hierarchical clustering tree which will help to determine the distance between these different sounds. Unlike k-means clustering, the hierarchical clustering method provides more information on the relatedness of the different points within the clusters and of the clusters themselves to each other.

There are two different approaches to hierarchical clustering, the agglomerative approach and the divisive approach. In the agglomerative approach, we build the trees from the bottom up by successively merging smaller clusters into larger ones. On the contrary, in the divisive approach, we work from the top down. This involves starting with one large cluster and gradually splitting the large cluster into smaller clusters.

For our analysis, we use the agglomerative approach, where we begin by first considering each point or sound as being in its own separate cluster. Each of the separate points or clusters are then successively merged with the closest pair of clusters into larger clusters using Euclidean distance as the distance metric. We continue this successive merging process until we are finally left with only one large cluster. The distance used to determine the closest pair of clusters can be defined in several different ways. The two most common ways to define this distance are known as single linkage clustering and complete linkage clustering. In single linkage clustering, the distance between two clusters is defined as the minimum distance between members of the cluster. It can be defined as shown in Equation 8 below.

$$d(A, B) \triangleq \min_{\overrightarrow{x} \in A, \overrightarrow{y} \in B} ||\overrightarrow{x} - \overrightarrow{y}||. \tag{8}$$

Complete linkage clustering, on the other hand, is the exact opposite of single linkage clustering. Here, the distance between the two clusters is defined as the maximum distance between members of the cluster. The complete linkage distance is defined in Equation 9.

$$d(A, B) \triangleq \max_{\overrightarrow{x} \in A, \overrightarrow{y} \in B} ||\overrightarrow{x} - \overrightarrow{y}||. \tag{9}$$

The hierarchical clustering trees and plots for both single and complete linkage clustering for the 3-D data are shown in Figures 33 and 34, and the tree and plots for the 4-D data are shown in Figures 35 and 36. The numbers on the clustering trees and plots each represent one of the 20 target sounds used for the experiment. We show the data grouped into three clusters, but, for our purposes, the exact number of clusters is less important.

As the structure of the hierarchical clustering trees are all very similar, we will start by discussing more in-depth the complete linkage tree for the 4-D data (Figure 35). Table 12 has a brief description of each of the 20 sounds used in the experiment. Here,

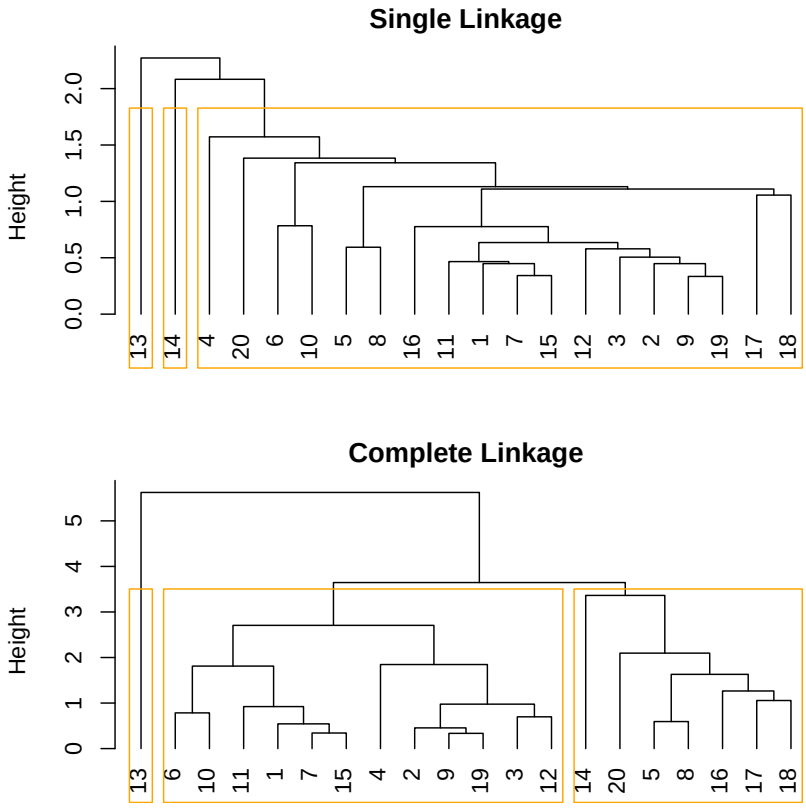


Figure 33: Hierarchical clustering trees for 3-D data.

we first notice that sounds 13 and 14 are each in their own separate cluster and are less similar or related to the rest of the sounds. Sound 13 is a bird chirping rapidly and sound 14 is a cymbal creating a bell-like sound. Both of these sound very different from all the other sounds in the set, in particular sound 14. Cymbals are known as non-pitched or indefinite pitched instruments as there is usually no discernible pitch. This is one reason that it may be different from the other sounds. Additionally, looking at the frequency spectrum for this sound, it covers a wide frequency range and has many random harmonics. Another consideration is that some of the sounds, including the cymbal, are salient to us due to a cognitive component that would not be captured by our bottom-up processing model.

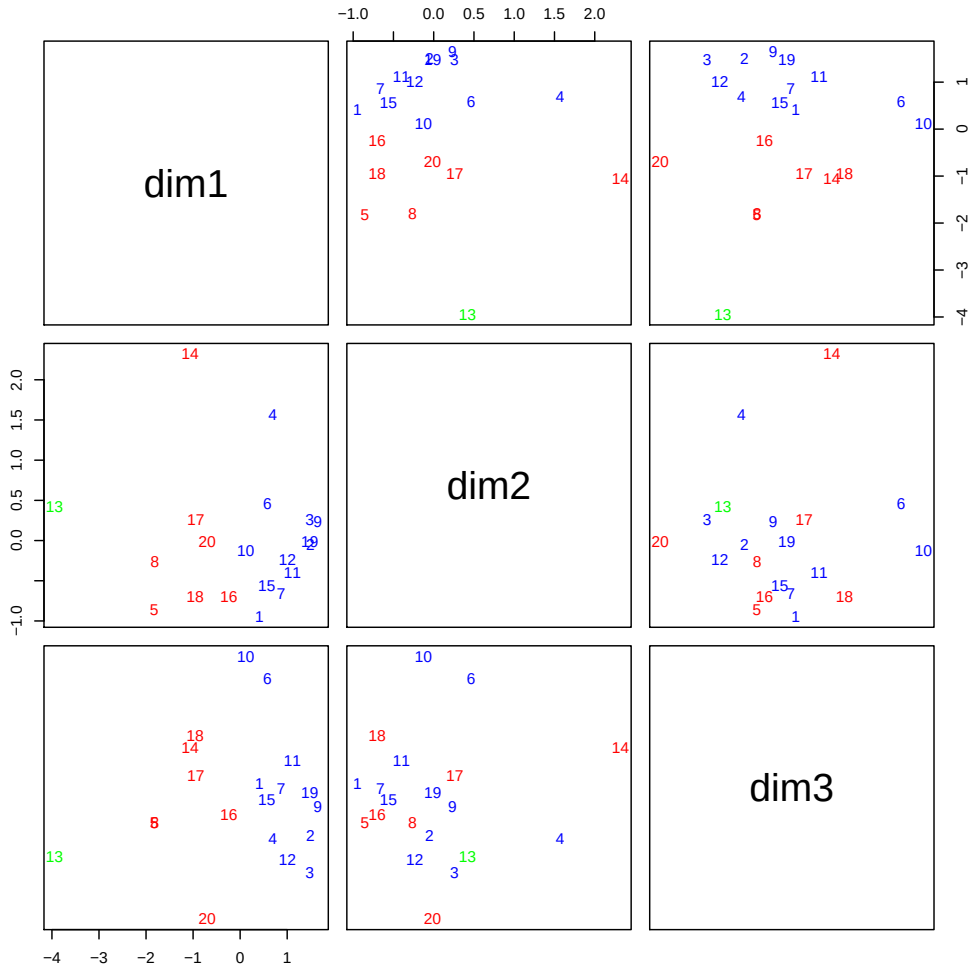


Figure 34: Plot of complete linkage clustering for 3-D data.

Next, we examine the relatedness of the other sounds. Sounds 3, 2, and 19 were found to be closely related. Listening to these sounds, all three are low frequency sounds. These three sounds are most closely related to sounds 4, 9, 12, which are all also low frequency sounds. In addition, sounds 4 and 9 are long, continuous sounds. Using this information, if we look at the plot in Figure 36, dimension 1 as you move up and down along the axis appears to be related to the frequency of the sounds with low frequency sounds grouped on one end, while at the other end, we have sound 13 which is a high frequency bird chirping sound.

Looking at the next grouping of sounds, we see that sounds 1, 15, and 11 are closely related to each other and are also related to the sounds 7 and 16. Both sounds 1 and

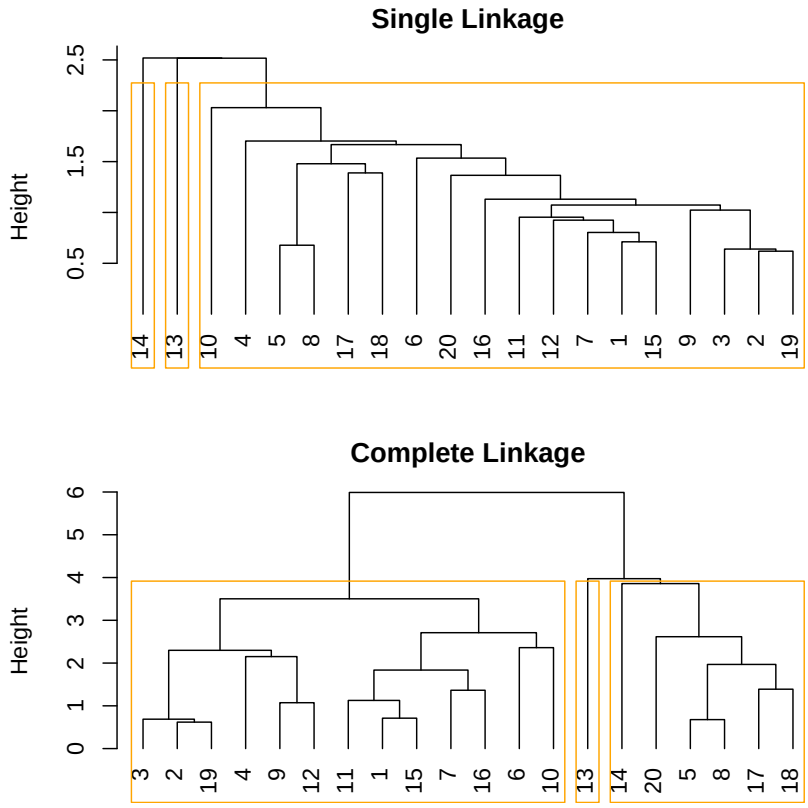


Figure 35: Hierarchical clustering trees for 4-D data.

15 have changes in pitch and repetitive elements. Sound 11 is a string instrument with harmonics. If we look at the frequency spectrum for these sounds, sounds 1, 7, and 15 all have very similar FFT plots with peaks around the same frequencies. Sounds 6 and 10 are the next most closely related cluster to these sounds. Sound 6 is the tone complex where one tone is amplitude modulated. This is the tone complex that was used in the peripheral task of the dual task experiment discussed in the previous chapter. Sound 10 is an animal making a knocking sound followed by a short animal call at the end. Listening to both sounds, you can hear similarities in the modulated component.

Looking at the cluster on the far right of the complete linkage tree in Figure 35, all

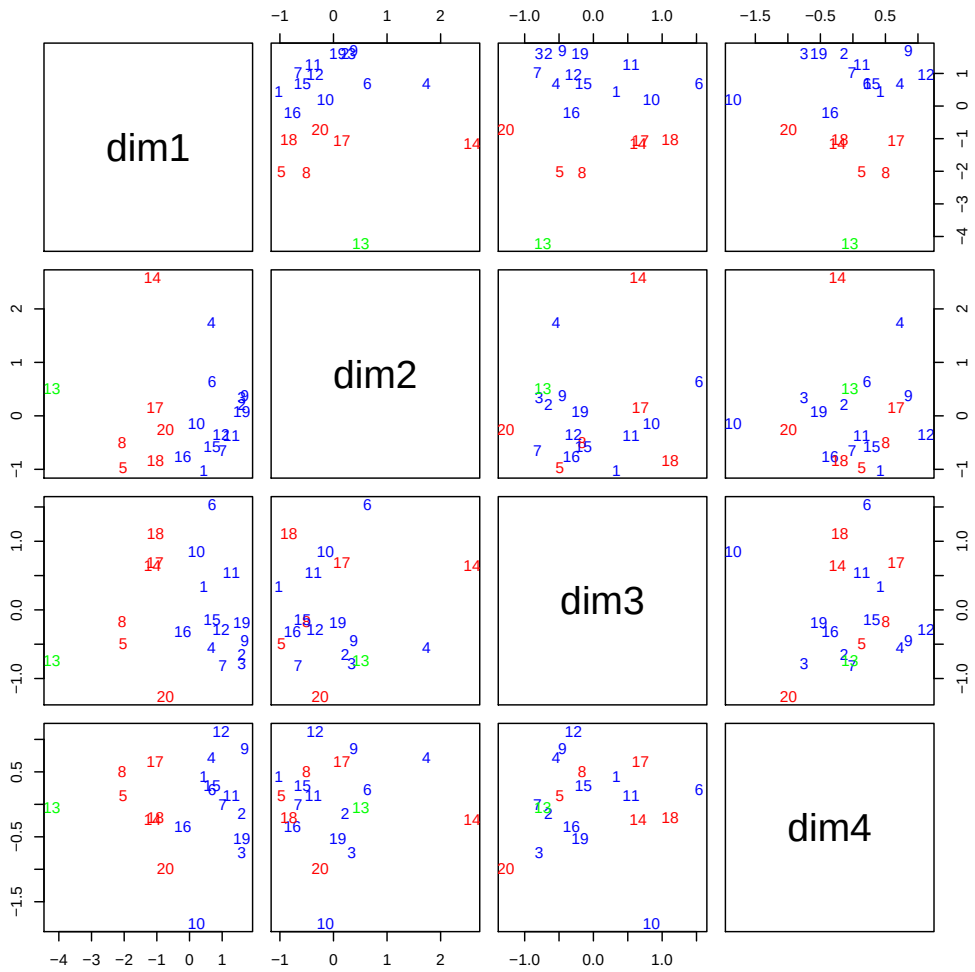


Figure 36: Plot of complete linkage clustering for 4-D data.

the sounds grouped within this cluster are animal sounds. Sounds 5 and 8 are closely related and are grouped together. Both of these two sounds are slightly repetitive sounds in which several pitch changes can be heard, going up and down several times throughout the duration of the signal. These two sounds are most closely related to sounds 17 and 18, which are both also animal calls, but are longer in duration compared to sounds 5 and 8. Sound 20 is the next sound that is most closely related to these two clusters. It is a screeching animal sound that is higher in pitch than the other four sounds. It also seems to have some variations in pitch, but they are more subtle than the ones that can be heard in the other sounds.

In setting up the experiment, we tried to limit the effect of loudness on saliency,

Table 12: Description of sounds used in MDS experiment.

Sound Number	Description
1	Monkey shrieking
2	Monkey call
3	Plane noise
4	Boat horn
5	Bird Chirp
6	Modulated tone complex
7	Bird call (not chirp)
8	Animal
9	Destroyer noise
10	Knocking sound
11	Musical instrument (strings)
12	Horn honking
13	Bird chirping
14	Cymbal
15	Bird call
16	Animal
17	Animal
18	Animal
19	Siren
20	Birds high pitched sound

but despite this, perceived loudness is still a subjective measurement and can vary between individuals. Some of the factors that can influence perceived loudness are the sound pressure, the duration, and the bandwidth of a sound. Additionally, it is known that sounds of the same intensity can still be perceived as louder or softer depending on the specific frequencies that are present in the sound. For example, sounds that are lower than 300 Hz are perceived as being softer than sounds at 800 Hz of the same intensity. This is due to the ear becoming less sensitive at the lower frequencies. The ear is especially sensitive to sounds in the 3000 to 4000 Hz range, therefore a sound at 3000 Hz will sound louder than a sound at 1000 Hz at the same intensity [26]. As we move into the higher frequencies, the ear again becomes less sensitive. In particular, as people age, they gradually lose the ability to hear some of the higher

frequencies [16, 62]. Looking at the clustering of the sounds along dimension 3, it appears that this dimension could be related to loudness perception.

The other two dimensions (dimensions 2 and 4) which may contribute to saliency are more ambiguous, and not as easily definable. They could represent a combination of several different attributes of sound, since each dimension can represent multiple attributes. Looking at the plot in Figure 35, some of the attributes dimension 2 could represent are the onset of sounds or the duration of the sound. It does not hold true for all the sounds, but in general, as we move along that dimension we tend to first encounter longer, more constant sounds. In addition, as we move down that axis, the sounds appear to have more repetitive elements and many onsets, as they have many starts and stops as opposed to being one constant sound of longer duration. These two dimensions may also be less clearly definable as they could also relate to aspects of timbre, which can be harder to define and identify.

5.1.3.3 Adding to the existing feature set

One of the goals of our MDS analysis is to identify what dimensions are most important in determining whether or not a sound is salient. The other objective is to use the information obtained about the factors contributing to sound saliency to explore whether or not there are any additional features that could be added to our existing feature set that would lead to an improvement in our model's ability to pick sounds that are salient to humans.

From our analysis, we found that two of the key dimensions contributing to saliency are related to the frequencies present within the sound and the perceived loudness of the sound. Using this information, we decided to add an additional stage to the model. This additional step would involve pre-processing the sound input by applying a weighting filter before the input is passed on to the feature extraction stage. Figure 37 shows an updated architecture of the computational auditory

saliency model. The filter would weight more heavily the frequencies that the human ear is more sensitive to, and as such, would be perceived as louder. On the other hand, it would give less weight or attenuate the frequencies (very low and very high) that the ear is less sensitive to and can be perceived as being softer.

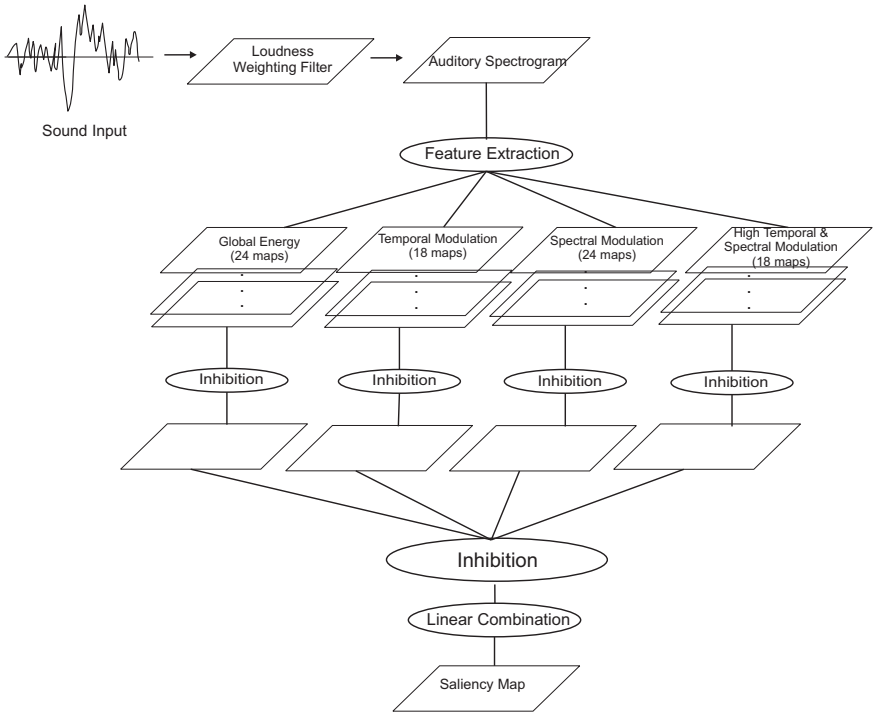


Figure 37: Architecture for enhanced computational auditory saliency model.

The filter used here is designed to emphasize the frequency ranges that the human ear is most sensitive to. We use the inverse of the 40 phon curve from the equal loudness contours first published by Fletcher and Munson in 1933. This type of filter is referred to as an A-weighting filter. Figure 38 shows the Fletcher-Munson loudness contours [26] as well the new standard equal loudness contours given by ISO 226:2003. We can see the 40-phon curve used for the filter matches well with the new standards.

While we base the filter on A-weighting filter specifications, the exact weighting filter used can be modified based on the type of stimuli or the task that the model is being applied to. For example, steeper attention can be used for the high frequencies as many of those frequencies cannot be heard by the ear and would therefore not be

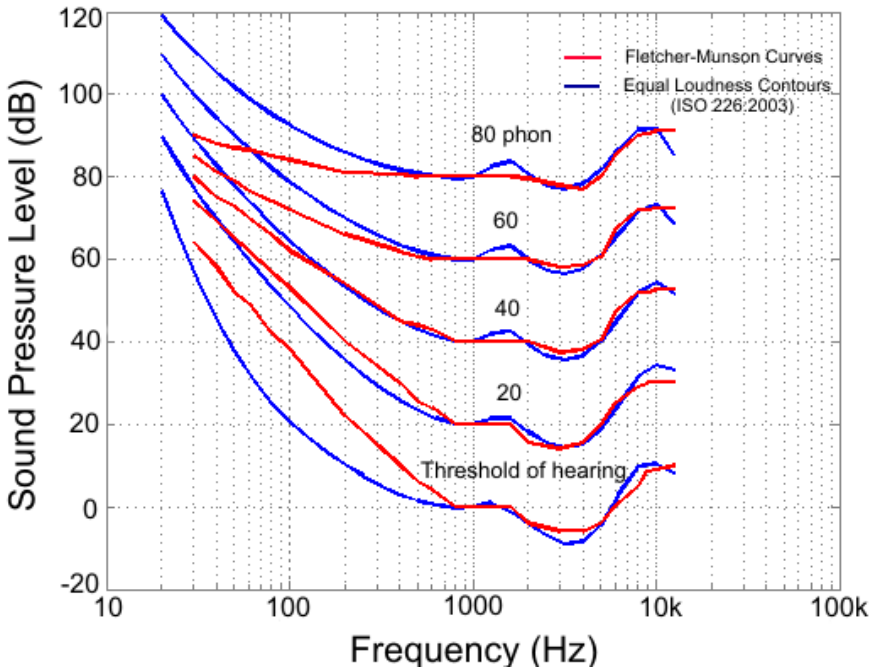


Figure 38: Comparison of Fletcher-Munson curves with new standard equal loudness contours (ISO 226:2003).

salient. A plot of the filter characteristic used is shown in Figure 39.

In Figure 40, we show an example of a sound where adding the pre-processing stage results in a large difference in the saliency map. The sound is similar to what a person waiting outside of a restaurant located along a busy street would hear. The background noise is street noise, and the target sound is a police siren. The signal-to-noise ratio of the siren to the noise is low (-10 dB), but despite this, the siren is very noticeable and would draw our attention.

The auditory spectrogram for this sound is given in Figure 40 (a) and the corresponding saliency map is shown in Figure 40 (b). Due to the low SNR, the siren sound is suppressed and the noise actually becomes the most salient element on the map. In Figure 41, we show the auditory saliency map for the model with the new filtering stage. The filter suppresses the lower frequencies where the noise is and emphasizes the frequencies of the siren, as it is within the frequency range that humans

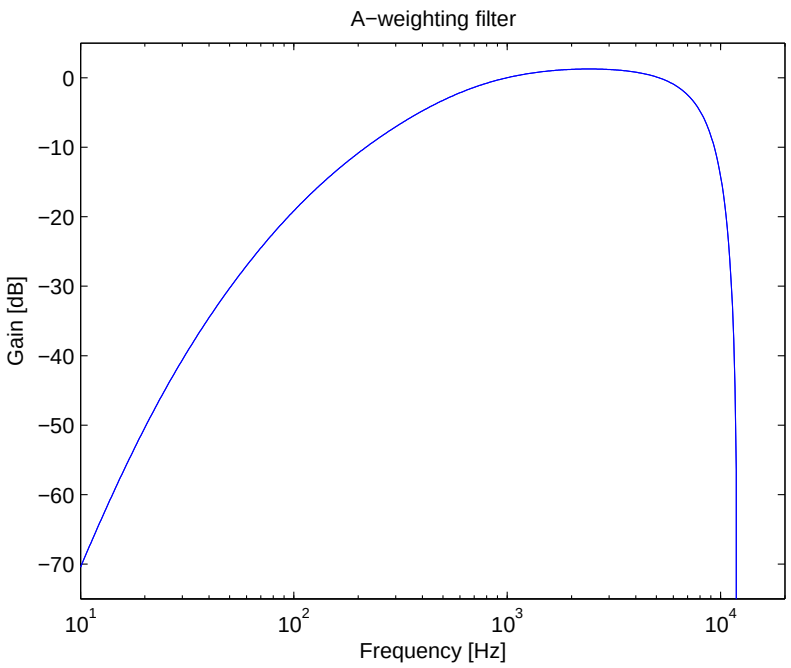
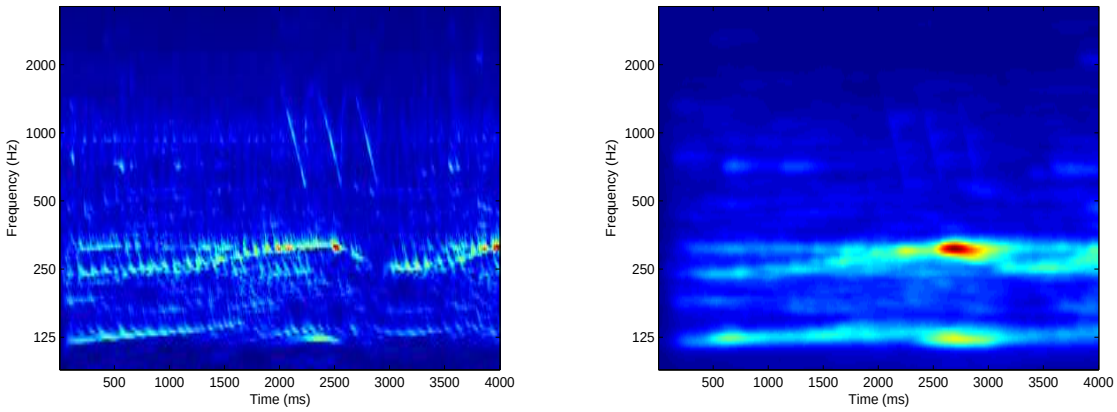


Figure 39: Pre-processing filter characteristic.

are more sensitive to. Thus, as shown in Figure 41, the siren sound becomes very salient, and the noise is suppressed, which is more consistent with what an observer would experience when listening to the sound.

Finally, we needed to determine if adding the new pre-processing stage to the current auditory saliency model resulted in any increase in the correlation between subjects' responses and the model's responses. We first generated the new model's responses for comparison with the subjects' responses from the experiment. The new correlation values for each subject to the model can be found in Table 13.

We found that adding the pre-processing stage to our model resulted in a large improvement in the correlation values of the model to the subjects, particularly in the case of Model 1 (global model). For Model 1, using a paired t- test, we found a statistically significant (p=3.35 * $10^{(}$ − 9)) increase in the average correlation of subjects' responses and the model's responses. For this case, the average correlation more than doubled, with an increase from 0.1291 (p=0.48) to 0.2749 (p<0.12). For



(a) (b)

Figure 40: a) Auditory spectrogram b) Auditory saliency map without new pre-processing stage.

Model 2, the local model, there was a slight increase in the correlation from 0.2406 (p=0.18) to 0.2708 (p=0.12). Here, the filtering did not have as much of an effect, since the filter is applied at the global level.

If we again examine only the scene pairs where more than half of the subjects agreed on the salience, there is a large improvement in the correlation between the model and subjects' responses. For these scene pairs, the model and subjects were strongly correlated with a correlation value of 0.71 for both Model 1 and Model 2.

Adding the new pre-processing stage to the model resulted in an increase in the correlation of the model and subjects' responses, such that the correlation of the model to the subjects is now slightly higher than the average correlation of the subjects to one another. The average correlation of the subjects to one another is one measure that we use to evaluate how well we can expect the model to match subjects' responses. We can use hypothesis testing to show that the model is not distinguishable from the human subjects, in that, the model is as correlated to the subjects as they are to each other. We can also show that prior to adding this new pre-processing stage, the model was distinguishable from the subjects, since the average correlation of the subjects to the model (particularly, Model 1) was significantly lower than the correlation to the

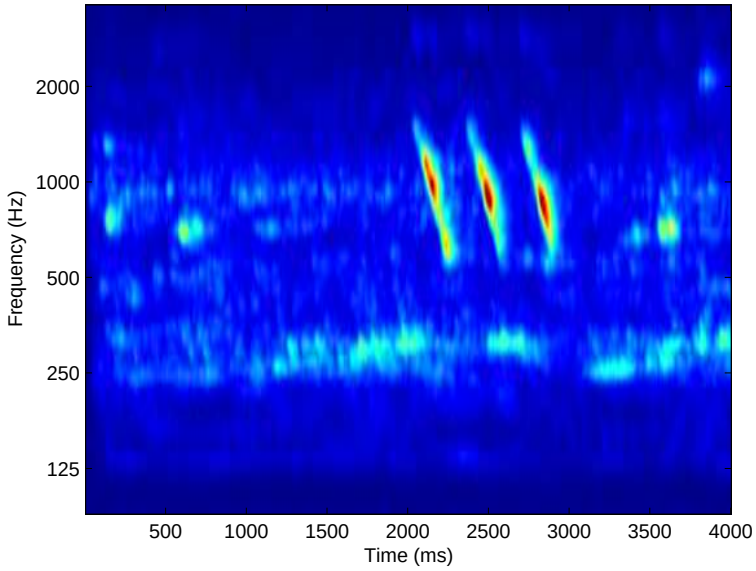


Figure 41: Auditory saliency map with new pre-processing stage.

subjects to each other.

Our null hypothesis is that there is no difference between the average correlation of the model with the subjects and the average correlation of the subjects to each other. The alternative hypothesis is the opposite, which is that we can distinguish the model from the subjects.

For the hypothesis testing, we use a z-test. First, for the global model without the new filtering stage, we found that the test statistic, z, is -4.27. Looking at the z distribution tables, the criteria for rejection of the null hypothesis is $z < -1.96$ and $z > 1.96$ for alpha = 0.05. Since z=-4.27 is greater than 1.96, we can reject the null hypothesis. Therefore, in this case, we can conclude that there is a statistically significant difference (p=0.00002) between the correlation of the global model to the subjects and the average correlation of the subjects to each other. In other words, the model is distinguishable from the subjects, due to the average correlation of the model being significantly lower than the average correlation of the subjects to one another.

Table 13: Correlation between subject and model (with new pre-processing stage) responses for MDS experiment.

Subject	Model 1	Model 2
1	0.3716	0.3037
2	0.4051	0.3956
3	-0.2524	-0.2838
4	-0.1120	-0.1020
5	0.4394	0.4515
6	0.3403	0.2830
7	0.1097	0.1245
8	0.4598	0.4386
9	0.3079	0.3200
10	0.5045	0.4738
11	0.5977	0.5851
12	0.3048	0.3759
13	0.2547	0.2893
14	0.1963	0.1820
15	0.4720	0.4174
16	0.4361	0.4200
17	0.4450	0.4840
18	0.0338	0.0353
19	0.3058	0.2606
20	0.1018	0.1349
21	0.4392	0.3890
22	-0.0186	0.0205
23	0.4137	0.4711
24	0.4337	0.4659
25	0.5395	0.5197
26	0.2253	0.2253
27	0.3123	0.3367
28	0.2313	0.2422
29	0.0459	-0.0413
30	0.3630	0.3727
31	0.3788	0.3306
32	0.4218	0.4551
33	-0.1212	-0.1041
34	-0.0394	-0.0656
Average	0.2749	0.2708
Std Dev	0.2117	0.2127

Next, we can do the same analysis for the global model with the new pre-processing stage added. Here, the average correlation of subjects to the model is closer to the average correlation of the subjects, and we expect that the model will not be distinguishable from the subjects. In this case, the test statistic, z, is 0.48. The critical z value for rejection of the null hypothesis is $z < -1.96$ and $z > 1.96$ for alpha = 0.05. Since z=0.48 is within the acceptable range, in this case, we can no longer reject the null hypothesis. We now find that there is no statistically significant difference between the correlation of the subjects to each other and the correlation of the subjects to the global model. Therefore, the model is as correlated to the subjects as they are to one another.

Looking at the local model, we find no statistically significant difference between the correlation of the subjects to each other and the correlation of the subjects to the

model both when the pre-processing stage is included and when it is excluded.

We also looked at whether the enhanced model resulted in any improvement in the correlation values between the model and subjects for the scene comparison experiment presented in the previous chapter. The results comparing the correlation values with and without the new pre-processing stage can be found in Table 14. Here, we found using the enhanced model did not result in any significant change to the average correlation values between the model and subjects, and the correlation values remained almost the same. The correlation here was 0.4878 ($p<0.08$) for the global model and 0.5256 ($p=0.05$) for the local model.

Table 14: Correlation values between the enhanced model's responses and subjects' responses for saliency scene comparison experiment.

	Correlation to	
Model	**Enhanced model**	**Original model**
1	0.4878	0.4745
2	0.5256	0.5272

One possible explanation for why we did not find any significant improvement with the enhanced model for the saliency scene comparison experiment is that the average correlation of the model to the subjects in this experiment was already higher than the average correlation of the subjects to each other. Thus, this may indicate that we were already close to or at the threshold for how closely we can expect the model to match the subjects' responses. Additionally, there were a total of 20 scene pairs where the enhanced model resulted in a different response from the previous model. If we take a closer look at these 20 scene pairs, more than half of these were scene pairs where there was no subject agreement on which of the two scenes was more salient. Thus, while there was a large increase in the correlation value for some subjects, other subjects had a slight decreases, resulting in very little change overall to the average correlation. For this same reason, there was also almost no change in

the correlation values for scene pairs where there was subject agreement.

In this chapter, we were able to use MDS analysis to find out more about the important dimensions contributing to saliency. Based on the results of the MDS analysis, we added a new stage to our model, which resulted in a large increase in the correlation values, in particular, for the global model. We found a strong correlation of 0.72 for the scenes where there was subject agreement on saliency. Although the correlation values for all the scenes was lower than the correlation values we found in the saliency scene comparison experiment, the correlation values of the subjects to each other was also lower for this experiment as well. As a result of this, we would expect to see lower correlation values, as how correlated the subjects are to each other provides us with an indication of how closely we can expect the model to match subjects' responses.

We were able to show that there was no difference between the average correlation of the subjects to the model and the average correlation of the subjects to each other once the new stage was added. Adding the new stage increased the correlation values of the model to the subjects to the limit that we would expect to match, as there are other top-down and subjective elements that cannot be accounted for with only the bottom-up model. In the next chapter, we will apply our model to scenes that have both an audio and video component. One thing we would like to investigate with this next experiment is whether or not the most salient audio segment of a movie scene can be representative of the entire scene.

CHAPTER VI

AUDITORY SALIENCY AND VIDEO

In this chapter, we apply the computational auditory saliency model to scenes with both an auditory and visual component. We were interested in determining whether or not sound, in particular, salient sounds could be used to create a summary of a movie scene by identifying the best "thumbnail" segment to represent each scene. There are two key differences in the auditory stimuli used for this experiment compared to those used for previous experiments. The first difference is that with this experiment, for the first time, we apply our model to scenes containing speech, so that we could evaluate how well the model performs in selecting salient elements from speech. Since the auditory saliency model is a bottom-up processing model, we expect that it may not perform as well on scenes that are mainly speech, due to the fact that speech is largely dependent on the context and would involve more of the higher level, top-down processing mechanisms that are not taken into account by the current model. The second difference in the stimuli for this experiment is that the soundtracks used are two-channel, stereo sounds, as opposed to the monaural stimuli used in previous experiments.

6.0.4 Subjects

Using the results supplied from the earlier power analysis, we wanted to have at least 32 subjects complete the experiment.

Results were obtained from 36 (19 female, 17 male) undergraduate students at the Georgia Institute of Technology under IRB approval. The students received course credit for participation in the experiment. Normal hearing was determined by self-report. Subjects were informed about the general aim of the experiment, but were

naive to the exact purpose of the study.

Subjects listened to the scenes through headphones (Sennheiser HD 280 pro). A MATLAB GUI was used to both present the scenes to the subjects and also to record subjects' responses. The experiment took each subject approximately 40 minutes to complete.

6.0.5 Experimental Procedures

Each subject was presented with 70 five-second audio and video scenes. The short movie scenes were created from 10 movies of different genres which varied from animated films to action films. We wanted to use movies with interesting audio soundtracks. The movies were first broken up into five-second scenes. We then separated out the scenes with interesting auditory elements, as there were many five-second segments that contained no sounds and consisted only of silence. For this experiment, we then randomly selected several scenes for each movie from all the scenes that were found to have interesting auditory elements.

The experiment itself was divided into two different parts. The first part of the experiment involved using only the audio soundtrack from each movie scene, while the second part involved showing the video along with the corresponding audio soundtrack. Each of the five-second scenes were divided into five non-overlapping one-second segments that were identified by a number from 1 to 5. This number was displayed at the top left corner of the screen. A sample of one of the movie scenes used for this experiment is shown in Figure 42.

Subjects were asked to answer a total of three questions, which are summarized in Table 15. In the first part of the experiment, subjects were presented with only the audio soundtrack. They were asked to listen to the soundtrack while being shown a black screen. At the top left corner of the screen, the numbers (1 to 5) identifying the segments were shown. They were then asked to indicate which number (1 to 5)

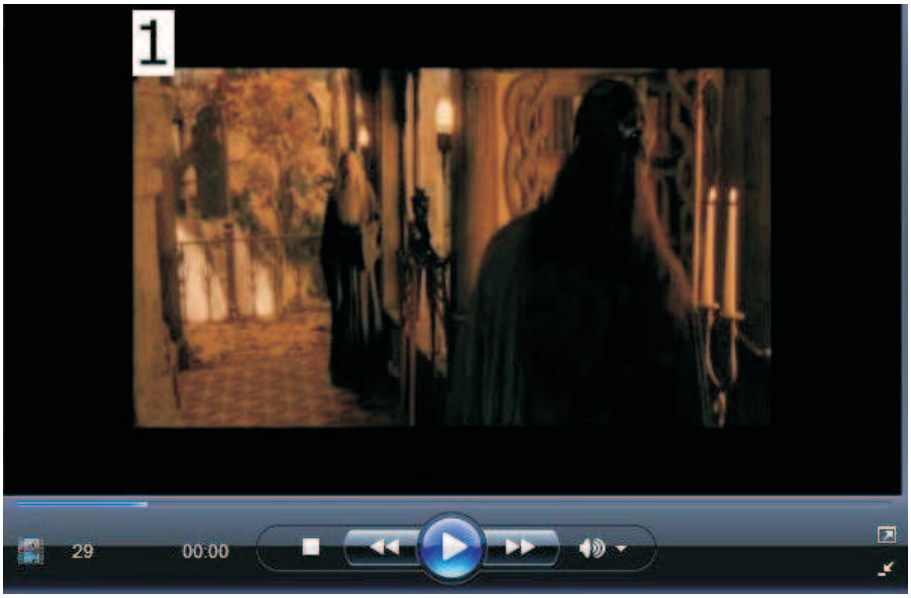


Figure 42: Sample of movie scene used in the experiment.

was displayed when they heard the sound they thought was the most salient sound. Subjects were allowed to replay each scene as many times as they needed to.

In the second part of the experiment, subjects were shown the video along with the corresponding soundtrack. Again, there was a number (1 to 5) displayed in the top left corner of the screen. In this part of the experiment, subjects were asked to respond to two separate questions. They were first asked again about which segment had the most salient sound considering both the video and corresponding audio. One reason we asked subjects to answer this question again is that we were interested in seeing if being shown the video with the corresponding audio had any effect on the sound they thought was the most salient. Finally, we also asked subjects to indicate which numbered segment they felt best represented the video scene. From the subjects' responses to this question, we wanted to investigate whether or not the most salient sound segment is representative of the entire video scene.

Prior to starting the experiment, subjects were also given a brief training session where they were shown sample scenes in order to acquaint themselves with the task. The training consisted of three sample audio scenes and three sample video and audio

Table 15: Summary of questions asked to subjects in the video experiment.

Experiment part	Question
Audio only	Which segment contains the most salient sound?
Audio and Video Q1	Considering both the video and corresponding audio, which segment contains the most salient sound?
Audio and Video Q2	Overall, which segment best represents the scene?

scenes.

6.0.6 Results and Discussion

For this experiment, two-channel, stereo audio inputs were used. Although the model was not created for binaural sounds, this experiment demonstrates how the model can possibly be useful in selecting salient stimuli from binaural sources as well. Figure 43 shows a diagram of how the saliency maps are obtained for binaural sounds. Inputs with multiple audio channels are evaluated by processing each of the channels separately. Therefore, we obtain one intermediate saliency map from each channel. In this case, we have two intermediate saliency maps which give the salient elements from each channel (ear), which are then combined into one final saliency map.

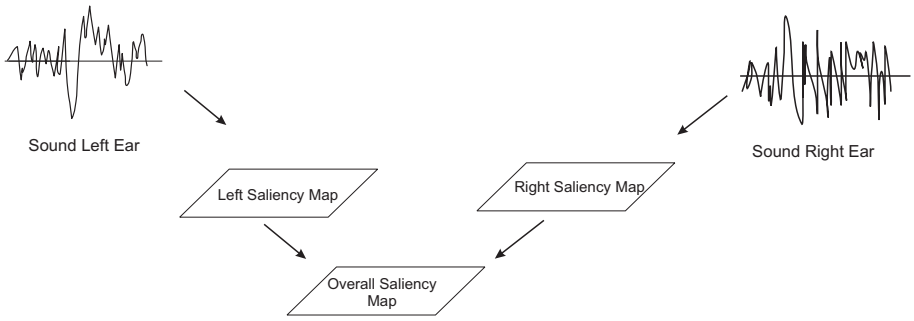


Figure 43: Saliency map processing for binaural sounds.

To demonstrate how the saliency maps are obtained for binaural sounds, we show in Figure 44, the saliency maps obtained for a five-second binaural recording of bird sounds. With this example, the final saliency map is very similar the the map obtained

from the right channel. For short natural scenes such as this one, it is often the case where one channel will make a larger contribution to the final saliency map, since there may not be salient sounds approaching simultaneously from both directions.

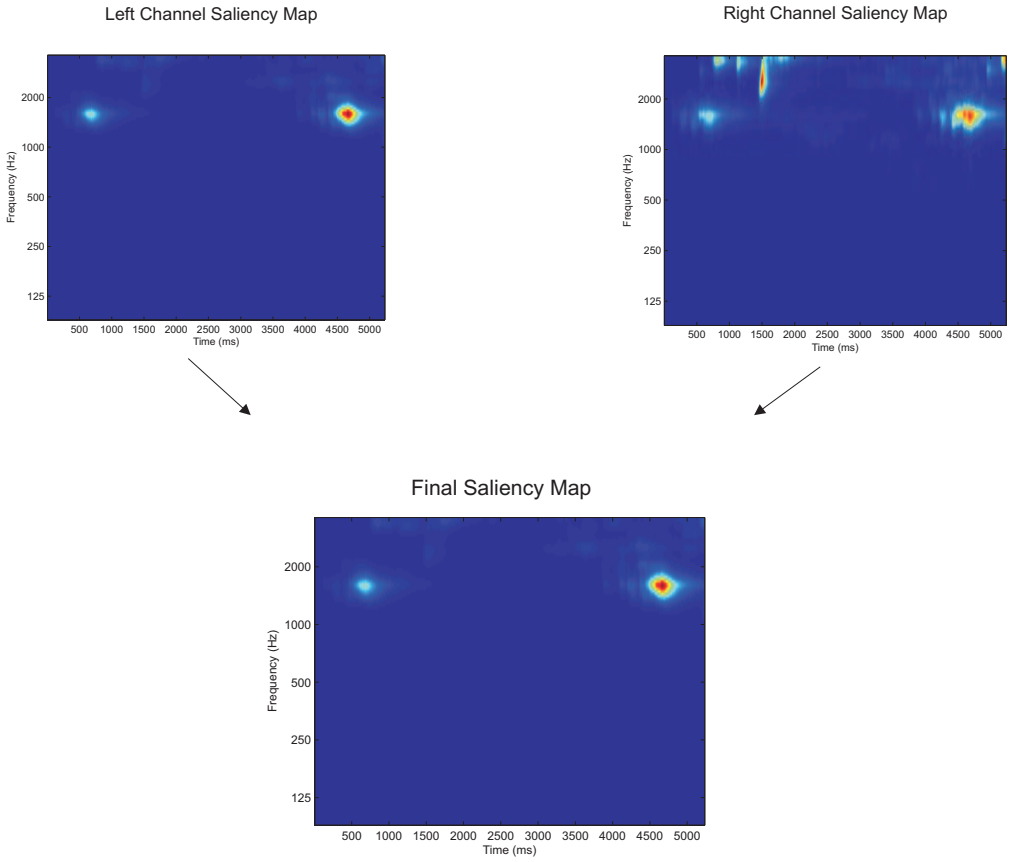


Figure 44: Example of combined saliency map for binaural bird sounds.

Next, we present an example of where the final saliency map differs from the two intermediate maps. In Figure 45, we show the saliency maps for a five-second binaural recording of clapping sounds. A total of five claps can be heard on the clip, and the claps occur at different positions around the listener. They start from the left of the listener, and they gradually move towards the right.

In Figure 45, looking at the left saliency map, the second clap is the most salient clap. The third clap occurs from in front of the listener and is not very salient on either the left or right saliency maps. Moving to the right saliency map, the two claps to the right of the listener can be seen on the map with the first clap occurring

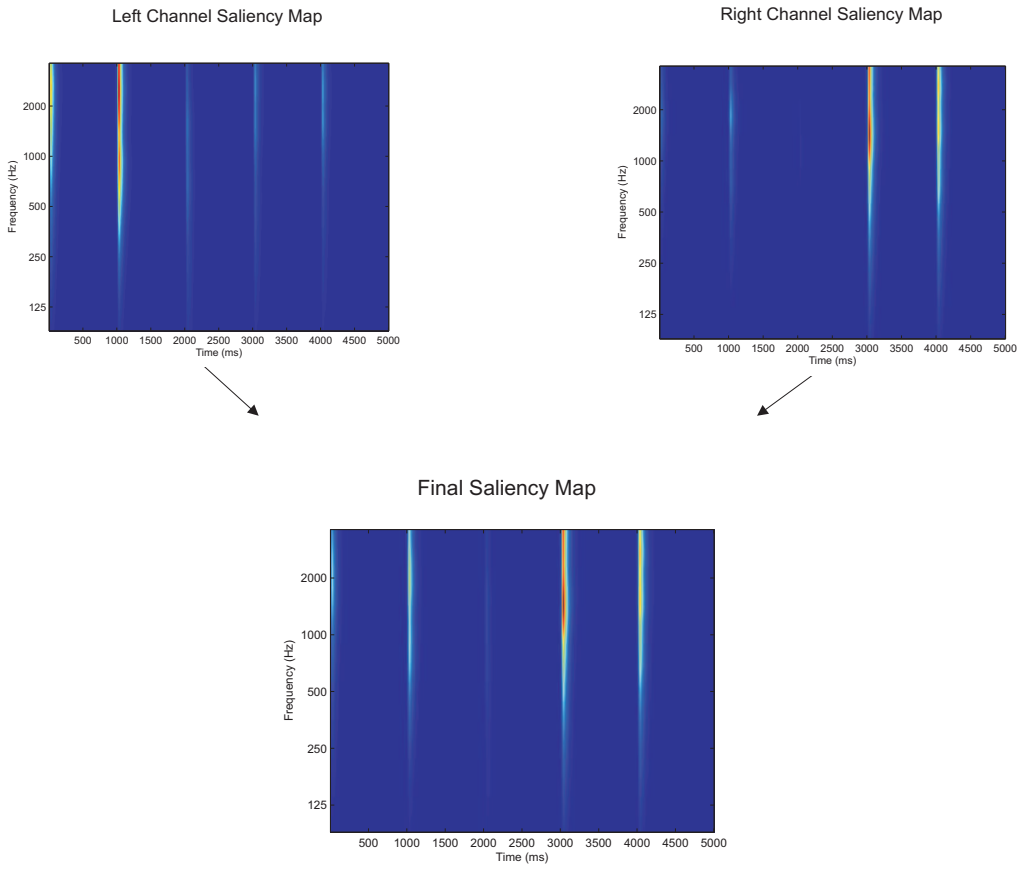


Figure 45: Example of combined saliency map for binaural clapping sounds.

from the right being more salient than the second one. The final saliency map shows all five claps, but the first clap when the sound is to the right of the listener is the most salient clap. This is expected as the clapping sound is now approaching from a new direction, and as a result, stands out more perceptually. This example also demonstrates how some information about the direction and location of sounds can potentially be incorporated when determining the saliency map.

We will first look at the results from the audio only portion of the experiment. Table 16 shows the correlation values between the segment that subjects selected as the most salient and the segment that the model selected as the most salient. The model used for comparison to the subjects in this experiment is the global model with the filtering stage discussed in the previous chapter. In general, the most salient segment selected was the same for both the global model and local model.

Table 16: Correlation between subject and model responses for audio only portion.

Subject	Model 1	Subject	Model 1
1	0.4402	19	0.4821
2	0.4313	20	0.5556
3	0.5611	21	0.4646
4	0.5507	22	0.5681
5	0.2800	23	0.5951
6	0.5883	24	0.3266
7	0.5121	25	0.6356
8	0.3525	26	0.5194
9	0.6300	27	0.5595
10	0.5263	28	0.3129
11	0.4292	29	0.4314
12	0.4242	30	0.5917
13	0.4760	31	0.4113
14	0.4672	32	0.3053
15	0.6520	33	0.2265
16	0.6092	34	0.4651
17	0.3986	35	0.3918
18	0.5582	36	0.3934

Average	0.4756
Std Dev	0.1087

When only the audio soundtrack is presented, we found the model does well in selecting the segment that subjects considered the most salient. The average correlation between the model's responses and subjects' responses was 0.4756, which is statistically significant (p=0.003). The average correlation found is similar to the correlation we found in the saliency scene experiment. We had expected that the correlation values might be lower for the video scenes, since they are dependent to some degree on the context of the scene which imply more involvement from the top-down processing mechanisms, but we found the bottom-up model performs well in selecting the salient segments.

As saliency can be subjective, it is possible that some of the scenes do not have a single segment that is clearly salient to listeners. For these scenes, there would be a

large variation in the subject response, and as a result, they would not be useful for validating or invalidating the computational saliency model. Therefore, in order to reduce some of the subject variation, we also looked at only the scenes where there was agreement amongst subjects on which segment was most salient. We used two different approaches for this analysis: 1) using the modal value and 2) using the scenes with majority agreement.

- Modal Value - Segment most frequently selected as salient by subjects for each scene was compared to the model's response
- Majority - Scenes where more than half of subjects agreed on which segment was salient were compared to the model's response

Table 17 shows the correlation values where there was agreed salience for both the audio only and audio and video portions of the experiment. Here, we found a strong correlation between the model's responses and subjects' responses. When only the audio was presented, the correlation between our model and the modal value of subjects' responses was 0.7237 and the correlation between the model and the majority response was 0.7041.

Table 17: Correlation of model to scenes with agreement on saliency.

Description	**Correlation Model to**	
	Mode	**Majority**
Audio Soundtrack Only	0.7237	0.7041
Audio and Video Question 1	0.6010	0.7320
Audio and Video Question 2	0.4553	0.8683

It should be noted also that the most salient segment for subjects was also typically considered to be very salient by the model, even if it was not the segment selected by the model as the most salient segment. One possible explanation for this is that in several of the scenes, a salient sound would often span across several one-second

segments, starting in one segment and ending in another segment. In these instances, the model and subjects could differ in which part of the sound (onset, middle, end) they found most salient. In order to see how salient the model found the segment that subjects selected as the most salient segment, we can look at where the model ranks (from 1 to 5) each segment. A ranking of 1 means that the model agreed with subjects on which segment was the most salient segment. A ranking of 5, on the other hand, means that the segment subjects chose as the most salient segment was the least salient segment for the model. Looking at this, we found that the average ranking was 2.09 with a standard deviation of 0.7. This means that the segment that was selected most salient by the subjects was on average either the same as the one the model selected or it was ranked as the second most salient segment according to the model. If we compare this to a random uniform model, the average ranking for that is 3.0 with a standard deviation of 1.41. To determine if there is a significantly significant difference between these two means, we perform a z-test. We found that the test statistic, z, is -4.92. Looking at the z distribution tables, the criteria for rejection of the null hypothesis is $z < -1.96$ and $z > 1.96$ for alpha = 0.05. Since z=-4.92 is less than -1.96, we can reject the null hypothesis. Thus, we did find a statistically significant difference ($p=9 * 10^{-}7$) between the two means and can conclude that our model performs better than a random model in selecting segments that subjects found salient.

Figure 46 shows an example of one of the movie scenes from the experiment where the subjects were divided on which segment contained the most salient audio. Here, we show the last frame from each of the 5 different segments for this video scene. This scene demonstrates how the model is able to select the segments that are found to be salient to subjects. In this example, subjects were almost equally split between whether segment 1 or segment 3 was the most salient segment. In the audio only portion, both segment 1 and segment 3 were each selected by 12 different subjects.

For this same question in the video portion, 11 subjects selected segment 1, and 12 subjects selected segment 3 as the most salient segment. Interestingly, for the second question regarding which segment was most representative of the scene, subjects were not divided, and they strongly agreed on segment 3 being most representative of the scene.

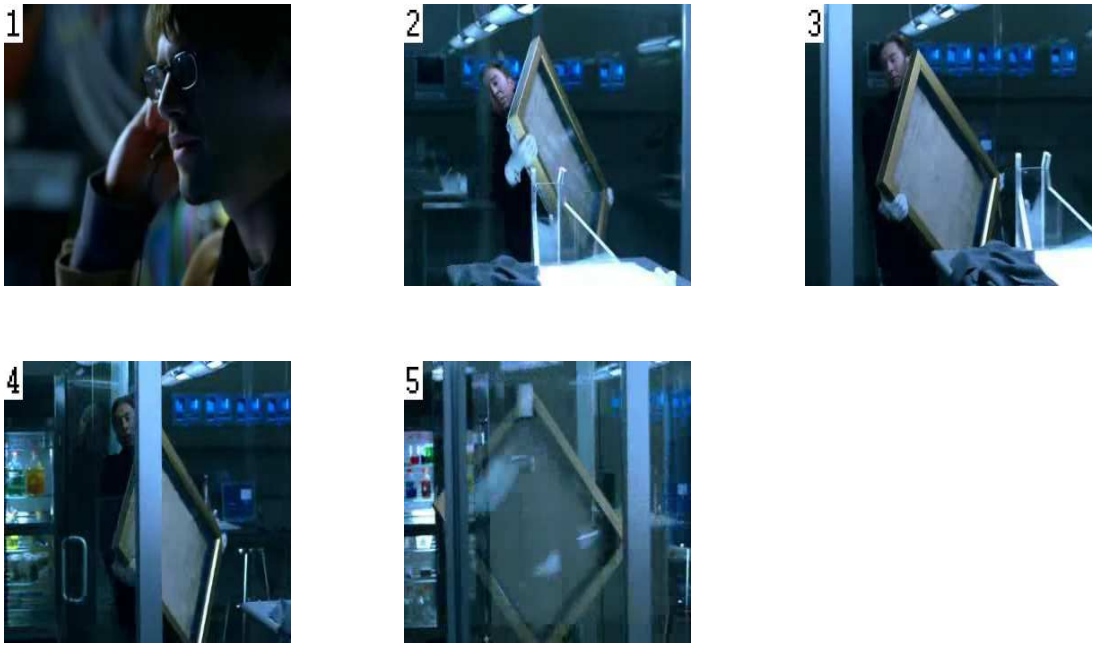


Figure 46: Sample video frames for one of the movie scenes used in the experiment.

The number of subjects that selected each segment as most salient can be used to create a ranking for the degree of saliency of the different segments. In this case, segments 3, 1, and 5 were the three more salient segments. Almost no subjects selected segments 2 or 4 as being the most salient segment, indicating that those two segments were of lower saliency. We can then compare this information to the auditory saliency map generated by the model and shown in Figure 47. Here, according to the model, segment 3 is clearly the most salient segment, but we can see from the saliency map that segments 1 and 5 are salient as well. Segment 5 was the third most salient segment for subjects, with 8 subjects selecting it as being the most salient segment

for them. This example shows how the model is able to match the rankings for the degree of saliency of various segments.

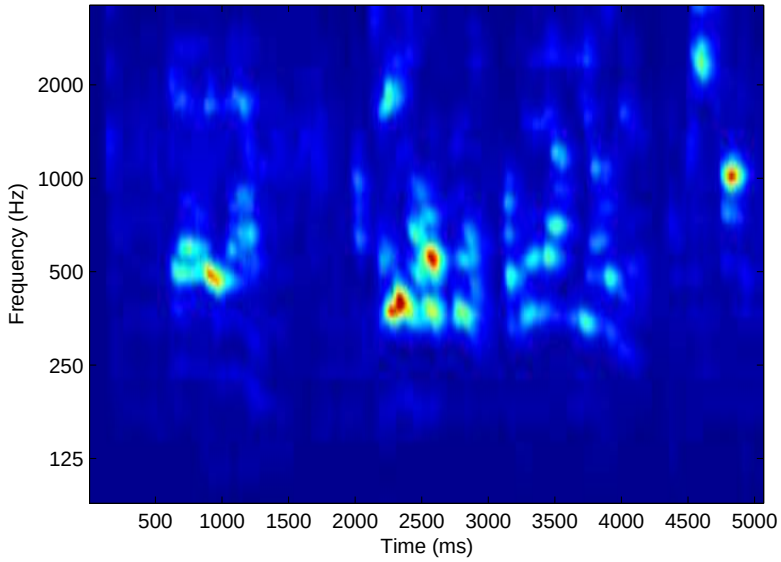


Figure 47: Auditory saliency map for the movie scene shown in Figure 46.

In the video portion of the experiment, subjects were presented with both the video and corresponding audio soundtrack for each movie scene. They were then required to answer two separate questions. Table 18 shows the correlation values found between the model's responses and subjects' responses for both of the questions asked. Looking at the responses to the first question, which asked subjects to indicate the segment with the most salient sound, the correlation value between the model and subject responses were slightly lower than the correlation found in the audio only portion. The average correlation found here was 0.4394 (p=0.007) compared to 0.4756 (p=0.003) when only the audio was presented. Similarly, if we look at the scenes where there was agreed salience, we again see a large increase in the correlation values. The correlation between the model and the modal value of subjects' responses was 0.6010 and the correlation between the model and the majority response was 0.7320.

The second question in the video portion of this experiment asked subjects to choose the segment they considered most representative of the video scene overall.

Table 18: Correlation between subjects' responses and the model's responses for both question 1 and question 2 of the video portion.

Subject	Video Q1	Video Q2
1	0.4378	0.0987
2	0.3457	0.3822
3	0.6457	0.4521
4	0.4915	0.3879
5	0.2312	0.2325
6	0.5399	0.1182
7	0.5989	-0.0088
8	0.3143	0.5338
9	0.5206	0.3429
10	0.5863	0.1035
11	0.3205	0.0639
12	0.4030	0.1620
13	0.3904	0.2203
14	0.4728	0.0941
15	0.4175	0.2730
16	0.6169	0.2524
17	0.3061	0.1804
18	0.3431	0.1382
19	0.5666	0.0310
20	0.4884	0.2958
21	0.4182	0.1090
22	0.5374	0.4119
23	0.4203	0.1467
24	0.3806	0.2217
25	0.2451	0.3818
26	0.5296	0.3648
27	0.4791	0.2151
28	0.4823	0.3971
29	0.2150	0.2000
30	0.5846	0.2978
31	0.4234	0.3161
32	0.4016	0.1893
33	0.3530	0.2029
34	0.4461	0.4055
35	0.4110	0.0132
36	0.4542	0.2405
Average	0.4394	0.2352
Std Dev	0.1093	0.1341

For this question, the average correlation found was much lower. Here, the correlation between the model's response and subjects' responses was 0.2352 (p=0.16). The lower correlation values here were expected, and one explanation for this is the increased reliance on top-down cues. To determine which segment is most representative of the scene, subjects will take into account what the scene is about or in other words, the context of the overall scene. In addition, this question is more subjective than the first question, and subjects can consider a segment to be representative of a scene for various individual reasons. This corresponds with the finding that, for this question, there was much less subject agreement amongst the subjects. They differed greatly on which segment they selected as being most representative of the scene, and out of the 70 scenes presented, there were only a total of 12 video scenes where the majority

of subjects agreed upon which segment was most representative of the scene. For these 12 scenes, where there was agreement, there was a strong correlation of 0.8683 between the model's response and subjects' responses. This result shows our model is able to choose the salient audio cues that can be used to summarize or represent a scene.

In order to determine how well we can expect the model to match subjects' responses, we examine the average correlation values for each subject to the rest of the subjects. Table 19 shows the average correlation values for each subject to all other subjects. We found that the correlation values for the second question of the video portion was much lower than the correlation values found for the first question from the video portion and from the audio only portion. Here, the average correlation of the subjects to one another was only 0.19 compared to approximately 0.42 for the responses to the question regarding which segment had the most salient sound. Due to the large amount of subject variation in the responses for this question, we did not perform any further analysis on the data collected from this portion of the experiment. The reason for this decision was that we wanted to see if the segment with the most salient sound would be representative of the scene overall. As the subjects themselves were not in agreement as to which segment was most representative for each scene, we could not expect the model to match subjects' responses.

Table 19: Average correlation of subjects to each other.

Experiment	**Correlation**
Audio soundtrack only	0.4227
Video and Audio Question 1	0.4238
Video and Audio Question 2	0.1850

In this experiment, speech was used as part of the stimuli set, and the model was evaluated for the first time on its ability to choose salient cues from speech. As speech can be very contextually dependent, we expected that with speech the saliency of the

stimulus would start to become increasingly dependent on top-down influences, but we found that the model performed surprisingly well in selecting salient cues from speech. There were several movie scenes (39 scenes) where the soundtracks were comprised of majority speech. Scenes were defined as majority speech if more than three seconds of a scene involved speech.

Looking at how the model performed on these scenes containing mainly speech, we found what can be considered a strong correlation by Cohen's guidelines [**?**]. In Table 20, we show the the average correlation between the model's responses and subjects' responses for both the audio and video portion of the experiment. When only the audio was presented, the average correlation was 0.5591, which we found to statistically significant (p=0.0004). We also found a similar correlation value when the same question was asked in the video portion. The correlation there was 0.5431 (p=0.0006). For the second question, regarding which segment was most representative of the scene, the correlation was again much lower. The correlation in this case was 0.1850 (p=0.19).

Table 20: Average correlation of model to subjects for majority speech scenes.

Experiment	**Correlation**
Audio soundtrack only	0.5591
Audio and Video Question 1	0.5431
Audio and Video Question 2	0.2211

We can also look at speech scenes where there is subject agreement. Table 21 shows the correlation of the model's responses to the modal value of the subjects' responses as well as the correlation for the scenes where there was majority agreement on the most salient segment. Looking at the correlation for scenes where the majority of subjects agreed upon which segment was most salient, the correlation value was extremely high at 0.8956 for the audio only portion and 0.8896 for the when both the audio and video was presented. For the question 2 of the video portion, there were

only 5 scenes where the majority of subjects agreed upon which segment was salient. For all 5 of these scenes, the model was able correctly match the subjects' responses.

Table 21: Correlation of model to scenes with agreement on saliency for majority speech scenes.

Description	Correlation Model to Mode	Correlation Model to Majority
Audio Soundtrack Only	0.8956	0.8857
Audio and Video Question 1	0.7894	0.8896
Audio and Video Question 2	0.4293	1.000

One explanation for why the top-down cues had little effect on the model's performance in selecting salient sounds is the relatively short duration of the scenes. With the scenes being only 5 seconds long and divided into 1 second segments, the context of the speech has less of an effect on saliency than if the scenes were of longer duration.

We showed in this experiment that the model is able to select salient auditory segments from video that match well with selections made by subjects. In the next chapter, we will conclude the thesis by summarizing the main contributions. We also provide a brief discussion on some areas for continued research into the topic of auditory saliency.

CHAPTER VII

CONCLUSIONS AND FURTHER EXTENSIONS OF THIS WORK

7.1 Thesis Overview

Auditory saliency research is still a relatively new area of research which has been rapidly gaining interest in recent years. In this thesis, we provide a comprehensive look at the factors that contribute to saliency for various sounds and present a bottom-up processing auditory saliency model that does well in selecting sounds that humans perceive as being salient.

The first contribution of the thesis is our novel computational auditory saliency model. In Chapter 3, we presented a bottom-up auditory saliency model, which uses cortical features similar to those our brain potentially uses in processing sounds. These features were generated using a cortical model that mimics the response properties of cortical neurons. The model also differs from existing auditory saliency models, in that it takes into consideration the fact the certain frequencies can be perceived as louder, and as result, be more salient. This is something not accounted for by the other auditory saliency models and can make a large impact on saliency. Additionally, our model is a general auditory saliency model that matches well with results from human subjects on selecting salient sounds from a scene. The model can also be used for a variety of applications, which we show in Chapter 6, where we apply it to video scenes. Many of the other currently existing auditory saliency models have been tailored to a specific task or application and have only been evaluated based on the specific tasks. As those models have only been tested on the one application, we do not have any knowledge of the model's performance on other tasks or how it

performs, in general, at selecting or ranking the saliency of different sounds from a complex auditory scene.

Another novel contribution, presented in Chapter 5 of the thesis, is the exploration into the dimensions that contribute to the salience of different sounds. This experiment and the analysis using multidimensional scaling (MDS) is the first to look specifically into determining what the key factors are that make a sound salient. The information from this analysis was then used to improve the auditory saliency model and create a more comprehensive feature set for the model. In particular, we were able to improve the model's performance by adding a pre-processing stage based on the information obtained through the MDS analysis. The new stage resulted in a significant improvement in the correlation values between the model's responses and subjects' responses for picking salient scenes.

Presenting the dual task paradigm as a way to evaluate auditory saliency is another contribution of this thesis. Evaluating or quantifying auditory saliency which was discussed in Chapter 4 remains a key issue. In auditory saliency, there is no easily tracked physical correlate that can be used to evaluate saliency the way eye tracking is used in visual saliency. In this thesis, we presented two different experiments that were used to evaluate the model, and both of these experiments can be used to measure the salience of different stimuli. In particular, we show how the dual task experiment paradigm can be used to evaluate auditory saliency.

Dual task experiments have been used in many different studies involving attention in the different modalities, but it has not yet been used to investigate auditory salience. The computational auditory saliency model presented in this thesis models the bottom-up, saliency-driven part of our attentional process. In this bottom-up process, sounds that are salient can be noticed without our attention. We show that the dual task paradigm is a useful way to evaluate this type of bottom-up auditory saliency. We also show how using a dual task experiment is helpful in determining

what changes can be made and the degree of change required to make a sound become salient to an observer. In the second experiment presented in Chapter 4, we use a pairwise presentation of sounds to evaluate the model's performance. We show that model does well in selecting sounds that subjects considered salient. There was a strong correlation between the model's responses and subjects' responses, especially for the sounds where there was agreed salience among the subjects.

Finally, the use of the auditory saliency model for selecting salient auditory segments from video is another contribution of this thesis. Here, we show the diverse applications the model can be used for. The experiment, as described in Chapter 6, is the first time auditory saliency maps have been used to evaluate salient audio in movie soundtracks. In addition, the stimuli used in this experiment were two-channel audio inputs. The use of the model on these two-channel stereo sounds as well as on binaural sounds is another key contribution of this thesis, since the existing auditory saliency models have all been used primarily for monaural sounds. In this experiment, we demonstrate how our model can be used on audio inputs with more than one channel, and we show that the model performs as well in selecting salient segments for these sounds as it did for the monaural sounds. With this experiment, we also applied the model to speech for the first time. The results obtained show that for short video scenes the model does well in selecting salient segments from speech.

7.2 Further Extensions of this Work

The computational auditory saliency model presented here performs well in detecting salient auditory events using only bottom-up processing cues. Although these bottom-up cues can account for a large portion of what makes a sound salient, there are still many instances and applications where the saliency of a stimulus will rely as much, if not more, on the top-down input. In addition, saliency can also be very task

dependent. The type of stimuli that the model is to be used on and the task or application that is being performed with the model can both influence which of the cues between the bottom-up and top-down are more important. Thus, the applications of the auditory saliency model is one area for future research. In terms of model development, the future direction of this work would be towards incorporating top-down input into the model depending upon the application the model is to be used for as well as to further expand the feature set. Below we present a brief discussion on two possible ways top-down feedback could be incorporated with the bottom-up model along with some preliminary findings using a boosting algorithm.

In Figure 3, we showed how top-down feedback could interact with the bottom-up cues in the feature selection and combination stage. One way top-down feedback could be incorporated here is to use learned weighting of the feature maps. The information obtained from learning could be used as weights for the individual feature maps or for the global feature maps prior to combining them into the final saliency map. In this way, instead of simply summing all the maps as is currently done with the bottom-up model, each individual map would be weighted differently based on the weighting parameters obtained from learning. The learned weights would then influence the contribution of each individual feature map to the overall saliency map. Using the learned information to as top-down weighting factors also allows us to tailor the model to promote certain types of sounds depending on the task being performed. One possible implementation is to use a boosting algorithm, such as Adaptive Boosting (AdaBoost) [63] to obtain the weights. The idea behind AdaBoost is for each subsequent weak classifier to focus on the instances where misclassifications occurred during the previous classifiers. This algorithm was modified by Tieu and Viola [75] for use as a feature selector by restricting the weak classifier to work on only one feature at at time [60].

The results from the scene comparison experiment presented in Chapter 3 were

used to train the AdaBoost based saliency model. The average of the responses from all subjects were used as the ground truth, and the difference between the features was used to represent each scene. This enables us to predict what features make one sound more salient as opposed to another sound. Preliminary runs were performed where the global feature maps were weighted prior to combination into one final saliency map. We found there was a slight improvement over the strictly bottom-up saliency model. One disadvantage to the AdaBoost-based system is that it would requires a large amount of training data, while the bottom-up model achieves similar results with no training required.

Another possible way to incorporate top-down input into the model is to use "template" matching. Here, we would have various "templates" for different situations. For example, depending on the task to be completed, we could train the auditory saliency model to "listen" or look for certain types of sounds or situations. In this way, we can bias the model such that depending on the task being performed, certain types of sounds would be much more salient than other sounds. For example, drums playing during a football game is less salient than drums playing elsewhere. We could use the template matching to identify the type of sounds or situations, and then depending on the identification, individual feature maps could be promoted or demoted.

Finally, as with the visual saliency models, there always remain areas for ongoing research into how the feature set can be improved to make a more comprehensive model. As more information becomes known about the primitives for audio, new features could be considered for the model.

REFERENCES

[1] Alain, C. and Arnott, S., "Selectively attending to auditory objects," *Frontiers in Bioscience*, vol. 5, pp. 202–212, 2000.

[2] Atlas, L. and Shamma, S., "Joint acoustic and modulation frequency," *Eurasip Journal on Applied Signal Processing*, vol. 2003, pp. 668–675, June 2003.

[3] Braun, J. and Julesz, B., "Withdrawing attention at little or no cost:detection and discrimination of tasks," *Percept Psychophys*, vol. 60, pp. 1–23, 1998.

[4] Bregman, A., *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1990.

[5] Calhoun, B. and Schreiner, C., "Spatial frequency filters in cat auditory cortex," in *Proceedings of the 23rd Annual Meeting Society of Neuroscience*, 1993.

[6] Chang, E., Rieger, J., Johnson, K., Berger, M., Barbaro, N., and Knight, R., "Categorical speech representation in human superior temporal gyrus.," *Nature Neuroscience*, vol. 13, no. 11, pp. 1428–1433, 2010.

[7] Cherry, E. C., "Some experiments on the recognition of speech with one and two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[8] Chi, T., Gao, Y., Guyton, M., Ru, P., and Shamma, S., "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.

[9] Chi, T., Ru, P., and Shamma, S., "Mulitresolution spectrotemporal analysis of complex sounds," *Speech Communication*, 2003.

[10] Chittka, L. and Brockmann, A., "Perception space–the final frontier.," *PLos Biol*, vol. 3, no. 4, p. e137. doi:10.1371/journal.pbio.0030137, 2005.

[11] Coath, M., Denham, S., Smith, L., Honing, H., Hazan, A., Holonowicz, P., and Purwins, H., "Model cortical responses for the detection of perceptual onsets and beat tracking in singing," *Connection Science*, vol. 21, no. 2, pp. 193–205, 2009.

[12] Coensel, B. and Botteldooren, D., "A model of saliency-based auditory attention to environmental sound.," *Proceedings of 20th International Congress on Acoustics*, 2010.

[13] Coensel, B., Botteldooren, D., Muer, T. D., and Nilsson, M., "A computational model for auditory saliency of environmental sound.," *Journal of the Acoustical Society of America*, vol. 125, no. 4, p. 2528, 2009.

[14] Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Eribaum Associates, 1988.

[15] Cohen, J., "A power primer.," *Psychological Bulletin*, vol. 112, no. 1, pp. 155–159, 1992.

[16] Corso, J., "Age and sex difference in pure tone thresholds," *Journal of the Acoustical Society of America*, vol. 31, no. 4, pp. 498–507, 1959.

[17] D. Parkhurst, K. L. and Nieber, E., "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[18] Dennis, D. and O'Leary, M., "Do cortical areas emerge from a protocortex," *Trends in Neurosciences*, vol. 12, no. 10, pp. 400–406, 1989.

[19] Depireux, D., Simon, J., Klein, D., and Shamma, S., "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 85, pp. 1220–1234, 2001.

[20] Duangudom, V. and Anderson, D., "Audio saliency," in *Workshop on Neuromorphic Engineering report*, pp. 62–70, 2004. http://ine.ini.unizh.ch/telluride/previous/report04.pdf.

[21] Duangudom, V. and Anderson, D., "Using auditory saliency to interpret complex auditory scenes," in *153rd Acoustical Society of America Meeting Abstract*, 2007. http://asa.aip.org/saltlakecity/Wednesdayam.pdf.

[22] Duangudom, V. and Anderson, D., "Using auditory saliency to understand complex auditory scenes," *Proceedings of the European Signal Processing Conference (EUSIPCO-2007)*, pp. 1206–1210, 2007.

[23] Duangudom, V., Francis, G., and Herzog, M., "What is the strength of a mask in visual metacontrast masking," *Journal of Vision*, vol. 7, no. 1, pp. 1–10, 2007.

[24] Elhilali, M., *Neural Basis and Computational Strategies for Auditory Processing*. PhD thesis, University of Maryland, 2004.

[25] Elhilali, M., Xiang, J., Shamma, S., and Simon, J., "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene.," *PLos Biol*, vol. 7, no. 6, p. e1000129, 2009.

[26] Fletcher, H. and Munson, W., "Loudness, its definition, measurement and calculation.," *Journal of the Acoustical Society of America*, vol. 5, pp. 82–108, 1933.

[27] Folkard, S., "Time of day and level of processing," *Memory and Cognition*, vol. 7, no. 4, pp. 247–252, 1979.

[28] Folkard, S. and Monk, T., "Circadian rhythms in human memory," *British Journal of Psychology*, vol. 71, pp. 295–307, 1979.

[29] Gottlieb, J., Kusunoki, M., and Goldberg, M., "The representation of visual salience in monkey parietal cortex," *Nature (London)*, vol. 391, pp. 481–484, 1952.

[30] Green, D., "Frequency and the detection of spectral shape change," in *Auditory Frequency Selectivity* (Moore, B. and Patterson, R., eds.), pp. 351–360, NATO ASI Series, 1986.

[31] Green, D., Kidd, G., and Picardi, M., "Successive versus simultaneous comparison in auditory intensity discrimination," *Journal of the Acoustical Society of America*, vol. 73, no. 2, pp. 639–643, 1983.

[32] Hartmann, W., "Pitch perception and the segragation and integration of auditory entities," in *Auditory Function* (Edelman, G., Gall, W., and Cowen, W., eds.), pp. 623–645, Wiley, 1988.

[33] Hartmann, W., McAdams, S., and Smith, B., "Hearing a mistuned harmonic in an otherwise periodic complex tone.," *Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1712–1724, 1990.

[34] Herzog, M., Duangudom, V., and Francis, G., "Spatial layout determines metacontrast masking [abstract]," vol. 7, p. 1015, 2007. http://journalofvision.org/7/9/1015/,doi:10.1167/7.9.1015.

[35] Hu, R., Hang, B., Ma, Y., and Dong, S., "Spatial audio cues based surveillance audio attention model," *icassp*, pp. 289–292, 2010.

[36] Itti, L., Dhavale, N., and Pighin, F., "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Proceedings SPIE 48th Annual International Symposium on Optical Science and Technology*, vol. 5200, pp. 64–78, 2003.

[37] Itti, L. and Koch, C., "Computational modelling of visual attention," *Nature*, vol. 2, pp. 194–203, 2001.

[38] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[39] J.C. Clarey, P. Barone, T. I., "Physiology of thalamus and cortex," in *The Mammalian Auditory Pathway: Neurophysiology* (Popper, A. and Fay, R., eds.), pp. 411–460, Springer-Verlag, 1992.

[40] Kalinli, O. and Narayanan, S., "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," *INTERSPEECH*, 2007.

[41] Kalinli, O. and Narayanan, S., "A top-down auditory attention model for learning task dependent influences on prominence detection in speech," *ICASSP*, pp. 3981–3984, 2008.

[42] Kayser, C., Petkov, C., Lippert, M., and Logothetis, N., "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, pp. 1943–1947, 2005.

[43] Koch, C. and Niebur, E., "Computational architectures for attention," in *The Attentive Brain* (Parasuraman, R., ed.), MIT PRESS, 1998.

[44] Koch, C. and Ullman, S., "Shifts in selective visual attention: Towards the underlying neurocircuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[45] Kowalski, N., Depireux, D., and Shamma, S., "Analysis of dynamic spectra in ferret primary auditory cortex. i. characteristics of single-unit responses to moving ripple spectra," *Journal of Neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[46] Kowalski, N., Depireux, D., and Shamma, S., "Analysis of dynamic spectra in ferret primary auditory cortex. ii. prediction of unit responses to arbitrary dynamic spectra," *Journal of Neurophysiology*, vol. 76, no. 5, pp. 3524–3534, 1996.

[47] Laird, D., "Relative performance of college students as conditioned by time of day and day of week," *Journal of Experimental Psychology*, vol. 3, pp. 50–63, 1925.

[48] McAdams, S., "Spectral fusion and the creation of auditory images," in *Music, Mind and Brain: The Neuropsychology of Music* (Clynes, M., ed.), Plenum, 1982.

[49] Mesgarani, N. and Chang, E., "Selective cortical representation of attended speaker in multi-talker speech perception.," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[50] Mesgarani, N. and Shamma, S., "Speech enhancement based on filtering the spectrotemporal modulations," in *"Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing"*, 2005.

[51] Mesgarani, N., Shamma, S., and Slaney, M., "Speech discrimination based on multiscale spectro-temporal modulations," in *"Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing"*, 2004.

[52] Mondor, T. and Bregman, A., "Allocating attention to frequency regions," *Percept. Psychophys*, vol. 56, pp. 268–276, 1994.

[53] Moore, B., *An introduction to the psychology of hearing*. London: Academic Press, 1989.

[54] Moore, B. and Bacon, S., "Identification of a single modulated component in a complex sound.," *Journal of the Acoustical Society of America*, vol. 93, no. 4, pp. 2325–2326, 1993.

[55] Moore, B., Peters, R., and Glasberg, B., "Thresholds for hearing mistuned partials as separate tones in harmonic complexes.," *Journal of the Acoustical Society of America*, vol. 80, pp. 479–483, 1986.

[56] Natale, V. and Lorenzetti, R., "Influence of morningness-eveningness and time of day on narrative comprehension," *Personality and Individual Differences*, vol. 23, no. 4, pp. 685–690, 1997.

[57] Niebur, E. and Koch, C., "Control of selective visual attention: Modeling the "where" pathway," *Neural Information Processing Systems*, vol. 8, pp. 802–808, 1996.

[58] Peters, R. and Itti, L., "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[59] Plomp, R., *Aspects of Tone Sensation: a psychophysical study*. Academic Press, 1976.

[60] Ravindran, S., *Physiologically motivated methods for audio pattern classification*. PhD thesis, Georgia Institute of Technology, 2006.

[61] Ravindran, S., Smith, P., Graham, D., Duangudom, V., Anderson, D., and Hasler, P., "Towards low-power on-chip auditory processing," *Eurosip Journal on Applied Signal Processing*, vol. 7, pp. 1082–1092, 2005.

[62] Robinson, D. and Sutton, G., "Age effect in hearing - a comparative analysis of published threshold data," *Audiology*, vol. 18, pp. 320–334, 1979.

[63] Schapire, R. E., "The boosting approach to machine learning: An overview," *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

[64] Schiffman, S., Reynolds, M., and Young, F., *Introduction to Multidimensional Scaling*. Academic Press, 1981.

[65] Schmidt, R. and Lee, T., *Motor Control And Learning: A Behavioral Emphasis*. Human Kinetics, 2005.

[66] Shamma, S., "Spatial and temporal processing in central auditory networks," in *Methods in Neuronal Modelling: From Ions to Networks* (Koch, C. and Segev, I., eds.), pp. 411–460, MIT PRESS, 1998.

[67] Shamma, S., "Physiological basis of timbre perception," in *The New Cognitive Neurosciences* (Gazziniga, M., ed.), pp. 411–423, MIT Press, 2009.

[68] Shamma, S., Fleshman, J., Wiser, P., and Versnel, H., "Organization of response areas in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 69, no. 2, pp. 367–383, 1993.

[69] Speigel, M., Picardi, M., and Green, D., "Signal and masker uncertainty in intensity discrimination," *Journal of the Acoustical Society of America*, vol. 70, no. 4, pp. 1015–1019, 1981.

[70] Sur, M., Garraghty, P., and Roe, A., "Experimentally induced visual projections into auditory thalamus and cortex," *Science*, vol. 242, pp. 1437–1441, 1988.

[71] Sutter, M. and Shamma, S., "The relationship of auditory cortical activity to perception and behavior," in *Auditory Cortex: Fundamental Neuroscience* (Winer, J. and Schreiner, C., eds.), pp. 617–642, Springer, 2011.

[72] Torgerson, W., "Multidimensional scaling i : Theory and method," *Psychometrika*, vol. 17, pp. 401–419, 1998.

[73] Treisman, A. and Gelade, G., "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[74] Tukey, J., *Exploratory Data Analysis*. Addison-Wesley Publishing Co, 1977.

[75] Viola, P. and Tieu, K., "Rapid object detection using a boosted cascade of simple features," *Proceedings of Computer Vision and Pattern Recognition*, pp. 511–518, 2001.

[76] Wang, K. and Shamma, S., "Self-normalization and noise robustness in early auditory representations," *IEEE Transactions on Speech and Audio Processing*, vol. 2, 1994.

[77] Wang, K. and Shamma, S., "Spectral shape analysis in the central auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 382–395, Sept 1995.

[78] Warren, R., Wrightson, J., and Puretz, J., "Illusory continuity of tonal and infratonal periodic sounds.," *Journal of the Acoustical Society of America*, vol. 84, no. 4, pp. 1338–1342, 1988.

[79] Yang, X., Wang, K., and Shamma, S., "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, pp. 824–839, March 1992.

[80] Zotkin, D., Shamma, S., Ru, P., Duraiswami, R., and Davis, L., "Pitch and timbre manipulations using cortical representation of sound," in *"Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing"*, vol. 5, pp. 517–520, 2003.