ABSTRACT

Title of Document:	Auditory streaming: behavior, physiology, and modeling
	Ling Ma, Doctor of Philosophy, 2011
Directed By:	Professor Shihab A. Shamma, Department of Electrical and computer Engineering, Institute of Systems Research

Auditory streaming is a fundamental aspect of auditory perception. It refers to the ability to parse mixed acoustic events into meaningful streams where each stream is assumed to originate from a separate source. Despite wide interest and increasing scientific investigations over the last decade, the neural mechanisms underlying streaming still remain largely unknown. A simple example of this mystery concerns the streaming of simple tone sequences, and the general assumption that separation along the tonotopic axis is sufficient for stream segregation. However, this dissertation research casts doubt on the validity of this assumption. First, behavioral measures of auditory streaming in ferrets prove that they can be used as an animal model to study auditory streaming. Second, responses from neurons in the primary auditory cortex (A1) of ferrets show that spectral components that are well-separated in frequency produce comparably segregated responses along the tonotopic axis, no matter whether presented synchronously or consecutively, despite the substantial differences in their streaming percepts when measured psychoacoustically in humans. These results argue against the notion that tonotopic separation *per se* is a sufficient neural correlate of stream segregation. Thirdly, comparing responses during behavior to those during the passive condition, the temporal correlations of spiking activity between neurons belonging to the same stream display an

increased correlation, while responses among neurons belonging to different streams become less correlated. Rapid task-related plasticity of neural receptive fields shows a pattern that is consistent with the changes in correlation. Taken together these results indicate that temporal coherence is a plausible neural correlate of auditory streaming. Finally, inspired by the above biological findings, we propose a computational model of auditory scene analysis, which uses temporal coherence as the primary criterion for predicting stream formation. The promising results of this dissertation research significantly advance our understanding of auditory streaming and perception.

AUDITORY STREAMING: BEHAVIOR, PHYSIOLOGY, AND MODELING

By

Ling Ma

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:

Professor Shihab A. Shamma, Chair
Professor Carol Espy-Wilson
Associate Professor Timothy K. Horiuchi
Assistant Professor Adam Hsieh
Professor Peter Kofinas
Associate Professor Jonathan Z. Simon

Acknowledgements

I would like to thank my Ph.D advisor, Professor Shihab A. Shamma, for his most valuable guidance, support, and mentorship during the entire process of conducting this dissertation research. I also would like to thank my Ph.D. committee, Doctors Jonathan Z. Simon, Timothy K. Horiuchi, Peter Kofinas, Carol Espy-Wilson, and Adam Hsieh. Special thanks go to Dr. Jonathan Simon, for his advice and encouragement along the way and to Dr. Timothy Horiuchi, for his inspiring questions. Many thanks to Dr. Mounya Elhilali, Dr. Christophe Micheyl, and Dr. Andrew J. Oxenham, for their contributions to this research and coauthorship on the journal publications that constitute part of this dissertation.

I also thank members of Neural Systems Laboratory for their friendship and support. Special thanks go to Dr. Pingbo Yin and Dr. Jonathan Fritz, for teaching me how to do physiological experiments and animal behavior. To Stephen, Majid, Dan, Nima, Mai, Serin, Kevin, and everybody else: thanks for your company and support to help me get through this long journey.

Last, but not least, I thank my family: my parents, Baotong Ma and Yuxiang Chen, for unconditional support and encouragement to pursue my interests. My husband, Gang Zi, for his love and for his believing in me. My son, Andrew L. Zi, for making my life happier. It is impossible to have this dissertation done without any of you.

College Park, March 2011

Ling

Table of Contents

List of Figures

Chapter 1 **Introduction**

It seems effortless for us to listen to someone at a crowded cocktail party, or try to follow the violin line in a symphonic orchestra. However, it is still a mystery how our brain parses these complex acoustic scenes into individual auditory "objects" or "streams". An auditory stream refers to sound elements coming from an individual sound source and perceived by listeners as a coherent entity. Despite much research to understand auditory streaming in psychoacoustics studies, electroencephalography (EEG) and magnetoencephalography (MEG) studies, brain imaging, and single/multi- unit recordings, the neurophysiological underpinnings of this process remain largely unknown. There are extensive debates about whether separation at the tonotopic axis is the principle involved in auditory streaming, how sound elements generated by the same sound source are bound, how attention affects the neural correlates of streaming, and what the role of the auditory cortex is in streaming. In another vein, many current models of auditory streaming rely on physiological observations. The performance of all these models still lags far behind that of the average human. It seems inescapable that unless we know more about the way the brain performs auditory scene analysis, our models are unlikely to go much further. Therefore, answering these questions not only helps us understand the fundamental aspect of hearing perception, but also provides the biological evidence and constraints for improving the current models.

In this thesis, I am going to tackle some of these issues. The thesis is organized in the following way. First, in chapter 2, we demonstrate the behavioral measures of auditory streaming in ferrets. We have adapted stimuli and tasks from two previous psychophysical studies, both of which

involved performance-based measures of auditory streaming and selective attention. We trained ferrets to perform the two auditory perception tasks. The behavioral performance of ferrets in the two tasks varied as a function of stimulus parameters in a way that is qualitatively consistent with the human data. The finding of similar trends in behavioral performance as a function of stimulus parameters in the two species indicates that the perceptual organization of these stimuli varies in qualitatively the same way in ferrets as it does in humans. Therefore, this shows that the ferret can be a useful animal model to study auditory streaming and provides a foundation for the neurophysiological studies in chapters 3 and 4.

Second, it is generally assumed that separation along the tonotopic axis is the principle involved in stream segregation. Current neurophysiological theories and computational models of auditory streaming rely heavily on tonotopic organization of the auditory system to explain the observation that sequential and spectrally distant sound elements tend to form separate perceptual streams. In chapter 3, the results from the physiological experiments in awake and naïve ferrets are in contradiction with those from the human psychophysical study. Responses from neurons in the primary auditory cortex (A1) of ferrets show that spectral components that are well-separated in frequency produce comparably segregated responses along the tonotopic axis, no matter whether presented synchronously or consecutively, despite the substantial differences in their streaming percepts. The results argue against tonotopic (spectral) separation *per se* as a neural correlate of stream segregation. Instead, we suggest that temporal coherence is the principle involved in streaming.

Thirdly, inspired by experimental results from Chapter 3, we postulate that temporal correlation

across auditory channels, and not the tonotopic separation per se, is the key neural correlate of auditory streaming. In chapter 4, we provide an evidence of temporal correlations between pairs of cells in the neural responses from A1 of a ferret during auditory streaming. Furthermore, comparing the temporal correlations between pairs of cells when the animal performed the task with those at passive condition, we found that attention modulates the correlation between pairs of cells in favor of the formation of the attended stream. We also found that rapid task-related plasticity of neural receptive fields shows a pattern that is consistent with the changes in correlation. The results confirm our hypothesis that temporal correlation mediates the perception of streaming.

Finally, inspired by the above neurobiological findings, in chapter 5, we propose a computational model of auditory scene analysis, which uses temporal coherence as the primary criterion for predicting stream formation. In the model, a multi-dimensional auditory representation of a feature vector that includes pitch, timber, and location information is extracted from the input mixture. Two-dimensional correlation analysis of the auditory representations is computed. A spatial-temporal mask is formed depending on attention or memory in order to filter out the attending stream. Channels highly correlated with the target stream are enhanced and the rest are suppressed.

In summary, auditory streaming is a fundamental aspect of auditory perception. Despite wide interest and increasing scientific investigations over the last decade, the neural mechanisms underlying streaming still remain largely unknown. In the literature, the stimuli used in streaming studies are mostly sequential tones. Most studies focused on the spectral separation

between tones and ignored another important factor, the temporal relation between tones, which is known to be able to mediate the streaming percepts as well. Therefore, the general conclusion drawn from the study of sequential tones only considering the spectral separation is that the tonotopic separation is the principle involved in stream segregation. However, when we took into account the temporal factor in chapter 3, our neurophysiological results from ferrets A1 do not support this conclusion. Instead, we postulate temporal coherence is the principle involved in stream segregation. In chapter 4, we provide an evidence to support our postulation. We found the temporal correlations of spiking activity between neurons belonging to the same stream display an increased correlation, while responses among neurons belonging to different streams become less correlated. Taken together these results indicate that temporal coherence is a plausible neural correlate of auditory streaming. And we also found that attention modulates this neural correlate in favor of the formation of the attended stream. In chapter 5, we propose a neurobiologically-inspired computational model of auditory scene analysis based on temporal coherence and attention/memory. Comparing with the conventional computational auditory scene analysis models (CASAs) which use different cues, such as pitch, location, common onset/offset, and frequency/amplitude modulation, to bind channels belonging to the same stream, our model provides an elegant way of solving the problem of integration of evidence derived from multiple cues. The promising results of this dissertation research significantly advance our understanding of auditory streaming and perception.

Chapter 2 **Behavioral Measures of Auditory Streaming in Ferrets**

The material contained in this chapter is published as L. Ma, C. Micheyl, P. Yin, A.J. Oxenham, and S.A. Shamma. (2010) Behavioral measures of auditory streaming in ferrets (Mustela putorius). Journal of Comparative Psychology. 124(3): 317-30.

2.1 Introduction

Humans and many other animal species are faced with the problem that the environments they inhabit often contain multiple sound sources. The sounds emanating from these sources mingle before reaching the listener's ears, resulting in potentially complex acoustic "scenes". The listener's brain must analyze these complex acoustic scenes in order to detect, identify, and track sounds of interest or importance, such as those coming from a mate, predator, or prey. This is known among auditory researchers as the "cocktail party" problem (Cherry, 1953) or, more generally, the "auditory scene analysis" problem (Bregman, 1990). One important aspect of the auditory system's solution to this problem relates to the formation of auditory "streams". An auditory stream refers to sound elements, or groups of sounds, which are usually associated with an individual sound source, and are perceived by the listener as a coherent entity. The sound of an oboe in the orchestra, a conspecific song in a bird chorus, the voice of a speaker in a crowd, and the light footfalls of a predator in the savanna, are all examples of auditory streams. The "auditory streaming" phenomenon can be demonstrated using sounds with very simple spectral and temporal characteristics, namely, sequences of tones that alternate between two frequencies (A and B) in a repeating ABAB or ABA-ABA pattern, where A and B denote tones of (usually) different frequencies, and the hyphen represents a silent gap. Such sequences have been found to

evoke two dramatically different percepts, depending on spectral and temporal stimulus parameters (Bregman & Campbell, 1971; Miller & Heise, 1950; van Noorden, 1975). When the tones are close in frequency, most listeners report hearing a single, coherent stream of tones with an alternating pitch; this percept is referred to as "stream integration". In contrast, when the tones are more widely spaced in frequency, and occur in relatively quick succession, the stimulus sequence "splits" perceptually into two streams, as if produced by two separate sound sources; this is referred to as "stream segregation". The formation of auditory streams has been the object of a large number of psychophysical studies over the past fifty years (for reviews, see Bregman, 1990; Carlyon & Gockel, 2008; Moore & Gockel, 2002). The neural basis of the phenomenon has also attracted considerable attention, inspiring studies with approaches ranging from single or multi-unit recordings in macaques (Fishman, Reser, Arezzo, & Steinschneider, 2001; Micheyl, Tian, Carlyon, & Rauschecker, 2005), bats (Kanwal, Medvedev, & Micheyl, 2003), birds (Bee & Klump, 2004; Itatani & Klump, 2009), guinea pigs (Pressnitzer, Sayles, Micheyl, & Winter, 2008), and ferrets (Elhilali, Ma, Micheyl, Oxenham, & Shamma, 2009) to electro- or magneto-encephalography and functional magnetic resonance imaging in humans (e.g., Gutschalk, Oxenham, Micheyl, Wilson, & Melcher, 2007; Gutschalk et al., 2005; Snyder, Alain, & Picton, 2006; Sussman, Ritter, & Vaughan, 1999; Wilson, Melcher, Micheyl, Gutschalk, & Oxenham, 2007).

While there exists a substantial body of experimental data on auditory streaming in humans, and while neuroscientists are starting to explore the neural basis of this phenomenon in both humans and non-human animals, the evidence for auditory streaming in animals remains limited (for recent reviews, see Bee & Micheyl, 2008; Fay, 2008). Measuring auditory streaming in non-

human species is not as easy as measuring it in humans, who can be asked directly what they perceive. This may explain why behavioral studies of auditory streaming in animals remain relatively few and far between. The earliest such study was performed by Hulse, MacDougall-Shackleton, and Wisniewski (1997). In this study, starlings were trained to discriminate 10-s excerpts of conspecific birdsongs, and subsequently tested for generalization with mixtures of two simultaneous birdsongs (conspecific plus heteropecific, or conspecific plus natural noises or chorus). Performance with two simultaneous birdsongs was still relatively high (about 85% correct), and animals readily generalized to mixtures of familiar songs in unfamiliar backgrounds, suggesting that they were able to segregate perceptually the target song from the background. This result was interpreted as evidence for auditory stream segregation in starlings. Further evidence that starlings experience stream segregation was obtained in an elegant study by MacDougall-Shackleton, Hulse, Gentner, and White (1998), using stimuli and a task that were perhaps less ecological, but more comparable to those used in human psychoacoustical studies. In this study, starlings were conditioned using sequences of constant-frequency tones arranged temporally into triplets (i.e., groups of three tones separated by a silent gap), which, in human listeners, yield a “galloping” percept (van Noorden, 1975). The birds were later tested for generalization to sequences of triplets in which the middle tone had a different frequency from the two outer tones. The results showed decreasing generalization with increasing frequency separation between the middle and outer tones. This effect is consistent with the results of psychoacoustical studies of auditory streaming in humans, which indicate that as frequency separation increases, the middle and outer tones are increasingly likely to be heard as separate streams, and that when this happens, the galloping rhythm is no longer heard.

Fay (1998) provided evidence that auditory streaming is also present in a species in which the phenomenon is perhaps less expected to play an important role. He conditioned goldfish on a mixture of two trains of acoustic pulses, which differed in spectral content (low vs. high center frequencies) and repetition rate (19 pulses/s for the low-frequency pulses, and 85 pulses/s for the higher-frequency pulses). Later, the fish were tested for generalization using single (low or high center frequency) pulse trains over a range of rates (between 19 and 85 pulses/s). In the group tested with the low center-frequency pulses, generalization decreased toward higher pulse rates; in the group tested with the high center-frequency pulses, the converse was observed. This pattern of results is consistent with the hypothesis that, during the conditioning phase, the fish heard the low-frequency and high-frequency tones as separate streams. In a subsequent study (Fay, 2000), the fish were conditioned using trains of pulses alternating between two center frequencies (a high frequency, 625 Hz, and a lower frequency drawn between 240 and 500 Hz) at an overall rate of 40 pulses/s (20 pulses/s at a given frequency). Subsequently, the fish were tested for generalization using only 625-Hz pulses presented at various rates (from 20 to 80 Hz). Generalization to rates near 20 pulses/s was stronger in the group conditioned with pulses at 240 and 625 Hz than in the other training groups. This outcome is consistent with the hypothesis that the mixture with the widest frequency spacing was heard as two separate streams, whereas the other mixtures were less easily segregated, due to the smaller frequency separation between the alternating tones.

Some evidence that auditory streaming is also present in species that are more closely related to humans has been provided by Izumi (2002). To test for auditory streaming in Japanese monkeys, Izumi used an approach inspired by psychoacoustical studies in which listeners had to recognize

familiar melodies, the notes of which were played in alternation – the interleaved melodies paradigm (e.g., Dowling, 1968, 1973). One of the main findings of these studies was that listeners' performance in the identification of such "interleaved melodies" usually increased with the mean frequency (or pitch) separation between the two melodies. This effect, which is well known to music composers, can be explained based on the observation that frequency separation facilitates stream segregation (van Noorden, 1977). Izumi (2002) replaced the melodies by short sequences of tones, which were either rising or falling in frequency. The monkeys were first trained to discriminate such sequences presented in isolation. Then, the sequences were interleaved temporally with a sequence of unrelated tones. Performance in this interleaved condition improved as the mean frequency separation between the tones in the two interleaved sequences increased, consistent with the results of interleaved melodies studies in humans, which have been interpreted in terms of stream segregation (Bey & McAdams, 2002, 2003; Dowling, 1968, 1973).

The behavioral findings reviewed above suggest that both auditory streaming and frequency-selective attention are relatively basic auditory abilities, shared by various animal species. The current experiments were performed in the context of a broader research project, the ultimate goal of which is to investigate neural correlates of auditory streaming and selective attention in the auditory and prefrontal cortices of behaving ferrets. One of the major sub-goals of this project involves devising behavioral tasks that can be used to manipulate—and at the same time, measure—auditory streaming and selective attention in ferrets. In particular, we were looking for behavioral tasks that could be used to encourage stream segregation and frequency-selective attention. The conditioning-generalization paradigms used by Fay (1998, 2008) and Mac-

Dougall-Shackleton et al. (1998) were “neutral”; the animals were not rewarded specifically for segregating (or for integrating) streams. From this point of view, these studies are comparable to human studies in which listeners are simply asked to report whether they hear a stimulus sequence as one stream or two streams, and not encouraged by instructions, or task demands, to try to hear the sequence in a specific way (see van Noorden, 1975). Here, we were specifically interested in manipulating the attentional and perceptual state of the animal in order to later measure the influence of such a manipulation on neural responses, compared to passive or “neutral” listening to the same stimuli. A second important constraint in the design of our experiments stemmed from our long-term objective of characterizing the influence of behavior in the task on neural responses, as measured using, e.g., “classic” frequency-response curves or spectro-temporal receptive fields. We reasoned that this, and the interpretation of the results in terms of sequential streaming and frequency-selective attention, would be facilitated by the use of stimuli with relatively simple and tightly-controlled spectro-temporal characteristics, in contrast to the use of natural sounds (e.g., bird songs) used by Hulse et al. (1997) and Wisniewski and Hulse (1997).

These considerations led us to adapt stimuli and tasks from two previous psychophysical studies, both of which involved performance-based measures of auditory streaming and selective attention. The stimuli and task that we used in our first experiment were adapted from an experiment in humans by Micheyl, Carlyon, Cusack, & Moore (2005), who found that thresholds for the discrimination of changes in the frequency of the last B tone in an ABA- sequence were influenced by stimulus parameters known to control the stream segregation of pure tone sequences. Specifically, they found that thresholds increased (i.e., worsened) as the frequency

separation between the A and B tones (ΔF_{AB}) decreased, and that they decreased (i.e., improved) as the tone-presentation rate and overall length of the sequence increased. Since large A-B separations, fast tone-presentation rates, and long sequence lengths are all facilitating factors of stream segregation, this pattern of results is consistent with a beneficial influence of stream segregation on the ability to discriminate changes in the frequency of the B tones. A likely explanation for the influence of stream segregation on frequency-discrimination performance in this experiment is in terms of selective attention. When the A and B tones are heard as separate streams, attention can more easily be focused selectively on the B tones. This limits potential interference from the A tones in the processing of the pitch of the B tones (Micheyl & Carlyon, 1998). In particular, when the A and B tones are heard within a single stream, the pitch "jumps" between the A and B tones may interfere with the detection of the (usually) smaller frequency shift in the last B tone (Watson, Kelly, & Wroton, 1976); when the A and B tones are perceived as separate streams, the pitch jumps are no longer heard, and listeners can focus solely on the B tones. In Experiment 1, we adapted the stimuli and task used by Micheyl, Carlyon et al. (2005) to measure stream segregation in ferrets. Based on our experience training ferrets in auditory-perception tasks (e.g., Atiani, Elhilali, David, Fritz, & Shamma, 2009; Fritz, Elhilali, David, & Shamma, 2007; Fritz, Shamma, Elhilali, & Klein, 2003; Kalluri, Depireux, & Shamma, 2008; Yin, Mishkin, Sutter, & Fritz, 2008), these animals can detect frequency differences, but they have difficulties making low-versus-high pitch judgments—an observation confirmed by recent results (Walker, Schnupp, Hart-Schnupp, King, & Bizley, 2009). Therefore, we changed the task from pitch-direction identification to simple pitch-change detection. Under the hypothesis that ferrets experience stream segregation, we predicted that their thresholds for the detection of a

change in the frequency of B tones in ABAB… sequences should decrease (improve) with increasing A-B frequency separations.

The stimuli and task in our second experiment were inspired by studies of "informational masking" in humans. The expression "informational masking" refers to masking effects that cannot be explained primarily in terms of peripheral interactions, and that do not depend critically on the energy ratio of the target and masker (for a recent review, see Kidd, Mason, Richards, Gallun, & Durlach, 2008). Informational-masking effects are especially large when the spectral characteristics of the masker vary randomly across presentations, and the target and masker are easily confusable. However, these effects can be dramatically reduced by stimulus manipulations that promote the perceptual segregation of the signal and masker. For instance, the detection threshold for a target tone of fixed frequency can be elevated by 40 dB or more if the tone is presented synchronously with a multi-tone masker, the frequencies of which are drawn randomly on each trial (Neff & Green, 1987); this is the case even if the masker frequencies are not allowed to fall within the same critical band as the target. However, if the constant-frequency target tones repeat at a rate sufficiently fast for them to form a stream, which separates ("pops out") from the randomly-varying masker tones, they become easily detectable again (Kidd, Mason, Deliwala, Woods, & Colburn, 1994; Kidd, Mason, & Richards, 2003; Micheyl, Shamma, & Oxenham, 2007). In general, performance in the detection of the target tones improves as the width of the protected region and the repetition rate of the target tones increase (Kidd et al., 1994; Kidd et al., 2003; Micheyl, Shamma et al., 2007). Under the hypothesis that ferrets experience the effect, we predicted that these two trends would be observed in experiment 2.

2.2 Methods

2.2.1 Subjects

Two female ferrets (*Mustela putorius*) obtained from Marshall Farms were used in these experiments. Both of them were young adults (about 2 years old) each about 780 g in weight. The ferrets were housed in pairs in a cage in facilities accredited by the American Association for Laboratory Animal Care and were maintained in a 12-h artificial-light cycle. They were only brought to the Neural Systems Lab during training and testing sessions. The ferrets had free access to dry food all the time but water access was restricted to water reward during task performance 5 d/week and on weekends, they had continuous access to water. Animal condition was carefully monitored on a daily basis, and weight was maintained above 80% of their *ad libitum* weight. The care and use of animals in this study was consistent with NIH Guidelines. All procedures for behavioral testing of ferrets were approved by the institutional animal care and use committee (IACUC) of the University of Maryland, College Park.

2.2.2 Experimental Design

Two domestic ferrets were trained to perform two different tasks, performance in which has been previously found to be related to stream segregation in humans. The stimuli and behavioral paradigms are detailed below.

2.2.2.1 Experiment 1: detection of a frequency shift within a stream.

On each trial, a sequence of pure tones alternating between two frequencies (A and B) in a repeating ABAB… pattern, as illustrated in Figure 2.1, was presented. On 78% of the trials, the

B frequency changed to a higher frequency, B'. On the remaining 22% of the trials, the B frequency did not change; these trials are hereafter referred to as “shams”. The task of the ferret was to detect the frequency change, when that change was present. Based on the findings of Micheyl, Carlyon et al. (2005) we predicted that if ferrets experience streaming, then their performance in the detection of a change in the frequency of the B tones (from B to B') should be higher when the A-B frequency separation is large (promoting stream segregation between A and B) than when it is small (making it difficult or impossible to hear the B tones stream as a separate entity).

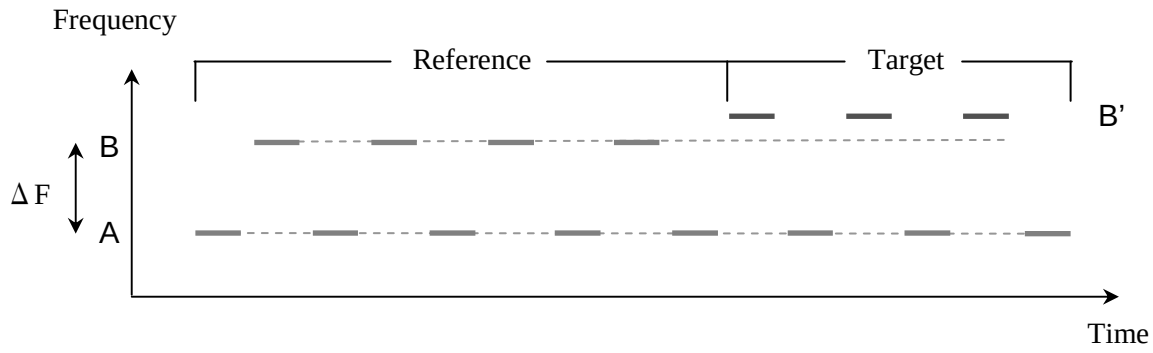

Figure 2.1 Spectro-temporal structure of the stimuli used in Experiment 1. This shows an example stimulus sequence on a trial containing target tones. The gray bars represent tones. The “reference” portion of the stimulus consisted of A and B tones alternating between two frequencies, A and B. The “target” tones had a higher frequency than the B tones, denoted as B'. Two stimulus parameters were varied: the frequency separation between the A and B tones, ΔF_{AB}, and the frequency separation between the B and B' tones, $\Delta F_{BB'}$ (See text for additional details).

The animal was trained to lick a waterspout during the “reference” sequence of AB tones, and to stop licking upon detecting the change from B to B'. Initially, the animal was trained to detect a

relatively large difference in the frequency of the B tones (from B to B') in the absence of any A tone. Once performance in this condition reached an asymptote, the A tones with a frequency 19 semitones (ST) below that of the B tones were introduced. Initially, the level of the A tones was set 50 dB below that of the B tones, which were always presented at 70 dB SPL. The level of the A tones was then raised progressively, over the course of several weeks, depending on the animal's performance. Eventually, the animal was able to perform the task relatively well with the A tones at the same level as the B tones. At that point, data collection began. Overall, training took about 7 months. The actual test phase lasted twelve days (four days for each ΔF_{AB}). During this test phase, the ferret performed at least 70 trials each day.

Detailed stimulus parameters were as follows. Each tone was 75 ms long, including 5 ms onset and offset cosine ramps. Consecutive tones were separated by a silent gap of 50 ms. Therefore, the repetition rate of the elementary AB pattern was equal to 4 Hz. The frequency of the B tone was fixed at 1500 Hz. The frequency of the A tone was constant within a block of trials, but varied across testing days in order to produce different frequency separations between the A and B tones, denoted here as ΔF_{AB}. Three ΔF_{AB}'s were tested: small (6 ST), medium (9 ST), and large (12 ST, i.e., one octave). Although the smallest ΔF_{AB} (6 ST) used here is relatively large, and would be considered "intermediate" in humans, we found while training the ferret that the animal could not do the task with smaller A-B separations; consequently, we decided to use larger separations. The frequency separation between the B and B' tones, denoted as $\Delta F_{B\,B'}$, varied randomly across trials within a test session in order to produce different levels of task difficulty, yielding different levels of performance. In all conditions in which the A tones were present, five values of $\Delta F_{BB'}$ were tested: 4%, 12%, 20%, 28%, and 36%. A larger number of

$\Delta F_{BB'}$s were tested during the initial training phase, in which the A tones were absent: 6, 8, 10, 12, 14, 16, 18, 21, and 27%.

The total trial duration varied randomly from 1.875 to 7.875 s, depending on the number of "reference" pairs (AB) presented before the introduction of the B' tones. This number was selected randomly between 4 and 28 on each trial. The number of "target" pairs (AB') was fixed at 3. If the animal stopped licking within 850 ms after the introduction of the first B' tone in the sequence, this was counted as a hit; otherwise, the trial was categorized as a miss. If a stop-lick response was produced on a sham trial, it was counted as a false alarm; otherwise, the trial was counted as a correct rejection. False alarms had no consequence. Following the 850-ms after the introduction of the first B' tone in the sequence, the spout became electrified, and hence the ferret received a mild shock if it continued licking afterwards, and the trial was labeled a miss. Each trial included silent periods of 400 ms pre-stimulus, and 600 ms post-stimulus.

2.2.2.2 Experiment 2: detection of regularly repeating target tones in a random multi-tone background.

This experiment was inspired by psychoacoustic experiments on informational masking (Kidd et al., 1994; Kidd et al., 2003; and Micheyl, Shamma et al., 2007). An example spectrogram of the stimuli used in this experiment is shown in Figure 2.2. On each trial, a sequence consisting of multiple tone pips with random frequencies and random onset times ("maskers") were presented. At some point in this random sequence, a regularly repeating sequence of constant-frequency tones ("targets") was introduced. The task of the ferret was to detect the target sequence amid the randomly varying masker tones. The animal was trained to withhold licking until the target was

introduced, and to start licking upon detecting the target. If a lick response occurred within 150 to 1050 ms after the onset of the first target tone, it was counted as a hit and reinforced with a 1/3 ml of water. These parameters were chosen based on the consideration that the quickest reaction time in ferrets is approximately 150 ms, and that the target was 900 ms long. Misses had no consequence. In this experiment, there were no "sham" trials; the target tones were presented on all trials. However, the start time of the target sequence varied randomly between 720 and 2160 ms after the onset of the masker sequence. When the animal produced a lick response before the onset of the target sequence, this was counted as a false alarm, the trial was aborted, and followed by a short timeout.

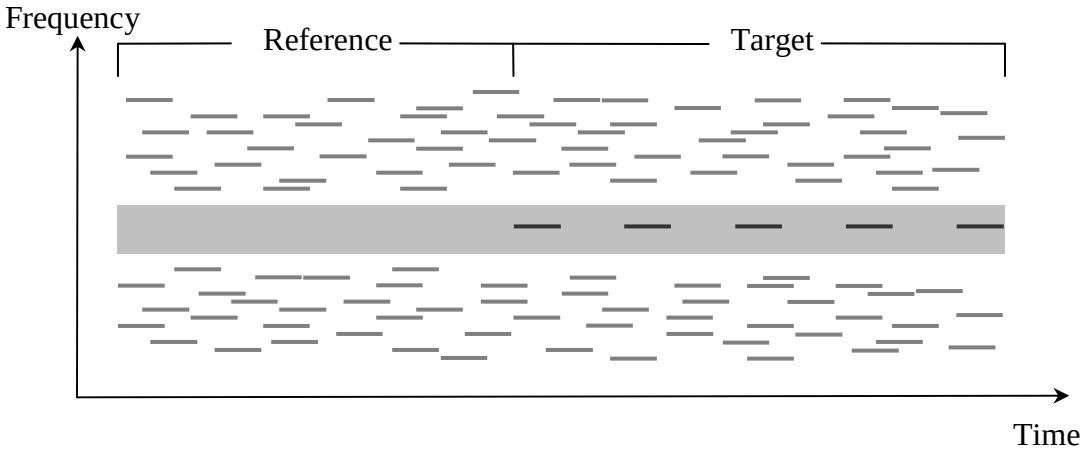

Figure 2.2 Schematic spectrogram of an example stimulus presented on a trial in Experiment 2. During the "reference" portion of the stimulus, only masker tones (gray bars) with random frequencies and onsets times were presented. During the "target" part, target tones (dark bars) repeating regularly at a constant frequency were introduced. The gray area around the target represents the "protected zone" (PZ), within which masker tones were not allowed to fall (See text for additional details).

Note that response contingencies are reversed here compared to Experiment 1, in that positive reinforcement is used (as opposed to negative reinforcement in Experiment 1). This control is specifically important for physiological experiments, because auditory cortical responses and adaptations can depend critically on whether the "target stimuli" in the tasks were aversively or positively reinforced (David, Fritz, & Shamma, 2008). Consequently, we felt it was important to demonstrate in this study that ferrets could perform both forms of the streaming tasks so as to facilitate recordings from their auditory areas during such behaviors.

The stimulus details were as follows. Each tone-pip (target or masker) was 70 ms long, including 5 ms onset and offset ramps. On each trial, 5 target tones were presented. Consecutive target tones were separated by a silent gap of 110 ms, yielding a repetition rate of about 5.6 Hz. Trial length varied randomly between 1.62 and 3.06 s across trials. These durations include the variable-length "reference" sequence (0.72 to 2.16 s) plus the fixed-duration "target" sequence. The masker tones occurred at an average rate of 89 tones per sec. The masker tones were generated as follows: first, eight different masker-tone frequencies were selected at random for every 90 ms; then, the masker tones were shifted pseudo-randomly in time, in such a way that they were not synchronous with the target, except by chance. The masker tone frequencies were drawn at random from a fixed list of values spaced one ST, approximately 6%, apart, excluding a "protected zone" (PZ) around the frequency of the target tones. The half-width of the PZ determines the minimum allowed frequency separation between the target and the closest masker component on either side. Three half-widths were tested: small (6 ST), medium (10 ST), and large (14 ST). Masker frequencies were selected from within a two-octave range on both sides of the PZ. The target frequency was roved daily from 3.5 to 4.1 kHz. PZ was varied randomly from

day to day, while target intensity and trial lengths varied within a session. The masker tones were presented at 50 dB SPL (each). Target-tone levels of -4, 0, +4, +8 and +12 dB relative to the level of the masker tones were tested. These values were chosen to produce different hit rates, allowing a psychometric function to be traced.

We also studied the influence of target repetition rate on performance. Three rates were tested: 3.7, 5.7, and 11.1 target tones/s. These different rates were produced by varying the duration of the silent interval between consecutive target tones (from 20 to 200 ms).

For this task, the training phase spanned 14 months, including sporadic intermissions of a few weeks during which the animal did not behave. Typically, the ferret was trained five days per week. Initially, the animal was trained with a very wide PZ (16 ST). The width of the PZ was progressively reduced, week after week (or sometimes, month after month). When the animal's performance reached an asymptote, training stopped and testing proper started. During the test phase, the animal performed a total of 24 sessions using the 6- and 10-ST PZ half-widths (twelve sessions for each of these two half-widths), at a pace of one session per day. For the 14-ST PZ half-width, the animal performed 11 daily sessions.

2.2.3 Apparatus

Ferrets were tested in a custom-designed cage (8 x 15 x 9 inch) mounted inside a Sonex-foam lined and single-walled soundproof booth (Industrial Acoustics Corporation). The stimuli were generated using Matlab (The MathWorks, Natick, MA). They were sampled at 40 kHz, played out at 16-bit resolution (NI-DAQ), amplified (Yamaha A520), and finally delivered through a

speaker (Manger) mounted in the front of the cage, at approximately the same height above the testing cage as the metal spout that delivered the water reward. Lick responses were registered by a custom “touch” circuit Acknowledgement.

2.2.4 Data Analysis

The behavioral data were analyzed using techniques from signal detection theory (Green & Swets, 1966). In particular, the responses of the animals were used to compute the area under the receiver operating characteristic (ROC). The area under the ROC provides an unbiased measure of performance, with 0.5 reflecting chance performance, and 1 reflecting perfect performance (Green & Swets, 1966; Hanley & McNeil, 1982). In Experiment 2, ROCs were derived by varying the duration of the response window, defined as the time interval within which a start-lick event was registered. The rationale for this analysis is that longer response times correspond to more liberal placements of the internal decision criterion (Luce, 1986; Yin, Fritz, & Shamma, submitted). The range of possible occurrence times of the first target tone in the stimulus sequence was from 0.72 to 2.16 s. On trials on which the target tones occurred relatively early (i.e., between 720 ms and 1620 ms), the response window started 150 ms after the onset of the first target, and a lick event occurring within the response window was counted as a hit. On trials on which the targets occurred relatively late (i.e., between 1.62 s and 2.16 s), the response window started 150 ms after the time at which the first target should have started had the target tones been early, and a lick occurring within the response window was counted as a false alarm. The duration of the response window was varied from 0 to 900 ms in 50 ms increments. Areas under the resulting ROCs were approximated using trapezoids. The advantage of this method is that it does not require specific assumptions regarding the underlying distributions.

In Experiment 1, a different data-analysis technique had to be employed in order to accommodate the different experimental design, and different response contingencies. First, we measured the duration for which the animal had made contact with the water spout within a 400-ms "reference" epoch, which just preceded the introduction of the target tones. If this duration was less than 20 ms (5% of the reference-epoch duration), we considered this as an indication that the animal was not ready for task performance, and data from the current trial were not included into subsequent analyses. In contrast, trials on which the animal licked the water spout for at least 20 ms during the reference period were retained for further analysis. These trials were divided into two groups, depending on whether the animal had come into contact with the water spout during the time period within which shocks could be delivered if the animal had not stopped licking. This "shock period" started 850 ms after the onset of the first target tone, and lasted for 400 ms. If the animal had made contact with the spout during the shock period, the trial was categorized as a "miss" or as a "correct rejection", depending on whether or not target tones were presented on that trial. If the animal had not made contact with the spout during the shock period, the trial was categorized as a "hit" or as a "false-alarm", depending on whether or not target tones were presented on that trial.

As a result, a single pair of hit and false-alarm rates was available for each condition. When the ROC contains a single point, approximation using trapezoids can lead to severe underestimation of the ROC area. Accordingly, in this experiment, we had to resort to parametric assumptions. Specifically, the ROC area was computed as the surface under a binormal curve passing through three points: (0,0), (1,1), and the point defined by the measured pair of hit and false-alarm rates.

Thresholds in both experiments were estimated by fitting the ROC-area data as a function of the relevant stimulus parameter ($\Delta F_{BB'}$ in % for Experiment 1, relative target level in dB for Experiment 2) with a sigmoid function defined by the following equation,

$$Pc(\Delta) = 0.5 + \rho\left[1 + e^{-(\Delta-\theta)/\sigma}\right]^{-1} \quad (1)$$

where *Pc* is the proportion of correct responses; Δ denotes the value of the stimulus parameter ($\Delta F_{BB'}$ in % for Experiment 1, relative target level in dB for Experiment 2); ρ is the dynamic range of the psychometric function, which corresponds to the difference between the guessing rate (0.5) and the lapse (i.e., miss) rate, λ; θ is the threshold, defined as the frequency difference (in Experiment 1), or target level (in Experiment 2) at which $Pc = (0.5+\lambda)/2$, i.e., the midpoint between the guessing rate and the lapse rate; σ is a "standard deviation" parameter, which corresponds to the reciprocal of the slope of the psychometric function. For Experiment 1, a data point corresponding to $Pc(0) = 0.5$ was introduced in order to reflect the fact that when $\Delta F_{BB'}$ was equal to 0% (i.e., the B and B' tones had the same frequency, and there was no signal for the animal to detect), performance should be at chance. In addition, the contribution of each mean data point to the overall fit was weighted by the inverse of the variance around the mean, and constraints were placed on the slope parameter in order to prevent unrealistically steep PFs in the 9-ST ΔF_{AB} condition. For each condition, 95%-confidence intervals (CIs) around the threshold estimates were computed using a statistical resampling-with-replacement technique (bootstrap with 1,000 draws) assuming binomial dispersion (Efron & Tibshirani, 1993).

2.3 Experiment 1

2.3.1 Results

The results of Experiment 1 are illustrated in Figure 2.3. Figure 2.3a shows an example of bi-normal ROC, determined as explained in the Data Analysis. This ROC was obtained based on a pair of hit and false-alarm rates measured using a 12-ST ΔF_{AB}, and a 36% $\Delta F_{BB'}$. In this example, the area under the ROC, which is shown in gray, was equal to 0.87, indicating good performance.

The ROC area was computed in a similar way for all other ΔF_{AB} and $\Delta F_{BB'}$ conditions. The resulting set of ROC areas are shown as data points in Figure 2.3b. These data were fitted using sigmoid psychometric functions for each ΔF_{AB} condition separately, and the best-fitting functions were used to estimate a threshold (defined as the $\Delta F_{BB'}$ value corresponding to the midpoint between chance performance and the upper asymptote) for each ΔF_{AB} condition.

The resulting threshold estimates are plotted in Figure 2.3c, along with the 95% CIs (computed using bootstrap). It can be seen that thresholds were highest for the lowest ΔF_{AB} tested (6 ST), and substantially smaller ($p < 0.05$) for larger separations (9 and 12 ST). In fact, the thresholds measured with separations of 9 and 12 ST were not significantly larger ($p > 0.05$) than those measured in the baseline condition, which did not contain any A tones. Comparing the thresholds for each ΔF_{AB} condition to those for the baseline condition, the effect sizes (Glass's Δ) were 6.09 with 95% CIs [3.91, 4.21] (6 ST vs. baseline condition), -0.83 with 95% CIs [-0.92, -0.74] (9 ST vs. baseline condition), and 0.48 with 95% CIs [0.39, 0.57] (12 ST vs. baseline condition). The only apparent difference in results between the 9 and 12 ST conditions was that the asymptotic

proportion of correct responses was somewhat higher in the latter condition (around 0.9) than in the other conditions. We have no explanation for this marginal observation. While asymptotic proportions of correct responses below 1 are usually regarded as indicative of attentional "lapses" (Klein, 2001), there was no a priori reason to expect a lower lapse-rate in the 12-ST condition than in the other conditions.

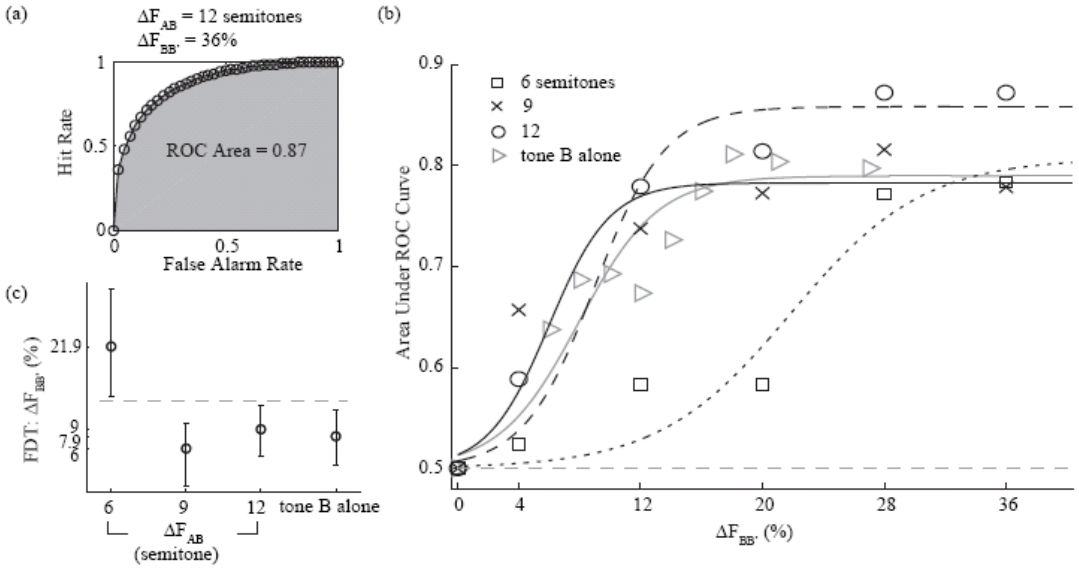

Figure 2.3 Performance measures in Experiment 1. (a) Example ROC curve obtained using the technique described in the main text. This curve was computed based on data obtained at 12-ST ΔF_{AB} and a 36% $\Delta F_{BB'}$. (b) ROC area as a function of $\Delta F_{BB'}$ and best-fitting psychometric functions. ROC areas are shown as symbols; the best-fitting psychometric functions as lines. The different symbols and line styles indicate different ΔF_{AB} conditions (6, 9, and 12 ST). (c) Frequency discrimination thresholds (FDTs) estimated based on the psychometric functions shown in panel b. The error bars indicate 95% CIs (bootstrap) around the mean FDTs. The dashed line indicates non-overlapped over 95% CIs.

2.3.2 Discussion

The pattern of results illustrated in Figure 2.3c is qualitatively consistent with the psychophysical data obtained by Micheyl, Carlyon et al. (2005) in human listeners. These authors found that frequency discrimination thresholds for target B tones inside repeating ABA sequences comparable to those used here improved as the frequency separation between the A and B tones increased. They explained this effect, and other effects of stimulus parameters (including rate and sequence length), in terms of stream segregation and selective attention. Specifically, they suggested that stimulus manipulations that promoted the perceptual segregation of the A and B tones into separate streams made it easier for listeners to attend selectively to the B tones, and to ignore the irrelevant but potentially interfering pitch information conveyed by the A tones.

The presented finding of smaller thresholds for the detection of frequency changes in the B tones for larger A-B frequency separations is consistent with the hypothesis that stream segregation facilitates selective attention in ferrets and leads consequently to improved detection thresholds. However, this interpretation can be further supported by other factors that are known to modulate streaming such as tone presentation rate and sequence length. Unlike our previous experiments in humans (Micheyl, Carlyon et al., 2005), we did not manipulate these parameters in the current experiments. Hence, it will be interesting to explore in the future the dependence of detection thresholds at a fixed ΔF_{AB} on presentation rate and sequence length, and whether this dependence is consistent with a beneficial influence of stream segregation.

Although the ferret data show trends that are qualitatively similar to those observed in psychophysical studies with human listeners, there are also several important differences

between the ferret and human data. Firstly, the frequency-discrimination thresholds that were measured in the ferret are substantially larger than those that have been measured in humans typically. In traditional 2I-2AFC experiments, highly trained human listeners can achieve frequency-discrimination thresholds of 0.1-0.2% (roughly 1.5-3 Hz) at 1.5 kHz (e.g., Moore, 1973; Wier, Jesteadt & Green, 1977; Micheyl, Delhommeau, Perrot, & Oxenham, 2006). In an ABA context, Micheyl, Carlyon et al. (2005) measured thresholds of less than 1% using test frequencies in the vicinity of 1 kHz—at least, when the A-B frequency separation was sufficiently large for listeners to hear the A and B tones as separate streams. These values are substantially smaller than the 8% or more average thresholds measured here under comparable (though not identical) stimulus conditions, using different procedures. On the other hand, the thresholds that were measured in this study compare well with those obtained in an earlier study (Sinnott, Brown, & Brown, 1992) in the gerbil (9% at 1 kHz, 10% at 2 kHz). These thresholds are also not very far off from those measured in the rats, Guinea pig, or chinchilla, where frequency discrimination thresholds ranging from 2 to 7% on average (with substantial inter-subject variability) have been obtained using test frequencies between 2 and 5 kHz (Heffner, Heffner, & Masterton, 1971; Kelly, 1970; Sloan, Dodd, and Rennaker, 2009; Nelson & Kriester, 1978; Syka, Rybalko, Brozek, & Jilek, 1996; Talwar & Gerstein, 1998, 1999). Thus, even though ferrets are not rodents (they are carnivores, most closely related to weasels), their frequency discrimination thresholds appear to be similar to those of rodents, which are generally much larger than those measured in humans. In all other species in which frequency discrimination thresholds have been measured, to our knowledge, the results indicate that these thresholds are not quite as low as those measured in highly trained human listeners–although for cats, they can be as low as 0.85% at 1 kHz and 0.75% at 2 kHz (Elliott, Stein, & Harrison, 1960); for dog,

roughly 0.9% at 1 and 2 kHz (Baru, 1967). This appears to be the case even for monkeys, in which frequency discrimination thresholds ranging between 1.6 and about 4% have been reported (Prosen, Moody, Sommer, & Stebbins, 1990; Sinnott & Brown, 1993; Sinnott, Petersen, & Hopp, 1985). It has been suggested that small frequency-discrimination thresholds below 4-5 kHz in humans reflect the use of temporal (i.e., phase-locking) information (Moore, 1973; Sek & Moore, 1995; Micheyl, Moore, & Carlyon, 1998), whereas monkeys and other animals may rely more heavily on tonotopic (i.e., rate-place) information (Prosen et al., 1990; Sinnott & Brown, 1993; Sinnott et al., 1985).

A second noteworthy difference between the ferret results and the human data is that, during the training phase, the ferret was found to be largely unable to perform consistently above chance when the A-B frequency separation was smaller than 6 ST; this is why smaller separations were not included into the design of Experiment 1. In contrast, in humans, thresholds could still be reliably measured for ΔF_{AB}'s as small as 1 ST (Micheyl, Carlyon et al., 2005). A possible explanation of this discrepancy is that, although human listeners almost certainly heard the tones as a single stream in these conditions, they could still perform the task above the chance level by comparing the frequency of the last B tone with that of a temporally adjacent A tone, or by sensing an overall increase or decrease in pitch between the last two triplets. The ferret was perhaps not able to adapt its listening strategy depending on ΔF_{AB} to take advantage of a different cue at small A-B separations than at larger ones. In this context, the observation that ferrets appear to need larger ΔF_{AB}'s than humans to perform reliably in the task could be due to larger frequency separations being needed to induce a percept of stream segregation in ferrets than in humans. For humans, the fission boundary, which corresponds to the smallest frequency

separation below which listeners are no longer able to hear two separate streams (van Noorden, 1975), is approximately equal to 0.4 times the equivalent rectangular bandwidth (ERB) of auditory filters (Rose & Moore, 2000; Rose & Moore, 2005). At 1 kHz, the ERB for normal-hearing listeners is 132 Hz (about 13% of the center frequency) (Moore, 2003), yielding a fission boundary of approximately 5% of the center frequency, or slightly less than 1 semitone. Micheyl, Carlyon et al.'s (2005) data indicate that the listeners in that study usually needed ΔF_{AB}'s larger than 1 semitone to be able to discriminate changes in the frequency of the B tone relatively accurately. To the extent that the fission boundary for stream segregation scales with auditory-filter bandwidths across species, one should expect this boundary to be larger in ferrets than humans. Even though, to our knowledge, auditory-filter bandwidths have not been measured in ferrets, the various other mammalian species in which they have been measured behaviorally, which include the cat (Pickles, 1979; Nienhuys & Clark, 1979) chinchilla (Seaton & Trahiotis, 1975) and macaque monkey (Gourevitch, 1970; for a review, see Fay, 1988), indicate somewhat larger bandwidths than in humans (Shera, Guinan & Oxenham, 2002).

Another factor, which may explain why ferrets require larger A-B frequency separations than humans, relates to frequency-selective attention bandwidths. Frequency-selective attention is likely to be critically involved for successful selective processing of changes in frequency in the presence of extraneous tones—in the present case, temporally adjacent (A) tones between the target (B or B') tones. In humans, frequency-selective attention has traditionally been measured using the "probe-signal" method (Greenberg & Larkin, 1968). In these experiments, the listener detects a tone close to its masked threshold in noise, and on a small proportion of randomly selected trials, the signal is presented at another frequency (probe). The results of these

experiments reveal that, as distance between the probe frequency and the signal frequency increases, the percentage of correct detections decreases, forming an inverted V-shaped selective-attention curve (Dai, Scharf, & Buus, 1991; Greenberg & Larkin, 1968). The width of the curve provides an indication of the bandwidth of the frequency-selective attention filter in the listener. In humans, this width is closely related to the bandwidth of auditory filters (Moore, Hafter, & Glasberg, 1996). To the extent that a similar relationship exists in ferrets, and that auditory-filter bandwidths are wider in ferrets than in humans, this could explain why ferrets need larger A-B separations to successfully detect changes in the frequency of specific tones in a stimulus sequence that contains other (irrelevant) frequencies and frequency changes.

2.4 Experiment 2

2.4.1 Results

Figure 4a illustrates how the hit and false alarm rates measured in this experiment increased with the duration of the response window. In this particular example, the PZ half-width was 14 ST, and the target tones were 12 dB above the masker tones. Both the hit rate and the false alarm rate tended to increase with the duration of the response window. However, the hit rate increased more steeply than the false alarm rate, indicating that in this condition, the ferret could reliably detect the target tones.

These pairs of hit and false-alarm rates were used to construct the ROC shown in Figure 2.4b. In this example, the ROC area was equal to 0.81. ROC areas were computed in this way for each PZ-width and target-level condition. The resulting ROC areas are plotted as a function of target level in Figure 2.4c. As expected, ROC area increased with target level. Overall, performance

was lower at the smallest PZ half-width (6 ST) than at larger half-widths (10 and 14 ST). This effect is illustrated in Figure 2.4d, which shows how thresholds (estimated based on the psychometric-function fits as explained in the Data Analysis section) improved as the PZ width increased. Comparing the thresholds among each PZ condition, the effect sizes (Hedges' g) were 4.41 with 95% CIs [4.25, 4.57] (6 vs. 10 ST condition), 3.03 with 95% CIs [2.90, 3.16] (6 vs. 14 ST condition), and 0.62 with 95% CIs [0.53, 0.71] (14 vs. 10 ST condition).

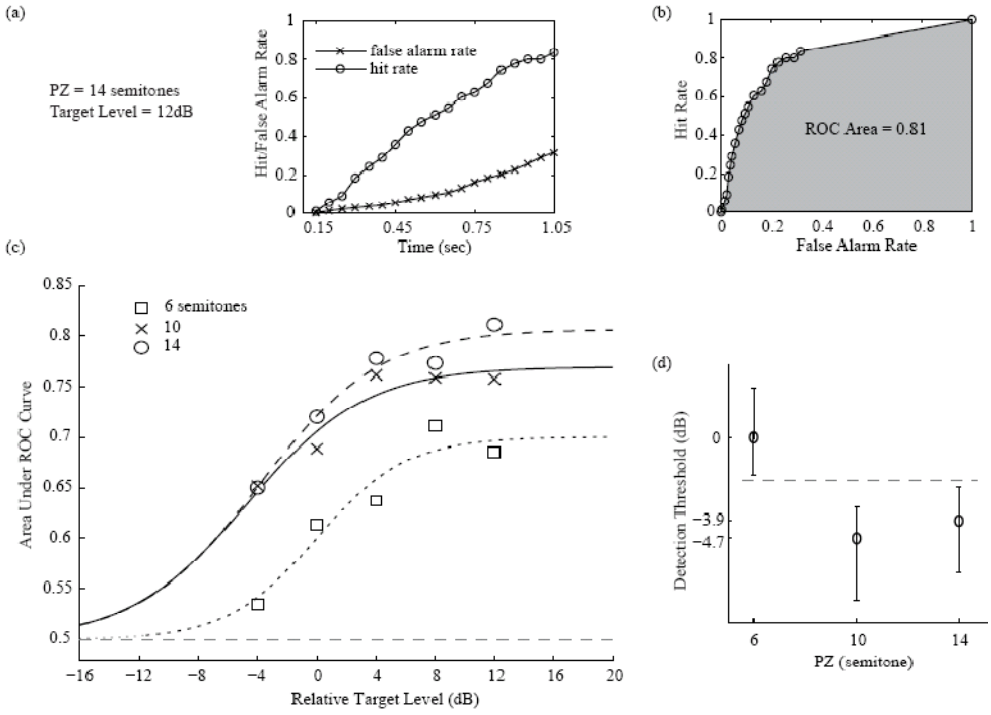

Figure 2.4 Performance measures in Experiment 2. (a) Example series of hit and false alarm rates generated by varying the response window from 150 to 1050 ms after the target onset. These example data were obtained using a PZ half-width of 14 ST and a relative target level of +12 dB (relative to the masker level). (b) Example ROC curve obtained by plotting the series of hit rates from panel a as a function of the corresponding false alarm rates. The ROC area, shown in gray, was estimated using a nonparametric technique (see text for details). (c) ROC area as a function of PZ half- width, and best-fitting psychometric functions. ROC areas are shown as symbols; the

best-fitting psychometric functions as lines. The different symbols and line styles indicate different PZ conditions (6, 10, and 14 ST). (d) Detection thresholds estimated based on the psychometric functions shown in panel c, for the different PZ half-widths. The error bars indicate 95% CIs around the mean. The dashed line indicates non-overlapped over 95% CIs.

Figure 2.5 illustrates how the ROC area varied over the time course of the stimulus sequence, from 150 ms after the onset of the first target tone until 150 ms after the offset of the last target tone. The different panels correspond to different target levels (relative to the masker), from low (left) to high (right). Within each panel, the different curves correspond to the different PZ half-widths that were tested. The different data points within each curve correspond to ROC areas based on pairs of hit and false-alarm rates computed using increasing response-window durations (in 50-ms increments). The ROC area generally increased over time following the introduction of the target tones, $F(18, 3040) = 100.17$, $p < 0.001$, $\eta_p^2 = 0.37$. This effect became more marked as the PZ became wider, $F(36,3040) = 2.52$, $p < 0.001$, $\eta_p^2 = 0.03$ and as the level of the target tone was raised from 4 dB below, to 12 dB above, the masker level, $F(72,3040) = 1.83$, $p < 0.001$, $\eta_p^2 = 0.04$.

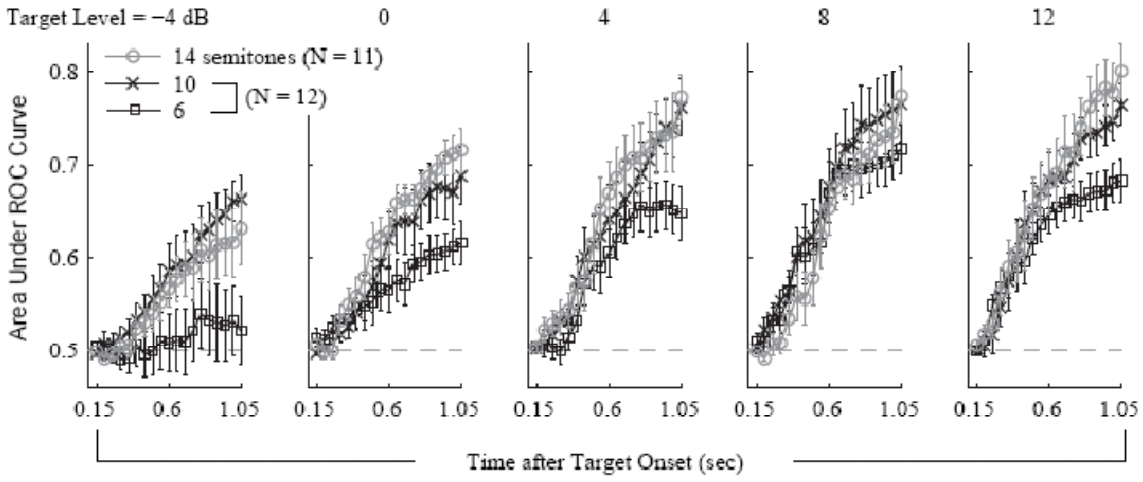

Figure 2.5 Area under the ROC curve as a function of time after target onset in Experiment 2. The different panels correspond to different target levels (relative to the masker), from low (left) to high (right). The error bars are standard errors of the mean across daily sessions.

Figure 2.6 illustrates the influence of the target repetition rate on detection performance. These data were obtained using a target level 4 dB below the masker level. The different line styles indicate different PZ widths. As can be seen, the ROC area was larger at the largest (14 ST) PZ half-width than at the two smaller widths, $F(2,27) = 8.58$, $p < 0.01$, $\eta_p^2 = 0.39$ and it increased with target repetition rate, $F(2,27) = 4.35$, $p < 0.05$, $\eta_p^2 = 0.25$.

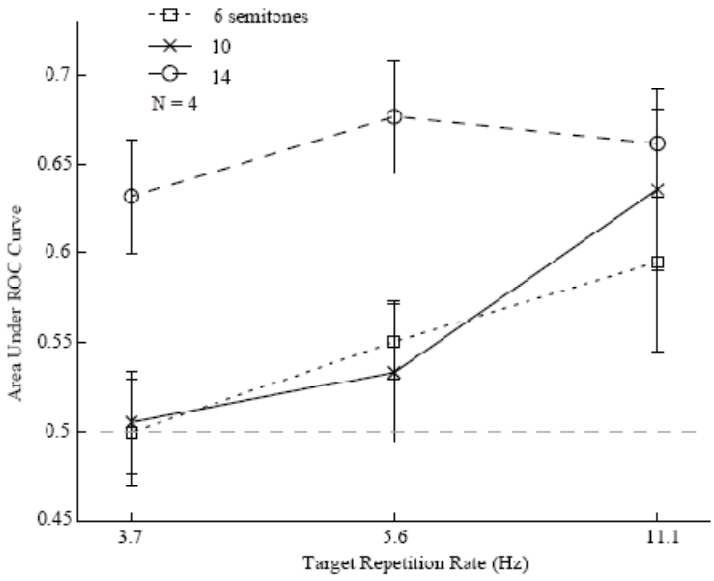

Figure 2.6 Area under the ROC curve as a function of target repetition rate in Experiment 2. The different line styles indicate different PZ widths. The error bars denote standard errors of the mean across daily sessions.

2.4.2 Discussion

The effects illustrated in Figure 2.4 are qualitatively consistent with earlier results in the human psychoacoustics literature, which show improvements in thresholds (Richards & Tang, 2006) or *d′* (Micheyl, Shamma et al., 2007) in a task involving the detection of regularly repeating target tones among randomly varying masker tones, as the width of the PZ around the target tones increases. However, there are some noteworthy differences between the ferret data and human data. For instance, Richards and Tang (2006) observed threshold improvements of 10 dB or more as they increased the half-width of the PZ around 1 kHz target tones from 20 to 350 Hz. These values correspond to half widths of approximately 0.34 and 5.2 ST, respectively. These values are substantially smaller than those used in the current study (6-14 ST), indicating that the ferret needed substantially larger PZ widths than human listeners in order to detect the target tones.

It is important to acknowledge that although the results have been discussed in terms of informational masking, a possible contribution of energetic masking cannot be completely eliminated. With the moderate stimulus level (50 dB SPL per tone), and wide protected-zone widths (12, 20, and 28 ST) used in this experiment, the contribution of energetic masking would probably have been minimal in humans, because even the smallest (12-ST) PZ width is roughly ten times the equivalent rectangular bandwidth (ERB) in normal-hearing listeners (Moore, 2003). As mentioned above, in the various animal species in which auditory-filter bandwidths have been measured behaviorally, these bandwidths have been found to be somewhat larger than in humans. However, for energetic masking to significantly limit the detection of the target tones in the current experiment, which involved PZ widths of one octave or more, frequency selectivity would have to be considerably less in ferrets than in humans.

The increases in ROC area over time following the onset of the stimulus sequence, which were seen in Figure 2.5, are qualitatively consistent with the results of psychophysical studies in humans, which indicate that the target tones become more detectable over time within the course of the stimulus sequence (Gutschalk, Micheyl, & Oxenham, 2008; Kidd et al., 2003; Micheyl, Shamma et al., 2007). This effect may be related to the phenomenon known as the "build-up of segregation", whereby the probability of hearing a sequence of alternating tones as two separate streams instead of a single coherent stream increases gradually (over several seconds) following the stimulus onset (Anstis & Saida, 1985; Bregman, 1978; van Noorden, 1975). The increasing probability of detecting the target tones amid the randomly varying maskers may be related to increasing segregation of the target tones from the background tones over time. From this point of view, the present findings suggest that stream segregation takes some time in ferrets, as it does in humans. However, because the changes in ROC area shown in Figure 2.5 occurred over approximately 1 s, this effect can also be explained in terms of response time—more specifically, decision time—without necessarily implicating the build-up of stream segregation. The animal may have needed more time to respond in conditions in which the target tones were harder to detect. Thus, the trends observed in Figure 2.5 could be reproduced, for instance, by a diffusion model of response time in which noisy sensory evidence accumulates toward a bound, and the rate of accumulation is determined by the strength of the sensory evidence (e.g., Ratcliff, Van Zandt, & McKoon, 1999). These trends could also be accounted for by probability-summation or multiple-looks model (Green & Swets, 1966; Viemeister & Wakefield, 1991), in which the probability of correct detection increases with the number of signals. Thus, these data do not allow us to conclusively dissociate components of response time that are unrelated to the build-up of segregation from components that are related to it.

The decrease in ROC area with decreasing target repetition rate in Figure 2.6 is qualitatively consistent with psychophysical data in humans (Kidd et al., 2003; Micheyl, Shamma et al., 2007). These data show a decrease in *d′* as the target presentation rate decreased from 16.7/s to 5.6/s, a range that partially overlaps with that tested in the ferret (11.1/s to 3.7/s). A similar effect was observed by Kidd et al. (1994), who measured masked detection thresholds rather than *d′*. These authors found that thresholds for the detection of 4 or 8 tone bursts inside a randomly varying multi-tone background improved by about 15 dB as the interval between consecutive target bursts decreased from 400 ms (which in that study yielded a target rate of 2.2/s) to 50 ms (a rate of 9.1/s). However, the trend illustrated in Figure 2.6 may also be explained by an increase in detectability of the target tones as a function of their number—an effect discussed in the preceding paragraph, and consistent with probability-summation or multiple-looks models. This is because, in the current experiment target sequence length was kept constant, independent from tone repetition rate, so that the total number of target tones in the stimulus sequence increased with the repetition rate. However, this probability-summation or multiple-looks models explanation is contradicted by Kidd et al.'s experiments (2003) in which they used similar stimuli to those in experiment 2, except for the masker components being synchronous with each target tone. They compared the target detection threshold under different repetition rates, but with the number of target tones *fixed*. They found that thresholds decreased with increasing rates, a finding inconsistent with a simple version of multiple-looks model, but instead in favor of a perceptual segregation of the signal from the masker. Accordingly, it is also unlikely that the effect seen in our experiment is purely due to probability-summation or multiple-looks models.

2.5 General Discussion

The results indicate that the behavioral performance of ferrets in two auditory perception tasks, which have been used to measure auditory streaming in humans, varies as a function of stimulus parameters in a way that is qualitatively consistent with the human data. Specifically, in Experiment 1, higher performance and lower thresholds in the detection of frequency shifts between targets tones at a given frequency were observed when temporally interleaved tones (interferers) were either absent, or at a remote frequency, compared to when the interfering tones were closer in frequency to the targets. This finding is qualitatively consistent with the human psychophysical data of Micheyl, Carlyon et al. (2005), which were explained as resulting from stream segregation allowing listeners to attend selectively to the target tones. Selective attention to the target sounds presumably allows the characteristics of these sounds (e.g., pitch or loudness) to be perceived more acutely, while other background sounds are analyzed less thoroughly by the auditory system. Accordingly, thresholds for the detection or discrimination of changes in the frequency (subjectively, pitch) of the B tones are expected to be smaller under stimulus conditions that promote stream segregation between the A and B tones. The increase in performance in the target-frequency discrimination task with increasing A-B frequency separation is consistent with stream segregation becoming easier as the A-B frequency separation increases. The finding of increasing performance, and decreasing thresholds, with increasing size of the protected width in Experiment 2, where the task was to detect regularly repeating target tones among randomly-varying masker tones, is also consistent with psychophysical data in humans (Micheyl, Shamma et al., 2007). Here the effect has been interpreted as resulting from wider frequency separations between the target and masker tones facilitating the perception of the target tones as a separate stream.

While we cannot ascertain that the ferrets experienced the stimuli in these experiments in the same way as humans do, the finding of similar trends in behavioral performance as a function of stimulus parameters in the two species indicates that the perceptual organization of these stimuli varies in qualitatively the same way in the animals as it does in humans. At the same time, the present data indicate important quantitative differences in the way in which performance, or thresholds, in the two considered tasks vary as a function of stimulus parameters in ferrets and humans. In general, the ferrets needed larger frequency separations in Experiment 1, and larger protected-zone widths in Experiment 2, in order to be able to perform the tasks above chance. Moreover, even under the most favorable stimulus conditions (i.e., very large spectral separations), performance in the ferrets was still well below ceiling, and thresholds were still considerably larger than those measured in humans. This cannot be due solely to insufficient training, because the ferrets received fairly extensive training, and performed these tasks or a simpler version of them repeatedly over the course of several months. This suggests that these tasks are intrinsically difficult for ferrets. A likely reason for this is that both tasks require selective attention, in addition to basic auditory detection and discrimination abilities, such as the ability to discriminate the frequencies of tones. Thus, while behavioral studies have found that the performance of various animal species in basic auditory detection or discrimination tasks can equal and sometimes exceed that of humans, these studies almost invariably used simpler stimuli than those used in the experiments described here. Importantly, both of the tasks that were used in this study required from the animal that it be able to sustain selective attention to a subset of stimuli within a sequence of sounds that contains other, irrelevant sounds. Unfortunately, such a selective-attention ability is required, to some extent, by any task that aims to measure performance in separating auditory streams.

Two limitations of the present study must be acknowledged. Firstly, while the patterns of results that were observed as a function of stimulus parameters in both experiments are qualitatively consistent with those that have been observed and attributed to auditory streaming in comparable experiments in humans, there remain numerous other parameters whose direct or indirect effects on streaming need to be investigated. For example, it is assumed that selective attention plays a key role in task performance in both of our tasks here, but it is of course difficult to assess precisely the role it plays in determining experimental thresholds. Instead, we are aware that it is extremely difficult, if not impossible, to measure auditory stream segregation without involving some form of selective attention. In audition, as in vision, the ability to attend selectively to certain dimensions or features of a stimulus is closely related to, as well as constrained by, perceptual grouping. Conversely, attention can influence auditory streaming—although the extent to which it does is still debated (see Carlyon, Cusack, Foxton, & Robertson, 2001; Sussman, Horvath, Winkler, & Orr, 2007). Selective attention in frequency or some other sound dimension almost certainly played a role in previous behavioral measures of auditory streaming as well. Nevertheless, we believe that, to the extent that the psychophysical results that have been obtained using comparable stimuli and tasks in humans are related to auditory streaming (which, introspectively, they appear to be), the observation of similar trends in performance as a function of various stimulus parameters in an animal is a good indication that the animal is experiencing a similar perceptual phenomenon.

A second limitation of the current study, which future studies should aim to overcome, relates to the fact that in both of the two tasks that were used in this study, stream segregation was beneficial to performance. It is known that auditory streaming depends on listener's intention or

"attentional set". In particular, the A-B frequency separation required for a listener to experience two streams is smaller if the listener is actively trying to hear two separate streams than if the listener is trying to "hold on" the percept of a single stream for as long as possible (van Noorden, 1975, 1977). Therefore, an important goal for future studies is to measure animals' performance in tasks that promote stream integration, rather than segregation. Such tasks have already been devised and tested in human listeners. In particular, performance in tasks in which listeners have to judge accurately the relative timing of sounds within a sequence appears to be dramatically affected by factors that promote stream segregation, and prevent stream integration. For instance, thresholds for the detection of a shift in the timing of the B tones relative to the temporally adjacent A tones in a repeating AB or ABA sequences have been found to be substantially higher when the A and B tones are widely separated in frequency, and are perceived as separate streams, than when frequency separation is small, and all tones are heard as part of the same stream (Micheyl, Hunter, & Oxenham, 2009; Roberts, Glasberg, & Moore, 2002, 2008; Vliegen, Moore, & Oxenham, 1999).

The development of behavioral tasks, which can be used to induce an animal to experience one of two bi-stable percepts (e.g., hear a sequence of tones as one stream or as two streams), and to indirectly verify that this percept was actually experienced through performance measures, has potentially important implications for studies of the neural correlates of perceptual experience (Logothetis & Schall, 1989). Over the past decade, a rapidly increasing number of studies have been devoted to unraveling the brain basis and neural mechanisms of auditory stream formation in both humans and animals (for reviews, see Carlyon, 2004; Micheyl, Carlyon et al., 2007; Snyder & Alain, 2007). In particular, recordings of neural responses to alternating-tone

sequences similar to those used in studies of auditory streaming in humans have started to reveal potential neural correlates of auditory streaming at the single-unit level (Bee & Klump, 2004; Elhilali et al., 2009; Fishman et al., 2001; Itatani & Klump, 2009; Kanwal et al., 2003; Micheyl, Tian et al., 2005; Pressnitzer et al., 2008). However, the conclusions of these studies are limited by the lack of behavioral data on auditory streaming under directly comparable stimulus conditions, in the same species. Therefore, the two tasks described here could prove particularly useful in investigations into the neural basis of auditory streaming in animals. In particular, one advantageous feature of randomly varying multi-tone stimuli such as those used in experiment 2 is that they can also be used to measure spectro-temporal receptive fields (e.g., Noreña, Gourévitch, Aizawa, & Eggermont, 2006; Noreña, Gourévitch, Pienkowski, Shaw, & Eggermont, 2008).

Chapter 3 Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes

The material contained in this chapter is published as M. Elhilali, L. Ma, C. Micheyl, A.J. Oxenham, and S.A. Shamma. (2009) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron. 61: 317-29. The first three authors contributed equally to this paper. Here, I only included the part contributed by me.

3.1 Introduction

When listening to someone at a crowded cocktail party, or trying to follow the second violin line in a symphonic orchestra, we rely on our ears' and brain's extraordinary ability to parse complex acoustic scenes into individual auditory "objects" or "streams" (Griffiths and Warren, 2004). Just as the decomposition of a visual scene into objects is a challenging and mathematically ill-posed problem, requiring both "top-down" and "bottom-up" information to solve (Marr, 1983; Zeki, 1993), the auditory system uses a combination of acoustic cues and prior experience to analyze the auditory scene. A simple example of "auditory streaming" (Bregman, 1990; Carlyon, 2004) can be demonstrated and explored in the laboratory using sound sequences like those illustrated in Figure 3.1. These sequences are produced by presenting two tones of different frequencies, A and B, repeatedly (Figure 3.1A). Many psychophysical studies have shown that this simple stimulus can evoke two very different percepts, depending on the frequency separation, ΔF, between the A and B tones, and the time interval, ΔT, between successive tones (for a review see Bregman, 1990). In particular, when ΔF is relatively small (< 10%), most listeners perceive and describe the stimulus as a single stream of tones alternating in frequency, like a musical trill.

However, when Δ*F* is large, the percept is that of two parallel but separate streams, each containing only tones of the same frequency (A-A- and B-B-) – see supplementary materials for an auditory demonstration. The perceptual separation of sound components into distinct streams is usually referred to as "stream segregation"; the converse process is variously known as "stream integration", "grouping", or "fusion". Manifestations of auditory streaming have been observed in various non-human species, including birds, fish, and monkeys, suggesting that streaming is a fundamental aspect of auditory perception, which plays a role in adaptation to diverse ecological environments (Bee and Micheyl, 2008; Fay, 1998, 2000; Hulse et al., 1997; Izumi, 2002; MacDougall-Shackleton et al., 1998).

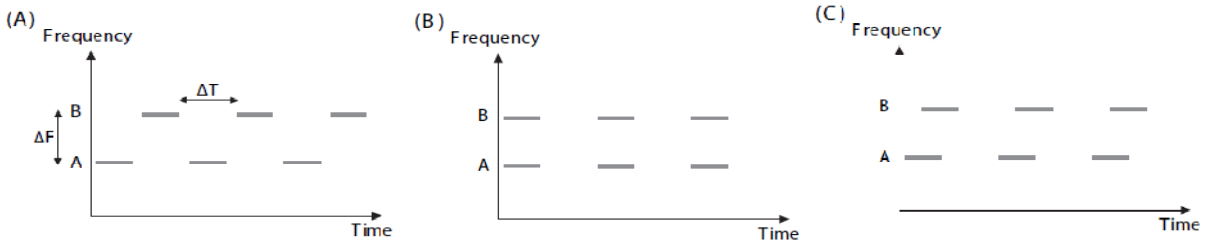

Figure 3.1 Schematic spectrograms of stimuli used to study the perceptual formation of auditory streams. **(A)** The typical stimulus used in the vast majority of psychophysical and physiological studies of auditory streaming: a sequence of tones alternating between two frequencies, A and B. The percept evoked by such sequences depends primarily on the frequency separation between the A and B tones, ΔF, and on the inter-tone interval, ΔT: for small ΔFs and relatively long ΔTs, the percept is that of a single stream of tone alternating in pitch (ABAB); for large ΔFs and relatively short ΔTs, the percept is that of two separate streams of tones of constant pitch (A-A vs. B-B). **(B)** A variation on the traditional stimulus, used in this study: here, the A and B tones are synchronous, rather than alternating. Such sequences usually evoke the percept of a single stream, regardless of ΔF and ΔT. **(C)** An alternating sequence of tones that is *partially*-

overlapped (40 ms onset asynchrony or about 50% overlap). This sequence is usually heard like the non-overlapping tone sequence (Figure 3.1A above).

Inspired by the observation that frequency-to-place mapping, or "tonotopy", is a guiding anatomical and functional principle throughout the auditory system (Eggermont, 2001; Pickles, 1988), current models of auditory streaming rely primarily on frequency separation for sound segregation (Beauvois and Meddis, 1991, 1996; Hartmann and Johnson, 1991; McCabe and Denham, 1997). These models predict that consecutive sounds will be grouped perceptually into a single auditory stream if they activate strongly overlapping tonotopic "channels" in the auditory system. In contrast, sounds that have widely different spectra will activate weakly overlapping (or non-overlapping) channels, and be perceptually segregated (i.e., heard as separate streams). In this way, models based on tonotopic separation can account for behavioral findings that show an increase in perceived segregation with increasing frequency separation (Hartmann, 1991). By additionally taking into account neural adaptation and forward suppression of responses to consecutive tones, these models can also account for the influence of temporal stimulus parameters, such as the inter-tone interval or the time since sequence onset, on auditory streaming (Beauvois and Meddis, 1991, 1996; Bee and Klump, 2004, 2005; Fishman et al., 2004; Fishman et al., 2001; Hartmann and Johnson, 1991; Kanwal et al., 2003; McCabe and Denham, 1997; Micheyl et al., 2007b; Micheyl et al., 2005; Pressnitzer et al., 2008).

Although tonotopic separation is important, it is clearly not the only determinant of auditory perceptual organization. Another factor is the relative timing of sounds. Sounds that start and end at the same time are more likely to be perceived as a single event than sounds whose onsets and

offsets are staggered by several tens or hundreds of milliseconds (Darwin and Carlyon, 1995a). Accordingly, if the AB tone pairs were presented synchronously (as in Figure 3.1B) instead of sequentially (as in Figure 3.1A), they might form a single perceptual stream, even at large frequency separations. This prediction poses a serious problem for purely tonotopic models of auditory streaming. Unfortunately, nearly all perceptual studies of auditory streaming so far have used strictly sequential, temporally non-overlapping, stimuli (Figure 3.1A), although one informal description of an experiment involving partially overlapping stimuli exists (Bregman, 1990, p. 213). On the physiological side, it is unclear how synchrony affects neural responses in the primary auditory cortex (AI), where previous studies have identified potential neural correlates of auditory streaming using purely non-overlapping stimuli (Fishman et al., 2004; Fishman et al., 2001; Gutschalk et al., 2005; Kanwal et al., 2003; Micheyl et al., 2007a; Micheyl et al., 2005; Snyder et al., 2006; Wilson et al., 2007). The complexity of auditory cortical responses makes it difficult to predict how responses of single AI units will be influenced by stimulus synchrony: depending on the position of the tones relative to the unit's best-frequency, responses might be facilitated (i.e. enhanced), inhibited (i.e. reduced), or left unchanged by the synchronous presentation of a second tone within the unit's excitatory receptive field.

Psychoacoustic findings reveal that synchronous and non-synchronous sound sequences are perceived very differently, with synchronous tone sequences heard as a single stream, even at very large frequency separations. Here we present physiological findings, which show that the synchronous and non-synchronous tone sequences evoke very similar tonotopic activation patterns in AI. Together, these findings challenge the current view that tonotopic separation in AI is necessary and sufficient for perceptual stream segregation. More generally, the present

findings suggest that the principle of grouping information across sensory channels based on temporal coherence may play a key role in auditory perceptual organization, just as it is proposed for visual scene analysis (Blake and Lee, 2005)

3.2 Methods

3.2.1 Experimental Design

The stimuli were sequences of A and B tones, where A and B represent different frequencies as illustrated in Figure 3.1. Both alternating (non-overlapping and partially-overlapping) and synchronous sequences were used (see details below). In *Experiment I,* Tones A and B were shifted equally in five steps relative to a unit's best frequency (BF, the frequency which a unit most responds to) as shown in Figure 3.2A, with tone B starting at the BF and tone A ending at the BF. ΔF between the tones was 0.25, or 0.5, or 1 octave, which was fixed within a trial and varied among different trials. The total number of conditions was 45 (5 positions x 3 ΔF x 3 modes). In *Experiment II,* tone A was set at the BF of the isolated unit, while tone B was placed to ±1/3, ±2/3, ±1, ±1.5, and ±2 octaves away from tone A if applicable, as illustrated in Figure 3.2B. The stimuli also included a single tone sequence to measure the frequency tuning of the unit.

In both experiments I and II, each trial included 0.4 sec pre-stimulus silence, 3 sec stimulus length, and 0.6 sec post-stimulus silence. Tone duration was 0.075 sec including 0.005 sec onset and offset ramps, and an inter-tone gap of 0.025 sec in the alternating sequence and 0.125 sec in the synchronous sequence. For the overlapped sequences, the tone onset asynchrony was 40 ms

(i.e., 50% overlap between the tones). All conditions were presented pseudo-randomly 10 times at 70 dB or at about 10 dB above threshold of the isolated units.

3.2.2 Data Analysis

For each unit and each condition, period histogram was constructed from the peri-stimulus time histograms (PSTH) by folded (averaged) responses to the two tones over the duration of the trial from 0.6 to 3 sec after the onset of the stimuli. Examples of such response histograms from a single unit the stimuli of Experiments I are shown in Figure A1.1 and A1.2 in Appendix 1. For each stimulus response, we excluded the first 0.6 sec so as to avoid adaptation effects. The mean firing rate (spikes/sec) was computed by taking the average value of the period histogram (averaged over 0.2 sec). The overall firing rate patterns were obtained by averaging the normalized responses from all isolated units. In order to compensate for inherent differences in the relative strength of tone responses across units, firing rates were first normalized by dividing them by the maximum rate at each ΔF and at each stimulus mode in experiment I and by the mean firing rate at BF in experiment II.

The magnitude of dip was determined according to the following equation:

$$(\text{Side} - \text{Center}) / (\text{Side} + \text{Center}) \%$$

where ‘Side’ is the maximum response at either of the “BF sites” (position 1 or 5); and ‘Center’ is the minimum response at any of the non-BF sites (positions 2, 3 or 4).

To measure the effective bandwidth of interaction between tones, the mean firing rate at the frequency closest to BF (i.e., 1/3 or -1/3 octave) was compared with those at the other frequencies on the same direction (i.e., below BF or above BF). The frequency showed the

significant difference (two-tailed t-test, $P < 0.05$) in mean firing rate from the frequency closest to BF was the effective bandwidth of interaction.

3.3 Results

The psychophysical results raise the question of whether neural responses to sequences of synchronous and sequential tones in the central auditory system differ in a way that can account for their very different percepts. To answer this question, we performed two experiments in which we recorded the single-unit responses in AI to sequences such as those illustrated in Figure 3.1 in the awake (non-behaving) ferret. In the first experiment, we explored directly the extent of *segregation* between the responses to the A and B tones. In the second experiment, we assessed the range of frequencies over which the tones *interacted* (or mutually influenced their responses).

3.3.1 Experiment I: Segregation Between Two-tone Responses

This experiment examined the distribution of responses to the two tones by translating them *together*, relative to the best frequency (BF) of an isolated single unit in AI of awake ferrets in five steps (labeled 1 - 5 in Figure 3.2A), where positions 1 and 5 correspond to one of the two tones being at BF of the unit. The frequency separation (ΔF) between the tones in each test was fixed at 1, 0.5, or 0.25 octaves, corresponding to 12, 6, and 3 semitones, respectively. As described above, alternating tone sequences are usually perceived as two streams at separations of 12 and 6 semitones (1 or 0.5 octaves), but are only marginally segregated at a separation of 3 semitones (0.25 octaves). In contrast, synchronous tone sequences are always heard as one stream. Therefore, if the “spatial segregation” hypothesis were valid, alternating sequences

should evoke well-segregated neural responses to the far-apart tones (1 and 0.5 octaves), whereas synchronous sequences should evoke spatially overlapping responses in all cases.

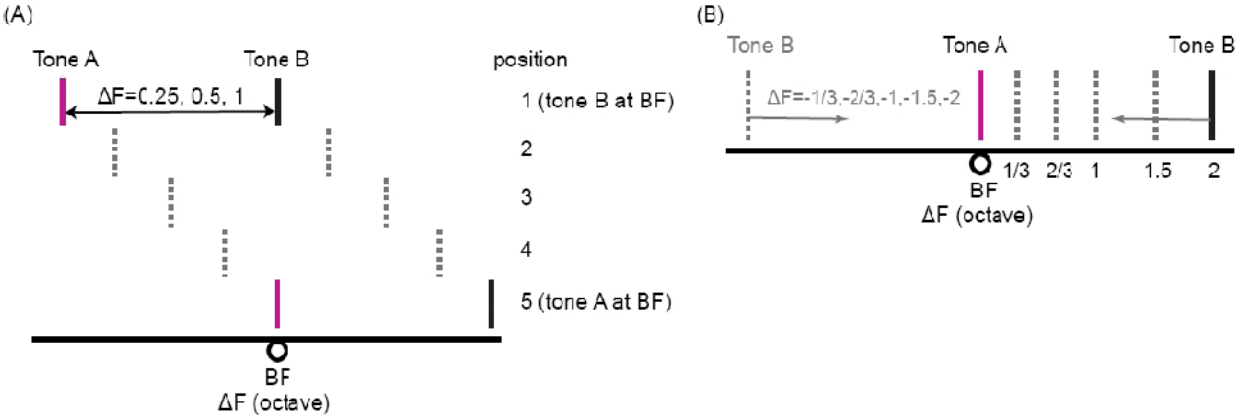

Figure 3.2 Schematic of the tone frequencies and conditions used in the physiological experiments. Both alternating and synchronous tone sequences were tested in all conditions. (A) Experiment I: The two-tone frequencies we held fixed at one of three intervals apart (ΔF = 0.25, 0.5, 1 octaves), and they were then shifted through five equally spaced positions relative to the BF of the isolated cell. (B) Experiment II: Tone-A is fixed at the BF of the isolated unit, while tone-B is shifted closer to BF in several steps.

The results from a population of 122 units in the AI of 4 ferrets are shown in Figure 3.3. In Figure 3.3A, the average rate profiles for the synchronous, overlapping, and alternating presentation modes are constructed from the responses as described in the Methods. All 122 units were tested with the synchronous and alternating modes; 75/122 units were also tested with the overlapping sequences. When the tones are far apart (ΔF = 1 octave; right panel of Figure 3.3A), responses are strongest when either tone is near BF (positions 1 and 5); they diminish considerably when the BF is midway between the tones (position 3), suggesting relatively good spatial separation between the representations of each tone. When the tones are closely spaced (ΔF = .25 octave; left panel of Figure 3.3A), the responses remain relatively strong at all

positions, suggesting that the representations of the two tones are not well separated. More importantly, the average rate profiles are similar for all presentation modes: in all cases the responses are well-segregated with significant dips when the tones are far apart (ΔF = 1 octave), and poorly separated (*no* dips) when the tones are closely-spaced (ΔF = 0.25 octaves). Thus, based on average rate responses, the neural data mimic the perception of the asynchronous but not the synchronous tone sequences. Therefore, the distribution of average rate responses does not appear to represent a general neural correlate of auditory streaming.

Instead of averaging the responses from all cells, we tabulated the number of cells indicating a significant segregation in the responses (implying a percept of 2 streams) or no segregation (a percept of 1 stream) by examining whether a significant dip occurred in each cell's profile during the two extreme presentation modes (synchronous *versus* alternating tones). The determination of a dip was derived for each condition by finding a significant difference (one-tailed t-test; $P < 0.025$) between the distributions of the maximum response at either of the "BF sites" (1 or 5) compared with the minimum response at any of the non-BF sites (2,3, or 4). For the purposes of this analysis, we used a population of 66 units for which positions 1 or 5 were "BF sites", and measurements were completed at all positions (1-5). In most experiments, several units with diverse BFs were recorded simultaneously with multiple electrodes, and hence it was only possible to match the tone frequencies to the BF of one or two of the cells. The percentage of cells with a significant dip in their profiles is shown in the histograms of Figure 3.3B. We also calculated the magnitude of the dip (see Method) for each unit, and established that there was no significant difference in neural responses between synchronous and alternating modes (two-tailed t-test, $P = 0.54$ at 0.25 octave, $P = 0.37$ at 0.5 octave, and $P = 0.42$ at 1 octave), and that

spatial segregation increases significantly with increasing ΔF (one-tailed t-test, shown in Figure 3.3B). The results show that (1) segregation is strongest at 1 octave separation, and weakest at 0.25 octaves, and that (2) there is little difference between the patterns of responses to the synchronous and alternating sequences. Thus, this alternative *individual-cell* response measure also fails to predict the different streaming percepts of the alternating and synchronous tones.

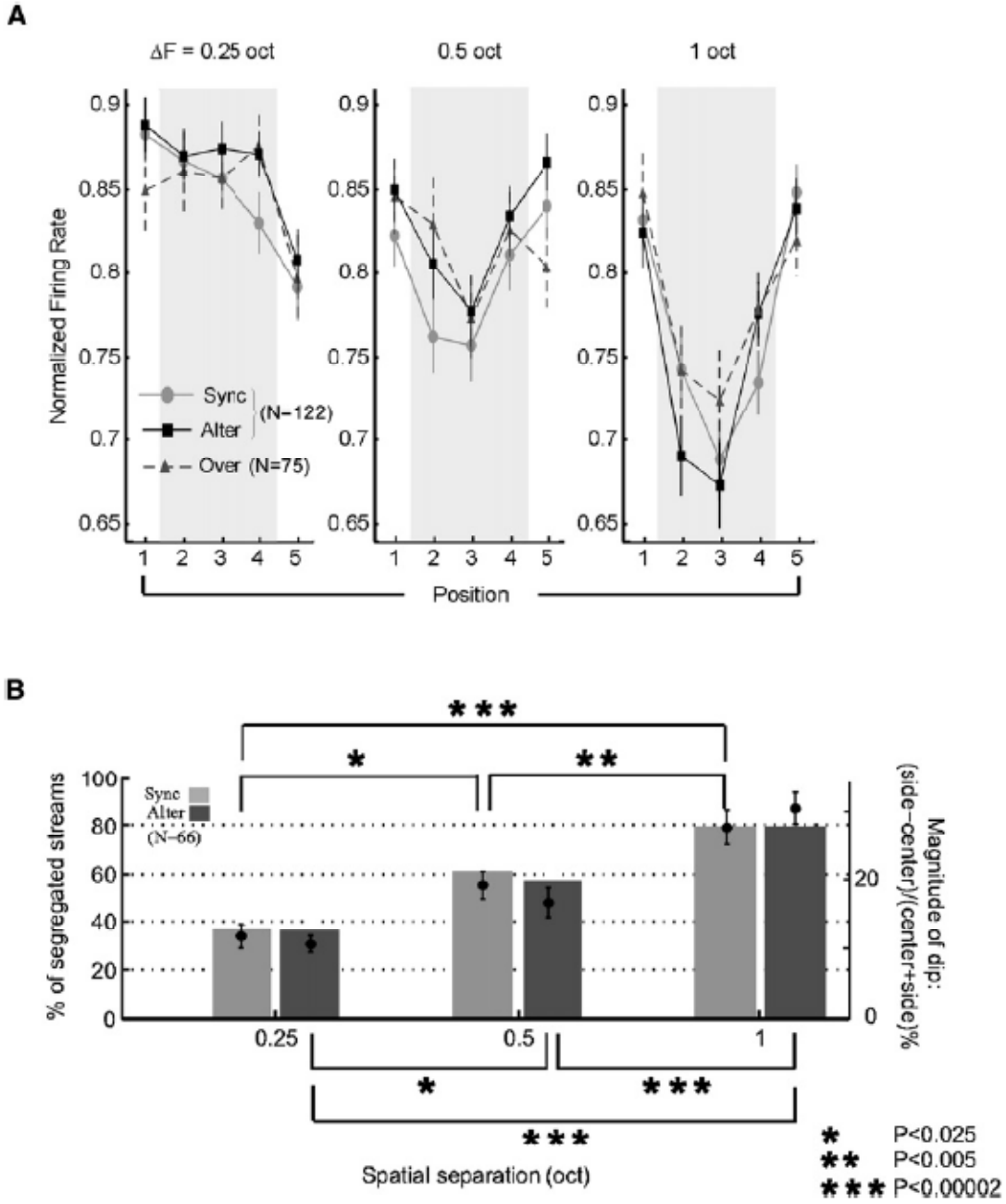

Figure 3.3 Responses of single-units to alternating (non-overlapping and partially-overlapping) and synchronous two-tone sequences at three different intervals (□F = 0.25, 0.5, 1 octaves). The two-tones were shifted relative to the BF of the cell in five equal steps, from tone-B being at BF (position 1) to tone-A at BF (position 5), as described in Experiment I paradigm. (A) Average firing rates from a total of 122 single-units in the five frequency positions in the synchronous and non-overlapping modes. Overlapping tones were tested in only 75/122 units. Responses in all presentation modes exhibited a significant dip in response when tones were further apart (0.5 and 1 octaves), and neither was at BF (shaded positions 2-4). (B) The percentage of cells that exhibited a significant dip in their responses were similar in the two extreme presentation modes (synchronous and non-overlapping alternating). Only the 66 single-units that were tested at all five positions were included in this analysis (since responses from all positions are necessary to compile such histograms). The magnitude of dip showed significant difference across □F, but nonsignificant difference across presentation mode.

3.3.2 Experiment II: Frequency Range of Interactions

The key question of interest in this experiment was whether the range of interactions between the two tones was significantly different in the three presentation modes (alternating, overlapping, or synchronous). We measured the frequency range of interactions between the two tones by fixing tone A at the BF of the isolated unit, while placing tone B at ±1/3, ±2/3, ±1, ±1.5, and ±2 octaves around the BF (Figure 3.2B). We also estimated the unit's frequency tuning by measuring the iso-intensity response curve with a single tone sequence (black curve in Figure 3.4A). Other methodological details can be found in the Methods.

The average spike counts are shown in Figure 3.4A from a population of 64 single units (in the synchronous and alternating modes) and 41 units (overlapping mode) that were recorded separately from Experiment I. All data were combined by computing the iso-intensity response curve of each unit, centering it around the BF of the unit, and normalizing it by the response of the unit to the single BF tone. We then kept only the half of the tuning curve above or below the BF from which the full 2 octave range was tested. Such (half-tuning) curves from all units were then averaged for each condition. The results highlight the interactions observed as the tones approached each other in frequency. For instance, when tone B was far from tone A at BF (e.g., at ± 2 octaves), the effects of the B tone on the cell are relatively small and the firing rate in all modes was similar to that of the single tone at BF (the normalized rate of 1, indicated by the dashed line). As tone B approached BF, the responses become modulated, first decreasing and then increasing steeply beyond about 1 octave on either side of the BF. Apart from differences in absolute firing rates, the pattern of interactions was similar in all three presentation modes. For example, the frequency separations at which significant interactions ensue are similar, implying that the units' receptive fields (or their tuning curves) are similar whether they are driven by synchronous, alternating, or partially overlapping sequences.

To further quantify the population responses, we computed the effective bandwidth of interactions for each unit, defined as the furthest frequency on either side of the BF at which response interactions between the two tones were significant (see Methods). The data from all units in the synchronous and alternating (non-overlapping) modes are displayed in the histogram of the differences between the two measured ranges in Figure 3.4B. The scatter is mostly symmetric, with a mean not significantly different from zero (two-tailed t-test, P = 1). Hence, the

bandwidth differences for individual units fail once more to account for the different streaming percepts evoked by the alternating and synchronous presentation modes. Similar comparisons were also performed for the overlapping vs. synchronous and overlapping vs. alternating modes. The bandwidth differences in both cases were also mostly symmetric, with a mean not significantly different from zero.

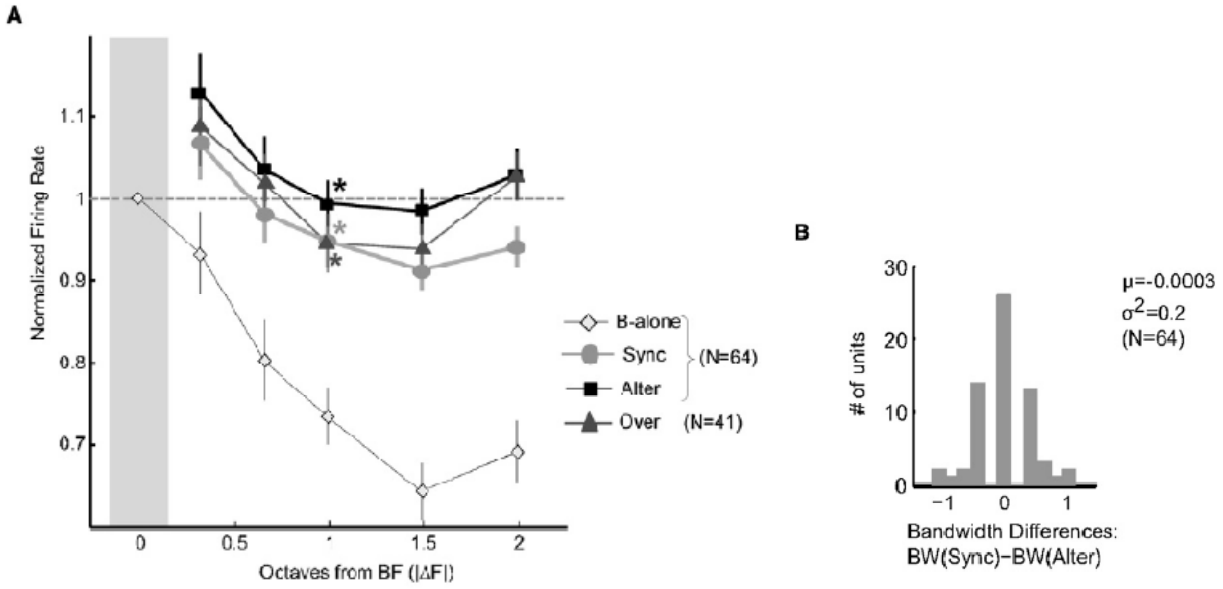

Figure 3.4 Averaged responses from a total of 64 units tested for alternating, synchronous and overlapping (tested in only 41/64 units) sequences using the paradigm of *Experiment II*. **(A)** The tuning near the BF averaged from all units. The average iso-intensity response curve is shown in black for comparison. To increase the number of cells included in the average, we folded the responses from above and below BF, but included only units that were tested with the entire 2 octave range from BF. All presentation modes show some suppression of responses as tone-A approaches the BF (1 to 1.5 octaves), and a significant increase closer to BF (about 1 octave; marked by the asterisks). **(B)** Histogram of the difference in bandwidth of interactions between the tones during the two extreme presentation modes (synchronous and alternating) is roughly symmetric indicating no systematic bias in the scatter.

3.4 Conclusions

The results from the two physiological experiments in awake ferrets contradict the hypothesis that segregation of AI responses to two-tone sequences is sufficient to predict their perceptual streaming. Instead, our findings reveal that synchronous and non-synchronous sequences do not differ appreciably in the spatial representations of their temporally-averaged responses in AI, despite the substantial differences in their streaming percepts. Clearly a model that is successfully able to predict perception from these neural data will need to incorporate the time dimension.

3.5 Discussion

3.5.1 Evidence against a purely tonotopic or "spatial" model of auditory streaming

We examined the hypothesis that acoustic stimuli exciting spatially segregated neural response patterns are necessarily perceived as belonging to different perceptual streams. This "spatial" hypothesis underlies (explicitly or implicitly) previous interpretations of the neural correlates of streaming in the physiological investigations and the computational models of streaming (Beauvois and Meddis, 1991, 1996; Bee and Klump, 2004, 2005; Fishman et al., 2004; Fishman et al., 2001; Hartmann and Johnson, 1991; Kanwal et al., 2003; McCabe and Denham, 1997; Micheyl et al., 2007b; Micheyl et al., 2005; Pressnitzer et al., 2008). One of the elegant aspects of the "spatial" hypothesis is that it can be generalized to predict that separate streams will be perceived whenever sounds evoke segregated responses along any of the representational dimensions in the auditory cortex, including not just the tonotopic axis, but also a fundamental-frequency (F0) or "virtual pitch" axis (Bendor and Wang, 2005, 2006; Gutschalk et al., 2007), as well as, perhaps, temporal- and spectral-modulation-rate axes (Bendor and Wang, 2007;

Kowalski et al., 1996a, b; Schreiner, 1998; Schreiner and Sutter, 1992; Sutter, 2005; Versnel et al., 1995), thereby accounting for psychophysical findings of stream segregation induced by differences in F0 or modulation rate in the absence of tonotopic cues (Grimault et al., 2002; Roberts et al., 2002; Vliegen and Oxenham, 1999).

However, the experimental data reported here cast doubt on the validity of an explanation of auditory streaming in terms of neural-response separation that ignores temporal coherence as an important determinant of perceived segregation. Our human psychophysical results show very different perceptual organization of synchronous and asynchronous tone sequences, whereas the extent of segregation of the neural responses in ferret AI was essentially independent of the temporal relationships within the sequences. This finding emphasizes the fundamental importance of the temporal dimension in the perceptual organization of sound, and reveals that tonotopic neural-response separation in auditory cortex alone cannot explain auditory streaming.

3.5.2 A spatio-temporal model of auditory streaming

Our alternative explanation augments the spatial (tonotopic) segregation hypothesis with a temporal dimension. It is a *spatiotemporal* view, wherein auditory stream segregation requires both separation into neural channels *and* temporal incoherence (or anti-coherence) between the responses of these channels. This *spatiotemporal* hypothesis predicts that if the evoked neural responses are temporally coherent, a single stream is perceived, regardless of the spatial distribution of the responses. This prediction is consistent with our psychophysical findings using synchronous tone sequences. The prediction is also consistent with the introspective observation, confirmed in psychophysical studies, that synchronous spectral components

generally fuse perceptually into a single coherent sound (e.g., a vowel or a musical chord), whereas the introduction of an asynchrony between one and the other components in a complex tone results in this component "popping out" perceptually (Ciocca and Darwin, 1993).

The present demonstration of a critical role of temporal coherence in the formation of auditory streams does not negate the role of spatial (tonotopic) separation as a factor in stream segregation. The extent to which neurons can signal temporal incoherence across frequency is determined in large part by their frequency selectivity. For example, the responses of two neurons tuned to the A and B tones in an alternating sequence (Figure 3.1A) can only show anti-coherence if the frequency-selectivity of the neurons is relatively high compared to the A-B frequency separation. If the neurons' frequency tuning is broader than the frequency separation, both neurons are excited by both tones (A and B), and respond in a temporally coherent fashion. In this sense, spatial separation of neural responses along the tonotopic axis may be *necessary* for stream segregation but, as this study shows, it is not *sufficient*.

The principle of channel coherence can be easily extended beyond the current stimuli and the tonotopic frequency axis to include other auditory organizational dimensions such as spectral shape, temporal modulations as well as binaural cues. Irrespective of the nature of the dimension explored, it is the temporal coherence between the responses along that dimension that determines the degree of their integration within one stream, or segregation into different streams.

Finally, there are interesting parallels between the present findings, which suggest an important role of temporal coherence across sensory channels in auditory scene analysis, and findings in other sensory modalities such as vision, where grouping based on coherence of temporal structure has been found to provide an elegant solution to the binding problem (e.g., (Alais et al., 1998; Blake and Lee, 2005; Fahle, 1993; Treisman, 1999). Together, these findings suggest that although the perceptual analysis of visual and auditory "scenes" pose (at least, superficially) very different problems, they may in fact be governed by common overarching principles. In this regard, parallels can be drawn between prominent characteristics of auditory stream formation, such as the buildup of streaming and its dependence on frequency separation, and processes involved in the visual perception of complex scenes.

3.5.3 Do percepts of auditory streams emerge in or beyond primary auditory cortex?

For neural activity in AI to be consistent with the psychophysical observation that synchronous tones with remote frequencies are grouped perceptually while alternating tones are not, there should be cells in AI whose output is strongly influenced by temporal coherence across distant frequencies. While such cells are likely to be present in AI (Barbour and Wang, 2002; Kowalski et al., 1996b; Nelken et al., 1999), we did not systematically find many that reliably exhibited the properties necessary to perform the coincidence operation. For example, all neurons sampled in this study followed the temporal course of the stimuli (with increased firing rates during epochs where at least one tone was present), the responses did not unambiguously increase in the presence of temporal coherence across tonotopic channels. Therefore, one possibility is that the percepts of stream segregation and stream integration are not determined in AI. Another possibility is that the coincidence and subsequent matrix decomposition described in the model

are realized in a different, less explicit, form. For instance, it is theoretically possible to replace the spectral decomposition of the coherence matrix by a singular-value-decomposition directly upon the arrayed cortical responses. The spectral decomposition of the coherence matrix is equivalent to PCA analysis of the covariance matrix of the channel responses. Equivalent results can be computed by a singular value decomposition directly on the channel temporal responses, i.e., without computing the covariance matrix, obviating the need for the coincidence detectors. This leaves open the question of how and *where*, in or beyond AI, the detection of temporal coincidences across remote frequency channels is neurally implemented (Nelken, 2004).

The auditory streaming paradigm, with its relatively simple and well-controlled stimuli and extensively characterized percepts, may in fact provide an excellent vehicle to explore a broader issue in brain function: that of the relationship between perception and neural oscillations, which reflects coherent responses across different regions in the brain. *Coherence* as an organizing principle of brain function has gained prominence in recent years with the demonstration that it could potentially play a role in mediating attention (Liang et al., 2003; Zeitler et al., 2006), in binding multimodal sensory features and responses (Lakatos et al., 2005; Schroeder et al., 2008), and in giving rise to conscious experiences (Fries et al., 1997; Gross et al., 2007; Meador et al., 2002; Melloni et al., 2007). Our results reinforce these ideas by emphasizing the importance of temporal coherence in explaining auditory perception. Specifically, the inclusion of the time dimension provides a general account of auditory perceptual organization that can in principle deal with any arbitrary combinations of sounds of time and frequency.

3.5.4 Attention and the neural correlates of streaming

Interpretations of neural responses recorded in *passive* animals as "correlates" of auditory percepts are necessarily speculative, since behavioral measures of the animal's percepts during the recordings are not available. Under such conditions, the experimenter can, at best, assert that the neural responses differ across experimental conditions (e.g., different stimuli) in a way that is consistent with behavioral measurements obtained in the same (or a different) animal (or species) under similar stimulus conditions. In this respect, the present study suffers from the same limitation as previous investigations of the neural basis of auditory streaming in awake animals that were either passive (Bee and Klump, 2004, 2005; Fishman et al., 2004; Fishman et al., 2001; Kanwal et al., 2003), or engaged in a task unrelated to streaming (Micheyl et al., 2005).

The possibility remains that AI responses to alternating and synchronous tone sequences in awake animals that are engaged in a task, which requires actively attending to the stimuli, might be substantially different from those recorded in passive animals. It is known that neural responses in AI are under attentional control, and can change rapidly as the task changes (Fritz et al., 2005a; Fritz et al., 2003; Fritz et al., 2005b). . Such attentionally driven changes in receptive fields might differentially affect the neural responses to alternating tones and those to synchronous tones, in a way that makes these responses more consistent with the percepts evoked by those sequences (Yin et al., 2007). On the other hand, the aspects of streaming investigated here – in particular the increased segregation with increasing frequency separation in asynchronous conditions – have been posited to be "automatic" or "primitive" and hence independent of attention (Macken et al., 2003; Sussman et al., 2007), although the matter is still debated (Carlyon et al., 2001).

The possible effects of attention could be investigated in future studies by controlling the attentional and behavioral state of the animal. Our model postulates the existence of units that should exhibit a dependence on temporal coherence. We have not found such units in AI, and therefore a future search may concentrate more fruitfully on other, supramodal, areas, such as the prefrontal cortex, where attentional modulation of AI responses may originate (Miller and Cohen, 2001).

Chapter 4 A neurophysiological Evidence of Temporal Correlation during Auditory Streaming in Ferret Primary Auditory Cortex

4.1 Introduction

Inspired by experimental results from Chapter 3, we postulate that temporal correlation across auditory channels, and not the tonotopic separation per se, is the key neural correlate of auditory streaming. According to this idea, cells which are simultaneously activated by one sound source (one stream), have coherent spiking activity and distinguish themselves from those activated asynchronously by other sound sources. This hypothesis requires that synaptic coupling strength between pairs of cells can be modulated at relatively faster timescale than conventionally assumed. That is, if the spiking activity between the two cells is synchronous, the synaptic strength is increased; if the spiking activity between the two cells is asynchronous, the synaptic strength is decreased. All are happening rapidly within a few seconds, or a fraction of a second.

Recently, rapid task-related plasticity of spectrotemporal receptive fields (STRFs, STRF of a neuron is the linear filter that, when convolved with the auditory spectrogram of an arbitrary stimulus, gives a linear estimate of the evoked firing rate) has been demonstrated in primary auditory cortex in a series of experiments by Fritz et al (2003, 2005, 2007a, 2007b). In these experiments, animals were trained to discriminate multiple spectral tasks with different tonal targets from a sequence of temporally orthogonal ripple combinations (TORCs), which are broadband noise bursts. In the single- (complex-) tones detection task, STRFs were enhanced at the target tone frequency (frequencies). In the two-tone discrimination task, an equally selective suppression at reference tone frequency was found in addition to the same change seen in the

single-tone detection task, the selective enhancement at target tone frequency. In all these tasks, the target tone was placed near a cell's receptive field and best frequency (BF). Furthermore, the task was modified to achieve a range of task difficulties (Atiani et al., 2009). The target tone was embedded in a TORC with different signal-to-noise ratio (SNR). When the target tone fell near the cell's BF at high SNR, the same enhancement was observed during behavior. When the tone was placed far from a cell's BF and receptive field at high SNR, or at low SNR, the STRF change became suppressive. This STRF plasticity can occur quite rapidly and fade quickly after task completion; in some cases, it persisted for minutes or hours.

In the current study, we recorded spiking activity with multiple electrodes while the same behaving animal described in Chapter 2 performed the streaming task of detecting regularly repeating target tones in a random multi-tone background (maskers/distracters).We placed either target or masker tones near a cell's BF. We called it a masker/target cell if masker/target tones were near the cell's BF (see Method for details). Following the above arguments, we hypothesized that (1) during streaming in behavioral contexts, the temporal correlation between spiking activity from pairs of simultaneously recorded target cells (or simultaneously recorded masker cells) would increase since these cells will be in the same stream; and (2) cells driven by different streams (target or maskers) would have decreased or no changed correlated spiking activity, e.g., as from pairs of simultaneously recorded target and masker cells. Clearly, the STRF gain or shape changes should reflect these changes in correlation. Since in appetitive tasks, reference stimuli are associated with the "warning" sounds when animals do not lick the spout, we anticipated that these responses would be enhanced (David et al, 2008). Therefore we predicted that for reference stimuli during behavior, the STRFs of masker cells would be

enhanced at BFs, while the STRF changes of target cells would become suppressive (David et al, 2008; Atiani et al., 2009).

4.2 Methods

4.2.1 Behavioral Task and Stimuli

The behavioral task was the same task as in experiment 2 in Chapter 2. The example spectrogram of the stimuli is shown in Figure 4.1 below and the description was repeated here for convenience and completion. Some parameters were changed slightly for consideration of neurophysiology data analysis. On each trial, a sequence consisting of multiple tone pips with random frequencies and random onset times ("maskers") were presented. At some point in this random sequence, a regularly repeating sequence of constant-frequency tones ("targets") was introduced. The task of the ferret was to detect the target sequence amid the randomly varying masker tones. The animal was trained to withhold licking until the target was introduced, and to start licking upon detecting the target. If a lick response occurred within 100 to 1360 ms after the onset of the first target tone, it was counted as a hit and reinforced with a 1/3 ml of water. Misses had no consequence. In this experiment, there were no "sham" trials; the target tones were presented on all trials. However, the start time of the target sequence varied randomly between 900 and 2700 ms after the onset of the masker sequence. When the animal produced a lick response before the onset of the target sequence, this was counted as a false alarm, the trial was aborted, and followed by a short timeout.

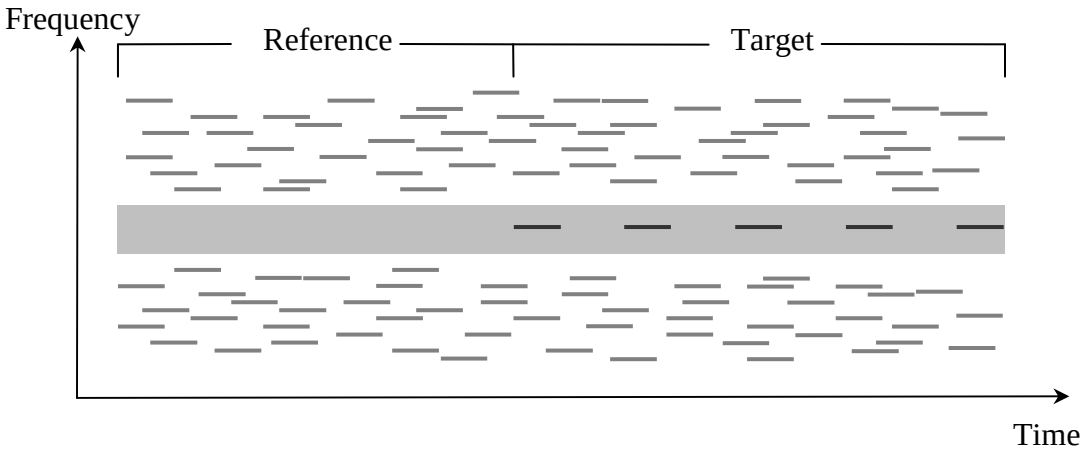

Figure 4.1 Schematic spectrogram of an example stimulus presented on a trial in Experiment 2. During the "reference" portion of the stimulus, only masker tones (gray bars) with random frequencies and onsets times were presented. During the "target" part, target tones (dark bars) repeating regularly at a constant frequency were introduced. The gray area around the target represents the "protected zone" (PZ), within which masker tones were not allowed to fall (See text for additional details).

The stimulus details were as follows. Each tone-pip (target or masker) was 70 ms long, including 5 ms onset and offset ramps. On each trial, 7 target tones were presented. Consecutive target tones were separated by a silent gap of 110 ms, yielding a repetition rate of about 5.6 Hz. Trial length varied randomly between 2.16 and 3.96 s across trials. These durations include the variable-length "reference" sequence (0.9 to 2.7 s) plus the fixed-duration "target" sequence. The masker tones occurred at an average rate of 89 tones per sec. The masker tones were generated as follows: first, eight different masker-tone frequencies were selected at random for every 90 ms; then, the masker tones were shifted pseudo-randomly in time, in such a way that they were not synchronous with the target, except by chance. The masker tone frequencies were drawn at random from a fixed list of values spaced one ST, approximately 6%, apart, excluding a

"protected zone" (PZ) around the frequency of the target tones. The half-width of the PZ determines the minimum allowed frequency separation between the target and the closest masker component on either side. Three half-widths were tested: small (6 ST), medium (9 ST), and large (12 ST). Masker frequencies were selected from within a two-octave range on both sides of the PZ. Target intensity and trial lengths varied within a session. The masker tones were presented at 50 dB SPL (each). Target-tone levels of 0, +4, +8 and +12 dB relative to the level of the masker tones were tested.

4.2.2 Neurophysiological Recording

To secure stability for electrophysiological recording, a stainless steel headpost was surgically implanted on the ferret's skull. Recordings were conducted in a double-walled, sound-attenuation chamber (IAC). Small craniotomies (1-2 mm in diameter) were made over A1 prior to recording sessions which lasted 6-8 hours. We used 2-4 independently moveable tungsten electrodes separated by ~500 µm from their nearest neighbor and AlphaOmega recording system. The range of best frequencies (BFs) in a given experiment varied from 0-2 octaves. During recording sessions, we stored all waveforms from each electrode. Offline, we sorted the multiunit waveforms into different single units using principal component analysis and rejected waveforms corresponding to movement artifacts, for example, licking. In this way, single units, typically 1-3 neurons per electrode, were isolated. This yielded typically 3-10 single units per recording, allowing us to examine firing synchrony (correlation) between each pair of units. Spikes were obtained by triggering at a level four standard deviations above baseline variation in the raw waveform.

On each recording day, electrodes were slowly advanced until we had isolated cells on all electrodes. Then, to assess the BFs and frequency tunings of the A1 neurons that we recorded, we measured the neuronal responses to TORCs and random tones. To make sure the level of BF tone was presented above the neuron's threshold, level tuning was obtained by presenting BF tone at different loudness. Finally, primary sets of task-related stimuli were presented with behavioral and passive conditions. Each set of stimuli is composed of pre (only reference stimuli in Figure 4.1 presented), task stimuli (both reference and target presented) and post (reference only). In the behavioral condition, the animal was required to perform the task, while in the passive condition, the animal just passively listened to the stimuli and no behavior was required. Either PZ or target frequency was changed across each set.

4.2.3 Data Analysis

All false alarm trials were excluded from data analysis. Spiking data were divided into two channels (groups) according to the position of each unit's BF. Units were labeled as target cells (channels) if the unit's BF was within the PZ and as masker cells (channels) otherwise.

4.2.3.1 Correlation Analysis

In order to measure cross-correlation (coherence) between each pair of single units recorded simultaneously from multiple electrodes, we computed the spike-triggered average correlation (STAC) of the spiking activity, also called cross-correlation histograms (CCHs), for all pairs of spike recordings at ±300 ms around all spikes recorded for each condition. Each STAC (or CCH) was normalized by the corresponding number of trigger spikes. If two units are activated synchronously, the spikes add up during the spike triggered averaging process, resulting in a

peak at STAC. On the other hand, if spiking activities between the pair have no reliable temporal relation, the spikes average out during the STAC process, resulting in a flat STAC. This trial-to-trial STAC includes both signal and noise correlations. To measure signal correlation, we computed the post-stimulus time histogram (PSTH) of each unit to the stimuli and computed the STAC of the PSTH. The difference between the signal correlation and the trial-to-trial correlation is noise correlation.

4.2.3.2 STRF Analysis

STRF for each unit was computed by reverse-correlating the spike responses with the spectrogram of the stimulus and then normalized by the autocorrelation of the stimulus (deCharms et al., 1998; Theunissen et al., 2000). The predictive power of computed STRF was estimated by calculating the correlation between the actual and the predicted responses to novel stimuli from the same ensemble. Each STRF is associated with a predictive power and those with a predictive power < 0.15 were excluded from further analysis.

To measure the population effect of the steaming related task on STRF change, we computed the difference between behavioral and passive STRFs ($STRF_{diff}$) for each unit. After normalizing each $STRF_{diff}$ by their individual r.m.s. power, we located the maximum point of each $STRF_{diff}$ in a band ±4 ST around the BF of the cell and within the first 40 ms of the STRF. Each $STRF_{diff}$ was aligned at the local maximum points and the average $STRF_{diff}$ was obtained for each condition. To quantify the $STRF_{diff}$ for each unit, we computed a local STRF change (ΔA_{local}). We defined the local difference as the average difference within ±4 ST and ±25 ms around the local maximum points.

4.3 Results

We recorded spiking activity from one ferret's A1, the same ferret from whom we collected behavioral data for Experiment 2 in Chapter 2. On most recording days, the ferret performed two sessions of the repeating tone sequence detection task. We refer to this condition with attention as behavioral condition. In between behavioral conditions, the similar stimuli with either different PZ or target frequency were presented and the animal was just passively listening. We refer to this condition without behavior as passive condition and compared pooled single units responses between the two attentional conditions.

4.3.1 Temporal Correlation

An example of STACs from four simultaneously recorded units under behavioral condition is shown in Figure 4.2. These are STACs from responses of two masker cells and two target cells at PZ = 6 ST. Masker/Target cells were those in which the BFs were near masker/target tones. The STACs of the spike trains between two masker/target cells shows a peak at 0 time relative to trigger spikes indicating synchronous firing activity between these two units; while STACs between the masker and target cells are relatively flat that indicates uncorrelated firing activity between these pairs of cells. Comparing the STACs between two masker cells during behavior versus the passive state, we note that the peak of the STAC during the reference stimuli becomes larger in the behavior condition compared to the pre-behavior period, indicating an increase in synchronous firing activity between these units during task performance. Comparing the STAC during pre-task, the correlation between target cells shows no change for reference stimuli during behavior. By contrast, there was no change in correlation between the masker and target cells

either during pre-task or task stimuli. These cells, as we discussed earlier, were driven by different stimulus streams, and had more separated STRFs than the other pairs.

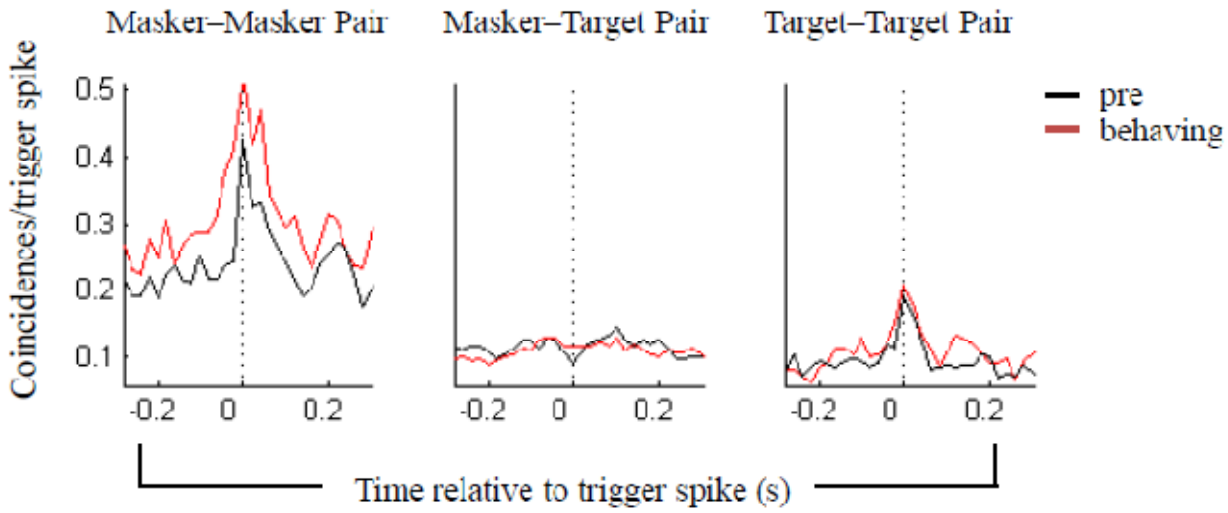

Figure 4.2 STACs of spike trains from simultaneously recorded four units: two masker cells and two target cells at PZ = 6 ST. STACs were computed from responses to reference stimuli. Left panel: STACs between pairs of masker cells; middle panel: STACs between pairs of masker and target cells; and right panel: STACs between pairs of target cells.

According to the behavioral performance in Chapter 2, at smaller PZ, we found significantly increased thresholds comparing with those at larger PZs, which indicates the perceptual difference under these two stimulus configurations. Therefore, to measure the population effect of the streaming related task on correlation changes, we pooled data for median and large width of PZ (PZ = 9 and 12 ST) and examined the correlation between spike trains to reference/target stimuli across pairs of cells under two attentional conditions. During the reference stimuli in the passive state (i.e., the pre and post-task conditions in Figures 4.3a and 4.3b), the STACs are essentially similar indicating that there was little persistence of any changes that might have occurred during the task. By contrast, during behavior, most STACs from reference responses

displayed increased pair-wise correlations relative to the passive conditions. These behavioral effects can be seen clearer in Figure 4.4 where we plot the difference between STACs ($STAC_{diff}$) during reference stimuli in the passive and behaving conditions. These data reveal a tendency of masker cells to have significantly more correlated firing during behavior (t test, $p < 0.01$) compared to the passive condition. By contrast, the correlations between masker and target cells, and target and target cells show no large or significant differences between the two behavioral conditions. Interestingly, there is also a significant peak at 0-lag in the larger PZ condition, whereas the increased correlation in the other case is primarily over all lags, indicating that it reflects a common increase in firing rates for the cell pairs. Similar results for $STAC_{diff}$ from behavior versus post-behavior are displayed in Figure A2.1 in Appendix 2.

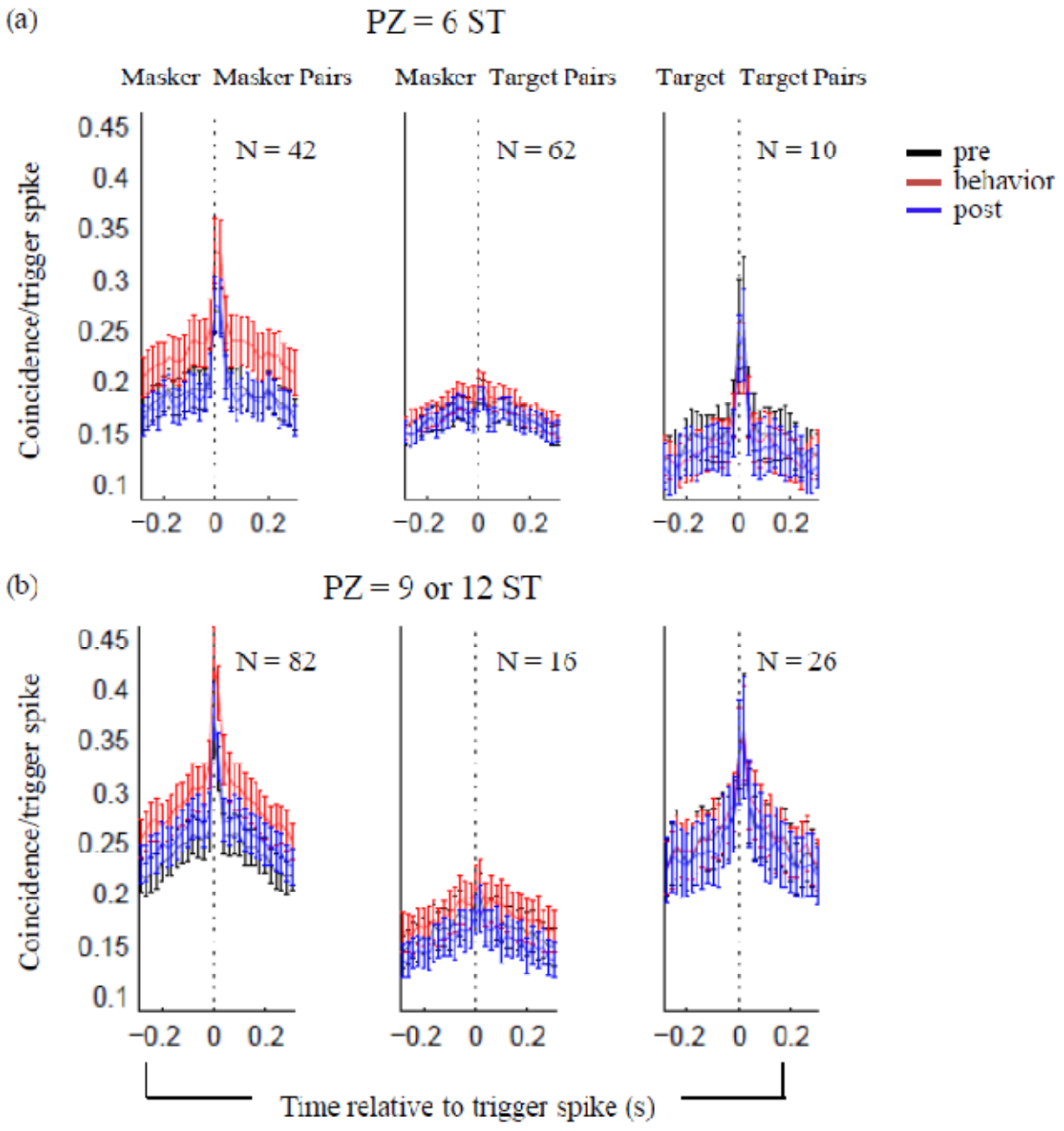

Figure 4.3 STACs from responses to reference stimuli (a) for smaller PZ, 6 ST and (b) for larger PZ, 9 or 12 ST. Error bars indicate standard error (SE).

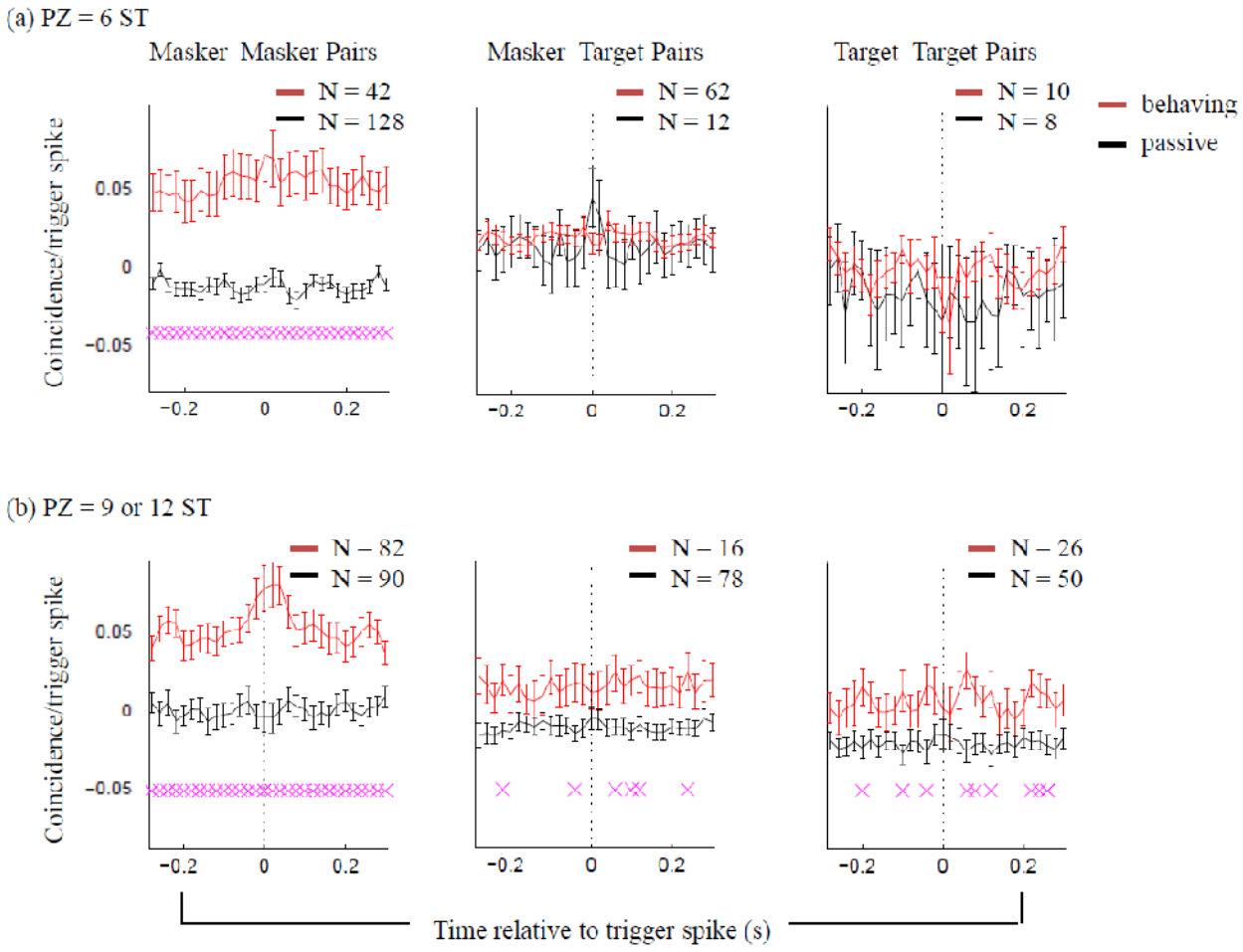

Figure 4.4 The difference between STACs ($STAC_{diff}$) from behavior versus pre-behavior conditions, respectively. $STAC_{diff}$ (a) at smaller PZ (6 ST); and (b) at larger PZ (9 or 12 ST). Magenta crosses indicate significant difference (t test, $p < 0.01$) at each time-lag between two conditions. Error bars indicate SE.

When target stimuli commence, the animal perceives two streams, target and maskers, and hence the responses during this period reflect its perception of the two streams and may demonstrate directly the changes related to auditory streaming (Figure 4.5). For the larger PZ condition, we found that the correlation among target cells increases significantly (t test, $p < 0.01$) during behavior compared to the passive state. Smaller increases are also seen among the masker cells

(the other stream). Interestingly, no such increases occur between target and masker cells. Consequently, one can say that in this condition, cells driven by the same stream experience an enhanced correlation of firing, but not the ones across streams. The same pattern is difficult to discern in the PZ = 6 ST cases because of the small number of target pairs recorded. But note specifically the significant *decrease* (t test, $p < 0.01$) in correlation between masker and target cells at PZ = 6 ST during behavior. In addition, we have found that attention not only modulated the correlation between cells that had close or overlapping receptive fields, but also between distantly related cells in the background stream. We computed the STACs between masker cells at opposite side of PZ and the same side of PZ respectively and found that there were increased correlations during behavior in both cases (Figure 4.6).

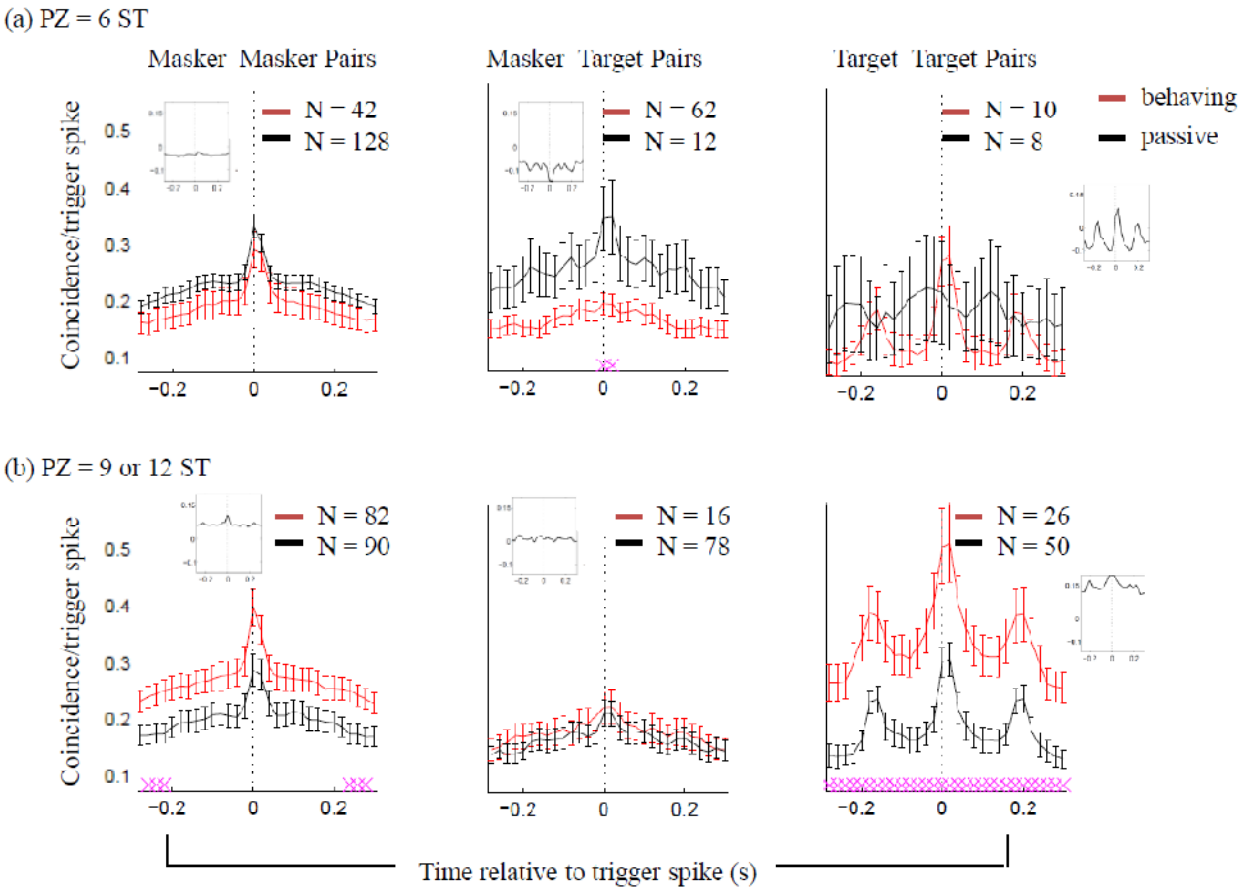

Figure 4.5 STACs from responses to target stimuli under two attentional conditions (a) for smaller PZ, 6 ST and (b) larger PZ, 9 or 12 ST. Insets display the difference between the mean STAC at each condition. Error bars indicate SE. Magenta crosses indicate significant difference (t test, $p < 0.01$) at the time-lag between the two conditions.

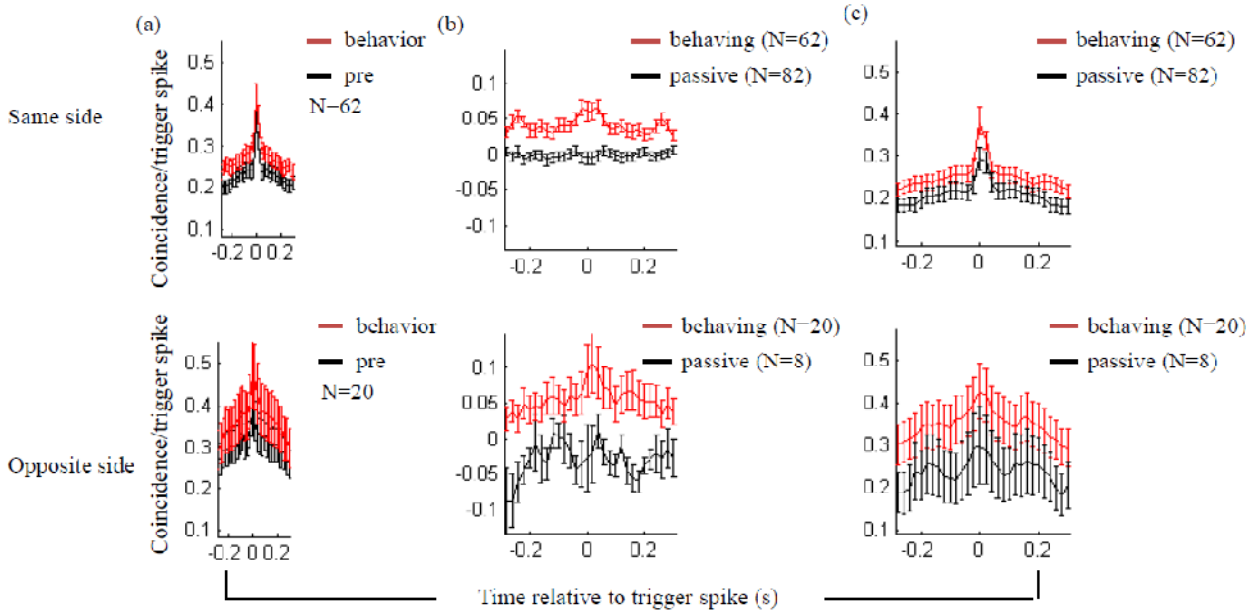

Figure 4.6 STACs of responses between distantly/closely related masker cells at larger PZ. (a) STACs from responses to reference stimuli; (b) The difference between STACs from behavior versus pre-behavior conditions to reference stimuli; and (c) STACs from responses to target stimuli.

Finally, to tease apart the sources of the correlations, we computed signal and noise correlations, respectively, and are given in Appendix 2 Figure A2.2. In most cases, correlations were due to the signal rather than the noise. The noise correlations computed from responses to reference stimuli show no significant difference between the two attentional conditions. The noise

correlations computed from responses to target stimuli show significantly reduced/increased correlations between masker cells at smaller/larger PZ.

In summary, during behavior, the STACs from responses to reference stimuli show that the correlations between masker cells are enhanced relative to the passive state. The STACs from responses to target stimuli (when both the target and maskers are present, but the animal's focus is on the target), correlations between target cells (and to a lesser extent among masker cells) are significantly increased, while those between cells belonging to the opposite streams (masker and target cells) are reduced or unchanged compared to the passive condition. These results are consistent with the assumptions of the coherence model that we describe in detail later.

4.3.2 Rapid STRF Plasticity

How are the correlations among different cell types related to the changes seen in their receptive fields during behavior? Here we examine examples in Figure 4.7 of masker and target cells for the 6 ST task. It is clear from raster and PSTH plots that the spiking activities of masker/target cells are driven by the masker/target tones. The target tones were always placed in the middle of the PZ. We computed reference-STRF from responses to reference stimuli. Consistent with the increased correlation among masker cells in the reference epoch during behavior (Figure 4.3 and 4.4) , we found that the masker cell's reference-STRFs were enhanced significantly during behavior, but reverted back to pre-behaivor shapes afterwards. By contrast, target cells (Figures 4.7c and 4.7d) experienced a relative "depression" in their reference-STRFs during behavior relative to the passive. These cells also do not show a significant or large increase in correlation among them or with masker cells during the reference period of the behavior. Other examples of

masker/target cells at 9 ST PZ demonstrate similar effects as those at 6 ST PZ (see Figure A2.3 in Appendix 2).

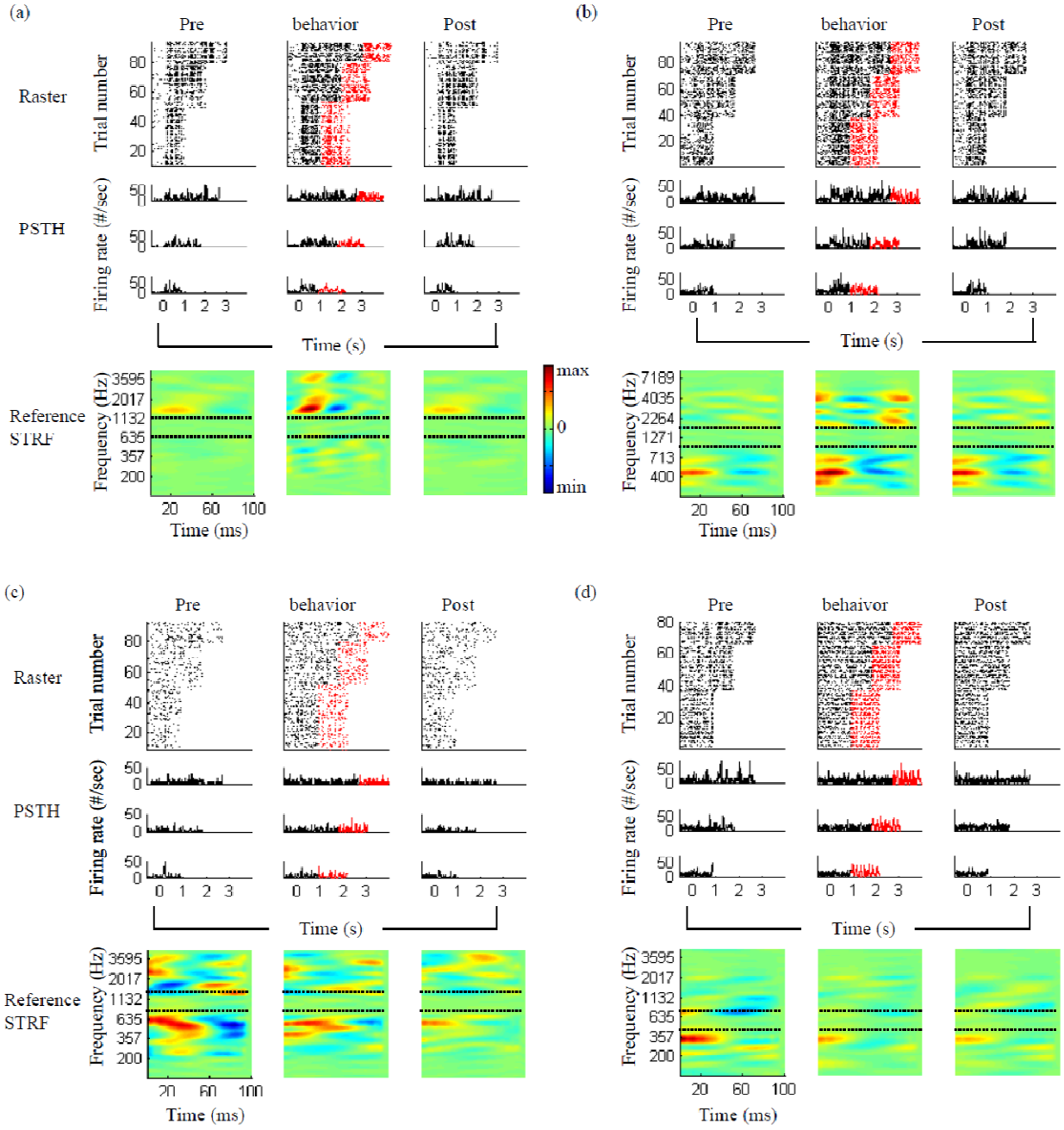

Figure 4.7 Examples of single units' raster plot, PSTH plot, and STRF at 6 ST PZ. (a) and (b) are from masker cells. (c) and (d) are from target cells. The area between two black dash lines represents PZ.

To examine the population effects of rapid STRF plasticity in streaming related task, we pooled in Figure 4.8 all units according to the stimulus conditions and computed the average STRF difference between the behavior and pre-behavior states by aligning each $STRF_{diff}$ to its unit's BF (see Methods for details). Those STRFs were computed only from responses to reference stimuli. STRF changes exhibit different patterns depending on cell groups and PZ width, but all are in line with the examples we show in Figure 4.7. For masker cells, we found the STRFs are enhanced in all PZ conditions. For target cells, we found that there was a net suppression which became weaker as the PZ width increased. In Figure 4.8b we illustrate a histogram of the local STRF changes from all masker and target cells. We also computed the difference between the average target STRF from behavior and passive conditions (see Figure A2.4 in Appendix 2). It is important to note in Figure 4.7 that while reference-STRFs are derived from exactly the same stimuli in passive conditions (pre and post-behavior) as during behavior, this is not the case for the "target-STRFs". They are computed from the target stimuli during behavior. Here, we compared these to target-STRFs computed from the same target stimuli but recorded later from a different population in passive condition. We found the STRFs are enhanced for target cells in all PZ conditions, which are consistent with increasing correlation during behavior for target cells. For masker cells, the enhancement is relatively small comparing with the changes in target cells. More careful control experiments need to be conducted before stronger conclusions can be drawn.

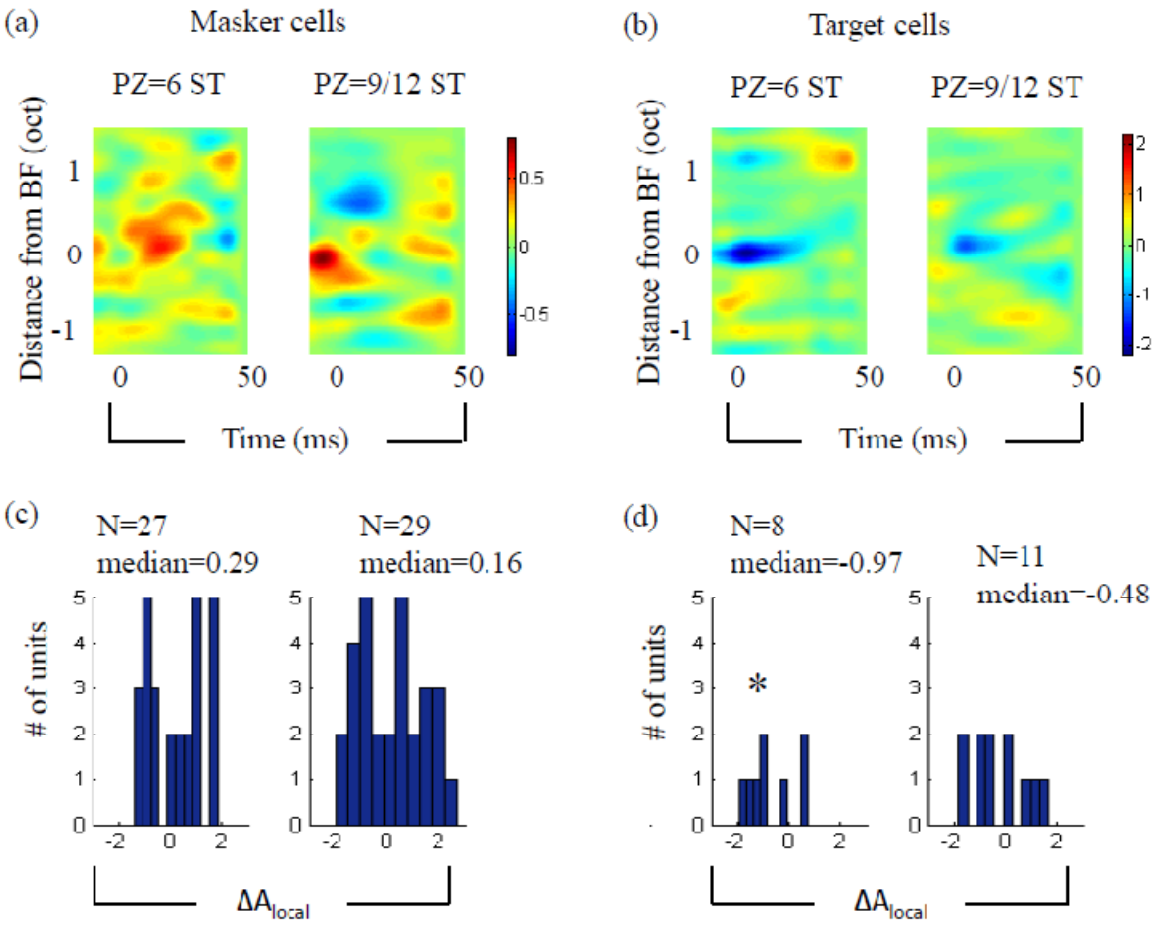

Figure 4.8 Population patterns of reference STRF plasticity. Average STRF difference between behavior and pre-behavior conditions (a) for masker cells; (b) for target cells. Histogram of STRF changes at BF (c) for masker cells; (d) for target cells. * represents that mean is significant different from 0 (t test, $p < 0.05$).

4.4 Summary and Discussion

In summary, we found changes in correlation and STRF in primary auditory cortex when an animal performed a task that involved streaming of the acoustic stimuli. That is the regularly repeating tone sequence embedded within multiple randomly varying tone bursts. Behavioral

results from Chapter 2 showed that ferrets could stream these stimuli and that the performance was improved with increasing PZ width. Here we found the following results:

1. Consistent with previous findings (David et al., 2008), responses and STRFs driven during the "reference" portion of an appetitive task became enhanced compared to the passive state. Cells that are not primarily driven (or have BF selectivity far from the reference stimuli such as the target cells) became suppressed.
2. In Figure 4.8a, there seems to be a bigger overall enhancement (and less suppression) in the 9/12 ST cases than in the 6ST case. Again, this is consistent with findings (Atiani et al., 2009) that easier tasks cause more overall enhancement and less suppression, compared to more difficult task. To elaborate further, in Atiani's experiments (2009), animals were trained to perform a detection task of a target tone embedded in noise using a conditioned avoidance procedure. They found at easy task (high SNR) cells tuned near the target tone frequency showed an enhancement at BF, while those tuned far from it showed suppression effects. In our experiment, the animal was trained to detect the target tone sequence using a positively reinforcement procedure. The masker cells in our case are the near cells in their experiments where enhanced sensitivity at BF was found. The target cells are equivalent to the far cells and the suppression at BF was displayed in our case. The suppression depended on PZ width, which is equivalent to task difficulty in their experiments.
3. During the target-phase, target and masker cells were driven by two streams, and the animal was attending to the target tone, so we found enhanced STRFs for both cell groups, but the effects were much larger for target cells.

4. The correlations in responses depended on the streaming and behavior. Basically, cells that belonged to the same stream were positively correlated, while those in different streams were uncorrelated or only weakly correlated. Furthermore, cells in the stream attended to by the animal during behavior were more positively correlated than the unattended stream.

Chapter 5 A Computational Model of Auditory Scene Analysis based on Temporal Coherence

5.1 Introduction

Auditory scene analysis (in analogy to vision) is to parse mixed acoustic events into meaningful streams where each stream is assumed to originate from a separate source. If the acoustic stimuli are speech, it is often known as cocktail-party effect (Cherry, 1953) or the speech segregation problem. As with the correspondence problem in vision scene analysis, the auditory scene analysis has to solve the binding/grouping problem for the cues that belong to each source. There are two levels of binding/grouping, namely simultaneous binding and sequential grouping. Simultaneous binding is to deal with at each time instant what channels (i.e. frequency, pitch, location, etc.) of auditory representations are dominated by one stream/source. Sequential grouping is to match the auditory representations of a stream at a particular time with the representations of the same stream at a later time. Pitch, common onset/offset, frequency/amplitude modulation, location, and frequency proximity are cues often used in the conventional computational auditory scene analysis (CASA) models. However, how these different cues/features are integrated has not been well addressed and remains a challenge. Here, we propose a novel CASA model based on temporal coherence and attention/memory. Temporal coherence solves the simultaneous binding problem and provides an elegant way of integrating different cues, for channels that belong to the same stream are activated coherently, no matter whether they represent pitch or location or timbre cues. Attention/memory is proposed to solve the sequential grouping problem.

5.1.1 Review of Existent Models

Because of its wide application in speaker separation, speech enhancement, and speech recognition, many computational models have been proposed to perform auditory scene analysis. Some of them, like blind source separation (BSS), are purely based on the statistics of the signals. In BSS, the goal is to reconstruct streams under the condition that their signals are independent and are linearly combined at multiple sensors. If the distribution of sources is hypothesized, the mapping between signals and sources can be found by minimizing the statistical distance between the hypothesized distribution of sources and the estimated distribution of the sources (Cardoso, J.F., 1998). BSS has been mathematically proven to be feasible for source separation (Belouchrani et al., 1997, Pham and Cardoso, 2001, and Fevotte and Doncarli, 2004) when mixtures are not too noisy and the number of sensors is equal or more than the number of sources. Recently, a Markov chain Monte Carlo implementation was proposed by Fevotte and Godsill (2006), which can deal with the noisy and underdetermined situation where sources exceed sensors. In this method, audio signals are first decomposed on a local cosine lapped transform basis, and are then sparsely represented. Separation is performed in the transform domain by using Gibbs sampler and minimum mean-square error estimates.

Instead of considering speech segregation under the multiple-sensor situation, some statistical methods are proposed to address the situation when only one channel is available, as when we listen monaurally. Raj and his colleagues (2006) proposed latent Dirichlet decomposition for single-channel speaker separation. Individual speaker's spectrogram at each time instant is modeled as a multinomial distribution. The spectrogram of the mixed speech is the linear combination of those from each speaker. The combination coefficients are modeled as a Dirichlet

density, and the parameters of the multinomial distribution and Dirichlet density are learned from the original unmixed speech. Reddy and Raj (2007) recently proposed two more methods for single-channel speaker separation MMSE algorithm and soft mask estimation. The distribution of the log spectral vectors for any speaker is modeled as a Gaussian density mixture. The MMSE algorithm attempts to minimize the mean squared error in the log spectrum. Soft mask estimation is to compute the probability of any time-frequency component belonging to the target speaker, instead of a binary mask. Both methods result in improving signal to interference ratio (SIR) and perform better than an equivalent binary-mask algorithm.

However, all these data-driven methods are essentially unrelated to the way that the best performers, human beings, do the task. In this vein, some auditory scene analysis models have been developed based on findings from psychoacoustical and neurophysiological studies. A number of ideas have been suggested to mediate scene analysis including pitch, common onset/offset, and location (Bregman, 1990; Bronkhorst, 2000; Shinn-cummingham, 2005). Almost half a century ago, pitch was already proposed as a way to perform monaural speech segregation probably because of its close relationship to voicing in speech. Parsons (1976) described a method to separate target speech from interfering speech based on harmonicity assumptions. First, the peaks of spectrum of two utterances are identified. Secondly pitches are extracted according to Schroeder's histogram which is generated by finding all integer submultiples of all the peaks. Then peaks are assigned to the corresponding pitch. Over time, pitches belonging to one speaker are tracked by fitting a least-squares straight line to the three most recent pitch samples and choosing the best match between the predicted and observed values. Finally, the spectrum of the target speaker is synthesized from its harmonics and the

inverse Fourier transform is performed to get a continuous speech waveform. Since then, much work has been done to develop better pitch estimation and tracking algorithms.

The classic pitch analysis is based on the auto-coincidence of the cochlear filter output proposed by Weintraub (1985), and which Slaney and Lyon (1990) call the *correlogram*. It is a short-time multi-channel auto-correlation from all channels of the auditory filterbank. Later, Karjalainen and Tolonen (1999) introduced a more efficient 2-channel auto-correlation analysis in which an enhanced summary auto-correlation function (ESACF) is generated by removing the repetitive peaks and the near zero time-lag part of the summary auto-correlation function (SACF) curve. Ottaviani and Rocchesso (2001) further improved the performance by multiplying the SACF and ESACF. Instead of summing over all channels in correlogram, Wu et al. (2002) used a statistical relationship between ideal pitch and the time lags of peaks in selected clean channels to estimate pitch. Inspired by findings from psychoacoustic studies which imply that auditory system processes the resolved and unresolved harmonics in different ways, Hu and Wang (2004) proposed an algorithm especially to improve the grouping of the unresolved harmonics based on common amplitude modulation and temporal continuity. Another alternative for pitch estimation was proposed by Quatieri (2002). Instead of computing a short-time autocorrelation analysis, the new method performs a 2-D Fourier transform of a narrowband spectrogram of the signal, which is called grating compression transform (GCT). Pitch is estimated by calculating the vertical distance of the peak of the GCT magnitude to the GCT origin. Its feasibility for estimating pitches of 2 speakers talking simultaneously has been applied to co-channel speaker separation (Wang and Quatieri, 2009a and 2009b). A totally different method for pitch perception is based on place, instead of temporal, representation of sound. In that model, pitch is estimated

instantaneously by cross-correlation the instant spectrum of input stimuli with harmonic pitch templates (Goldstein, 1973; Shamma and Klein, 2000). Most CASA approaches segregate speech from target and interfering speakers by assigning all of the energy to the dominant speaker after examining the energy in each time-frequency unit, which reduces the models performance when the energy from both speakers overlap in a particular time-frequency unit. An algorithm proposed by Vishnubhotla and Espy-Wilson (2009) is different from traditional CASA approaches. The algorithm separates the participating speaker streams using a Lease-Squares fitting approach to model the speech mixture as a sum of complex exponentials.

All above pitch-based methods are limited to voiced speech. For separation of the non-voiced speech, several other cues have been proposed, most common among them are the binaural cues. The first biologically-inspired computational model of binaural localization and separation was proposed by Lyon (1983). Applying the Jeffress model (Jeffress, 1948), the cross-correlation between auditory spectrogram coming from two ears is calculated. Sources are localized based on peak-picking in the correlation functions and different gains are assigned to the corresponding sources. Improving of the Jeffress model whose performance degrades significantly in more than 2 sources, Liu et al. (2000) proposed to incorporate a “stencil” filter which can reduce the high-frequency ambiguity for ITD estimate and enhance the localization of sound sources. The system performs well in detecting the source locations with four talkers in experiments or six talkers in computer simulation. Palomaki et al. (2004) applied a skeleton cross-correlation function to improve the ITD estimation. Interaural level difference (ILD) is also incorporated to refine speech segregation by comparing the measured ILD with the ideal ILD at the particular location estimated from the ITD.

Roman et al. (2002 and 2003) presented a CASA model to segregate speech based on sound localization. They suggested a binary mask for auditory spectrogram of the mixtures to select the target if it was stronger than the interference in each time-frequency unit. The mask is generated based on joint information of both ITD and ILD. Although the performance has been demonstrated better than an existing approach, the proposed model does not address how to define a target in a multi-source situation. In a similar way, a soft time-frequency mask is derived based on the joint distribution of ITD and ILD cues (Brown et al., 2006; Srinivasan et al., 2006). In addition, common onset/offset (Brown and Cooke, 1994; Hu and Wang, 2007) and amplitude modulation (Kollmeier and Koch, 1994) have been implemented in several CASA models. However, it remains a problem to integrate these different cues into a CASA model.

5.2 The Temporal Coherence Model

The model we propose uses a novel cue, temporal coherence, to perform the simultaneous binding. The idea is that highly correlated channels over a short time period represent a common fate which has been well known as one of the Gestalt Principles guiding scene analysis. Temporal coherence provides an elegant way of solving the problem of integration of evidence derived from multiple cues. For example, channels, that have a common onset and are co-modulated are coherently activated. Such temporal coherence correctly associates all different types of features (e.g. pitch, location, loudness, color, texture, etc.) within and even across modalities.

The correlation theory of brain function was described by von der Malsburg in 1986 to address how neural assemblies communicate across distance. In these theories, temporal oscillations are

assumed to arise from spontaneous sources that are not related to the stimulus dynamics per se. Evidence from visual cortex shows that neural populations responding to the same object tend to fire synchronously and are desynchronized from those responding to different objects. Stimulus-specific neuronal oscillatory responses have been found in the cat visual cortex (Gray and Singer, 1989; Gray et al., 1989). This correlation in brain has been simulated in the auditory modality to illustrate cocktail-party effect by an oscillatory network (von der Malsburg and Schneider, 1986). Wang and Brown (1999) also proposed a CASA model using the principle of oscillatory correlation. Segments derived from harmonically related channels are input into an oscillatory network. Binding across these frequency channels at each time frame of the representation is formed by the oscillatory correlation.

Our model is completely different in that the temporal characteristics are those of the stimulus and are not intrinsic. Furthermore, sequential grouping depends on attention or memory. ASA has been hypothesized to consist of two processes (Bregman, 1990). The first process is pre-attentive/data-driven, which forms low level auditory representations and only involves peripheral processing. The second one is schema-based, which involves high level, central/top-down processing. The role of attention in stream formation has been a debatable question for a long time. Results from EEG studies have argued that attention is not always required for stream formation, but can limit the processing of unattended input in favor of attended sensory inputs (Sussman et al., 1999 and 2007). However, experiments from psychoacoustical studies have shown that temporal coherence/fission boundary is influenced by attention (van Noorden, 1975) and recent binaural stream segregation experiments by Carlyon and his colleagues (2001) support that attention plays a key role in streaming by showing that effective build-up of

streaming is significantly affected by attention. Besides, evidence from MEG studies (Gutschalk et al., 2008; Elhilali et al., 2009), and neurophysiological study (Yin et al., 2007) has demonstrated that attention can modulate neural responses associated with streaming percepts. Although top-down processing is an important part of ASA, computational modeling of this processing is a challenging problem, which may explain why very few models in the literature consider the top-down influence. However, a computational model of auditory selective attention for stream segregation has been presented by Wrigley and Brown (2004) in which a network of neural oscillators performs stream segregation based on oscillatory correlation proposed by Wang and Brown (1999). The attentional process is modeled as an attentional leaky integrator, which determines the connection weights between oscillators and an attentional unit. The attentional stream is those auditory representations, the activity of whose oscillators are synchronized with the attentional leaky integrator. Godsmark and Brown (1999) employed a schema-driven organization in their multi-layer blackboard architecture for CASA, which allowed high-level predictions based on a previously observed pattern to influence the organization at lower levels of the blackboard. Elhilali and Shamma (2008) presented an online clustering method for stream segregation by comparing predictions from Kalman filter with the incoming sensory input. In our model, the role of attention is simply simulated by assigning channel binary weights based on the energy of each channel (i.e. variance of each channel) within a range of selected channels. In the CASA literature reviewed in Introduction, the speaker ID task has been performed using pitch tracking or assumed known spatial location for target. But this has not addressed the issue of scene analysis based on long term memory. However, it is trivial in our model to use location or pitch as a cue guiding the grouping over time, again, because of binding variant attributes based on temporal coherence.

In the temporal coherence model, the first stage is to compute a coherence matrix by calculating correlation at zero-lag across channels in the multi-dimensional auditory representations. The pair-wise correlation measures the degree of synchrony between channels. Highly and positively correlated channels are synchronized and desynchronized away from the rest channels. Coherence forms a basis for organizing features belonging to one stream and detaching those belonging to the interference. The second process is either employing attention to select the correlation coefficients for target stream or finding the closest matched correlation coefficients with memory. A pre-trained support vector machine (SVM) is used to mimic the function of memory. The selected correlation coefficients act as a mask to enhance the auditory representations activated by target stream and suppress those activated by the interference.

5.3 Computational Model for Auditory Scene Analysis

A biologically-inspired computational auditory scene analysis model, based on temporal coherence and attention/memory, is proposed in this study. The diagram of the model is shown in Figure 1. The model is comprised of four main stages. First, sound waveforms are projected into multi-dimensional feature space (i.e. frequency, bandwidth, pitch, and location). These features are rapidly extracted (e.g. in the order of 10 ms or less). Second, after multi-resolution filtering over time, a windowed pair-wise correlation is computed across all feature channels. Channels with high correlation coefficients and the same sign tend to evolve together over time and belong to one stream. Each row of the coherence matrix tells the degree of coherence of this channel with the rest, or from this channel point of view, how the auditory elements are organized. This stage requires integrating a relatively long time periods (e.g. up to ~500 ms). Thirdly, a mask is formed by selecting channels and correlation coefficients associated with

those channels (i.e. rows of the coherence matrix) from the coherence matrix according to either attention focusing or memory matching. Finally, the mask is multiplied with the input features to separate target from the interference by filtering out the target representations and suppressing the interference. Then, the filtering representations are converted back to the acoustic domain.

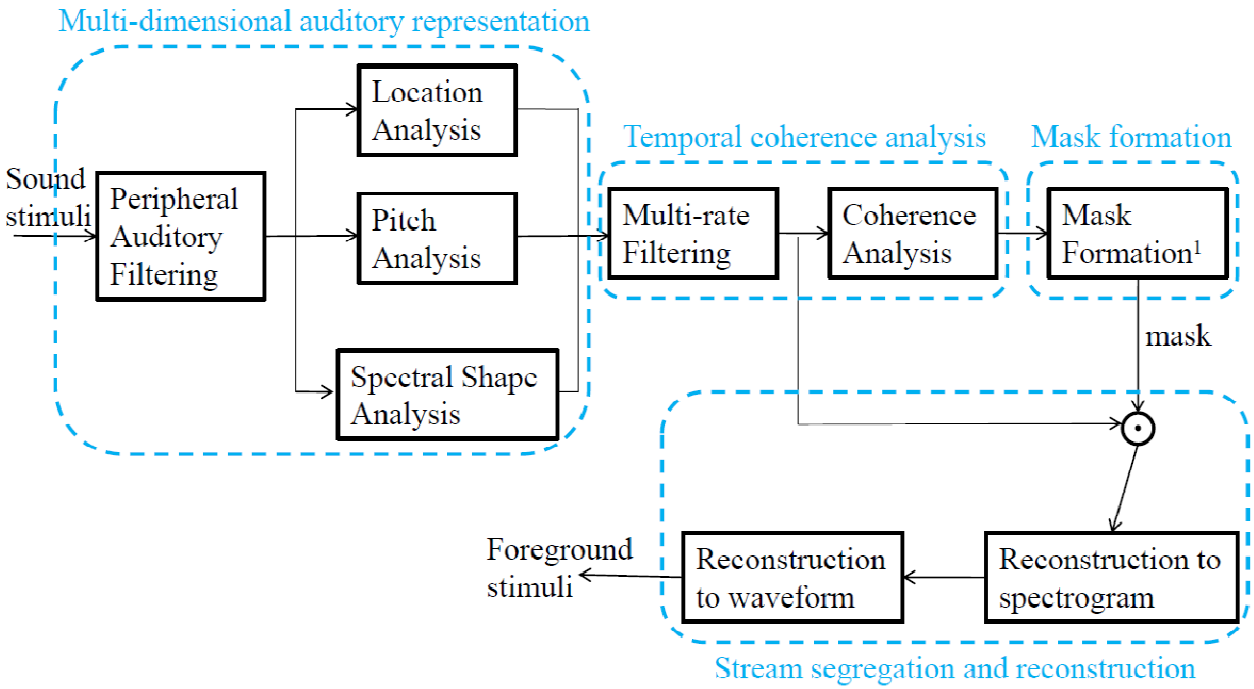

Figure 5.1 The CASA Model diagram. 1: selecting mask on feature domain according to either attention or memory.

5.3.1 Stage 1: Multi-dimensional Auditory Representation

This stage basically performs simultaneous feature extraction of acoustic signals. Here we only consider four primary features (i.e. frequency, pitch, bandwidth, and location) promoting auditory stream segregation; however, it is straight-forward to add other features into this model.

5.3.1.1 Peripheral auditory processing

After travelling through the inner ear, the input signal, s(t), is decomposed into a two-dimensional time-frequency domain through a series of peripheral auditory processing (Yang et al, 1992; Wang and Shamma, 1994): cochlear filter bank decomposition, hair cell filtering, and spectral sharpening and rectification.

Cochlear filtering is modeled by a bank of 128 overlapping constant Q bandpass filters whose center frequencies are uniformly distributed along a tonotopic/logarithmic frequency axis (x) over 5.3 octaves and impulse response of each filter is denoted by h(t;x). The cochlear filter is implemented by a minimum-phase signal h(t) with magnitude frequency response

$$|H(x)| = \begin{cases} (x_h - x)^{\alpha} e^{-\beta(x_h - x)}, & 0 \le x \le x_h \\ 0, & x > x_h \end{cases}$$
$$y_{coch}(t,x) = s(t) *_t h(t,x) \qquad (1)$$

where x_h is the cutoff frequency, $\alpha = 0.3$, $\beta = 8$, and $*_t$ denotes convolution operation in the time domain. For details of cochlear filter implementations, see Ru (2000).

The responses of these cochlear filters are further transduced by inner hair cells through a high-pass filter mimicking the fluid-cilia coupling, a nonlinear compression, g, modeling the function of ionic channels, and a low-pass filter, w(t), accounting for hair cell membrane leakage.

$$y_{AN}(t,x) = g(\partial_t y_{coch}(t,x)) *_t w(t) \qquad (2)$$

Then the auditory nerve responses are transmitted to the cochlear nucleus, where a lateral inhibitory network is applied to enhance the frequency selectivity of the cochlear filter bank. The

lateral inhibition is approximated by a first-order derivative with respect to the tonotopic axis and then half-wave rectifying.

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

Finally, the output of the lateral inhibition network is integrated over a short period to effectively extract the envelope of the channel outputs. The final output is called auditory spectrogram shown in Figure 5.2a.

$$y(t,x) = y_{LIN}(t,x) *_t \mu(t,\tau) \quad (4)$$

5.3.1.2 Spectral shape analysis

The auditory spectrogram is further transmitted to higher central auditory stages to extract cues/features. Neurophysiological findings in primary auditory cortex (Kowalski et al, 1996; Miller et al., 2001; Elhilali et al., 2007) and human psychoacoustical experiments (Eddins and Bero, 2007; Green, 1986; Viemeister, 1979) suggest that the central auditory system performs a spectral shape analysis which is an effective physical correlate of the percept of timbre. The spectral shape analysis is implemented in the model by wavelet decomposition along the tonotopic axis (Wang and Shamma, 1995; Chi et al., 2005). Each slice of the auditory spectrogram at a given time instant (t) is convolved with a bank of scale filters, $\mathcal{G}_{SF}$, that range from narrow to broad bandwidths. This multi-scale analysis captures the local and global spectral modulation of the auditory spectrogram. For example, in Figure 5.2b, the output of the broadband scale filters shows speech formants, while the output of the narrowband scale filters shows the resolved harmonics.

where $*_x$ denotes convolution operation with respect to the tonotopic axis x and and denote Hilbert transform pairs. The scale filter is a complex spectral "impulse response" function. It is chose to be a Gabor-like function, which is defined as the second derivative of a Gaussian function. is the spectral density of the filter, covering the range from 1/8 to 8 cycles/octave where the cortical neurons most sensitive to.

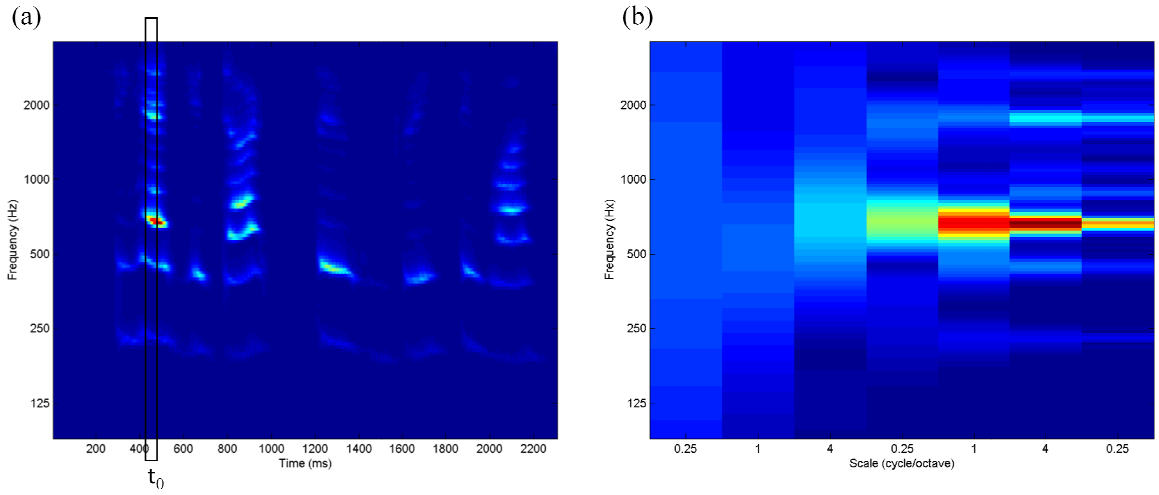

Figure 5.2 Examples of the outputs of peripheral auditory processing and cortical spectral shape analysis. (a) The auditory spectrogram of an utterance from a female speaker, and (b) the output of cortical spectral shape analysis for the slice of the auditory spectrogram at time instant t_0.

5.3.1.3 Pitch analysis

Pitch is an important cue in segregation of harmonic sounds (Moore et al., 1986), for example speech. Pitch information is implicitly represented in the harmonic structure in the auditory

spectrogram. Here, we extract pitch information from the auditory spectrogram using a template matching model proposed by Goldstein (1973) and Shamma and Klein (2000).

First a harmonic template representative of any harmonic series is generated by cochlear filtering. On the logarithmic frequency axis of the cochlea, this template remains the canonical template and unchanged since the harmonic series for any fundamental is simply a translation along the frequency axis. At each time instant t, this template is convolved with the input spectrum, y(t,x). and the similarity at each shift is scored by cross-correlation between them. Pitch values are given by peak-picking from the output of the cross-correlation indexed by tonotopic frequency x. The pitch strength at a given fundamental frequency is based on Euclidean distance between the spectrum and the corresponding template. One octave confusion introduced by this template matching method is solved according to the relative pitch strengths. Figure 5.3 displays the pitch estimates over time for a mixture of male and female speech.

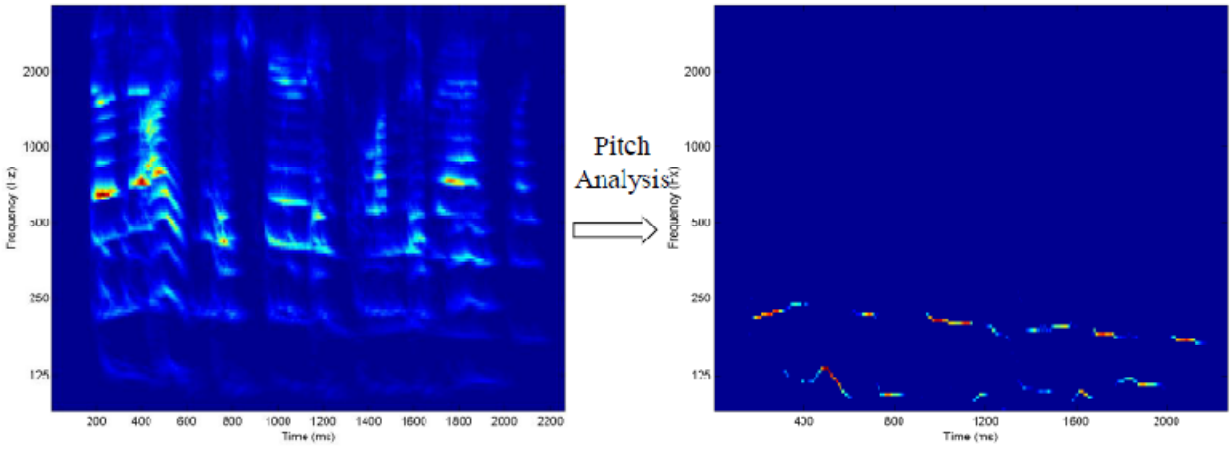

Figure 5.3 An example of output of pitch analysis. Left panel: the auditory spectrogram of the mixture from a male speaker and a female speaker. Right panel: the output of pitch extraction for both speakers.

5.3.1.4 Location analysis

Humans and many animals determine the location of a sound source by comparing the responses between two ears. For example, sound waves arrive at the ear closer to the source slightly earlier than the farther ear, which causes an interaural time difference (ITD). Meanwhile, an interaural intensity/level difference (IID/ILD) is caused by the sound intensity difference between the closer and the farther ears. In this study, we only consider ITD as the location cue. A biologically plausible model for ITD was described by Jeffress in 1947. The signals at the two ears are transmitted to the higher central auditory system with a delay because of ITD. The corresponding coincidence detector represents this delay. We implement this descriptive model using the algorithm proposed by Lyon (1983).

The algorithm begins by computing a cross-correlation between the auditory spectrograms at the two ears.

$$L(t,x,\tau) = \sum_{n=0}^{N-1} y_L(t-n,x)\,y_R(t-n-\tau,x)\,w(n) \qquad (6)$$

where $y_L(.)$ and $y_R(.)$ are the auditory spectrograms at the left and right ears, respectively. w(n) is a window of size N samples. We use a rectangular window with 50 samples (i.e. 6.25 ms). τ is between -1 and 1 ms. L(t,x, τ) is called cross-correlogram and Figure 5.4a shows an example of two speakers with -0.375 ms and 0.5ms ITD, respectively. To improve the robustness of the ITD estimation, we sum the cross-correlation over frequency channels shown in Figure 5.4b. By peak-picking from the summary cross-correlogram, ITDs are estimated (Figure 5.4c). Since we only consider stimuli in which ITDs are synthetically generated as those in Shackleton (1992), there are no diffraction effects which introduce a weak frequency-dependence to ITDs and no reverberant conditions. Therefore, this simple algorithm already gives reasonable results.

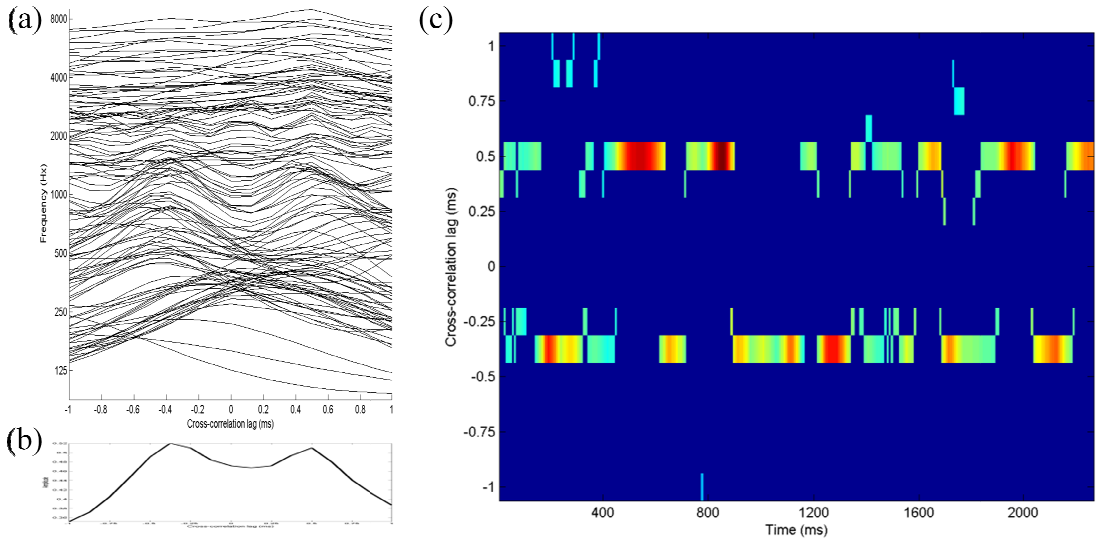

Figure 5.4 An example of output of ITD analysis. (a) The cross-correlogram and (b) the corresponding summary cross-correlogram for one time instant of the mixture in Figure 5.3, and (c) the ITDs for the two speakers over time.

5.3.2 Stage 2: Temporal Coherence Analysis

The analysis of this stage proceeds in two steps: multi-rate filtering and pair-wise correlation.

5.3.2.1 Multi-rate filtering

Evidence shows that cortical neurons tune to a limited range of temporal modulations (Kowalski et al., 1996; Lu et al., 2001). At this stage, features extracted from the early stages are gone through a temporal analysis with multi-rate dynamics covering from 2 to 32 Hz. The multi-rate analysis integrates the history of neuron responses and is used in the next step to compute a temporal coherence matrix. Similar to the multi-scale analysis, this multi-rate analysis is implemented by wavelet decomposition along the time axis (Wang and Shamma, 1995; Chi et

al., 2005). Specifically, the temporal analysis is implemented by convolving the input (i.e. each frequency-scale-pitch-location channel), I, at each time instant t with a bank of rate filters, $\mathcal{G}_{RF}$.

$$\begin{gathered} R(t,x,\Omega,\mathrm{p},\mathrm{l},\mathrm{w}) = I(t,x,\Omega,p,l) *_t \mathcal{G}_{RF}(t,w) \\ \mathcal{G}_{RF}(t,w) = \mathrm{w}g_t(wt) + j\mathrm{w}\hat{\mathrm{g}}_{\mathrm{t}}(\mathrm{wt}) \\ g_t(t) = t^2 e^{-3.5t}\sin(2\pi t) \end{gathered} \qquad (7)$$

where $\mathcal{G}_{RF}(.)$ is assumed to be a gamma function parameterized by the temporal modulation, w, which are [2 4 8 16 32] Hz.

5.3.2.2 Pair-wise correlation

This correlation analysis postulates that cortical neurons express relations between active cells representing parts of the same object through temporal coherence (Shamma et al., 2010). It measures the similarity of auditory responses across channels. This correlation is used to bind coherent channels and separate them away from those incoherent ones.

The correlation is computed over relatively long time windows based on the rate filters chosen in the multi-rate analysis, ranging from about 30 to 500 ms. This is consistent with the typical range of phase-locking rates in the cortical responses and stimulus presentation rates over which the formation of streams usually occurs. We only consider the instantaneous coincidence (i.e. correlation at zero lag) across all pairs of channels (i.e. frequency, scale, pitch, and location channels) integrated over time, which is roughly equal to instantaneous correlation between pairs of channels summed over rate filters.

$$C = \int I_i(t) I_j(t) dt \cong \sum_w R_i(w) R_j^*(w) \qquad (8)$$

where (*) represents the complex-conjugate.

This coherence matrix consists of a map of weights indicating the degree of coherence between pairs of channels. For example, the correlation coefficient near 1 indicates highly coherent pair of channels; the correlation coefficient near -1 indicates highly anti-coherent channels. For stationary stimuli, the matrix reaches a stable point after a build-up period, while for non-stationary stimuli, it dynamically evolves over time.

5.3.3 Stage 3: Mask Formation

Each row/column of the coherence matrix can be viewed as a "mask", which indicates from this channel's (i.e. this neural cluster) point of view how the auditory responses are organized, inferring the percepts of the stimulus. Presumably, channels belonging to the same source are co-modulated over time. Therefore, they are highly correlated differentiating them from those belonging to interference.

We postulate that a mask can be formed in two ways. In one way, when attention is applied to a range of feature channels, a binary weight is generated for each channel according to energy (i.e. variance) at those channels and is applied to all correlation coefficients paired with that channel. The diagonal of the coherence matrix indicates variance of each corresponding channel. Within the attentional focus, top N channels with highest energy are chosen. A target mask is formed by taking the average of the correlation coefficients paired with the chosen channels as indicated in Figure 5.5. In Figure 5.5, for illustration purpose, 2 channels within the attentional focus are chosen. During simulation, we chose N=5.

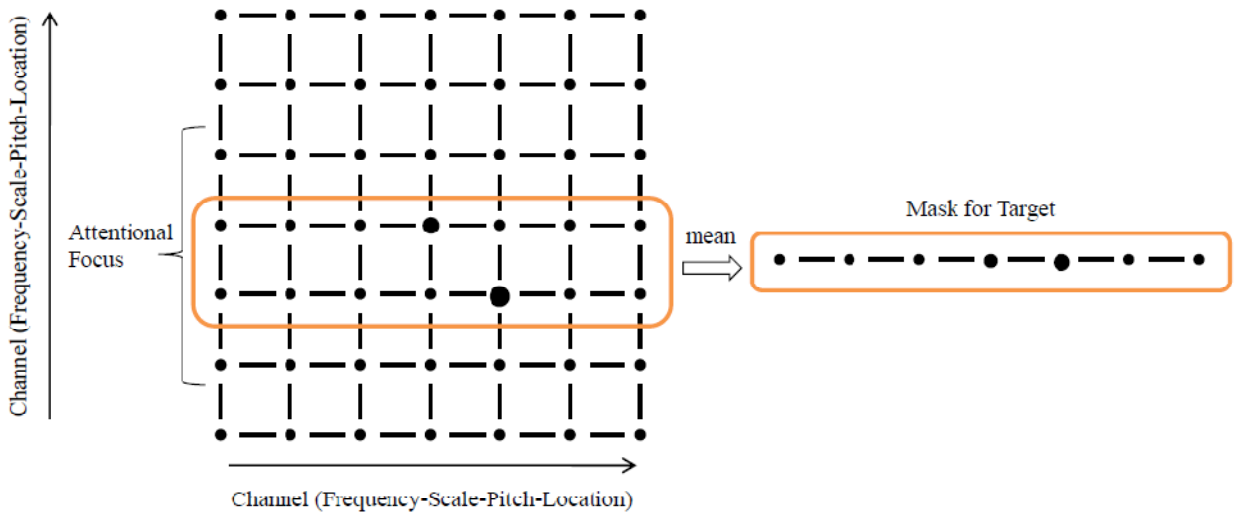

Figure 5.5 Schematic of coherence matrix and mask formation. The size of a circle indicates correlation coefficient between the corresponding pair of channels. A circle at the diagonal indicates variance of the corresponding channel. Within the attentional focus, N channels with highest energy are chosen. Here N=2 for illustration purpose. During simulation, we chose N=5. The average coefficients of the chosen channels form the mask for target.

In another way, a mask can be formed by selecting multi-rows from the coherence matrix according to memory. A boundary between masks for target and for non-target is pre-defined by a support vector machine (SVM). A SVM performs classification by constructing a hyperplane that separates the data into two categories (Theodoridis and Koutroumbas, 2006). The hyperplane is computed by optimizing the margin between separating boundary and support vectors. Radial basis function is used as the SVM kernel.

5.3.4 Stage 4: Stream Segregation and Reconstruction

Finally, the cortical representations of the target stream are segregated from those of the background by point to point multiplication of the formed mask with the output of multi-rate analysis. The same mask is applied to the output of each rate filter. The inverse wavelet transform and an iterative method based on the convex projection algorithm proposed by Yang et al. (1992) and Chi et al. (2005) are used to reconstruct the signal from the streamed cortical representations to the time domain.

5.4 Simulation Results

To demonstrate the performance of the model, we first test the model on the classic stimuli widely used to study the perceptual formation of auditory streams. Then we show the simulations of the model on speech segregation (or speaker separation).

5.4.1 Segregation of Tone Sequences

A sequence of tones alternating between two frequencies, A and B, is the typical stimulus used in many psychophysical and physiological studies of auditory streaming. The percept evoked by such sequences depends primarily on the frequency separation between the two tones, ΔF, and on the inter-tone interval, ΔT. For small ΔFs and relatively long ΔTs, the percept is that of a single stream of tone alternating in frequency (ABAB); for large ΔFs and relatively short ΔTs, the percept is that of two separate streams of tones of constant frequency (A-A vs. B-B). In the example shown in Figure 5.6a, the frequency separation is 1 octave and the representation rate is about 4 Hz for each tone. Under this condition, the percept is that of two separate streams. Figure 5.6a illustrates the model simulation of streaming of the two alternating tone sequences.

Another example is a sequence of harmonics alternating between two fundamental frequencies, 310 and 200 Hz (Figure 5.6b). The representation rate is about 4 Hz for each fundamental frequency. It is normally reported as two streams as well in perception under this condition and the simulation result is depicted in Figure 5.6b.

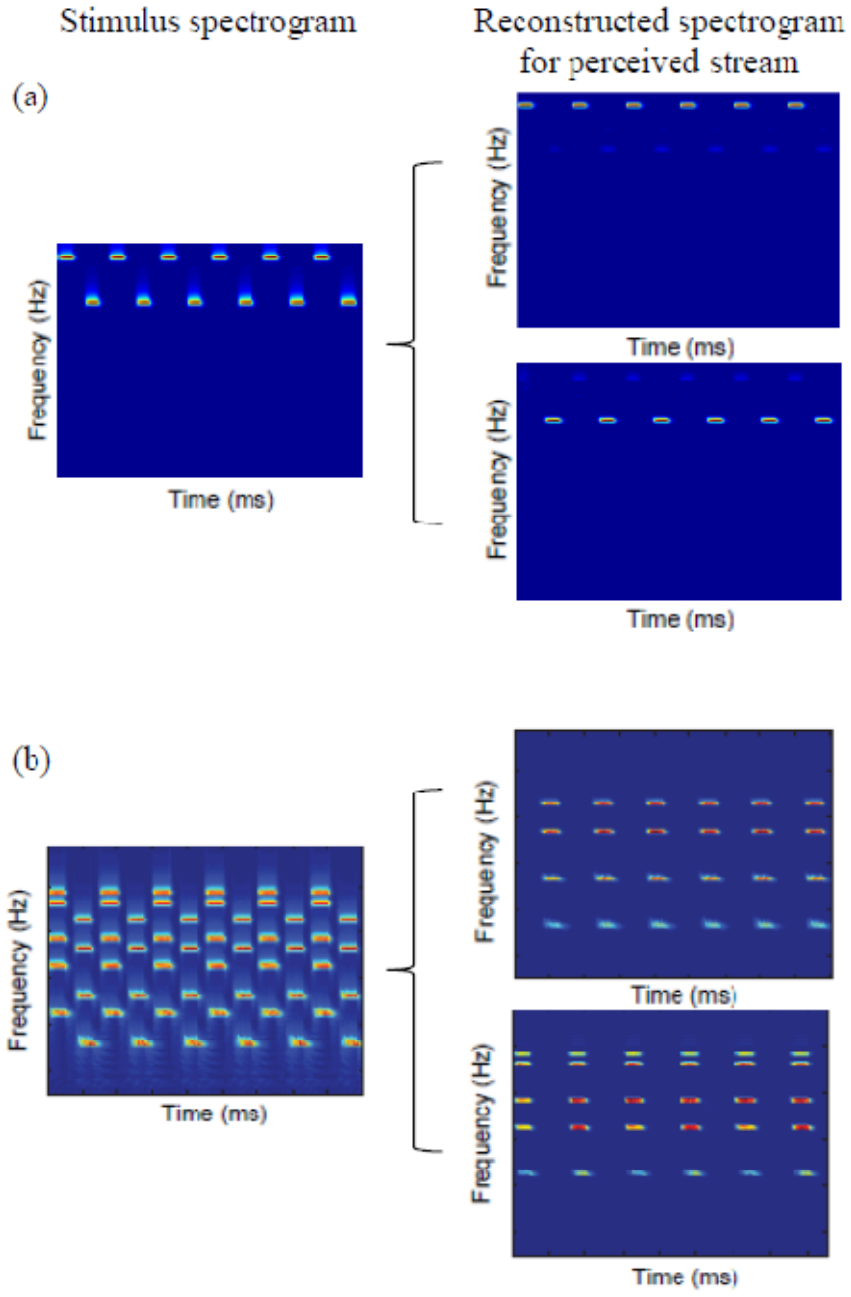

Figure 5.6 Model simulation of commonly used stimuli in auditory scene analysis. (a). A sequence of tones alternating between two frequencies (left panel) and the simulation results of perceived streams (right panels); (b) A sequence of harmonics alternating between two fundamental frequencies (left panel) and the simulation results of perceived streams (right panels).

5.4.2 Segregation of Speech Sounds

Experiments are conducted on synthetic mixtures of signals from different speakers to evaluate the model performance using both mask formation methods proposed in the previous section. Utterances of male and female speakers from the TIMIT database are used. Prior to addition, signals are resampled to 8 kHz and scaled to create speaker-to-interference ratios (SIRs) at 0 and 6 dB. Mixed signals are obtained by digital addition of utterances from individual speakers. The length of the mixed signal is set to the shorter of the two signals. We have three sets of two-speaker mixtures: female-female, male-male, and female-male. In this study, we only test our model on 2-speaker mixtures.

5.4.2.1 Attention-based mask formation

In this set of experiments, a mask is computed by applying attention to a range of channels in a specific feature domain. For instance, for mixtures from female and male speakers, we attend to pitch channels ranging from 150 to 300 Hz (from 70 to 150 Hz) to segregate utterances from the female (male) speaker. For mixtures from the same gender speakers, we attend to location channels corresponding to the target speaker's position. In this study, we assume sound sources to be stationary in space. Figure 5.7 shows an example of the original, mixed, and the segregated

spectrogram of the utterances from a female and a male speaker. It can be seen from the figure that considerable separation has been achieved for the target speaker. To quantify the model performance in speech segregation, we use the same metric, correlation between the original and segregated spectrograms, as that in Elhilali et al. (2008).

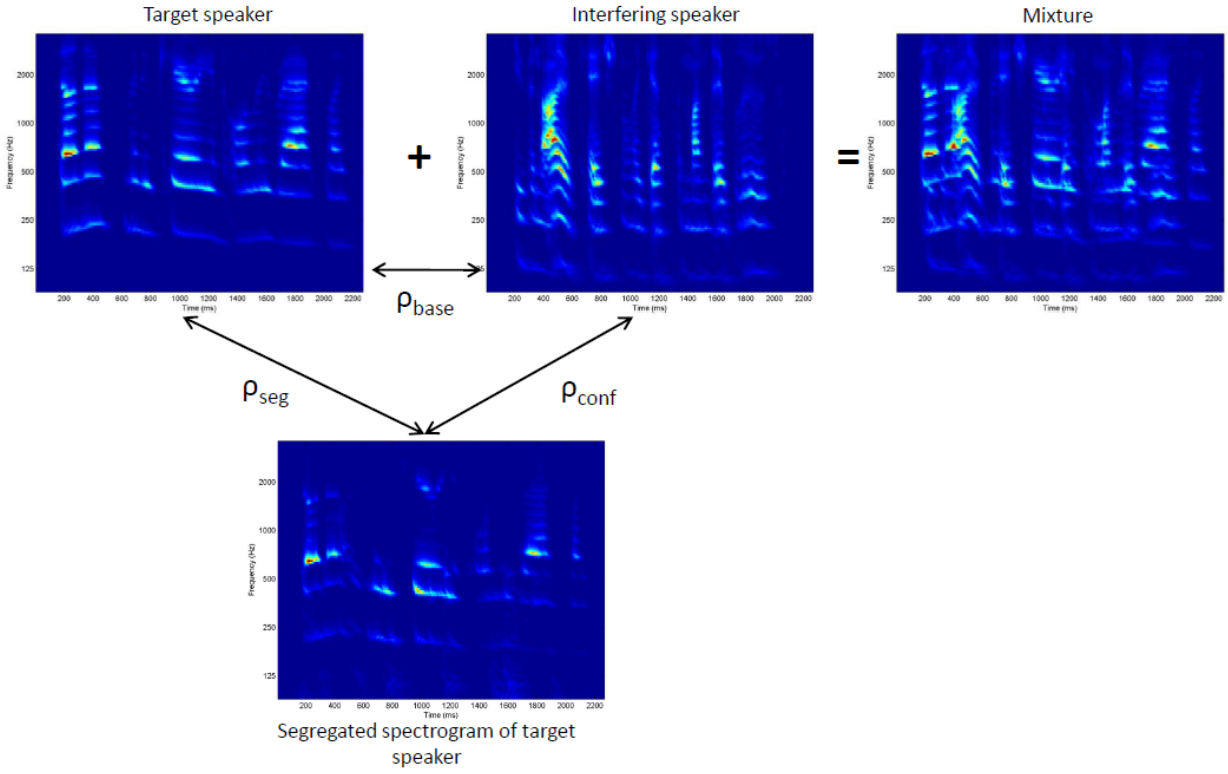

Figure 5.7 Model simulation with speech-on-speech mixtures. Model performance is evaluated using correlation coefficients between (1) the original and segregated spectrograms, ρ_{seg}, inferring how well the target signals are extracted; (2) the two original spectrograms, ρ_{base}, providing a baseline; and (3) the segregated spectrogram of the target signal against the spectrogram of the original competing signal, ρ_{conf}, indicating how well the interference is suppressed.

Several correlation coefficients are computed between: (1) the original and segregated spectrograms, ρ_{seg} (segregation correlations), inferring how well the target signals are extracted; (2) the two original spectrograms, ρ_{base} (baseline correlations), providing a baseline; and (3) the segregated spectrogram of the target signal against the spectrogram of the original competing signal, ρ_{conf} (confusion correlations), indicating how well the interference is suppressed. In the ideal condition, ρ_{conf} is equal to ρ_{base}, both of them much lower than ρ_{self}, and ρ_{seg} is equal to 1. To compensate for amplification and distortion effects introduced in the resynthesis process, we use resynthesized spectrograms for the two original signals to compute the correlation coefficients. The histograms of correlation coefficients (Figure 5.8a) at 0 dB SIR demonstrate that for mixtures from different gender speakers, segregation occurs with an accuracy of ρ_{seg} =0.73, which is significantly higher than the baseline, ρ_{base}=0.04 and the confusion, ρ_{conf}=0.27. The performances show further improvement at 6 dB SIR (Figure 5.8b) with an accuracy of ρ_{seg} =0.82 and ρ_{conf} = 0.16. For mixtures from the same gender speakers, performances are similar as those from different gender speakers, except that the confusion correlations increase slightly as expected (Figure 5.8c and d). The numbers reported here are the median values of the correlation coefficients.

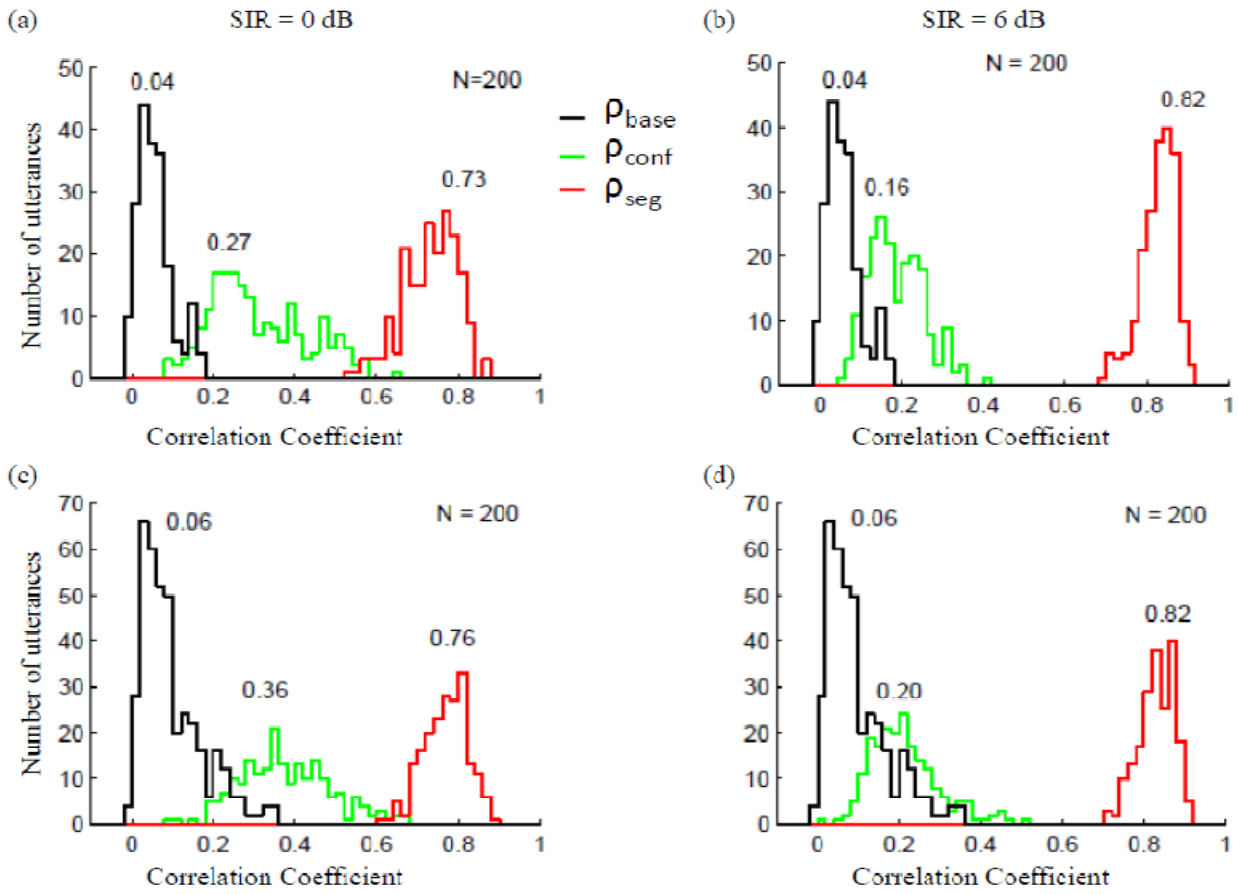

Figure 5.8 Histogram of speech segregation performance for SIR = 0 and 6 dB, respectively. (a) and (b) are results for male+female speech mixtures; (c) and (d) are results for male+male and female+female mixtures. The number indicates the median value of each distribution.

5.4.2.2 Memory-based mask formation

In this experiment, a mask is derived by a pre-trained SVM for each speaker. We train a SVM to classify masks for target against non-target speakers. Utterances of target and multi-interfering speakers from the TIMIT database comprising approximately 3 minutes of speech are used as training data for each target speaker. Mixed utterances are used to train the SVM instead of the original utterances from individual speakers. Mixed utterances are obtained by combining

utterances from the target speaker with those from non-target speakers. In order to make the target and intrusion at the same signal level, all training utterances are normalized to have 0 mean and unit variance before addition. To avoid over fitting, cross-validation is used to evaluate the fitting. Different data sets are used for training, validating, and testing the SVM.

At each time instant, two rows selected by peak-picking of variances (i.e. the diagonal of the coherence matrix) within a feature domain are fed to the SVM. The outputs from the SVM are the class label and the distance to the hyperplane for each input. The SVM gives an error rate of about 6% in classification. The performance of the model from a SVM trained for a female speaker is shown in Figure 5.9 (SIR = 0 dB). Segregation occurs with an accuracy of ρ_{seg}=0.72, which is significantly higher than the baseline, ρ_{base}=0.03 and the confusion, ρ_{conf}=0.25. These results are comparable with those using attention-based mask.

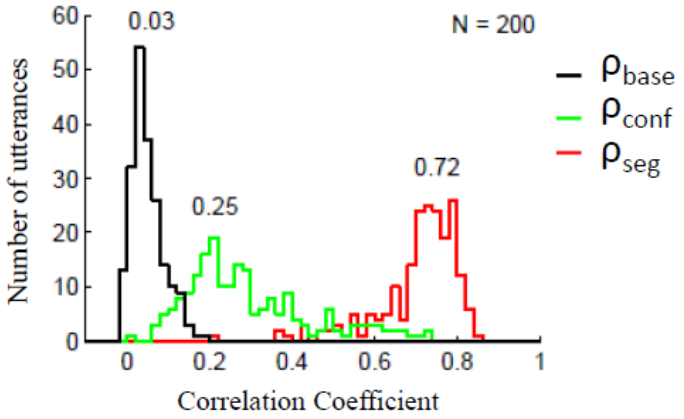

Figure 5.9. Simulation results using a SVM classifier for a female target speaker at 0 dB SIR.

5.5 Summary and Discussion

Inspired by neurobiological findings, we proposed a computational model of auditory scene analysis based on temporal coherence across multi-dimensional auditory representations. The

model can account for the percepts of the commonly used stimuli in auditory scene analysis and successfully perform speech segregation at two-speaker conditions. Our results are comparable with those presented by Elhilali and Shamma (2008). Temporal coherence is the foundation of the model, which provides a binding cue not only for auditory representations linking with a particular attribute, but for various attributes belonging to the same source/stream within and across modalities as well. For instance, paying attention to one pitch, bind not only all harmonics associated with the pitch, but also all coherent attributes (e.g. spatial location, timbre, loudness etc.) belonging to the same source. The model has the flexibility to integrate any other sensible attributes known to promote stream segregation.

Like conventional CASA models, we compute auditory representations of various sound attributes such as frequency, pitch, and spatial location. However, our model is substantially different from previous studies in the way such ASA cues are integrated. The model presented by Tessier and Berthommier (1997) performs double vowel segregation based on pitch and ITD cues. But by selecting the segments generated according to either pitch or ITD cues, the model really does not combine both cues. Several CASA systems are proposed to segregate speech utterances based on harmonicity or amplitude modulation, while binaural cues are used to guide the grouping over time (Kollmeier and Koch, 1994; Okuno et al., 1999; Shamsoddini and Denbigh, 2001). In our model, temporal coherence automatically binds the auditory representations of the diverse attributes of a stream, which allows some features of the stream outside of the attentional focus to contribute to the stream by simply attending to one particular feature.

The idea of temporal coherence as a binding cue in perception has been presented earlier by Malsburg in 1981 and has been implemented in CASA using neural and oscillatory networks (Malsburg and Schneider, 1986; Wang and Brown, 1999). However, in these models, the correlations are induced by *intrinsic* oscillatory activity at the cellular level such as a tendency of cells to form bursts of spikes. By contrast, in our model, the correlations are stimulus-driven, that is caused by the slow phase-locking rates in the cortical responses to sensory signals.

In our model, attention plays a crucial role in the stream formation. Prior to attentional selection, the coherence matrix may be computed by pre-attentive (data-driven) process, providing flexibility of potential decompositions of auditory scenes into streams. A mask corresponding to a particular scene is only formed when attention is applied. However, mask can also be formed based on memory. In this model, we use a SVM to model the process of speaker identification (ID). Basically, a pre-trained SVM of a target speaker classifies the potential masks from the coherence matrix into target and interfering speakers. Finally, for extracting ITD cue, the Jeffress model used in this study is the simplest one, which is adequate for synthetically generated stimuli without the presence of noise and room reverberation. However, in free-field listening conditions, the head-related impulse responses and the precedence effect need to be considered in the model. Further modification of the Jeffress model for improving the accuracy of ITD estimates, such as "stencil" approach (Liu et al., 2000) and "skeleton" cross-correlogram (Palomaki et al., 2004), is required. As in the CASA literature, we assume that sound sources have stationary locations.

Chapter 6 **Conclusions**

6.1 Thesis Overview

This thesis is an attempt to answer the questions stated in Chapter 1: what are the neural correlates of auditory streaming in A1and how attention can modulate the correlates? And furthermore, we propose a neuro-biologically inspired computational model of auditory scene analysis. First, we adapted two auditory perception tasks, used in recent human psychophysical studies, to obtain behavioral measures of auditory streaming in ferrets. One task involved the detection of shifts in the frequency of tones within an alternating tone sequence. The other task involved the detection of a stream of regularly repeating target tones embedded within a randomly varying multi-tone background. In both tasks, performance was measured as a function of various stimulus parameters, which previous psychophysical studies in humans have shown to influence auditory streaming. Ferret performance in the two tasks was found to vary as a function of these parameters, in a way that is qualitatively consistent with the human data. These results suggest that auditory streaming occurs in ferrets, and that the two tasks provide a valuable tool in neurophysiological studies of the phenomenon.

Second, current neuro-computational theories of auditory streaming rely on tonotopic organization of the auditory system to explain the observation that sequential and spectrally distant sound elements tend to form separate perceptual streams. Here we show that spectral components that are well separated in frequency are no longer heard as separate streams if presented synchronously, rather than consecutively. In contrast, responses from neurons in the primary auditory cortex of awake passive ferrets show that both synchronous and asynchronous

tone sequences produce comparably segregated responses along the tonotopic axis. The results argue against tonotopic (spectral) separation *per se* as a neural correlate of stream segregation.

Thirdly, to explore attention effects on streaming, we recorded spiking activity in ferrets A1 under two attentional states: passively listening to the stimuli and attending to the target stream. Attention modulates the correlation of spike trains from pairs of cells in favor of stream segregation. The correlation between cells belonging to the same stream is increased, while the correlation between cells responding to different streams becomes reduced. Furthermore, STRF plasticity reflects those changes in correlation. The strength of the STRF changes is modulated by task difficulty.

Finally, taking into account the above biological findings, we propose a computational model of stream segregation that uses temporal coherence as the primary criterion for predicting stream formation. Channels with high correlations and the same sign coefficients are grouped together. The new model provides a framework which can be used to study and predict the perceptual organization of arbitrary sound combinations, such as speech from multiple talkers or polyphonic music.

6.2 Future Research

We have focused here on a neurophysiological study of selective attention to one stream and showed that neural responses were modulated by attention in ferrets A1. To extend this study, we can explore the effects of switching attention between two streams and how this switching affects the representation of streams. Also we have found that attention can not only modulate

the correlation between cells that had close or overlapping receptive fields, but between distantly related cells as well in the background stream (Chapter 4). Therefore, it would be interesting to see if we can replicate this result for distantly related cells but are still carrying features within the foreground stream.

Noninvasive studies, EEG, MEG, and fMRI, with human subjects indicate that auditory streaming may be strongly related to responses in Heschl's gyrus which incorporate primary and nonprimary areas of the auditory cortex. Neurophysiological experiments in non-human primate A1 and in songbird auditory forebrain, an area that is the homologous to the mammalian A1, have found that neural responses were modulated by the frequency separation of a two-tone alternating sequence. Our experiments thus far have focused on A1. But it has been found that A1 is more stimulus-driven, and areas further down the auditory pathway such as secondary auditory cortices and the prefrontal cortex, may be more selective to categories, concepts, and cognition, more sensitive to streaming, and hence would show stronger modulation by attention.

Finally, to truly combine behavior and physiology in realistic tasks, we need to use more convenient recording technology such as chronic multielectrode recordings that would allow us to use more freely behaving animals, and monitor changes in unit responses rapidly. Only then will it be possible to establish a truly workable model for the study of the neural correlates of streaming in animals.

Appendix 1 Supplementary Figures for Chapter 3

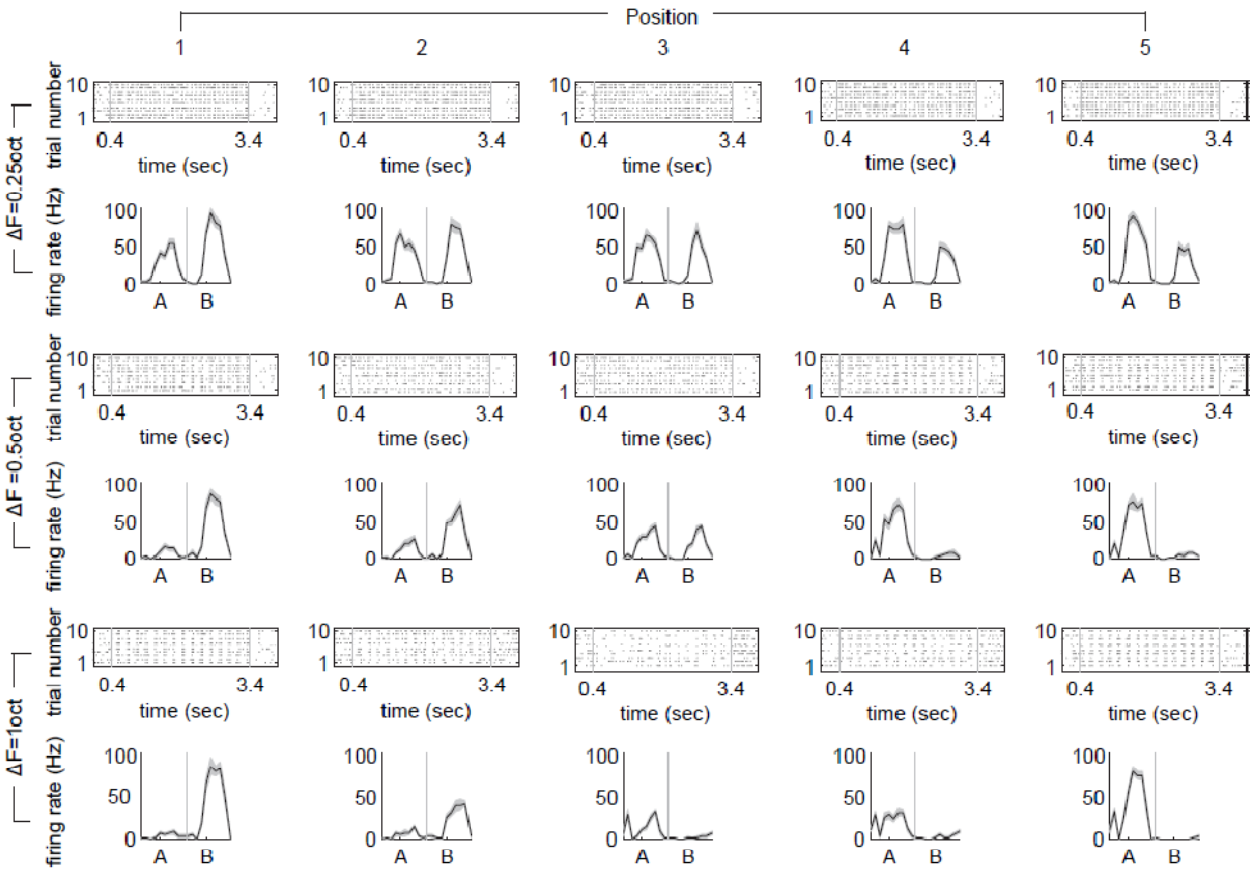

Figure A1.10 Single unit example for alternating sequence in experiment 1 in Chapter 3. Raster and period histogram in two-tone alternating mode for all conditions (3 ΔF x 5 positions). Each condition has 10 repetitions. Each trial includes 0.4 second pre-stimulus silence, 3 second two-tone sequences, and 0.6 second post-stimulus silence. Grey lines in raster indicate stimuli onset and offset. Grey area in period histogram indicates standard error.

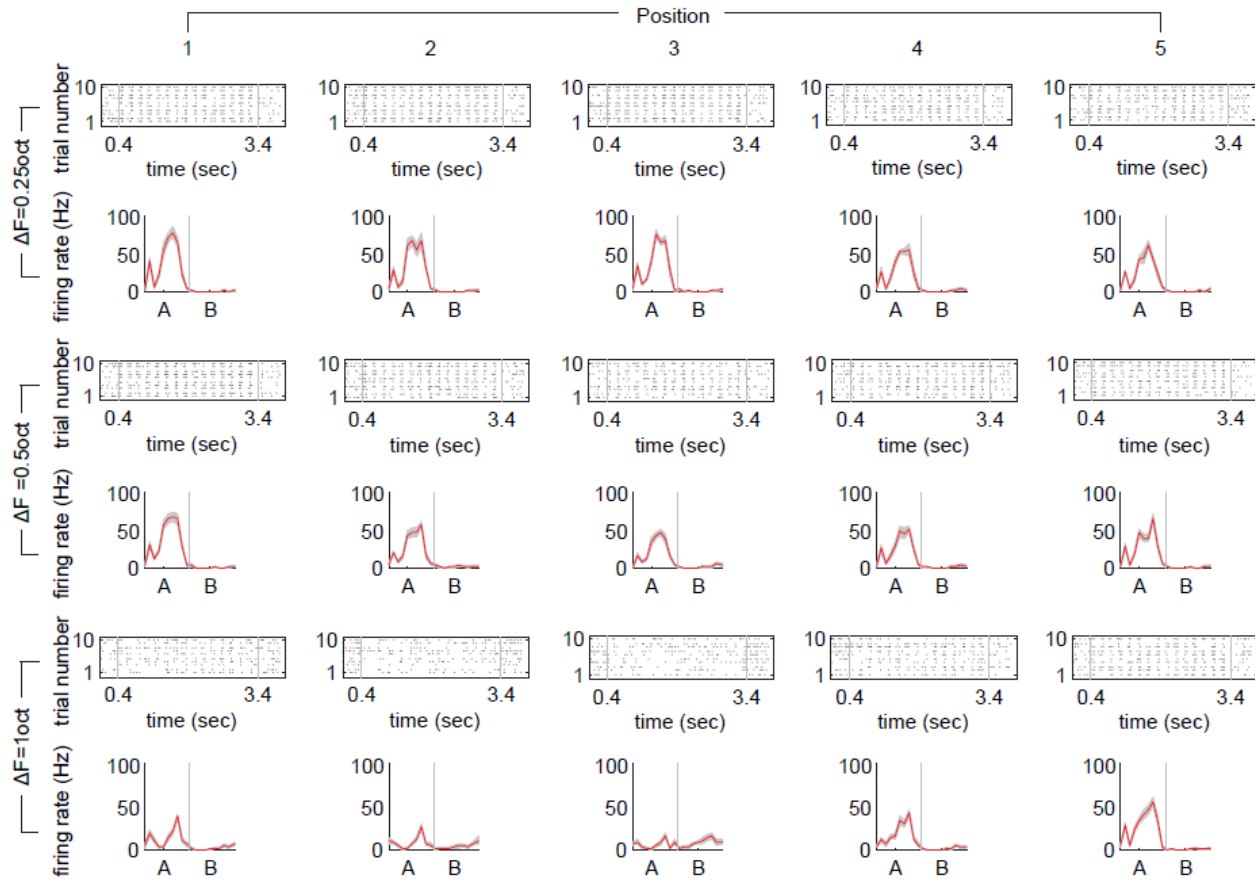

Figure A1.2 Single unit example for synchronous sequence in experiment 1 in Chapter 3. Raster and period histogram in two tone synchronous mode for all conditions (3 ΔF x 5 positions). Each condition has 10 repetitions. Each trial includes 0.4 second pre-stimulus silence, 3 second two-tone sequences, and 0.6 second post-stimulus silence. Grey lines in raster indicate stimuli onset and offset. Grey area in period histogram indicates standard error.

Appendix 2 Supplementary Figures for Chapter 4

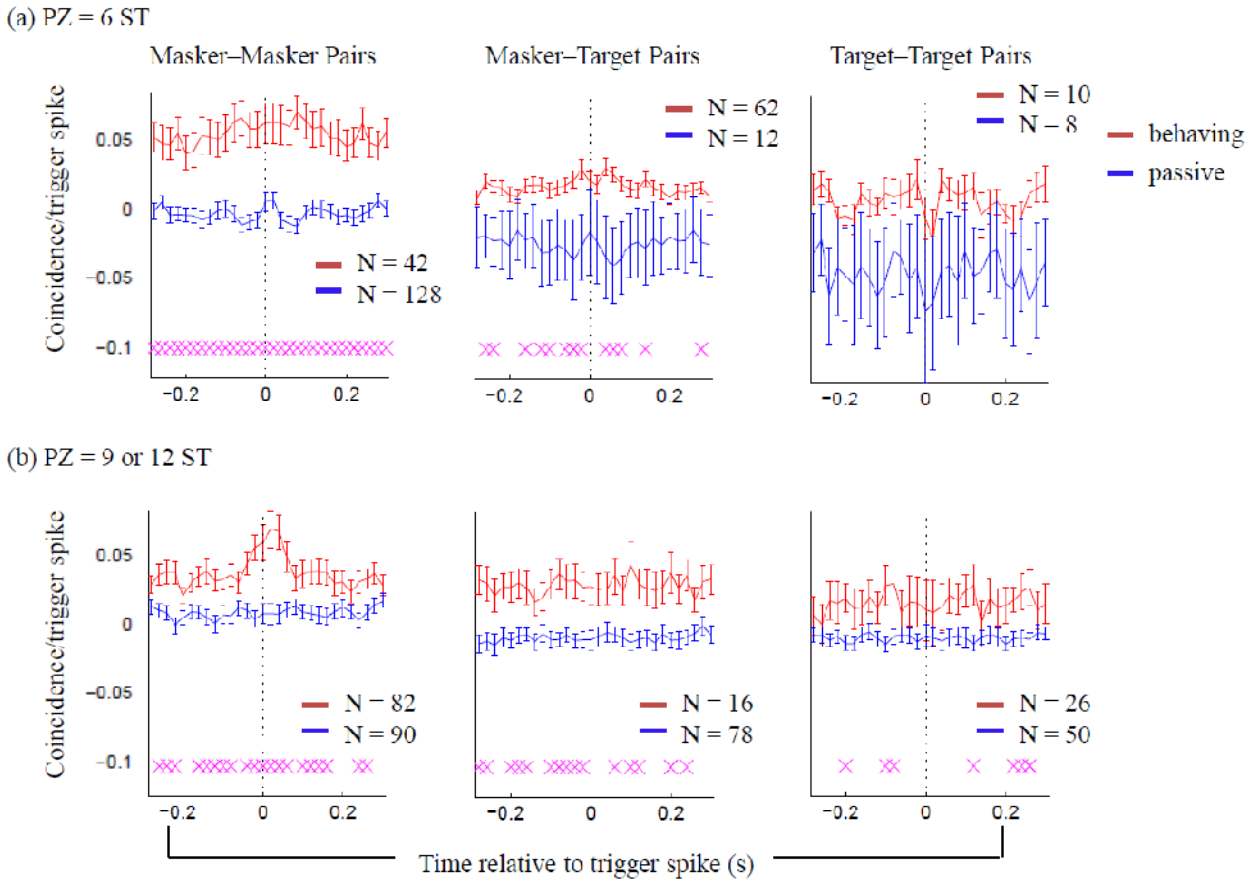

Figure A2.1. The difference between STACs ($STAC_{diff}$) from task stimuli and post during behavioral and passive conditions, respectively. $STAC_{diff}$ (a) at smaller PZ (6 ST); and (b) at larger PZ (9 or 12 ST). Magenta crosses indicate significant difference (t test, $p < 0.01$) at the time between two attentional conditions.

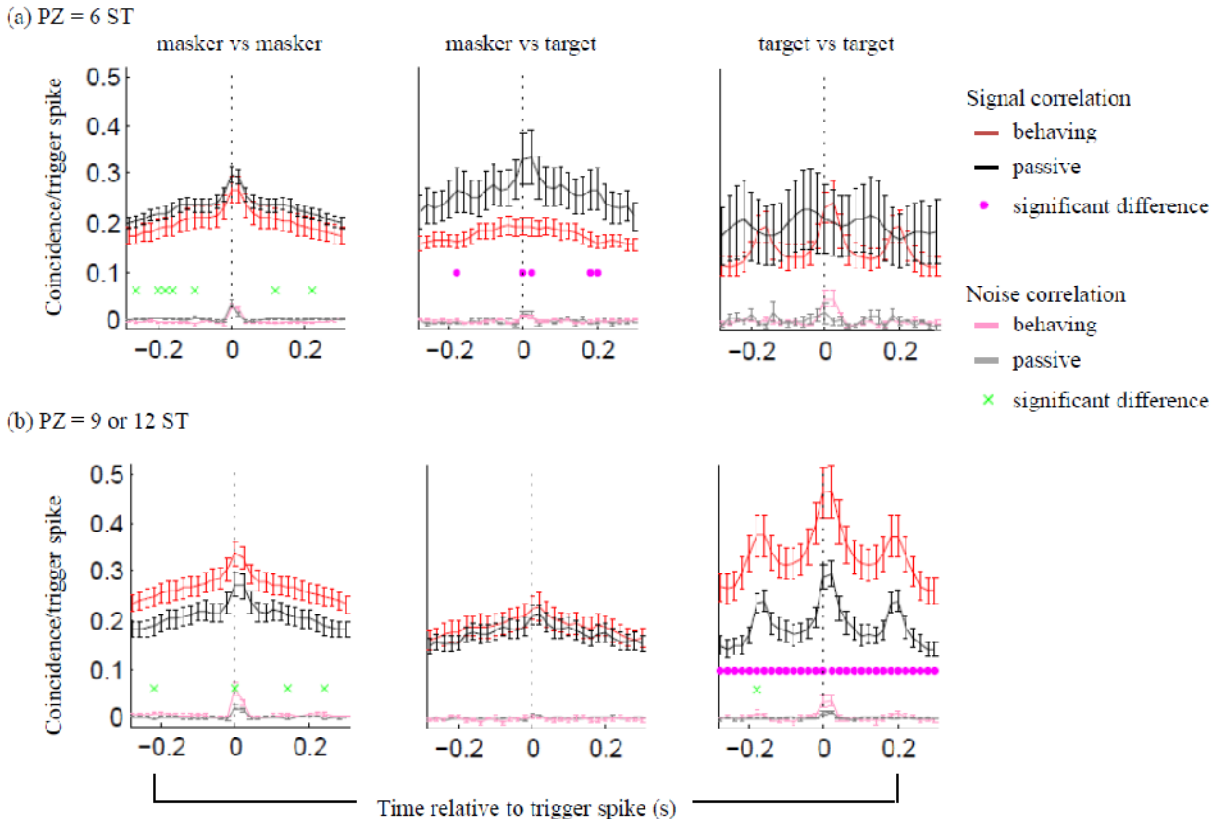

Figure A2.2 Signal and noise correlations during target stimuli under two attentional conditions (a) for smaller PZ, 6 ST and (b) larger PZ, 9 or 12 ST. Error bars indicate standard error (SE). Magenta dots/green crosses indicate significant difference (t test, $p < 0.01$) at the time between the two conditions for signal/noise correlation.

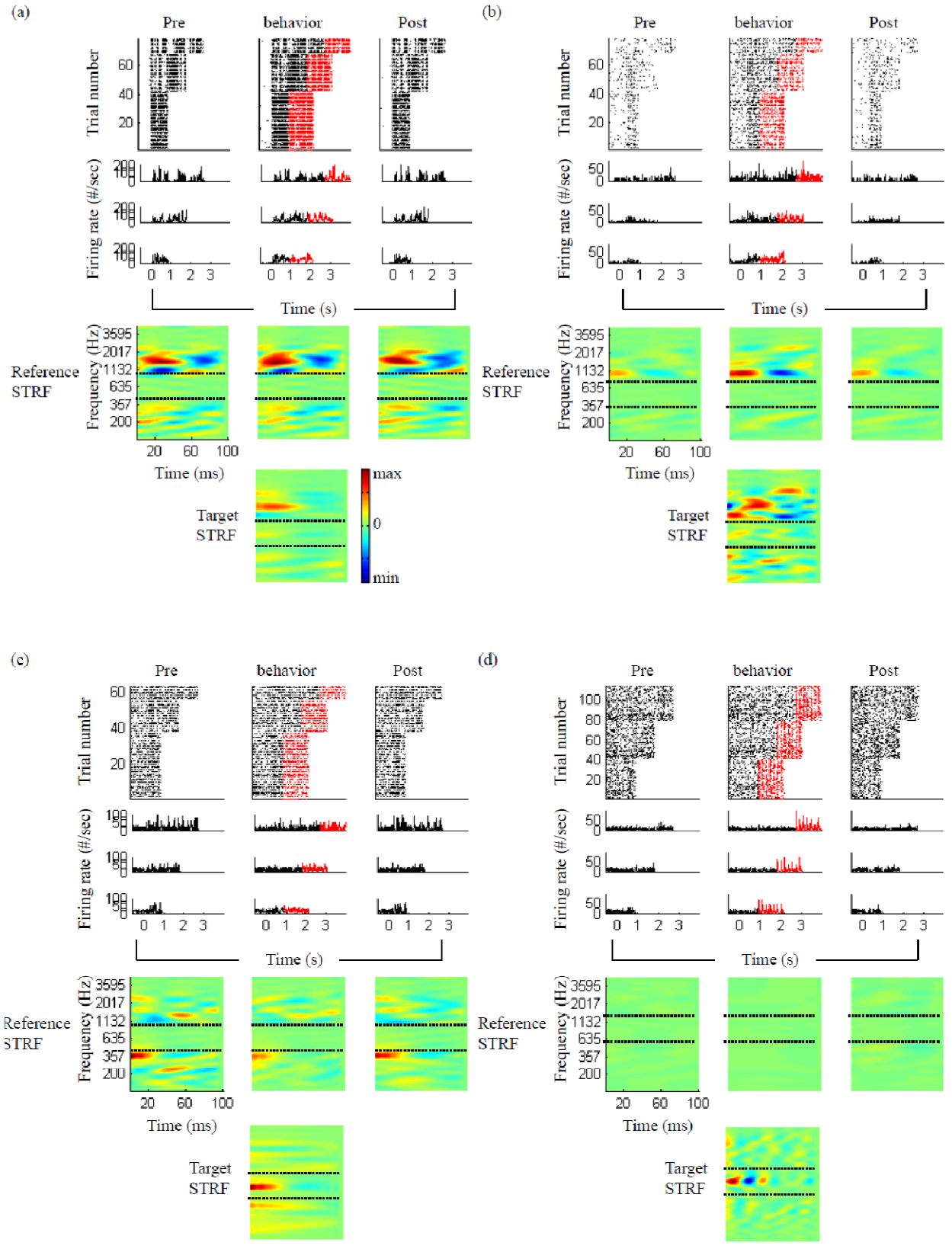
(a)
Pre
behavior
Post
Trial number
Firing rate (#/sec)
Time (s)
Reference STRF
Frequency (Hz)
Time (ms)
Target STRF
max
0
min
(b)
(c)
(d)

Figure A2.3 Examples of single units' raster plot, PSTH plot, and STRF at 9 ST PZ. (a) and (b) are two simultaneously recorded masker cells. (c) and (d) are target cells. The area between two black dash lines represents PZ.

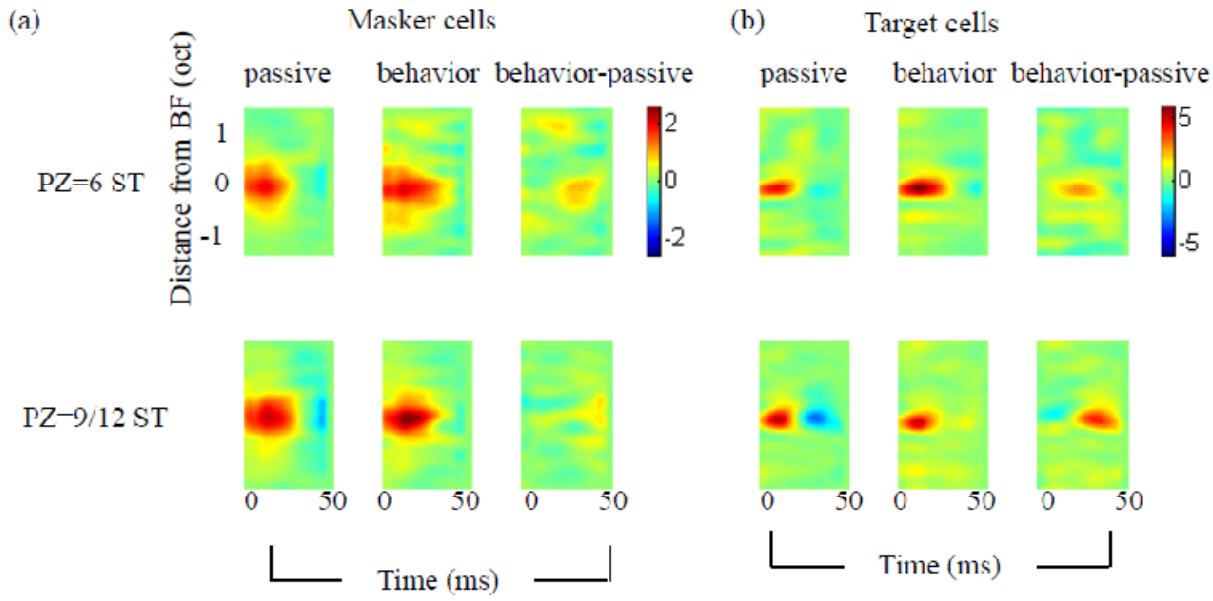

Figure A2.4 Population patterns of target STRF plasticity. Difference between the average STRF from behavior and passive conditions (a) for masker cells; and (b) for target cells.

Bibliography

Alais, D., Blake, R., and Lee, S.H. (1998). Visual features that vary together over time group together over space. *Nat Neurosci 1*, 160-164.

Anstis, S., & Saida, S. (1985). Adaptation to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 257-271.

Atiani, S., Elhilali, M., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron, 61*(3), 467-280.

Barbour, D.L., and Wang, X. (2002). Temporal coherence sensitivity in auditory cortex. *J Neurophysiol 88*, 2684-2699.

Baru, A. V. (1967). *Mechanisms of Hearing (in Russian).* Nauka, Leningrad.

Beauvois, M.W., and Meddis, R. (1991). A computer model of auditory stream segregation. *Q J Exp Psychol A 43*, 517-541.

Beauvois, M.W., and Meddis, R. (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. *J Acoust Soc Am 99*, 2270-2280.

Bee, M. A., & Klump, G. M. (2004). Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. *Journal of Neurophysiology, 92*, 1088-1104.

Bee, M.A., and Klump, G.M. (2005). Auditory stream segregation in the songbird forebrain: effects of time intervals on responses to interleaved tone sequences. *Brain Behav Evol 66*, 197-214.

Bee, M. A., & Micheyl, C. (2008). The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? *Journal of Comparative Psychology, 122*, 235-251.

Belouchrani, A., Abed-Meraim, K., and Cardoso, J. F. (1997). A blind source separation technique based on second order statistics. *IEEE Trans. Signal Process.*, 45(2), 434-444.

Bendor, D., and Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex. *Nature 436*, 1161-1165.

Bendor, D., and Wang, X. (2006). Cortical representations of pitch in monkeys and humans. Curr Opin Neurobiol *16*, 391-399.

Bendor, D., and Wang, X. (2007). Differential neural coding of acoustic flutter within primate auditory cortex. *Nat Neurosci 10*, 763-771.

Bey, C., & McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Perception & Psychophysics, 64*, 844-854.

Bey, C., & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology: Human Perception & Performance, 29*, 267-279.

Blake, R., and Lee, S.H. (2005). The role of temporal structure in human vision. *Behavioral and cognitive neuroscience reviews 4*, 21-42.

Bracewell, R. (1999). *The Fourier Transform and Its Applications*, third edn (, McGraw-Hill).

Bregman, A. S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception & Performance, 4*, 380-387.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organisation of Sound*. Cambridge, MA: MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89, 244-249.

Broadbent, D.E., and Ladefoged, P. (1959). Auditory Perception of Temporal Order. *J Acoust Soc Am 31*, 1539.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. *Acustica*, 86, 117–128.

Brown, G. J. and Cooke M. P.(1994). Computational auditory scene analysis. *Comput. Speech Lang.*, 8, 297–336.

Brown, G. J., Harding, S., and Barker J. P. (2006). Speech separation based on the statistics of binaural auditory features. *Proc. ICASSP '06*, 5.

Cardoso, J. F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, 9(10), 2009-2025.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends Cogn Sci*, *8*(10), 465-471.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception & Performance, 27*, 115-127.

Carlyon, R. P., & Gockel, H. (2008). Effects of harmonicity and regularity on the perception of sound sources. In W. A., Yost R. R. Fay, & A. N. Popper (Eds.), *Auditory perception of sound sources*. New York: Springer.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America, 25*, 975-979.

Chi, T., Gao, Y., Guyton, M.C., Ru, P., and Shamma, S. (1999). Spectro-temporal modulation transfer functions and speech intelligibility. *J Acoust Soc Am 106*, 2719-2732.

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am*. 118, 887-906.

Ciocca, V., and Darwin, C.J. (1993). Effects of onset asynchrony on pitch perception: adaptation or grouping? *J Acoust Soc Am 93*, 2870-2878.

Dai, H. P., Scharf, B., & Buus, S. (1991). Effective attenuation of signals in noise under focused attention. *Journal of the Acoustical Society of America, 89*, 2837-42.

Darwin, C.J., and Carlyon, R.P. (1995a). Auditory grouping. In *Hearing*, B.C.J. Moore, ed. (Orlando, FL: Academic Press), pp. 387-424.

Darwin, C.J., and Carlyon, R.P. (1995b). Auditory grouping. In *Hearing*, B. Moore, ed. (London: Academic Press), pp. 387-424.

David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Dynamics of rapid plasticity during positively and negatively reinforced behavior in auditory cortex. *Proceedings of the 31th Mid-winter meeting of the Association for Research in Otolaryngology*, pp. 115.

Dowling, W. J. (1968). Rhythmic fission and perceptual organization. *Journal of the Acoustical Society of America, 44*, 369.

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology, 5*, 322-337.

Eddins, D. A., and Bero, E. M. (2007). Spectral modulation detection as a function of modulation frequency, carrier bandwidth, and carrier frequency region. *J. Acoust. Soc. Am.* 121, 363-372.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Eggermont, J.J. (2001). Between sound and perception: reviewing the search for a neural code. *Hear Res 157*, 1-42.

Elhilali, M., Fritz, J. B., Chi, T., and Shamma, S. A. (2007). Auditory cortical receptive fields: stable entities with plastic abilities. *J. Neurosci.* 27, 10372-10382.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron, 61*, 317-329.

Elhilali, M. and Shamma, S. A. (2008). A Cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *J. Acoust. Soc. Am.* 124(6), 3751-3771.

Elliott, D. N., Stein, L. & Harrison, M. J. (1960). Determination of absolute-intensity thresholds and frequency-difference thresholds in cat. *Journal of the Acoustical Society of America, 32*, 380-384.

Fahle, M. (1993). Figure-ground discrimination from temporal information. *Proc. R. Soc. London Ser. B 254,* 199-203.

Fay, R. R. (1988). Comparative psychoacoustics. *Hearing Research, 34,* 295-306.

Fay, R. R. (1998). Auditory stream segregation in goldfish (Carassius auratus). *Hearing Research, 120*(1-2), 69-76.

Fay, R. R. (2000). Spectral contrasts underlying auditory stream segregation in goldfish. *Journal of the Association for Research in Otolaryngology, 1,* 120-128.

Fay, R. R. (2008). Sound source perception and stream segregation in nonhuman vertebrate animals. In W. A. Yost, R. R. Fay, & A. N. Popper (Eds.), *Auditory perception of sound sources* (pp. 307-323). New York: Springer.

Fevotte, C. and Doncarli, C. (2004). Two contributions to blind source separation using time-frequency distributions. *IEEE Signal Process. Lett.,* 11(3), 386-389.

Fevotte, C. and Godsill, S. J. (2006). A Bayesian approach for blind separation of sparse sources. *IEEE Trans. On Audio, Speech, and Language Process.,* 14(6), 2174-2188.

Fishman, Y.I., Arezzo, J.C., and Steinschneider, M. (2004). Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. *J Acoust Soc Am 116,* 1656-1670.

Fishman, Y. I., Reser, D. H., Arezzo, J. C., & Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research, 151*(1-2), 167-187.

Formby, C., Sherlock, L.P., and Li, S. (1998). Temporal gap detection measured with multiple sinusoidal markers: effects of marker number, frequency, and temporal position. *J Acoust Soc Am 104*, 984-998.

Fries, P., Roelfsema, P.R., Engel, A.K., Konig, P., and Singer, W. (1997). Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America 94*, 12699-12704.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention--focusing the searchlight on sound. *Curren Opin Neurobiol, 17*(4), 437-455.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007a). Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1? *Hear. Res.* 229, 186-203.

Fritz, J. B., Elhilali, M., & Shamma, S. A. (2007b). Adaptive changes in cortical receptive fields induced by attention to complex sounds. *J. Neurophysiol.* 98, 2337-2346.

Fritz, J., Elhilali, M., and Shamma, S. (2005a). Active listening: task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Hear Res 206*, 159-176.

Fritz, J. B., Shamma, S. A., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience, 6*(11), 1216-1223.

Fritz, J.B., Elhilali, M., and Shamma, S.A. (2005b). Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. *J Neurosci 25*, 7623-7635.

Godsmark, D. J. and Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis.*Speech Communication*, 27, 351–366.

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *J. Acoust. Soc. Am*. 54, 1496-1516.

Golub, G.H., and Van Loan, C.F. (1996). *Matrix Computations*, 3 edn (Baltimore, MD: Johns Hopkins University Press).

Gourevitch, G. (1970). Delectability of tones in quiet and in noise by rats and monkeys. In W. C. Stebbins (Ed.), *Animal Psychophysics: The Design and Conduct of Sensory Experiments* (pp. 67-97). New York: Plenum.

Gray, C. M. and Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Nat. Acad. Sci*. USA, 86, 1698-1702.

Gray, C. M., Koenig, P., Engel, A. K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338, 334-337.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Krieger.

Green, D. M. (1986). *Auditory frequency selectivity*, NATO ASI Series, Series A: Life Sciences (Plenum, New York, NY), 351-359.

Greenberg, G. S., & Larkin, W. D. (1968). Frequency-response characteristics of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *Journal of the Acoustical Society of America, 44*, 1513-1523.

Griffiths, T.D., and Warren, J.D. (2004). What is an auditory object? *Nat Rev Neurosci 5*, 887-892.

Grimault, N., Bacon, S.P., and Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *J Acoust Soc Am 111*, 1340-1348.

Gross, J., Schnitzler, A., Timmermann, L., and Ploner, M. (2007). Gamma oscillations in human primary somatosensory cortex reflect pain perception. *PLoS biology 5*, e133.

Gutschalk, A., Oxenham, A. J., Micheyl, C., Wilson, E. C., & Melcher, J. R. (2007). Human cortical activity during streaming without spectral cues suggests a general substrate for auditory stream segregation. *The Journal of Neuroscience, 27*, 13074-13081.

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M., & Oxenham, A. J. (2005). Neuromagnetic correlates of streaming in human auditory cortex. *The Journal of Neuroscience*, *25*, 5382-5388.

Gutschalk, A., Micheyl, C., & Oxenham, A. J. (2008). Neural correlates of auditory perceptual awareness under informational masking. *PLoS Biology*, *6*, e138.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

Hartmann, W., and Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception 9*, 155-184.

Hartmann, W., Johnson, D (1991). Stream segregation and peripheral channeling. *Music Perception 9*, 155-184.

Heffner, R. S., Heffner, H. E., & Masterton, B. (197l). Behavioral measurements of absolute and frequency difference thresholds in guinea pig. *Journal of the Acoustical Society of America, 49,* 1888-1895.

Hu, G. and Wang, D. L. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Networks*, 15 (5) 1135–1150.

Hu, G. and Wang, D. L. (2007). Auditory segmentation based on onset and offset analysis. *IEEE Trans. Audio, Speech and Language Process.*, 15, 396–405.

Hulse, S. H., MacDougall-Shackleton, S. A., & Wisniewski, A. B. (1997). Auditory scene analysis by songbirds: stream segregation of birdsong by European starlings (Sturnus vulgaris). *Journal of Comparative Psychology, 111,* 3-13.

Itatani, N., & Klump, G. M. (2009). Auditory streaming of amplitude-modulated sounds in the songbird forebrain. *Journal of Neurophysiology, 101*, 3212-3225.

Izumi, A. (2002). Auditory stream segregation in Japanese monkeys. *Cognition, 82*, 113-122.

Jeffress, L. A. (1948). A place theory of sound localization. Journal of Comparative and *Physiological Psychology*, 41, 35–39.

Kalluri, S., Depireux, D. A., & Shamma, S. A. (2008). Perception and cortical neural coding of harmonic fusion in ferrets. *Journal of the Acoustical Society of America, 123*(5), 2701-2716.

Kanwal, J. S., Medvedev, A. V., & Micheyl, C. (2003). Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment. *Network, 14*, 413-435.

Karjalainen, M. and Tolonen, T. (1999). Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis. *Proc. IEEE ICASSP*, 2, 929–932.

Kelly, J. B. (1970). Effects *of lateral lemniscal and neocortical lesions on auditory absolute thresholds and frequency difference thresholds of the rat*, Ph.D. Thesis, Vanderbilt University.

Kidd, G. J., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *Journal of the Acoustical Society of America, 95*, 3475-3480.

Kidd, G. J., Mason, C. R., & Richards, V. M. (2003). Multiple bursts, multiple looks, and stream coherence in the release from informational masking. *Journal of the Acoustical Society of America, 114*, 2835-2845.

Kidd, G. J., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost, R. R. Fay, & A. N. Popper (Eds.), *Auditory Perception of Sound Sources* (pp. 143-189). New York: Springer.

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: a commentary. *Perception & Psychophysics, 63*, 1421-55.

Kollmeier, B. and Koch, R. (1994). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *J. Acoust. Soc. Am.*, 95, 1593–1602.

Kowalski, N., Depireux, D.A., and Shamma, S.A. (1996a). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *J Neurophysiol 76*, 3503-3523.

Kowalski, N., Depireux, D.A., and Shamma, S.A. (1996b). Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. *J Neurophysiol 76*, 3524-3534.

Lakatos, P., Shah, A.S., Knuth, K.H., Ulbert, I., Karmos, G., and Schroeder, C.E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J Neurophysiol 94*, 1904-1911.

Liang, H., Bressler, S.L., Ding, M., Desimone, R., and Fries, P. (2003). Temporal dynamics of attention-modulated neuronal synchronization in macaque V4. *Neurocomputing 52-54*, 481 - 487.

Liu, C., Wheeler, B. C. O'Brien, W. D., Bilger, R. C., Lansing, C. R., and.Feng, A. S (2000).Localization of multiple sound sources with two microphones. *J. Acoust. Soc. Am.*, 108 (4), 1888–1905.

Logothetis, N. K., & Schall, J. D. (1989). Neuronal correlates of subjective visual perception. *Science, 245*(4919), 761-763.

Lu, T., Liang, L., and Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat. Neurosci*. 11, 1131-1138.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Lyon, R. F. (1983). A computational model of binaural localization and separation. *Proc. IEEE ICASSP*, 1983, pp. 1148–1151.

Ma, L., Micheyl, C., Yin, P. Oxenham, A. J., and Shamma, S. A. (2010). Behavioral measures of auditory streaming in ferrets (Mustela putorius). *Journal of Comparative Psychology*. 124(3): 317-30.

MacDougall-Shackleton, S. A., Hulse, S. H., Gentner, T. Q., & White, W. (1998). Auditory scene analysis by European starlings (Sturnus vulgaris): perceptual segregation of tone sequences. *Journal of the Acoustical Society of America*, *103*, 3581-3587.

Macken, W.J., Tremblay, S., Houghton, R.J., Nicholls, A.P., and Jones, D.M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of experimental psychology 29*, 43-51.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman.

McCabe, S., and Denham, M.J. (1997). A model of auditory streaming. *J Acoust Soc Am 101*, 1611-1621.

Meador, K.J., Ray, P.G., Echauz, J.R., Loring, D.W., and Vachtsevanos, G.J. (2002). Gamma coherence and conscious perception. *Neurology 59*, 847-854.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., and Rodriguez, E. (2007). Synchronization of neural activity across cortical areas correlates with conscious perception. *J Neurosci 27*, 2858-2865.

Micheyl, C., & Carlyon, R. P. (1998). Effects of temporal fringes on fundamental-frequency discrimination. *Journal of the Acoustical Society of America, 104,* 3006-3018 .

Micheyl, C., Carlyon, R. P., Cusack, R., & Moore, B. C. J. (2005). Performance measures of auditory organization. In D. Pressnitzer, A. de Cheveigné, S. McAdams & L. Collet (Eds.), *Auditory Signal Processing: Physiology, Psychoacoustics, and Models* (pp. 203-211). New York: Springer.

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., Oxenham, A. J., Rauschecker, J. P., et al. (2007). The role of auditory cortex in the formation of auditory streams. *Hearing Research, 229,* 116-131.

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219,* 36-47.

Micheyl, C., Hunter, C., & Oxenham, A. J. (2009). Auditory stream segregation and the perception of across-frequency synchrony. *Journal of Experimental Psychology: Human Perception and Performance,* in press.

Micheyl, C., Moore, B. C., & Carlyon, R. P. (1998). The role of excitation-pattern cues and temporal cues in the frequency and modulation-rate discrimination of amplitude-modulated tones. *Journal of the Acoustical Society of America, 104,* 1039-1050.

Micheyl, C., Shamma, S., & Oxenham, A. J. (2007). Hearing out repeating elements in randomly varying multitone sequences: a case of streaming? In B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp & J. Verhey (Eds.), *Hearing - from basic research to applications* (pp. 267-274), Berlin: Springer.

Micheyl, C., Tian, B., Carlyon, R. P., & Rauschecker, J. P. (2005). Perceptual organization of tone sequences in the auditory cortex of awake macaques. *Neuron, 48,* 139-148.

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annu Rev Neurosci 24*, 167-202.

Miller, G. A., & Heise, G. A. (1950). The trill threshold. *Journal of the Acoustical Society of America, 22*, 637-638.

Miller, L.M., Escabi, M.A., Read, H.L., and Schreiner, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J Neurophysiol 87*, 516-527.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing*. 5th ed. London: Academic Press.

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica united with Acustica, 88,* 320-333.

Moore, B. C. J. (1973). Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America*, *54,* 610-619.

Moore, B. C. J., Hafter, E. R., & Glasberg, B. R. (1996). The probe-signal method and auditory-filter shape: results from normal- and hearing-impaired subjects. *Journal of the Acoustical Society of America*, *99,* 542-552.

Neff, D. L., & Green, D. M. (1987). Masking produced by spectral uncertainty with multicomponent maskers. *Perception & Psychophysics*, *41,* 409-415.

Neff, D.L., Jesteadt, W., and Brown, E.L. (1982). The relation between gap discrimination and auditory stream segregation. *Percept Psychophys 31,* 493-501.

Nelson, D. A., & Kiester, T. E. (1978). Frequency discrimination in the chinchilla. *Journal of the Acoustical Society of America, 64,* 114-126.

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Curr Opin Neurobiol 14*, 474-480.

Nelken, I., Rotman, Y., and Bar Yosef, O. (1999). Responses of auditory-cortex neurons to structural features of natural sounds. *Nature 397*, 154-157.

Nienhuys, T. W., & Clark, G. M. (1979). Critical bands following the selective destruction of cochlear inner and outer hair cells. *Acta Oto-Laryngol, 88,* 350-358.

Noreña, A. J., Gourévitch, B., Pienkowski, M., Shaw, G., & Eggermont, J. J. (2008). Increasing spectrotemporal sound density reveals an octave-based organization in cat primary auditory cortex. *The Journal of Neuroscience*, 28, 8885-8896.

Noreña, A. J, Gourévitch, B., Aizawa, N., & Eggermont, J. J. (2006). Spectrally enhanced acoustic environment disrupts frequency representation in cat auditory cortex. *Nature Neuroscience, 9,* 932-939.

Okuno, H. G., Nakatani, T., and Kawabata, T. (1999). Listening to two simultaneous speeches. *Speech Communication*, 27, 299–310.

Ottaviani, L. and Rocchesso, D. (2001). Separation of speech signal from complex auditory scenes. *Proceedings of the Conference on Digital Audio Effects*, 2001, Limerick, Ireland.

Palomaki, K. J., Brown, G. J., and.Wang, D. L. (2004). A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication,* 43 (4), 273–398.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am., 60*(4), 911–918.

Pham, D. T. and Cardoso, J. F. (2001). Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Process.,* 49(9), 1837-1848.

Pickles, J. O. (1979). Psychophysical frequency resolution in the cat as determined by simultaneous masking and its relation to auditory-nerve resolution. *Journal of the Acoustical Society of America, 66,* 1725-1732.

Pickles, J.O. (1988). *An Introduction to the Physiology of Hearing*, 2 edn (Academic Press).

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Curren Biology, 18*, 1124-1128.

Prosen, C. A., Moody, D. B., Sommers, M. S., & Stebbins, W. C. (1990). Frequency discrimination in the monkey. *Journal of the Acoustical Society of America, 88*(5), 2152-2158.

Quatieri, T. F. (2002). 2-D Processing of speech with application to pitch estimation. *ICLSP*, September 2001.

Raj, B., Shashanka, M. V. S., and Smaragdis, P. (2006). Latent dirichlet decomposition for single channel speaker separation. *IEEE ICASSP*, V-821-824.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 106,* 261-300.

Reddy, A. M. and Raj, B. (2007). Soft mask methods for single-channel speaker separation. *IEEE Trans. On Audio, Speech, and Language Process.,* 15(6), 1766-1776.

Richards, V. M., & Tang, Z. (2006). Estimates of effective frequency selectivity based on the detection of a tone added to complex maskers. *Journal of the Acoustical Society of America, 119,* 1574-1584.

Roberts, B., Glasberg, B. R., & Moore, B. C. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *Journal of the Acoustical Society of America, 112,* 2074-2085.

Roberts, B., Glasberg, B. R., & Moore, B. C. (2008). Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. *Journal of Experimental Psychology: Human Perception & Performance, 34*, 992-1006.

Roman, N. Wang D. L. and Brown, G. J. (2003). Speech segregation based on sound localization. *J. Acoust. Soc. Am.,* 114 (4), 2236–2252.

Rose, M. M., & Moore, B. C. J. (2000). Effects of frequency and level on auditory stream segregation. *Journal of the Acoustical Society of America, 108,* 1209-14.

Rose, M. M., & Moore, B. C. J. (2005). The relationship between stream segregation and frequency discrimination in normally hearing and hearing-impaired subjects. *Hearing Research, 204,* 16-28.

Ru, P. (2000). *Perception-based multi-resolution auditory processing of acoustic signal*. Ph.D. Thesis, University of Maryland, College Park, MD.

Schreiner, C.E. (1998). Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. *Audiol Neurootol 3*, 104-122.

Schreiner, C.E., and Sutter, M.L. (1992). Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings. *J Neurophysiol 68*, 1487-1502.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends Cogn Sci 12*, 106-113.

Seaton, W. H., & Trahotis, C. (1975). Comparison of critical ratios and critical bands in the monaural chinchilla. *Journal of the Acoustical Society of America*, *57,* 193-199.

Sek, A., & Moore, B. C. J. (1995). Frequency discrimination as a function of frequency, measured in several ways. *Journal of the Acoustical Society of America*, *97*, 2479–2486.

Shackleton, T. M., Meddis, R., Hewitt, M. J. (1992). Across frequency integration in a model of lateralization. *J. Acoust. Soc. Amer*. 91(4), 2276-2279.

Shamma, S. A., Elhilali, M., and Micheyl, C. (2010). Temporal coherence and attention on auditory scene analysis. *Trends in Neurosciences*, 34(3), 114-123.

Shamma, S. A. and Klein, D. J. (2000). The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *J. Acoust. Soc. Am.* 107, 2631-2644.

Shamsoddini, A. and Denbigh, P. N. (2001). A sound segregation algorithm for reverberant conditions. *Speech Communication*, 33, 179–196.

Shera, C. A., Guinan, J. J., & Oxenham, A. J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proceedings of the National Academy of Sciences USA, 99*, 3318-3323.

Shinn-Cunningham, B. G. (2005). Influences of spatial cues on grouping and understanding sound. *Proceedings of the Forum Acusticum*.

Sinnott, J., & Brown, C. (1993). Effects of varying signal and noise levels on pure tone frequency discrimination in humans and monkeys. *Journal of the Acoustical Society of America, 93*, 1535-1540.

Sinnott, J. M., Brown, C. H., & Brown, F. E. (1992). Frequency and intensity discrimination in Mongolian gerbils, African monkeys and humans. *Hearing Research, 59*(2), 205–212.

Sinnott, J. M., Petersen, M. R., & Hopp S. L. (1985). Frequency and intensity discrimination in humans and monkeys. *Journal of the Acoustical Society of America, 78*(6), 1977-1985.

Slaney, M. and Lyon, R. F. (1990). A perceptual pitch detector. *Proc. IEEE ICASSP*, 1, 357–360.

Sloan, A. M., Dodd, O. T., & Rennaker, R. L. (2009). Frequency discrimination in rats measured with tone-step stimuli and discrete pure tones. *Hearing Research, 251*(1-2), 60-69.

Snyder, J. S., & Alain, C. (2007). Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin, 133*, 780-799.

Snyder, J. S., Alain, C., & Picton, T. W. (2006). Effects of attention on neuroelectric correlates of auditory stream segregation. *Journal of Cognitive Neuroscience, 18*, 1-13.

Srinivasan, S., Roman, N., and Wang, D. L. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Commun.*, 48, 1486–1501.

Sussman, E., Ritter, W., & Vaughan, H. G. J. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology, 36*(1), 22-34.

Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception and Psychophysics, 69*, 136-152.

Sutter, M.L. (2005). Spectral processing in the auditory cortex. Int Rev Neurobiol *70*, 253-298.

Syka, S., Rybalko, N., Brozˇek, G., & Jilek, M., (1996). Auditory frequency and intensity discrimination in pigmented rats. *Hearing Research, 100*(1-2), 107–113.

Talwar, S. K., & Gerstein, G. L. (1998). Auditory frequency discrimination in the white rat. *Hearing Research, 126*(1-2), 135-150.

Talwar, S. K., & Gerstein, G. L. (1999). A signal detection analysis of auditory-frequency discrimination in the rat. *Journal of the Acoustical Society of America, 105*(3), 1784-1800.

Tessier, E., and Berthommier, F. (1997). A Model of the Cumulative Effect of Pitch and Interaural Delay Differences for Double Vowel Segregation. *ICSP'97*, pp. 753-758, Seoul.

Theodoridis, S. and Koutroumbas, K. (2006). *Pattern Recognition*, Elsevier Academic Press.

Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron 24*, 105-110, 111-125.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*, Ph.D. Thesis, University of Technology, Eindhoven.

van Noorden, L. P. A. S. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences ABAB. *Journal of the Acoustical Society of America, 61*, 1041-1045.

Versnel, H., Shamma, S.A., and Kowalski, N. (1995). Ripple Analysis in the Ferret Primary Auditory Cortex. III. Topographic and Columnar Distribution of Ripple Response. *J Aud Neurosc*, 271-285.

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *J. Acoust. Soc. Am.* 66, 1364-1380.

Viemeister, N. F., & Wakefield, G. H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America, 90*, 858-865.

Vishnubhotla, S. and Espy-Wilson, C. Y. (2009). An algorithm for speech segregation of co-channel speech. *IEEE ICASSP,09,* 109-112.

Vliegen, J., Moore, B. C., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *Journal of the Acoustical Society of America, 106*, 938-945.

Vliegen, J., and Oxenham, A.J. (1999). Sequential stream segregation in the absence of spectral cues. *J Acoust Soc Am 105*, 339-346.

Von der Malsburg, C. (1981). The correlation theory of brain function. *Max Planck Inst. Biophys. Chem.*, Gottingen, Germany, Rep. 81-2.

Von der Malsburg, C. and Schneider, W. (1986). A neural cocktail-party processor. *Biol. Cybern.*, 54, 29-40.

Walker, K. M., Schnupp, J. W., Hart-Schnupp, S. M., King, A. J., & Bizley, J. K. (2009). Pitch discrimination by ferrets for simple and complex sounds. *Journal of the Acoustical Society of America* 126, 1321-1335.

Wang, D. L. and Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Networks*, 10 (3), 684–697.

Wang, K. and Shamma, S. A. (1994). Self-normalization and noise-robustness in early auditory representations. *IEEE Trans. on Speech and Audio Process.,* 2(3), 421-435.

Wang, K. and Shamma, S. A. (1995). Spectral shape analysis in the central auditory system. *IEEE Trans. Speech Audio Process.,* 3, 382-395.

Wang, T. T. and Quatieri, T. F. (2009a). 2-D Processing of speech for multi-pitch analysis. *Interspeech ISCA* September 6-10, 2009 Brighton, UK. 2827-2830.

Wang, T. T. and Quatieri, T. F. (2009b). Towards co-channel speaker separation by 2-D demodulation of spectrograms. *IEEE workshop on Applications of Signal Process. To Audio and Acoustics.* October 18-21, 2009, New Paltz, NY.

Warren, R.M., Obusek, C.J., Farmer, R.M., and Warren, R.P. (1969). Auditory sequence: confusion of patterns other than speech or music. *Science 164*, 586-587.

Watson, C. S., Kelly, W. J., & Wroton, H. W. (1976). 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*, 1176-1186.

Weintraub, M. (1985). *A theory and computational model of monaural auditory sound separation.* Ph. D. Thesis, Standford University.

Wier, C. C., Jesteadt, W., & Green, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America, 61*, 178-184.

Wilson, E. C., Melcher, J. R., Micheyl, C., Gutschalk, A., & Oxenham, A. J. (2007). Cortical FMRI activation to sequences of tones alternating in frequency: relationship to perceived rate and streaming. *Journal of Neurophysiology, 97*, 2230-2238.

Wisniewski, A. B., & Hulse, S. H. (1997). Auditory scene analysis in European starlings (Sturnusvulgaris): Discrimination of song segments, their segregation from multiple and reversed conspecific songs, and evidence for conspecific song categorization. *Journal of Comparative Psychology*, 111, 337-350.

Wu, M., Wang, D. L., and Brown, G. J. (2003). A multipitch tracking algorithm for noisy speech. *IEEE Trans. Speech Audio Proc*., 11 (3), 229–241.

Wrigley, S. N. and Brown, G. J. (2004). A computational model of auditory selective attention. *IEEE Trans. Neural Networks*, 15 (5), 1151–1163.

Yang, X., Wang, K. and Shamma, S. A. (1992). Auditory representations f acoustic signals. *IEEE Trans. on Information Theory*, 38(2), 824-839.

Yin, P., Fritz, J. B., & Shamma, S. A. (2010) Do ferrets perceive relative pitch. *J Acoust Soc Am* 127(3), 1673-1680..

Yin, P., Ma, L., Elhilali, M., Fritz, J., and Shamma, S.A. (2007). Primary auditory cortical responses while attending to different streams. In Hearing: from Sensory processing to perception, B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, eds. (Springer-Verlag).

Yin, P., Mishkin, M., Sutter, M., & Fritz, J. B. (2008). Early stages of melody processing: stimulus-sequence and task-dependent neuronal activity in monkey auditory cortical fields A1 and R. *Journal of Neurophysiology*, *100*(6), 3009-29.

Zeitler, M., Fries, P., and Gielen, S. (2006). Assessing neuronal coherence with single-unit, multi-unit, and local field potentials. *Neural computation 18*, 2256-2281.

Zeki, S. (1993). Vision of the brain (Oxford: Wiley-Blackwell).

Zera, J., and Green, D.M. (1993a). Detecting temporal asynchrony with asynchronous standards. *J Acoust Soc Am 93*, 1571-1579.

Zera, J., and Green, D.M. (1993b). Detecting temporal onset and offset asynchrony in multicomponent complexes. *J Acoust Soc Am 93*, 1038-1052.

Zera, J., and Green, D.M. (1995). Effect of signal component phase on asynchrony discrimination. *J Acoust Soc Am 98*, 817-827.