EXPLORATION OF BEHAVIORAL, PHYSIOLOGICAL, AND COMPUTATIONAL APPROACHES TO AUDITORY SCENE ANALYSIS

A THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Peter Sou-Kong Chang, B.S., B.A.

The Ohio State University

2004

Master's Examination Committee:

Professor DeLiang Wang, Advisor

Professor Eric Fosler-Lussier

Dr. Douglas Brungart

Approved by

_____

Advisor

Graduate Program in Computer and Information Science

ABSTRACT

We present an overview for the study of auditory perception and scene analysis through the three main approaches researchers have used to study perception in general: behavioral, physiological, and computational. At the behavioral level, we discuss the principles and origins of auditory scene analysis, and establish the relationship between auditory scene analysis and auditory masking. Within auditory masking, we note the coexistence of informational and energetic masking, and utilize the ideal time-frequency binary masks in a series of speech intelligibility experiments to isolate the energetic component of speech-on-speech masking. At the physiological level, we propose the adoption of the two-dimensional time-frequency oscillatory correlation representation as a main representation in auditory perception, after reviewing several of the theories and experiments in neurophysiology in effort to find its support. Finally, at the computational level, we extend an existing implementation of oscillatory correlation, LEGION [144], to simulate the major behavioral principles in alternating-tone sequences. Most notably, the decision boundaries of the temporal coherence boundary (TCB) and fission boundary (FB) first observed by Van Noorden [135] are automatically generated by the model. The results are compared to several existing implementations designed to simulate alternating-tone sequences [11, 104, 139]. Throughout this thesis, we use the three levels of analysis proposed by Marr in vision [89]. We emphasize the importance of balance at each level of analysis, and their relationship with the three approaches in the study auditory perception.

Dedicated to my Parents

and to bb

ACKNOWLDGMENTS

I want to thank my research advisor, Prof. DeLiang Wang, for his constant guidance throughout my graduate study at Ohio State. His knack for explaining the most difficult scientific concepts in a clear and concise manner is unparalleled. His extensive knowledge in so many scientific fields and his patience in helping me understand them have not only driven me to complete my research projects but also fueled my passion for pursuing future scientific endeavors in a variety of subjects.

I am also grateful to have the opportunity to work with Dr. Douglas Brungart, who has built for me the bridge between behavioral and engineering methodologies through psychophysics. I will always be awed by his amazingly sharp mind to constantly generate new and creative ideas for experiments, and his remarkable capability in sorting through and analyzing the most complicated experimental data. I gain a great deal of insights every time I speak with him, and I am indebted to his patience in always taking the time to explain a problem in detail so that I really understand each issue.

Thanks are due to Brian Simpson, who has given me much assistance while working in the Air Force Research Laboratory. He has shown me a great deal on the methods to conduct psychophysical experiments. Coming from an interesting combination of backgrounds in psychology and music, I definitely have enjoyed many interesting discussions with him.

I would also like to thank Prof. Eric Fosler-Lussier for being on my committee and providing insightful and objective critiques to my thesis. His advice and opinions have

allowed me to think of my research through different points of view, and help me to make my work relevant for more people.

I want to also thank my undergraduate research advisor, Prof. Vera Maljkovic. She not only introduced to me the wonderful field of psychology and perception, but always gives me words of encouragement and valuable advice. Her belief in my ability has given me the confidence to pursue graduate school and tackle challenging problems, and will continue to inspire me in the future.

I wish to thank my lab mates from the Perception and Neurodynamics Lab, who are always around to help and discuss interesting topics. I am grateful to Soundarajan Srinivasan, who often puts his own work aside to help others, and always provides insightful answers regardless of the subject matter of my concerns. I will certainly miss his good-natured and approachable personality plus the breadth of knowledge he possesses. I want to thank Yipeng Li, whom I also frequently approach with my academic dilemmas, and would always patiently explain to me fundamental issues that boggle my mind. Guoning Hu continues to impress me with his strong scientific knowledge, self-discpline, and a balanced lifestyle. I thank Yang Shao for his technical assistance with a variety of software and hardware that I depend on dearly. I have enjoyed my conversations on life and family with Nicoleta Roman, and also want to thank her for the help she gave me in getting started with the ideal binary mask. Zhaozhang Jin immediately impacted me in the short time that I have known him, by gladly answering my questions on signal processing fundamentals that I have always been afraid to ask. Lab alumnus Mingyang Wu has shown me the importance of maintaining social networks and good communication skills even in academic pursuits; his advice on careers will long be remembered.

I would like to thank my family and friends because they have supported me all these years and have kept life interesting, especially my parents whom I will forever look to as my role models.

Last but most, I am grateful to have met my girlfriend Jessica Yi Jin during my graduate study, because the endless hours of hard work sifting through hundreds of textbook pages, journal articles, astronomical equations, and problem sets are quickly forgotten when I remember that we have always been side by side throughout this amazing journey, and the bond we have built is in itself worth the most novel scientific discovery.

On a final note, I want to give credit to the financial support from the Air Force Research Laboratory that has made my research possible.

VITA

December 17, 1980 ……..	Born in Taipei, Taiwan
September, 1998 – June, 2002 …….	B.S., Computer Science B.A., Psychology The University of Chicago, Chicago, IL
September, 2002 – Present ……..	Graduate Teaching Associate Graduate Research Associate The Ohio State University, Columbus, OH

FIELDS OF STUDY

Major Field: Computer and Information Science

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 MOTIVATION: FROM SENSATION TO PERCEPTION

Audition, vision, and other senses are all important modalities for humans and other animals. For people with normal eyesight, hearing, and other sensory functionalities, they often take these senses for granted and fail to realize the complex processes that take place inside our brain. Consider the following situation where a child watches TV in a quiet room. While giggling loudly to the goofy cartoon characters on the TV screen, she is also made aware of the pleasing aroma of her favorite hot cocoa brewing in the kitchen's microwave. When the microwave timer beeps, she steps to the kitchen and tries to take the cup out of the microwave, only to drop the cup on the floor because of the unbearable heat exerted by the cup. Hearing the cup shattering into bits from upstairs, the child's mother rushes down to the kitchen, at which point the child breaks into tears and rushes into her mother's arms. For the participants of this situation, the boundary between sensation and perception can easily be blurred. In fact, behind this unrehearsed and seemingly natural sequence of events exists numerous perceptual processes scientists only begin to explain. Physiologists, psychologists, and many other scientists and philosophers have marveled over the complexity of perception for centuries. Recently, a new group of scientists, the computer scientists, have also been humbled by the complexity of perception, when they find all sorts of challenges implementing feasible models of machine perception. Regardless, most agree that perception occurs through the processing of the stimuli gathered by our sensory organs to form some form of mental representation. Unfortunately, the exact relationship between the sensory input and the perceptual representation is not clear. Many theories and methodologies do exist from

years of research by scientists in multiple fields, and it is through these existing methods we will begin to explore this issue.

1.2 LINKING THE STIMULUS AND PERCEPTION: BEHAVIORAL, PHYSIOLOGICAL, AND COMPUTATIONAL APPROACHES

To decipher the complexity of the perceptual process, perception researchers have been approaching the problem by studying various relationships and links between the events in the overall process. The approaches can be broadly summarized into two categories. The behavioral approach links the stimulus directly with the resulting perception, while the physiological approach links the stimulus with neural responses [61].

With the behavioral approach, scientists attempt to explain the people's perceptual responses based on some physical properties of the stimulus. The best way to measure such relationships is often through conducting experiments and asking people the relevant questions. A common method, the phenomenological method, is often used to directly ask the participants to describe what they perceive. This is a very useful method to understand qualitatively the participants' perceptual responses, but cannot establish a rigorous quantitative relationship between the stimulus and perception. Psychophysical methods belong to another type of behavior approach that can be more rigorously quantified and often require scrutiny in the measurement of responses. First developed in the 19th century, there now exists numerous methods to measure accurately the participants' responses to the stimulus. In particular, psychophysical methods exist that measure thresholds, the smallest detectable units the participants can detect, recognize, or identify for a particular stimulus. Along the same line, psychophysical experiments can also be used to measure the just noticeable difference (JND), which is the threshold of difference between two stimuli that a person can detect. Furthermore, the techniques can be used to conduct magnitude estimation, which may lead to proposals of equations relating perceived magnitude to stimulus intensity. The most notable of relationships derived from this technique is Steven's power law, $P=kS^n$, where perceived magnitude, P, is equal to a constant, k, times the stimulus intensity, S, raised to a power, n [125]. This

is the law that explains how all our senses follow the basic relationship, such as our perception of brightness or loudness.

The physiological approach has focused on the relationship between the stimulus and the nerve impulses, and how these nerve impulses lead to perception. Structures called receptors exist to receive stimuli from the environment and transduce them into electrical signals, which are then transmitted along neurons to different areas of the brain. The electrical signal propagating along a neuron is caused by rapid increase in positive charge inside the cell membrane, and this is called the action potential. Because the action potential is an all-or-none response within each neuron, once it is triggered, it will stay about the same size and propagate all the way down the axon. Because of this property, it is found that the rate of nerve firing directly relates to information about the stimulus, such as the intensity of the stimulus. The way these nerve impulses lead to perception is to help communicate such information to other neurons. We will not go into detail here about the chemical and electrical events that take place to allow the information carried by the action potential to generate a signal in another neuron; the reader may find such information in a neurobiology textbook. However, it is important to note that the action potentials do not travel from one neuron (called the presynaptic neuron) across to another neuron (called the postsynaptic neuron), but instead generate a chemical or electrical process that, in turn, triggers a voltage change in the postsynaptic neuron. Furthermore, the transmitters responsible for this communication process may cause two effects. They may cause an excitation or inhibition of the postsynaptic neuron, which will increase or decrease the rate of nerve firing, respectively. As a neuron often receives many excitatory or inhibitory inputs, such a vast number of interconnections in the brain can lead to the final percept. Finally, we must note that as an effort to understand these processes, scientists have devised methods to measure properties of the nerve impulses, rate of nerve firing, and many other physiological phenomena. The methods generally fall into invasive or noninvasive methods. Invasive methods require inserting instruments into the brain, such as the commonly used technique of implanting electrodes to measure a neuron's electrical activity. Noninvasive methods do not require introducing an instrument into the brain, and include several imaging techniques like X-

rays, magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET). These methods are useful to identify brain structures and functional areas of the brain.

From the brief discussion above, we note several properties from both the behavioral approach and the physiological approach that seem to relate well to a computational approach of perception. First of all, it is commonly accepted that the brain performs computation to achieve perception, thus it is reasonable to utilize the computer to simulate the perceptual processes based on what we know about the mechanisms. While we do not argue in this thesis whether the computer can duplicate human perception, we do assume that building computer models is useful in advancing the understanding of perception. That said, many results from psychophysical experiments naturally become the basis of computational models because these experimental results often give tangible data about the perceptual process, and any feasible model should conform to these results in order to be plausible. For example, consider the following simplified scenario where experimental results show that speaking at a far distance from the listener would reduce the listener's comprehension compared to if the speaker is very close to the listener. The corresponding computer model must have the mechanism to simulate such an effect or its value would likely be limited. On the other hand, the computations should also adhere as best as possible to the underlying physiological processes to maintain biological plausibility. Continuing the example just given above, one can build a computer model simply by programming a function that takes as input the speaker's distance from the listener, and outputs some metric of the listener's comprehension by reading from a table of experimental data. While the model would be accurate given those conditions, it does not reveal anything insightful of how this effect is actually processed in the brain. Furthermore, it loses the ability to generalize when placed inside any other environments. This approach is consistent with David Marr's framework of complex information processing [86]. A most influential scholar of computational vision, Marr [86] defined three levels of analysis of an informational processing system. The first, and the most abstract, level is the *computational theory*. It describes *what* the functions of the system are and also formulates the constraints that tell us *why* the system is behaving in certain

ways. The second level involves choosing a *representation* for the data and the *algorithm* that describes *how* the proper functions are accomplished. The third, and the most concrete, level is the *implementation* of the system physically. Based on our discussion, the behavior approach can tell us a lot about the computational theory level of perception, as it directly links the stimulus with the resulting percept. Then further examination of the physiological processes in the brain can help us select the best representation and algorithm involved. Fortunately, there exist many similarities between the structure of our nervous system with modern computing architectures that allow the representation of neurocomputations by computing concepts quite naturally. Perhaps most notable is the similarity of the all-or-none property of the action potential with the binary property found in most digital computer designs. In fact, even before modern computers existed, McCulloch and Pitts were inspired by the all-or-none feature of a real neuron to propose the first artificial neural network designed to compute like the real brain [89]. To be sure, there exist many stark contrasts as well between the brain and the computer [113]. Therefore this thesis is not advocating that computers will explain everything about perception, but simply that many computing concepts seem to represent and compute perceptual processes well. The goals of the computational study of perception should be attempting to better understand how perception works as well as to build artifacts such as "mobile robots" that can artificially perceive events in the environment. The latter is an example of Marr's implementation level of analysis, and is an immediately useful application to such research efforts. The methods of such computational studies should maintain consistency with both the behavioral approach and the physiological approach in the study of perception.

Hopefully, our above discussion about the relevance of computational studies in perception has also motivated the reader to study perception with a balance between behavior and physiology. We have seen with the example given above about how speaker distance can influence listening comprehension, and that building a computational model of this process by utilizing both approaches together gives us a more complete picture of the perceptual effect. Looking from another angle, studying a phenomenon behaviorally can lead to insights about the underlying physiological

mechanisms. Continuing with the same example, the perceptual behavior of losing comprehension from a speaker far away tells us that stimulus intensity is important in perceiving speech. While the result is obvious, it demonstrates the important interrelations of the behavioral and physiological approaches.

1.3 A CLOSER LOOK AT AUDITION

The main focus of this thesis is on audition. As audition is a specific modality of perception, all our discussion about approaches to study perception so far is relevant under the assumption that properties of audition are directly inheritable from human perception in general. In perceptual research, vision dominates the literature. Computer vision has also led the way in the computational modeling of the sensory systems. However, one must not take the importance of hearing lightly. Audition is especially important in facilitating human communication, through the development of speech and language. Helen Keller, who was both deaf and blind, had expressed that being deaf was worse than being blind, because blindness isolates one from inanimate things, while deafness isolates one from the vital signs of people [84]. Recently, the increasing usage of automatic speech recognition tools in commercial products has brought computational audition under the spotlight. However, fundamental problems still exist at both the level of the computational theory and the level of representation. One particular problem is the difficulty to bridge the gap between the stimulus and the representation. A listener in a realistic environment is always presented with stimulus from different sound sources. Yet humans with normal hearing have no problems in organizing the bits of input together into the right sources to perceive a meaningful sound. This process is part of a general phenomenon called *auditory scene analysis (ASA)* [18], which pertains to the perceptual grouping of auditory stimuli in complex auditory environments into coherent streams. Because neither the behavioral approach of understanding the relationship between sound stimuli and what humans can hear nor the physiological approach of understanding the biological mechanisms of hearing is clearly resolved, utilizing computational models to link these mechanisms together may prove to be the best approach.

Auditory scene analysis is a truly remarkable accomplishment by the auditory system. One of the first experiments that have demonstrated the robustness of our auditory system is Cherry's study on the listeners' ability to separate concurrent mixtures of speech recorded from the same speaker [35]. The author termed this phenomenon the *cocktail party problem* and since then numerous researchers have worked to understand this type of problems and auditory perception in general [18, 143, 150]. Furthermore, there have also been numerous computational models of audition developed as ways to represent the known computational theories of auditory perception [8, 23, 71, 135, 139, 149]. We will continue to follow the same framework of study in this thesis, by incorporating behavioral studies to understand the underlying theory, and complementing with computational modeling as possible representational schemes.

1.4 THESIS OVERVIEW

During the introduction of this thesis, we have outlined the different approaches to study perception, emphasizing the combination of behavioral and physiological approaches coupled with computational frameworks. As we focus on auditory perception, Chapter 2 goes in details about the principles of auditory scene analysis as proposed by Bregman (1990). The insights we gain from these principles will help us understand the basic computational theories and goals. Chapter 3 delves deeper into the behavioral approach, by studying auditory perception in multi-talker speech environments and describes some psychophysical experiments of multi-talker intelligibility. These experiments allow us to better understand the complex theory of auditory perception as well as offer insights on the relationship between computational goals and representation of auditory perception. Chapter 4 gives an introduction to the physiological process of audition and explores existing biological plausible computational models, with particular focus on Wang's LEGION architecture [127, 135, 140]. Then we conclude the computational approach by extending LEGION to implement a particularly interesting behavioral phenomenon of ASA, called the segregation of alternating-tone sequences. Chapter 5 summarizes the

insights gained from this thesis, discusses the state of the study of auditory perception in terms of the different approaches, and outlines future research.

CHAPTER 2
AUDITORY SCENE ANALYSIS

2.1 INTRODUCTION

We have discussed in Chapter 1 that behavioral studies using psychophysical experiments reveal the direct relationship between the stimulus and percept, which would fittingly become the basis for the computational theory level according to Marr's framework [86]. Since Cherry's experiments in 1953 introduced the *cocktail party effect*, numerous auditory experiments involving stimuli ranging from simple tones and noise bursts to complete speech mixtures have been investigated. Many important conclusions about auditory perception have been drawn from these efforts. Bregman's proposed theory of auditory scene analysis (ASA) covers a wide range of related issues, and he summarized many important results in his book [18]. We look at some main results and discuss how they can fit into our analysis.

ASA is concerned with perceptual questions such as the number, characteristics, and locations of the sound sources. The auditory system can solve these questions by breaking down the complex sound mixture received from the environment into smaller *segments* and grouping the segments into *streams*. The grouping mechanism determines which segments belong to the same sound source and thus each stream formed is a complete perceptual representation of a sound source. Bregman noted in the book that "the stream plays the same role in auditory mental experience as the object does in visual." [18] He also noted the difference between the word "stream" from the word "sound" or "acoustic event", as the "sound" refers to the physical signal while the "stream" is the perceptual result interpreted mentally. This process, which is the main

process responsible for auditory scene analysis, has been called *auditory stream segregation*.

Many principles from ASA can be traced back to the studies in Gestalt psychology [78, 110]. Gestalt principles of grouping explained the ways visual inputs connect to each other in the mental percept, despite the fact that the input stimuli triggered discrete receptors in the visual system. The main characteristics for grouping are summarized as follows:

- Proximity: nearby elements tend to be grouped together
- Similarity: elements similar in attributes tend to be grouped together
- Good Continuation: continuous elements resulting in straight or smoothly curving form tend to be grouped together
- Connectedness: dots, lines, areas, and other elements that are physically connected tend to be perceived as a unit
- Closure: elements that form an enclosed object even with some gaps tend to be grouped
- Common Fate: elements moving in the same direction over time tend to be grouped

Gestalt psychology was formed in the early 20th century. Before that, most psychologists believed in the *structuralistic* approach, which states that perception of the whole is made up entirely of the sum of its parts. Challenging structuralism, Gestalt psychologists believed that perception was much more than just summing up individual elements, and thus formulated several principles of perceptual organization described above to refute structuralism. While these principles may seem obvious to us now because we are so used to perceiving these regularities in our daily environments, they are well formulated phenomena that can be good guidelines for more complex analysis. In fact, we will soon see that many of these principles apply to the auditory domain as well.

2.2 PRIMITIVE AUDITORY SCENE ANALYSIS

Bregman [18] described two mechanisms involved in ASA, primitive and schema-based, that are responsible for auditory stream segregation. The majority of the work in the book is on the primitive mechanisms. He believed that primitive segregation and grouping are innate and pertain more closely with the Gestalt principles described above. These mechanisms are bottom-up and involve the breaking down of sound signal into many elements for analyses. The grouping is done in two dimensions across time and frequency (although Bregman noted that more than two dimensions may be possible). Grouping across time involves *sequential integration* and grouping across frequency involves *simultaneous integration*. Grouping is done based on different *cues* produced by the analyses of the elements. The following is a summary of some important cues in the auditory domain. Many of them are similar to or are a subset of the Gestalt principles.

- Frequency/pitch proximity: When listening to two tones of different pitches, the tones tend to be grouped into the same stream when they are close together in frequency (as well as if they are presented close together in time)[19]. This is known to be a dominant cue in the auditory domain, and is directly related to the Gestalt principles of proximity and similarity. One of the earliest experiments related to auditory streaming was performed by Miller and Heise [93]. They identified the "trill threshold" where the listeners listening to a rapidly alternating pair of tones form two streams perceptually. It is when the difference between the two tones is at 2-3 semitones apart, with an alternation rate of 100ms per tone. Van Noorden [131] conducted further tests on streaming with alternating tones. Because the study of streaming with alternating tones is a fundamental percept in auditory streaming, and because the frequency/pitch cue for perceptual organization is very important in auditory perception, we will study the results in detail in Chapter 4.

- Presentation rate: Faster presentation rates allow the tone sequences to be grouped into the same stream. Because in the time-frequency representation, faster presentation rate means shorter interval between tones, this cue is consistent

with the principle of proximity. Presentation rate has also been found to alter the effect of the frequency proximity cue we have just discussed. In Miller and Heise's [93] study, they also found that the "trill threshold" of frequency difference that determines whether the listener can segregate the alternating tones into two different streams also depends on the presentation rate. Again, we will look at this effect in more details in Chapter 4, when we specifically study the modeling of alternating tones.

- Similarity of timbre: Sounds of the same timbre tend to be grouped to the same source. This is directly related to the principle of similarity. A common occurrence of this cue is found in musical performances, where different musicians playing the same type of instrument tend to be grouped.

- Spatial location: In a realistic listening environment, sound sources usually are located at different locations in the environment space. The auditory system tends to group those sound elements coming from the same location. Bregman (1990) pointed out that spatial cues have been extensively used in early engineering work on automatic speech segregation. While the spatial cues seem to be a dominant cue as well, it is clear that they are not necessary for ASA, as Cherry (1953) and many others have shown that we can segregate sounds presented monaurally as well. This cue is also directly related to the Gestalt principle of proximity.

- Spatial continuity: Sound sources usually originate from people's voices or sounds emitted by certain objects. These sources tend to move in space contiguously and not too rapidly. The spatial continuity of sound elements from the same source can strengthen grouping. This cue is related to the principle of good continuation.

- Sound continuity and smooth transition: Similar to the previous cue, sounds can stay continuous in other ways besides in space. Continuity in fundamental frequency, time, spectral shape, and intensity can all help the elements to group

together [4, 17]. Any sudden changes in these attributes should set off an alarm that they are from different sources.

- Onset/Offset: If two sound elements have the same onset or offset time, they are more likely to be grouped as one sound stream [20]. This cue is related to several of the Gestalt principles. Proximity and similarity should play a role since the elements share similarity in the time domain. This may also relate to common fate, since the elements have the same temporal patterns.

- Loudness differences: Significant differences in loudness of sound elements may help the sounds to be segregated [130].

- Common amplitude and frequency modulation: Simultaneous tones that undergo the same kind of amplitude or frequency modulation at the same time tend to be grouped into the same stream. Furthermore, notice that a complex tone is really composed of several simultaneous harmonics, but they are grouped into one sound percept. When the pitch of the complex tone rises, all the harmonics rise by the same proportion as well, and thus remain grouped as one sound. Harmonic grouping is well supported by experimental research, where listeners can decide which harmonics belong to each sound even if the harmonics are interleaved [5, 6, 21]. This cue is directly related to the Gestalt principle of common fate.

- Cumulative effect: Whether a sequence of sound is segregated into separate streams or remains one stream is often a cumulative effect. Bregman [16] found that the effects of segregation of sounds can be influenced by sounds heard a few seconds preceding those sounds. Furthermore, a few seconds are also necessary after a period of silence before stream segregation occurs. Bregman [18] thus also believes that coherence is a property that has to be "lost" by the auditory system. In other words the system starts with the perceptual state of one single stream and gradually becomes segregated after auditory streaming takes place.

- Collaboration and competition: This effect is not really addressed by the Gestalt principles. Based on all of the cues above, some will dominate over others depending on the stimuli, and some of the cues will strengthen grouping or segregation with other cues present. This effect may be quite useful in building a computational system that takes account of these cues for processing.

Among all of these cues, we will delve into the pitch/frequency cue and presentation rate further in Chapter 4, when computational models of ASA will be explored. We will then also be able to look at an example of how collaboration and competition of grouping cues will affect the final grouping.

2.3 SCHEMA-BASED INTEGRATION

Bregman (1990) described the other mechanism involved in ASA as schama-based. The schema-based integration involves the listener to utilize attention to "listen for" a sound, and requires using prior learned knowledge or familiarity about the sounds to aid integration [4, 147]. Primitive processes *partition* the inputs according to evidence the system receives while the schema-based processes *select* from the evidence directly. Thus the mechanism is top-down. In fact, schemas can be very flexible in exactly what is encoded; some can represent sounds ranging from phonemes to words and melodies [46, 64]. In phonemic restoration, when the experimenter replaces a phoneme in a speech utterance with some noise burst, the listeners can still perceive the deleted phoneme based on the context of the sentence [124, 142]. This example of schema-based integration shows that the physical signal does not even need to be present to be integrated into the perceptual stream! Many other experiments show that schema-based integration coexists with primitive scene analysis. With schema-based integration, prior knowledge and familiarity with certain sound sequences can aid the segregation of known sound from the other sound sources, even if the sounds are in overlapping frequency range [46]. Attention is also believed to be very important for schema-based integration, even with very simple sound sequences like alternating tones. Van Noorden [131] has

found that at a certain range of inter-onset-interval (IOI) and frequency difference, the listener can actually utilize their selective attention to hear the sequence as one stream or segregate into two streams.

2.4 SPEECH SCENE ANALYSIS

While Bregman (1990) did not classify speech scene analysis as a particular mechanism in ASA, we find it very important to explore the additional issues involved with ASA specific to speech scenes. We assume that speech scene analysis consists of both primitive processes and schema-based integration. However, due to the complexity of speech signals, it is difficult to determine the boundary between primitive processes from schemas in the perception of speech. In fact, Bregman suggests that contributions of the primitive processes are obscured by the contribution of schemas in speech perception. Furthermore, there seem to be some special cues that help group speech from one source speaker together. In terms of primitive scene analysis, speech is often made up of succession of very different sounds, such as low frequency tones for vowels and high frequency noise for fricatives, thus it seems that there need to be schemas in order to group these very different kinds of sounds together. However, grouping of speech sounds is also not solely due to schemas, as Bregman's experiments show that one can still discover coherent words exist by playing words backwards and thus destroying the schema. On the other hand, one cannot discover coherent sounds as words when the words used are artificial and mixed with some mechanical sounds. So he concluded there was something inherently different between a real speech sequence and a sequence of artificial sounds. The following is a summary of some more sophisticated cues directly relevant in grouping of speech sequences:

- Pitch trajectory: In general, the pitch of a human changes over time, but changes slowly. The pitch also follows certain melodies consistent with the language being spoken. Listeners can to be forced to follow the pitch contour due to its control of the listeners' attention [41, 128].
- Spectral continuity: The spectral formants in a spectrogram in successive sounds tend to be continuous with one another, because the formants are

caused by the filtering from the vocal tract which does not move instantly from one position to another. This helps to provide spectral continuity in a speech sequence.

- Common fate cues: Besides utilizing continuities, a class of cues used in speech integration also involves the principles of common fate through correlated changes in different parts of the spectrum. In terms of frequency, for example, speech harmonics move in parallel on a log-frequency scale as the pitch of the speech sequence changes. The parallel movements would indicate that the harmonics belong to the same sound source. The changing harmonics can also be used to help the listener to "trace out the spectral envelope" that may benefit integration [4].

2.5 COMPUTATIONAL AUDITORY SCENE ANALYSIS

A comprehensive computational model of audition, for our purpose, is a system that can account for all three levels of analysis in Marr's framework, such that the system is functionally and biologically consistent with the current knowledge of auditory perception. As we have mentioned before, much of the experimental work involving ASA that reveals all of the grouping cues discussed above, helps us narrow down the computational goals of ASA. Each cue points to certain function or constraint that we should take into account to achieve the holy grail of ASA. Some of these cues may also indicate the kind of representation or algorithm that should be used, but only to a limited extent, because the majority of these cues directly link the stimuli with the final percept. Therefore, the next computational challenge after studying ASA would be to determine the best representation and algorithm to reach the goals. Before we discuss that issue further, I should discuss that even at the first level of analysis, many differ in what the computational goals of ASA should be and in turn what the best representation should follow.

Traditionally, computational solutions to tasks related with ASA have been motivated by two types of researchers. On one side, practical application such as improvements to

automatic speech recognition (ASR) systems, audio information retrieval, and "intelligent" hearing devices motivates the design of a system to accomplish ASA tasks. The main tasks under interest for these applications, even before the term ASA was formulated by Bregman in 1990, were related to source segregation, one of the major issues in ASA. Here, the primary goal is to separate the target sound from the rest of the interfering sounds in a mixture of sound signals. This effort has focused on engineering techniques which does not utilize many insights in experimental and physiological studies about the auditory system. The three main approaches were speech enhancement [81, 102], spatial filtering with a microphone array [79, 132], and blind source separation using independent component analysis (ICA) [73, 80]. These approaches have significant limitations such as dealing with unpredictable nature of a variety of nonstationary interfering signals, tracking of a moving target source, and unrealistic assumptions for ICA [138]. As the field advanced, these attempts have led to the creation of models that draw insights from psychophysical principles [23, 52, 90, 139, 145]. The field of study termed computational auditory scene analysis (CASA) was then formulated. More recently, there have been several computational models that have considered many of the organizational cues described in the previous sections. Hu and Wang [71] have utilized not only the pitch/frequency cue but also the amplitude modulation cue coupled with optimization techniques. Roman et al. [111] extensively utilized the spatial location cues of interaural time difference (ITD) and interaural intensity difference (IID), and formulated the computation as a binary Bayesian classification problem to determine where the target sound is stronger than the interfering sound. Srinivasan and Wang [124] have actually utilized a schema-based approach to process speech sound streams, simulating the phonemic restoration phenomenon observed in speech perception [142]. Furthermore, more evidence has shown that attention plays a significant role in ASA, and Wrigley and Brown [147] have taken that into account by building a computational model that includes attentional leaky integrator (ALI) and a representation of attention allocation across frequency.

On the other side, computational researchers in collaboration with neuroscientists and psychologists have been interested with producing biological plausible models of ASA to

not only improve the performance CASA systems but also gain insights into the mechanisms of the auditory system [88, 101, 135, 147]. The evaluation criteria for these models need to account for known psychophysical or physiological data. While these systems have been built in terms of the interaction of neurons in the brain, each of the models is limited in the breadth of auditory phenomena that can be handled. The difficulty lies in that the underlying mechanisms of the human auditory system are complicated and many regions in the brain that are involved are not well understood [42]. Therefore, building a single computational model that can simulate the entire auditory system is unlikely in the near future. Norris [101] proposed that it can be feasible to build computational models that simulate a limited range of phenomena, but flexible enough to be integrated into a master system as part of a module. That is the same view we hold in this thesis, and we will explore this issue in Chapter 4.

As we have discussed, the computational goals for CASA can be quite different depending on the research interests. However, there does exist much overlap and most researchers' ultimate goals are more or less the same: to build models that not only faithfully simulate the human auditory system, but can produce good and robust ASA performance with natural speech sounds in real-life environments. After all, humans can perform many ASA tasks with ease and thus an accurate computational model should be able to do everything humans can. Let us now return to the question we have left off regarding Marr's three levels of analysis for the problem of ASA. We have outlined in this chapter several important experimental results regarding auditory perception, and have seen the formation of perceptual theories based on the results. Even with these computational theories in place, challenges still exist linking the behavioral results with the underlying physiological mechanisms that will be critical for the representational and algorithmic level of analysis. Depending on the computational goals, considerations for the representation and computation methods may be very different. Because we are focusing on studying biologically plausible computational models, we will try to study and design representations that are consistent with neurobiological studies relating to the auditory system. In particular, many recent neural network solutions for CASA have been based on neurobiological studies that focus on emergent properties of

interconnected neural oscillators utilizing synchronization to achieve ASA [24, 101, 135, 139, 147]. We will study these and related models and utilize similar representations in chapter 4.

2.6 SUMMARY

We have seen in this chapter that auditory scene analysis is a fundamental issue in auditory perception. Being able to take a mixture of incoming sounds from different sources and analyze and identify details about each sound source is not a trivial task, yet anyone with normal hearing can accomplish this without any trouble. Experiments have been done extensively in the past fifty years to identify the capability and limitation of the auditory system, and theories have been proposed regarding our ASA capability. From these, we have outlined several cues that the auditory system uses to help the segregation and grouping of different sound elements into streams. These cues, mostly based on primitive mechanisms, are traced back to the Gestalt organizational principles. Besides primitive mechanisms, the auditory system also utilizes schema-based mechanisms by integrating prior knowledge for analyses. The problem becomes more interesting when dealing with a mixture of speech signals. While many of the primitive scene analysis cues apply to speech mixtures, the primitive processes tend to be obscured by top-down schemas [4, 18]. Thus it becomes difficult to study perception of speech mixtures because the effects of bottom-up cues and top-down schemas cannot be separately analyzed easily. Furthermore, speech signal is naturally subject to masking by competing signals in a mixture. Traditional "energetic" masking occurs when both the target and the interfering utterances contain energy in the same critical bands at the same time, where the target becomes inaudible [25]. Another known type of masking, "informational" masking, occurs when the signal and masker are audible but the listener cannot correctly separate the target signal from the interfering signal [44, 76, 144]. Informational masking has been likened to the listener's ability to segregate perceptual similar sounds, and is sensitive to differences in the perceived location of target from that of the interferers [25]. This shows direct relationship with the organizational principles for ASA we have been discussing in this chapter. However, because it has proven very

difficult to isolate the informational and energetic portions of speech-on-speech masking, no previous studies have directly examined the contribution of each kind of masking effects. Without such mechanisms, we cannot fully understand the underlying process of ASA when dealing with speech mixtures. Because speech mixtures are the most natural stimuli humans utilize to communicate, we find it extremely important to further explore these and related issues, which will be the focus of the next chapter.

CHAPTER 3
ON THE IDEAL BINARY MASK AND ITS EFFECTS ON MASKING IN MULTITALKER SPEECH MIXTURES

3.1 INTRODUCTION

3.1.1 ENERGETIC AND INFORMATIONAL MASKING

The discussion about auditory scene analysis (ASA) in the previous chapter has given us a good idea of the fundamental theories relating auditory stimuli to the final percept. The organizational cues proposed by Bregman [18] regarding primitive scene analysis based upon the Gestalt organizational principles are certainly useful in formulating the computational theories and goals of ASA, the first level of analysis in our quest to understand the underlying mechanisms of ASA and auditory perception using Marr's framework. However, the organizational principles alone quickly become inadequate when presented in a naturally complex environment involving multiple speech sounds, such as the "cocktail party" type environment. Under complex environments, a variety of effects can complicate our understanding of the primitive organizational principles, including masking effects and schema-based mechanisms. Thus it is necessary to take our behavioral approach in depth and explore these issues further in order to have a more complete understanding of the computational theories involved in ASA. In particular, we will focus on some issues central to the complication manifested through the masking effects in the perception of two or more simultaneous talkers. To isolate the problem, we have limited ourselves in dealing with only the monaural case, which means that the same sound will be presented to both ears, and location or spatial based cues are not an issue. We also limit ourselves to the discussion of simultaneous target and interference.

However, we hope that the insights gained from our discussion and study can be extended to other complicated situations (e.g. sequential masking) as well when utilizing the behavioral approach.

As discussed previously, when a target speech signal is obscured by one or more simultaneous competing talkers, at least two types of masking interfere with the listener's ability to comprehend the target speech. The first type of masking is classical "energetic" masking, which occurs when target speech and masking speech overlap in time and frequency in such a way that portions of the target speech signal are rendered acoustically undetectable at the periphery. This type of masking depends only on the spectral-temporal content of the interfering sound, and it occurs for all types of speech and non-speech masking signals. The second type of masking, often referred to as "informational masking," is higher level and occurs when the listener is unable to segregate the acoustically detectable portions of the target speech signal from the similar-sounding acoustically detectable portions of the interfering speech [25, 32, 58, 77, 107]. This type of masking occurs most predominantly with interfering speech signals, but may also occur with other types of speech-like maskers, including time-reversed speech signals and multitalker babble signals [26]. The effects of this type of masking have also been termed non-energetic masking, and are directly related to the organizational principles of ASA and may also be influenced by attention and schema-based mechanisms as well.

With the coexistence of energetic and informational masking in most speech mixtures of more than one talker, it is very important to be able to measure the relative contribution of both types of masking. The reason is that both the physiology and the effects of these two masking phenomena are very different, as energetic masking occurs at the periphery, while informational masking is believed to be centrally processed. Indeed, different areas of the brain are activated in speech-on-speech and speech-on-noise masking [115]. Therefore, experimentation involving multi-talker speech mixtures will give misleading and confusing results if the two effects are not controlled properly. However, it has been difficult to directly measure the relative effects of each type of masking in multitalker listening experiments. There have been several studies that utilize the characteristic

effects of informational masking and make the effects more prominent during experiments by stimulus manipulations. The major effects of informational masking come from similarity between the target and interfering speech signals, and thus the effects are greatly relieved when the target and interfering signals are very distinguishable. For example, intelligibility performance greatly improves when the target and interfering talkers are of opposite sex instead of the same sex [28]. Target and interference located at different spatial locations are also useful in releasing from the masking caused by informational masking, and lead to a greater release from masking with speech maskers than with noise maskers [22, 25, 58, 65, 100, 105]. Very interestingly, lowering the target voice below the level of the masking voice can actually improve intelligibility for the target signal, as the listener can use selective attention to attend to the "quiet" talker in the mixture [25, 43, 50]. As a whole, these results are all consistent with ASA principles, which due to the dissimilarity between the target and interfering speech, the auditory system can successfully segregate the mixture into two separate streams and thus identify the target stream for the recognition task.

3.1.2 ISOLATING THE INFORMATIONAL COMPONENT OF SPEECH-ON-SPEECH MASKING

The examples just described show how informational effects can become more prominent in speech mixtures, but do not actually isolate either kinds of masking. Isolation of the masking effects has proven to be quite difficult in speech mixtures, but recently there have been some research studies aimed to accomplish this, especially in the isolation of informational masking without effects from energetic masking. For example, Brungart and Simpson [26], have shown that listeners who are attempting to listen to one of two simultaneous talkers presented in one ear are susceptible to informational masking from a third independent talker presented in the opposite ear. Because little or no energetic masking occurs when speech-shaped noise is presented in the ear opposite the target speech, any across-ear masking effect can be attributed almost entirely to the central speech segregation processes that are generally associated with informational masking.

Another interesting technique to isolate informational masking is to divide a speech signal into 15 logarithmically-spaced envelope modulated sinewaves, and then randomly assign 8 of the bands as the target signal and 6 other bands as the masking signal [2]. This technique gives two intelligible speech signals without any spectral overlap, effectively removing energetic masking. What Arbogast et al. [2] found was that informational masking effects can still occur with these stimuli, just like other normal multitalker stimuli. In addition to this isolation technique, they also demonstrated an application for isolating information masking by testing the effects of spatial separation on informational masking. Indeed, they found that spatially separating the stimuli helps release informational masking effects much greater than it helps to release energetic masking from stimuli that contain noise in the same bands as the target speech (approximately a 20 dB release in informational masking compared to only a 5 dB release in energetic masking).

This difference between the spatial unmasking that occurs with speech and noise maskers is consistent with, but much larger in magnitude than, the differences found in other studies that have compared spatial unmasking for interfering speech signals and for interfering noise signals [22, 25, 58, 65, 100, 105]. This discrepancy seems to reflect the fact that the energetic masking component that occurs in natural interfering speech limits the size of the advantage that can be gained by spatially separating a target speech signal from a normal interfering speech masker. While this is an interesting result in itself, this also shows us the importance of the isolation of different kinds of masking effects.

3.1.3 ISOLATING THE ENERGETIC COMPONENT OF SPEECH-ON-SPEECH MASKING

Besides isolating informational masking, one can also try to isolate energetic component in speech masking without significant contributions from informational masking. A simple approach to this problem is to create a continuous noise signal that matches the long-term average spectrum of speech. This is called a “speech-spectrum-shaped” noise masker and studies have shown they do not represent speech signals accurately. For

example, at the same signal-to-noise ratio (SNR), when target and interfering speech signals are very similar, a large amount of informational masking exists and the interfering speech signal reduces performance more than the equivalent speech-shaped noise masker [25, 50]. However, when the target and interfering speech are very different, such as voices of the opposite sex at different locations, and informational masking effects are less emphasized, then in general the interfering speech degrades performance less than the equivalent speech shaped noise masker [48, 55, 65, 72, 105]. However, note that this is true for normal hearing listeners only. Hearing impaired listeners generally do not do better with a speech masker than a noise masker and in general seem to be less able to take advantage of the dips that occur in a fluctuating masker than normal listeners.

The above technique seems to overestimate the amount of energetic masking present in a speech signal. Indeed, speech has been known to be a inefficient energetic masker because there exists spectral fluctuations in natural speech and they allow listeners to hear "glimpses" of the target speech even when the SNR is relatively low [7, 36, 40, 92]. Therefore, an extension to the previous technique to characterize only the energetic effects of a speech masker can be obtained by amplitude modulating a speech-spectrum-shaped noise with the overall envelope of a natural speech masker signal. As expected, such maskers produce significantly less masking than continuous speech-shaped noise without the modulation [22, 25], but still produce more masking than an equivalent natural speech masker [7, 36, 40, 92]. Because speech envelope fluctuates differently at different frequency regions, there are dips at different regions of the interfering signal that the listener can catch more "glimpses" of the target. The natural next step is to divide the speech-shaped noise signal into two bands and use different envelopes to amplitude modulate the two bands, which is what Festen and Plomp [55] did. While their results do not show significant difference from before, there is plenty of evidence that listeners can extract information by obtaining "glimpses" from the target varying in both time and frequency [30, 69].

We can continue to extend the technique for isolating energetic masking by dividing the speech-spectrum-shaped noise into more bands and modulate them with envelopes derived from different frequency regions of the target speech. However, the more bands there are, the more "speech-like" the speech-shaped noise would be, which may cause informational masking to be introduced again in the sound mixture. In that case it seems that we are going in circles and not accomplishing what we set out to do, which is to eliminate informational masking. Indeed, several researchers have argued that some amount of informational masking may occur with any kind of noise masker because even just random fluctuations in the noise waveform can generate similar acoustic elements as that in natural speech [47]. If the speech-shaped noise masker is modulated to match that of natural speech, the modulated masker can indeed sound like speech and be partially intelligible even with only a few independently modulated bands present [45]. Brungart et al. [27] showed that these modulated speech-shaped noise signals can generate across-ear informational masking in dichotic cocktail-party listening tasks. Based on these results, it appears that the modulated speech-shaped noise approach cannot isolate the energetic component of speech-on-speech masking because it forces the listener to segregate a target voice from an intelligible masking voice which would likely introduce informational masking to the task. It seems that in order to avoid introducing informational masking effects into the stimulus, we should somehow take away the segregation task in the stimulus from the listener, so that the listener would not be confused of the identity of the stimulus. That is exactly the approach we take in this chapter.

In this chapter, we propose an approach that can retain the energetic masking effects produced by a natural speech interferer without introducing informational masking effects into the stimulus. This approach, based on the "ideal binary mask" procedure that has been proposed and used as a performance standard in computational auditory scene analysis (CASA) [138], uses information about the time-frequency composition of the target and interfering signals to eliminate just those portions of the stimulus that are dominated by the interfering speech. This allows speech intelligibility to be evaluated in a stimulus that contains only those portions of the target speech that should be

acoustically detectable in the presence of the interferer, but none of the components of the masking speech that might otherwise be confused with the target speech and thus generate informational masking. The resulting signal will be highly distorted compared to the original, but evidence has shown that listeners can adapt to speech stimuli with temporal gaps and spectral holes [75, 83, 92]. Therefore, the expected intelligibility performance should be higher than with using the original speech mixtures. Any decrease in performance due to the distortion of the target signal can be attributed to the loss of energetic portions of the target speech eliminated from the stimulus in regions where the interfering signal dominates. Thus, the ideal time-frequency binary mask approach would provide an upper bound on the portion of speech-on-speech masking attributable to energetic masking effects.

3.2. THE IDEAL BINARY MASK

3.2.1. BACKGROUND

As we have discussed in chapter 2.5, for a typical problem in computational auditory scene analysis (CASA), one is usually given only the information available in the mixed audio signal, with the ultimate goal of returning the perfectly reconstructed target signal. However, keeping the limitation caused by energetic masking in mind, it is doubtful that our auditory system retains such accurate information while sensing the input stimuli. The objective then becomes retaining regions of the target sound that are stronger than the interference. In concrete terms, "ideal" binary masking is a signal processing approach that segregates audio mixtures by retaining those time-frequency regions of a combined signal that contain useful information about the "target" sound source and eliminating those time-frequency regions that are dominated by "interfering" sound sources. This processing is typically based on a two-dimensional time-frequency audio representation where the time dimension consists of a sequence of time frames and the frequency dimension consists of a bank of auditory filters (e.g. gammatone filters). Thus the basic element in the binary mask paradigm is a time-frequency (T-F) unit corresponding to a specific filter at a

particular time frame, and the binary mask itself is a two-dimensional matrix where each cell corresponds to a single T-F unit within the mixture audio signal. In the ideal case, those time-frequency units where the mixture signal is dominated by the target are assigned a "1" in the binary mask, and those where the mixture signal is dominated by an interfering sound are assigned a "0" in the binary mask. This binary mask can be used to reconstruct a subset of the combined signal (corresponding to the "1's" in the mask) that consists of only those T-F units that contain useful information about the target sound. These ideal binary masks are generated by converting both the target and interfering signals into the time-frequency domain and directly calculating the SNR in each T-F unit. Those T-F units where the SNR is greater than a predefined Local SNR Criterion (LC) Value are assigned a 1 in the ideal binary mask, and those with an LC value less than the predefined LC Value are assigned a 0 in the ideal binary mask. This ideal binary mask can then be used to resynthesize a subset of the mixture audio signal containing only those T-F units where the local SNR value exceeds the predefined LC value.

Figure 3.1 illustrates the ideal binary mask using a 0 dB LC value for a mixture of the male utterance "Ready Baron go to blue one now" and the female utterance "Ready Ringo go to white four now", where the male utterance is regarded as target. The overall SNR of the mixture (measured from the RMS energy in each utterance) is 0 dB. In the figure, the top left panel shows the T-F representation of the target utterance, the top right panel the representation of the interfering utterance, and the bottom left panel the representation of the mixture. The middle panel shows the ideal mask for the mixture based on 0 dB LC, and the bottom right panel the output signal that has been resynthesized by applying the binary mask to the mixture. Note that the masked mixture appears much closer to the clean target than the original mixture.

Figure 3.2 illustrates the effect that varying the LC values has on the ideal binary mask for the two-talker speech mixture shown in Figure 3.1. The left and right panels show the ideal mask and resulting resynthesized mixture with the LC value set at -12 dB (top row), 0 dB (middle row), and +12 dB (bottom row). As can be seen from the figure,

increasing the LC value requires a higher local SNR value to retain a particular unit and reduces the total number of T-F units retained in the resynthized mixture.

Although very few psychoacoustic experiments have been conducted with ideal binary masks, they have been used quite extensively in CASA. The notion of the ideal binary mask was first proposed by Hu and Wang [70] as a computational goal of CASA, and further developed by Roman et al. [111] and Hu and Wang [71]. Binary masks had been used as an output representation in the CASA literature [23, 139]. Cooke et al. [37] used the *a priori* mask - defined according to whether the mixture energy is within 3 dB of the target energy – in the context of robust speech recognition. Roman et al. [111] conducted speech intelligibility tests and found that estimated masks that are very close to ideal ones yield substantial speech intelligibility improvements compared to unprocessed mixtures. Furthermore, Wang [138] gave an extensive discussion on the use of the ideal binary mask as the computational goal of CASA. Using the results from our study in this chapter, we will provide further evidence for the validity of this proposal at the end of the chapter.

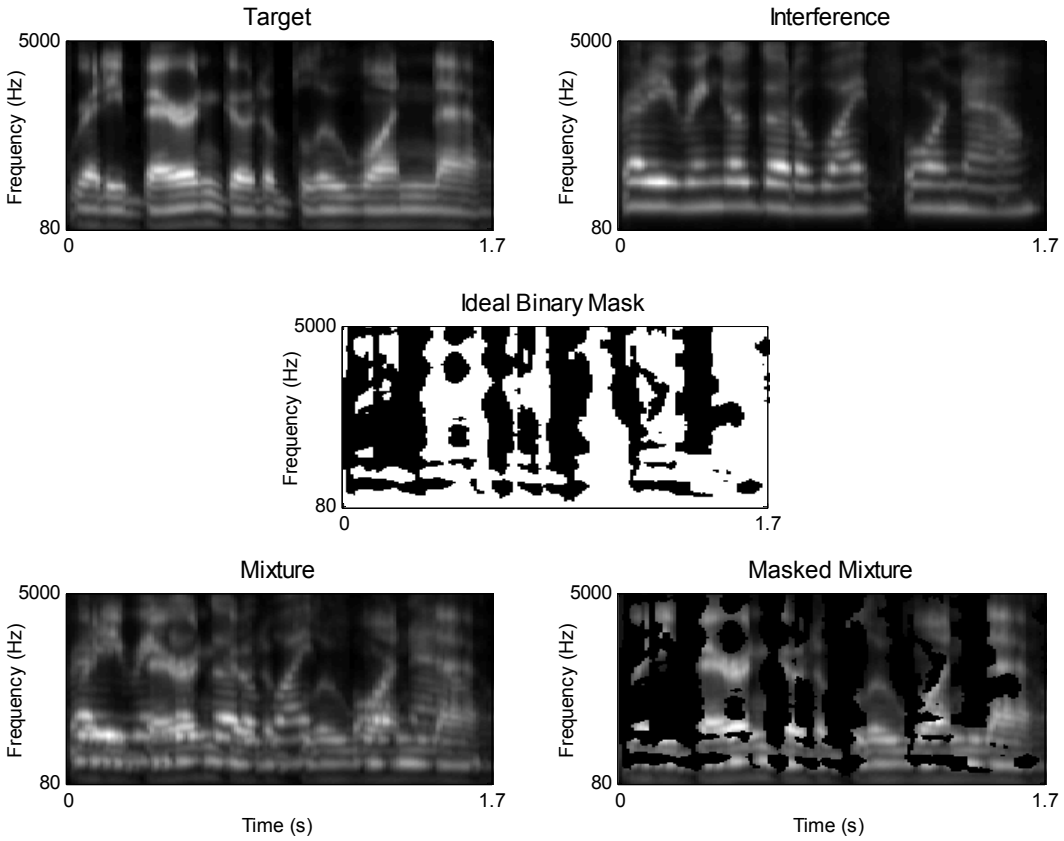


Figure 3.1. An illustration of the ideal binary mask for a mixture of two utterances. (**Top left**) Two-dimensional time-frequency representation of a target male utterance ("Ready Baron go to blue one now"). The figure displays the rectified responses of the gammatone filterbank with 128 channels. (**Top right**) Corresponding representation of an interfering female utterance ("Ready Ringo go to white four now"). (**Middle**) Ideal binary mask generated at 0 dB LC, where white pixels indicate 1 and black pixels 0. (**Bottom left**) Corresponding representation of the mixture. (**Bottom right**) Masked mixture using the ideal mask.

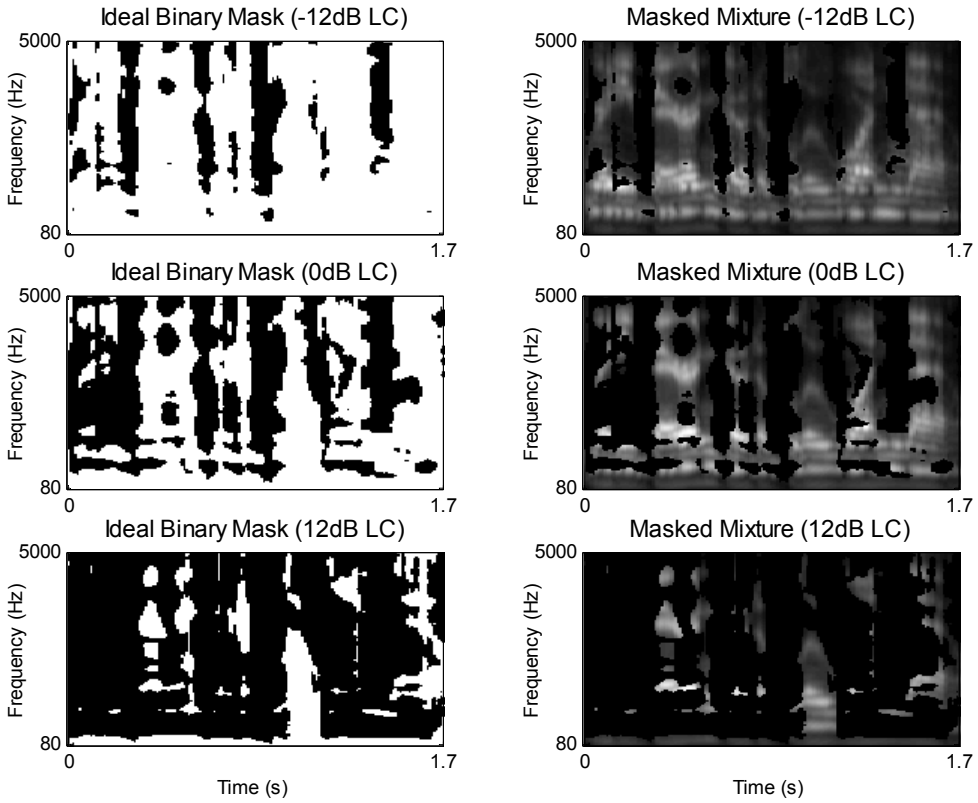


Figure 3.2. An illustration of ideal binary masking at different LC values, using the same speech mixture of two utterances from Figure 3.1. The three rows show three different LC Values (-12 dB, 0 dB, and +12 dB from top to bottom). The left column shows the ideal binary mask in each condition. The right column shows the corresponding masked mixture using these ideal masks. Note that increasing the LC value makes the binary masking procedure more conservative and thus decreases the number of retained units in the resynthsized signal.

3.2.2. IMPLEMENTATION

The techniques used to generate ideal binary masks in this study are very similar to those used in previous studies (e.g. [71]). For each stimulus presentation, the binary mask processing is based on a total of three input signals: a *target* signal, an *interfering* signal, and a *mixture* signal consisting of the sum or the target and the interference. Each of these signals is first processed through a bank of 128 fourth-order gammatone filters with overlapping passbands [104] and with center frequencies that are quasi-logarithmically spaced from 80Hz to 5000Hz. Quasi-logarithmically spaced means that the frequency spacing is designed to be more linear in the low frequency range and logarithmic in the higher frequency range to match the auditory system's structure. The gammatone filterbank design is selected because the magnitude characteristics of the fourth-order gammatone filter are approximately equivalent to the frequency resolution characteristics of the human auditory filter, thus making the gammatone filterbank a reasonable representation of the cochlear filtering that occurs in the human auditory system [94].

The gammatone filterbank effectively decomposes the signals into arrays of 128 narrowband signals, which are further divided into 20-ms time frames with 10 ms overlaps in order to produce a matrix of T-F units for each of the input signals. Then, within each of the T-F units, a comparison is made between the energy of the target signal and that of the interfering signal. The resulting local SNR for each T-F unit is compared to a predefined LC value to determine whether that particular unit should be retained or removed from the resynthesized signal: in T-F units where the local SNR is greater than or equal to the LC value, the binary mask is assigned a "1" for unit and the corresponding T-F unit in combined signal was included in the resynthesized signal; in T-F units where the local SNR is less than the LC value, the binary mask is assigned a "0" for that unit and the corresponding T-F unit in the combined signal is excluded from the resynthesized signal. Once this binary mask is defined, the output signal is resynthesized from the mixture signal using the same method that was described by Weintraub [145] (see also [23, 139]). Specifically, the binary mask is used to weight the filter output in

response to a mixture and the weighted filter outputs are summed across all frequency channels to yield the resynthesized waveform.

In the rest of this chapter, we describe the results of several experiments that are designed to examine the effect that ideal binary masking with different LC values has on the perception of multi-talker speech stimuli by human listeners. These experiments were conducted with the Coordinate Response Measure (CRM) corpus [15], a call-sign based color and number identification task that has been shown to produce a great deal of informational masking in cocktail party listening [25, 28].

3.3. EXPERIMENT 1: EFFECTS OF THE IDEAL BINARY MASK ON SPEECH INTELLIGIBILITY

3.3.1. METHODS

A. Listeners

Nine paid listeners have participated in the experiment. All have normal hearing and their ages range from 18-54. Most have participated in previous auditory experiments, and all are familiarized with similar experiments prior to conducting this experiment.

B. Speech Stimuli

The target and masking phrases used in the experiment are derived from the publicly available CRM speech corpus for multitalker communications research [15]. This corpus, which is based on a speech intelligibility test first developed by Moore [95], consists of phrases of the form "Ready (call sign) go to (color) (number) now" spoken with all possible combinations of eight call signs ("Arrow," "Baron," "Charlie," "Eagle," "Hopper," "Laker," "Ringo," "Tiger"); four colors ("blue," "green," "red," "white"); and eight numbers (1-8). Thus, a typical utterance in the corpus would be "Ready Baron go

to blue five now." Eight talkers - four males and four females – have been used to record each of the 256 possible phrases, so a total of 2048 phrases are available in the corpus.

For each trial in the experiment, a total of three audio signals are randomly generated and stored for offline binary mask processing prior to their presentation to the listeners. The first audio signal (the "target" signal) consists of a CRM phrase randomly selected from the corpus containing the target call sign "Baron." The second audio signal (the "interfering" signal) consists of one, two, or three different phrases randomly selected from the CRM corpus that are spoken by the same talker used in the target phrase but contains call-signs, color coordinates and number coordinates that are different from the target phrase and different from each other. Each of these interfering phrases is scaled to have the same overall RMS power as the target phrase, and then all of the interfering phrases are summed together to generate the overall "interfering" signal used for the binary mask processing. The third audio signal (the "mixture" signal) is simply the sum of the target and interfering signals for that particular stimulus presentation.

Note that, although all of the individual masking talkers in the combined signal are scaled to have the same RMS power, the overall SNR is less than 0 dB in the conditions with more than two simultaneous talkers. We clarify this distinction by referring to the ratio of the target speech signal to each individual interfering talker as the Target-to-Masker Ratio (TMR), and referring to the ratio of the target talker to the mixture of interfering signal(s) as the SNR [28]. Under this terminology, a mixture signal with two equal-level interfering talkers would have a TMR value of 0 dB and an SNR value of approximately –3 dB.

C. Ideal Mask Processing

Prior to the start of data collection, each set of three audio signals ("target", "interference" and "mixture") have been used to generate a single ideally-masked stimulus signal at a predetermined LC value. This ideal binary masking extracts only those time frequency units in the combined signal where the local SNR is greater than or

equal to the LC value. A total of 29 different LC values, in 3 dB increments ranging from –60 dB to +30 dB, are tested in the experiment. In addition, a control condition is used where the signals are simply processed and then reconstructed using an all-1 mask (or, equivalently, an LC value of negative infinity). This control condition is essentially equivalent to simply presenting the mixture of target and interfering signal to the listener.

D. Procedure

The listeners participate in the experiment while seated at a control computer in one of three quiet listening rooms. On each trial, the speech stimulus is generated by a sound card in the control computer (Soundblaster Audigy) and presented to the listener diotically over headphones (Sennheiser HD-520). Then an eight-column, four-row array of colored digits corresponding to the response set of the CRM is displayed on the CRT, and the listener is instructed to use the mouse to select the colored digit corresponding to the color and number used in the target phrase containing the call sign "Baron".

The trials are divided into blocks of 50, each taking approximately five minutes to complete. Each subject participates in 90 blocks for a total of 4500 trials per subject. These include 150 trial combinations for each of the 30 LC values (including the "no masking" condition) evenly divided among three talker conditions (2-talker, 3-talker and 4-talker). The trials are also balanced to divide the eight target speakers as evenly as possible across the trials being collected in each condition for each subject.

3.3.2. RESULTS AND DISCUSSION

Figure 3.3 shows the percentage of trials where the listeners correctly identify both the color and the number in the target phrase as a function of LC for each of two, three, and four simultaneous talkers configurations in the experiment. The data are averaged across the listeners in the experiment, and the error bars in the figure represent the 95% confidence interval of each data point. The points at the far left of the figure (labeled "no mask") represent the control conditions where all of the time-frequency units were

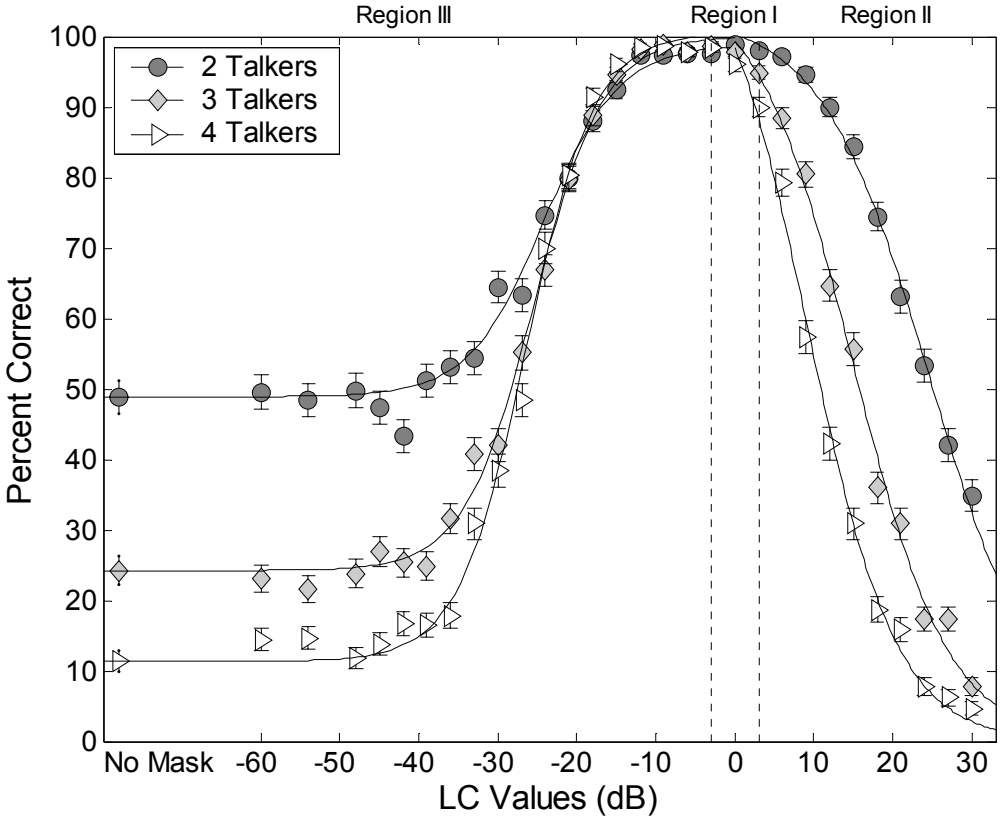


Figure 3.3. Percentage of trials in Experiment 1 in which the listeners correctly identified both the color and number coordinates in the target phrase as a function of the LC values. A T-F unit corresponding to the mixture is retained in the final output stimulus only if the target energy at the T-F unit is greater than the various specified intensity levels in dB relative to the combined masking energy. The legend indicates the number of simultaneous talkers tested in the experiment. The error bars represent 95% confidence intervals (±1.96 standard errors) in each condition. Because of the discontinuity in the performance curve at 0 dB, two different logistic curves were used to fit the positive and negative LC values in each talker condition [3, 34]. At negative LC values, this logistic curve was set to asymptote at the performance value achieved in the "no mask" control condition.

retained in the resynthesized signal. This is also called a signal resynthesized from an all-1 mask. These points indicate that the listeners are able to correctly identify the color and number coordinates in the stimulus in approximately 50% of the trials with 2 simultaneous talkers, 25% of the trials with 3 simultaneous talkers, and 12% of the trials with four simultaneous talkers. These results are consistent with previous experiments that have examined performance in the CRM task with two, three, or four identical talkers [28].

Although the overall number of correct color and number identifications clearly decreases as the number of talkers in the stimulus increased, the general pattern of performance is similar for all three of the talker conditions tested. For purposes of discussion, it is easiest to divide these performance curves into three distinct regions that are clearly defined in all three of these performance curves shown in Figure 3.3 and have very different interpretations with respect to the effect of the LC value. Each of these regions is discussed in detail below.

A. Region I: Performance for LC values near 0 dB

In the CASA applications that have utilized the ideal mask as a performance standard, the LC value of 0 dB has traditionally been used and the region of values near that region most closely matches the concept of the "ideal binary mask". Assuming that sound is energetically audible at or above 0 dB local SNR, then the auditory stimuli in this region consist of all the time-frequency (T-F) units that contain useful information about the target signal, and none of the T-F units that contain useful information about the interfering signal. In other words, the stimuli incorporate the same loss of information from the target signal that would occur due to energetic masking, but contain none of the acoustic elements of the interfering voices that might cause confusion with the target signal and in turn generate informational masking. Thus, we claim that the performance results shown for Region I in Figure 3.3 represent the degradation in multi-talker speech perception tasks that can be attributable to energetic masking effects, particularly for the

CRM corpus. As the figure clearly shows, there is not much degradation at all, implying the insignificant effect of the energetic component in this task.

In fact, examining the results of Figure 3.3 more closely, we can see that the subjects perform near 100% in all of the tasks, including the task with 4 competing voices. Comparing the performance in Region I with the no-mask control conditions at the far left of the graph, energetic masking effects do not account for any of the 50% to 90% drop in performance that occur when the competing talkers are added to the no-mask stimulus. To be fair, this surprising finding can be partially attributed to the use of the CRM corpus, which contains only a small number of possible responses (four colors and eight numbers). With fewer possible responses, the listener can take advantage of the redundant phonetic in speech and the task is thus more resistant to masking by noise effects than many other intelligibility tests [25]. Nevertheless, it should still be acknowledged that energetic masking plays a very insignificant role in multi-talker mixtures, at least in the CRM task with up to 4 simultaneous voices. In these tasks, with the target and interference presented at approximately the same level (0 dB TMR), informational masking is the dominant source of masking effects.

Observing Region I again in Figure 3.3, all of the conditions tested exhibit a sharp discontinuity towards the positive LC range, while exhibiting a plateau in performance near 100% going from 0dB LC to the negative LC range. This result indicates that 0 dB LC is indeed the "ideal" criterion where the listener can extract useful information about the target signal locally at any T-F unit. With an LC value of 3dB, all of the T-F units where the local SNR value lies between 0-3dB would be removed, and the sharp drop-off in performance indicate the listeners need the information available in those units to achieve optimal performance. As the performance has already peaked around 0 dB, changing the LC toward the negative side would not give any additional useful information about the target needed for identification and recognition, even though more units containing some target energy would be retained in the resynthesis. Drullman et al. [47] did find that listeners can still extract useful information from some portions of the time-frequency speech signal at a little lower level compared to the noise. That cannot be

verified with this task because the listeners have already reached ceiling performance at 0 dB LC. With future experiments, we will look at whether using other kinds of tasks will still produce such optimal performance at 0 dB LC or perhaps keeping more units might be more useful, even though the target level is lower than the interferer's level. If that is the case, then the "ideal" LC value for ideal masking may need to be reconsidered.

B. Region II: Energetic masking effects for LC values greater than 0 dB

All three of the performance curves in Figure 3.3 show a sharp drop-off in performance when the LC value is increased above 0 dB. As discussed before, increasing *x* LC value in this region will eliminate all the T-F units where the target signal is not at least *x* dB more intense than the interfering signal. It can be argued that a 1 dB increase in the LC value in this region is equivalent to a 1 dB decrease in the effective TMR of the stimulus (see Experiment 4 for a detailed discussion of this approximation). In each case, the listener loses the target signal's information if it is not at least 1 dB more intense than the masker.

Based on that interpretation, increasing the LC values above 0 dB effectively demonstrates the increase in the energetic masking effect of multi-talker speech perception as the TMR of the stimulus decreases. The effective TMR at each point in Region II is simply assigned to be the negative of the LC value, as we have approximated in the previous paragraph. Thus, clearly the effects of energetic masking increase dramatically as the effective levels of the interfering voices are increased above the level of the target voice, and that energetic masking also increases significantly with each additional competing voice. Nevertheless, the results also show that even at very low effective TMR, such as at LC value of 20 dB, performance is well above chance for all conditions. Even at an LC value of 30 dB, where the effective TMR is -30 dB, performance is still at about 35% for the 2-talker task. In any case, Region II represents the errors that are made when the phonetic elements of the target signal are removed from the stimulus due to spectral overlap in the target and interfering signals (i.e. masking by subtraction).

C. Region III: Informational masking effects for LC values less than 0 dB

As stated before, assuming that all of the useful information of the target signal is contained in units where the local SNR is greater than 0, decreasing the LC value below 0 dB has no impact on the amount of target phonetic information available to the listener. What does impact the listener when the LC value becomes negative is that the units where the interfering signal dominates are gradually added back in the output. When that happens, the likelihood that the interfering signal would confuse the listener increases. So the performance degradation occurring at negative LC values in Region III can be attributed to this confusion about which elements in the mixture belongs to the target, which is directly related with the concept of informational masking. Region III represents the errors that are made when the phonetic elements of the interfering signal are *added* to the stimulus and the listener becomes confused as to which elements belong to the target speech signal (masking by addition).

Now that we have established that Region III is dominated by informational masking effects, we can take a look at the effects of informational masking effects in multi-talker speech perception. The first observation is that informational masking effects increase gradually. At LC value of -12 dB, where the output contains interfering units as much as 12 dB more intense than the target signal, the performance is still near 100% for all conditions. This result suggests that the content of these interfering units were consistent with the target utterance, minimizing the opportunity for the listener to be confused about the target speech.

As the LC values become more negative, the stimulus starts to include more T-F units where very little target energy remains, and thus starts to differ from the characteristics of the target speech, and increases the opportunity to cause informational masking to the target speech. At around -12dB LC, the performance starts to drop rapidly and steadily from nearly 100% performance down to the same performance as the no-mask condition at -40 dB LC. Presumably, decreasing the LC value further caused no further

degradation in performance because all of the useful phonetic elements in the interfering signal have already been retained.

D. Error Analysis

Figure 3.3 gives the overall error rates in intelligibility for stimuli with 2, 3, and 4 simultaneous talkers at different LC values. However, it does not tell us what type of errors the listeners have made. For that purpose, Figure 3.4 shows an area graph of the listeners' responses for the color and number of the target signal for the 2-talker condition. The responses are divided into three categories: the darkly shaded region on the bottom represents the listeners' correct responses that match the color or number of the target phrase. The lightly shaded region in the middle represents the listeners' incorrect responses that match the color or number of the interfering phrase. The white region on the top represents the listeners' incorrect responses that do not match the color or number of the interfering phrase. The results in this figure match nicely with our previous interpretations. In Region I, performance peaks at nearly 100%. In Region II, where degradation in performance is dominated by energetic masking, we see indeed that the incorrect responses were randomly distributed. For the top panel representing the number response, the lightly shaded area indeed covers only about 1/7 of the total errors, as one out of the seven possible numbers. For the bottom panel representing the color response, the lightly shaded area covers about 1/3 of the total errors, as one out of the three possible colors that can be incorrect. This is consistent with the concept of energetic masking, as the masked elements of the target due to elimination of critical information do not cause bias for the listener to select any particular color or number.

The opposite is true for Region III, where the performance degradation is due to increasing informational masking, there exists a clear trend of confusion between target and interference. With the addition of more T-F units into the stimulus, the addition of elements dominated by the interference introduces a second voice into the stimulus and causes significant confusion for the listener on which voice the target signal corresponds to. The result is quite compelling, as nearly 100% of the incorrect responses for both the

number and color panel are from the number or color in the interference. This is precisely one of the major characteristics of informational masking, where as the target and interference are both audible, some central processing mechanisms and inability to segregate cause confusion to the listener and in turn select the response in the interfering signal. One more interesting observation with Region III in Figure 3.4 is that at no point does the listener exhibit a significant percentage of errors but does not match the number of color of the interfering phrase. At the LC values where the listeners are just starting to degrade in performance, such as in the -12 to -24 dB LC range, there might be cases where there are enough low-level interfering elements in the stimulus to confuse the target identification, but not enough elements to actually bias the listener to select the responses given in the interfering phrase. This implies that, for the CRM task at least, informational masking occurs at the word level, and not at the sub-word phoneme level. The phonemic elements in the interference added into the stimulus at the negative LC range have no effect on intelligibility until they themselves become intelligible and give the listener an alternative response. This also explains why not until past -12 dB does the intelligibility start to drop off. Finally, the results seen here may be partly caused by the CRM corpus. Informational masking may occur at a phoneme level with another corpus.

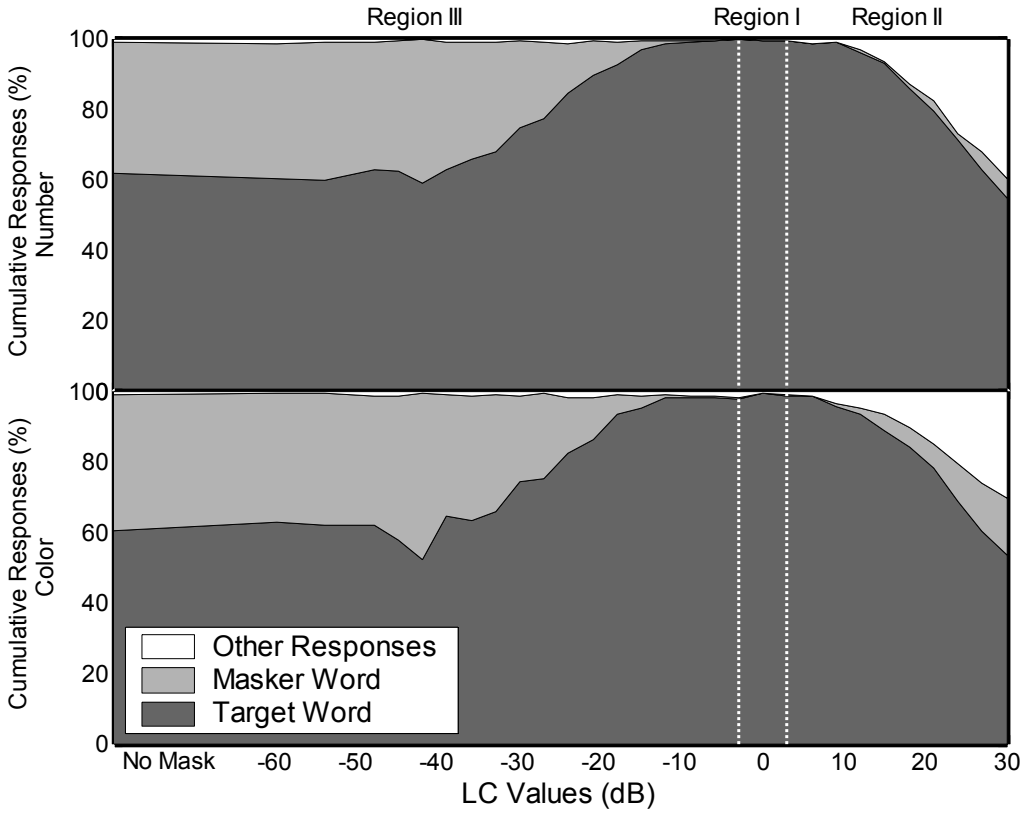


Figure 3.4. Distribution of listener color and number responses in the 2-Talker condition of Experiment 1. The top panel shows the distribution of listener number responses in the experiment: the darkly shaded area indicates correct responses that matched the number word in the target phrase; the lightly shaded area indicates incorrect responses that matched the number word in the masking phrase; the white area indicates responses that didn't match either of the number words contained in the stimulus. The bottom panel shows the same information for the color responses in Experiment 1. See text for details.

3.4. EXPERIMENT 2: EFFECTS OF SEX AND CHARACTERISTICS OF INTERFERING SPEAKERS WITH IDEAL BINARY MASKING

As discussed in the introduction of this chapter, listeners often find release from masking using the differences in the acoustic characteristics of the competing voices as a means to segregate the target voice from the speech mixture. In a past study, Brungart [25] found in monaural listening tasks as much as a 9 dB increase in speech performance when the two competing voice are composed of the opposite sex rather than composed of the same voices. The advantage can be attributed to differences in the fundamental frequency, differences in speaking style and intonation, differences in the vocal tract lengths, or a combination of the different factors.

As the coexistence of energetic and informational masking is inevitable in multi-talker speech stimuli, the natural question is how much effect does energetic and informational masking have in stimuli with voices of different sexes compared to stimuli with voices of the same sex or even of the same talker. It is quite reasonable to believe that energetic masking should play a prominent role in this issue because the difference in vocal tract length and fundamental frequency between a male and a female talker should produce some difference in the amount of acoustic overlap between the two voices. Using the isolation technique we have described earlier on isolating energetic masking using the speech-shaped noise matching the long-term average spectrum of natural speech, Plomp [55] found that changing from a noise masker shaped with the long-term average spectrum of female speech to a noise masker shaped with the long-term average spectrum of male speech produce a 2-3 dB decrease in the speech reception threshold (SRT) of a female target speech signal, but that changing from a male speech-shaped noise to a female speech-shaped noise actually producee a slight *increase* in the SRT for a male target speech signal. Since we now have a technique to isolate energetic masking, we can process the different stimuli with the ideal binary mask and see what kind of results would occur at the different regions dominated by informational and energetic masking.

Therefore, a second experiment is conducted to further examine the influence that target and masker similarity has on the energetic component of speech-on-speech masking. We use the ideal binary masking approach to compare multitalker listening performance with three levels of similarity between the target and masking voices: a same-talker condition (similar to Experiment 1), where the target and interfering phrases are spoken by the same talker; a same-sex condition, where the target and masking phrases are spoken by different talkers of the same sex, and a different-sex condition, where the interfering phrases are spoken by talkers who are of the opposite sex of the target talker.

3.4.1. METHODS

The procedures used in the second experiment were essentially identical to those used in the first experiment. The target and interfering speech signals were selected randomly from the CRM speech corpus, mixed together at a 0 dB target-to-masker ratio, and used to determine the ideal binary mask for a predetermined LC value. This binary mask was then used to resynthesize the combined audio signal, which was then presented to the listeners diotically over headphones. The listeners responded with the color and number coordinates contained in the target phrase addressed to the call sign "Baron."

The experiment was divided into three sub-experiments, with each sub-experiment examining performance with two, three, or four competing talkers over a specific range of LC values. Because some LC values were repeated across sub-experiments, this resulted in an uneven distribution of data collection across the fifteen LC conditions tested in the experiment (14 LC values ranging from –48 dB to +30 dB, plus a "no masking" control condition). Table 1 shows the number of trials collected in each LC condition for each of the listeners in the experiment.

Each stimulus presentation contained either two, three, or four simultaneous talkers, which could be composed of same-talker, same-sex, or different-sex masking phrase(s) depending on the particular condition of the experiment. As in Experiment 1, the stimuli for each listener were selected randomly prior to the start of the experiment, processed

offline, and stored on a PC for later presentation to the listeners. Again, the stimuli were selected randomly with the restriction that the interfering phrases contained call signs, colors, and numbers that differed from those in the target phrase and each other.

Nine paid volunteer listeners participated in the experiment. All nine were also participants in the first experiment. Combining the three sub-experiments, each listener participated in a total of 4500 trials, divided into blocks of 50 trials. The different number-of-speakers configurations and the different characteristics of the interfering speaker(s) were exactly distributed across all the trials. Thus there were 500 trials for each of the nine configurations (2,3, and 4 talkers by same-talker, same-sex, and different-sex configurations).

LC values	Trials per subject
No Masking	300
-48 dB	120
-36 dB	120
-24 dB	120
-18 dB	300
-12 dB	300
-6 dB	300
0 dB	300
+6 dB	450
+12 dB	450
+15 dB	270
+18 dB	450
+21 dB	270
+24 dB	450
+30 dB	300

Table 1. Number of trials collected in each LC condition

3.4.2. RESULTS AND DISCUSSION

Figure 3.5 shows the percentage of trials where the listener correctly identified both the color and the number in the target phrase as a function of LC values for all the configurations in the experiment. The figure is divided into three panels to separate the results for two, three, and four simultaneous talkers. Within each panel, the three curves show performance for the three different levels of similarity between the target and interfering voices (same-talker, same-sex, and different-sex). The data in the curves are averaged across the listeners have used in the experiment, and the error bars in the figure represent the 95% confidence interval for each data point. Just like in Experiment 1, each curve is also fitted with two logistic functions: one for LC values greater than 0 dB, and one for LC values less than 0 dB. At negative LC values, this logistic curve is set to asymptote at the performance level achieved in the no-mask control condition. At positive LC values, the curve is set to asymptote at zero percent correct responses, and it is additionally used to calculate the threshold value for 60% correct responses for each condition of the experiment (indicated by the horizontal dashed line in Figure 2). The 60% threshold values for each stimulus condition of the experiment are presented in Tabular form in Table 2.

As a whole, the results from Experiment 2 shown in Figure 3.5 are similar to those obtained in the first experiment. In the no-mask control conditions (left-most points on the graphs), performance in the same-talker conditions is comparable to that achieved in the no-mask conditions of Experiment 1 (roughly 12% correct responses in the 4-Talker condition, 25% in the 3-Talker condition, and 50% in the 2-Talker condition). Switching from same-talker interfering voices to different same-sex interfering voices improves performance substantially in the 2-Talker condition, but produces only a modest improvement in performance in the 3- and 4-talker conditions. Switching from same-sex interfering voices to different-sex interfering voices, however, produces substantial performance improvements in all the configurations tested. These results are consistent with those that have been obtained in other experiments that have examined the effect of

target-masker similarity on 2-, 3-, and 4-talker listening with the CRM at a 0 dB target-to-masker ratio (TMR) [28].

The curves for the 2-, 3-, and 4-talker conditions in Figure 5 also exhibit the same overall pattern of performance seen in the curves from that Experiment 1 shown in Figure 3. In Region I, where the LC value is near 0 dB, performance is near 100% correct in every condition tested. Again, this indicates that the effects of energetic masking are essentially negligible in the CRM task with 2-4 simultaneous talkers. The patterns are also the same at the two regions, where performance drops off sharply at the positive LC values in Region II, and drops off more gradually at the negative LC values in Region III.

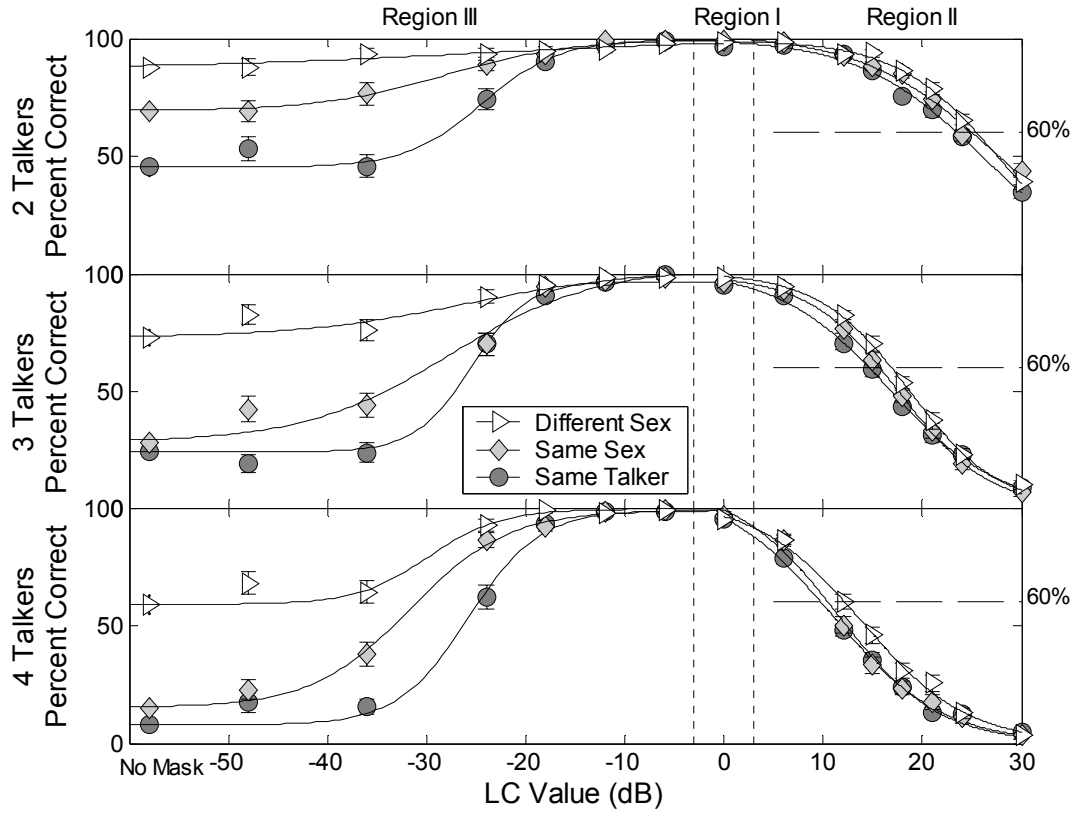
In Region II, where the LC values are greater than 0 dB, this drop off can again be argued to be the effect of energetic masking when the TMR of the stimulus is below 0 dB, specifically where the TMR is the negative of the LC value. Confining ourselves to this region, we can see that energetic masking has a much stronger effect on the number of talkers compared to the similarity of voice characteristics of the competing talkers. Observing the 60% threshold LC values shown in Table 2, the average across the conditions with different voice characteristics shows significant difference among the number of speakers. The threshold for 60% correct performance occurred at 24.69 dB in the 2-Talker task, 15.73 dB for the 3-Talker task, and 10.68 dB in the 4-Talker task. Assuming that this region is dominated by energetic masking as we have claimed, increasing from 2 to 3 competing talkers has an effective increase of 9 dB increase in the level of the interfering talker. Increasing from 2 to 4 competing talkers has an effective increase of about 14 dB increase in the level of the interfering talker. This increase is quite dramatic compared to the 3 dB and 5 dB increase in overall masking that actually occur when a second and third interfering signal is added to the 2 competing talkers' mixture. This result supports the fact that adding more speakers to the mixture tend to "fill in the gaps" of speech with natural fluctuations, and become more evenly distributed in the time-frequency domain, causing more effective energetic masking than would be expected from the TMR (or SNR) of the stimulus [22, 55, 91]. This is a very interesting occurrence and we will examine this effect in more detail in Experiment 3.

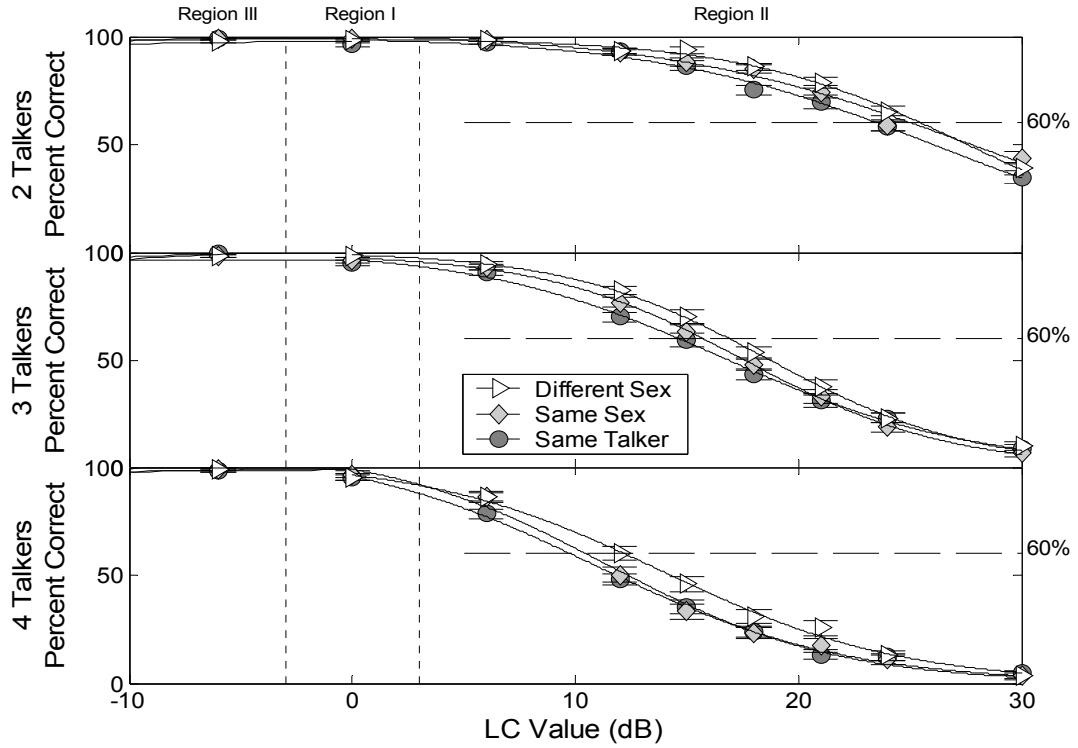
Getting back to Region II in Figure 3.5, we can see that changing the voice characteristics of the competing talkers seems to have negligible effect on intelligibility. Therefore we can argue target-interference similarity has negligible effect on the energetic component of speech-on-speech masking. Quantitatively, at the 60% threshold LC, Table 2 shows that changing from a same-talker interfering stimulus to a same-sex interfering stimulus merely produces an average of 1.25 dB increase, and changing from same-sex interfering stimulus to a different-sex interfering stimulus only produces an average of 1.12 dB increase in the 60% threshold LC (The higher the increase in LC value, the better the performance). Therefore, as a whole the voice differences between different-sex interfering talkers produce only about a 2.5 dB to 3 dB change in the effective energetic masking of the interfering signal. Furthermore, there does not seem to be much difference when switching from the same-talker to the same-sex condition, compared to switching from the same-sex to the different sex condition, which is observed in the no-mask condition. It is also worth noting that this estimate is similar to the 2-3 dB change reported by Festen and Plomp [55] for the SRT of a female voice masked by female-speech-shaped noise versus the SRT for a female voice masked by male-speech-shaped noise.

In Region III, as in Experiment 1, all the conditions tested produce a plateau in performance near 100% correct responses for LC values between 0 dB and -12 dB, a gradual drop-off in performance as LC values below -12 dB, and an asymptote in performance near the level achieved in the no-mask control condition at the lowest LC value tested (-48 dB). Again, this drop-off in performance can be attributed to informational masking that occur when the listener confuse the interferer-dominated T-F units that are increasingly retained as the LC values become more negative. Using the same argument as in Experiment 1, the plateau in performance from 0 dB LC up to about -12 dB LC is because the interferer-dominated T-F units that are retained still contain a lot of the target energy, and also may be consistent with the target energy as well. As the LC values become more negative, the target energy contained in the added T-F units are

starting to become negligible, and the interferer starts to cause confusion for the listener, causing substantial informational masking as a result.

Figure 3.5. (**Top**) Percentage of correct color and number identifications in Experiment 2 as a function of the LC value. The top panel shows results for the 2-talker conditions, the middle panel shows results for the 3-talker conditions, and the bottom panel shows results for the 4-talker conditions. The error bars represent 95% confidence intervals (±1.96 standard errors) in each condition. As in Figure 3, two separate logistic curves have been fitted to the data at positive and negative LC values, with the logistic curve at negative LC values set to asymptote at the performance level obtained in the "no-masking" control condition. Note that a dashed line has been drawn to indicate the 60% threshold level of performance at positive LC values that was used to produce the LC threshold values shown in Table 2. (**Bottom**) A more detailed view of the same graph at the positive range of LC values in order to emphasize on the energetic masking portion.


Region III
Region I
Region II
2 Talkers
Percent Correct
3 Talkers
Percent Correct
4 Talkers
Percent Correct
100
50
0
60%
Different Sex
Same Sex
Same Talker
No Mask
-50
-40
-30
-20
-10
0
10
20
30
LC Value (dB)


Region III
Region I
Region II
2 Talkers
Percent Correct
3 Talkers
Percent Correct
4 Talkers
Percent Correct
100
50
0
60%
Different Sex
Same Sex
Same Talker
-10
0
10
20
30
LC Value (dB)

	Interferer(s) voice characteristics					Mean
	Same Voice	Same-Sex		Different-Sex		
2 Talkers	23.50 dB	25.07 dB	Δ +1.59 dB	25.49 dB	Δ +0.42 dB	24.69 dB
3 Talkers	14.61 dB	15.69 dB	Δ +1.08 dB	16.89 dB	Δ +1.20 dB	15.73 dB
4 Talkers	9.37 dB	10.46 dB	Δ +1.09 dB	12.20 dB	Δ +1.74 dB	10.68 dB
Δ Mean			Δ +1.25 dB		Δ +1.12 dB	

Table 2: Maximum LC for 60% performance or better

3.5. EXPERIMENT 3: EFFECTS OF NUMBER OF COMPETING SPEAKERS WITH IDEAL BINARY MASKING

One interesting result from Experiments 1 and 2 is that none of the conditions indicates that any significant energetic masking occur when the effective TMR is set to 0 dB (LC value of 0 dB). Even in the most difficult case tested, the 4-talker same-talker condition, performance is near 100% when the LC Value is set to 0 dB (as compared to about 12% correct in the no-mask condition that includes the effects of both informational and energetic masking). It raises the interesting question of how many simultaneous overlapping equal-level speech signals *are* necessary to produce a significant amount of energetic masking in the CRM listening task. Finally, we will also bring back results from previous experiments to compare some of the effective energetic masking effects caused by the decrease in the effective SNR of the stimulus. In order to further explore these questions, a third experiment is conducted to examine performance in the CRM listening task as a function of the number of competing talkers when the LC value is fixed at 0 dB.

3.5.1. METHODS

The procedures in Experiment 3 are again similar to those used in the first two experiments. For each trial, the target and interferer phrase(s) are randomly selected from the CRM corpus, scaled to have the same overall RMS levels, and summed together to produce target, interferer, and combined signals. The procedures outlined in Section 3.2 and 3.3 are then used to generate a binary mask with the LC Value set to 0 dB, and this binary mask is used to produce resynthesized combined output signal that have been stored off-line on a PC for later presentation to the listeners in the experiment. The primary difference between Experiment 3 and the earlier experiments is the number of interfering talkers: in Experiment 3, the number of interfering talkers in each trial is randomly selected from one of 10 values ranging from 1 to 18 (1, 2, 4, 6, 8, 10, 12, 14, 16, and 18), with all of the interfering voices same as the target talker.

Due to the large number of simultaneous talkers for some cases, as well as the limiting variety of unique call-sign/color/number combinations in the CRM corpus, extra care is undertaken to ensure that the phrases presented within each trial overlap as little as possible. Previously, call signs, colors, and numbers cannot repeat within the same trial. In this experiment, while the target call sign remains unique with "Baron," the masker call signs, colors, and numbers can duplicate within each trial. The distribution of these conditions within each trial is assigned to minimize such duplication. The same nine listeners who have participated in Experiment 2 also participate in Experiment 3, with each subject conducting 10 blocks of 50 trials.

3.5.2. RESULTS AND DISCUSSION

Figure 3.6 shows the percentage of trials where the listeners correctly identify both the color and the number in the target phrase as a function of the number of simultaneous talkers in the experiment. Note that the LC is fixed at 0 dB, and the speech mixture is presented at 0 dB TMR. As in Experiments 1 and 2, these results show the color and number identification performance near 100% when the stimulus contains only a small number of competing talkers. As the number of competing talkers increases beyond four, there is a gradual but steady decrease in performance. However, even in the worst condition tested, where the target speech signal is masked by 18 different competing speech signals spoken at the same level by the same talker, the application of an ideal binary mask with a 0 dB LC criterion results in a performance level well above chance (35% versus about 3% for chance performance). Theses results suggest that indeed in normal multi-talker speech perception tasks, the degradation in performance is mostly due to informational masking as opposed to energetic masking.

While the energetic degradation caused by increasing the number of talkers does not reach chance level performance, energetic masking certainly produces a substantial effect as more competing voices are added into the stimulus. A number of researchers already noted that this kind of phenomenon is attributed to the speech signals "fill in the gaps" to prevent listeners from catching glimpses of the target speech signal [22, 55, 91].

However, with previous techniques it would have been hard to accurately the size of this effect because we cannot isolate the energetic effects from the informational effects in multitalker speech mixtures. Now, with the ideal binary masking methodology, we can directly measure effects of energetic masking as a function of the number of interfering talkers in the stimulus and the effective overall SNR of the signal.

The results of our findings are shown in Figure 3.7. We have compiled results from Experiment 1 and Experiment 3 to generate this plot for comparison. In particular, the shaded circles in the figure show the results for the 10 different numbers of competing talkers from Experiment 3, with the number of competing talkers for each condition indicated in the center of each data point. We have plotted this data as a function of the mean overall SNR, which is simply calculated from the ratio of the total RMS energy in the target to the total RMS energy in the combined multi-talker interfering speech signal, depending on the number of talkers. For example, for the 3-talker case, the two statistically-independent interfering talkers are presented at the same level as the target speech, so the total mixed interfering signal in that condition is around 3 dB more intense than the target and thus produce an overall approximate SNR of -3 dB. For the 19-talker case, the 18 interfering talkers mixture produce a RMS energy level around 13 dB higher than the target phrase, so it has been plotted at an SNR value of approximately -13 dB. Because in the experiment many trials are conducted forming different combinations of speech stimuli, we have averaged the SNR value for all of the stimuli for each of the number-of-talker condition to get an approximate value for the graph.

The open and numbered circles in Figure 3.7 re-plot the data obtained at the different positive LC values tested in Experiment 1 to show performance as a function of the approximate effective overall SNR in a stimulus containing a fixed number of interfering talkers. For example, in the 4-talker condition of Experiment 1, the binary-masked condition with an LC value of 0 dB produce results that effectively simulate the energetic masking for a 4-talker interfering speech signal presented at a TMR of 0 dB. Such an interfering signal, which contains three independent speech signals with the same RMS power as target speech, has a RMS power that is approximately 4.8 dB higher than the

target. Consequently, the 4-talker condition with a 0 dB LC value is plotted at an SNR value of -4.8 dB in Figure 7. Similar procedures are used to plot performance from the 2-, 3-, and 4-talker conditions of Experiment 1 as a function of overall effective SNR for all SNR values from 0 dB to around -14 dB.

Finally, the closed symbols in Figure 7 show performance in the CRM task as a function of overall SNR for a continuous speech-shaped noise masker that is shaped to match the overall average spectrum of all of the phrases in the CRM corpus. It can be argued that this kind of speech-shaped noise masker is equivalent to an interfering signal comprised of an infinite number of independent interfering talkers. The data in the speech-shaped noise curve is re-plotted from [28]. They were collected with the same speech materials used in this experiment, but with a different set of listeners.

Looking at the different curves from Figure 3.7, it is clear that the effects of energetic masking increases with additional talkers at the same overall effective SNR. We can see a progressive increase of energetic masking from the 2, 3, 4-talker lines from Experiment 1, and the "multi-talker" line from Experiment 3 shows worse performance than the 2,3, and 4-talker lines starting when the number of talkers becomes 5. Finally, we can see that the result from the speech-shaped noise masker gives the worst performance at almost all of the effective SNR values. These results indicate the effects of energetic masking increase substantially when additional interfering talkers are added to a multitalker stimulus at a fixed overall SNR. The results from the speech-shaped noise masker also suggest that a very large number of simultaneous speech signals are needed to reach the asymptotic level where a multitalker speech signal produces the same amount of energetic masking as a random-phase speech-shaped noise.

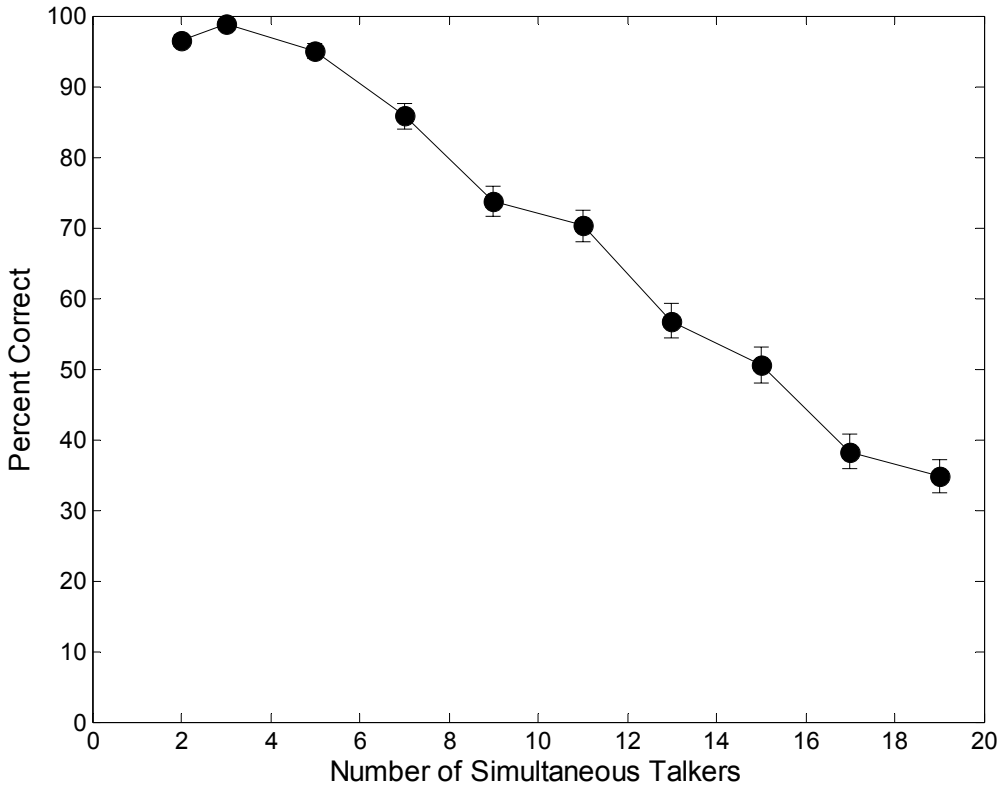


Figure 3.6. Percentage of correct color and number identifications in Experiment 3 as a function of the number of simultaneous talkers. The error bars represent 95% confidence intervals (±1.96 standard errors) in each condition. Up to 19 simultaneous phrases were presented to the listeners at once in this experiment.

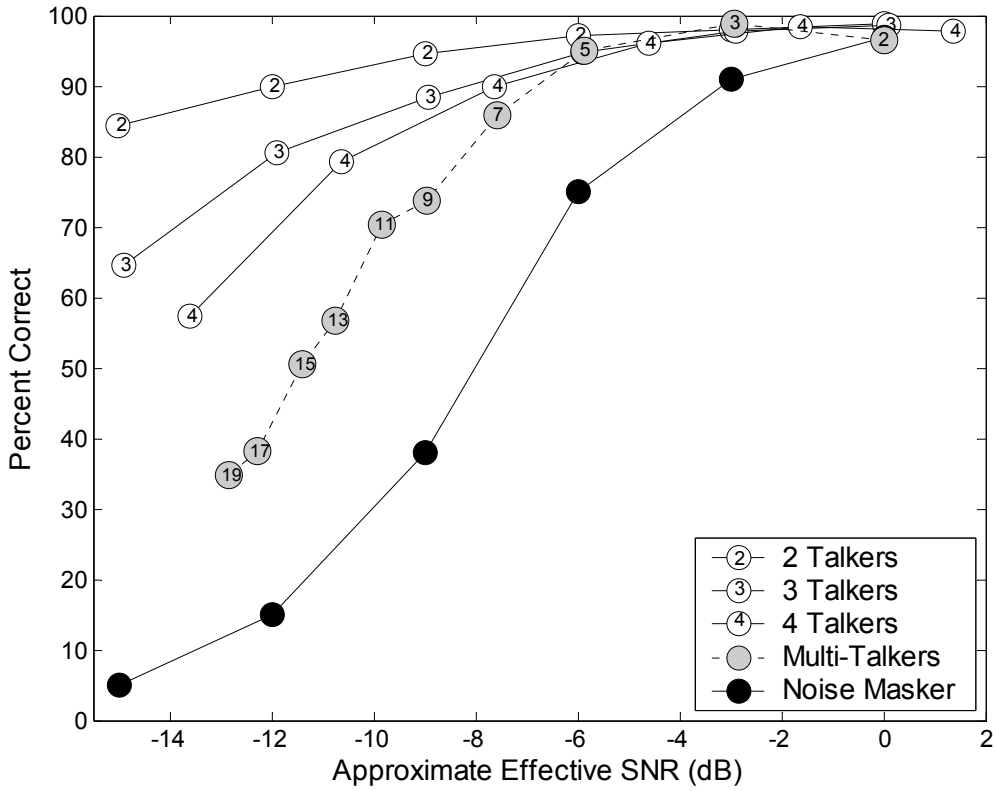


Figure 3.7. Percentage of correct color and number identifications in Experiment 3 as a function of the overall SNR. The shaded circles in the figure show the performance for each of the 10 different numbers of competing talkers tested in Experiment 3, with the number of competing talkers in each condition indicated in the center of each data point. In each case, the data have been plotted as a function of the mean overall SNR, calculated from the ratio of the total RMS energy in the target phrase to the total RMS energy in the mixture of multitalker interfering speech signal. The open and numbered circles re-plot the data obtained at the different positive LC values tested in Experiment 1 to show performance as a function of the approximate effective overall SNR in a stimulus containing a fixed number of interfering talkers. The closed symbols show the performance in the CRM task as a function of overall SNR for a continuous speech-shaped noise masker that matches the overall average spectrum of all of the phrases in the CRM corpus. The data in the speech-shaped noise curve have been re-plotted from Brungart [25]

3.6. EXPERIMENT 4: EFFECTS OF TMR ON RESYNTHESIS OF MIXTURE SIGNAL

In Section 3.2.2.B, we have noted that to a first approximation, a 1 dB increase in the LC value in Region II (See Figure 3.3) is equivalent to the 1 dB decrease in the effective target-to-masker ratio (TMR) of the stimulus, which is also the effective change in energetic masking. While we get the same binary mask for a signal with -3 dB TMR at an LC value of 0 dB as the signal with 0 dB TMR at an LC value of 3 dB, the resynthesized output is different. One would expect the performance for the 0 dB TMR and 3 dB LC signal to be better since at each unit the local SNR will be at least 3 dB for every unit that is retained.

We now wish to verify that the approximation is valid. Experiment 4 is conducted in which binary masks calculated for an input signal with a 0-dB TMR are used to resynthesize combined signals with TMR values that vary from -6 dB to +6 dB. Figure 3.8 shows the results of this experiment. Each panel in the figure represents a different number of competing talkers, and each line represents a different binary mask calculated with one of three different LC values (0, +3, or +6 dB) on an input mixture with a 0 dB TMR. The abscissa in each case represents the TMR of the combined signal that is used to resynthesize the binary masked signal. Thus, the different lines in each panel of the figure indicate the effects that changes in the binary mask (i.e. the number set of retained units) have on performance, while the points within each line represent the effects that changes in the local SNR values within the same set of retained time-frequency units have on performance. These results clearly indicate that the number of units retained in the binary mask, which is determined by the LC value, has a much greater impact on performance than the local SNR values within the retained units. Indeed, the local SNR within the retained units has very little impact on performance in the CRM task. This result argues the validity of the way we have calculated the effect of TMR on energetic masking.

It may be surprising that local SNR at each unit does not have the effect on intelligibility as the number of units has. Two factors may explain this result. Increasing the local SNR does not affect performance much because the listeners can extract useful information from the units that are much higher than the threshold values we have used in this experiment, so no marginal increase in performance is obtained when the local SNR of these units increases. The second factor is the observation that much of the information about the target speech signal is determined more by *where* the energy in the target speech signal is located in the spectrogram than the specific details of *what kind* of energy is present there. As we have mentioned in the introduction, modulated speech-shaped noise maskers matching natural speech envelopes can give a lot of information. Synthetic speech signals that consist of five or more frequency bands of sinewaves or narrowband tones modulated to match the envelopes of a normal speech signal in those frequency regions can be partially intelligible [117]. It has also been found listeners can understand synthetic speech signals consisting of a handful of sine waves that are amplitude and frequency modulated to track the first few formants of a natural speech signal [109]. So these results show it does not really matter what the underlying signal that is present within the different T-F units, as long as there are audible units at specific regions. Applying binary masking to a signal with a lower TMR gives the same units, and the underlying energy might be more interferer dominated, but the result will have a similar distribution of energy as the original. These kinds of signals is often referred to as "auditory chimeras", and have been shown to produce intelligible speech signals with a wide variety of underlying modulated signals [121]. Thus it is reasonable that the target speech intelligibility is well maintained when ideal binary mask calculated with 0 dB TMR is used to resynthesize a signal with a lower TMR. In terms of a scene analysis explanation for this phenomenon, depending on *where* the energy in the target speech is located, it may strengthen the good continuation cue as a mechanism to help group the target signal together into a recognizable sequence. Furthermore, schema-based integration mechanisms may also be involved to give an effect similar to the phonemic restoration where portions of the target signal replaced by noise can still be perceptually "heard" and thus successfully integrated [142].

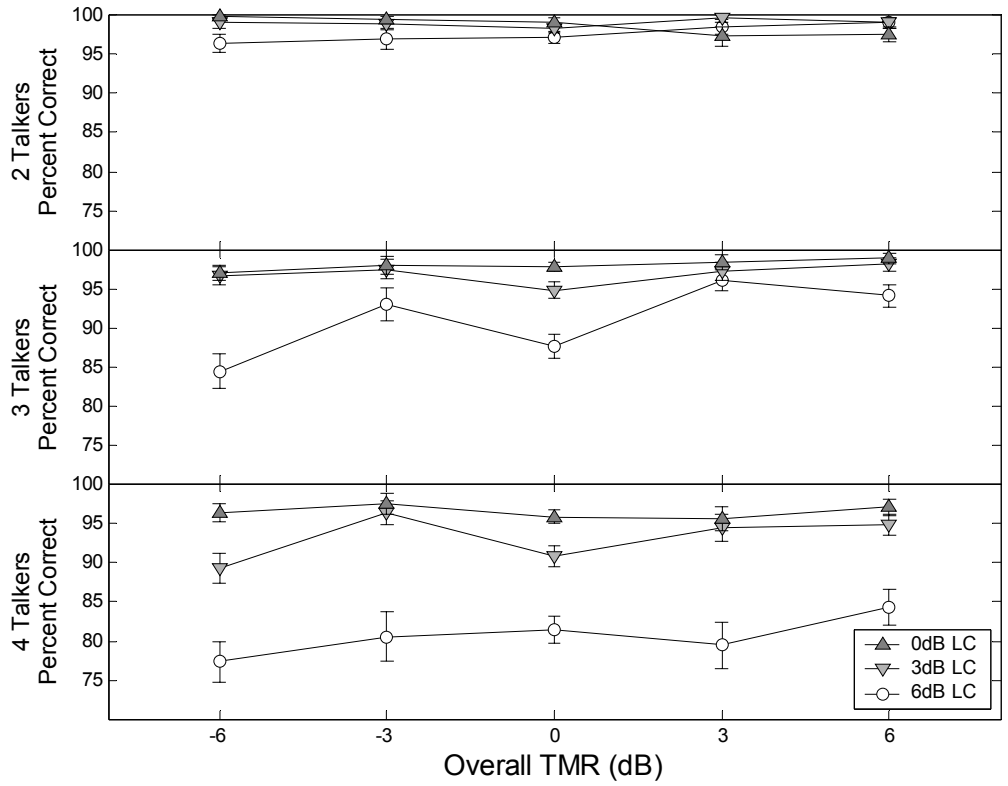


Figure 3.8. Percentage of correct color and number identifications as a function of the TMR of the combined signal used to resynthesize the binary masked signal. Each panel in the figure represents a different number of competing talkers, and each line represents a different binary mask calculated with one of three different LC values (0, +3, or +6 dB) on an input mixture with a 0 dB TMR. The error bars represent 95% confidence intervals (±1.96 standard errors) in each condition. The experimental procedures used for this experiment were similar to the previous experiments. As in the first experiment, all masking phrase(s) were the same voice as the target voice within each trial. For each TMR, each listener participated in 900 trials, divided in blocks of 50. Four different levels of TMR (-6 dB, -3 dB, +3 dB, +6 dB) were used to resynthesize the binary masked signal, and the data for 0dB TMR is actually taken from Experiment 1. Data for eight listeners who have also participated in Experiment 1 and 2 are shown here.

3.7. SUMMARY

3.7.1. CONCLUSIONS FROM THE EXPERIMENTS

As informational and energetic masking coexist in almost all natural speech mixtures, the isolation of each of these effects in multitalker speech listening environments is quite important to the behavioral study of auditory perception. Previous researchers have developed multitalker listening tasks that reproduce the informational masking effects occurring in natural speech perception without any significant contributions from energetic masking [2, 26]. In this chapter, the ideal binary masking methodology has been used to isolate the effects of energetic masking component of speech-on-speech mixtures. In general, our results maintain consistency with past results relating to informational and energetic masking. In addition, because we can now use ideal masking to isolate the energetic masking effects, we have also found several interesting results relating to the effect of energetic and informational masking that could not have been shown before because of the difficulty in quantitatively measuring the contribution of each effect. The major results presented in this chapter are summarized below.

A. Informational masking effects dominate performance when TMR is near 0 dB

In Experiment 1, the most difficult task is to respond to the target signal with 3 other interfering CMR phrases with the same voice as the target talker. After applying ideal binary masking at 0 dB LC, performance still reaches nearly 100%. This shows that the energetic effects due to temporal and spectral overlap in multitalker speech stimulus are not significant enough to degrade the performance in response to the color and number spoken by the target in the CRM corpus. Significant reduction in performance does not occur until the stimulus contains many more talkers, as shown in Expeiment 3, where the performance is still well above chance even with 19 competing talkers.

B. Similarities between the voice characteristics of the target and interfering talkers have a relatively minor effect on energetic masking

As shown in Table 2, changing from a stimulus with same-talker interfering phrases to a stimulus with same-sex interfering talkers results in a large release from informational masking, but only about a 1.25 dB release from energetic masking. Changing from same-sex interfering talkers to different-sex interfering talkers produces only an additional 1.1 dB release from energetic masking. Thus it seems that target-interferer voice similarity has only a relatively minor effect on the energetic masking effects that occur in multi-talker listening.

C. Adding additional talkers to a multitalker stimulus results in a greater increase in energetic masking than would be predicted from the overall decrease in the SNR of the stimulus

Researchers have argued that as more speech signals are added to the mixture, the "gaps" that exist in natural speech signals can be filled in to produce more effective energetic masking for the listener. Using ideal binary masking we are able to quantitatively measure the energetic effects. In the CRM task, the shift from 2 to 3 competing talkers decrease the overall SNR of the signal by about 3 dB but produce as much as 9 dB increase in effective energetic masking as we have discussed. Shifting from 3 to 4 competing talkers decreases the overall SNR by about 1.8 dB but increases in effective energetic masking by about 5 dB (demonstrated in Table 2). Nevertheless, while increasing talkers significantly increases energetic masking, the results of Experiment 3 suggest that significantly more than 18 interfering talkers are required to reach the asymptotic point where the energetic masking effects of adding an additional talker to the stimulus will match the results from a speech-shaped noise masker.

3.7.2 LIMITATIONS AND FURTURE RESEARCH

While the ideal binary mask can certainly be used to isolate energetic masking and help us understand the effects of masking in multitalker speech perception, there are some limitations to our current methodology that will need to be addressed in the future. The first is the impact of other kinds of masking effects critical to the proper measurement of

energetic masking. They include upward and downward spread of masking, as well as non-simultaneous masking that spread the energetic masking effects across T-F units. We have generated the T-F units in this experiment with the fourth-order gammatone filterbank, which provides a rough estimate of the frequency resolution of the human cochlea for relatively low-level sounds. However, the problem arises with the effect of energetic masking when the overall level of the stimulus varies. For example, the spread of masking across frequencies is much greater when the interfering signal is presented at a high pressure level, and thus we should expect more energetic masking at the very low effective TMR regions of our experiments.

Forward and backward temporal masking can also occur when a relatively strong interfering signal occurs in the same spectro-temporal region as a relative weak target signal [e.g 103]. It can have a significant impact on speech perception even in a signal containing only a single talker's voice [122]. The ideal binary masking model used in this experiment does not account for either of these effects, so again we might expect more energetic masking when applying our measurements to scenarios with very low effective TMRs, using the data we have with the relatively large LC values in Figures 3.3 and 3.5. In future studies, we should develop a more sophisticated binary masking model that would take these factors into account and produce a more accurate estimate of the effects of energetic masking for listening situations with a relatively low target-to-masker ratio.

Finally, we have frequently seen the insignificant effects of energetic masking in our experiments. While we believe that indeed multitalker speech perception is dominated by informational masking rather than energetic masking, the results might also partially depend on the corpus used for the experiments. Since the CRM corpus only consist of four phonetically-distinct colors and eight number alternatives, the task is relatively easy compared to other possible data sets with more response alternatives. Even in very noisy environments, performance has been good with the CRM corpus [28]. Thus, it is reasonable to believe the effects of energetic masking would be stronger even at 0 dB LC with other corpora and tasks, especially if the response set contains words that are more phonetically similar, such as the Modified Rhyme Test [68], or if the data simply contains

more alternative word responses. A recent study by Roman et al. [111] used unrestricted and semantically predictable sentences from the Bamford-Kowal-Bench corpus [13] and found that the binary masks that are very close to ideal binary masks with the 0 dB LC are very effective in removing interfering sounds (competing speech or babble noise) and improving speech intelligibility with low SNR conditions. Further research is still needed in this area to determine the energetic masking effects for different kinds of tasks in multitalker speech perception.

3.7.3 IDEAL BINARY MASK AS A COMPUTATIONAL GOAL OF CASA

One of the main evaluation criteria of a CASA system is evaluation with human listening [138]. The obvious motivation for this is that one can objectively evaluate the results of the system simply by listening to the system output and make the proper judgments. In order to have more quantitative results, intelligibility experiments such as the ones conducted in this chapter would be suitable to evaluate the system. Another motivation for using human evaluation is that a major practical application of CASA is to improve hearing prosthesis for listeners with hearing impairment, or listeners with normal hearing in very noisy environments. Based on these motivations, the intelligibility tests conducted in this chapter provides overwhelming support for the usage of the ideal binary mask as the standard of CASA. Our experiments have shown that for a range of different LC near 0 dB, ideal masking produces intelligibility results near 100%, certainly satisfying the human listening evaluation and setting the ceiling performance for all binary masks. Furthermore, the ideal mask is consistent with ASA constraints in terms of what can be heard and segregated, in direct correspondence with the auditory masking phenomenon. Recall from Figure 3, performance at the most negative LC value of region III is strongly dominated by informational masking, where all of the T-F units are assigned "1". One may think of this as the default state of the auditory system, where the auditory scene is in perceptual fusion [18]. As the LC values becomes more positive, the informational masking effects lessen, causing the target stream more likely to be segregated from the mixture. In region I, where the LC values are near 0 dB, informational masking is essentially gone, and in this region the target stream can be

most optimally segregated from the mixture, subject to only energetic masking by the interference. Therefore, as far as ASA is concerned, the range of LC values where informational masking has disappeared should be the optimal case for segregation of the target stream from the mixture. In addition, with its flexibility in specifying which stream is the target stream and its well-definedness regardless of the number of interferences in the scene, the ideal binary mask is a very desirable CASA criterion.

Our study in this chapter also demonstrates that in realistic environments, the effects of sounds from different sources have complicated relationships when input to our auditory system. While ASA principles provide well formulated theories on the direct relationships between the stimulus and the percept, the relationships quickly become complicated in real speech environments where one effect may obscure another and become difficult to isolate. Our methodologies help show some ways to separate different effects, in particular the energetic and informational masking in monaural speech mixtures. We believe that understanding the contributions of the different mechanisms, whether they are related with auditory masking, primitive segregation, or schema-based integration, will be key in providing the most accurate theories that can be utilized most effectively alongside other approaches, such as the physiological and the computational approaches.

CHAPTER 4
AN OSCILLATORY CORRELATION APPROACH TO ASA AND THE COMPUTATIONAL SEGREGATION OF ALTERNATING-TONE SEQUENCES

4.1 INTRODUCTION

In the previous chapters, we have discussed the behavioral approach to issues in auditory scene analysis, and covered in depth about specific complications involving speech-on-speech mixtures. While the theories derived from experimentation have revealed much about the various percepts of the auditory system given certain stimuli, they give very limited insights on how the brain represents these features, bind them together, and form stream segregation. This is where we need to consider Marr's second level of analysis, the analysis of representations and algorithms for ASA and auditory perception in general. The most relevant approach is the detailed analysis of the physiological processes of the auditory system, in order to understand the underlying mechanisms for the process. While the understanding of neural mechanisms for the organizational principles we have discussed is just at its infancy [54], there have been many studies and evidence available that allow us to ponder on the possible mechanisms. We explore these issues in this chapter and they will help us to develop sensible representations and algorithms for ASA and also help us transition into the implementation stage, the third level of analysis in Marr's framework.

4.2 NEUROPHYSIOLOGICAL MECHANISMS FOR AUDITORY STREAMING

4.2.1 CARRYING THE SOUND FROM THE EAR TO THE BRAIN

As we have discussed in the first chapter, the stimulus is received by the body's sensors and transduced into electrical signals, or neural activity, which are then transmitted to different areas of the brain. In audition, the transduction is accomplished in the cochlea in the inner ear, which contains hair cells with cilia on them whose bending leads to transduction [61, pg 321]. The electrical signal is then transmitted to fibers in the auditory nerve, and the auditory nerve carries the signals toward the receiving area in the auditory cortex. Along the way, the signal passes through structures of the cochlear nucleus such as the superior olive (SO), inferior colliculus (IC), and the medial geniculate nucleus (MGN). Each structure of this pathway is known to be responsible for some functions of control [14, 98]. In terms of the auditory cortex, it is composed of two-dimensional tonotopic maps one of which preserves the frequency response arranged in ascending order [74]. The other dimension is possibly a *latency* dimension formed by a sequence of delay lines [53, 126].

4.2.2 THEORIES OF AUDITORY NEURAL CODING AND REPRESENTATIONS

The foundation of the study of neural representations is based on studies in neural coding. For example, there has long been a theory for the place code for the ear's frequency analysis, where different frequency band's information is signaled by activity in the neurons at different places in the auditory system [12, 66]. There have also been theories of rate coding related to the rate of firing of individual neurons. However, the auditory stimulus tends to have much higher frequencies than individual neurons can fire, so direct rate-coding of individual neurons to signal a percept is not feasible. Temporal patterns of firing from a group of neurons can have aggregate firing of higher frequencies, called *volleying* [146]. Combining with the idea of *phase locking* [112], where neurons fire at particular phases of the stimulus, this allows the neurons to signal frequency through some temporal pattern of firing [38, 118].

In another form of temporal coding, multiple auditory objects can be represented by a group of neurons firing out of phase in response to one object with another group of neurons representing another object. This is thus consistent with our behavioral investigation in auditory stream segregation. Building from the theory of temporal binding, there is also evidence that a general increase in synchrony among neurons can occur independent of average firing rates. This synchrony can coincide with oscillations in cortical regions. In physiological experiments, cortical activity can be recorded by an electroencephalogram (EEG) in the form of event-related potential (ERP). 40Hz oscillations, known as 40Hz ERP or Gamma Synchronous Oscillations have been found in physiological experiments, and they are strongly related to responses of the sensory systems, including hearing [60, 85]. In general, oscillations in the frequency range of 20Hz to 70Hz are seen in the cortical and subcortical regions and exhibit synchronization when stimulated by a stimulus pattern [31, 49, 57, 63, 99]. Von der Malsburg [133] proposes that correlation among temporal responses of neurons are the mechanisms for binding features of the percept. Other experimental support exists as well [9, 51, 116]. The flexibility of the temporal synchronization theory is quite desirable because the coding is relational, depending entirely on the context, and each neuron can represent many different perceptual objects depending on the stimulus and how the cells are synchronized with each other.

Thus far, we have discussed a possible representational scheme that can connect the computational theories from behavioral studies with the underlying neural mechanisms during auditory stream segregation. The idea is to use the synchronization among a group of neural oscillators to represent one stream, and desynchronization among different groups of neurons to represent different streams. The oscillatory pattern has an extra degree of freedom in its *phase*, and can thus be used to represent synchronization and desynchronization. The proposed mechanism is that a set of auditory features forms a stream if the corresponding oscillators oscillate in phase, and different sets of oscillators oscillate out of phase to represent different streams [135]. Based both on topographical relationships among different tones in the Gestalt principles, and the two-dimensional composition of the auditory cortex in time and frequency, the representation is also two-

dimensional in effort to retain the topographical relationships in the set of features the system is representing.

4.3 BIOLOGICALLY PLAUSIBLE IMPLEMENTATIONS

4.3.1. IMPLEMENTATION BASED ON OSCILLATORY CORRELATION

With the representation and basic algorithmic requirements in place, we can now discuss the specifics of implementation for the final level analysis in the computational study of ASA. We have decided to utilize the time-frequency, oscillatory correlation representation for auditory streaming. One of the first models to implement these ideas was proposed by Von der Malsburg and Schneider [134]. Their mechanisms utilize the oscillator to represent the mean discharge response of a pool of cells. They constructed a network of fully connected oscillators, each receiving input from a frequency band and a global inhibitor that inhibits the synchronization of weakly coupled groups. This model simulates stream segregation based on onset synchrony.

4.3.2. LOCALLY EXCITATORY GLOBALLY INHIBITORY OSCILLATOR NETWORKS (LEGION)

One of the major limitations of previous models of oscillatory correlation is that they rely on long-range or fully connected connections to achieve phase synchrony. However, more local connections would be important because critical information of topological relations between objects is lost with a globally connected network. This is important because from the organizational principles we have studied so far, stream segregation demonstrates a clear dependency on the distances among tones in the time and frequency dimension, and thus this requirement needs to be represented in the model. Furthermore, previous models demonstrate lack of effective mechanism for desynchronization [114]. Thus, Wang and Terman [140] proposed a novel class of oscillator networks called LEGION (locally excitatory, globally inhibitory oscillator networks), which can both rapidly achieve synchronization within a group of locally coupled oscillatory units

representing the same object, and desynchronization among oscillator groups representing different objects. Based on local cooperation and global competition, LEGION is composed of the following elements [140]:

- Use of relaxation oscillators
- Local excitatory connections to achieve phase synchrony within an oscillator group representing the same object
- Global inhibitor to receive input from the network and in turn inhibits every oscillator to produce desynchronization of oscillator groups representing different objects

The building block of such a network is defined in the form of a feedback loop between an excitatory x_i and an inhibitory unit y_i as follows [127, 140]:

$$\frac{dx_i}{dt} = f(x_i, y_i) + I_i + S_i + \rho \tag{4.1a}$$

$$\frac{dy_i}{dt} = \varepsilon g(x_i, y_i) \tag{4.1b}$$

Where $f(x_i, y_i) = 3x_i - x_i^3 + 2 - y_i$ and $g(x_i, y_i) = \gamma(1 + \tanh(x_i / \beta)) - y_i$ in this particular implementation. I_i represents external stimulation to the oscillator and S_i represents the overall coupling from other oscillators in the network, which also includes the coupling from the global inhibitor. Symbol ρ denotes the amplitude of a Gaussian noise term. The x-nullcline and y-nullcline of Equation (4.1) is a cubic curve and a sigmoid function, respectively, as shown in Figure 4.1. For $I > 0$, the two nullclines intersect only on the middle branch of the cubic, and Equation (4.1) gives rise to a stable periodic orbit for all values of ε sufficiently small. The periodic solution alternates between an *active phase* of relatively high values of x, and a *silent phase* of relatively low values of x. The transition between the phases takes place on a fast time scale compared to the behavior within each of the two phases. When $I > 0$, the oscillator is *enabled*. For $I < 0$, the two nullclines intersect on the left branch of the cubic, and Equation (4.1)

produces a stable fixed point at a low value of x. In this case, the oscillator is *disabled*. Equation (4.1) has been interpreted biologically as a model of action potential generation of a single neuron [96], or as a mean field approximation to an interacting network of excitatory and inhibitory neurons [29, 123]. The building block of LEGION has two timescales, and thus belongs to a family of relaxation oscillators [136]. It is related to the van der Pol oscillator and the Hodgkin-Huxley equations for action potential generation [56, 67, 129].

The network exhibits the *selective gating* property, where an enabled oscillator jumping up to the active phase rapidly recruits the oscillators stimulated by the same object, or pattern, while preventing units representing other patterns to jump up to the active phase. The network can then achieve synchronization rapidly and desynchronization among different patterns. Desynchronized blocks never contain units that stay in active phase at the same time. Once an oscillator jumps up, it also triggers the global inhibitor which then attempts to inhibit the entire network and successfully inhibits weakly coupled units from jumping up at the same time.

LEGION can be extended to arbitrarily many dimensions, and has been successfully implemented as a two-dimensional network to achieve scene segmentation in vision [140, 141]. Regardless of the dimensionality of the network or the sensory modality of the application, LEGION can do scene segmentation or segregation when multiple objects or streams are mapped locally in some manner, and the local connectivity will group together the oscillators stimulated by a single object. The grouping is manifested by the phase synchrony within the oscillator group. Desynchronization is achieved by the global inhibitor that influences the network with inhibitory responses, causing weakly connected oscillator units to be out of phase. The reader can refer to Wang and Terman [140] for details of the computer simulations that demonstrate these effects. Terman and Wang [127] have analyzed and proven mathematically many of the computational mechanisms we have discussed so far.

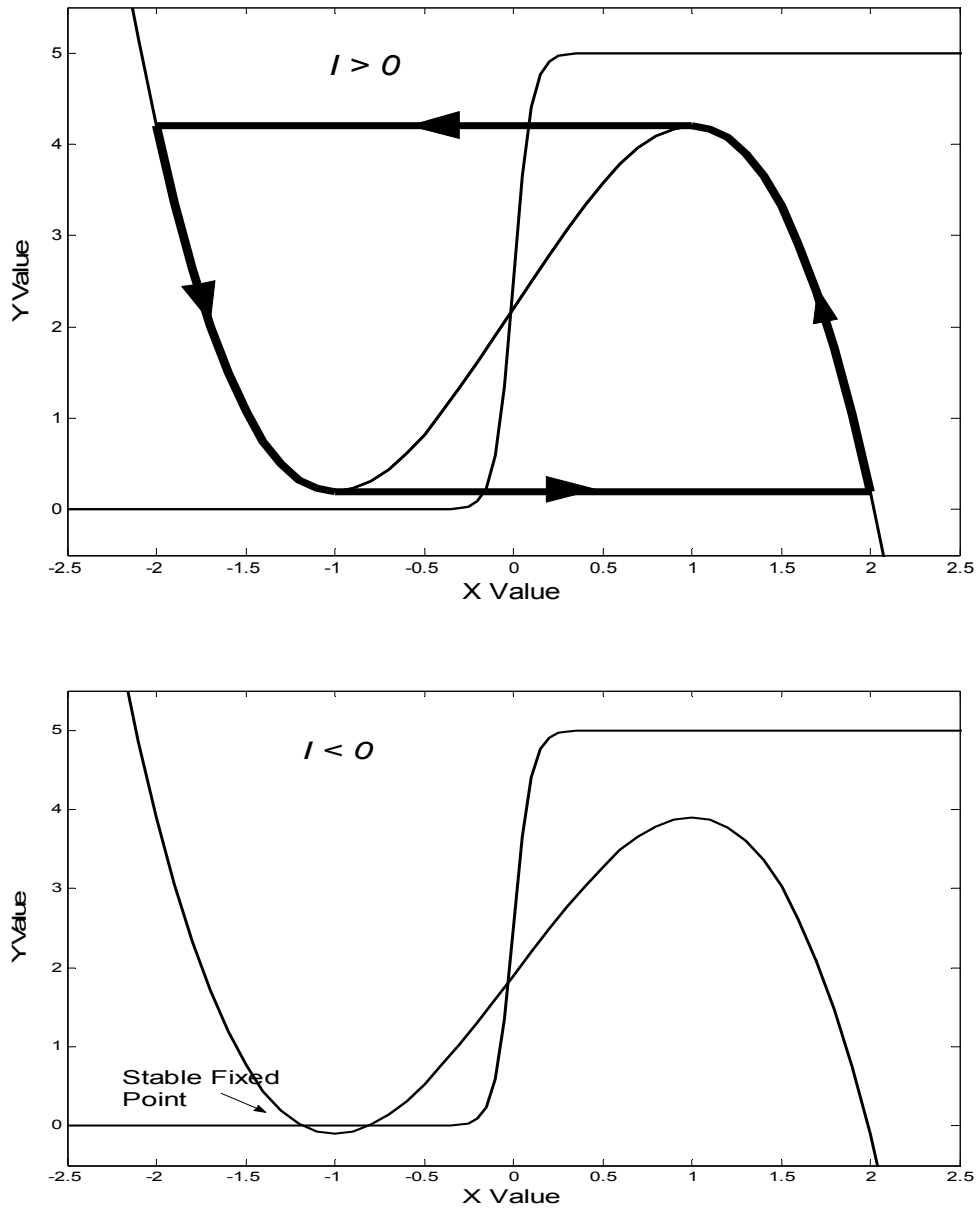


Figure 4.1. Similar to Figure 2 in Wang [135], this figure shows the nullcline for a single oscillator (nullcline is defined in Equation (4.1) when $dx/dt = 0$ and $dy/dt = 0$). (**Top**) For $I > 0$, Equation (4.1) gives rise to a stable periodic orbit for all values of ε sufficiently small. The periodic orbit is shown in bold with the direction of movement indicated by the arrows. The periodic solution alternates between an *active phase* of relatively high values of x, and a *silent phase* of relatively low values of x. (**Bottom**) For $I < 0$, the two nullclines intersect on the left branch of the cubic, and Equation (4.1) produces a stable fixed point at a low value of x.

4.3.3 LEGION IMPLEMENTATION DETAILS

With all of the essential elements of LEGION defined, we now examine some implementation details involved in order to use this network to simulate stream segregation. This LEGION implementation is represented as a two-dimensional matrix with constant size in the frequency dimension and constant time steps in the time delay dimension. Within each unit in the matrix, the presence or absence of the input stimulus is recorded as a binary record. The delay lines serve to help information be correlated across time, just as it does across frequency. Wang and Terman [140] originally implemented LEGION by numerically integrate the differential equations through time using the Runge-Kutta integration method, which is very accurate. However, this method requires significant computation time, and thus Wang [135] developed a set of algorithms to implement all of the essential behaviors of the dynamical systems. The algorithmic steps are quoted below ([135] p. 427), as they are similar to the algorithms we will utilize in the next section when we extend LEGION to model a specific auditory phenomenon.

- When no oscillator is in the active phase, the one closest to the jumping point among all enabled oscillators is selected to jump up to the active phase.
- An oscillator jumps up to the active phase immediately if it receives an excitatory input from its neighbors and the net input it receives from external input, neighboring oscillators, and the global inhibitor is positive
- The alternation between the active phase and the silent phase of a single oscillator take one time step only
- All of the oscillators in the active phase jump down if no more oscillators can jump up. The situation occurs when the oscillators stimulated by the same stream have all jumped up.

In this algorithm, all of the essential properties of relaxation oscillators are preserved, as well as the properties of synchrony and desynchrony. Nevertheless, these are approximating algorithms and caution needs to be taken when used or extended as to not trivialize the computation. Linsay and Wang [82] developed another approximating

methodology as an alternative to traditional methods of numerical integration, based on analysis of relaxation oscillations in the singular limit, thus called the *singular limit method*. The idea is to solve the system in the singular limit when the system evolves in the slow time scale, and to approximate the system in the fast time scale. The resulting computations become much faster to calculate.

4.4 COMPUTATIONAL SEGREGATION OF ALTERNATING-TONE SEQUENCES

4.4.1 INTRODUCTION

As a final effort in our brief, but complete, exploration of the computational approach to auditory perception in this thesis, we now go in depth to implement the perception of alternating-tone sequences. This is a very interesting phenomenon to study because while perceiving alternating-tone sequences seems quite simple at the surface, in fact its relationship to ASA can be associated with complex combinations of primitive scene analysis principles we have discussed in Chapter 2 as well as top-down integration such as the use of selective attention. Therefore, we must proceed carefully by understanding the specific behavioral results pertaining to alternating-tone sequences, in order to both build upon our existing knowledge of organizational principles in general auditory perception and to focus on some specific computational goals pertinent to alternating-tone sequences. Then, we can determine how to best achieve the computational goals by integrating the biological plausible oscillatory correlation representation and the LEGION implementation into the system. Through this exercise, we can see how the computational approach can link the behavioral and physiological results and help us perform the analyses needed to understand the underlying mechanisms of a complex phenomenon.

4.4.2 THE PHENOMENON

Extending the experiments performed by Miller and Heise [93], Van Noorden [131] exhaustively studied the perception of tone sequences in his dissertation. While alternating-tone sequences are simply pure tones of different frequencies repeated continuously in time, he found that the perceptual effects they have on the auditory system are far from simply a sequence of tones. In tone sequences that follow one another in quick succession, his experiments have shown that the listener does not process each tone individually. Instead, listeners perceive alternating tones with a small frequency separation to be coherent, while alternating tones with large frequency separation to form two separate streams perceptually segregated in that one can pay attention to only one stream at a time. It turns out that the frequency at which a coherent stream segregates also depends on the Tone-Repetition-Time (TRT), also known as the Inter-Onset-Interval (IOI), which is the time interval between the onset of two successive tones in a sequence. Furthermore, depending on TRT, there also exists a frequency range in which the listeners can switch between the percept of a coherent stream and segregated streams, using selective attention. Based on these results, Van Noorden went on to describe boundaries called the Temporal Coherence Boundary (TCB) and the Fission Boundary (FB), which show quantitatively the frequency differences of alternating tones that cause the listeners to perceive one percept over another, along different TRTs. His experimental results can be seen from Figure 4.2. The TCB is characterized by the curve with circles as data points. Given a TRT, above the frequency separation denoted by the TCB curve, the listener inevitably loses the perception of temporal coherence, and has to segregate the alternating tones into two streams. The FB is characterized by the line with x's as data points. Given a TRT, below the frequency separation denoted by the FB line, the listener inevitably loses the perception of fission, and has to perceive the alternating tones as one coherent stream. Between the two boundaries is the *ambiguous* region, where the listener can selectively attend to either the perception of one temporal coherent stream or two segregated streams. From the experimental results, we can see a clear dependence of TCB on TRT, with a weaker dependence of FB on TRT.

In order to computationally model this phenomenon, we must constrain our computational goals to fit this particular phenomenon. Besides achieving the automatic generation of the TCB and FB boundaries, the main constraint is to achieve these results using biologically plausible representations and implementations that could possibly be utilized in the future to describe the general behavior of the auditory system, and perhaps even all of the perceptual systems. Furthermore, as we have mentioned in Section 2.5, because the phenomenon described here is only one out of the many behavioral phenomena known in auditory perception, the model must be flexible enough that it can be integrated with other implementations to describe the general system. To this end, we decide to extend LEGION for auditory stream segregation, as LEGION's biological plausibility based on the oscillatory correlation representation was well established in the previous section. In addition, LEGION's simplicity and flexibility makes it a good candidate for extension to and integration with other perceptual phenomena.

While this study does not take account of the mechanisms of attention explicitly, the LEGION architecture can also be extended to incorporate components involving attentional mechanisms. In fact, LEGION has already been utilized in the study of computational modeling of selective attention (Wrigley and Brown, 2004). It is indeed one of our goals to apply similar models of selective attention into our model in the future.

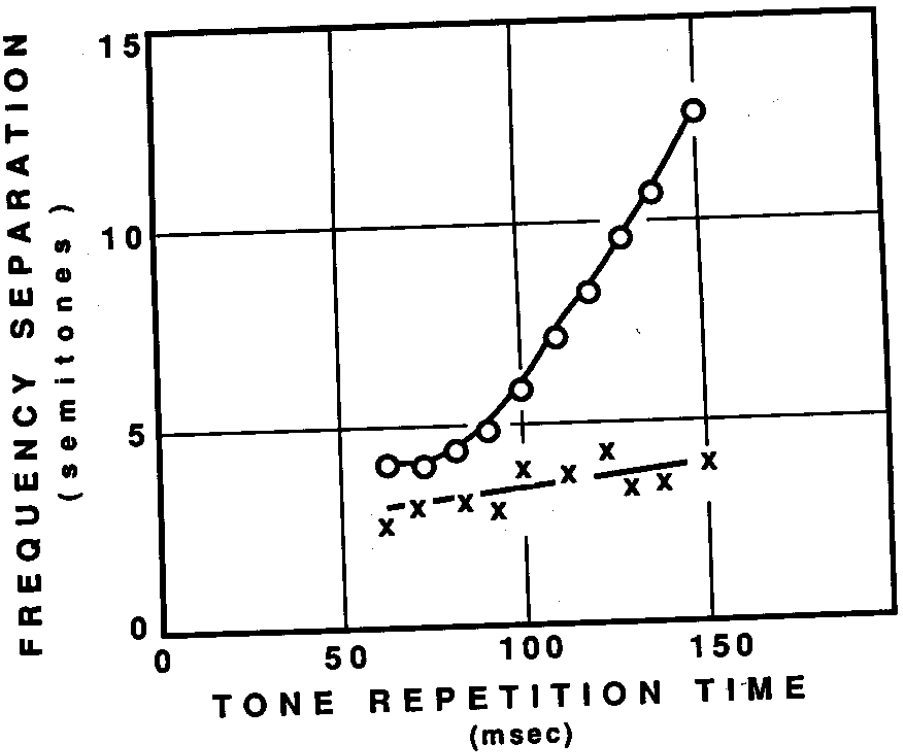


Figure 4.2: Extracted directly from Van Noorden ([131], p. 13). These are the psychophysical experimental results collected by Van Noorden in his dissertation.

4.4.3 THE PROPOSED MODEL

A. Oscillatory units and neural architecture

Because the foundation of our study is based on LEGION, the building block of the network is defined exactly the same as in Equation (4.1). The periodic solution oscillates between the active and silence phase as described before. The oscillator is *enabled* only if $I > 0$. As discussed before for the requirements of a plausible representation for auditory perception, the two-dimensional time-frequency representation is used, where the input to the network consists of units representing distinct frequencies, called the *input channels*. Each input channel connects to a corresponding row of oscillators, called a *frequency channel*, by a system of delay lines that map the sequential input presentation into the architecture in order to correlate information across time. A global inhibitor is present as well that has excitatory connections from all units and feeds back inhibitory connections to all units.

B. Lateral connections

One of the main features of LEGION is the ability to synchronize units representing the same pattern utilizing the locally excitatory connection weights. Wang and Terman [140] used two kinds of excitatory connection weights, permanent and dynamic weights. Dynamic weights are normalized and updated based on the permanent connection and play an effective role in influencing neighboring units. The original LEGION implementation used nearest-neighbor connectivity, which is too rigid for a complex system. Instead, we adopted the architecture where the connection strength between two oscillators falls off exponentially, by using the Gaussian distribution, as suggested by Wang and Terman [140] (see also [135]). Furthermore, we extend the idea that dynamic connections allow other kinds of connections to form depending on the state of the network in time, by making our system's connection weights to dynamically change intrinsically. Looking at Figure 4.2 reveals that the upper bound of frequency separation

to maintain temporal coherence increases dramatically with increasing TRT. This suggests a change in the shape of the Gaussian model of intrinsic connection according to a change in the TRT. In particular, the width of the Gaussian distribution along the frequency dimension seems to widen dramatically with high TRT, while the width along the time dimension does not seem to change much in the TRT range that we are interested in (less than 300 ms). Therefore, in our model, we have combined the original permanent connection representation and the dynamic connections into one intrinsically dynamic form:

$$J_{ij} = \eta \exp[-(\frac{(t_j - t_i)^2}{w_t^{\,2}} + \frac{(f_j - f_i)^2}{w_f^{\,2}})]H(x_i)H(x_j) \qquad (4.2)$$

Where t_i and f_i are the positions of oscillator in the time-frequency dimension, and w_t and w_f are the dynamic widths of the Gaussian distribution along the time and frequency axis, respectively. The function $H(x_i)$ acts as a heaviside function, where it equals 1 if oscillator i is recently enabled and its activity x is greather than a specified threshold (.05 in our implementation), and equals 0 otherwise. Note that the frequency values have been transformed into logarithmic values in Hz. It turns out that with absolute frequency values in Hz, frequency ratio is more fitting in the calculations rather than frequency difference between two different tones. Thus, once the frequency domain is converted to the logarithmic domain, as typical in auditory peripheral processing, division of frequency becomes equivalent to subtraction of frequency. Similarly, in the time dimension, the actual calculations for the time delay difference between t_i and t_j is compressed by several factors in msec (10 is used in our model), in order to emphasize the strength of the connection in time based on our stimuli and task. Self connectivity J_{ij} is set to 0. w_f represents the dynamic change of the Gaussian width along the frequency dimension which we designate to be directly dependent on TRT. This suggests that we would also need an additional map of tone onsets available to the network. The idea of having a map of tone onsets has been proposed by Norris [101] to model a related phenomenon called stream bias adaptation, which we will also discuss in the discussion

section. w_t represents the Gaussian width along the time dimension. Since connection strength along the time dimension does not change according to our earlier observations, we keep the value constant for our model. We propose the following relationship between w_f and TRT:

$$w_f = L_f + \frac{U_f - L_f}{1 + \exp(-\kappa_f (t - \theta_f))} \tag{4.3}$$

Thus this is a sigmoid function, where t denotes TRT, U and L denotes the parameters for the upper and lower bounds of the Gaussian width along the frequency axis, and κ_f and θ_f are the parameters of the sigmoid function. We generated a preliminary set of data points for w_f through simulation trials to achieve the desired outcome and then adopted the curve fitting algorithm for logistic functions described in Arnold [3] (see also [34]) to describe a concrete relationship between w_f and TRT. The final values we used were: U_f = 11, L_f = 2.3, κ_f =0.03, θ_f =226. In our case w_t = 50 and the upper bound of w_f is limited to U_f.

With this extension, the connection strength in frequency now dynamically changes according to the stimulus presentation rate, and thus is an example of stimulus-dependent rapid changes in the network. This is consistent with recent physiological studies relating to rapid task-dependent plasticity of spectrotemporal receptive fields [59], and provides evidence for the biological plausibility of our extension to the network implementation. One other critical modification to the system is the requirement for a map of tone onsets to be available for input to the network alongside the existing representation of time-frequency map that forms the basic architecture in LEGION. This map is necessary in order to take account of the presentation rate, and Norris ([101] p. 137) claimed that having a separate onset map may actually improve the biological plausibility of the overall system. In his case it is a partial solution to the phenomenon of the stream bias adaptation. Studies have shown that onsets are extracted from the signal at the peripheral level, and passed to the central auditory system [97]. Beauvois and Meddis [10] also

claimed that bias adaptation depends on the density of onsets. Furthermore, various onset detecting neurons have actually been identified in the auditory system [106, 108]. This evidence shows that having a tone onset map is supported physiologically and does not compromise the biological plausibility of our model. For our implementation, we have simply assumed the existence of an onset map. How the auditory system arrive at such a map has not been considered, although Wang [135] suggested that SegNet can simply be extended to include another layer of neural network for detecting stimulus onset. Smith [119] (see also [120]) have actually built computational models that specifically incorporated the detection of onsets and offsets into sound segregation, and he proposed to use different filters such as the difference of Gaussians to detect the onsets of sound.

C. Global inhibition

As discussed before, an important part of the overall coupling is the global inhibitor. Its main purpose is to prevent synchrony in disjoint blocks of units that do not have enough neighboring excitatory connections (Wang and Terman 1995). Therefore, the dynamical activity of the global inhibition z can be defined as in Wang and Terman[140]:

$$\frac{dz}{dt} = \phi(\sigma_\infty - z) \tag{4.4}$$

In this formulation, $\sigma_\infty = 0$ if the x value of every oscillator is under an activity threshold, and $\sigma_\infty = 1$ if at least one oscillator i is greater than or equal to an activity threshold. If every oscillator is below that threshold, then the global inhibitor will not receive any input and z will approach 0 rapidly, and therefore the network would not receive any inhibition. However, if at least one oscillator is above that threshold, then global inhibitor will receive input and x will rapidly approach 1, causing the entire network to receive inhibition.

As discussed in the Section 4.4.2, listeners are able to switch between the percepts of temporal coherence and segregated streams using selective attention. Depending on the

TRT, the range of frequency difference that the listener can hear both percepts can vary quite dramatically due to the ability to maintain temporal coherence with large frequency separation at higher TRTs. This observation suggests that with high TRT, where the rate between onsets of consecutive tones is relatively slow, the listeners have the time and capacity to direct their attention to a particular percept. Our idea to this attentional mechanism is through the random activity of the global inhibition that causes synchrony or desynchrony between blocks of oscillators in a random manner. The global inhibitor is especially effective for this situation because it can exert control to the entire network. In fact, Wang [140] claimed that the global inhibitor may indeed be regarded as an attentional control unit. There are structural and functional similarities between the global inhibitor with the thalamus, since the thalamus both sends and receive input from almost the entire cortex. Crick [38, 39] has suggested that parts of the thalamus may be critical in exerting attentional control. With our existing model, the global inhibitor naturally becomes the structure to take account of this phenomenon. The random activity of the global inhibition depends on TRT as:

$$r_z = random\left(0\,,\ \frac{U_r}{1+\exp(-\kappa_r(t-\theta_r))}\right) \tag{4.5}$$

Again, we utilize the sigmoid function to define the random activity, because the sigmoid function is bounded from above and below, which is only natural with the amount of randomness a system may have. U_r is the parameter for the upper bound on activity, and we default the lower bound to 0 as it is reasonable to assume no random inhibition in human perception in certain situations, such as when there arguably is no perceptible stimulus present. κ_r and θ_r denote the parameters for the sigmoid function. The *random* function simulates the uniformly random attentional shift that occurs in a range from 0 to the value specified by the sigmoidal function depending on TRT, which is denoted by t. The finals values we used were: $U_r = 0.27$, $\kappa_r = 0.03$, $\theta_r = 166$.

Our implementation for this attentional mechanism may be simple minded compared to some dedicated models relating to attention (e.g. Wrigley [147]). Nevertheless because

of the flexibility in the core architecture of LEGION and its usage of a global inhibitor, we believe that this basic analysis of inevitably a much more complicated process still has a lot of room for further extensions, and our computations provide a foundation for further analyses in this area.

D. Overall coupling

After defining the oscillator units, the lateral connections, and the global inhibition, we now discuss the overall coupling of the oscillator network to an oscillator *i*, the term S_i from Equation (4.1). In Wang [135], the connection weights in SegNet were dynamically normalized utilizing methods developed earlier. The purpose was to ensure that each oscillator has equal connection weights from its neighbors. While we do not normalize weights based on that purpose, we do normalization of the connection strength depending on the number of oscillators in the active phase relative to the total number of oscillators. The reason is to keep the connection strengths bounded from above and to steadily grow the overall connection strength of the network as more units become activated. We define the overall coupling as follows:

$$S_i = \frac{\sum_j J_{ij} S_\infty(x_j, \theta_x)}{n_z} \left(1 + \frac{n_z}{N_i}\alpha\right) - (1 + r_z) W_z z \qquad (4.6)$$

Here $S_\infty(x, \theta)$ is 1 if $x > \theta$ and 0 otherwise. z is as defined in Equation (4.4), and r_z is as defined in Equation (4.5). W_z is the weight for the global inhibition, and is set to .96. N_i denotes the number of oscillators that are enabled in row *i*. α is a constant that is used as a normalizing constant, and it is set to 0.2 in this study. Finally, n_z is the number of oscillators whose *x* activities are greater than the activity threshold, similar to the measurement from Equation (4.4). This formulation allows the connection strength to increase as more units are activated, which increases the possibility for units at different frequencies from the leader that have not been recruited to be recruited into the same stream. Note that the formulation here is consistent with the dynamics of LEGION,

because while the system needs to possess great precision over which units are activated, Linsay and Wang [82] stated that the coupling term changes at each jumping instant. Thus, at each jumping instant the system can recalculate the data it possesses about the state of the network as a whole. The algorithm implemented here ensures that the oscillator unit will not jump up to the active phase without S_i being a positive value.

E. Input, time representation, and implementation issues

Time has been recognized as an important property in auditory segregation by Wang [135] (see also [137]). As stated earlier in this section, the input is represented as a shifting binary matrix with each input unit of equal size in the time and frequency dimension. The input is then presented to the network, in which the delay lines in the network provide a form of short-term memory that has a recent history of external stimulation. In order to relate to real time, we assume the difference between two neighboring isofrequency oscillators is 10 ms, and thus that is the way the input is divided in time. This is consistent with many spectrogram representations where the input waveform is divided into time frames of 10 or 20 ms in the T-F domain. The number of oscillators in a row needs to be large enough so that the sequential tones with slow patterns can be presented simultaneously in the matrix. Looking at Figure 4.2, the slowest pattern in the data has a TRT of around 200 ms, so at least 400 ms of delay across the matrix is necessary to represent a set of alternating tones. We use 600 ms in all the simulations of this study, same as Norris [101]. To be consistent with biological studies of synchronous oscillations where the pattern of firing has around a 40 Hz oscillation as we have previously discussed, we use a similar oscillation frequency for the units we define. Because we are using the algorithmic approximations in Section 4.3.3 and the granularity in time is not as fine, we make our oscillations 50 Hz, translating to a period of 20 ms. This means that an input divided into 10 ms frames would be updated exactly twice by the network during the loop of steps 1 through 4 in the algorithm described in Section 4.3.3. Therefore, each time through the loop, the appropriate oscillators will be stimulated depending on the input presentations. Note that the number of frequency bands and bandwidth are not explicitly addressed in this study, because specifying the

distance between tones of different frequencies is sufficient in implementing the system. However, future work will certainly formalize the frequency representation in detail, in order to achieve demonstrations of different phenomena at a larger scale.

4.5 COMPUTATIONAL SIMULATIONS AND EVALUATIONS

4.5.1 OUTPUT MEASUREMENT

Measuring the effects of auditory streaming in humans can be tricky, in terms of what exactly it means to perceive one coherent stream or segregated streams. In behavioral experiments, one can only assess the streaming effects through a listener's verbal report ([18], p. 53). Besides asking the listeners whether the sounds are heard as two separate streams or only one, other measures such as loss of order and changes in perceived rhythm can be useful in determining the loss of temporal coherence. The idea is that one can usually hear and report the order of the tones in a single stream of sounds, while they cannot identify the order of tones when they are segregated into two streams ([18], p. 57). Physiological methods such as measuring event-related potentials (ERPs) mentioned earlier in the chapter have also been used to measure streaming [1]. When a sequence of sounds are heard as two segregated streams, the tones heard as one stream will usually have stronger ERP than the tones for the other stream. With tones forming one coherent stream, they tend to produce ERPs of similar strength. The observations for humans we have discussed in this paragraph can be carried over as measurements for the computational model as well.

As LEGION does not have an explicit output, the behavior of the network is entirely measured by the synchrony and desynchrony of the oscillatory units. While visualizing the evolution of synchrony by utilizing an animation as shown in Wang [135] is perhaps the most effective way in measuring the performance of a network, a more quantitative measurement is necessary in evaluating whether the system successfully simulates the psychophysical results shown in Figure 4.2. In the current study we first assume that each presentation input contains only representations for two alternating tones of

different frequencies. Looking at the algorithm described in Section 4.3.3, we can assume that as long as external input is continuously provided, the system loops indefinitely between steps 1 to 4. The x value of an oscillator when it is active is predictable, especially because the current study is based on the approximating algorithms. During the execution of the loop each time, we can record whether all of the stimulated oscillators in the entire matrix became active, in the case of temporal coherence, or whether all of the stimulated oscillators in an entire row representing a frequency band became active, in the case of segregated streaming. In part B of this section, we have discussed that in all of our examples the connection strength in time is relatively independent of the TRT, and therefore should remain strong at each condition. In fact, based on experimental studies, with inputs where the TRT is at most 200 ms and a representational matrix of 600 ms, the connection weight in time should remain strong regardless of change in TRT [131], thus we do not account for the situation where each tone takes on its own phase. Therefore no other situations should occur during the simulation that implies retuning of the parameters for the system. By varying TRT and frequency separation of the tones, we run the simulation for a finite number of times through the loop provided with constant input of the same tone sequences, we can enumerate the total number of times that the system is in the temporal coherence state compared to the segregated streaming state. If the system is in temporal coherent state for at least θ percentage of the time, we say the system is temporal coherent for that specific TRT and frequency separation. Likewise, if the system is in segregated streaming state for at least θ percentage of the time, we say the system is segregated for that TRT and frequency separation. Otherwise, we say the system is in the ambiguous state. θ is set to be 95 percent in our stimulations. In addition, in order to let the system stabilize first, we run the algorithm in Section 4.3.3 for about 10 cycles before starting to record the results.

4.5.2 SIMULATION RESULTS

We simulate the system using the same parameters reported in Section 4.4.3, and vary frequency difference and TRT between simulations. All combinations of frequency

difference ratio of 1.1 to 4 in steps of .02 and TRT of 50 to 200 ms in steps of 50 ms are simulated. The tone duration is fixed at 40 ms, which means each tone can influence 4 oscillators. The simulation is run for 100 loops (~2 seconds in real time) in the algorithm and the decision of temporal coherence, segregated, or ambiguous is calculated for each simulation. Then, simulated fission and temporal coherence boundaries are calculated by interpolating synchrony values between points simulated in the space formed by TRT and frequency ratio. The simulated result can be seen in Figure 4.3.

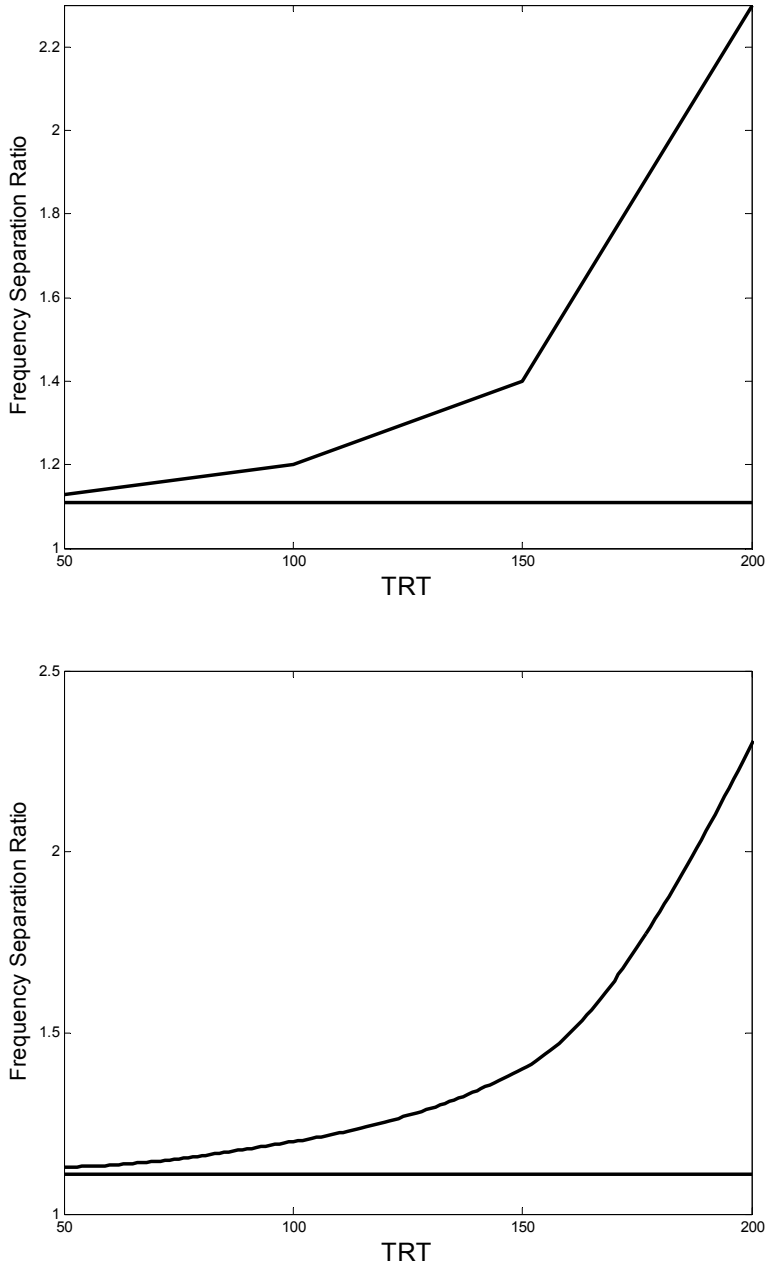


Figure 4.3: The simulated results from the proposed model. The boundaries for TCB and FB are simulated at four different TRT's of 50, 100, 150, and 200 ms. (**Top**) A linear interpolation of the resulting frequency separation ratio. (**Bottom**) A Spline interpolation created by Matlab. Thus the top curve represents the proposed model's TCB and the bottom line represents the proposed model's FB.

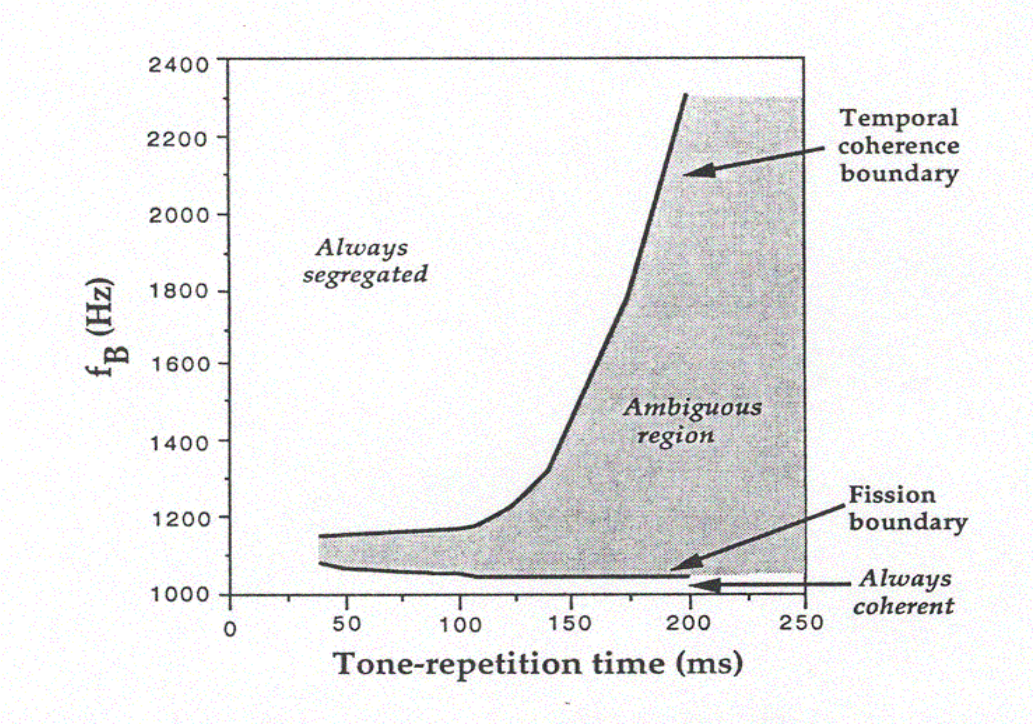


Figure 4.4: This figure is extracted directly from McAdams and Bregman [87]. These are their experimental results on the TCB and FB boundaries for alternating tones of 40 ms durations. The reason this figure is used as comparison is because the frequency separation scale used in their experiment is more similar to our proposed model's scale, where the differences in frequency separation in our case are expressed as frequency ratio of one tone over another.

The result is compared to the psychophysical results obtained from McAdams and Bregman [87] shown in Figure 4.4, because their experimental conditions and representations of the tones were more similar to our model than the original set of conditions used by Van Noorden [131]. It is clear that our model can simulate the boundaries quite accurately. In fact, as Van Noorden found out that individual results can vary in a wide range, the parameters of our model can easily be adjusted to fit to any reasonable individual differences.

4.5.3 COMPARISONS WITH OTHER IMPLEMENTATIONS

Wang [135] was the first to extend the LEGION architecture for auditory stream segregation, called the *segregation network* (SegNet). SegNet simulates alternating-tone sequences in a few conditions, such as small frequency difference, large frequency difference with small TRT, and large frequency difference with large TRT. However, because SegNet's connections are based on synchrony by time-frequency proximity alone, it cannot generate the three different boundaries shown on Figure 4.2. In particular, the major limitation is that using only local connectivity inherent in LEGION cannot account for the nontrivial dependence of TCB and FB on TRT. To be sure, Norris [101] claimed to have replicated the phenomenon without major changes to SegNet. Nevertheless, based on our understanding and conception of this representation, we cannot explain this phenomenon completely with the core architecture alone. Therefore, we have made significant extensions, not on the core architecture of LEGION, but on the structure of the network connections, in order to rigorously simulate the perceptual phenomenon observed through previous behavioral experiments relating to alternating-tone sequences [87, 93, 131]. Specifically, we allow the strength of local connections to change dynamically depending on the TRT, and utilize a different dynamic normalization approach than that in [135] where the local connection at the oscillators' jumping instants are normalized in order to take account of the activity of the oscillators. Furthermore, we specifically account for the ambiguous region by using a simple model of attentional control exhibited through the global inhibitor. The resulting implementation as shown in

Section 4.5.2 can faithfully simulate the psychophysical boundaries and has the flexibility to account for experimental differences as well.

Beauvois and Meddis (1996) have also successfully modeled the experimental results using a different architecture, but their model is entirely based on the peripheral auditory system, and does not seem to leave room to explain the relationship between clear effects of selective attention with this phenomenon [33], or any other ASA principles that require central processing in the brain.

In addition to achieving more accurate simulation results than SegNet, our study also incorporated some structural and conceptual differences to make the implementation more quantitative and simpler. For example, Wang [135] recognized the importance of time as a property in auditory stream segregation, but its representation was not explicitly defined in SegNet due to the approximating nature of the extracted algorithms. We have successfully incorporated the time factor into the approximations in Part E of Section 4.4.3, and ensured the consistency of the oscillation frequency relative to the system architecture.

Another noteworthy difference from SegNet is that Wang's study actually demonstrated the case where each tone takes on its own phase because they are so separated in the time domain. While that is certainly consistent with the time-frequency proximity principle, we have noted several times that in all of the experiments we consider here the connection strength in time remains relatively constant and strong regardless of TRT. Therefore it is reasonable for us to keep the isofrequency tones synchronized across the entire time domain.

Finally, SegNet's global inhibitor is defined as a pair of oscillatory units described by differential equations where each unit is turned on at different times according to the activity of the oscillatory units [135]. In the original model of LEGION, there was only a single unit of global inhibition that exerts effect as soon as one single unit in the network is in the active stage. In our model, we are able to do away with the second inhibitor, and

retain only the one global inhibition unit for simplicity and consistency with biological mechanisms.

4.6 GENERAL DISCUSSION

In this chapter, we have shown how a set of behavioral principles of ASA and evidence from physiological experiments can form theories of representation and algorithms for a computational approach to achieve ASA, the second and third level of analysis in Marr's framework. The approach of using a two-dimensional representation of time and frequency, knowledge about neural coding based on temporal correlation, and studies in experimental results of gamma oscillations, have led us to explore the representation of two-dimensional oscillatory correlation as the key representation and mechanism for auditory stream segregation.

With the representation and basic set of algorithmic requirements in place, a computational implementation for the segregation of alternating tone sequences has been proposed in this chapter. The model is based upon the LEGION's inherent ability to synchronize neighboring oscillators and desynchronize different blocks of oscillators. The model is further refined to take advantage of SegNet's time-frequency representation and its ability to segregate tones using dynamic connections covered by a Gaussian function. In addition, the model implements mechanisms to account for the dependence of TRT on the perception of alternating tone sequences, as well as the varying ranges of frequency separation in the ambiguous region. The shifting of perception between temporal coherence and segregated streaming at the varying frequency ranges has been attributed to the listener's attentional mechanisms, and we have utilized the global inhibitor to act as such an attentional control. Furthermore, this study formalizes the measurements of streaming to produce the quantitative results and also carefully considers the relationship of real-time versus simulation-time, which have all been ambiguous in SegNet.

Given these results, we must emphasize that this is only a beginning study in the attempt to model the streaming of alternating tones. We now outline some of the important issues that we have not yet considered and map a guideline on how to proceed from here. First, as mentioned before, all of our simulations have been done based on the approximating algorithms extracted from the analysis of a set of differential equations in Terman and Wang [127] (see also [135]). We have not matched all of the extensions we made to LEGION to the original sets of differential equations, but instead extended the model to directly correspond with the approximating algorithms. While all of the extensions and modifications we have made are consistent with the dynamic analysis and behavior of LEGION, such a relationship must be established in the future to strengthen the biological plausibility of our model. Second, we have not considered rhythmic information and also have not varied the tone durations of the alternating tone sequences, both of which have been shown to affect the perception of streaming. However, we believe that the robustness of our extended model will be able to handle the change in rhythm and tone durations. Third, we also have not considered a common issue related to streaming, called Stream Bias Adaptation (SBA) [101]. SBA is the build up and decay of streaming; it is known that streams do not form immediately after a long silence, but takes seconds to form [11, 16, 18]. We plan to handle this issue as we make the dependence on TRT a more dynamical process, which should be relatively straightforward based on the inherent dynamic connection weights in our system. Fourth, we have not tested the scalability of our model with more enabled frequency channels and different patterns of tone sequences. In particular, some demonstrations shown in Wang [135] such as sequential capturing and frequency modulation that require a more complicated input space have not been tested. Ultimately the model should be scalable to demonstrate the correct behaviors with speech signals processed using peripheral analysis and quantized in the time-frequency domain as we have done in Chapter 3 when we introduced the ideal-mask paradigm. Finally, and very importantly, we do not consider attention explicitly in the current model. So far, we have arbitrarily dictated attention to be a random process controlled by the global inhibitor and dependent upon the TRT. While attentional mechanisms may very well be controlled by the global inhibitor and depend on TRT, the specific ways in which it is implemented need to be further analyzed.

As mentioned before, studies such as Wrigley [147] (see also [148]) might prove useful in incorporating a more biologically plausible attentional model into our system.

Finally, we want to remark that the similarity in the shape of the human and modeled boundaries has strengthened the biological plausibility of LEGION and our modifications and extensions applied to perception of alternating tones. With oscillatory correlation as a dominant theory of perceptual feature binding [62], and our promising results in applying it to auditory streaming, it is certainly worthwhile to consider further extensions to this architecture in order to achieve an accurate representation of our auditory system.

With everything above in mind, we want to take a step back from our modeling and take a higher level look at the computational approach to auditory scene analysis and auditory streaming. From the behavioral studies documented in this thesis, we have seen that even some seemingly simple phenomenon such as the streaming of alternating tones can lead to very complicated percepts that involve both primitive scene analysis and schema-based integration, including active selective attention. Thus we must not underestimate any problem in auditory scene analysis and should adhere to a structured plan of analysis to tackle the problems computationally if we want to eventually understand the underlying mechanisms. Our adherence to Marr's three levels of analysis gives us the opportunity to independently build knowledge at the different approaches we have explored, and also a framework to integrate the findings at the different levels into a complete process. Therefore, the most important aspect of a computational model of perception is its ability to be extended and integrated with other existing systems while satisfying all of the constraints given by behavioral and physiological studies as well as conforming to the representational and algorithmic requirements.

CHAPTER 5
CONCLUSIONS

In this thesis, we have touched upon a wide range of issues relating to auditory perception. We have lain out on the table the complexity of processes relating to any modality of perception which we often take for granted as an active participant in the environment. We have also outlined the main approaches that scientists use to study perception in general, which originate from behavioral studies that directly link the stimulus to the final percept, and further analyzed with physiological studies that link the stimulus with the underlying mechanisms inside our brain that we cannot simply observe with our bare senses. However, it has been difficult to connect the results from one approach with another, in that how the sensory organs take the stimulus and send to the brain and form the final percept is difficult to understand. A natural approach that has emerged from these challenges is from computational scientists who aim to structurally and algorithmically describe exactly how humans perceive. While many have received their motivation from building engineering applications that can take advantage of humans' robustness and capability in dealing with tasks in a complex environment, many are also motivated by the need to faithfully adhere to the biological principles in order to understand the natural systems using computational tools. One such influential figure, David Marr [86], devised a framework that instructs us in how to proceed using the computational approach when faced with a complex biological process. His framework divides the process into three levels of analysis, each at it own level of abstraction, from the philosophical theories and goals of the problem, through the representational and algorithmic requirements, all the way down to the physical realization of the process. As we have demonstrated in this thesis, at different levels of the analysis we can draw many

fitting parallels from the behavioral and physiological approaches that keep a complex problem well organized.

Because this thesis has focused on auditory perception, in particular auditory scene analysis, we have lain out the problem using Marr's framework, and then have gone on to explore some results researchers have found using different approaches, at each level of analysis. In addition, we have focused on several areas and made some original contributions to the study of study. The first major contribution is the set of psychophysical studies conducted in Chapter 3 that help to further the understanding of speech-on-speech energetic and informational masking, the relationship between masking and auditory scene analysis, and provide support for the ideal binary mask as a computational goal of auditory scene analysis. While experimental and theoretical studies in auditory scene analysis conducted by Bregman [18] and other researchers have established a solid foundation relating the characteristics of the stimuli and our "behaviors" in response to those stimuli, they did not account for the fact that many different effects coexist in sound stimuli. In order to further study the issues rigorously using the behavioral approach, we must be able to isolate the effects and understand the relative contribution of each kind of effect in a phenomenon. One prominent example is the coexistence of energetic and informational masking in perceiving sound mixtures, and we have utilized ideal time-frequency binary masks to isolate the effects of energetic masking on mutitalker speech perception by retaining the T-F regions dominated by the target signal and eliminated those regions dominated by the interfering signal. Through a set of carefully designed experiments, we have shown the dominance of informational masking in multitalker speech mixtures. We have also demonstrated valuable applications for isolating energetic masking by determining the contribution of each kind of masking from differences in voice characteristics and number of interfering voices.

The second major contribution is the computational implementation of alternating tones described in Chapter 4. The model can be seen as the culmination of behavioral and physiological approaches at the theoretical and representational levels of analysis leading to its realization. The model's computational goals are derived from the organizational

principles of auditory scene analysis, such as the principle of time-frequency proximity, as well as schema-based integration requiring selective attention, such as the ambiguous switch of percept between temporal coherence and segregated streaming. The model utilizes physiological evidence to arrive at a choice of representation, such as the time-frequency representation based on the oscillatory correlation theory, derived as a special form from the general neurocomputational theory of temporal correlation [133, 134]. The model makes extensions to previous implementations in effort to simulate more computational goals from the behavioral results, as well as adhering to the biological plausibility at different levels of analyses. Our model has thus successfully implemented the major behavioral results relating to the perception of alternating-tone sequences, most notably the computation of decision boundaries for the TCB and FB regions that previous oscillatory correlation implementations cannot.

The contributions of this thesis do not only help to further the understanding of ASA in order to satisfy the scholars' curious minds. Many practical applications could benefit from the research in this subject. As we have discussed in Section 2.5, many CASA researchers that aimed to study ASA solely for the engineering applications have recently realized the need to understand the underlying mechanisms of the human auditory system because of limited success utilizing traditional engineering methods. Based on that reason alone, the reader should find the framework presented in this thesis useful in generating ideas to take advantage of the auditory mechanisms when constructing an engineering application. However, the ideas proposed in this thesis can be of use in subtler ways. Engineering applications are built to improve people's lives. So whether one is building a computerized hearing aid or an automatic audio information retriever, the bottom line is to improve the user experience. We believe that there is no better way to improve the experience other than understanding the perceptual experience itself, by evaluating the percepts behaviorally, investigating the mechanisms physiologically, and in turn building computational tools that faithfully simulate the entire process, providing the user with the most natural experience with optimal performance.

In conclusion, audition is a complex modality in our sensory system, and in a fundamental way auditory scene analysis is at the heart of auditory perception. As we have just touched the surface of a variety of issues, the studies and contributions in this thesis are by no means complete. With each approach and at each level of analysis, a variety of future directions have been outlined and discussed in the previous chapters. After reading this exploratory overview, one may feel overwhelmed by the amount of further research needed to really understand the underlying mechanisms of our auditory system. Fortunately, Marr's framework provides guidelines for how to organize the analysis for the problems, hence facilitating constructions for the solutions. The computational approach to ASA has slowly emerged, and it seems especially promising in integrating results from the behavioral and physiological approaches. However, we must continue to improve all of the approaches discussed and achieve a proper balance. At the behavioral level, while we now have numerous data describing our responses to the stimuli, in a complex environment different effects often coexist and we must aim to quantify the relative contribution of each effect in order to have the most comprehensive results. At the physiological level, while it has been difficult to come up with an all-encompassing representation on how exactly the brain processes the stimuli to form so many interesting percepts, we must continue to uncover neurobiological evidence that can provide invaluable constraints on other levels. Finally, at the implementation level, the computational models should maintain consistency with the results at other levels and also use efficient computations. As mentioned before, the computational approach to auditory scene analysis is still in its infancy. By continuing to build biologically plausible and computationally realizable models, the behavioral, physiological, and computational approaches will continue to advance together to help us understand the underlying process of auditory scene analysis and perception in general.

BIBLIOGRAPHY

[1] C. Alain and D. Woods, "Signal clustering modulates cortical activity in humans". Perception and Psychophysics. 56(5): pp. 501-516. 1994.

[2] T. Arbogast, C. Mason, and G. Kidd, "The effect of spatial separation on information and energetic masking of speech". Journal of the Acoustical Society of America. 112: pp. 2086-2098. 2002.

[3] D. Arnold, "Fitting a Logistic Curve to Data": College of the Redwoods. 2002.

[4] B. Arons, "A review of the cocktail party effect". MIT: MIT Media Lab.

[5] P. Assmann and Q. Summerfield, "Perceptual segregation of concurrent vowels". Journal of the Acoustical Society of America. 100(2): pp. 1141-1152. 1987.

[6] P. Assmann and Q. Summerfield, "The contribution of waveform interactions to the perception of concurrent vowels". Journal of the Acoustical Society of America. 95(1): pp. 471-484. 1994.

[7] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions", in *Speech Processing in the Auditory System*, P. Greenberg, Ainsworth, and Fay, Editors, Springer-Verlag: New York, NY, 2004.

[8] B. Baird, "A cortical model of cognitive 40 hz attentional streams, rhythmic expectation, and auditory stream segregation". Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society: pp. 365-371. 1997.

[9] D.S. Barth and K.D. MacDonald, "Thalamic modulation of high-frequency oscillating potentials in auditory cortex". Nature. 383: pp. 78-81. 1996.

[10] M. Beauvois and R. Meddis, "Time decay of auditory stream biasing". Perception and Psychophysics. 59(1): pp. 81-86. 1997.

[11] M.W. Beauvois and R. Meddis, "Computer simulation of auditory stream segregation in alternating-tone sequences". Journal of the Acoustical Society of America. 99: pp. 2270-2280. 1996.

[12] G.v. Bekesy, *Experiments in hearing*, New York: McGraw-Hill. 1960.

[13] J. Bench and J. Bamford, *Speech hearing tests and the spoken language of hearing-impaired children*, London: Academic Press. 1979.

[14] K. Binns and T. Salt, "The importance of NMDA receptors for multimodal integration in the deep layers of the cat superior colliculus". Journal of Neurophysiology. 75: pp. 920-930. 1995.

[15] R. Bolia, et al., "A speech corpus for multitalker communications research". Journal of the Acoustical Society of America. 107: pp. 1065-1066. 2000.

[16] A. Bregman, "Auditory streaming is cumulative". Journal of Experimental Psychology: Human Perception and Performance. 4(3): pp. 380-387. 1978.

[17] A. Bregman and G. Dannenbring, "The effect of continuity of auditory stream segregation". Perception and Psychophysics. 13(308-312). 1973.

[18] A.S. Bregman, *Auditory scene analysis*, Cambridge MA: MIT Press. 1990.

[19] A.S. Bregman and J. Campbell, "Primary auditory stream segregation and perception of order in rapid sequences of tones". Journal of Experimental Psychology. 89: pp. 244-249. 1971.

[20] A.S. Bregman and S. Pinker, "Auditory streaming and the building of timbre". Canadian Journal of Psychology. 32: pp. 19-31. 1978.

[21] D. Broadbent and P. Ladefoged, "On the fusion of sounds reaching different sense organs". Journal of the Acoustical Society of America. 29: pp. 708-710. 1957.

[22] A. Bronkhorst and R. Plomp, "Effects of multiple speechlike maskers on binaural speech recognitions in normal and impaired listening". Journal of the Acoustical Society of America. 92: pp. 3132-3139. 1992.

[23] G.J. Brown and M. Cooke, "Computational auditory scene analysis". Computer Speech and Language. 8: pp. 297-336. 1994.

[24] G.J. Brown and M.P. Cooke, "Temporal synchronisation in a neural oscillator model of primitive auditory stream segregation", in *Computational auditory scene analysis*, D. Rosenthal and H. Okuno, Editors, Lawrence Erlbaum: Mahwah NJ. pp. 87-103, 1998.

[25] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers". Journal of the Acoustical Society of America. 109: pp. 1101-1109. 2001.

[26] D. Brungart and B. Simpson, "Within-channel and across-channel interference in the cocktail-party listening task". Journal of the Acoustical Society of America. 112: pp. 2985-2995. 2002.

[27] D. Brungart and B. Simpson, "Within-ear and across-ear interference in a dochotic cocktail party listening task: effects of masker uncertainty". Journal of the Acoustical Society of America. 115: pp. 301-310. 2004.

[28] D. Brungart, et al., "Informational and energetic masking effects in the perception of multiple simultaneous talkers". Journal of the Acoustical Society of America. 110: pp. 2527-2538. 2001.

[29] J. Buhmann, "Oscillations and low firing rates in associative memory neural networks". Physics Review A. 40: pp. 4145-4148. 1989.

[30] E. Buss, J. Hall, and J. Grose, "Spectral integration of synchronous and asynchronous cues to consonant identification". Journal of the Acoustical Society of America. 115: pp. 2278-2285. 2004.

[31] G. Buzsáki, et al., eds. *Temporal coding in the brain*. c, Springer-Verlag: Berlin, 1994.

[32] R. Cahart and T. Tillman, "Perceptual Masking in Multiple Sound Backgrounds". Journal of the Acoustical Society of America. 45: pp. 694-703. 1969.

[33] R.P. Carlyon, et al., "Effects of attention and unilateral neglect on auditory stream segregation". Journal of Experimental Psychology: Human Perception and Performance. 27: pp. 115-127. 2001.

[34] F. Cavallini, "Fitting a Logistic Curve to Data". College Mathematics Journal. 24(3): pp. 247-253. 1993.

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

[36] M. Cooke, "Making sense of everyday speech: a glimpsing account", in *Speech separation by humans and machines*, P. Divenyi, Editor, Kluwer Academic: Norwell, MA. pp. 305-314, 2004.

[37] M. Cooke, et al., "Robust automatic speech recognition with missing and unreliable acoustic data". Speech Communication. 34: pp. 267-285. 2001.

[38] F. Crick, "Function of the thalamic reticular complex: The searchlight hypothesis". Proceedings of the National Academy of Sciences of USA. 81: pp. 4586-4590. 1984.

[39] F. Crick, *The astonishing hypothesis*, New York: Scribner. 1994.

[40] J. Culling and C.J. Darwin, "Perception and computational separation of simultaneous vowels: cues arising from low frequency beating". Journal of the Acoustical Society of America. 95: pp. 1559-1569. 1994.

[41] C.J. Darwin. "On the dynamic use of prosody in speech perception". in *Structure and Process in Speech Perception: Proceedings of the Symposium on Dynamic Aspects of Speech Perception*. I. P. O., Eindhoven, The Netherlands, 1975.

[42] H. Dinse and C. Schreiner, "Do primary sensory areas play homologous roles in different sensory modalities?" in *Cortical Areas: Unity and Diversity: Conceptual Advances in Brain Research*, A. AMiller, Editor, Taylor and Francis: London, New York. pp. 273-310, 2002.

[43] D. Dirks and D. Bower, "Masking effects of speech competing messages". Journal of Speech and Hearing Research. 12: pp. 229-245. 1969.

[44] T. Doll and T. Hanna, "Directional cueing effects in auditory recognition", in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. Gilkey and T. Anderson, Editors, Erlbaum: Hillsdale, NJ, 1997.

[45] M. Dorman, P. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs." Journal of the Acoustical Society of America. 102: pp. 2403-2411. 1997.

[46] W.J. Dowling, "The perception of interleaved melodies". Cognitive Psychology. 5: pp. 322-337. 1973.

[47] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception". Journal of the Acoustical Society of America. 95: pp. 2670. 1994.

[48] A. Duquesnoy, "Effect of a single interfering noise or speech source on the binaural sentence intelligibility of aged persons". Journal of the Acoustical Society of America. 74: pp. 739-743. 1983.

[49] R. Eckhorn, et al., "Coherent oscillations: A mechanism of feature linking in the visual cortex". Biological Cybernetics. 60: pp. 121-130. 1988.

[50] J. Egan and E. Cartererette, "Factors affecting multi-channel listening". Journal of the Acoustical Society of America. 26: pp. 774-782. 1954.

[51] J. Eggermont, "Firing rate and firing synchrony distinguish dynamic from steady state sound". Neuroreport. 8(12): pp. 2709-2713. 1997.

[52] D.P.W. Ellis, "Prediction-driven computational auditory scene analysis". Ph.D. Dissertation in MIT Department of Electrical Engineering and Computer Science, 1996.

[53] U. Eysel, "Latency as additional dimension for object encoding in sensory systems: a commentary on the paper by T.P.L. Roberts and D. Poeppel". Neuroreport. 7: pp. 1113. 1996.

[54] A. Feng and R. Ratnam, "Neural basis of hearing in real-world situations." Annual Review of Psychology. 51: pp. 699-725. 2000.

[55] J. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing". Journal of the Acoustical Society of America. 88: pp. 1725-1736. 1990.

[56] R. FitzHugh, "Impulses and physiological states in models of nerve membrane". Biophysical Journal. 1: pp. 445-466. 1961.

[57] W.J. Freeman, "Spatial properties of an EEG event in the olfactory bulb and cortex". Electroencephalographical and Clinical Neurophysiology. 44: pp. 586-605. 1978.

[58] R. Freyman, et al., "The role of perceived spatial separation in the unmasking of speech". Journal of the Acoustical Society of America. 106: pp. 3578-3587. 1999.

[59] J. Fritz, et al., "Rapid task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex". Nature Neuroscience. 6(11): pp. 1216-1223. 2003.

[60] R. Galambos, S. Makeig, and P.J. Talmachoff, "A 40-Hz auditory potential recorded from the human scalp". Proceedings of the National Academy of Sciences of USA. 78: pp. 2643-2647. 1981.

[61] E.B. Goldstein, *Sensation & Perception*. 5th ed, Pacific Grove, CA: Brooks/Cole Publishing Company. 661. 1999.

[62] C.M. Gray, "The temporal correlation hypothesis of visual feature integration: still alive and well". Neuron. 24: pp. 31-47. 1999.

[63] C.M. Gray, et al., "Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties". Nature. 338: pp. 334-337. 1989.

[64] W. Hartmann and D. Johnson, "Stream segregation and peripheral channeling". Music Perception. 9(2): pp. 155-184. 1991.

[65] M. Hawley, R. Litovsky, and J. Culling, "The benefit of binaural hearing in a cocktail party: Effects of location and type of interferer". Journal of the Acoustical Society of America. 115: pp. 833-843. 2004.

[66] H. Helmholtz, *On the sensation of tone*. Second English ed, New York: Dover Publishers. 1863.

[67] A.L. Hodgkin and A.F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve". Journal of Physiology (London). 117: pp. 500-544. 1952.

[68] A. House, et al., "Articulation testing methods: Consonantal differentiation with a closed response set". Journal of the Acoustical Society of America. 37: pp. 158-166. 1965.

[69] P. Howard-Jones and S. Rosen, "Uncomodulated glimpsing in "checkerboard" noise". Journal of the Acoustical Society of America. 93: pp. 2915-2922. 1993.

[70] G. Hu and D.L. Wang. "Speech segregation based on pitch tracking and amplitude modulation". in *WASPAA 2001*, 2001.

[71] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation". IEEE Transactions on Neural Networks. 15: pp. 1135-1150. 2004.

[72] S. Hygger, et al., "Normal-hearing and hearing-impaired subject's ability to just follow conversation in competing speech, reversed speech, and noise backgrounds". Journal of Speech and Hearing Research. 35: pp. 208-215. 1992.

[73] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, New York: Wiley. 2001.

[74] J. Kaas and T. Hackett, "Auditory processing in primate cerebal cortex". Current opinion in Neurobiology. 9: pp. 164-170. 1999.

[75] K. Kasturi, et al., "The intelligibility of speech with "holes" in the spectrum". Journal of the Acoustical Society of America. 112(1102-1111). 2002.

[76] G. Kidd, et al., "Reducing information masking by sound segregation". Journal of the Acoustical Society of America. 95: pp. 3475-3480. 1994.

[77] G. Kidd, et al., "Release from informational masking due to the spatial separation of sources in the identification of nonspeech auditory patterns". Journal of the Acoustical Society of America. 104: pp. 422-431. 1998.

[78] K. Koffka, *Principles of Gestalt psychology*, New York: Harcourt. 1935.

[79] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach". IEEE Signal Processing Magazine. 13: pp. 67-94. 1996.

[80] T.-W. Lee, *Independent component analysis: theory and applications*, Boston: Kluwer Academic. 1998.

[81] J. Lim, ed. *Speech enhancement*. c, Prentice Hall: Englewood Cliffs NJ, 1983.

[82] P.S. Linsay and D.L. Wang, "Fast numerical integration of relaxation oscillator networks based on singular limit solutions". IEEE Transactions on Neural Networks. 9: pp. 523-532. 1998.

[83] R. Lippman, "Accurate consonant perception without mid-frequency speech energy". IEEE Transactions on Speech and Audio Processing. 4: pp. 66-69. 1996.

[84] J.K. Love, ed. *Helen Keller in Scotland: a personal record written by herself*. c, Methuen & Co.: London, 1933.

[85] C. Madler and E. Pöppel, "Auditory evoked potentials indicate the loss of neuronal oscillations during general anesthesia". Naturwissenschaften. 74: pp. 42-43. 1987.

[86] D. Marr, *Vision*, New York: Freeman. 1982.

[87] S. McAdams and A. Bregma, "Hearing musical streams". Computer Music Journal. 3: pp. 26-43. 1979.

[88] S.L. McCabe and M.J. Denham, "A model of auditory streaming". Journal of the Acoustical Society of America. 101: pp. 1611-1621. 1997.

[89] W.S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity". Bulletin of Mathematical Biophysics. 5: pp. 115-133. 1943.

[90] D.K. Mellinger, "Event formation and separation in musical sound". Ph.D. Dissertation in Stanford University Department of Computer Science, 1992.

[91] G. Miller, "Sensitivity to changes in the intensity of white Gaussian noise and its relation to masking and loudness". Journal of the Acoustical Society of America. 191: pp. 609-619. 1947.

[92] G. Miller and J. Licklider, "The intelligibility of interrupted speech". Journal of the Acoustical Society of America. Supp. 1: pp. 20. 1950.

[93] G.A. Miller and G.A. Heise, "The trill threshold". Journal of the Acoustical Society of America. 22: pp. 637-638. 1950.

[94] B.C.J. Moore, *An introduction to the psychology of hearing*. 5th ed, San Diego, CA: Academic Press. 2003.

[95] T. Moore. "Voice communication jamming research". in *AGARD Conference Proceedings 331: Aural Communication in Aviation*. Neuilly-SurSeine, France, 1981.

[96] C. Morris and H. Lecar, "Voltage oscillations in the barnacle giant muscle fiber". Biophysical Journal. 35: pp. 193-213. 1981.

[97] D. Mountain, "Auditory periphery and cochlear nucleus", in *Handbook of Brain Theory and Neural Networks*, M. Arbib, Editor, MIT Press: Cambridge, MA. pp. 115-119, 1995.

[98] D. Mumford, "Thalamus", in *Handbook fo Brain Theory and Neural Networks*, M. Arbib, Editor, MIT Press: Cambridge. pp. 981-984, 1995.

[99] V.N. Murthy and E.E. Fetz, "Coherent 25- to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys". Proceedings of the National Academy of Sciences of USA. 89: pp. 5670-5674. 1992.

[100] W. Noble and S. Perret, "Hearing speech against spatially separate competing speech versus competing noise". Perception and Psychophysics. 64: pp. 1325-1336. 2002.

[101] M. Norris, "Assessment and extension of Wang's oscillator model of auditory stream segregation". Ph.D. Dissertation in School of Information Technology and Elecrical Engineering. The University of Queensland, Queensland. 2003.

[102] D. O'Shaughnessy, *Speech communications: Human and machine*. 2nd Ed. ed, Piscataway NJ: IEEE Press. 2000.

[103] A. Oxenham, "Forward Masking: Adaption or Integration". Journal of the Acoustical Society of America. 109: pp. 732-741. 2001.

[104] R.D. Patterson, et al., "SVOS final report, part B: Implementing a gammatone filterbank". MRC Applied Psychology Unit. 1988.

[105] J. Peissig and B. Kollmeier, "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners". Journal of the Acoustical Society of America. 35: pp. 1660-1670. 1997.

[106] J.O. Pickles, *An introduction to the physiology of hearing (2nd edition)*, London: Academic Press. 1988.

[107] I. Pollack, "Auditory Informational Masking". Journal of the Acoustical Society of America. Supp. 1: pp. 82. 1975.

[108] A.N. Popper and R.R. Fay, eds. *The mammalian auditory pathway: Neurophysiology*. c, Springer-Verlag: New York, 1992.

[109] R. Remez, et al., "Speech perception without traditional speech cues". Science. 212: pp. 947-950. 1981.

[110] I. Rock and S. Palmer, "The legacy of Gestalt psychology". Scientific American. 263: pp. 84-90. 1990.

[111] N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization". Journal of the Acoustical Society of America. 114: pp. 2236-2252. 2003.

[112] J.E. Rose, et al., "Phase locked response to low frequency tones in single auditory nerve fibers of the squirrel monkey". Journal of Neurophysiology. 30: pp. 769-793. 1967.

[113] S. Russell and P. Norvig, *Artificial intelligence: A modern approach*. 2nd ed, Upper Saddle River, NJ: Prentice Hall. 2003.

[114] T.B. Schillen and P. König, "Stimulus-dependent assembly formation of oscillatory responses: II. Desynchronization". Neural Computation. 3: pp. 155-166. 1991.

[115] S. Scott, et al., "A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception". Journal of the Acoustical Society of America. 115: pp. 813-821. 2004.

[116] W. Senn, I. Segev, and M. Tsodyks, "Reading neuronal synchrony with depressing synapses". Neural Computation. 10(815-819). 1998.

[117] R.V. Shannon, et al., "Speech recognition with primarily temporal cues". Science. 270: pp. 303-304. 1995.

[118] W. Singer and C.M. Gray, "Visual feature integration and the temporal correlation hypothesis". Annual Review of Neuroscience. 18: pp. 555-586. 1995.

[119] L.S. Smith, "Sound segmentation using onsets and offsets". Journal of New Music Research. 23: pp. 11-23. 1994.

[120] L.S. Smith and D.S. Fraser, "Robust sound onset detection using leaky integrate-and-fire neurons with depressing synapses". IEEE Transactions on Neural Networks. 15(5): pp. 1124-1134. 2004.

[121] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception". Nature. 416: pp. 87-90. 2002.

[122] M. Spiegel, "Speech masking: I. Simultaneous and nonsimultaneous masking within stop /d/ and flap /J/ closures". Journal of the Acoustical Society of America. 82(5): pp. 1592-1502. 1987.

[123] O. Sporns, et al., "Reentrant signaling among simulated neuronal groups leads to coherency in their oscillatory activity". Proceedings of the National Academy of Sciences of USA. 86: pp. 7265-7269. 1989.

[124] S. Srinivasan and D.L. Wang, "Schema based modeling of phonemic restoration". Proc. Eurospeech - 2003: pp. 2053-2056. 2003.

[125] S.S. Stevens, "The surprising simplicity of sensory metrics". American Psychologist, (17): pp. 29-39. 1962.

[126] N. Suga and J. Kanwal, "Echolocation: Creating computational maps", in *Handbook of Brain Theory and Neural Networks*, M. Arbib, Editor, MIT Press: Cambridge. pp. 344-348, 1995.

[127] D. Terman and D.L. Wang, "Global competition and local cooperation in a network of neural oscillators". Physica D. 81: pp. 148-176. 1995.

[128] A.M. Treisman, "Contextual cues in selective listening". Quarterly Journal of Experimental Psychology. 12: pp. 242-248. 1960.

[129] B. van der Pol, "On 'relaxation oscillations'". Philosophical Magazine. 2(11): pp. 978-992. 1926.

[130] L. van Noorden, "Minimum differences of level and frequency for perceptual fission of tone sequence ABAB*". Journal of the Acoustical Society of America. 61(4): pp. 1041-1045. 1977.

[131] L.P.A.S. van Noorden, "Temporal coherence in the perception of tone sequences". Ph.D. Dissertation in Eindhoven University of Technology, 1975.

[132] B. van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering". IEEE ASSP Magazine: pp. 4-24. 1988.

[133] C. von der Malsburg, "The correlation theory of brain function". Max-Planck-Institute for Biophysical Chemistry (Reprinted in *Models of neural networks* II, E. Domany, J.L. van Hemmen, and K. Schulten, eds., Berlin: Springer, 1994). 1981.

[134] C. von der Malsburg and W. Schneider, "A neural cocktail-party processor". Biological Cybernetics. 54: pp. 29-40. 1986.

[135] D.L. Wang, "Primitive auditory segregation based on oscillatory correlation". Cognitive Science. 20: pp. 409-456. 1996.

[136] D.L. Wang, "Relaxation oscillators and networks", in *Encyclopedia of electrical and electronic engineers*, J. Webster, Editor, Wiley: New York. pp. 396-405 (also available on the web at www.cse.ohio-state.edu/~dwang), 1999.

[137] D.L. Wang, "The time dimension for neural computation". OSU Department of Computer & Information Science: Columbus OH. 2002.

[138] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis", in *Speech Separation by Humans and Machines*, P. Divenyi, Editor, Kluwer Academic: Norwell, MA. pp. 181-197, 2004.

[139] D.L. Wang and G.J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation". IEEE Transactions on Neural Networks. 10: pp. 684-697. 1999.

[140] D.L. Wang and D. Terman, "Locally excitatory globally inhibitory oscillator networks". IEEE Transactions on Neural Networks. 6(1): pp. 283-286. 1995.

[141] D.L. Wang and D. Terman, "Image segmentation based on oscillatory correlation". Neural Computation. 9: pp. 805-836 (for errata see *Neural Comp.*, vol. 9, pp. 1623-1626, 1997). 1997.

[142] R.M. Warren, "Perceptual restoration of missing speech sounds". Science. 167: pp. 392-393. 1970.

[143] R.M. Warren, *Auditory perception: a new analysis and synthesis*, New York: Cambridge University Press. 1999.

[144] C. Watson, W. Kelly, and H. Wroton, "Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty". Journal of the Acoustical Society of America. 60: pp. 1176-1185. 1976.

[145] M. Weintraub, "A theory and computational model of auditory monaural sound separation". Ph.D. Dissertation in Electrical Engineering. Stanford University, 1985.

[146] E.G. Wever and C.W. Bray, "The perception of low tones and the resonance-volley theory". Journal of Psychology. 3(101-114). 1937.

[147] S. Wrigley, "A theory and computational model of auditory selective attention". Ph.D. Dissertation in Computer Science. University of Sheffield, Sheffield. 2002.

[148] S. Wrigley and G. Brown, "A computational model of auditory selective attention". IEEE Transactions on Neural Networks. 15(5): pp. 1151-1163. 2004.

[149] S.N. Wrigley and G.J. Brown, "A neural oscillator model of auditory attention". Lecture Notes in Computer Science. 2130: pp. 1163-1170. 2001.

[150] W.A. Yost, "The cocktail party problem: Forty years later", in *Binaural and spatial hearing in real and virtual environments*, R.H. Gilkey and T.R. Anderson, Editors, Lawrence Erlbaum: Mahwah, NJ. pp. 329-347, 1997.