Influence of spatial cues on the identification and the localization of objects in the auditory foreground

by

Adrian Kuo Ching Lee

B.E. Electrical Engineering (2003)
University of New South Wales, Sydney, Australia

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE IN HEALTH SCIENCES AND TECHNOLOGY

AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SEPTEMBER 2007

Signature of Author: ______________________________
Harvard-MIT Division of Health Sciences and Technology
July 26, 2007

Certified by: ______________________________
Barbara G Shinn-Cunningham
Associate Professor of Cognitive and Neural Systems, Boston University
Lecturer, Harvard-MIT Division of Health Sciences and Technology
Thesis Supervisor

Accepted by: ______________________________
Martha L Gray
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

Biographical note

Adrian KC Lee attended the University of New South Wales, Sydney, Australia (1998-2002) and earned a Bachelor of Engineering in Electrical Engineering with first class honors in May 2003. He received the UNSW Co-Op scholarships for his undergraduate studies and gained industrial trainings at Nokia, Alcatel and EnergyAustralia. In 1999, he also obtained the Licentiate Diploma in piano performance from Trinity College London, UK.

In September 2003, he began his doctoral studies at the Harvard-MIT Division of Health Sciences and Technology, Speech and Hearing Bioscience and Technology (SHBT) program. He completed his doctoral thesis work under the supervision of Dr. Barbara Shinn-Cunningham supported by a research assistantship in the Auditory Neuroscience Laboratory at the Department of Cognitive and Neural Systems at Boston University (2003-2007). He presented his research at seven international conferences and portions of this work earned him the Graduate Student Travel Award from the Association for Research in Otolaryngology (2007). He has also been invited to give both technical and non-technical talks on psychoacoustics in the US and in the UK. During his time as a doctorate student, he was a Visiting Scholar at the Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, UK, under the supervision of Drs. Robert Carlyon and Rhodri Cusack (summer 2005), and he also served as a teaching assistant at MIT for the graduate course HST.714J Acoustics of Speech and Hearing (Fall, 2005) and the undergraduate course 6.011 Introduction to Communication, Control and Signal Processing (2006 and Spring 2007).

He is a co-author on the following publication in auditory stream segregation (Appendix of this thesis):

Shinn-Cunningham, B.G., Lee, A.K.C. and Oxenham, A.J (**2007**). A sound element gets lost in perceptual competition. *Proc. Nat. Acad. Sci.* **104**:12223-12227.

AD MAIOREM DEI GLORIAM

Influence of spatial cues on the identification and the
localization of objects in the auditory foreground

by

Adrian Kuo Ching Lee

B.E. Electrical Engineering, University of New South Wales, 2003

Submitted to the Harvard-MIT Division of Health Sciences & Technology
on July 26, 2007 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Science in
Health Sciences and Technology

Abstract

The ability to form auditory objects is important in the natural environment where sounds arriving at our ears are a resultant of all spectro-temporal components that may have arisen from different auditory events. It has been shown that auditory spatial cues are effective for grouping acoustical energy across time or across frequency. However, little is known about the effect of spatial cues on scene analysis when more than one auditory object is being presented.

In this thesis dissertation, a novel two-object paradigm was used to investigate how spatial cues influence the identification and the localization of object in the auditory foreground. Specifically, the effect of both spatial and non-spatial cues on auditory grouping and object identification was ascertained. Using an acoustic pointer and the same stimuli for the object identification task, the apparent spatial location of these objects was measured to test the hypothesis that only the spatial attributes of the components grouped to form an object influences the localization of the same object. A conceptual model was generated to highlight the role of spatial cues in object formation, and the dissociation between the auditory computation of "what" and "where" was further investigated.

In current technology, object segregation presents a fundamental challenge for the hearing impaired, hearing aid design and speech recognition algorithms. It is hopeful that the findings in this dissertation will inspire new biologically-based algorithms for auditory scene analysis and in turn, influence designs in assistive hearing devices and other technological development that is dependent on multi-source segregation.

Thesis Supervisor: Barbara G Shinn-Cunningham

Title: Associate Professor of Cognitive and Neural Systems, Boston University
Lecturer, Harvard-MIT Division of Health Sciences and Technology

Acknowledgements

I have been blessed with so many wonderful people who have shaped me both professionally and personally in the last four years. I would especially like to thank the following people in guiding me to successfully completing my doctoral training:

Dr. Barbara Shinn-Cunningham, my supervisor, for giving me the freedom to grow as a scientist. I am also extremely grateful for her taking me up as a research assistant at the beginning without any prior experience.

Other faculty members at the Boston University Hearing Research Center: Drs. Steve Colburn, Gerald Kidd, Chris Mason and Nat Durlach for cultivating an environment that makes hearing research fun but yet scientifically critical.

Dr. Louis Braida, chair of my thesis committee, especially for his meticulous and helpful comments on my thesis dissertation.

Drs. Edward Adelson and Pawan Sinha, readers of my thesis committee, for their insightful questions that helped me make important links between visual and auditory scene analysis.

Dr. Andrew Oxenham for his generous help in the initial experimental designs.

Dr. Jennifer Melcher for encouraging me to seek inputs from MIT-BCS.

Dr. Alan Oppenheim for offering me the opportunity to serve as a teaching assistant for 3 semesters in MIT-EECS, ensuring that signal processing is now engrained in my bones.

Dr. John Rosowski for his generous time to listen to me, and especially in the initial few months of joining the program.

Drs. Robert Carlyon and Rhodri Cusack for my summer research project at MRC-CBU.

Tim Streeter, lab manager, for helping me through all the technical hurdles.

The ANL crew, especially Erick Gallun, Gin Best, Antje Ihlefeld, Erol Ozmeral, and Scott Bressler for all the lively discussions and beverages we shared.

Ann Dix, audiologist, for helping with the subject screening process.

Ade Dean-Pratt from UCL, for her help in the localization experimental designs.

Sigrid Nasser, our lab assistant, for helping me recruit over 40 subjects involving more than 100 hours of subject testing, even when I was researching in England; Steve Babcock, whose senior project made a guest appearance in the supplementary material in our recent publication.

The entire SHBT faculty and student body for making me feel belonged, especially my fellow classmates (SHBT 03: Anne Dreyer, Anton Peng, Daryush Mehta, Sasha Devore, S.R. Prakash, Yoko Saikachi and Sherry Zhao) for the camaraderie inside and outside the classroom.

The HST family, especially the student lounge crew and the admin staff.

All the friends who have kept me sane throughout the years: people I met at S&P, the 376 housemates, mates back home in Oz and especially J. Alberto Ortega for keeping me balanced physically, mentally and spiritually.

Last, but not least, the endless love and support of my parents.

Table of contents

CHAPTER 1 INTRODUCTION

1.1 Auditory scene analysis: the "cocktail party problem"

> *Imagine you are dining in a crowded restaurant with your parents. Amongst the bangs of cutlery and dishes, the sound of patrons toasting, clinking their glasses and the cry of a baby across the room for attention, you are able to effortlessly concentrate on your mother's voice. Meanwhile, your father switches off his hearing aid because everything sounds like noise to him.*

How does one focus a specific voice from a cacophony of conversation and other environmental sounds? Cherry [1] first coined this question as the "cocktail party problem." While we can seamlessly perform this task in our daily lives, one of the chief complaints for listeners with sensorineural hearing loss is the difficulty of communications in such complex acoustical environments [2].

In general, sound arriving at our ears is a sum of acoustical energy from all the auditory sources in the environment. However, to derive the actual waveforms from the sum is intrinsically an ill-posed problem.[1] Albert Bregman, in his seminal book [3], systematically described the principles underlying the perception of sounds in complex mixtures and posited what information our brain might use to derive the original waveforms. This branch of research is now commonly known as auditory scene analysis. By understanding the underlying mechanisms into how our brain achieves such a feat not only sheds new light into our fundamental understandings of human perception, it can also, in turn, inspire new algorithms that can be used in assistive hearing devices, such as hearing aids [4] and cochlear implants [5].

[1] Similarly in arithmetic, given that the resultant sum is 8, the problem of working out the original addends is ill-posed, since it could have been a result of 2+6, $\pi+(8-\pi)$, or any other of an infinite number of ways how the addends could have been combined.

1.2 Perceptual organization and early Gestalt influence

The process of separating the relevant information related to the stream or object of interest from all other information in the environment is not exclusive to audition. In general, information available at our sensory epithelia, e.g., cochlea and retina, is a chaotic juxtaposition of different elementary sensations [6-9], and the term scene analysis describes the cognitive process that groups elements together into discrete perceptual objects independent of modalities. Gestalt theorists [9, 10] were particularly interested in the problem and, by predominantly observing how scenes are parsed in the visual domain, they proposed rules to qualitatively address the principles of perceptual organization. These rules were known as the law of Prägnanz [8]. For example, the law of proximity and the law of similarity, respectively, govern how the center square in Figure 1-1 is parsed to be grouped horizontally or vertically with the scene in the environment. The Gestalten school, emphasizing the importance of *Gestaltqualität* (German for *organized whole*) whereby the whole is more than the sum of its parts, had a big influence in the visual scene analysis community in the first half of last century, and only later championed by Marr's school of computational scene analysis [11].

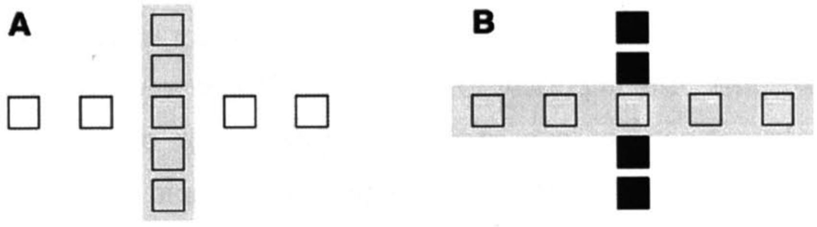


Figure 1-1 Examples of classical Gestalt perceptual organization principles. **(A)** The center square is perceptually grouped vertically by proximity. **(B)** The center square is perceptually grouped horizontally by similarity.

Even though Wertheimer originally described how the law of Prägnanz can be applied to both visual and acoustic stimuli, only after the aforementioned influential publication by Bregman that the Gestalten school has a presence in the psychoacoustics literature. Borrowing from the well-documented examples in vision, Bregman translated them back to the auditory domain, and posited two different dimensions for grouping: across time (or sequential) and across frequency (or simultaneous). Common onset from different harmonic components, for example, is an important cue to promote simultaneous grouping while co-modulation of stimuli across time is an important cue for sequential grouping.

1.3 Role of spatial cues in stream segregation

In a natural auditory environment, acoustic sources are generally displaced in space. Our ability to utilize auditory spatial cues for perceptual organization poses a paradox [12]. Daily experiences suggest that we have a robust spatial percept of our auditory surrounding [13], and spatial cues have also been empirically shown to be a very important cue for grouping spectro-temporal elements across time [7]. However, spatial cues alone cannot be used to segregate speech of a single talker from similar simultaneous sounds [14].

Spatial cues not only help us parse the auditory scene to better understand what information is conveyed from the source of interest in a sound mixture [15-17], they also help us to locate these sources in space. However, while the visual system is capable of distinguishing spatial changes of angle less than one minute of arc, the spatial resolution limit of the auditory system is about two orders of magnitude higher. It is important to note that unlike in a single retina, a single cochlea does not have an explicit spatial representation of sound sources and this must be calculated using spectro-temporal elements available at the two cochleae further down in the central nervous system [18]. Therefore, while there are many common principles that govern perceptual organization both in vision and in audition, the role of spatial cues in auditory scene analysis is particularly interesting, since not only is it different from other modalities like vision, it also influences grouping differently within audition depending on whether the elements are grouped across time or across frequency.

1.4 Aims of thesis dissertation

The goals of this thesis dissertation were twofold. The first goal was to investigate how spatial and non-spatial cues influence the identification of the object in the auditory foreground. When a spectro-temporal element can logically belong to more than one object in the auditory scene, the energy of this spectro-temporal element perceived in each object is expected to sum to the total energy of that element in the auditory scene if veridical parsing occurred. In Chapter 2, Chapter 3 and the Appendix [56], this hypothesis was explicitly tested.

The second goal of the dissertation was to ascertain how spatial cues influence the localization of the attended object. Using the same two-object stimuli similar to that used in the identification experiment, the hypothesis that only spatial attributes of the components grouped to form an object would influence the localization of the same object was tested, and the results were presented in Chapter 4. Interactions in localization between objects in the auditory scene were also characterized in this chapter.

In Chapter 5, the exact stimuli employed in the Appendix [56] were used for a localization experiment. The results for the two studies (investigating the identity and localization of objects influenced by spatial cues) were compared and the dissociation between the auditory computation of "what" and "where" was investigated. Finally, Chapter 6 summarized all the results in the aforementioned experiments and offered a preliminary conceptual framework to account how spatial cues affect auditory scene analysis.

1.4.1 Stylistic notes and chapter organizations

Chapter 2, Chapter 3 and Chapter 4 were written for technical journal submissions, while Chapter 5 was written similar in style to Appendix, having a wider scientific community in mind.

CHAPTER 2 TRADING HYPOTHESIS OF AMBIGUOUS TARGET

Portion of the work described in this chapter has been submitted to the *Journal of the Acoustical Society of America*.

Lee, A.K.C. and B. Shinn-Cunningham, *Effects of frequency disparities on the trading of an ambiguous tone between two competing auditory objects*, submitted.

2.1 Abstract

Listeners are relatively good at estimating the true content of each physical source in a sound mixture in most everyday situations. However, if there is a spectro-temporal element that logically could belong to more than one object, the correct grouping of sound can be ambiguous. To be veridical, estimates of the amount of energy that an ambiguous element contributes to each perceptual object must sum to equal the physical energy in that element in the physical stimulus. This kind of energy conservation is implicitly assumed in many psychoacoustic experiments, even though few have directly tested this relationship. The current experiments test whether energy conservation holds in a two-object sound mixture containing an ambiguous element. The relative strength of simultaneous and sequential grouping cues was varied to alter the allocation of the ambiguous tone to the two competing objects. The contribution of the tone to the two objects varies systematically as the relative strength of the grouping cues changes; however, the total perceived tone energy present in the two objects sums to less than the physical tone energy in the mixture. Thus, although there is an energy trading relationship, energy conservation fails for these ambiguous stimuli.

2.2 Introduction

Sound arriving at our ears is a sum of acoustical energy from all the auditory sources in the environment. In order to make sense of what we hear, we must group related components from a source of interest and perceptually separate these elements from the elements originating from other sources. Auditory scene analysis depends upon segregating sound-energy from each source from the mixture of sound (concurrent segregation), and grouping together the energy from each source across time (streaming or sequential grouping).

The process of sequential grouping has been studied by many investigators. The perceptual organization of a sequence of tones depends on the frequency proximity of the component tones [19], the tone repetition rate, and the attentional state of the observer. For example, in an isochronous (evenly spaced) sequence of tones whose pitch alternates between two values, a single auditory stream is generally perceived when the frequency difference is small. The likelihood of perceiving two separate streams (corresponding to the two pitches) increases as the frequency separation and / or the tone repetition rate increases. While there have been attempts to explain sequential streaming based on peripheral processing of the auditory system [20-22], such explanations cannot fully account for how even simple tone sequences are perceptually organized [23].

For some stimuli, perceptual organization is relatively unambiguous. However, for some mixtures, the perceived organization can be affected by how a listener is instructed to hear the mixture. For such ambiguous stimuli, the perceptual organization tends to change over time [24, 25]. Usually, such stimuli are first heard as one stream, but as time evolves, listeners become more likely to hear two separate streams. Moreover, attention switches can reset this buildup of streaming [26, 27]. The role of attention in the formation of auditory streams is under debate [28]. Possible neural correlates in the buildup of streaming have been found by single-unit recordings in animals [29-32]. Recent imaging studies in humans also show a neural correlate of this kind of buildup [33-35].

If two spectrally non-overlapping sounds are simultaneously presented from two loudspeakers at different locations, one might expect the difference in the spatial cues in the sounds to cause the two sources to be segregated into independent objects. Surprisingly, there is little evidence to support the use of auditory spatial cues to segregate simultaneous sounds, at least when other grouping cues support hearing the sound as coming from a single source. In particular, nearly any spectro-temporal features that push sources to be integrated into one object override differences in spatial cues [14, 36, 37], including harmonicity [38, 39] and onset / offset synchrony [40, 41].

The relative potency of acoustical cues influencing sequential and simultaneous grouping has been measured by pitting cues against one another to determine which cue dominates perceptually. However, few studies have investigated the interplay between sequential and simultaneous grouping cues. Bregman [42] showed that frequency separation strongly influences sequential grouping while onset / offset synchrony strongly affects simultaneous grouping. Moreover, there is an interaction between sequential and simultaneous grouping [43-45]. For instance, a repeating tone sequence that matches the frequency of one harmonic in a harmonic tone complex (henceforth, the target tone) can cause that harmonic to contribute less to the perceived content of the complex [36, 44, 46, 47]. While there are many studies exploring how a sequential stream influences a simultaneously grouped harmonic complex, we know of no reciprocal studies exploring how the complex influences perception of the across-time stream. Intuitively, one might expect the contribution of the target tone to the harmonic complex to be inversely related to its contribution to the tone stream. Specifically, one might expect the contribution of the ambiguous target to one stream to increase when its contribution to the other stream decreases, consistent with the way energy trades physically. Acoustically, if two sources (S_1 and S_2) are uncorrelated, then at each frequency, the total energy at the frequency $\overline{I}_S(f,\tau)$ is the sum of the energies in S_1 and S_2 at that frequency, $\overline{I}_{S_1}(f,\tau)$ and $\overline{I}_{S_2}(f,\tau)$ respectively:

$$\overline{I}_S(f,\tau) = \overline{I}_{S_1}(f,\tau) + \overline{I}_{S_2}(f,\tau)\,. \qquad (2.1)$$

(Section 2.7 provides a formal mathematical proof of this trading relationship).

Using a rhythmic masking release paradigm, recent work explored how frequency proximity (which strongly influences sequential grouping) interacts with harmonicity and common onset / offset (strong cues for simultaneous grouping) to influence the perceived content of a harmonic complex [48, 49]. However, in interpreting these results, it was explicitly assumed that a trading relationship exists. In particular, when a tone did not contribute strongly to the simultaneously grouped object, it was assumed to strongly contribute to the ongoing stream. In the current study, we constructed a stimulus containing a fixed ambiguous spectro-temporal element (the target) that could logically be heard as one tone in an isochronous stream of repeating tones and as part of a more slowly repeating harmonic complex. The frequency of the repeating tones was varied systematically from below to above the frequency of the target, which was at the frequency of the fourth harmonic of the complex. A control experiment allowed us to compute the effective level of the target perceived in the two objects (tone stream and complex) to see if there is a systematic trading relationship between the perceived energies of the target in each of the two competing objects. We also performed another experiment that assessed the degree to which the perceptual organization of the tone stream was influenced by the presence of the repeating harmonic complex.

2.3 Experiment 1: Competing streams

Stimuli consisted of a repeating tone stream and a harmonic complex that repeated at one-third the rate. Unlike most streaming experiments, which compare when a stimulus is heard as one versus two streams, our two-object stimuli were always perceived as containing two objects. The target tone could logically belong to either stream. We manipulated the frequency content of the repeating tones to explore how this influenced the grouping of the target. We investigated whether a frequency difference between the repeating tones and target altered the degree to which the target was heard as a part of this sequence. We hypothesized that as the frequency difference increased, the contribution of the target to the tones sequence would decrease. We further hypothesized that as the target contributed less to the tones stream, it would contribute more to the harmonic complex (i.e., we hypothesized that there would be a trading relationship between the contribution of the target to the two streams that competed for ownership of the target).

2.3.1 Methods

2.3.1.1 Stimuli

Stimuli generally consisted of a repeating sequence of a pair of tones followed by a harmonic complex (Figure 2-1A). The frequency of the pair of tones varied from trial to trial from two semitones below 500 Hz to two semitones above 500 Hz (i.e., 445.4 Hz, 471.9 Hz, 485.8 Hz, 500 Hz, 514.7 Hz, 529.7 Hz and 561.2 Hz). Each tone was 60 ms in duration, gated with a Blackman window of the same length.

The harmonic complex contained the first 39 harmonics of the fundamental frequency (F0) of 125 Hz, excluding the fourth harmonic (500 Hz). The phase of each component was chosen randomly on each trial. The magnitudes of the harmonics were shaped to simulate the filtering of the vocal tract [50]. The first formant frequency (F1) was set to 490 Hz, close to the expected value for the American-English vowel /ɛ/ [51]. The second and the third formants were fixed at 2100 and 2900 Hz, respectively. The half-power bandwidths of the three formants were 90, 110 and 170 Hz (parameters were chosen based on studies by [52]). Each harmonic complex was 60 ms in duration, gated with a Blackman window of the same length.

The target was a 500 Hz tone that was gated on and off with the harmonic complex (60 ms in duration, gated with a 60-ms-Blackman window). As a result of this structure, the target could logically be grouped as the third tone in the repeating tone stream or as the fourth harmonic in the harmonic complex.

The amplitude of the target and the tones was equal, and matched the formant envelope of the vowel. There was a 40 ms silent gap between each tone and harmonic complex to create a regular rhythmic pattern with an event occurring every 100 ms. This basic pattern, a pair of repeating tones followed by the vowel complex / target, was repeated ten times per trial to produce a three-second stimulus that was perceived as two streams: an ongoing stream of tones and a repeating vowel occurring at a rate one-third as rapid.

The rhythm of the tone sequence depends on the degree to which the target is perceived in the tone stream (Figure 2-1B, top panel). Specifically, the tone stream is heard as "even" when the target is heard in the stream and "galloping" when the target is not perceived in the stream. Similarly, the perceived content of the harmonic complex depends on whether or not the target is heard as part of the complex (Figure 2-1B, bottom panel). Specifically, when the target is perceived in the complex, the first formant peak (F1) of the complex is perceived as slightly higher when the target tone is heard as part of the complex compared to when the target is not part of the complex. This slight shift of F1 causes the complex to be heard more like /ɛ/ when the target is part of the complex, and more like /ɪ/ when it is not part of the complex.

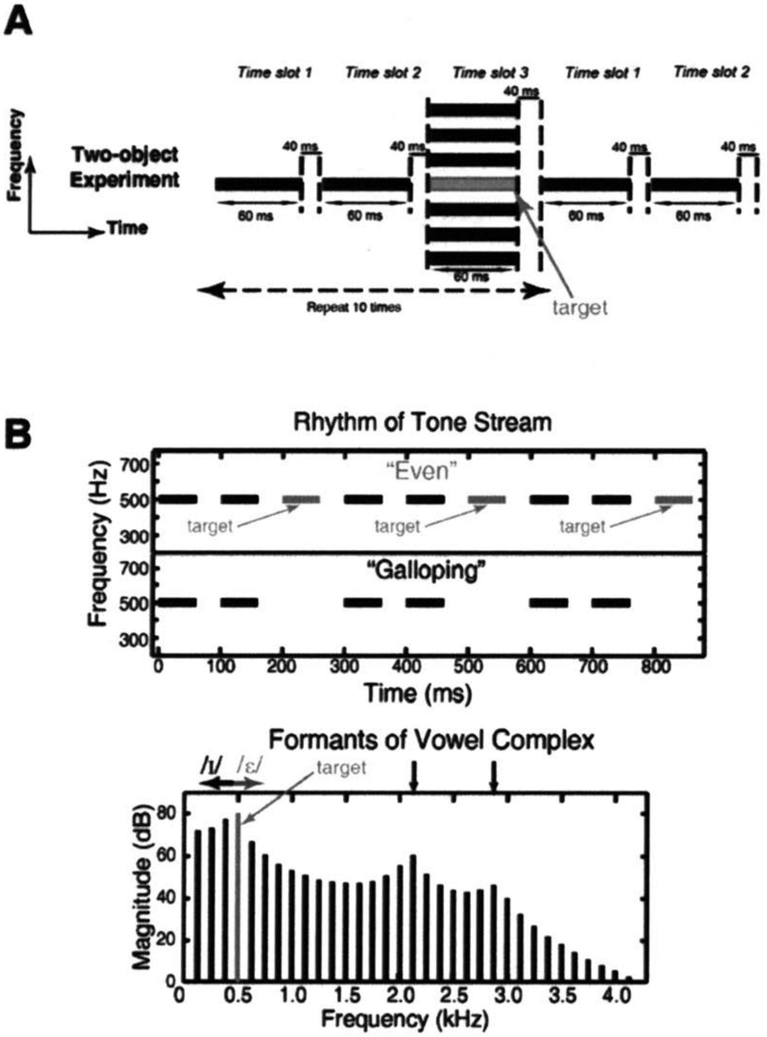


Figure 2-1(A): Two-object stimuli were created by repeating a three-item sequence that consisted of a pair of pure tones followed by a harmonic complex. In the reference configuration, the pure tones in time slots 1 and 2 are at 500 Hz. Time slot 3 is made up of two components: a target tone at 500 Hz and a tone complex with an F0 of 125 Hz (with the fourth harmonic 500 Hz omitted). The tone complex is shaped by a synthetic vowel spectral envelope to make it sound like a short vowel [53]. Because the first formant of the vowel complex is near 500 Hz, the relative level of the target tone perceived in the vowel complex affects perception of the first formant frequency, which affects the perceived identity of the vowel. **(B)** (Top Panel): The perceived rhythm depends on whether or not the 500 Hz target tone is perceived in the sequential tone stream: if the target is grouped with the repeated tones, the resulting rhythmic percept is "even;" if the target is not grouped with the pair of tones, the resulting perceived rhythm is "galloping." (Bottom Panel): The synthetic vowel spectral envelope is similar to that used by Hukin and Darwin [54]. The identity of the perceived vowel depends on whether or not the 500-Hz target is perceived in the complex: the vowel shifts to be more like /ε/ when the target is perceived as part of the complex and more like /ɪ/ when the target is not perceived in the complex. The arrows indicate the approximate locations of the first three formants of the perceived vowel.

Control stimuli consisted of single-object presentations with only the tones or only the harmonic complex, either with the target (“target-present” prototype) or without the target (“target-absent” prototype). Finally a two-object control was generated in which the repeating tones and complex were presented together, but there was no target (“no-target” control).

2.3.2 Task

In order to assess perceptual organization of the two-object mixture and how it affected the perceived content of both stream of tones and vowel, the same physical stimuli were presented in two separate experimental blocks. In one block subjects judged the rhythm of the tone sequence (“galloping” or “even”), in a one-interval, two-alternative-forced-choice design. In the other block, the same physical stimuli were presented in a different random order, and subjects judged the vowel identity (“/ɪ/ as in ‘bit’” or ”/ɛ/ as in ‘bet’”).

2.3.3 Environment

All stimuli were generated offline using MATLAB software (Mathworks Inc.). Signals were processed with pseudo-anechoic head-related transfer functions (HRTFs; see [55] for details) in order to make the stimuli similar to those used in a companion study (Appendix [56]) that varied source location. In the current experiment, all components of the stimuli were processed by the same HRTFs, corresponding to a position straight ahead of and at a distance of one meter from the listener.

Digital stimuli were generated at a sampling rate of 25 kHz and sent to Tucker-Davis Technologies hardware for D/A conversion and attenuation before presentation over headphones. Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. A randomized roving attenuation level between 0 and 14 dB was applied to each stimulus before presentation in order to reduce the reliability of absolute presentation level as a cue in the identification task. Subjects were seated in a sound-treated booth and responded via a graphical user interface. Stimuli were presented over insertion headphones (Etymotic ER-1). All signals were presented at a listener controlled, comfortable level that had a maximum value of 80 dB SPL.

2.3.4 Experimental procedures

2.3.4.1 Participants

Fourteen subjects (eight male, six female, aged 18-31) took part in this experiment. All participants had pure-tone thresholds of 20 dB HL or better at all frequencies in the range from 250-8000 Hz, in both ears, and their threshold at 500 Hz was 15 dB HL or better. All subjects gave informed consent to participate in the study, as overseen by the Boston University Charles River Campus Institutional Review Board and the Committee On the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology.

2.3.4.2 Training with single-object prototypes

In each session of testing, each subject was familiarized with the single-object prototypes with and without the target. In the rhythmic block of the experiment, subjects were trained to label a stream of 500-Hz tones with the target present as "even," and to label the tones without the target present as "galloping." In the corresponding vowel training runs, subjects were trained to label the harmonic complex with the target present as /ɛ/ (as in 'bet') and the harmonic complex without the target as /ɪ/ (as in 'bit').

During training, subjects were given feedback to verify that they learned to correctly label the single-object, target-present and target-absent prototypes. This feedback ensured that subjects could accurately label the tone-stream rhythm and the harmonic complex identity for unambiguous, single-object stimuli. Subjects had to achieve at least 90% correct when identifying the two prototypes in the single-object pre-test before proceeding to the two-object experiment.

2.3.4.3 Two-object experiments

Following training, listeners judged the tone-stream rhythm and the vowel identity for stimuli that had both objects present (see left side of Figure 2-2). Intermingled with these trials were single-object control trials (which allowed us to assess whether listeners maintained the ability to label the unambiguous stimuli without feedback throughout the run; see right side of Figure 2-2). From trial to trial, the frequency of the repeating tones in the two-object stimuli varied randomly relative to the target (Figure 2-2). Seven two-object conditions were tested, with the frequency of the repeated tones ranging from two semitones below to two semitones above the target frequency (Δf = 0, ±0.5, ±1, ±2 semitones). A control two-object condition was included in which the target was not presented. In this control, the frequency of the repeated tones were randomly selected from the seven possible frequencies used in the other two-object conditions (i.e., 0, ±0.5, ±1, ±2 semitones from 500 Hz; Figure 2-2) to ensure that subjects did not make rhythmic or vowel judgments based on the absolute frequency of the repeated tones.

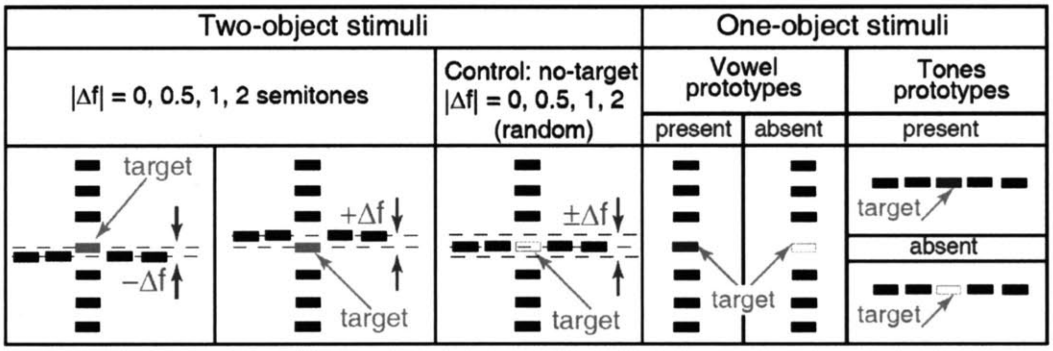


Figure 2-2: Experimental conditions. Each block consists of seven two-object conditions with the target present, a two-object control condition without the target present, and two single-object prototype conditions (see text for more details).

In one block of the experiment, we presented eight two-object stimuli and single-object prototypes containing no vowel. In this block, we asked the subjects to report the perceived rhythm of the tones. In a separate block of the experiment, we presented the same eight two-object stimuli intermingled with the two single-object vowel prototypes and asked the subjects to report the perceived vowel. Both blocks consisted of 30 repetitions of each stimulus in random order, for a total of 300 trials per block. We used the response to the prototype stimuli both for screening and in interpreting the results to the ambiguous two-object stimuli, as discussed below.

2.3.4.4 Single-object control experiments

Two companion single-object control experiments tested the subjective impressions of the tone-stream rhythm and the vowel identity when there were no other objects present and the physical intensity of the target varied from trial to trial. In these control experiments, subjects were presented with single-object stimuli (tones in one experiment, harmonic complex in the other) with a variable-level target. From trial to trial, the intensity of the target was attenuated by a randomly chosen amount ranging between 0 dB and 14 dB (in 2 dB steps) relative to the level of the target in the two-object experiments. In the single-object tone task, subjects reported whether the rhythm on a given trial was "galloping" or "even." In the single-object harmonic complex task, subjects reported whether the complex was /ɪ/ or /ɛ/.

For both experiments, the percent responses (y) for each subject was related to the target attenuation (x) by fitting to a sigmoidal function of the form:

$$\hat{y} = \frac{1}{1+e^{-a(x-x_0)}}, \qquad (2.2)$$

where $\hat{y}$ is the estimated percent response, a is the best-fit slope parameter and x_0 is the best-fit threshold constant (50% of maximum). If 95% or more of a subject's responses for a given condition were target-present (i.e., "even" or "/ɛ/ as in 'bet'") or target-absent (i.e., "galloping" or "/ɪ/ as in 'bit'"), the effective attenuation was set to 0 dB or 16 dB, respectively. The corresponding psychometric functions for each subject allowed us to map the percent response in the two-object experiment onto an effective target attenuation, based on the mapping between physical target attenuation and response percentages in the single-object control experiment.

2.3.4.5 Data analysis

Raw percent correct "target-present" responses ("even" for the tones, /ε/ for the vowel) were computed for each subject and condition. These results were then averaged across subjects to see overall trends. Individual raw percentages were also used to quantify the influence of the stimulus conditions on the perceived content of the tone and vowel streams. The percentage of "target-present" responses for each stimulus condition was used to estimate the perceptual distance between the stimulus and the single-object target-absent prototypes (see Appendix A.4), as summarized below.

In each block of the two-object experiment, single-object objective control conditions (the prototypes with and without the target) were randomly intermingled with the ambiguous, two-object stimuli. We assumed that in judging the content of an object, listeners used an internal Gaussian-distributed decision variable whose mean depended on the stimulus and whose variance was independent of the stimulus. This internal decision variable was assumed to represent the perceptual continuum from target-absent to target-present. Listener responses on a given trial (either target-absent or target-present) were assumed to be the result of a comparison of a sample of this internal decision variable to a criterion that was constant throughout the block, enabling us to compute the relative perceptual separation of the means of the conditional distributions for the different stimulus conditions. In particular, conditioned on which stimulus was presented, the percent target-present responses were assumed to equal the portion of the conditional distribution of the decision variable falling to the appropriate side of an internal decision criterion. Differences in these conditional probabilities then can be used to compute the perceptual distances (d') between the distributions (see Figure 2-3)

We use $d'_{present:absent}$ to denote the perceptual distance between the target-present and target-absent prototypes. Assuming the above decision model, $d'_{present:absent}$ is given as [57, 58]:

$$d'_{present:absent} = \Phi^{-1}[\Pr(\text{"target present"} \mid \text{target present})] - \Phi^{-1}[\Pr(\text{"target present"} \mid \text{target absent})], \qquad (2.3)$$

where Φ^{-1} denotes the inverse of the cumulative Gaussian distribution, and the double quotation marks denote the response of the subject. In order to avoid an incalculably large value of d' due to sampling issues, 0.5 was added to all data cells prior to percentage conversion for all d' calculations. As a result of this adjustment, a d' of 4.28 corresponds to perfect performance when discriminating between the two prototypes in these experiments. For each independent subject, $d'_{present:absent}$ was calculated from response percentages.

The perceptual distance between any stimulus and the target-absent single-object controls was then calculated individually for each subject as:

$$d'_{condition:absent} = \Phi^{-1}[\Pr(\text{"target present"}|\text{condition})] - \Phi^{-1}[\Pr(\text{"target present"}|\text{target absent})]. \qquad (2.4)$$

In order to directly determine whether a particular stimulus was perceived as more similar to the target-present prototype or more like the target-absent prototype, for each subject, we computed a normalized $d'_{present:absent}$ for each condition from the raw sensitivities in equation (2.4) as:

$$\delta'_{condition:absent} = \frac{d'_{condition:absent}}{d'_{present:absent}}. \qquad (2.5)$$

A value of $\delta'_{condition:absent} < 0.5$ indicates that the stimulus was perceived as more like the prototype with the target not present while $\delta'_{condition:absent} > 0.5$ indicates the condition was perceived more like a target-present than a target-absent prototype. A value near zero indicates the condition was perceived like that of target-not-present prototype. A score close to 1 occurs when the response probabilities for that condition are nearly identical to those for target-present prototype (Figure 2-3).

To quantify the effective level that the target contributed to each object, we analyzed the psychometric functions fit to the responses from the single-object control experiment (Section 2.3.5.4), which relate the percentage of "target-present" responses to the physical intensity of the target actually present in the stimuli. Using the psychometric functions obtained for each individual subject, we mapped the percent response in the two-object experiment to the target intensity that would have produced that percentage of responses for the single-object stimuli.

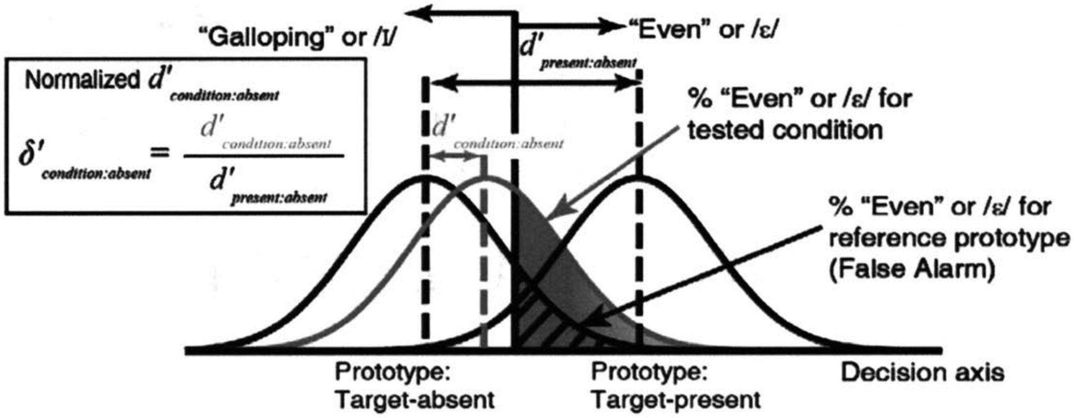

"Galloping" or /ɪ/
"Even" or /ɛ/
$d'_{present:absent}$
Normalized $d'_{condition:absent}$
$\delta'_{condition:absent} = \frac{d'_{condition:absent}}{d'_{present:absent}}$
$d'_{condition:absent}$
% "Even" or /ɛ/ for tested condition
% "Even" or /ɛ/ for reference prototype (False Alarm)
Prototype: Target-absent
Prototype: Target-present
Decision axis

Figure 2-3: Schematics of the decision model assumed in computing $\delta'_{condition:absent}$. The decision axis for either the rhythmic or vowel identification space is shown along the abscissa. The distributions show the probabilities of observing different values of the decision variable for the target-absent and target-present prototypes (left and right distributions) as well as for a particular stimulus (middle distribution).

2.3.5 Results

2.3.5.1 *Subject screening*

Despite training, not all subjects could reliably label the single-object vowel prototype stimuli when they were presented in the two-object experiment (i.e., some subjects could not maintain a consistent decision criterion for labeling the vowels). We adopted a screening protocol to exclude any subjects who could not accurately label the prototype stimuli during the two-object experiment.

Specifically, we excluded data from subjects who failed to achieve $d'_{present:absent}$ > 1.0.

We also excluded any subject for whom there was a weak dependence on their responses percentages as a function of target attenuation in the single-object control experiment. Specifically, if the fit slope parameter a in equation (2.2) was less than 10 percent / dB, the subject was excluded from further analysis. We also excluded any subject for whom the correlation coefficient (ρ) between the observed data (y) and the data fit ($\hat{y}$) was less than 0.9.

The top panel in Figure 2-4 shows example psychometric functions for the tone experiment for subjects S18 and S34, who are representative of the kinds of results we saw across the set of subjects. Participants responded “galloping” in conditions where the target tone was highly attenuated (i.e., at or above 12 dB of attenuation) and “even” when the intensity of the repeating tones matched that of the target (i.e., 0 dB attenuation). In other words, all subjects perceived changes in the rhythm of the tone stream with attenuation of the target. Similarly, all subjects could maintain a consistent decision criterion for labeling the rhythm of the tones, even without feedback, when the prototypes were presented alongside ambiguous two-object stimuli. No subjects were excluded from the experiment based on poor performance in the tones task.

The bottom panel in Figure 2-4 shows the psychometric functions for the same two subjects for the vowel experiment. S18 shows a steeply increasing psychometric function relating percent correct responses to the target attenuation. In contrast, S34 has a very shallow function, demonstrating poor sensitivity to changes in the target attenuation as measured by vowel identification. Consistent with this, S34 also had a low $d'_{condition:absent}$ in the vowel task (1.16 compared to 3.80 for S18). Despite the relatively poor sensitivity of S34, this subject did meet the liberal criteria required for our study.

Six out of the fourteen subjects (three male, three female) failed these criteria and had their results excluded from further analysis. All data analyzed below is from the eight subjects who passed all criteria for both tone and vowel screenings

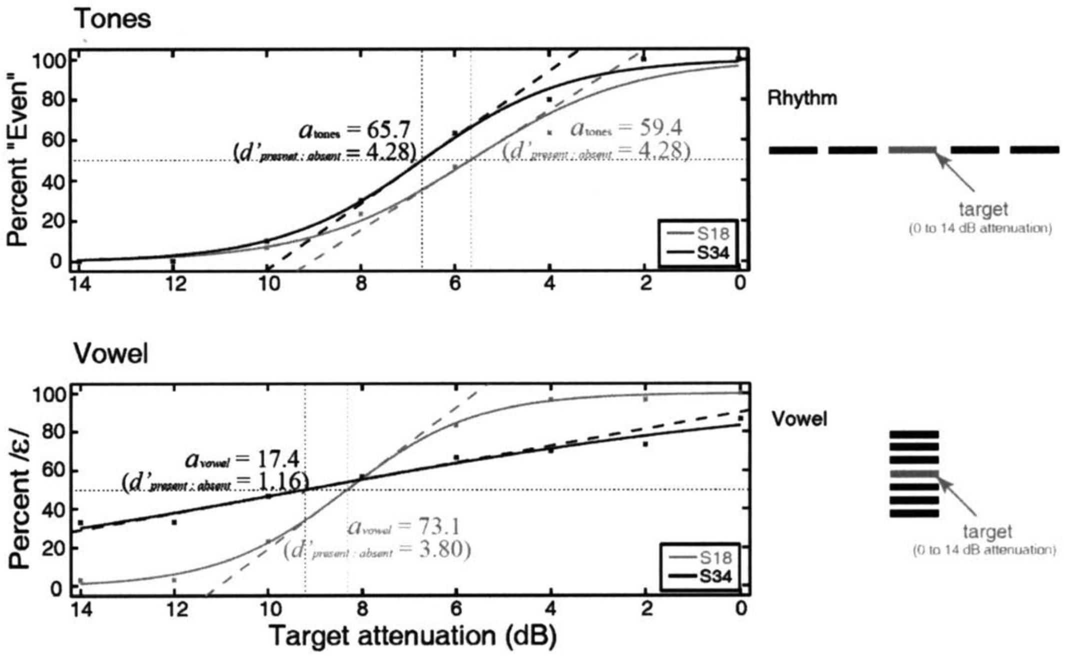


Figure 2-4: Example psychometric functions for results of two single-object experiments in which the target attenuation varied from 0 to 14 dB (in 2 dB steps), for two subjects (S18 and S34). Dotted lines show the slope of each of the psychometric function at the 50% point. The gradient at the 50% maximum point for the sigmoidal function is a quarter of the a value. The raw percent responses ("even," top; /ɛ/, bottom) are shown for each subject as a function of target attenuation. The $d'_{present:absent}$ results for each subject illustrate the close relationship between the slope of the psychometric function and the perceptual distance between the prototypes measured in the two-object experiment.

2.3.5.2 Rhythmic judgments

Figure 2-5 summarizes results of the main two-object experiment for the rhythm judgments (left column of Figure 2-5; results for the vowel, in the right column, are considered in the next section. Figure 2-5C and 5F reanalyze these results using the results of the single-object experiment and are considered in Section 2.3.5.4).

Figure 2-5A shows the across-subject mean and the standard error of raw percentage "even" response to the tone stream. Nearly all subjects correctly identified the "even" and "galloping" single-object prototypes with 100% accuracy. When the frequency of the repeating tones matched that of the target in the two-object condition (i.e., $\Delta f = 0$), subjects generally reported an "even" percept (average "target-present" response rate was greater than 90%). As the frequency difference between the repeating tones and target increased, the probability of responding as if the target was present in the tone stream decreased. As expected, there was a very low probability of reporting the target-present in the two-object condition with no target present (i.e., the average percentage of target-present responses was 1.7%). Replotting the results as $d'_{condition:absent}$ shows patterns that are virtually identical to the raw responses (not shown). In the tone-rhythm experiment, these values range from a low of 0.136 (for the target-absent two-object stimuli) to a high of more than 3.5 (for the $|\Delta f| = 0$ stimulus).

Figure 2-5B shows $\delta'_{condition:absent}$, which quantifies the perceptual distances of each stimulus relative to the single-object prototypes (0, near the galloping-prototype; 1 , near the even-prototype). As all subjects are nearly equal in their ability to properly label the two prototypes, mean δ' results look very similar to raw percentage responses. A repeated measures analysis of variance (ANOVA) on the $\delta'_{condition:absent}$ scores found a significant main effect of the absolute magnitude of the frequency difference [$F_{GG}(1.06,7.42) = 48.8$, $p_{GG} < 1.47 \times 10^{-4}$, using the Greenhouse-Geisser corrected degrees of freedom to account for violations of the sphericity assumption under the Mauchly's test. Henceforth, $F_{G\text{-}G}$ and $p_{G\text{-}G}$ denote statistics corrected using the Greenhouse-Geisser correction; where the subscript is left off, it signifies a condition for which the sphericity assumption was met]. This suggests that the frequency separation between the tones and the target has a strong influence on the rhythmic perception in this two-object paradigm. There was also a significant interaction between $|\Delta f|$ and the sign of the frequency difference [$|\Delta f|$ X sign(Δf), $F(2,14)=8.85$, $p< 3.28 \times 10^{-3}$], suggesting that the effect of having the repeating tones higher than the target is not equivalent to the effect of having the repeating tones lower by an equal amount. However, paired-sample *t*-tests (two-tailed with Dunn-Sidak post-hoc adjustments for three planned comparisons) revealed no significant differences in the means of all the condition pairs that had the same magnitude frequency difference but differed in sign ($\Delta f = \pm 0.5$: $t_7 = -2.614$, $p_{DS} = 0.101$; $\Delta f = \pm 1$: $t_7 = 2.317$, $p_{DS} = 0.152$; and $\Delta f = \pm 2$ semitones: $t_7 = 1.688$, $p_{DS} = 0.353$).

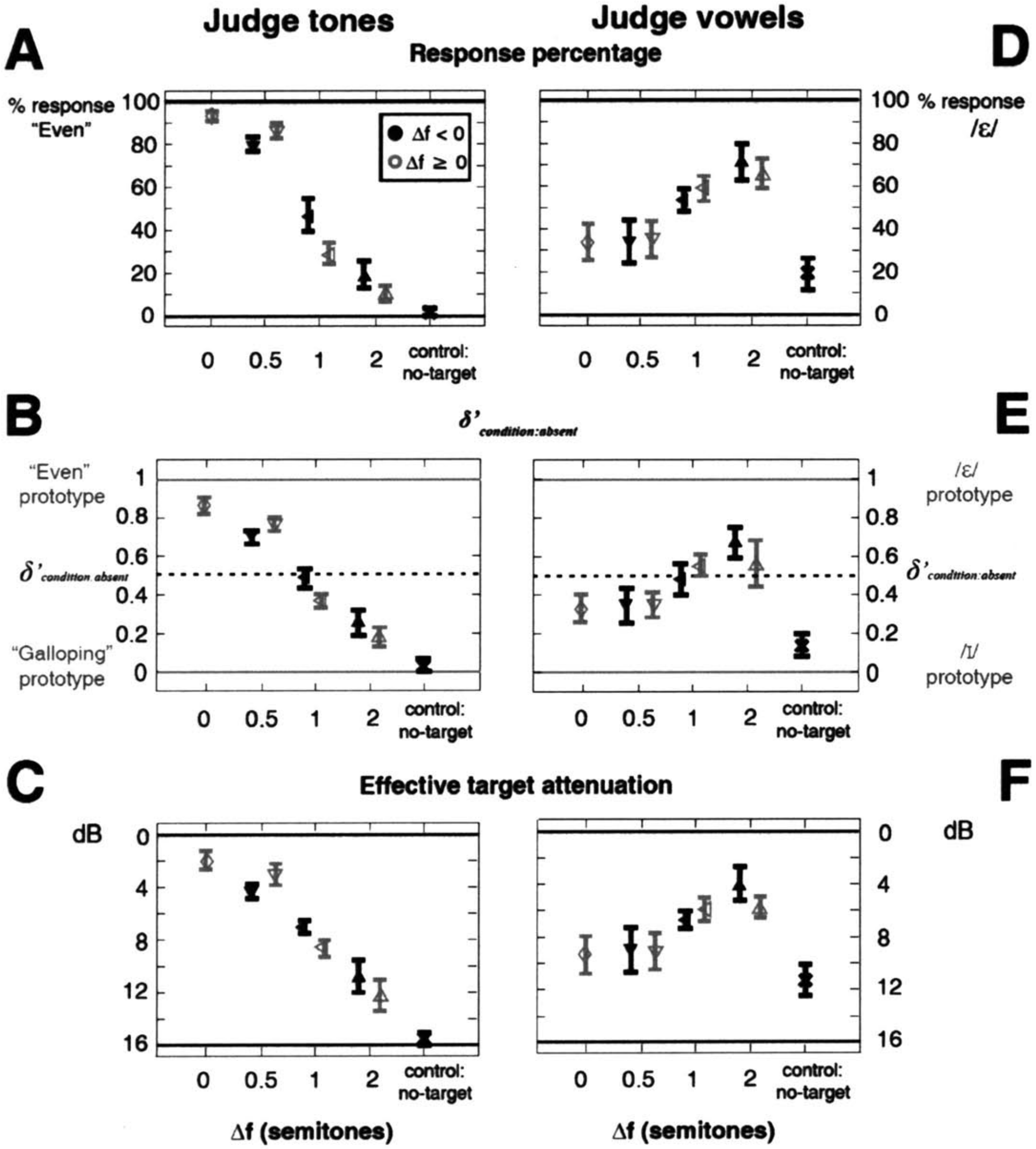


Figure 2-5: Results of both rhythm judgments (left column) and vowel judgments (right column). **(A)** & **(D)**: Raw response percentages. **(B)** & **(E)**: $\delta'_{condition:absent}$ derived from raw results. **(C)** & **(F)**: Effective target attenuation derived from the psychometric functions relating raw responses to effective target attenuations Each marker represents the across-subject mean estimate and the error bar shows ±1 standard error of the mean.

2.3.5.3 *Vowel judgments*

Figure 2-5D shows the across-subject mean and the standard error of the raw response percentages for the vowel judgments. Unlike in the rhythmic judgment block of the experiment, there was a non-zero likelihood of subjects responding /ɛ/ when an /ɪ/ prototype was presented; similarly, subjects sometimes responded /ɪ/ when an /ɛ/ prototype was presented. When the frequency difference between the tones and the target frequencies was zero, subjects were more likely to respond /ɪ/ (as if the target was not part of the vowel) than /ɛ/ (as if the target was part of the vowel). As the magnitude of the frequency difference between the tones and the target increases, the probability of reporting an /ɛ/ increases. As expected for the control "no-target" two-object stimulus, subjects frequently responded /ɪ/, as if the target was not present.

Because there were large subject differences in how consistently prototypes were labeled, transforming the data into d' scores increases the across-subject variability (not shown). Average d' values were lower overall than in the tone-rhythm experiment. The average d' values are roughly half the magnitude that they were for the tone-rhythm, with values ranging from a mean of 0.51 (for the target-absent two-object stimulus) to a high of roughly 2 (for the $|\Delta f| = 2$ semitones stimulus). In general, listeners have more difficulty in identifying the vowel than in labeling the tone rhythm. Transforming the results to δ' (the normalized d') reduces the across-subject variability by normalizing d' by the differences in overall performances (Figure 2-5E). In general, as the frequency difference between the repeated tones and the target increases, the likelihood that responses are like those to the /ɛ/ prototype increases (i.e., the target contributes more to the vowel percept). A repeated measures ANOVA on the $\delta'_{condition:absent}$ results reveals a significant main effect of the magnitude of the frequency difference [$|\Delta f|$, $F_{GG}(1.06,7.39) = 6.37$, $p_{GG} < 0.0368$]. However, unlike in the rhythmic-identification task, there were no significant interaction between $|\Delta f|$ and the sign of the frequency difference [$|\Delta f|$ X sign(Δf): $F(2, 14) = 1.554$, $p = 0.246$]. In pair-wise two-tailed *t*-tests, there were no significant differences between any of the two conditions with the same $|\Delta f|$. There was also no significant difference between the Δf = 0 condition and the control no-target condition.

2.3.5.4 Target *intensity trading*

The percentage responses found in the single-object control experiment provide mappings that allow us to evaluate directly whether there is a trading relationship between the level of target perceived in the tone stream and in the vowel. The individual psychometric functions that relate the target attenuation in a single-object stimulus to a percent response were used to find the equivalent target attenuation for each subject and condition. The across-subject means and standard errors of these mapped equivalent attenuations are plotted in Figure 2-5C (for the tones) and Figure 2-5F (for the vowel).

Overall, these results are similar to the percent response and δ' results for Figure 2-5. This shows that this remapping is relatively similar for all subjects and across conditions. However, this analysis has the advantage of enabling a direct comparison between the energy that the target contributions to the tones and that it contributes to the vowel and to check to see whether there is energy trading of the target.

Figure 2-6A plots the across-subject mean effective attenuation of the target in the tone stream against the mean attenuation of the target in the vowel. The plot shows all conditions that were common to the two experiments including the two-object target-absent control (see ex symbol). The solid curve in the figure plots the trading relationship that would be observed if the effective energy the target heard in the two objects summed to the physical energy in the actual target stimulus.

Results for the target-absent control fall near the upper-right corner of the plot, as expected, indicating that the perceived qualities of the tone stream and vowel were consistent with a target that was highly attenuated. For the ambiguous stimuli, trading did occur: the effective attenuation of the target in the tone stream is larger for stimuli that produce less attenuation in the vowel. However, the trading does not strictly follow the predicted absolute physical energy summation rule: in general, the total of the sum of the effective energies of the target in the two streams is less than that actual present in the target (this can be seen in the fact that the data points fall above and right of the solid line in the figure).

To quantify the "trading relationship" observed in Figure 2-6A, we computed the total effective energy of the target by summing, for each condition, its effective energy when attending the tones and its effective energy when attending the vowel. We then computed the "lost" target energy by subtracting the total effective target energy from the physical energy of the target. The across-subject means of these values are shown in Figure 2-6B.

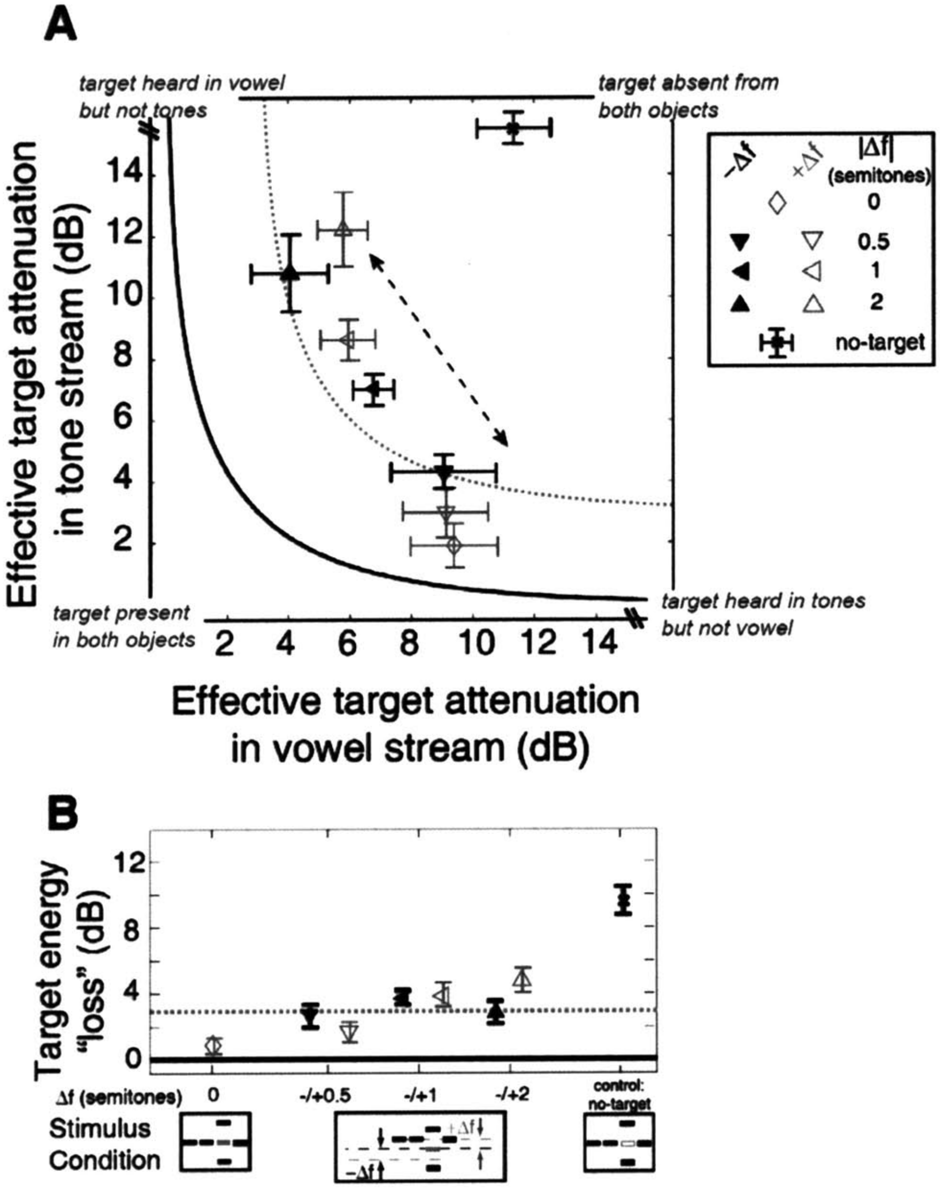


Figure 2-6(A): Scatter plot of the effective target attenuation in the tones versus the effective target attenuation in the vowel. Data would fall on the solid line if energy conservation holds. An energy trading relationship in which the total perceived target energy is 3 dB less than the physical target energy would fall on the dashed line (equivalent to conservation of amplitude, rather than energy; see [53]. **(B)**: The lost energy of the target for each condition, equal to the difference between the physical target energy and the sum of the perceived target energy in the tones and vowel. The solid line (zero dB lost energy) shows where results would fall if energy conservation holds. The dashed line shows where results would fall if amplitude, rather than energy, conservation holds. All error bars show ±1 standard error of the means.

In general, the total perceived target energy was less than the physical target energy in the stimuli. The lost energy was near 3 dB for many of the stimuli (see dashed line at 3 dB, which is equivalent to trading amplitude, rather than intensity; see also [53, 59]). A repeated measures ANOVA on the total effective target energy lost found no effects of frequency difference [$|\Delta f|$, $F(2,14) = 4.29$, $p = 0.0544$], the sign of the frequency difference [sign(Δf), $F(1,7) = 2.062$, $p= 0.194$], or their interaction [$|\Delta f|$ X sign(Δf), $F(2,14) = 3.466$, $p = 0.0599$] on the amount of energy lost. One-sample *t*-tests performed separately on the lost target energy values explored whether the lost energy was statistically significantly different from zero dB (with Dunn-Sidak post-hoc adjustments for seven planned comparisons). For conditions $\Delta f = -0.5$ ($t_7 = -4.06$, $p_{DS} < 0.0333$), $\Delta f = -1$ ($t_7 = -9.30$, $p_{DS} < 2.42 \times 10^{-4}$), $\Delta f = +1$ ($t_7 = -5.37$, $p_{DS} < 0.00729$), and $\Delta f = +2$ semitones ($t_7 = -6.13$, $p_{DS} < 0.00333$), the lost energy was significant greater than zero, supporting the conclusion that, in general, energy conservation fails.

2.3.6 Discussion

In many past studies of this sort, adaptation in the periphery has been brought up as a possible explanation for the reduced contribution of the target to the harmonic complex [44]. Peripheral adaptation could contribute to the "lost" target energy here, as well. However, such adaptation would be greatest when $|\Delta f| = 0$ and the tones have the greatest effect on the target. Instead, the amount of target energy that is lost is smallest when $|\Delta f| = 0$. If peripheral adaptation were the only factor contributing to the lost energy, the lost energy should be greatest, not least, for this condition. Thus, adaptation is not able to account for the perceptual loss of target energy observed heard (also see the Discussion in the Appendix [56]).

While some target energy is not accounted for, we found a trading relationship in that the greater the contribution of the target to the tone stream, the smaller its contribution to the vowel. This finding is similar to past studies [53, 59] that found a lossy trading of an ambiguous element between two competing objects.

However, this result contrasts with results using similar stimuli with $|\Delta f| = 0$ but in which the spatial cues of the tones and target were manipulated to change the relative strength of simultaneous and sequential grouping, rather than the frequency difference between tones and the ambiguous target (Appendix [56]). Taken together, these results suggest that lossy trading occurs under many circumstances, but not all. Moreover, none of these studies finds energy conservation, where the perceived target energy in the competing objects sums to the physical energy of the target. In this respect, the current results support the idea that the way in which the acoustic mixture is broken into objects does not follow energy trading, despite the intuitive appeal of this idea (see Appendix [56]).

2.4 Experiment 2: No competing streams

In the first experiment, subjects were more likely to hear the target as part of the vowel complex as the frequency separation between the repeating tones and the target increased. In conditions where $|\Delta f|$ = 2 semitones, subjects reported a strong "galloping" percept. However, studies using single-object tone stimuli generally find that a two-semitone difference is not enough to cause a single object to break apart into two objects [23, 24, 26, 34]. This suggests that the presence of the vowel, which competes for ownership of the target, influenced the effectiveness of frequency separation in reducing the contribution of the target to the tone stream. A follow-up experiment was conducted to assess directly whether the strong effect of a relatively small frequency separation on the perceived tone-stream rhythm was a consequence of the competition between tones and vowel for ownership of the target, rather than due to some procedural or stimulus differences between the current and past studies.

2.4.1 Methods

Stimuli were similar to the single-object tone stimuli used in the previous experiment. In the two-object experiment, frequency separations of only up to two semitones were tested, as any bigger separation would make the repeating tones closer to neighboring harmonics than to the target (which could cause the repeating tones to capture those harmonics rather than the target). Here, we tested separations between the repeating tones and the target of up to eight semitones (i.e., 397, 445, 472, 486, 500, 515, 530, 561, and 630 Hz) to make results more comparable to previous single-object experiments.

Subjects were instructed to judge the repeating tone-stream rhythm ("galloping" versus "even") after 10 presentations of the repeating-tones-target triplet. Eight subjects participated in this experiment, six of whom participated in the previous experiment and two who had previously participated in and passed the screening criteria in other experiments conducted in our laboratory.

2.4.2 Results

All subjects could distinguish the "galloping" from the "even" prototypes nearly perfectly. Figure 2-7A shows the raw percentage response scores, and Figure 2-7B shows the results $\delta'_{condition:absent}$ results, averaged across subjects.

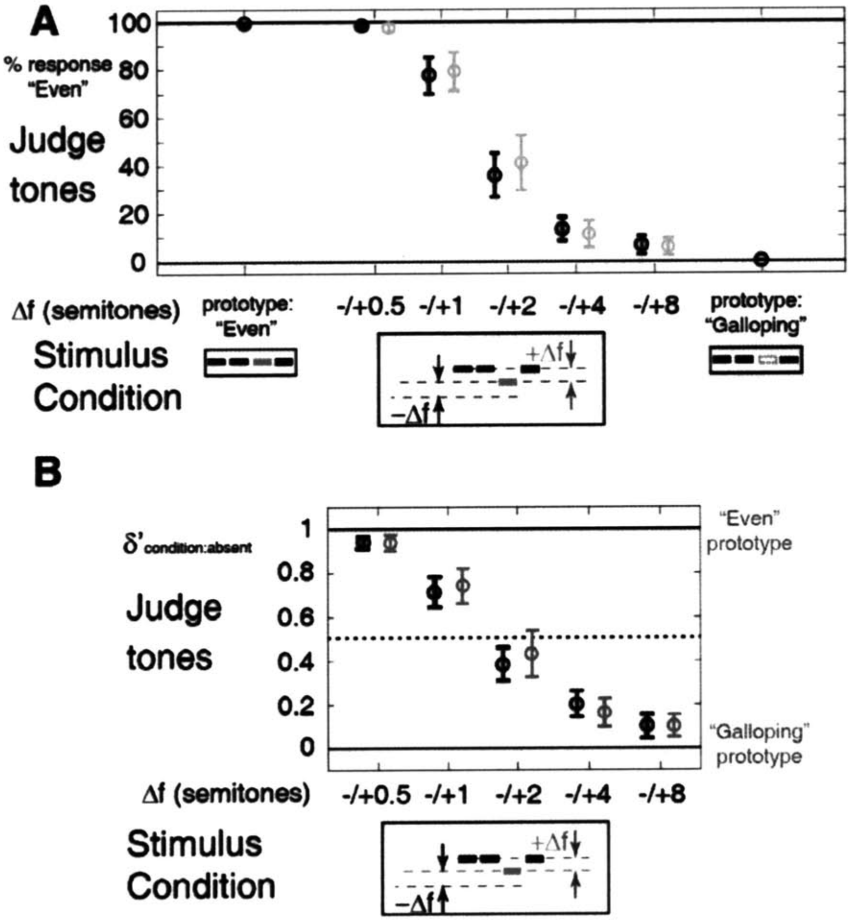


Figure 2-7(A): Raw percent responses for single-object stimuli as a function of the physical attenuation of the target. The target-present prototype is equivalent to the Δf = 0 condition. Note that there is no equivalent vowel manipulation in this experiment. **(B)**: Normalized $\delta'_{condition:\,absent}$ results, derived from the raw percent-category responses in (A).

A repeated measures ANOVA on the one-stream tone rhythm $\delta'_{condition:\,absent}$ results found a significant main effect of the absolute magnitude of the frequency difference [|Δf|: $F_{GG}(1.74,12.2) = 67.0$, $p_{GG} < 4.27 \times 10^{-7}$], but did not find a significant effect either for the sign of the frequency difference [sign(Δf): $F(1,7) = 0.078$, $p = 0.789$] or the interaction between |Δf| and the sign of the frequency difference [|Δf| X sign(Δf): $F(4, 28) = 0.455$, $p = 0.768$]. These results show that for our rapidly repeating stimuli, separations of two semitones increase the likelihood of hearing the repeating tones as "galloping." However, responses are perceptually only halfway between "galloping" and "even" for separations of two semitones. In contrast, when the vowel was present, this same $|\Delta f|$ was enough to cause all responses to be "galloping." For the repeating-tones-only stimuli, the stimulus is heard as "galloping" with comparable probability only when the frequency separation is eight semitones.

2.4.3 Discussion

Compared to previous sequential streaming investigations, we found that the target and tones were heard in distinct streams for relative small frequency separations. Past studies show that the potency of a particular frequency separation on streaming depends on the repetition rate, subject instructions and even the musical training of the subjects [23]. Therefore, the most important comparison is with the results of the tone stream judgments in the two-object experiment (Experiment 1), using identical stimulus parameters, instructions, and conditions tested were identical.

In Experiment 1, subjects judged the tone stream as "galloping" when the absolute frequency separation between the tones and the target was 2 semitones apart. In the absence of the competing harmonic complex, subjects only judged the tone stream to be strongly "galloping" when the absolute frequency separation between the tones and the target were 4 or more semitones apart. These results[2] show that sequential streaming is affected by the presence of competing auditory objects.

2.5 General discussion

Many past studies of auditory object and stream formation investigated the potency of different acoustical grouping cues; however, the majority focused exclusively on either sequential grouping [19, 23, 24, 60] or simultaneous grouping [14, 36-38]. Most of these studies explored what acoustical parameters

[2] aside from any contextual effects that might be associated with this follow-up experiment

would lead sound elements to be heard as one object and what would lead the stimulus to break apart into two perceptual objects. However, in everyday complex settings, multiple acoustical objects often coexist. In such situations, it is more natural to ask how simultaneous objects interact and influence grouping of ambiguous elements, like our target, rather than whether a mixture is heard as one or two objects.

2.5.1 Scene analysis in audition and analogies drawn from vision

The problem of how to organize sensory inputs into perceptual entities arises in all modalities. In the visual scene analysis literature, Gestalt theory has been used to describe perceptual organization in vision and audition evolved to address processing of visual and auditory scenes [3, 9, 10, 61, 62]. Many Gestalt laws of grouping, such as proximity and common fate, were recognized to operate both on the auditory and the visual modalities.

In auditory scene analysis, parallels to the visual literature are involved to develop concepts like the principle of figure-ground organization [63, 64]. Neural correlates of such processes have been identified recently [65, 66]. In vision, the foreground object generally occludes background objects. Therefore, edges are important in figure-ground analysis of visual scenes [67]. While visual edges are well defined [68], the proper definition of an auditory edge is not clear cut [69-72]. Generally speaking, the critical structure in sound that enables object formation is in the spectro-temporal correlations that arise, including common amplitude modulation, onsets, offsets, and harmonic relationships across sound elements [6, 12]. Thus, there is no direct analogy to an "edge" in vision in the auditory scene: however, a generalized form of across element correlation exists.

Another important difference between the auditory and visual scenes is that a sound in the auditory foreground does not occlude a sound with the same spectro-temporal content in the background [69]. From this observation, the acoustic scene is often described as transparent [3]. Consider the spectro-temporal signal that reaches our ears. Assume all auditory sources are co-located and that the source separation is based upon the spectro-temporal information at one receiver, denoted by $R(f,t)$ with the two parameters f and t representing frequency and time, respectively. Each source S_i can be expressed as a sum of frequency components:

$$S_i(f,t) = \sum_{j=1}^{F_i} A_{ij} \cos(2\pi f_{ij} t + \psi_{ij}), \qquad (2.6)$$

Similarly, the total signal at the receiver is the superposition of these sources.

$$R(f,t) = \sum_{i=1}^{s} S_i(f,t) = \sum_{i=1}^{s} \sum_{j=1}^{F_i} A_{ij} \cos(2\pi f_{ij} t + \psi_{ij}). \qquad (2.7)$$

In order to understand and process the sources in the environment, the listener must decompose $R(f,t)$ to try to recover the original sources. This process may be imperfect. The resulting estimates of the sources can be written as $\hat{S}_i(f,t)$. Perfect scene analysis would correspond to:

$$\hat{S}_i(f,t) = S_i(f,t) \quad \forall i. \qquad (2.8).$$

Combing the expression for $R(f,t)$ in equation (2.7) with the analysis in Section 2.7 shows that for uncorrelated sources, the expected intensity of the sound energy of frequency f at the receiver, $\overline{I}_S(f,t)$, is given by

$$\overline{I}_S(f,t) = \sum_{i=1}^{N} \overline{I}_{S_i}(f,t), \qquad (2.9),$$

where $\overline{I}_{S_i}(f,t)$ is the intensity of the component of sound of frequency f in the physical source i at time instant t. Similarly, if the listener is able to separate the sources perfectly, then the perceived intensity of sound energy of frequency f in each perceived object will sum to equal the true energy in the original source so that at each time instant t:

$$\overline{I}_S(f,t) = \sum_{i=1}^{N} \hat{\overline{I}}_{S_i}(f,t)\,. \qquad (2.10).$$

While it is nearly impossible to perfectly estimate the spectral content of the original sources making up a mixture, equation (2.10) provides a constraint on the true solution. Moreover, this constraint is relatively simple to meet. Put into words, equation (2.10) states that the sum of perceived energies at frequency f in all of the sources should, on average, equal the physical energy of that frequency in the signal reaching the ear. Thus, equation (2.10) quantifies the trading relationship that intuitively seems logical.

This trading relationship highlights the transparency of the auditory scene. In particular, the content of the signal of a given frequency reaching the sensory epithelium is often a sum of contributions from multiple sources. In most computational approaches to auditory scene analysis, this trading relationship is implicitly assumed [73, 74]. However, to the best of our knowledge, the intensity trading hypothesis has not been explicitly tested. By studying how assignment of the spectro-temporal of sound in a mixture is allocated across perceived streams, new insight into scene analysis can be gained.

2.5.2 Conceptual models for streaming

Let $\mathcal{D}(\cdot)$ represent an operator that, utilizing all available information from the signal reaching the receiver, $R(f,t)$, produces as output the best estimates of the content of all the auditory sources in the scene $\{\hat{S}_i(f,t)\}$. As psychophysicists, we wish to understand the operator $\mathcal{D}(\cdot)$, but cannot directly observe its operation. One method for building insight into the operator is to ask the subject to make a subjective judgment about the spectro-temporal content of an auditory stream of interest. It is worth noting that using this method, the nature of the task necessarily brings the stream of interest into the auditory foreground. In order to assess the detailed spectro-temporal content of the other stream, one must direct the subject to attend to the other object, bringing it into the attentional foreground. If the perceptional organization changes, depending on what object lies in the attentional foreground, then the total perceived energies in objects when they are in the foreground may not obey the energy conservation rule in equation (2.10).

Figure 2-8A shows a realization of such a process with the decomposing function when $\mathcal{D}(\cdot)$ is independent of what object is attended. In this process, it is assumed that 1) perceptual grouping is a bottom-up process, purely based on the statistics of the signal $R(f,t)$, and 2) a subject can select or reject a perceptual group of spectro-temporal elements as a whole (c.f., [27] Figure 1b). In this realization, the spectro-temporal content of the background does not change when the background becomes the foreground, and it is more likely that the total perceived energy will be conserved. Figure 2-8B depicts the alternative in which perceptual organization changes with attentional focus. $\mathcal{D}(\cdot)$ can differ depending on the object being attended, one can no longer directly infer what energy will be in the competing object when it is the focus of attention. In order to distinguish from the system denoted by Figure 2-8A, a new operator, $\mathcal{D}_{TDA|S_i}(\cdot)$, is used to denote that the decomposing process is influenced by the object being attended. In general, $\mathcal{D}_{TDA|S_i}(\cdot) \neq \mathcal{D}_{TDA|S_j}(\cdot)$ for $i \neq j$. Evidence that perceptual organization develops over time [24, 26, 27, 42] further suggests that the operator $\mathcal{D}(\cdot)$ is not time-invariant.

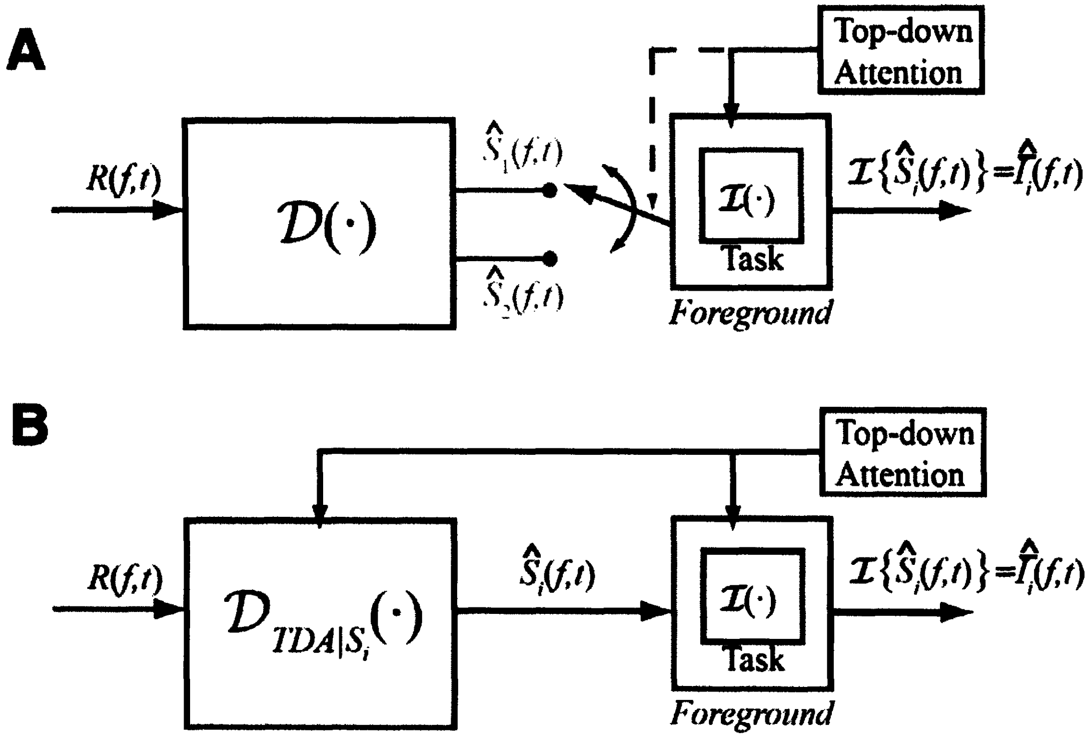


Figure 2-8 (A): A realization of the decomposing process that is not influenced by top-down attention. **(B)**: A realization of the decomposing process that is driven by attention. The decomposing process depicted in **(B)** contains non-linearity introduced by top-down processes and therefore energy trading hypothesis is not expected to hold. Note that the operator $\mathcal{I}(\cdot)$ denotes the listener's task of reporting the intensity of the spectro-temporal content.

2.5.3 Attention-dependent segmentation will cause failures in the trading hypothesis

The validity of the intensity trading hypothesis depends on whether the process, $\mathcal{D}(\cdot)$ is linear[3]. Moreover, one source of possible non-linearity is in the auditory periphery. For example, studies into the macro-mechanics [75, 76] and the electrical transduction [77, 78] have characterized many sources of non-linearity in the cochlear. Using models such as the auditory image model [79], the effects of possible non-linearity processing through the level of the hair cell level can be,

[3] i.e., it satisfies both the additivity and homogeneity properties.

by and large, taken into account (e.g. redefining $R(f,\tau)$ at the hair-cell transduction level, rather than at the ear drum). However, given that the stimuli used in this and many other studies investigating auditory scene analysis designed to rule out known effects of non-linearity in the peripheral system (e.g. [44], any failure in the intensity trading hypothesis strongly implicates central processing mechanisms, possibly involving top-down processes mediated by task-switching or attention.

Consider the bottom-up system in Figure 2-8A, whereby $\mathcal{D}(\cdot)$ is independent of the task. If the segregation function $\mathcal{D}(\cdot)$ is linear and time-invariant, then the intensity trading hypothesis must hold. Even if the segregation process allows for build-up of streaming, (i.e., is time-varying), the energy trading hypothesis is still valid at any time t, since the build-up of the process affects the formation of all perceptual objects at the same rate.

In contrast, consider the task-driven model in Figure 2-8B. Top-down processes potentially influence the transformation, $\mathcal{D}_{TDA|S_i}(\cdot)$, depending on the object being attended. In such a model, the intensity trading hypothesis is not expected to hold because the intensity observed depends on the object being attended, and that in general $\mathcal{D}_{TDA|S_i}(\cdot) \neq \mathcal{D}_{TDA|S_j}(\cdot)$ for $i \neq j$. The dependency of attention on the transformation can exist at many different stages along the auditory pathway. At the highest cognitive level, one might expect the task to influence how different cues in the available spectro-temporal cues are weighted. For example, spatial cues only weakly affect across-frequency grouping, but have a strong influence on across-time streaming [36, 37].

Depending on whether the task focuses attention on an object whose primary organization is across time or across frequency, spatial cues may have either a strong or weak effect on object formation, and the organization of the scene may therefore change, depending on what object is attended. In the visual modality, it has been suggested that attention modulates the response along the visual pathway (e.g., [80-82] including at the peripheral level [83]). In audition, there is also evidence suggesting that spectro-temporal receptive fields in the primary auditory cortex of behaving ferrets [84, 85] change depending on the behavioral task. Similarly attention has been implicated in corticofugal modulation of cochlear function in awake mustached bats during vocalization [86]. If the behavioral task modulates the spectro-temporal peripheral response, or influences how different cues are weighted in the formation of objects, such task-dependent effects would be expected to break the trading hypothesis.

2.5.4 The effects of competing objects on perceptual organization

In our two-object experiment, the presence of the tones at the same frequency as the target is enough to substantially remove the contribution of the target in the harmonic complex. This result supports observation that pitch perception is influenced by sequential streaming [44]. We further extended this observation by systematically weakening the frequency proximity of tones and target, thereby changing the relative potency of the across time streaming in perceptual organization. Results of Experiment 2 showed that this change in frequency proximity is not sufficient to pull the target out of the tones stream in the absence of competition for the target. Thus, sequential grouping is influenced by simultaneous grouping.

At first glance, there appears to be a trading relationship in these experiments, supporting the bottom-up streaming model depicted in Figure 2-8A. However, there is a systematic 3 dB of target energy that is not accounted for by first examining the contribution of the target to the tones and then considering its contribution to the vowel. This discrepancy may be due to a non-linearity in peripheral processing. For instance, if there is adaptation in the perceived level of the target due to the presence of the repeated tones, the total perceived energy in the two object mixture could be less than the true physical energy of the target. Of course, peripheral adaptation cannot be the full explanation for the current results. Peripheral adaptation of the target should be maximal when the tones and target have the same frequency. Instead, the energy that is "lost" through competition for the target is lowest when the target frequency matches that of the repeated tones. However, some other nonlinearity in the target level represent in the auditory periphery that depended on |Δf| could explain the current results.

From the current experiments, we cannot reject the bottom-up-only model. An interesting experiment to definitively rule out a pure peripheral argument for the lack of a trading relationship would test perception with using stimuli that had identical spectro-temporal contents but differed in some other way that affected grouping (e.g., in interaural time differences). If the trading relationship was violated for such stimuli, it suggests that a more central process, such as attention-dependent segmentation, influences object formation.

2.6 Conclusions

Results are consistent with there being a trading relationship that governs how an ambiguous spectro-temporal element is allocated between two competing auditory objects, one grouped across time, (i.e., the stream of tones) and one grouped across frequency, (i.e., the vowel complex,). However, the observed trading does not account for all of the physical energy present in the target. Whenever the frequency separation between target and repeated tones was non-zero, the total perceived target energy was roughly 3 dB less than the physical target energy. Such a result may be explained by some peripheral nonlinearity in the represented target level (e.g., at the level of the auditory nerve). However, the most obvious non-linearity (adaptation) predicts a result in the opposite direction from what was observed. Alternatively results may be a manifestation of a system in which object formation is task and / or attention dependent. Regardless of what nonlinearity causes the failure of a strict energy trading of the target, these results conclusively prove that the organization of an across-time object influences the organization of simultaneous elements just as the simultaneous elements influence the formation of an across-time stream.

2.7 Formalizing the trading relationship

Assume that a single receiver receives $R(f,t)$, the superposition of two independent sound sources, S_1 and S_2, representing tones and vowel, respectively. The ambiguous target (f = 500 Hz) can logically belong to S_1 and S_2 or can be linear combination of elements of frequency f in the two objects:

$$R(500,t)=\sum_{i=1}^{2}A_i(500,t)e^{-j(2\pi 500t+\psi_i(500,t))}, \qquad (2.11)$$

where $A_i(500,t)$ is the amplitude of the 500 Hz component originated from the i^{th} source, and $\psi_i(500,t)$ is the phase component of that frequency associated with the i^{th} source at time instant t. Assuming that all ψ_i's are uncorrelated and identically, uniformly distributed between 0 and 2π, the expected value of $R^2(500,t)$ is:

$$E\left\{R^2(500,t)\right\}=\frac{1}{2}\sum_{i=1}^{2}A_i^2(500,t)=\sum_{i=1}^{2}A_{i,rms}^2(500,t). \qquad (2.12)$$

To simplify the analysis, assume that plane wave approximation is valid and thus let us define the specific acoustic impedance to be a real quantity, $z_0=\rho c$, where ρ is the density of the air (or other medium) and c is the velocity of sound in the medium.

Let us also define the average intensity as:

$$\overline{I}_R(f,\tau)=\frac{|R(500,\tau)|^2}{\rho c}, \qquad (2.13)$$

then the expected average intensity of the 500 Hz will obey the trading relationship:

$$\overline{I}_S(500,t)=\overline{I}_{S_1}(500,t)+\overline{I}_{S_2}(500,t). \qquad (2.14)$$

In general, if there are N uncorrelated sources with the frequency component f, the expected average intensity of that component is:

$$\bar{I}_S(f,t)=\sum_{i=1}^{N}\bar{I}_{S_i}(f,t)\,. \qquad (2.15)$$

If veridical parsing occurs, then

$$\bar{I}_S(f,t)=\sum_{i=1}^{N}\hat{\bar{I}}_{S_i}(f,t)\,. \qquad (2.16)$$

CHAPTER 3 VIOLATION OF TRADING HYPOTHESIS

Portion of the work described in this chapter has been submitted to the *Journal of the Association for Research in Otolaryngology*.

Lee, A.K.C. and B. Shinn-Cunningham, *Effects of reverberant spatial cues on attention-dependent object formation*, submitted.

3.1 Abstract

A recent study showed that when a sound mixture has ambiguous spectro-temporal structure, spatial cues alone are sufficient to change the balance of grouping cues and affect the perceptual organization of the auditory scene. The current study synthesizes similar stimuli in a reverberant setting to see whether the interaural decorrelation caused by reverberant energy reduces the influence of spatial cues on perceptual organization. Results suggest that reverberant spatial cues are less influential on perceptual segregation than anechoic cues. In addition, results replicate an interesting finding from the earlier study, where an ambiguous tone that could logically belong to either a repeating tone sequence or a simultaneous harmonic complex can sometimes "disappear" and never be heard as part of the perceptual foreground, no matter which object a listener attends. As in the previous study, the perceived energy that the ambiguous element contributes to objects in a complex scene does not equal the physical energy of that element in the sound mixture. This loss of perceived target energy depends on an interaction between spatial and non-spatial grouping cues, consistent with the idea that the perceptual organization of an acoustic mixture depends on what object a listener attends.

3.2 Introduction

In our everyday perceptual experiences, objects rarely occur in isolation. In all sensory modalities, the information available at our sensory epithelia is a chaotic mixture of different elementary sensations arising from separate physical sources in the environment [9]. In order to make sense of these mixtures of signals, a cognitive process known as scene analysis must group sensory elements together into *objects* (estimates of what sensory inputs coming from a single physical source in the external world). Gestalt theory has been used to describe this perceptual organization [10], and many Gestalt laws of grouping, such as proximity and common fate, are known to influence object formation both in the visual [62] and auditory [3] modalities.

While there are many similarities between visual and auditory scene analysis, differences in the physical properties of light and sound and how they propagate to our eyes and ears [69, 71] as well as the organization of the sensory epithelia [70, 72] lead to differences in the heuristics that the brain uses to estimate the content of visual and auditory objects. An important difference between the visual and auditory scenes, for example, is that a visual object that is closer to the observer generally occludes an object that is further away. In contrast, two sounds that contain energy in the same frequencies at the same time sum acoustically before entering the ear. As a result, the auditory scene is often described as 'transparent" [3].

Physically, the total energy a listener perceives at a given time and frequency can be broken down into components from different objects. Intuitively, this suggests that if there is an ambiguous element that could logically belong to more than one competing object in the scene, the listener should allocate the physical energy in the ambiguous element to these objects in a way that obeys "energy conservation:" the sum of the energies that the ambiguous element contributes to the objects perceived in the scene should equal the physical energy in the sound mixture.

While the idea of energy conservation is intuitively appealing, only a handful of studies [53, 56, 59, 87] have explicitly tested whether it holds. Moreover, the results of these studies are mixed. While none of the studies found evidence that energy conservation occurs, three of the four studies found support for "energy trading" (i.e., when the ambiguous element was heard more prominently in one competing object, it was less prominent in the other object [53, 59, 87]. However, one study, in which spatial cues were manipulated to affect the relative strength of different grouping rules, found a stimulus in which an ambiguous tone was never heard prominently in either of the objects competing for its ownership (Appendix [56]). The experimenters pointed out that if perceptual organization depends on what object is attended, there is no reason to expect energy conservation, or even energy trading, to hold. The experimenters suggested that energy trading fails because which object is attended determines the relative importance of various grouping cues, causing the perceptual organization to shift, depending on which object is in the attentional foreground.

Due to the transparent nature of the auditory scene, distinct objects can come from the same location in space (e.g., a single loudspeaker can simultaneously emit the sound of a violin and a piano). In addition, unlike in the retina, the cochlea does not have an explicit spatial representation of sound sources. Auditory spatial information must be calculated neurally, based on differences in the signals reaching the two ears and in the spectral content of the signals received [88]. Interaural time differences (ITDs) and interaural intensity differences (IIDs) between the signals at the two ears are arguably the most robust cues for source location. Perhaps as a result, and in contrast to their role in visual object formation, spatial cues generally affect auditory object formation over short time scales only weakly. Instead, local spectro-temporal cues such as harmonicity and common onsets generally determine how simultaneous sounds are grouped into objects. While spatial cues only weakly influence simultaneous grouping, they play a prominent role in sequential grouping and selective attention [7, 12, 55, 89, 90].

These differences in how spatial cues affect simultaneous and sequential grouping build intuition into why attention may alter perceptual organization of a scene and why energy trading is not always observed. In particular, in the "orphan" condition in which the ambiguous target element was lost (Appendix [56]), the objects competing for the target element were a sequential tone stream and a simultaneous harmonic complex. In the "orphan" condition, the spatial cues supported grouping the target with the simultaneous harmonic complex, while the overall spectro-temporal structure generally supported hearing the target as part of the sequential tone stream. Thus, when listeners focused attention on the sequential stream, where sequential grouping cues might be expected to determine how the foreground object is grouped, listeners may have weighted spatial cues heavily, and relegated the target to the perceptual background. In contrast, when attending to the simultaneous harmonic complex, listeners may have weighted spectro-temporal cues heavily and been less influenced by spatial cues. Again, this choice would have relegated the target to the perceptual background.

The current study tests whether energy trading fails for stimuli like those in the previous study, but for which spatial cues are made more ambiguous. In particular, natural reverberant energy degrades the fidelity of ongoing interaural time differences by decorrelating the left and right ear signals [91-94]. We hypothesized that the failure of energy trading depends on there existing a fragile balance between the relative strengths of simultaneous and sequential grouping cues. Therefore, we speculated that weakening the spatial cues might shift the balance to favor spectro-temporal structure and reduce the influence of spatial cues on perceptual organization. Depending on the extent to which the influence of spatial cues was reduced, this might produce stimuli for which energy trading occurs. The stimuli used here are identical to those of the previous study (Appendix [56]), except that stimuli were convolved with binaural room impulse responses (BRIRs) that contained natural room reverberation [94].

3.3 Methods

3.3.1 Subjects

Nine subjects (eight male, one female, aged 18-32) took part in this experiment. All participants had pure-tone thresholds of 20 dB HL or better at all frequencies in the range from 250-8000 Hz, in both ears, and their threshold at 500 Hz was 15 dB HL or better. All subjects gave informed consent to participate in the study, as overseen by the Boston University Charles River Campus Institutional Review Board and the Committee On the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology.

3.3.2 Stimuli

Stimuli consisted of a repeating sequence of a pair of tones followed by a harmonic complex (Figure 3-1A), similar to those used by (Appendix [56]). The pair of tones had a frequency of 500 Hz. Each tone was 60 ms in duration, gated with a Blackman window of the same length. The harmonic complex was filtered with a formant filter to simulate the spectral content of a short vowel [95]; see also Chapter 2 [87]). The first, second and third formant peaks were set to 490, 2100 and 2900 Hz, respectively (similar to [95]). Each harmonic of the simultaneous complex was also 60 ms in duration, gated by the same Blackman window used for the repeating tones. The target was a 500-Hz tone temporally aligned with and with the same onset / offset as the harmonic complex (60 ms in duration, gated with a 60-ms-Blackman window). As a result of this structure, the target could logically be heard as the third tone in the repeating tone stream or as the fourth harmonic in the harmonic complex.

The magnitude of the target matched that of the repeating tones and the formant envelope of the vowel. There was a 40-ms-long silent gap between each tone and harmonic complex, creating a regular rhythmic pattern with an event occurring every 100 ms. This basic pattern, a pair of repeating tones followed by the vowel complex / target, was repeated ten times per trial to produce a three-second-long stimulus. This produced the percept of two objects: an ongoing stream of tones and a repeating vowel occurring at a rate one-third as rapid.

The rhythm of the tone sequence and the identity of the vowel depend on whether or not the target is perceived as part of the respective object. Specifically, the tone stream is heard as "even" when the target is heard in the stream and "galloping" when the target is not perceived in the stream. The complex is heard more like /ɛ/ when the target is perceived as part of the vowel and more like /ɪ/ when it is not part of the vowel (Figure 3-1B).

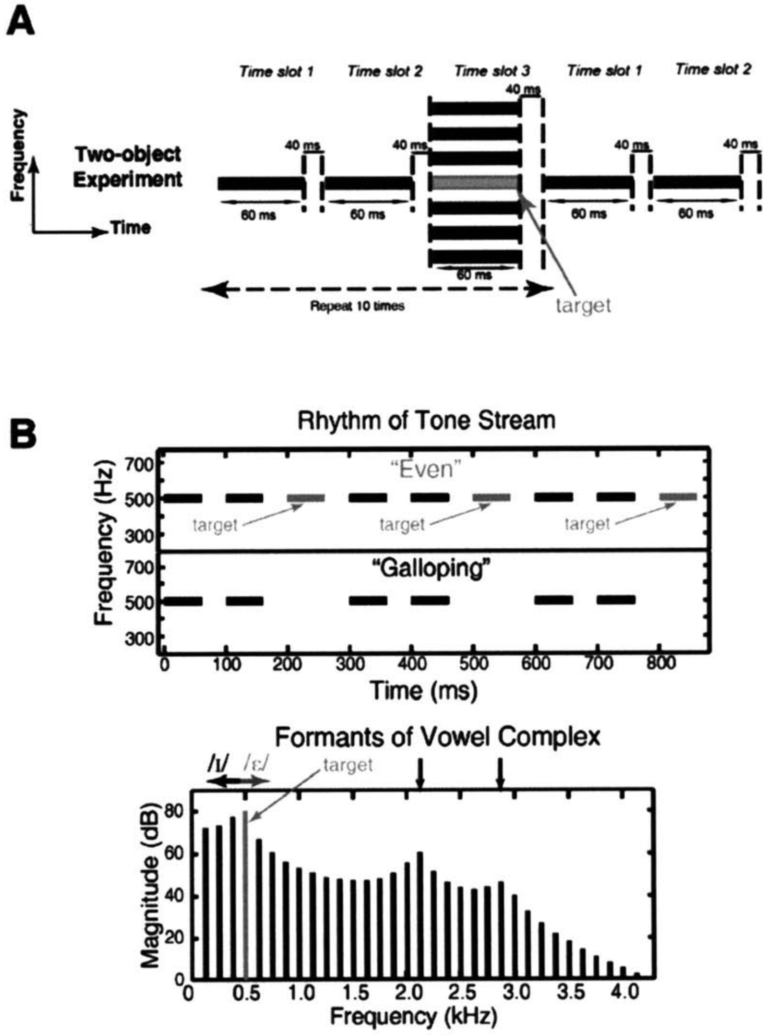


Figure 3-1(A): Two-object stimuli were created by repeating a three-item sequence that consisted of a pair of pure tones followed by a harmonic complex. In the reference configuration, the pure tones in time slots 1 and 2 are at 500 Hz. Time slot 3 is made up of two components: a target tone at 500 Hz and a tone complex with an F0 of 125 Hz (with the fourth harmonic 500 Hz omitted). The tone complex is shaped by a synthetic vowel spectral envelope to make it sound like a short vowel [53]. Because the first formant of the vowel complex is near 500 Hz, the relative level of the target tone perceived in the vowel complex affects perception of the first formant frequency, which affects the perceived identity of the vowel. **(B)** (Top Panel): The perceived rhythm depends on whether or not the 500 Hz target tone is perceived in the sequential tone stream: if the target is grouped with the repeated tones, the resulting rhythmic percept is "even;" if the target is not grouped with the pair of tones, the resulting perceived rhythm is "galloping." (Bottom Panel): The synthetic vowel spectral envelope is similar to that used by Hukin and Darwin [54]. The identity of the perceived vowel depends on whether or not the 500-Hz target is perceived in the complex: the vowel shifts to be more like /ɛ/ when the target is perceived as part of the complex and more like /ɪ/ when the target is not perceived in the complex. The arrows indicate the approximate locations of the first three formants of the perceived vowel.

Control stimuli consisted of single-object presentations (only the tones or only the harmonic complex) either with the target ("target-present" prototype) or without the target ("target-absent" prototype). Finally, a two-object control was generated in which the repeating tones and the complex were presented together, but in which there was no target ("no-target" control).

3.3.3 Environment

All stimuli were generated offline using MATLAB software (Mathworks Inc.). Signals were processed with BRIRs measured in a classroom[4] with a manikin head located in the center of the room and the sources one meter away, either originating from 0° or 45° to the right of the manikin [94].

Digital stimuli were generated at a sampling rate of 25 kHz and sent to Tucker-Davis Technologies hardware for D/A conversion and attenuation before presentation over headphones. Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. A randomized roving attenuation level between 0 and 14 dB was applied to the stimulus for each trial before presentation in order to reduce the reliability of absolute presentation level as a cue in the identification task. Subjects were seated in a sound-treated booth and responded via a graphical user interface. Stimuli were presented over insertion headphones (Etymotic ER-1). All signals were presented at a listener controlled, comfortable level that had a maximum value of 80 dB SPL.

3.3.4 Task

In order to assess perceptual organization of the two-object mixture and how it affected the perceived content of both the tone stream and the vowel, the same physical stimuli were presented in two separate experimental blocks. In one

[4] $T_{60} \simeq 600$ ms [94].

block, subjects judged the rhythm of the tone sequence ("galloping" or "even") using a one-interval, two-alternative-forced-choice design. In the other block, the same physical stimuli were presented in a different random order, and subjects judged the vowel identity ("/ɪ/ as in 'bit'" or "/ɛ/ as in 'bet'").

3.3.5 Training procedure with single-object prototypes

In each session of testing, each subject was familiarized with the single-object prototypes with and without the target. In the rhythmic block of the experiment, subjects were trained to label a stream of 500-Hz tones with the target present as "even," and to label the tones without the target present as "galloping." In the corresponding vowel training runs, subjects were trained to label the harmonic complex with the target present as /ɛ/ (as in 'bet') and the harmonic complex without the target as /ɪ/ (as in 'bit').

In the training phase of the experiment, subjects were given feedback to ensure that they learned to correctly label the single-object, target-present and target-absent prototypes. This feedback ensured that subjects could accurately label the tone stream rhythm and the harmonic complex identity for unambiguous, single-object stimuli. Subjects had to achieve at least 90% accuracy when discriminating between the two prototypes in the single-object pre-test before proceeding to the two-object experiment.

3.3.6 Procedures for the main, two-object experiment

Following training on single-object prototype stimuli, listeners judged the tone stream rhythm and the vowel identity for stimuli that had both objects present. The spatial configuration of the repeating tones and the target (either consistent with a source from 0° or 45°) was varied to ascertain how spatial cues influenced the perceptual grouping of the stimuli in a reverberant environment. In all two-object trials, the vowel was always presented at 0° azimuth. Four different spatial configurations were tested, differing in whether the spatial cues of the vowel and

/ or the repeating tones matching that of the target (Figure 3-2). A control two-object condition was also included in which the target was not presented.

Intermingled with the two-object trials were single-object control trials (which allowed us to assess whether listeners maintained the ability to label the unambiguous stimuli without feedback throughout the run). There were three single-object trials in each block of the experiment, with the attended object (either the stream of tones or the vowel) processed to have spatial cues consistent with a source from straight ahead (azimuth = 0°) and the target processed to have spatial cues consistent with a source from 0° or 45° (Figure 3-2).

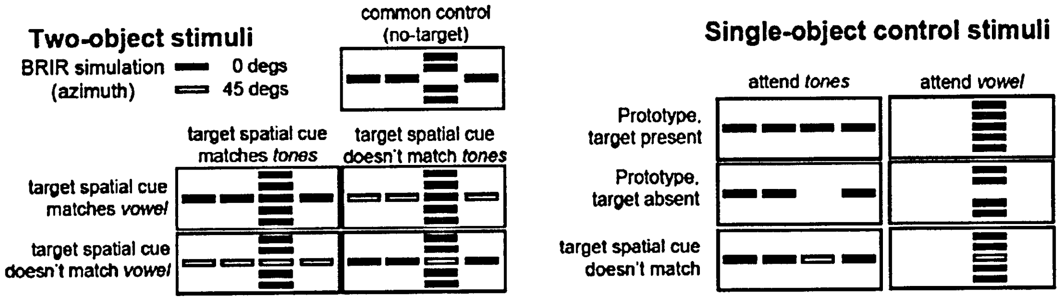

Two-object stimuli
BRIR simulation (azimuth)
0 degs
45 degs
common control (no-target)
target spatial cue matches *tones*
target spatial cue doesn't match *tones*
target spatial cue matches *vowel*
target spatial cue doesn't match *vowel*
Single-object control stimuli
attend *tones*
attend *vowel*
Prototype, target present
Prototype, target absent
target spatial cue doesn't match

Figure 3-2: The simulated spatial locations of the target and repeating tones varied across conditions. All two-object stimuli were common to both experimental blocks (rhythm and vowel identification). Control conditions included a two-object stimulus without a target presented and conditions in which only one object was presented.

In one block of the experiment, subjects reported the perceived rhythm of the two-object stimuli and the single-object tone stimuli. In a separate block, subjects reported the perceived vowel for the same two-object stimuli and one-object vowel stimuli. Both blocks consisted of 30 repetitions of each stimulus in random order, for a total of 240 trials per block. We used the response to the intermingled prototype stimuli both for screening and in interpreting the results to the ambiguous two-object stimuli, as discussed below.

3.3.7 Data analysis

We processed our data using the procedures described in our companion studies [56, 87]. Raw percent correct "target-present" responses ("even" for the tones, /ɛ/ for the vowel) were computed for each subject and condition. These results were then averaged across subjects to see overall trends (individual subject data were summarized well by the across-subject averages, so no individual results are shown here). The percentage of "target-present" responses to each stimulus condition for each subject was used to estimate the perceptual distance between the stimulus and the single-object target-absent prototypes. For each subject, we computed a normalized d' score, $\delta'_{condition:absent}$, defined as:

$$\delta'_{condition:absent} = \frac{d'_{condition:absent}}{d'_{present:absent}}, \qquad (3.1)$$

where $d'_{present:absent}$ is the standard psychophysical measure of the perceptual distance between the two prototypes in each of the experiment ("even"-"galloping" prototypes for tones and /ɛ/-/ɪ/ prototypes for vowel), and $d'_{condition:absent}$ is the perceptual distance between any stimulus and the target-absent, single-object controls (see [58]). These values were calculated individually for each subject as:

$$d'_{condition:absent} = \Phi^{-1}[\Pr(\text{"target present"}|\text{condition})] - \Phi^{-1}[\Pr(\text{"target present"}|\text{target absent})]. \qquad (3.2)$$

A value of $\delta'_{condition:absent} < 0.5$ indicates that the stimulus was perceived as more like the prototype with the target not present while $\delta'_{condition:absent} > 0.5$ indicates that responses were more like those for the target-present than for the target-absent prototype.

3.3.8 Mapping percent response to effective attenuation for each object

Single-object control experiments, described in detail in companion studies (see [56, 87]), were used to construct individual psychometric functions for each subject that related the physical intensity of the target in unambiguous, single-object conditions to the raw percentage of responses in the categorization tasks ("even" vs. "galloping", /ε/ vs. /ɪ/).

Briefly, in these control experiments, subjects were presented with a single object (tones in one experimental block, harmonic complex in the other) with a variable-level target. From trial to trial, the intensity of the target was attenuated by a randomly selected value between 0 dB and 14 dB (in 2 dB steps) relative to the level of the target in the two-object experiments. For both experiments, the percent response relating to the target attenuation of each subject was fit to a logistic function of the form:

$$\hat{y} = \frac{1}{1+e^{-a(x-x_0)}}, \qquad (3.3)$$

where $\hat{y}$ is the predicted percentage of "target-present" response, and the free parameters are: a, a slope parameter, and x_0, a threshold constant (50% of maximum). If 95% or more of a subject's responses to a given condition were either target-present (i.e., "even" or /ε/) or target-absent (i.e., "galloping" or /ɪ/), the effective attenuation was set to 0 dB or 16 dB, respectively. The corresponding psychometric functions for each subject were used to map the percent response in the two-object experiment onto the effective target attenuation in the two-object conditions.

3.4 Results

3.4.1 Subject screening

To ensure that subjects were able to accurately label the prototype stimuli during the two-object experiment, we excluded from all subsequent analysis data from any subject who failed to reach a criterion level of perceptual sensitivity to the prototypes when they were intermingled with ambiguous stimuli in the main, two-object experiment ($d'_{present:absent}$ > 1.0; see also [56, 87]). Two out of the nine subjects were unable to reliably label the vowel in the two-object experiment [i.e., $d'_{present:absent}$ (vowel) < 1.0].

For similar reasons, we also excluded any subject for whom the psychometric function relating response to the target attenuation had a very shallow slope or for whom the psychometric function did not fit responses well. Specifically, any subject for whom the slope parameter a (equation 3.3) was less than 10 percent / dB or the correlation coefficient (ρ) between the observed data (y) and the data fit ($\hat{y}$) was less than 0.9 was excluded. One out of the nine subjects was excluded based on these criteria.

Given the two screening criteria, all subsequent results are from six of the original nine subjects.

3.4.2 Rhythmic judgments (tone stream)

Figure 3-3 summarizes results of the main two-object experiment for both the rhythm judgments (top row; Figure 3-3A and B) and vowel identity (bottom row; Figure 3-3D and E, considered in the next section). Figure 3-3C and F reanalyze the raw results using the results of the single-object experiment and are considered in Section 3.4.4.

Figure 3-3A shows the across-subject mean and the standard error of the raw percentage "even" response to the tone stream. Subjects were generally accurate in identifying prototypes (accuracy: 87.78 ± 6.36% for "even" and 97.22 ± 2.18% for "galloping" prototypes), although this accuracy was reduced compared to results in a similar experiment using anechoic spatial simulation (see Appendix [56]). The spatial cues had a large effect on the rhythm judgments in the presence of the vowels, in line with previous studies [7, 56, 96]. When the simulated target location matched that of the tones, regardless of the vowel location (filled triangles and filled circles in Figure 3-3A), the target was perceived to be part of the rhythmic stream. When the target location matched neither that of the tones nor of the vowel, subjects still perceived the target as part of the tones sequence (open triangles in Figure 3-3A). However, when the target location matched that of the vowel but not the tones, the rhythmic stream was heard as "galloping" (open circles in Figure 3-3A), suggesting that the target was perceptually removed from the across-time stream. When the target was not presented (in the two-object no-target control condition), subjects generally heard the rhythm as "galloping" (exes in Figure 3-3A). In the one-object condition with the vowel absent, even though the spatial location of the target did not match that of the tones, subjects generally perceived an "even" rhythm (asterisks in Figure 3-3A).

Results in Figure 3-3B, which map the raw responses to relative perceptual distances from responses to the prototype stimuli, show the same trends as the raw response results. The rhythm is generally heard as "even," except when spatial cues of the target match those of the vowel and not the tones and for the target-absent, two-object stimulus.

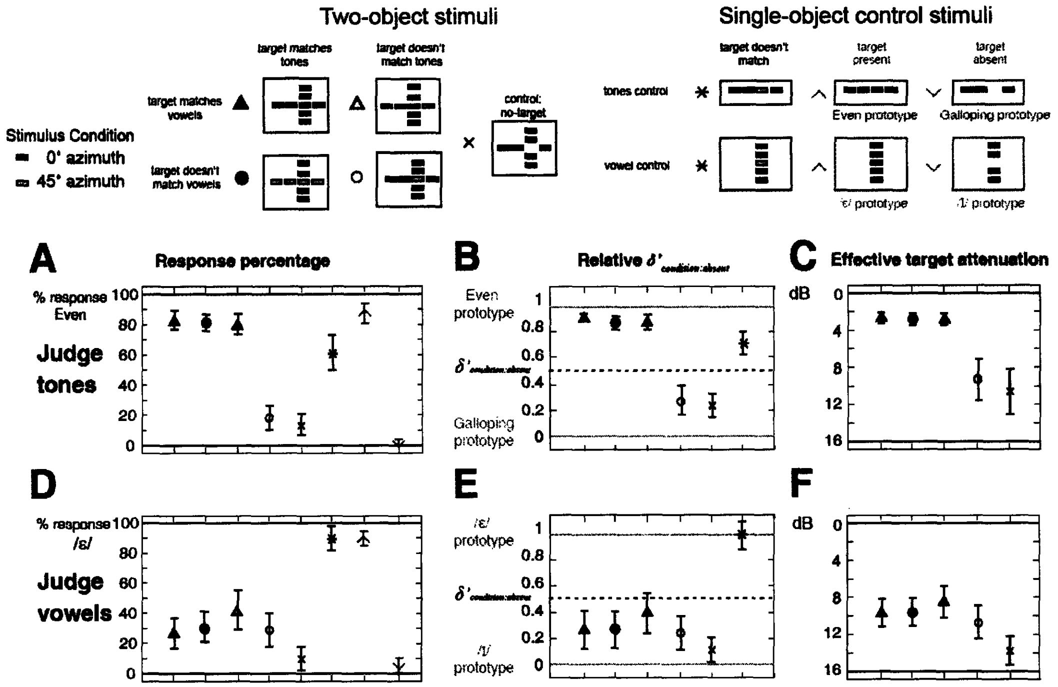


Figure 3-3: Results for tones (top row) and vowel (bottom row). **(A)** and **(D)** Raw percentage of "target present" responses. **(B)** and **(E)** Relative sensitivity (δ'; 1.0 indicates raw results identical to those with the target present; 0.0 indicates raw results identical to those with the target absent). **(C)** and **(F)** Effective target attenuation, based on results from the single-object control experiment (see Figure 3-5 for an illustration of how these results are derived). Each marker represents the across-subject mean and the error bar shows ±1 standard error of the mean.

3.4.3 Vowel judgments (vowel stream)

Figure 3-3D shows the across-subject mean and the standard error of the raw response percentages for the vowel judgments. There was a non-zero likelihood of subjects responding /ɛ/ when presented with an /ɪ/ prototype; similarly, subjects sometimes responded /ɪ/ when an /ɛ/ was presented (accuracy: 89.44 ± 4.92% for /ɛ/ and 94.46 ± 8.07% for /ɪ/ prototypes). Unlike in the rhythmic judgment, spatial cues had only a weak effect on the perceived identity of the vowel in the two-object mixtures. Moreover, as in the companion study using anechoic spatial cues (Appendix [56]), listeners were more likely to respond that the vowel in the two-object conditions was /ɪ/ rather than /ɛ/ for all spatial configurations. In the one-object condition in which only the vowel was present and the target location did not match that of the vowel, subjects nonetheless responded as if the target was part of the vowel, responding /ɛ/ roughly 90% of the time.

Replotting the data in terms of the relative perceptual distance to the prototypes (Figure 3-3E) shows similar patterns. In all two-object configurations, regardless of spatial cues, responses were more like /ɪ/ (target not present in the vowel) than /ɛ/ (target present).

3.4.4 Mapping raw responses to equivalent attenuations

For all subjects who passed our screening, presenting an unambiguous single-object stimulus with different target intensities produced a well-behaved psychometric function. An example of these functions is shown in Figure 3-4 for one subject (S18) for both tones and vowel. To the left of each psychometric function, the same subject's raw percent responses "even" (Figure 3-4A) and /ɪ/ (Figure 3-4B) are plotted. These response percentages can be mapped to the equivalent target attenuations, as illustrated by the dashed lines in the figure.

For each subject and condition, raw results from the two-object experiments were mapped to equivalent target attenuations. These mapped values were then averaged across subjects to produce the plots in Figure 3C (tones) and 3F (vowel). These results, in turn, allow us to quantify the degree of energy trading of the target that occurs for two-object stimuli.

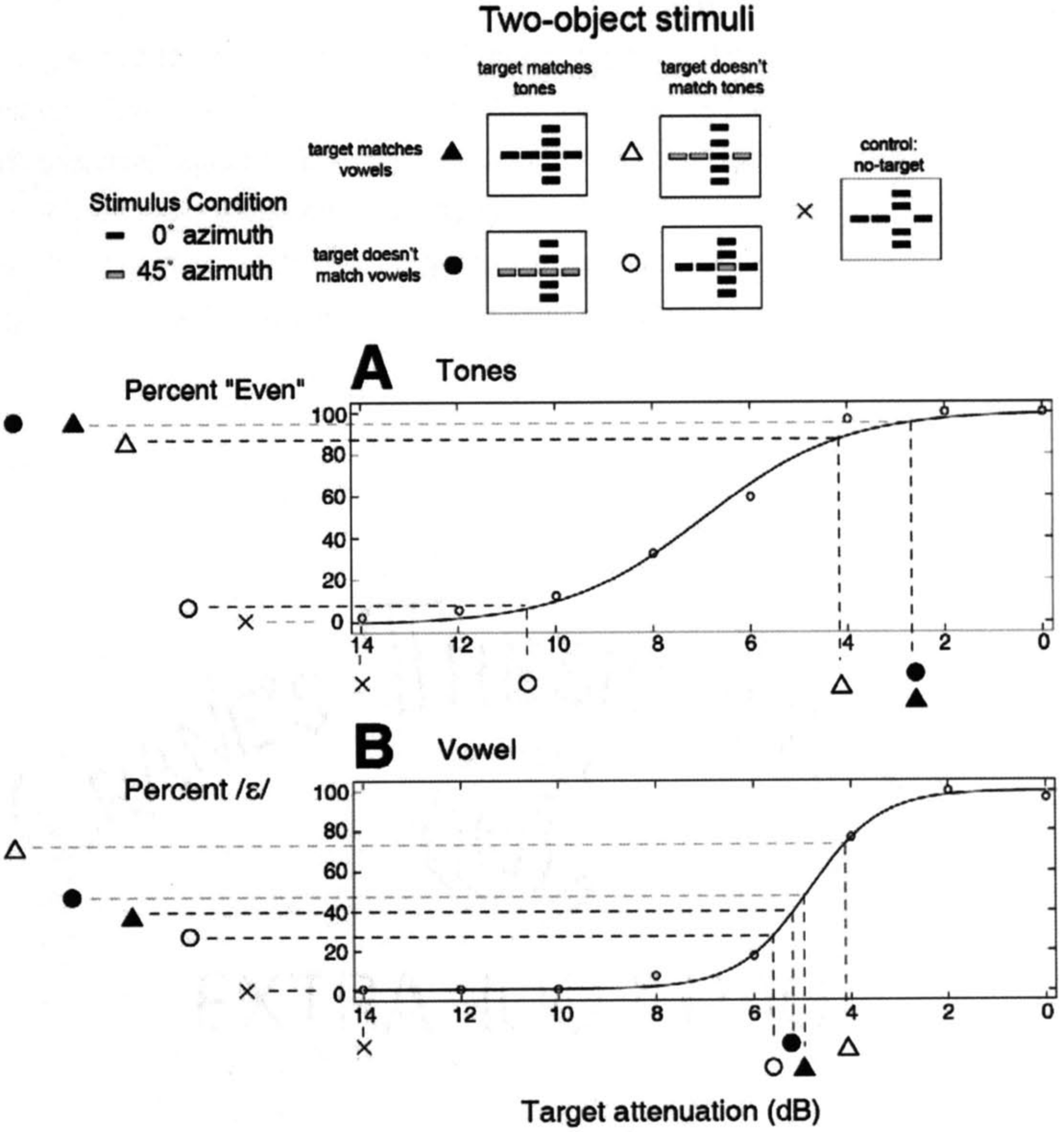


Figure 3-4: Illustration of how the results from the single-object control experiment results are used to map a raw category response for the tones **(A)** and vowel **(B)** to an effective target attenuation in the two-object control conditions. The psychometric functions in the right of each panel show the raw results (circles) and the fit psychometric function (solid line) for subject S18 in the corresponding single-object experiments, plotted as a function of the physical attenuation of the target. The symbols to the left of the panel show the raw percent category responses for this subject in the main two-object experiment. Dashed lines map these raw categorization responses to the equivalent target attenuations.

3.4.5 Target intensity trading

Figure 3-5A plots the across-subject mean effective attenuation of the target in the tone stream against the mean attenuation of the target in the vowel. The plot shows all conditions that were common to the two experiments, including the two-object, target-absent control. The solid curve in the figure plots the trading relationship that would be observed if energy conservation holds. The dotted curve in the figure shows where data would fall if amplitude, rather than energy, traded between objects (see [53, 59]).

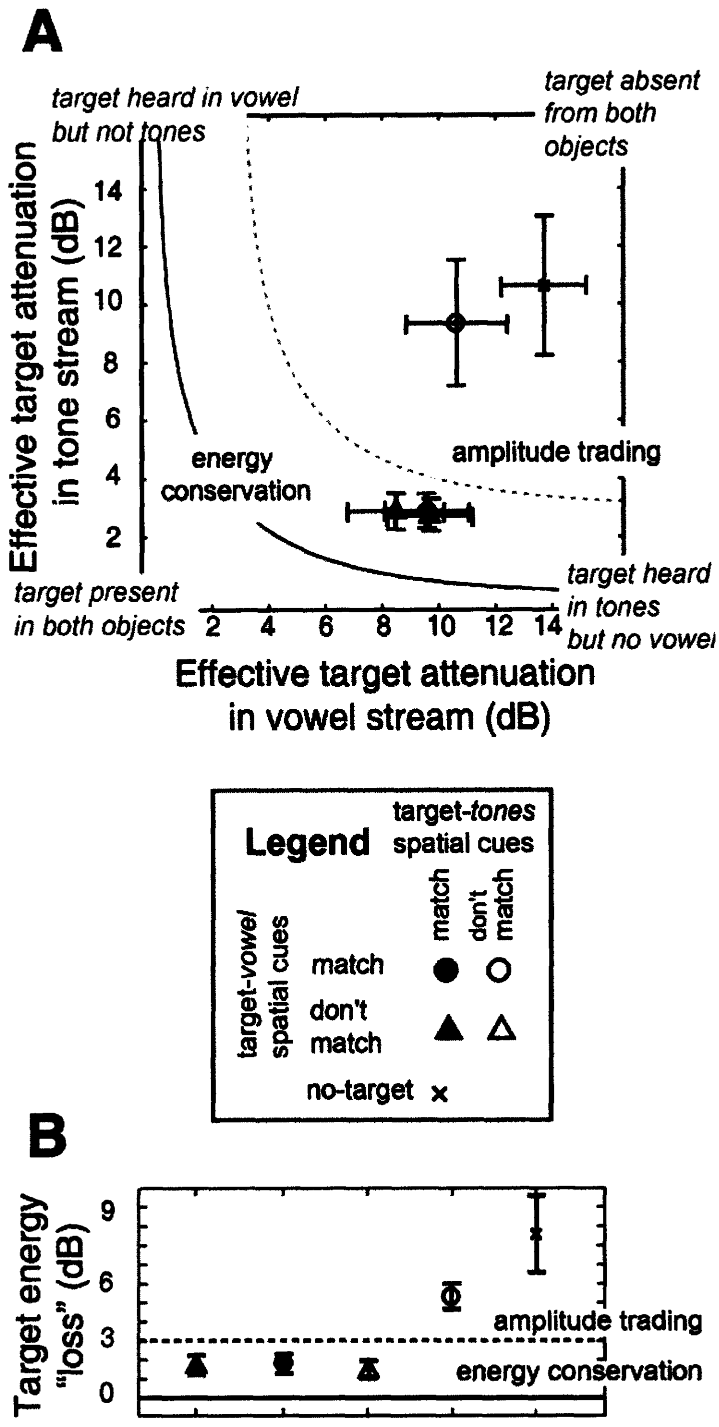


Figure 3-5 (A): Scatter plot of spatial cues influencing target level attenuation trading between tone and vowel streams. Solid line denotes the locus of a possible energy trading relation between the two streams. Dotted line denotes a 3 dB total target loss. **(B)**: Perceptual target energy not accounted for after the summation of the perceived energy of the target from the two streams.

As expected, results for the target-absent control fall near the upper-right corner of the plot, indicating that the perceived qualities of the tone and vowel in the target-absent, two-object stimulus produced percepts with a very weak target (ex in Figure 3-5A). When the spatial location of the target matched that of the vowel but not the tones (open circle), the effective level of the target was attenuated by an average of 9 dB or more both when the listeners attended to the tones and when they attended to the vowel. In all of the other two-object spatial configurations, the target was generally perceived as part of the tone stream and not the vowel.

We further quantified the "trading relationship" by computing the total effective energy of the target, summing its effective energy when attending the tones and its effective energy when attending the vowel, for each condition. The across-subject means for the "lost" target energy in each condition was then found by subtracting the total effective target energy from the physical energy of the target. These values are shown in Figure 3-5B.

Consistent with past studies, energy conservation is not found in any two-object stimulus. Instead, the total target energy in the two objects sums to less than the total energy, producing positive values for the "lost" target energy in Figure 3-5B. Results for three of the four ambiguous stimuli are consistent with past results in finding a loss of target energy between 0-3 dB [53, 59, 87]. However, when the spatial cues of the target match those of the vowel but not those of the tones, the lost energy is significantly larger than in the other conditions (more than 5 dB, on average).

3.5 Discussion

3.5.1 Comparison with anechoic results

We expected reverberant energy to reduce the saliency of spatial cues and therefore to reduce the effects of spatial cues on perceptual organization, compared to our companion study using anechoic cues (Appendix [56]).

Spatial cues caused changes in perceptual organization that were qualitatively similar to our previous results. However, consistent with our hypothesis, the magnitude of the effects of spatial cues on perceptual organization was smaller. In particular, in our anechoic study, there were larger shifts in the perceptual organization with changes in spatial configuration than in the current study. For instance, the equivalent target attenuations in two-object conditions varied from 6-10 dB for vowel judgments and from 1-13 dB for tone judgments. Adding reverberation to the stimuli in the current experiment reduced these ranges to roughly 8-10 dB (vowels) and 3-10 dB (tones).

Despite the fact that reverberant energy reduced the influence of spatial cues, the current results find evidence for a "lost" target, just as in our previous study (Appendix [56]). In particular, when the target spatial cues match those of the vowel but not the tones, the effective target attenuation is large for both the vowel and tones. While the same spatial configuration produced effective attenuations that were comparable for the vowel in anechoic condition (roughly 9 dB), the effective attenuation for the tone condition was larger in anechoic space (~13 dB) than here (~10 dB). Thus, our current results show that even in reverberation, spatial cues play an important role in perceptual organization, even though they exert less influence on perceptual organization in reverberant than in anechoic conditions.

Reverberation not only affected the strength of spatial cues on object formation, but also altered the reliability of judgments about tone rhythm. In anechoic conditions, most subjects performed with 100% accuracy when identifying the prototype rhythms of single-object tone stimuli. In the current study, accuracy was reduced, with overall percent correct around 92%. This result undoubtedly reflects the fact that the reverberant energy not only reduces the reliability of spatial cues, but smears out the spectro-temporal content of the stimuli. Thus, in the current reverberant stimuli, the energy from the proceeding tones is extended temporally into the nominal gaps in between the tones. As a result, even in the absence of the target, there will be some residual energy at 500 Hz during the time that the target might be present, whether or not the target is part of the stimulus. This causes a small but noticeable degradation in the ability to label the tone rhythms for the prototype, single-object tone streams. This temporal smearing caused listeners to be more likely to report a galloping prototype as even (~12% of the time in the current study, compared to < 1% of the time in the anechoic study; Appendix [56]).

The uncertainty about spectro-temporal structure of the control stimuli caused by reverberation was even more pronounced from the "no-target" two-object stimulus. In the anechoic study, this control almost always produced "galloping" responses, while in the current reverberant study, the control was heard as "even" on about 10 % of all trials. As a result, the effective attenuation of target perceived in the tone object for the no-target control was only 11 dB in the current study, whereas it was near 14 dB in anechoic conditions (Appendix [56]).

3.5.2 Conceptual model for perceptual organization in a multi-source reverberant environment

Recently, there have been attempts to develop a scene-analysis model to account for the perceived intensity (Chapter 2) and the perceived location (Chapter 4) of spectro-temporal contents in a sound mixture. The task of perceptual organization is based on the spectro-temporal signal that reaches our ears, $R^L(f,t)$ and $R^R(f,t)$, with two parameters, f and t, representing frequency and time, respectively. The spectro-temporal content for each source S_i can be described by the following expression:

$$S_i(f,t) = A_i(f,t)e^{j(2\pi f\,t+\psi_i(f,t))}, \qquad (3.4)$$

where $A_i(f,t)$ and $\psi_i(f,t)$ represent the amplitude and the phase variation in both frequency and time associated with source i, respectively. If the frequency components of each source contain different spatial attributes and that each source can also move in time on the azimuthal plane relative to the observer, the spatial attributes associated with source i can also be summarized by a parameter $\theta_i(f,t)$.

Since the receivers are spatially displaced and may experience different acoustical filtering, the signals at the two receivers can be expressed as:

$$R^L(f,t)=\sum_{i=1}^{N} DTF^L_{\theta_i(f,t)}\{S_i(f,t)\}=\sum_{i=1}^{N} A_i^L(f,t)e^{j\left(2\pi f\,t+\psi_i^L(f,t)\right)}; \quad (3.5)$$

$$R^R(f,t)=\sum_{i=1}^{N} DTF^R_{\theta_i(f,t)}\{S_i(f,t)\}=\sum_{i=1}^{N} A_i^R(f,t)e^{j\left(2\pi f\,t+\psi_i^R(f,t)\right)}, \quad (3.6)$$

where there are N sources in the scene and the signals arriving at each receiver have been filtered by the directional transfer functions (DTFs, which are receiver and angle dependent[5]), and subsequent filtering also makes both $A_i(f,t)$ and $\psi_i(f,t)$ receiver dependent, i.e., $A_i^{L/R}(f,t)$, $\psi_i^{L/R}(f,t)$.

In order to understand and process the sources in the environment, the listener must combine the information at the two receivers, $R^L(f,t)$ and $R^R(f,t)$, to extract the spatial information exploiting the interaural delay, interaural intensity and spectral cue differences between the two receivers. Furthermore, in order to understand and locate the sources in the environment, the listener must decompose $R^L(f,t)$ and $R^R(f,t)$ to try to recover the content and the location of these original sources. Let us for simplicity assume henceforth that the monaural and binaural systems can extract the spectro-temporal as well as the spatial information of the whole scene perfectly, i.e.

$$\hat{S}(f,t)=S(f,t)=\sum_{i=1}^{N} S_i(f,t); \qquad (3.7)$$

[5] Without loss of generality, we assume all auditory sources distributed spatially only along on the azimuthal plane.

$$\hat{\theta}(f,t)=\theta(f,t)=\sum_{i=1}^{N}\theta_i(f,t). \qquad (3.8)$$

The process of decomposing the whole scene to its constituent sources may be imperfect. The resulting estimates of the spectro-temporal content for i^{th} source can be written as $\hat{S}_i(f,t)$, while their spatial content estimates can be expressed as $\hat{\theta}_i(f,t)$. Perfect scene analysis would correspond to perfect recovery of the spectro-temporal content for each source:

$$\hat{S}_i(f,t)=S_i(f,t) \quad \forall i\ , \qquad (3.9)$$

and perfect recovery of the location estimate associated with each source would correspond to:

$$\hat{\theta}_i(f,t)=\theta_i(f,t) \quad \forall i\,. \qquad (3.10)$$

3.5.3 Failure in the intensity trading hypothesis

Chapter 2 introduced an operator $\mathcal{D}(\cdot)$ to describe the scene analysis process. It was argued that if veridical parsing of the scene occurred, the perceived intensity of sound energy in each object must sum to equal the true energy in the original source at any given frequency f. However, the validity of this trading relationship is expected to fail if the decomposing process, $\mathcal{D}(\cdot)$, is dependent on the object being attended to, i.e., $\mathcal{D}_{TDA|S_i}(\cdot)$. By manipulating the frequency of the repeating tones, we found a trading relationship that governs how an ambiguous target tone is allocated between two competing target. Therefore, they cannot reject a scene analysis model that is only based on a bottom-up process.

Our data, in support of previous experiment (Appendix [56]), show that the intensity trading hypothesis does not hold (Figure 3-5B). In all four two-object conditions, the spectro-temporal contents were identical and they only differ by their spatial configurations. Bottom-up influences, such as peripheral adaptation, cannot adequately explain the substantial loss of target energy in only one of the four conditions since all the stimuli would have caused the same amount of adaptation. Therefore, extending the decomposing process proposed in Chapter 2 to include binaural processing, Figure 3-6A shows a realization of a non-linear process that implicates the influence of top-down attention. Such top-down influences in resolving perceptual competition have long been established in the vision literature [97, 98].

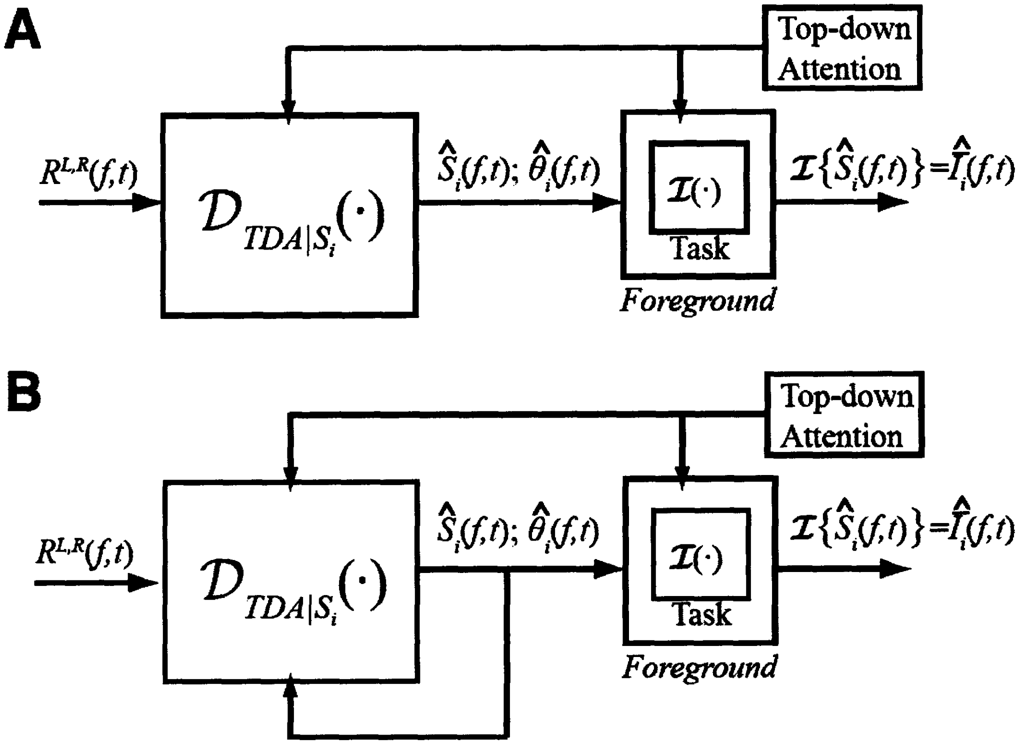

A
Top-down Attention
$R^{L,R}(f,t)$
$\mathcal{D}_{TDA|S_i}(\cdot)$
$\hat{S}_i(f,t);\ \hat{\theta}_i(f,t)$
$\mathcal{I}(\cdot)$
Task
Foreground
$\mathcal{I}\{\hat{S}_i(f,t)\}=\hat{I}_i(f,t)$
B
Top-down Attention
$R^{L,R}(f,t)$
$\mathcal{D}_{TDA|S_i}(\cdot)$
$\hat{S}_i(f,t);\ \hat{\theta}_i(f,t)$
$\mathcal{I}(\cdot)$
Task
Foreground
$\mathcal{I}\{\hat{S}_i(f,t)\}=\hat{I}_i(f,t)$

Figure 3-6 (A): A realization of the decomposing function that is attention dependent. **(B)**: A realization of the decomposing function that is attention dependent feeding the spatial cues for each object back as a closed-loop system.

The ability to utilize auditory spatial cues for perceptual organization poses an interesting paradox [12]. In previous psychoacoustical studies, it has been established that ITD cues can be used for sequential grouping [21]. Darwin and Hukin [7] showed that listeners use ITD as a cue when parsing ambiguous mixtures of sounds across time. However, sequential grouping by ITD cues is less dominant than frequency proximity cues when multiple sources are present

[36]. Furthermore, despite evidence that spatial cues can influence simultaneous grouping [37], ITD cues are relatively ineffective for such across-frequency grouping [14, 52].

Given that spatial cues have a strong influence in across-time but not across-frequency grouping, a possible explanation for the failure in the trading relationship is that the decomposing function $\mathcal{D}_{TDA|S_i}(\cdot)$ only takes into account the spatial information of the scene for rhythmic identification task (i.e. across-time grouping) but not for the vowel identification task (i.e. across-frequency grouping). Interpreted this way, it also suggests that the auditory system favors efficient processing over veridical parsing of the scene (Appendix [56]). The process uses different strategies in order to disambiguate signal that is the most pertinent, and weight evidence differently according to the object attended.

3.5.4 Object formation influencing spatial processing

Even though the interaural differences caused by a pure tone (like that of our repeating tones) converge to constant values over time in a reverberant environment, a broadband signal (like that of our vowel) contains fluctuations in the short-term spectrum [92, 94]. Therefore distortion from reverberant energy, especially for a substantial distance between the source and the receiver, can cause quite severe interaural decorrelation that can potentially influence our abilities to organize the auditory scene. However, despite the frequency-to-frequency fluctuation in ITD, integrating binaural information across frequency can give a reliable ITD estimate when the range of the candidate ITD values is limited to physiological range[6].

[6] In our pilot study, a programming error accidentally caused the manipulation of the ITD cues to be 3,000 μs instead of the intended ITD (300 μs) that is within physiological range. The experimental setup was identical to this experiment, except the spatial configurations of stimuli were solely manipulated by ITD cues. Despite the ITD cue falling outside of physiological range, the results obtained were comparable to all that have been reported in this paper.

For laterality judgments of broadband sounds in some reverberant settings, it might be especially beneficial to integrate binaural information across frequency [94, 99, 100]. Many studies also suggest that there is an obligatory combination of binaural information across frequencies for binaural parameters discriminations ([101-103]. However, recent studies ([13], Chapter 4) suggest that localization of auditory sources is intimately linked with the perceptual organization of the auditory scene. In the current experiment, the spatial location of the target has a big impact on the perceptual organization of the tones across time (i.e., a "galloping" percept is only heard when the target location matches that of the vowel and not the tones).

Introspection also suggests that the rhythmic percept takes a finite amount of time to build-up to a stable identity (from "even" to "galloping") in the two-object condition with the target matching the location of the vowel but not the tones. However, by the end of the three-second presentation of the stimulus, the build-up phase has completed and the percept converged to a stable rhythmic identity. The change in percept over time such as the build-up of streaming [24-27] and the build-up of echo suppression [104, 105] have been well documented. However, how these two effects are interlinked have not been extensively studied and clearly more research is needed. Kidd *et al.* [106] and Freyman and Keen [107] postulated that listeners may dynamically map the reflected sounds in space in order to learn the characteristics of the reverberant environments. Figure 3-6B shows a possible realization of such a decomposing system that combine spatial information selectively depending on object grouping to generate location estimates and adaptively using this information for perceptual organization computation.

3.6 Conclusions

Results are consistent with the findings by Shinn-Cunningham *et al.* (Appendix [56]) that by manipulating the spatial attributes of spectro-temporally identical stimuli, the assignment of an ambiguous element does not always follow a trading relationship. In one condition, the target energy unaccounted for was comparable to a control condition in which the target was not physically present. This observation strongly lends support to a scene analysis system that changes its "strategies" according to the object being attended.

CHAPTER 4 LOCALIZATION DUE TO STREAMING

The work described in this chapter is currently in preparation for journal submission.

4.1 Abstract

The perceived location of an auditory object can be influenced by the binaural information of components that are spectrally remote and / or temporally incoherent to the spectro-temporal elements of interest. Integration (*pulling*) and repulsion (*pushing*) effects are often described in different localization experiments under different contexts, e.g. binaural interference, precedence effect. In this paper, we present an experiment that investigated how auditory grouping influences object localization. Results are consistent with the hypothesis that the integration of binaural information, either across time or across frequency, is observed when the spectro-temporal elements are grouped as one object, while repulsion effects are observed between two separate streams. A generalized model is proposed, not only to account for the data obtained in the present experiment, but also to describe many seemingly disparate observations in localization experiments under the framework of auditory scene analysis.

4.2 Introduction

In everyday life, we constantly analyze sounds arriving at our ears that originate from multiple acoustical events in the environment and are spatially displaced from each other. The task of auditory scene analysis is to segregate the sound-energy belonging to an auditory event of interest from the cacophony of sounds accompanying it [3]. While it is important to be able to understand the spectro-temporal content of the signal of interest (i.e., 'what' you are listening to), the spatial information associated with the source is also an important attribute for scene analysis (i.e., 'where' the auditory event comes from). For example, in a cocktail party, not only are you able to hear out your name when multiple speech sources are present [1], you are also interested in the location of the person calling for your attention.

Many studies have investigated the influence of spatial cues in facilitating sound source segregation and increasing the intelligibility of an attended signal [7, 14]. However, little is known on how auditory scene analysis influences the localization of auditory events in the environment. Although simultaneous sounds can generally be localized quite accurately [108-110], there is a vast body of literature describing how the sensitivity to binaural parameters may be degraded by the presence of simultaneous energy, even if they are in remote spectral regions [13, 111].

Gardner considered several localization phenomena which arise as a result of interaction between sources, e.g.,: 1) "fusion of spatially separated signals" into an apparent single image located at the position of the earlier signal; 2) "displacement of the resulting image" to a position that does not coincide, in general, with the location of either of the real sources [112]. The first phenomenon described has since evolved to an extensive branch of research known as the precedence effect (c.f. [113] for a comprehensive review). The second phenomenon can be divided into two observations: i) *pulling* of binaural information across frequency into one fused image and as a result the spatial displacement between sources is reduced compared to the perceived location of each source otherwise presented separately (also known as 'integration' or 'attraction'); ii) *pushing* effect against two auditory sources such that the displacement between sources is increased (also known as 'repulsion'). It is interesting to note that if the *pulling* of binaural information is not restricted to be across frequency and includes the temporal integration of binaural information [114-117], Gardner's first description for localization phenomenon can be subsumed into the second.

Butler *et al.* described a *pulling* effect when subjects were asked to localize a source in the presence of an interference stimulus delivered monaurally [118].

However, a more general description of the *pulling* effect is often linked to discriminability of different binaural cues across frequency rather than the perceived location of auditory sources *per se* and is commonly known as "binaural interference," which was first formally described by McFadden and Pasanen [111]. This obligatory combination of binaural information from spectrally remote region generally causes degradation in discrimination of the binaural parameters [101-103, 119-121] but it was also later discovered to influence the perceived location of the attended spectro-temporal elements [13, 122, 123].

Best *et al.* [13] reviewed many of the aforementioned studies and rekindled the idea that binaural interference is a by-product of grouping processes (see [122]). In this idea, *pulling* effects are likely to come from combining information that is perceived as the same auditory object. They also showed that sequential streaming can play a strong role in how spatial information is combined across frequency. While these and other authors have discussed binaural interference and more generally binaural perception [124] in terms of object formation, there has yet been a fully developed scene-analysis-based model to account for perceived location of objects in a mixture.

The *pushing* effect observed in many experiments [110, 125] is not as well documented compared to the different manifestations of the *pulling* effect. Best *et al.* [108] postulated that *pushing* effects observed with concurrent sounds might be the results of the formation of two auditory objects. However, to the best of our knowledge, there have not been any localization studies to characterize the *pushing* effect.

In the current study, we constructed a stimulus that contained an ambiguous spectro-temporal element (the target) that could logically belong to an isochronous stream of repeating tones or to a more slowly repeating harmonic complex. The spatial attributes of the repeating tones, the harmonic complex and the target were systematically varied. In different experimental blocks using the same stimuli, subjects were asked to make localization judgments on the repeating tones or the complex. This enabled us to observe the effect of streaming on object localization. We also assessed the degree to which the *pulling* and *pushing* effects are influenced by manipulating the inter-stimulus rate, with the hypothesis that the across-object interactions would decrease as the temporal separation between objects increases.

4.3 Experiment

We used a novel two-object stimulus, which consisted of a stream of two repeating tones (S) and a harmonic complex (C) that repeated one-third the rate. A target tone (T) could logically belong to the stream of tones or the harmonic complex. We manipulated the spatial content of the repeating tone stream, the complex and the target to explore how spatial cues influence the localization of the perceived objects. In a series of companion studies (Chapter 2, Chapter 3 and Appendix [56]), we used the same two-object stimulus structure to investigate how spatial cues influence the identity of the perceived objects. We hypothesized that if the target is perceptually grouped with the stream, the perceived location of the stream would be influenced by the spatial cues of the target independent of the stimulus rate. We further hypothesized that there would be an interaction between the perceived location and the inter-stimulus repetition rate when multiple objects are present in the auditory scene.

4.3.1 Methods

4.3.1.1 Stimuli

Stimuli generally consisted of a repeating sequence of a pair of tones followed by a harmonic complex (Figure 4-1A). The frequency of the pair of tones was 500 Hz and each tone was 60 ms in duration, gated with a Blackman window of the same length. The harmonic complex was filtered with the same formant filter described in Chapter 2. Briefly, the first formant frequency (F1) was set to 490 Hz, with the second and the third formants fixed at 2100 and 2900 Hz respectively, similar to the parameters used by [52]. Each harmonic was also 60 ms in duration, gated by the same Blackman window used for the repeating tones.

The target was a 500 Hz tone that had the same onset / offset as the harmonic complex (60 ms in duration, gated with a 60-ms-Blackman window). As a result of this structure, the target could be heard as a third tone in the repeating tone stream or as the fourth harmonic in the harmonic complex (Figure 4-1B).

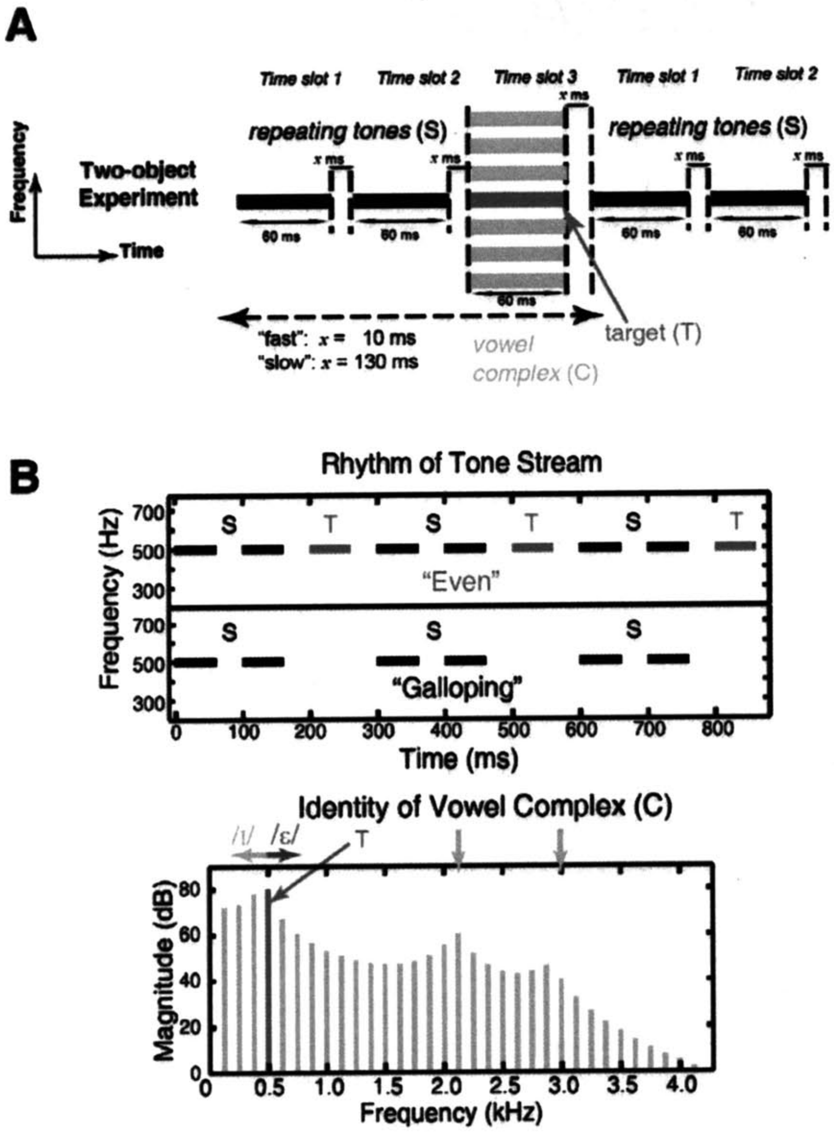


Figure 4-1 (A): The two-object stimulus consists of a three-part sequence: a pair of pure tones followed by a harmonic complex (in the form of AAB). In the basic configuration, the pure tones in time slots 1 and 2 (S) are at 500 Hz. Time slot 3 is made up of two components: a target tone at 500 Hz (T) and a tone complex (C) with a F0 of 125 Hz (with the fourth harmonic 500 Hz omitted). This tone complex is shaped with a synthetic vowel spectral envelope to make it sound like a short vowel. The stimulus has two repetition rates controlled by the silent gap of x ms ("fast"; x = 10 ms, "slow"; x = 130 ms). **(B)** (top panel): The perceived rhythm depends whether the target (T) is grouped with the repeating tones (S). The resulting rhythmic perception would be "even" if the target (T) is grouped with the repeating tones (S), otherwise a "galloping" rhythm would be perceived. (bottom panel): The synthetic vowel spectral envelope was similar to the shape used by Hukin and Darwin (1995a). The perceived vowel shifts from /ɛ/ if the target (T) is grouped with the complex (C) to /I/ if the target (T) is not grouped with the complex (C). The arrows indicate the approximate locations of the first three formants of the perceived vowel.

The target and the tones were also filtered by the same formant filter such that the amplitude of the target and the tones would match the formant envelope of the harmonic complex. There was a silent gap between each tone and the harmonic complex and its duration controlled the inter-stimulus-time interval. In the "fast" block of the experiment, the silent gap was 10 ms and thus an acoustic event occurred every 70 ms. In the "slow" block of the experiment, the silent gap was set at 130 ms, with the effective inter-stimulus-time interval being 190 ms. This basic pattern, a pair of repeating tones followed by the vowel complex and target, was repeated to produce a stimulus that was perceived as two streams: an ongoing stream of tones and a repeating vowel occurring at a rate one-third as rapid.

In order to reduce build-up of streaming, which might affect perceptual grouping [24-27], we kept the presentation time of these stimuli to three seconds. In the "fast" condition, the three-second stimuli consisted of 14 repetitions of the repeating-tones-complex triplet, while in the "slow" condition it consisted of five repetitions.

4.3.1.2 Acoustic pointer

The acoustic pointer used in this experiment was a 200-Hz-wide band of noise centered at 2 kHz. Subjects had control of the interaural intensity difference (IID) of the pointer by means of two buttons (right or left) and thereby could change its perceived location along the intracranial axis.

4.3.1.3 Task

In order to assess the perceived location of the attended stream and how it interacts with the alternate stream, the same physical stimuli were presented in two experimental blocks. In one block, subjects matched the perceived location of the repeated tones with the acoustic pointer. In the other block, the same physical stimuli were presented in a different random order, and subjects matched the perceived location of the harmonic complex.

The stimuli were presented at two repetition rates ("fast" or "slow"). Therefore, in total, there were four experimental blocks (localization of the repeating tones or the complex at two different repetition rates). Each subject completed the four blocks on two separate days. On any given day, a subject completed one block of complex localization and one block of repeating tones localization presented at different inter-stimulus time rates ("fast" or "slow"). The order of the stimulus rate and the order of the task were all counter-balanced across subjects.

4.3.1.4 Environment

All stimuli were generated offline using MATLAB software (Mathworks Inc.). Signals were processed with pseudo-anechoic head-related transfer functions (HRTFs) measured on a KEMAR manikin at a distance of 1 m in the horizontal plane (see [94] for details). Sources were processed to have spatial cues consistent with a source either from a position straight ahead (0° azimuth), or 45° to the left or right of the listener.

Digital stimuli were generated at a sampling rate of 25 kHz and sent to the Tucker-Davis Technologies hardware for D/A conversion and attenuation before presentation over headphones. Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. A randomized roving attenuation level between 0 and 14 dB was applied to the stimulus and the acoustic target in each trial before presentation in order to reduce the possibility of presentation level influencing the localization tasks. Subjects were seated in a sound-treated booth and responded via a button-box (TDT Bbox) which was directly connected to the hardware. Stimuli were presented over insertion headphones (Etymotic ER-1). All signals were presented at a listener controlled, comfortable level that had a maximum value of 80 dB SPL.

4.3.2 Experimental procedure

4.3.2.1 Participants

Nine subjects (four male, five female, aged 18-31) took part in the experiment. All participants had pure-tone thresholds of 20 dB HL or better in the range from 250 to 8000 Hz, in both ears, and their threshold at 500 Hz was 15 dB HL or better. All subjects gave informed consent to participate in the study, as overseen by the Boston University Charles River Campus Institutional Review Board and the Committee On the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology.

4.3.2.2 *Training*

At the beginning of each experimental block, all listeners received fifteen minutes of practice to familiarize with the equipment setups. During each practice session, subjects were encouraged to explore the full range of the acoustic pointer, and diagrams were presented on screen to help emphasize the difference between the repeating tones and the harmonic complexes. Feedback was not provided during the practice session or in the main experiment.

4.3.2.3 *Perceived location matching procedure*

Each trial began with presentation of a three-second stimulus (listening phase). This was followed by a three-second matching phase in which subjects had control of the IID of the acoustic pointer. At each clock-cycle during the matching phase, button presses were acquired. If the right button was pressed, the intensity of the right headphone increased by a constant Δ dB step and the intensity of the left headphone decreased by the same amount, resulting in an IID change of 2Δ dB. Button presses were sampled at 25 kHz and the constant step-size was set to be very small ($|\Delta| = 1.5 \times 10^{-3}$ dB) such that listeners essentially heard a punctate image traversing continuously along the intracranial axis. At the conclusion of the three-second matching phase, the three-second stimulus was presented again and the buttons became inactive. When the matching phase restarted, the pointer started with the last IID value, and the listeners regained control of the pointer. Alternation between the listening and the matching phase repeated until the subject was satisfied that the pointer matched the perceived location of the attended stream, indicated by pressing a third button.

The initial IID of the pointer was assigned randomly from the range of +20 dB (full right) to -20 dB (full left) to discourage a counting strategy. Listeners positioned their headphones only once per session. Typically subjects cycled through three to four iterations of the listening-matching sequence for each trial, and each session lasted no longer than 1.5 hours.

4.3.2.4 Spatial configuration of stimuli

Each block of the experiment consisted of seven single-object conditions. In the repeating tones localization block, there were three single-object (control) conditions consisting of the repeating tones and the target coming from the same location, i.e., 0° (S_0T_0), +45° (S_RT_R) and -45° (S_LT_L) in azimuth, and four single-object conditions consisting of the stream of tones and the target originating from different location (i.e., S_0T_R, S_0T_L, S_RT_0 and S_LT_0, see Figure 4-2, top). In the harmonic complex localization block, there were three (control) single-object conditions consisting of the target and the complex being co-located at 0° (T_0C_0), +45° (T_RC_R) and -45° (T_LC_L) azimuth, and four single-object conditions consisting of the target and the complex originating from different locations (T_RC_0, T_LC_0, T_0C_R and T_0C_L, see Figure 4-2, bottom).

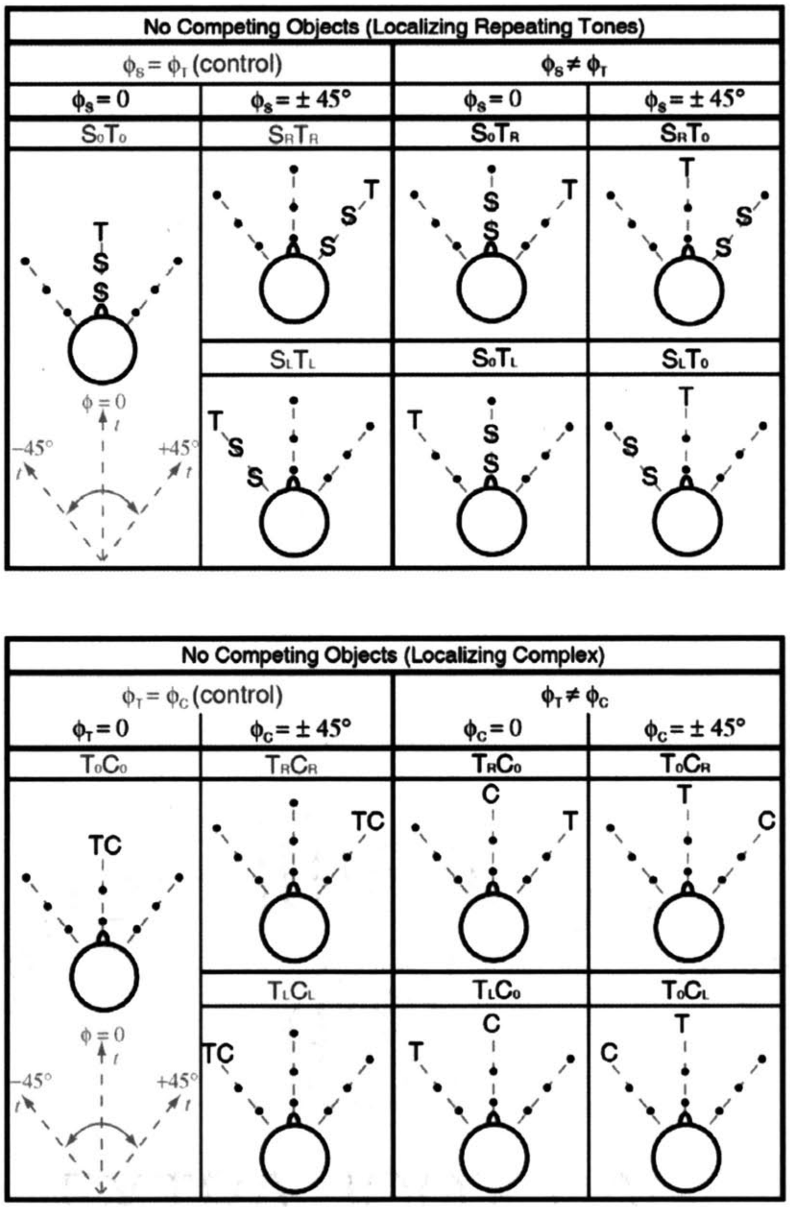


Figure 4-2: Spatial configuration for all the single-object conditions tested with no competing objects. (Top panel): seven conditions consisted of only the repeating tones and the target tone. The three control conditions: S_0T_0, S_RT_R, S_LT_L have consistent spatial information and are shown on the left half of this panel. Since the repeating tones and the target are temporally disjoint, there are no conflicting spatial cues overlapped in time. (Bottom panel): seven conditions consisted of only the target and the complex which are gated simultaneously. The three control conditions: T_0C_0, T_RC_R, T_LC_L are spectrally coherent in their spatial cues and are shown on the left half of this panel.

Intermixed with the seven single-object conditions in each block of the experiment were seven two-object conditions that were common across experimental blocks (Figure 4-3). In these conditions, the complex always originated from 0° azimuth. In one condition, the target and the repeating tones were co-located with the complex ($S_0T_0C_0$). This served as our two-object control condition. The other six conditions were made up of two conditions with only the target coming from the side ($S_0T_RC_0$ / $S_0T_LC_0$); two with only the repeating tones coming from the side ($S_RT_0C_0$ / $S_LT_0C_0$); and two with the repeating tones and the target both coming from the side ($S_RT_RC_0$ / $S_LT_LC_0$).

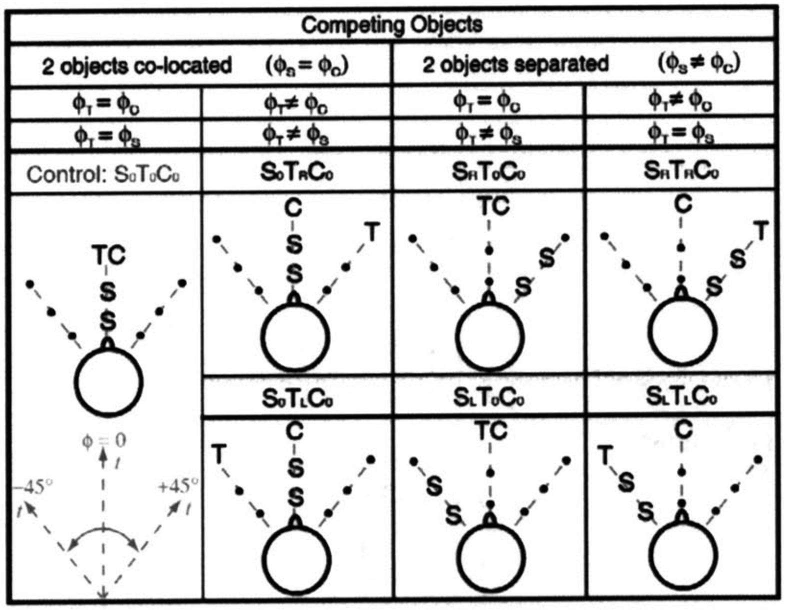


Figure 4-3: Spatial configuration for all two-object conditions. For each attended stream (either the repeating tones or the complex), there is a competing stream also present in these conditions. The control condition, $S_0T_0C_0$ has all components originating from 0° azimuth. In all of these conditions, the complex (C) always originated from the center.

Each subject completed the four experimental blocks, each consisting of 8 repetitions of each of the 14 stimuli in a random order, for a total of 448 trials.

4.3.3 Results

4.3.3.1 Localization of single streams

Figure 4-4 shows the mean and the standard error of the IID pointer in dB pooled across all subjects for all the single-object conditions (top, perceived location of the repeating tones; bottom, the complex). In each block of the experiment, subjects were able to match the IID pointer close to 0 dB when either the repeating tones and the target or the complex and the target were co-located at 0° [S_0T_0 (tones, "fast" / "slow") = -1.11 ± 0.761 / -0.101 ± 0.971 and T_0C_0 (complex, "fast" / "slow") = -0.572 ± 0.502 / -0.441 ± 0.589, mean (dB) ± 1 s.e.m.]. When the repeating tones or the complex were co-located with the target at a lateralized position (±45°), the absolute IID response was close to the maximum value of the acoustic pointer (all conditions exceeded 15 dB in magnitude with a maximum value of 20 dB).

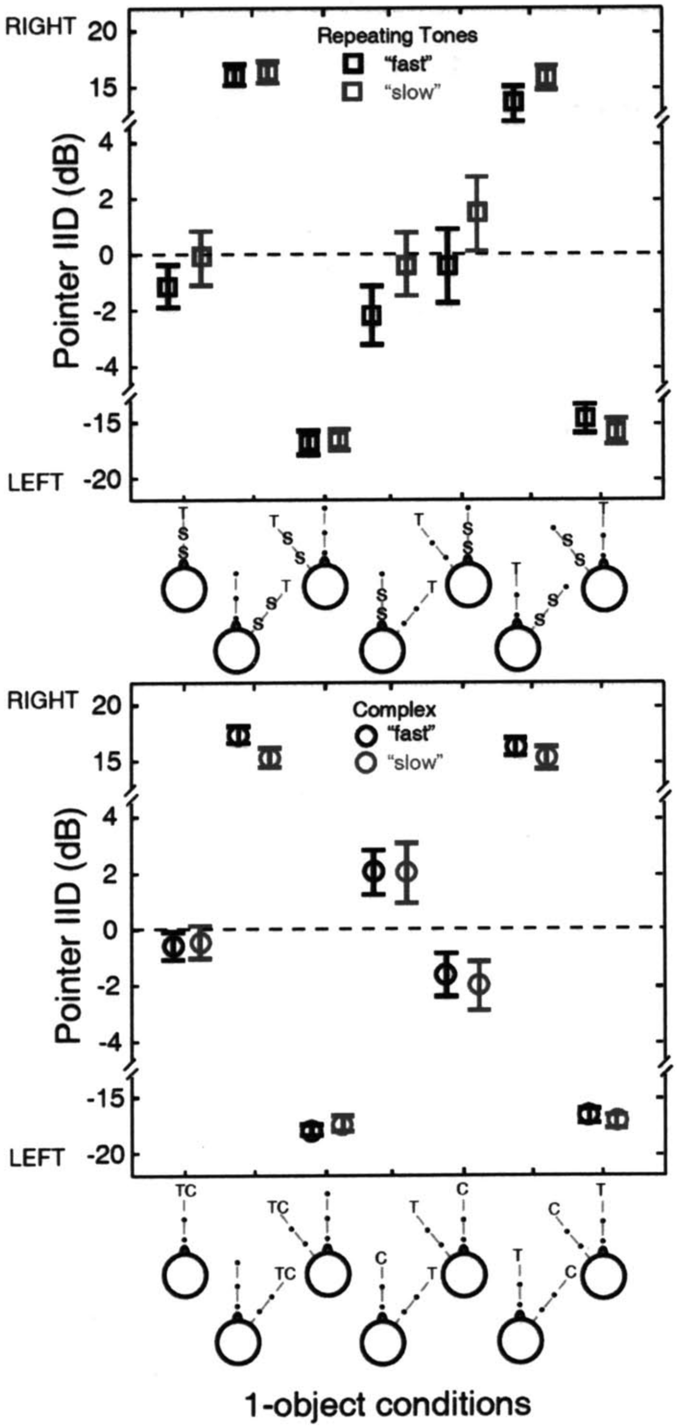
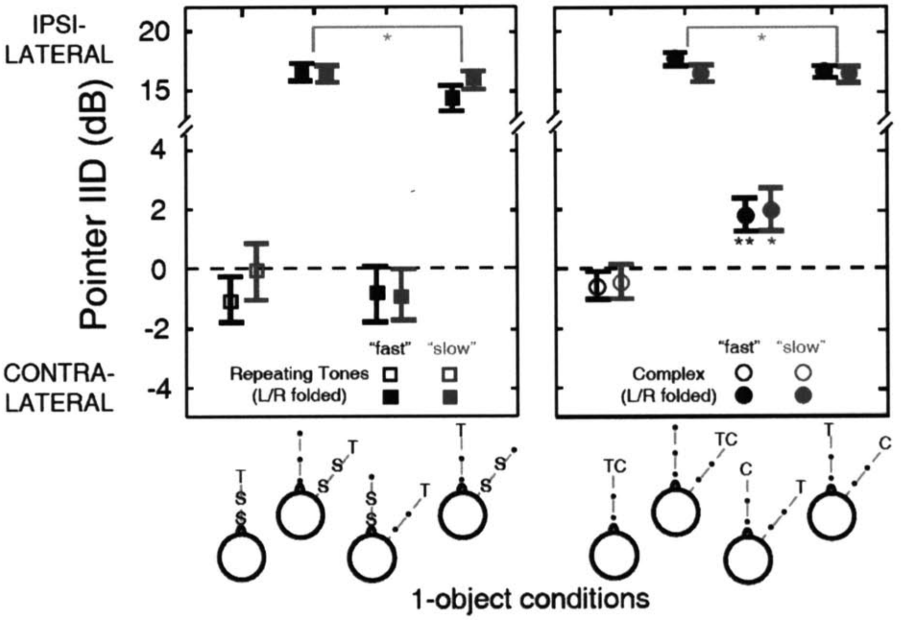


Figure 4-4: Perceived location of the repeating tones (left) and the complex (right) quantified by the mean value of the IID pointer, with positive values referenced to the right hemi-field. The darker symbols denote the IID values obtained in the "fast" block, while the lighter symbols denote the IID values obtained in the "slow" block.

In order to look at the effect of the perceived location for each condition independent of side, responses to conditions with any component (S, T or C) originating from -45° azimuth were mirror-flipped around 0 dB and combined with responses to conditions of their symmetric counterparts. Figure 4-5 shows the combined mean (independent of side) and the standard error of the IID pointer in dB pooled across all subjects for all single-object conditions tested (left: localization of repeating tones; right: complex).



Localization effects on

Repeating Tones		
$\phi_S = \phi_T$	$\phi_S \neq \phi_T$	
Controls No Effects	No effects	Integration*
No effects on repetition rate		

Complex		
$\phi_T = \phi_C$	$\phi_T \neq \phi_C$	
Controls No Effects	Integration**	Integration *
No effects on repetition rate		

Figure 4-5: Perceived location of the repeating tones (left) and the complex (right) quantified by the mean value of the IID pointer, with positive values referenced to the ipsi-lateral side of the stimulus origin for all the 1-object conditions. The darker symbols denote the IID values obtained in the "fast" block, while the lighter symbols denote the IID values obtained in the "slow" block. Asterisks denote the mean for the tested condition is significantly different from 0 dB (double asterisks, $p < 0.01$). Bracketed asterisks denote the significance level of a pair-sampled *t*-test between the tested conditions (single asterisk, $p<0.05$).

When subjects were asked to localize the repeating tones, the target did not influence the perceived location of the repeating tones, even at a faster inter-stimulus time-interval. In the condition where the target originated from the side, the perceived location of the repeating tones originated centrally was not affected by the target regardless of the stimulus repetition rate [Figure 4-5: $S_0T_{R/L}$ (tones, "fast" / "slow") = -0.881 ± 0.876 / -0.893 ± 0.856]. One-sample *t*-tests were performed to test whether their means were significantly different from zero dB (i.e., the nominal midline with Dunn-Sidak post-hoc adjustments for 2 planned comparisons), and no statistical differences were found [$S_0T_{R/L}$ (tones, "fast" / "slow"): t_{17} = 1.005 / -1.043; p_{DS} = 0.550 / 0.526]. If the repeating tones originated from the side and the target from the center, the perceived location of the repeating tones was closer to the midline when the repetition rate was faster [Figure 4-5: $T_0S_{R/L}$ (tones, "fast" / "slow") = 14.04 ± 1.00 / 15.80 ± 0.760]. A two-way repeated measures ANOVA was conducted on the mean data with factors of condition ($T_{R/L}S_{R/L}$ and $T_0S_{R/L}$) and inter-stimulus rate ("fast" and "slow"). The main effect of condition was significant [F(1,17) = 5.540, $p < 0.031$], but the main effect of repetition rate [F(1,17) = 1.291, p = 0.272] and the two-way interaction [F(1,17) = 3.223, p = 0.090] were not significant. This suggests that the target had an influence on the perceived location of the repeated tones in these two conditions, but the amount of influence was not significantly different at different repetition rates.

In the complex localization blocks, the perceived location of the complex was *pulled* by the target regardless of the repetition rate (Figure 4-5) When the target originated from the sides it significantly *pulled* the perceived location of the complex away from the midline [$T_{R/L}C_0$ (complex, "fast" / "slow") = 1.833 ± 0.540 / 2.022 ± 0.674]. One-sample *t*-tests were performed to test whether their means were significantly different from zero dB (with Dunn-Sidak post-hoc adjustments for 2 planned comparisons), and found that the perceived location of the complex was influenced by the target at both repetition rates [$T_{R/L}C_0$ (complex, "fast" / "slow"): t_{17} = 3.393 / 3.001; p_{DS} = 6.91 x 10^{-3} / 0.0160]. A centrally located target also *pulled* the complex originated from the sides towards the midline. A two-way repeated measures ANOVA was conducted on the mean data with factors of condition ($T_{R/L}C_{R/L}$ and $T_0C_{R/L}$) and repetition rates ("fast" and "slow"). The main effect of condition [F(1,17) = 4.681, p < 0.0450] was significant, but the main effect of repetition rate [F(1,17) = 2.618, p = 0.124] and the two-way interaction [F(1,17) = 4.312, p=0.0533] were not significant, suggesting that the target had an influence on the perceived location of the complex, but the amount of influence did not co-vary with the inter-stimulus-time interval.

4.3.3.2 *Localization in the presence of a competing stream*

Figure 4-6 shows the mean value and the standard error of the IID pointer in dB for all two-object stimuli (top, perceived location of the repeating tones; bottom, complex). For the stimulus condition in which all the components (S, T, C) were co-located at 0° azimuth, subjects were able to localize the repeating tones [$S_0T_0C_0$ (tones, "fast" / "slow") = -0.528 ± 0.761 / 0.144 ± 1.086] and the complex [$S_0T_0C_0$ (complex, "fast" / "slow") = -0.118 ± 0.672 / -0.775 ± 0.690] close to the midline. One-sample *t*-tests were performed to test whether these means were significantly different from zero dB with Dunn-Sidak post-hoc adjustments for 2 planned comparisons for tones localization and 4 planned comparisons for complex localization, and no statistical differences were found [$S_0T_0C_0$ (tones, "fast" / "slow"): t_8 = -0.694 / 0.133; p_{DS} = 0.757 / 0.133; $S_0T_0C_0$ (vowel, "fast" / "slow"): t_8 = -0.175 / -1.122; p_{DS} = 1.000 / 0.752]. This confirms that the subjects were able to localize the object of interest (i.e., repeating tones, or complex) close to the midline in the presence of a competing stream.

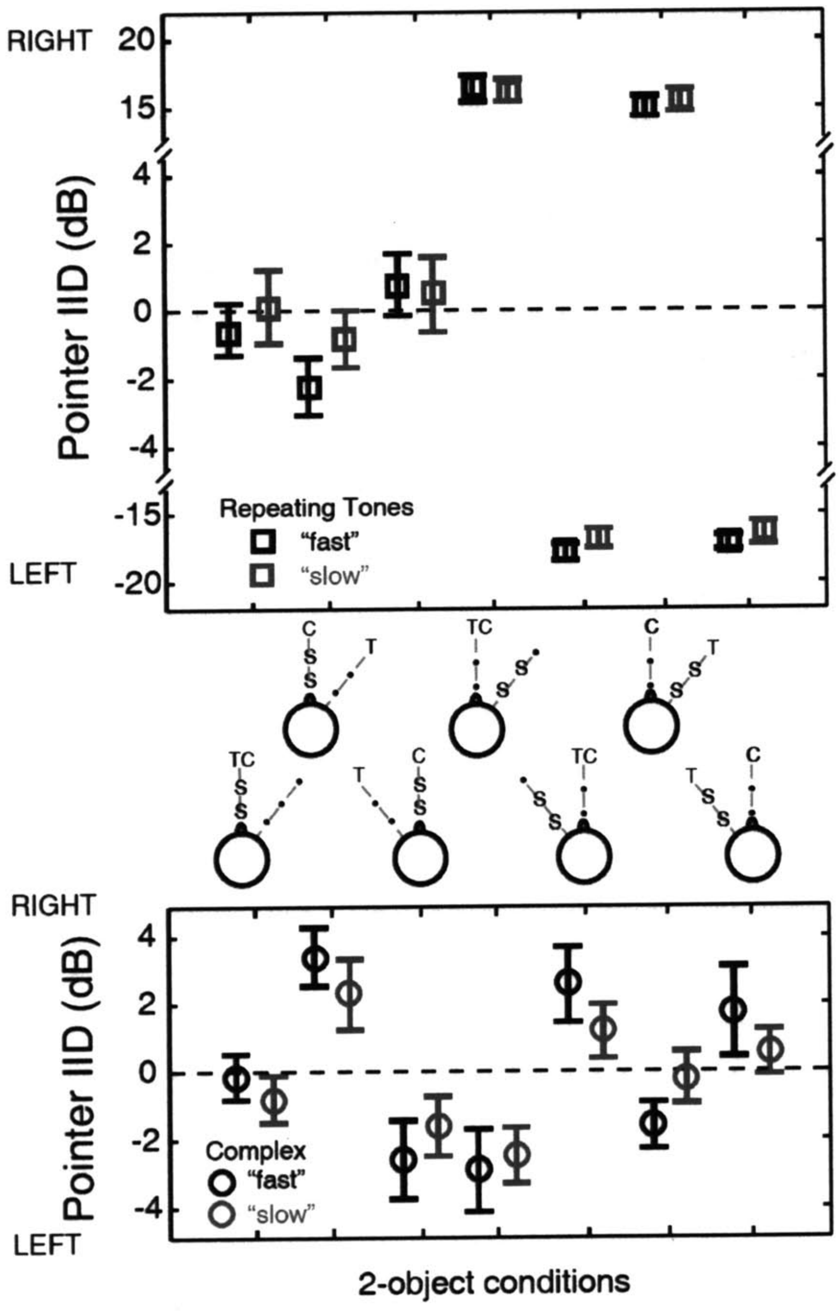
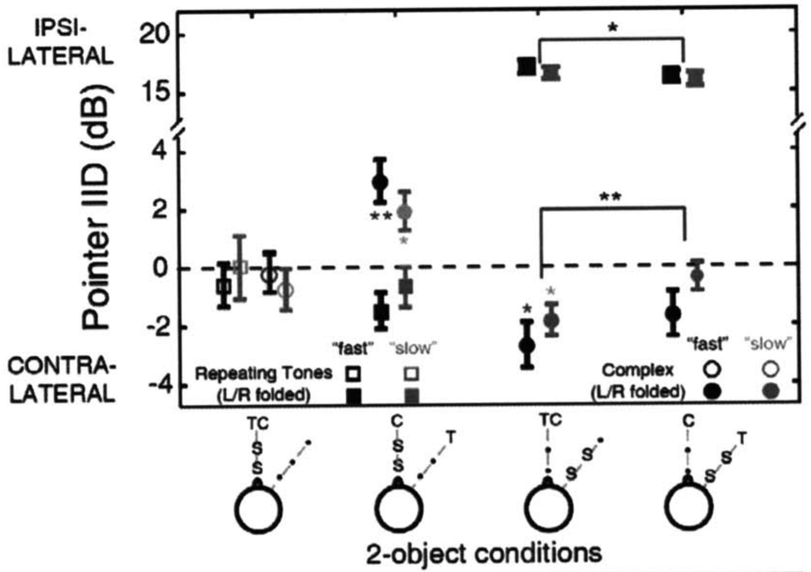


Figure 4-6: Perceived location of the repeating tones (squares, top panel) and the complex (circles, bottom panel) quantified by the mean value of the IID pointer, with positive values referenced to the right. The darker symbols denote the IID values obtained in the "fast" block, while the lighter symbols denote the IID values obtained in the "slow" block.

In order to investigate the influence of spatial cues on the perceived location of the repeating tones and complex independent of side, we mirror-flipped the IID pointer values around 0 dB for all the conditions with either the repeating tones and / or the target originating from the left. Figure 4-7 shows the combined mean (independent of the side of origin for the components S or T) and the standard error of the IID pointer in dB for all two-object tested at two repetition rates.



	2 objects collocated ($\phi_S = \phi_C$)		2 objects separated ($\phi_S \neq \phi_C$)	
	$\phi_T = \phi_C$	$\phi_T \neq \phi_C$	$\phi_T = \phi_C$	$\phi_T \neq \phi_C$
Localization effects of	$\phi_S = \phi_T$	$\phi_S \neq \phi_T$	$\phi_S \neq \phi_T$	$\phi_S = \phi_T$
● Complex	Controls No Effects	Integration** + (Repulsion)	Repulsion**	(Repulsion) + Integration**
■ Repeating Tones		(Repulsion)	---	Less Repulsion*
Rate of Repeltion		Less Repulsion**	Less repulsion for slower rate	

Figure 4-7: Perceived location of the repeating tones (squares) and the complex (circles) quantified by the mean value of the IID pointer, with positive values referenced to the ipsi-lateral side of the stimulus origin for all the 2-object conditions presented at two repetition rates ("fast", darker symbols; "slow", lighter symbols). Asterisks denote the mean for the tested condition is significantly different from 0 dB (single asterisk, $p < 0.05$; double asterisks, $p < 0.01$). Bracketed asterisks denote the significance level of a pair-sampled *t*-test between the tested conditions (single asterisk, $p<0.05$; double asterisks, $p < 0.01$).

The perceived location of the complex had a strong dependence on the spatial attribute of the repeating tones and the target. When subjects were asked to localize the complex co-located with the target at 0° azimuth with the stream of repeating tones originating from the side, the complex was perceived in the contra-lateral hemi-field relative to the repeating tones. One-sample *t*-tests showed that these means were significantly different from zero dB with Dunn-Sidak post-hoc adjustments for 4 planned comparisons [$S_{R/L}T_0C_0$ (complex, "fast" / "slow"): t_{17} = -3.409 / -3.187; p_{DS} < 0.0133 / 0.0214]. This suggests that the perceived location of the complex was *pushed* to the contra-lateral hemi-field at both repetition rates. When the target is co-located with the repeating tones on the side with the complex originating from 0° azimuth ($S_{R/L}T_{R/L}C_0$), there was a trend for the perceived location of the complex to be in the contra-lateral hemi-field with respect to the repeating tones, but the means were not significantly different from the midline [$S_{R/L}T_{R/L}C_0$ (complex, "fast" / "slow"): t_{17} = -2.294 / -0.752; p_{DS} = 0.132 / 0.916]. A two-way repeated measures ANOVA was also conducted on the mean IID value for the complex localization with factors of these two conditions [$S_{R/L}T_0C_0$ (complex) and $S_{R/L}T_{R/L}C_0$ (complex)] and inter-stimulus interval ("fast" and "slow"). The main effect of condition was significant [F(1,17) = 14.482, p < 0.00141], but the main effect of inter-stimulus interval [F(1,17) = 3.928, p = 0.0639] and the two-way interaction [F(1,17) = 0.262, p=0.615] were not significant. This suggests that the spatial configuration of the repeating tones and the target had a strong influence on the perceived location of the complex in these two conditions. However, even though that there was a general trend of a reduced "pushing" effect observed for the slower repetition rate on the perceived location of the complex [$S_{R/L}T_0C_0$ (complex, "fast" / "slow") = 17.043 ± 0.593 / 16.489 ± 0.528; $S_{R/L}T_{R/L}C_0$ (complex, "fast" / "slow") = 16.042 ± 0.525 / 15.929 ± 0.596], this was not statistically significant.

Subjects were also asked to report the perceived location of the repeating tones under these two exact spatial configurations. There was again a general trend that the perceived location of the repeating tones was less *pushed* by the complex at slower rates [$S_{R/L}T_0C_0$ (tones, “fast” / “slow”) = -2.741 ± 0.804 / -1.827 ± 0.573; $S_{R/L}T_{R/L}C_0$ (tones, “fast” / “slow”) = -1.665 ± 0.726 / -0.371 ± 0.494]. However, analogous to the perceived location of the complex, when a two-way repeated measures ANOVA was conducted on the mean IID value for the repeating tones localization with factors of these two conditions [$S_{R/L}T_0C_0$ (tones) and $S_{R/L}T_{R/L}C_0$ (tones)] and inter-stimulus interval (“fast” and “slow”), the main effect of condition was significant [$F(1,17) = 6.310$, $p < 0.0224$], but the main effect of inter-stimulus interval [$F(1,17) = 0.430$, $p = 0.521$] and the two-way interaction [$F(1,17) = 0.962$, $p = 0.340$] were not significant.

In the spatial configuration where the complex and the repeating tones originated from 0° azimuth and the target originated from ±45° ($S_0T_{R/L}C_0$), the perceived locations of the repeating tones and the complex *pushed* each other into different hemi-fields. There was a trend that the tones were perceived on the contra-lateral hemi-field with respect to the target [$S_0T_{R/L}C_0$ (tones, “fast” / “slow”) = -1.498 ± 0.627 / -0.653 ± 0.674], however, this was not statistically significant away from the midline [$S_0T_{R/L}C_0$ (tones, “fast” / “slow”): t_{17} = -2.388 / -0.969; p_{DS} = 0.0568 / 0.572]. The target, however, significantly *pulled* the complex towards its ipsi-lateral side at both repetition rates [$S_0T_{R/L}C_0$ (vowel, “fast” / “slow”) = 3.000 ± 0.711 / 1.948 ± 0.664; t_{17} = 4.213 / 2.932; p_{DS} = 2.34×10^{-3} / 0.0367]. A two-way repeated measures ANOVA was conducted on the mean data with factors of task (repeating tones or complex localization) and repetition rate (“fast” and “slow”). The two-way interaction [$F(1,17) = 11.493$, $p < 3.48 \times 10^{-3}$] and the main effect of task [$F(1,17) = 20.529$, $p < 2.96 \times 10^{-4}$] were significant, suggesting that the amount of repulsion between these two perceived locations was significantly influenced by the stimulus repetition rate.

4.3.4 Discussion

4.3.4.1 One-object localization and the effects of repetition rate

The one-object localization data show that the perceived location of an object can be influenced by spectro-temporal elements containing different spatial information in a number of different ways. First, the perceived location of the complex was influenced by the spatial content of the target which can logically be grouped as its fourth harmonic. This observation is consistent with the elevation of just-noticeable differences in interaural time difference (ITD) when a diotic spectral element was presented simultaneously [102, 103, 111, 120-122]. Binaural interference has also been shown to influence localization. Heller and Trahiotis [122] found that a synchronous diotic low-frequency interferer *pulls* the lateral position of a high frequency target towards the midline. Best *et al.* [13] also measured the reduction in laterality of a high-frequency sinusoidally amplitude-modulated target tone with a lateral ITD in the presence of a simultaneously gated diotic low-frequency interferer. Hill and Darwin [123] found that a complex has a *pulling* effect on a 500-Hz component but the 500-Hz component was not strong enough to influence the perceived location of the complex. They concluded that the contribution of the 500-Hz component to the lateral position of the complex was so minimal that a reliable assessment of the perceived location of the complex could not be made. The data in this experiment extended Hill and Darwin's [123] observations in that, given a different spectral envelope and fundamental frequency, the 500-Hz component can *pull* the perceived location of the complex as well.

Second, we observed that the localization of the repeating tones in the one-object condition was influenced by the location of the target, with all the elements being temporally disjoint from each other. This observation is an asymmetric one, with the position of the target only influencing the perceived location of the repeating tones when they originated from the side, but not centrally. The influence of the target location on the repeating tones is consistent with the view that the binaural system is "sluggish" [126-128]. Several studies have sought to measure the shape of the temporal integration window associated with binaural processing, and the estimated equivalent rectangular duration of the window varies from 40-200 ms [129, 130]. It is commonly accepted that the temporal characteristic of the binaural system is an order of magnitude more sluggish than those of the monaural system. It has also been suggested that the minimum audible angle (MAA) (i.e., the minimum angle of arc which must separate two sources emitting sounds in succession for a listener to discriminate them from a single source emitting the same sounds), is also influenced by the sluggishness of the binaural system. Perrot and Pacheco [131] observed that the MAA was significantly larger for inter-stimulus-onset-intervals of less than 75 ms for discriminating between two spatially displaced short bursts of pink noise.

From the estimated equivalent rectangular duration of the binaural window [129, 130], it is likely that at the “fast” repetition rate (i.e., with only 10 ms silence gap between sound elements), the rapid change in the spatial content of the 500-Hz tone should present significant challenges for the sluggish binaural system to resolve the spatial information. However, at the “slow” repetition rate (i.e., with more than 100 ms of silence gap between sound elements), the binaural system should be able to track the changes in the spatial content of spectro-temporal elements. Therefore, we hypothesized that for all the one-object conditions, repetition rate would have a strong influence on the localization of the repeating tones, but not for the complex.

The one-object complex localization data showed a consistent *pulling* effect of spatial information across-frequency for simultaneously gated spectro-temporal elements. This was also shown not to co-vary with the repetition rate, as hypothesized. However, the influence of repetition rate on the localization of temporally disjoint elements was elusive, despite the short silent gap in the “fast” condition which might have been influenced by the sluggishness of the binaural system. Subjects were able to accurately localize the repeating tones originating from the center with the target coming from the side regardless of the repetition rate. When the repeating tones originated from the side, the target *pulled* the perceived location of the repeating tones towards the midline, but this effect was not significantly different between the two repetition rates. This result suggests that binaural information might be integrated over time, but not all portions of the stimuli are weighted equally [132]. Bernstein *et al.* [133] and Kollmeier and Gilkey [130] suggested that different tasks may tap different aspects of binaural temporal processing. From the data obtained in this experiment, we did not observe a substantial influence of binaural sluggishness, at least not as pronounced as the integration of spatial cues across frequencies for spectral elements that were gated simultaneously.

In the absence of competition, the anecdotal reports of the subjects suggested that they only heard one stream in all the single-object conditions, despite some elements in that stream originating from different locations. Since ITD is well-known to be a weak cue for simultaneous sound segregation [14, 44], it is not surprising that the complex and the target are perceived as one object, even in cases where these two elements are spatially inconsistent. Therefore, the *pulling* effects that we observed in all the single-object conditions could be explained from an object formation point of view, suggesting that this spatial integration occurs within an object.

4.3.4.2 Localization interactions between streams

Unlike the localization data for the one-object conditions, the repetition rate has a significant influence on the localization of the repeating tones and the complex in the presence of competition in the auditory scene. We observed a significant reduction in the *pushing* effect between the streams in the condition $C_0T_{R/L}S_0$ presented at a "slow" repetition rate. While the comparisons between the conditions $C_0T_0S_{R/L}$ and $C_0T_{R/L}S_{R/L}$ did not show an effect of repetition rate, the observed trend was that the perceived location of both the repeating tones and the complex were closer to the midline in these conditions presented at the "slow" repetition rate (i.e. less pushing).

In the two-object conditions, the perceived location of the repeating tones and the complex were affected by the interaction between the competing streams. When the target and the complex were co-located centrally, subjects were able to localize it very close to the midline. However, in the presence of the repeating tones originating from the sides, the perceived location of the complex was substantially *pushed* by the repeating tones. Lorenzi *et al.* [110] observed both *pushing* and *pulling* effects and Braasch *et al.* [125] also noted that the presence of a distracter shifted the perceived location of the target in the opposite direction of the distracter if the target-distracter ratio is at 0 dB. Best *et al.* [134] postulated that the *pulling* effect of simultaneous stimuli is linked with across-frequency grouping phenomenon, while the strong *pushing* effect that they observed originated from the streaming of the simultaneously presented stimuli due to their the distinct temporal envelopes. Since the complex and the repeating tones were temporally disjoint in this experiment, and their spectral contents were also very different, there is little reason to doubt that they were perceived as two separate streams. Our observations add evidence that *pushing* is a phenomenon due to an interaction between two segregated streams. However, the *pushing* and *pulling* effects are not exclusive to one another. Comparing the perceived location of the complex in conditions $S_{R/L}T_{R/L}C_0$ and $S_{R/L}T_0C_0$, we showed that the complex originating centrally could both be *pushed* by the repeating tones originated from the side, and *pulled* (or a reduction of repulsion) by the lateral target.

4.4 General discussion

The results from the current experiment extended previous studies on perceived location under binaural interference conditions. Data from Heller and Trahiotis [122] and later replicated by Best *et al.* [13] show that a diotically presented sinusoidally amplitude-modulated (SAM) tone *pulls* a spectrally remote, simultaneously-gated SAM tone (centered at 4 kHz) with interaural time difference (ITD) of up to 600 μs towards the midline. A pure-tone target with ITD of 1.5 ms would normally be heard toward the ear receiving the lagging signal due to phase ambiguity and the auditory system's preference for physiologically plausible delays [135, 136]. However, Hill and Darwin [123] showed that when a pure tone with an ITD of 1.5 ms was embedded in a seven-component harmonic complex, the perceived target location of this tone was also *pulled* by the complex, although the contribution of the 500-Hz component to the lateral position of the complex was so minimal that they could not see an effect of this tone influencing the perceived location of the complex.

Two critical differences between the present experiments and those reported by Hill and Darwin [123] are worth highlighting. First, in our experiment, the 500 Hz target was much higher in amplitude (Figure 4-1B, bottom panel) compared to the rest of the harmonics due to the vowel spectral envelope. This is in contrast to Hill and Darwin's seven-component complex with equal amplitude for each harmonic. Second, subjects were not asked to report the perceived location of the 500-Hz component in the present study. Darwin *et al.* [44] showed that an ITD cue is insufficient to segregate the 500-Hz tone from the vowel complex, in line with the view that ITD cues are relatively ineffective for simultaneous grouping [7, 14, 36]. Since we did not manipulate the onset synchrony between the target and the complex in these experiments, and consistent with anecdotal reports by subjects, only one object was perceived regardless of the spatial configuration of the target and the complex.

The present data, extending the results of Hill and Darwin [123], show that the spatial location of the target can influence the perceived location of the complex. The perceived location of the complex, when originating from the side, was *pulled* towards the midline by the target originating from 0° azimuth. This extends the observation made by Heller and Trahiotis [122] and Best *et al.* [13] to instances of harmonic complexes spatialized using HRTFs instead of just ITD cues. Although lateralized spectro-temporal elements have been shown to reduce ITD sensitivity of diotic spectro-temporal elements and vice-versa [101], we know of no other investigations that have examined the perceived location of a centrally located object under binaural interference conditions. In the present experiment, we showed that the spatial location of the target, perceived as part of the complex located centrally, was able to *pull* the perceived location of the complex towards the ipsi-lateral side of the target away from the midline.

4.4.1 Perceived location and auditory grouping

Best *et al.* [13] noted that, only a few researchers have discussed binaural interference results in the framework of auditory object formation [121]. They reinterpreted many studies and suggested that the conditions under which binaural interference occur could be explained in terms of auditory grouping mechanisms.

Harmonicity and common onsets are strong cues that promote simultaneous grouping while inharmonicity and asynchronous gating increases the likelihood of hearing more than one object [3]. Studies like Trahiotis and Bernstien [120] and Heller and Trahiotis [122] showed that binaural interference was almost eliminated when asynchronous gating was used. Buell and Hafter [101] reported that when two low-frequency tones were gated simultaneously, binaural interference only occurred when the tones were harmonically related. Similarly, Hill and Darwin [123] reported that the perceived location of a target tone is affected by a simultaneously gated complex only when the tone is harmonically related to the rest of the complex. They concluded that spatial information may be combined selectively across frequency consistent with an auditory object formation framework proposed by Woods and Colburn [121].

Best *et al.* [13] also showed that sequential grouping cues can reduce the amount of binaural interference. In the binaural interference condition, a low frequency diotic SAM tone (centered at 500 Hz) was able to *pull* a spectrally remote SAM tone with ITD up to 600 μs towards the midline. However, this *pulling* effect was substantially reduced when the low SAM tone was embedded in a stream of diotic isochronous SAM tones of the same frequency. Furthermore, a similar reduction of the *pulling* effect was not observed when the low SAM tone was flanked by a narrowband noise instead of the isochronous SAM tones, suggesting that the reduction in the *pulling* cannot be explained simply from a peripheral adaptation argument.

Given the evidence suggesting that auditory grouping cues influence how spatial information is combined across frequency, is there a parsimonious interpretation of the perceived location of auditory objects in an auditory scene? Best *et al.* [134] postulated that integration of binaural information is a grouping phenomenon while repulsion suggests two objects are perceived. In the current experiment, subjects could easily distinguish the two objects in the two-object conditions: a fast stream of repeating tones and a stream of harmonic complexes repeated one-third the rate. These two distinct objects interact with each other in their perceived locations, exhibiting a strong *pushing* effect that effectively magnifies their perceived lateral separation from each other. The target, which was subjectively perceived as part of the complex in all the one-object conditions, also *pulled* the perceived location of the complex. The data in the present study are consistent with the postulation of a *pulling* effect within a stream and a *pushing* effect between streams.

The one-object condition in which the repeating tones have different spatial information than the target warrants further discussion. We observed that the target originating centrally *pulled* the perceived location of the repeating tones originating from the sides towards midline. However, when the repeating tones originated from the center, the target originating from the sides was not able to *pull* the perceived location of the repeating tones away from the midline. Even though the binaural system is often described as sluggish, we did not find convincing evidence that this observation can be attributed to the binaural temporal window with a time constant that is an order of magnitude longer than that of the monaural one. Subjects anecdotally reported that they generally heard one stream but with two separate locations. It might be the case that subjects perceive a stream of tones moving in time. Investigation into the apparent movement of an across-time tone stream is beyond the scope of this current study. In the present experiment, subjects were explicitly told to listen to the repeating tones and matched the perceived location of these tones. The measured perceived location of the repeating tones originating from the side was *pulled* by the centrally located target, but the target originating from the side did not have the same *pulling* effect on the repeating tones coming from 0° azimuth. Clearly, more research needs to be done in ascertaining the apparent motion of sequential stream with different spatial information across time.

4.4.2 Conceptual model for localization in a multi-source environment

There has been no attempt to develop a scene-analysis based explanation for localization of auditory objects in a multi-source environment [13]. Chapter 2 proposed a conceptual model accounting for streaming and the perception of spectro-temporal elements in each object. Extending this model to incorporate multiple sources displaced in space, we offer a preliminary framework to not only account for the observations made in the present localization experiment, but also reconcile other seemingly disparate localization effects, such as binaural interference, *pushing* and *pulling* effects observed in localization experiments as well as the buildup of the precedence effect.

Consider the spectro-temporal signal that reaches our ears. Without loss of generality, assume all auditory sources distributed spatially only along on the azimuthal plane and that the location of the source is based upon the spectro-temporal information received by two receivers, $R^L(f,t)$ and $R^R(f,t)$, with two parameters, f and t, representing frequency and time, respectively. The spectro-temporal content for each source S_i can be described by the following expression:

$$S_i(f,t) = A_i(f,t)e^{j(2\pi f\, t+\psi_i(f,t))}, \qquad (4.1)$$

where $A_i(f,t)$ and $\psi_i(f,t)$ represent the amplitude and the phase variation in both frequency and time associated with source i, respectively.

Assume that within each source the frequency components might contain different spatial attributes. Furthermore, assume that each source can move in time on the azimuthal plane relative to the observer. In this generalized case, the location of each source S_i relative to the observer is summarized by an azimuth angle function, $\theta_i(f,t)$, associated with each source i at each frequency f and time t.

The total signals at the receivers are the superposition of all the sources in the scene. Since the receivers are spatially displaced and may experience different acoustical filtering, e.g., head shadowing effects, the signals at the two receivers can be expressed as follows:

$$R^L(f,t)=\sum_{i=1}^{N} DTF^L_{\theta_i(f,t)}\{S_i(f,t)\}=\sum_{i=1}^{N} A_i^L(f,t)e^{j\left(2\pi f t+\psi_i^L(f,t)\right)}\text{; (4.2)}$$

$$R^R(f,t)=\sum_{i=1}^{N} DTF^R_{\theta_i(f,t)}\{S_i(f,t)\}=\sum_{i=1}^{N} A_i^R(f,t)e^{j\left(2\pi f t+\psi_i^R(f,t)\right)}\text{, (4.3)}$$

where there are N sources in the scene and the signals arriving at each receiver have been filtered by the directional transfer functions (which are receiver and angle dependent[1]), and subsequent filtering also makes both $A_i(f,t)$ and $\psi_i(f,t)$ receiver dependent, i.e., $A_i^{L/R}(f,t)$, $\psi_i^{L/R}(f,t)$.

In order to understand and process the sources in the environment, the listener must combine the information at the two receivers, $R^L(f,t)$ and $R^R(f,t)$, to extract the spatial information exploiting the interaural delay, interaural intensity and spectral cue differences between the two receivers. Furthermore, in order to understand and locate the sources in the environment, the listener must decompose $R^L(f,t)$ and $R^R(f,t)$ to try to recover the content and the location of these original sources. This process may be imperfect. The resulting estimates of the spectro-temporal content of the sources can be written as $\hat{S}_i(f,t)$, while their location estimates can be expressed as $\hat{\theta}_i(f,t)$. Perfect scene analysis would correspond to perfect recovery of the spectro-temporal content for each source:

$$\hat{S}_i(f,t) = S_i(f,t) \quad \forall i \ , \qquad (4.4)$$

and perfect recovery of the location estimate associated with each source would correspond to:

$$\hat{\theta}_i(f,t) = \theta_i(f,t) \quad \forall i \, . \qquad (4.5)^7$$

In order to focus on the effect of localization at an object-formation level, let us for simplicity assume henceforth that the monaural and binaural systems can extract the spectro-temporal as well as the spatial information of the whole scene perfectly, i.e.,

$$\hat{S}(f,t) = S(f,t) = \sum_{i=1}^{N} S_i(f,t); \qquad (4.6)$$

$$\hat{\theta}(f,t) = \theta(f,t) = \sum_{i=1}^{N} \theta_i(f,t) \, . \qquad (4.7)$$

Chapter 2 introduced an operator $\mathcal{D}(\cdot)$ to describe the scene analysis process. By extending this operator to incorporate binaural information in this model, $\mathcal{D}(\cdot)$ can potentially utilize all available information derived from the signal reaching the two receivers, i.e. $S(f,t)$ and $\theta(f,t)$, in order to produce the best estimates of the content of all the auditory events in the scene, i.e., $\left\{\hat{S}_i(f,t)\right\}$ and $\left\{\hat{\theta}_i(f,t)\right\}$.

As pointed out in Chapter 2, we wish to understand the operator $\mathcal{D}(\cdot)$, but as psychophysicists, we cannot directly observe its operation. By observing the apparent location of objects in a scene we may gain insight into this operator. Let

[7] Although equation (4.5) represents veridical localizing of the source locations, this is not necessarily possible, especially in reverberant situation, due to the degraded spatial information reaching to our receivers. Fidelity of source location is also sometimes not important, especially for source reconstruction in consumer entertainment systems [137].

$\mathcal{L}(\cdot)$ represent an operator that, utilizing all available information from the spectro-temporal and spatial information of the scene, produces an estimate of the spatial information about the object of interest. It is worth noting that while an object containing spatial attributes that are consistent across all frequencies generally forms a well-localized punctate image that can be ascertained unambiguously using traditional psychophysical pointing tasks, Licklider [138], Gardner [112] and Blauert *et al.* [139] described the image width influenced by auditory event with the phase content at the two receivers being interaurally incoherent. Therefore in general, the output of the operator $\mathcal{L}(\cdot)$ can be dependent on the task, especially for spectrally incoherent objects, in which $\hat{\theta}_i(f,t)$ could potentially be in a multi-dimensional space.

Along with other investigations on perceived location under binaural interference conditions [12, 13, 122, 123], we employed a psychoacoustic pointing task to ascertain the apparent location of the perceived objects. Under binaural interference conditions observed in these conditions, the *pulling* effect of different spatial components within the same object can be postulated as a spatial estimate averaged across frequency. One realization of the operator $\mathcal{L}(\cdot)$ could be of the form

$$\hat{L}_i(t) = \mathcal{L}\left\{\hat{S}_i(f,t),\hat{\theta}_i(f,t)\right\} = \frac{\int_f \left|\hat{S}_i(f,t)\right|^2 w(f)\,\hat{\theta}_i(f,t)\,df}{\int_f \left|\hat{S}_i(f,t)\right|^2 w(f)\,df}, \qquad (4.8)$$

where $w(f)$ denotes some spectral weighting of spatial information that is frequency dependent. In other words, the apparent location of an object under binaural interference conditions, as a function of time, is the spatial average weighted across frequency by the intensity of the spectral constituents in the perceived object. This is only one of many possible realizations of the proposed $\mathcal{L}(\cdot)$ operator, but it is hoped that such a postulation can be used as a starting hypothesis for testing a scene-analysis-based localization model.

Figure 4-8A illustrates a possible realization of a scene-analysis-based localization model, utilizing the two proposed operators, $\mathcal{D}(\cdot)$ and $\mathcal{L}(\cdot)$. In this realization, acoustic object $\hat{S}_i(f,t)$ is first promoted to the foreground, and its apparent location is then computed by the $\mathcal{L}(\cdot)$ operator. Therefore the system assumes that the localization process only operates on the object in the foreground and does not take into account of the spatial information available in the background.

However, from the observation that the apparent locations of two objects experience repulsion, modifications must be made to this serial approach. In the present experiment, we observed that the repulsion effect was observed when subjects were asked to localize the repeating tones co-located centrally with the complex, with the target originating from the side (condition $S_0T_{R/L}C_0$). The perceived location of the complex in this condition suggests that the target was grouped with the complex. However, turning to the results for the repeating tones localization, we also observed a hint of repulsion. If the localization operator $\mathcal{L}(\cdot)$ only takes the foreground object as input, combined with the localization results of the repeating tones in single-object conditions ($S_0T_{R/L}$), we would not observe repulsion. Therefore, the results from our experiment suggest that the localization process also takes the location of the unattended object into account. Figure 4-8B depicts such a system, that incorporates the location of the unattended object.

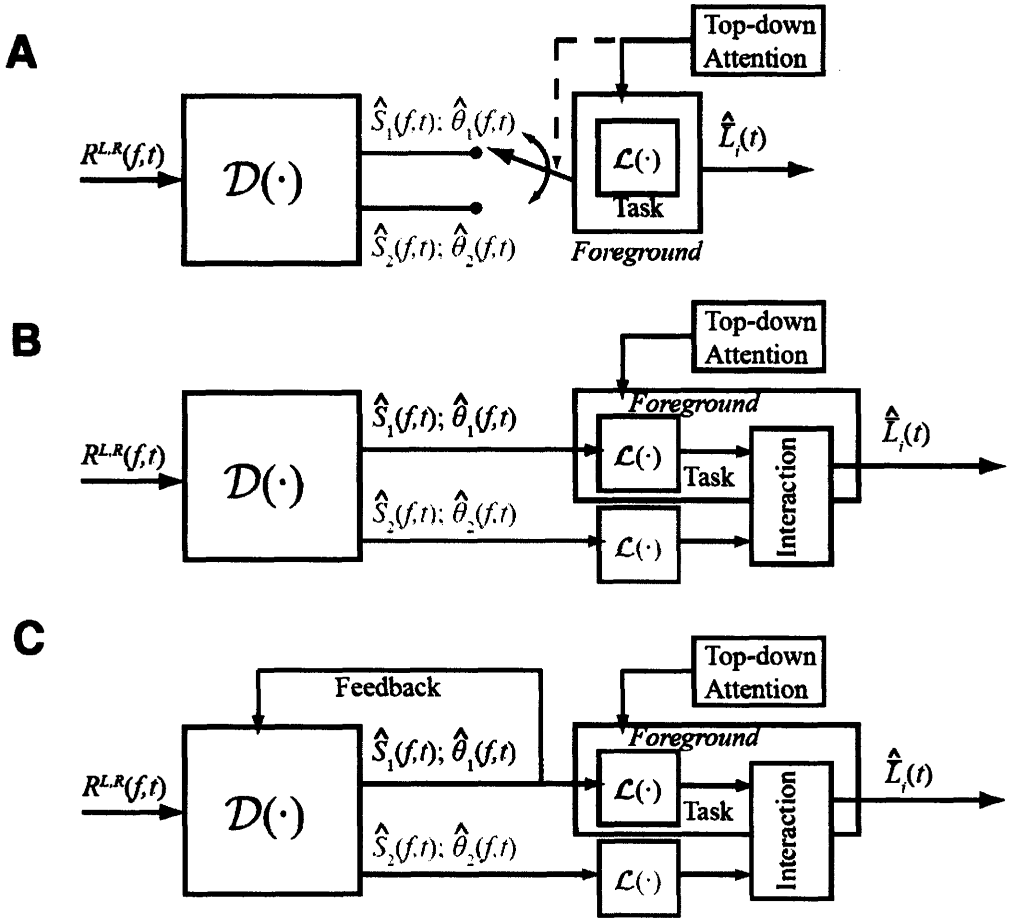


Figure 4-8: Realizations of a scene-analysis based localization model, using the two proposed operators, $\mathcal{D}(\cdot)$ and $\mathcal{L}(\cdot)$ (see text for detail). **(A)**: The location estimate of the object in the auditory foreground is the only input to the localization operator $\mathcal{L}(\cdot)$. **(B)**: The location estimate incorporates the location of other objects in the auditory scene besides the foreground object. A non-linearity block accounts for the *pushing* effect observed in the experiment. **(C)**: A realization of a feedback network that utilizes the location estimate of objects and in turn influences how the scene is parsed in time, i.e. the operator $\mathcal{D}(\cdot)$ is time-varying and takes the ongoing location estimate as a possible input.

4.4.3 Pulling and pushing in an object perception framework

The *pushing* effect for objects temporally displaced, as observed in these experiments, is a robust phenomenon and to a first order of approximation, the amount of *pushing* is a function of stimulus repetition rate. The *pulling* effect is also observed in the one-object complex localization data. These two effects are almost certainly not mutually exclusive. In the condition where the repeating tones originate from the side and the complex and the target are co-located at 0° azimuth (condition $S_{R/L}T_0C_0$), there is a strong *pushing* effect. However, if the target is co-located with the repeating tones rather than with the complex (condition $S_{R/L}T_{R/L}C_0$), the *pushing* effect is substantially weakened. This suggests that the target is grouped with the complex due to its *pulling* effect on the complex. Similarly, although we observed primarily a *pushing* effect in the condition where the repeating tones and the complex were co-located at 0° azimuth and the target alone originated from the side (condition $S_0T_{R/L}C_0$), it must also necessarily imply that the target has already been grouped with the complex such that it is spatially displaced from the repeating tones and thus causing a *pushing* effect. To what extent we can infer from the localization data the underlying perceptual organization of the scene is an interesting debate. It seems that such a postulation requires a leap of faith. However, since the perceived location of an object is influenced by the spatial information of the components that are grouped, it is equally as disturbing to hypothesize otherwise.

If the *pushing* effect holds more generally for the perception of distinct objects, even if they are not temporally displaced, then an interesting corollary can potentially be exploited to quantitatively assess the degree of perceptual segregation. Kubovy *et al.* [69] argued that perceptual boundaries are important for the formation of auditory objects. However, the exact nature of the auditory boundaries can only be speculated thus far [71]. In vision, perceptual boundaries, or edges, are determined by spatiotemporal discontinuities [68], while in audition, it has been postulated that boundaries are determined by spectro-temporal structures [71]. By parametrically varying spectro-temporal elements (e.g., in the dimension of onset synchrony, harmonicity, common modulation), the degree of perceptual segregation can be indirectly assessed by their perceived locations. If two objects were displaced in space, veridical parsing would imply that the perceived locations of the objects would repel one another. Conversely, if the auditory system combined these two objects into one fused image, then the spatial information of each object would be averaged and a *pulling* effect would be observed. There are two current methods of assessing the number of streams present in an auditory scene. One is to rely on a subjective response, asking whether the subjects perceive one or two streams. The other method is an arguably more objective method, exploiting the fact that temporal order can be easily detected within streams but not across streams. However, both methods, are discrete in nature (e.g. an integer number of objects perceived). Auditory objects are not as distinct as visual objects, consistent with the notion of the auditory scene being transparent [3]. Using a continuous measure such as perceived location to assess the degree of perceptual segregation might provide more insight into the nature of auditory objects and their spectro-temporal boundaries.

4.4.4 Precedence effect in an object perception framework

The precedence effect generally refers to a group of phenomena that are thought to be involved in resolving competition for perception and localization between a direct sound and a reflection (c.f., [113] for an extended review). In order to simplify the psychophysical experiments, investigators generally use a pair of stimuli, one leading and one lagging, to model the direct sound and a single reflection, respectively. Four main perceptual phenomena have been extensively quantified over the years as a function of the lead-lag delay. At short delays (1-5 ms for clicks), the lead and lag stimuli are heard as a fused image. As the delay increases, the fused image moves toward the direction of the lead signal, and this observation is often referred to as the localization dominance effect. At short delays, changes in the location of the lag stimulus are also harder to discriminate, and this is often known as the discrimination suppression. Finally, as the delay keeps increasing, subjects tend to report two sounds, and the sharp transition between the perception of one and two sounds is known as the echo threshold, but this parameter usually varies across individuals [140].

There is clearly a strong relationship among these aforementioned perceptual observations. However, to the best of our knowledge, there has not been a model proposed to account for all these effects in a cohesive framework, especially in terms of object formation. Consistent with the localization model based on an object perception framework, *pulling* (or fusion) effect is observed within an object. Echo threshold, reinterpreted in the object perception framework, can describe the amount of onset asynchrony needed for the operator $\mathcal{D}(\cdot)$ to parse the scene into two separate objects. An interesting corollary of this interpretation would be that the perceived locations of the lead and the lag stimuli, when their lead-lag interval passes the echo threshold, would experience a *pushing* effect. However, we know of no investigations into this hypothesis.

The echo threshold can also be elevated by exposing the listener to repeated presentations of lead and lag stimuli, and this phenomenon is often referred to as the buildup of echo suppression [104, 105]. Freyman and Keen [107] suggested that this increase of echo threshold might be due to the listener constructing a model of auditory space by rapidly mapping the reflected sounds in space as information comes in. In the object perception framework, such buildup effect can easily be explained by connecting the output of the localization operator, $\mathcal{L}(\cdot)$, as a feedback input to the time-varying decomposing operator, $\mathcal{D}(\cdot)$. Recently Dizon *et al.* [114] highlighted that precedence is also present in the ongoing portion of a long-duration source-reflection stimulus pair, without their initial onset time-of-arrival cues. They suggested that precedence should be viewed as a localization phenomenon related to a source and its reflections, rather than a mechanism solely associated with the onset of the stimulus. This suggests that the ongoing localization estimation of the scene can be used in time in a feedback network to influence how the scene is parsed (Figure 4-8C).

Yang and Grantham [105] also suggested that the buildup of echo suppression and the buildup of discrimination suppression may have different mechanisms due to the substantial difference in the increase of threshold under the buildup conditions. While we certainly do not believe that all of the phenomena originate from the same physiological mechanisms [141], such discrepancies can be incorporated in the object perception framework since the decomposing operator can also be task dependent (Chapter 2). By a first approximation, many phenomena related to the precedence effect can be explained in a cohesive framework, but clearly, much more research needs to be done to substantiate the validity of applying the object perception framework in the analysis of the precedence effect.

4.5 Conclusions

Results are consistent with the hypothesis that there is a *pulling* effect of spatial cues when the associated spectro-temporal elements are perceived within an object and a *pushing* effect is observed if the spectro-temporal elements are parsed as separate objects. Binaural sluggishness does not seem to account for the *pulling* effect. The strength of the *pushing* effect is, to a first-order approximation, a function of the inter-stimulus interval. In general, the shorter the inter-stimulus interval, the stronger the repulsion effects observed between objects. An auditory-grouping based localization model has been proposed to account for localization results observed in these experiments, as well as other observations ranging from binaural interference to the precedence effect.

CHAPTER 5 "WHAT" / "WHERE" INCONSISTENT READOUTS

The work described in this chapter is currently in preparation for journal submission.

5.1 Abstract

Descriptions of our surroundings are made up of the identity ('what') and the location ('where') of objects we perceive in one or more senses. However, the information available at our sensory epithelia is a chaotic juxtaposition of different elementary sensations [8, 9, 12] and we rely on a cognitive process, known as scene analysis, to group elements together into perceptual objects. Two cortical pathways have been identified in vision with a ventral pathway responsible for 'what' information, and a dorsal pathway, 'where', and a prevalent working model for auditory [142, 143] and somatosensory [144] processing suggests that this principle is modality independent. However, how the brain re-integrates information is still unknown [145]. Here we show that an auditory element can contribute to the perceived location but not the identity of an object. We found that when an object was presented in isolation, the identity and the perceived location were consistent with the constituents of the object. However in the presence of a competing stream, the localization computation was inconsistent with the identity of the object reported. Our results not only add evidence to the separate 'what'/'where' pathways but also challenge both scientifically and philosophically our view on object re-integration as a result of parallel processing.

5.2 Experiment

Modern philosophy defines an individual object as a bundle of intrinsic properties at a certain position [146]. An example of a visual object description would be: "I saw a bright red circle on the right half of a black screen", while an example of an auditory object description would be: "I heard a bowed violin playing a melody from my left". Each description consists of the 'what' (e.g. color and form in vision or timber and pitch in audition) and the 'where' (i.e. perceived location referenced to the observer) associated with the attended object.

In our common perceptual experiences, however, objects rarely occur in isolation. When multiple objects are present, object analysis must involve separating the relevant information to the object of interest from all other information in the environment [72]. This grouping of information into discrete perceptual entities is known as scene analysis. Gestalt psychologists use principles for perceptual organization, such as similarity, proximity and common fate to describe both visual [8] and auditory [3] scene analysis. It is noteworthy to highlight one fundamental difference between how information is arranged on these two sensory epithelia. Unlike in the retina, the cochlea does not have an explicit spatial representation of sound sources and this must be calculated using the spectro-temporal elements available at the two cochleae further down in the central nervous system [18]. While the visual system is capable of distinguishing changes in spatial changes of angle less than one minute or arc when two retinae are available, the spatial resolution limit of the auditory system is about two orders of magnitude higher [88].

Our ability to utilize auditory spatial cues for perceptual organization poses a paradox [12]. Daily experiences suggest that we have a robust spatial percept of our auditory surroundings [13], and spatial cues have also been empirically shown to be important for stream segregation across time [7]. However, spatial cues alone cannot be used to segregate speech of a single talker from similar simultaneous sounds [14]. The aim of the current study is to ascertain the extent spatial cues influence our perceptual organization. We presented a rhythmically repeating sound mixture identical to that used by Shinn-Cunningham *et al.* (see Appendix [56] and Section 5.3, Figure 5-1a). The feature of this sound mixture is the ambiguous tone, known as the target (T), that could logically be a member of either of the perceived objects, either as another tone in the rhythmic sequence (S) or as a harmonic in the vowel-shaped complex (C). While their experiment addressed the identity of the objects ('what'), we asked our listeners to report the perceived location of the objects ('where'). Using the exact stimuli allows us to compare the reported identity and perceived location of these objects given the same auditory scene.

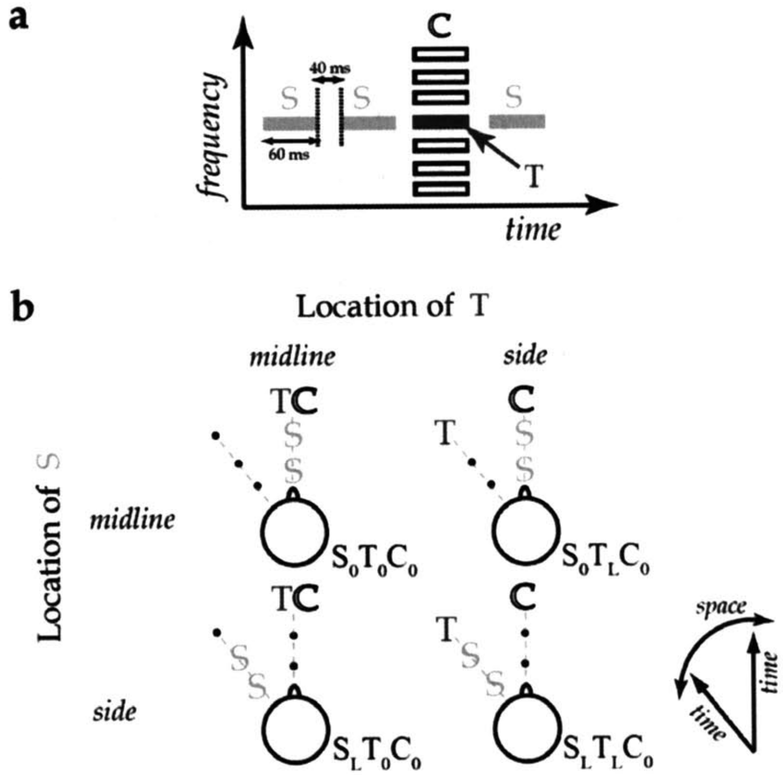
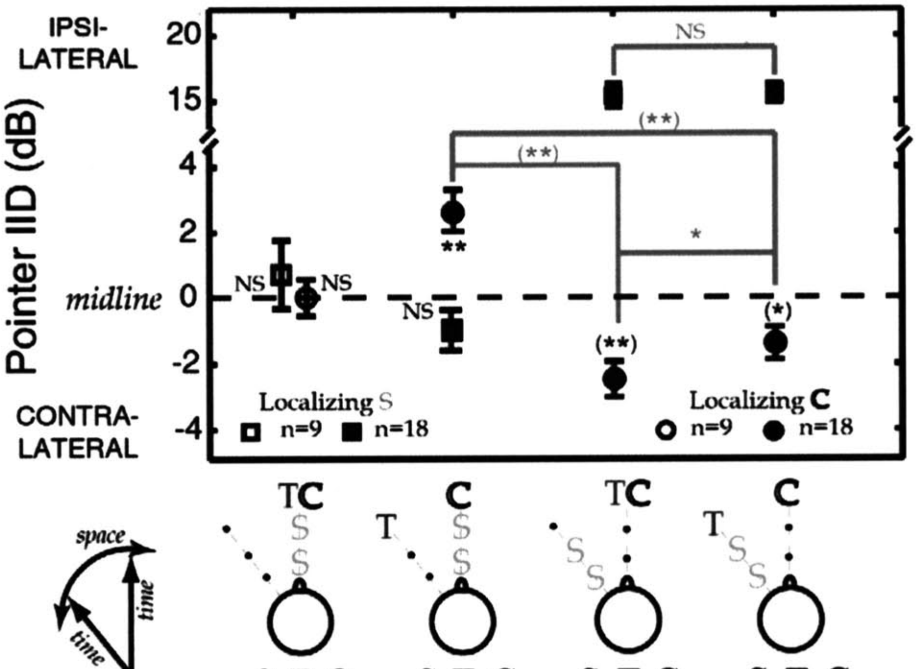


Figure 5-1 Time-frequency and time-space configurations of the two-object stimuli. a, Time-frequency diagram of the stimuli. **b,** Time-space configuration of the stimuli. The complex (C) originated from the midline in all these 2-object conditions. The location of the repeating tones (S) and target (T) can be originated from the midline or from the sides (±45° azimuth). If S and T were both originated from the sides, they always originated from the same hemi-field (either both from the right or left).

In the identity experiment (Appendix [56]), it was concluded that spatial configuration of the repeating tones and the target did not influence the vowel identity (/I/ or /ε/), while the spatial agreement between the complex and the target has a dramatic effect on the perceived identity of the rhythmic stream ("galloping" or "even"). Therefore, if the 'what' and 'where' computation is self-consistent in perceptual organization, we would expect the same asymmetric influence of the target on the perceived location: spatial attribute of the target should dramatically influence the localization of the repeating tones but not the complex.

Identical stimuli were presented to the listeners in two separate blocks, one in which listeners matched the perceived location of the repeating tones and one in which they matched the perceived location of the complex (see Section 5.3 and Figure 5-1b). Intermingled control conditions presented single-object stimuli in which only the attended object was presented (see Figure 5-4) to ascertain listener's object localization in the absence of a competing object.

While the localization results from the single-object stimuli generally concur with the identification experiment (see Appendix [56], Section 5.4 and Figure 5-5), we observed inconsistencies in the perceived location as implied by the identity of the objects (Figure 5-2 summarizes all the two-object stimuli localization results). The spatial attributes of the complex and the target did not have an influence on the perceived location of the repeating tones, despite their strong influence on the rhythmic identity. When the repeating tones were located from the side, the target had a strong influence on the rhythmic identity (Appendix [56]). The normalized d-prime score for the target present in the rhythmic stream for condition $S_LT_LC_0$ ['what'/tones: 0.96 ± 0.021; mean ± s.e.m.] was much higher than in the condition $S_LT_0C_0$ ['what'/tones: 0.12 ± 0.056 ± s.e.m.]. This suggests that the target was strongly grouped with the rest of the repeating tones if the spatial attribute of the target matches the repeating tones and not the complex (see Methods for the interpretation of the normalized scores). However, we did not observe the same influence in our localization experiment [paired *t*-test between $S_LT_LC_0$ and $S_LT_0C_0$ ('where'/tones), $t_{17} = 0.501$, $p = 0.623$].



	$S_0T_0C_0$	$S_0T_LC_0$	$S_LT_0C_0$	$S_LT_LC_0$
Location of T	*midline*	*side*	*midline*	*side*
Location of S	*midline*	*midline*	*side*	*side*
S, T localization effects on **C**	Controls. No Effects	Integration**	Repulsion**	Repulsion* + Integration *
T, **C** localization effects on S		No effects	—	No effects

Figure 5-2 Perceived location of the complex and the repeating tones summarized by the IID pointer (mean ± s.e.m.). For the control condition, the perceived location was not significantly different from the nominal 0 dB midline reference [$S_0T_0C_0$ ('where'/tones), t_8 = 0.701, $p_{DS,2}$ = 0.753; $S_0T_0C_0$ ('where'/vowel), t_8 = 0.045, $p_{DS,4}$ = 1.000]. The perceived location of the repeating tones were also not influenced by the target in condition $S_0T_LC_0$ ('where'/tones) [t_{17} = -1.616, $p_{DS,2}$ = 0.234]. Detailed statistics for comparisons denoted by brackets are included in the supplementary materials. Black denotes one-sample *t*-test statistics against 0 dB (midline) and grey denotes paired *t*-test between bracketed conditions.

More dramatic differences between the 'what' and 'where' readouts lie in the localization results of the complex in the presence of a competing stream tones. In the condition where the target was spatially displaced from both the repeating tones and the complex, the target was not perceived as part of the vowel [$S_0T_LC_0$ ('what'/vowel): 0.26 ± 0.15; mean ± s.e.m.], but the spatial attribute of the target had a significant effect on the perceived location of the vowel [one-sample *t*-test $S_0T_LC_0$ ('where'/vowel) against midline (0 dB), $t_{17} = 4.138$, $p_{DS,4} = 2.75 \times 10^{-3}$, with a Dunn-Sidak post-hoc tests for 4 family-wise comparisons, and henceforth all reported post-hoc adjustments employed the Dunn-Sidak tests and the number of family-wise comparisons is denoted in the subscript, also see Section 5.3.1]. When the repeating tones originated from the side, listeners reported the percept similar to the target not being grouped with the vowel regardless of whether the spatial attribute of the target matches the tones [$S_LT_LC_0$ ('what'/vowel): -0.094 ± 0.16; mean ± s.e.m.] or the complex [$S_LT_0C_0$ ('what'/vowel): 0.31 ± 0.087; mean ± s.e.m.]. However, we observed a differential effect in the localization experiment between these two conditions [paired *t*-test between $S_LT_LC_0$ and $S_LT_0C_0$ ('where'/vowel), $t_{17} = 3.358$, $p_{DS,3} = 0.0112$], suggesting that the target had a significant effect on the localization of the complex in these conditions.

By using the identical stimuli, results from our localization study and In conjunction with the identity study by Shinn-Cunningham *et al.* (Appendix [56]), we showed a double dissociation on the effects of auditory spatial cues in scene analysis between localization and identification tasks: the perceived location of the vowel but not its identity was affected by the spatial attribute of the target and the perceived identity of the rhythmic stream but not its perceived location was affected by the spatial attribute of the target. Furthermore, there were two instances in which the target was not perceived as part of the vowel complex when asked about its identity, yet the perceived location of the same object was significantly influenced by the target (see Figure 5-3 for a schematic summary).

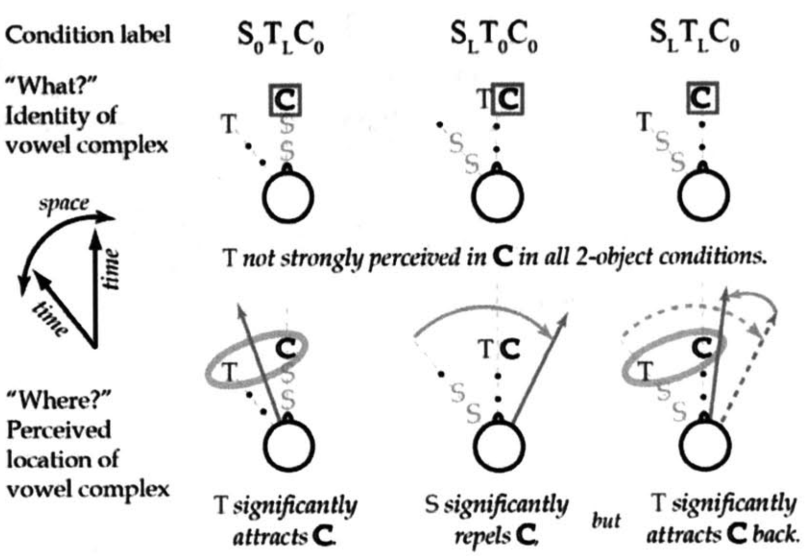

Condition label
$S_0T_LC_0$
$S_LT_0C_0$
$S_LT_LC_0$
"What?" Identity of vowel complex
T
C
S
S
space
time
time
T not strongly perceived in C in all 2-object conditions.
"Where?" Perceived location of vowel complex
T significantly attracts C.
S significantly repels C,
but
T significantly attracts C back.

Figure 5-3 Schematic summary of the inconsistent readouts of the identity and the perceived location of the vowel complex. While the target significantly influences the perceived location of the complex, it did not strongly affect the identity of the complex. The mean normalized score for the identity of the vowel complex in each of these conditions were all less than 0.5 (see text for details). Repulsion effect observed in the perceived location of the complex in conditions $S_LT_0C_0$ and $S_LT_LC_0$ are qualitatively similar to other localization studies reported in vision[147, 148] and audition [134].

Segregated 'what' and 'where' processing pathways in the visual cortex is well established [149, 150], and a prevalent working model for dual-object processing is also emerging in the auditory [142, 143, 151-154] as well as somatosensory [144] neuroscience literature. Although the pre-frontal cortex has generally been implicated for the task of re-integration of information [145], less is known about how these streams are combined, both within and across modalities. In audition, spatial information is first computed in the subcortical superior olivary structure [18]. Indeed, the inconsistent readouts for the 'what' and 'where' information in the auditory objects can potentially be traced to the different resolution limitation attributed to the monaural and binaural systems. Nonetheless, from our own perceptual experiences, the constituents of an object is generally consistent for both the 'what' and 'where' computation. The inconsistent 'what' / 'where' readouts presented here not only challenges philosophically how we should reconcile an element not contributing to the identity of an object yet influencing its perceived location, but also highlight the importance, yet still poorly understood, of the re-integration process of 'what' / 'where' information in object perception as a result of parallel cortical processing independent of the sensory modality.

5.3 Methods

5.3.1 Stimuli

Stimuli used in this experiment were identical with that used in the companion identity experiment (Appendix [56]). Briefly, they consisted of a 3-s long sequence, composed of three 100-ms-long elements: two repeating 500-Hz tone bursts (S) followed by a harmonic complex (C) of 125 Hz fundamental frequency, and synthetically shaped like a vowel. The target (T) was a 500-Hz tone presented simultaneously with the complex. All three components (S, T and C) were gated with a 60-ms long Blackman window followed by a 40-ms silent gap. This sound mixture caused a percept of two distinct auditory objects (rapid repeating tones and a slower sequence of repeating complex; see Figure 5-1a).

The complex consisted of first 40 individual random-phase harmonics of 125 Hz fundamental frequency and it was filtered with the same formant filter described in previous experiments (Chapter 2, Chapter 3, Appendix [56]). The complex did not contain any energy in the fourth harmonic, the frequency of the target. However, the identity of the vowel was dependent on whether the presented target being perceptually grouped with the rest of the complex [95].

Spatial cues in all three components (S, T, C) were controlled by processing sounds with head-related transfer functions (HRTFs) measured on a manikin at a distance of 1 meter in the horizontal plane [94]. Sources were processed to have spatial cues consistent with a source either from straight ahead (azimuth = 0°), 45° to the right, or 45° to the left (obtained by interchanging the two HRTFs associated with 45° to the right). In the two-object stimuli, the simulated complex azimuth was always zero, and four different spatial configurations were tested (Figure 5-1b). Symmetric spatial configurations were also tested, such that for each spatial configuration, each component that was not originated from zero azimuths had equal probability of originating from the right or from the left. In the single-object control trials, two of the three components (either S and T in the tones-localization block, or T and C in the complex-localization block) presented could independently be either originating from zero azimuths or from the sides (45° right or left).

5.3.2 Procedures

An acoustic pointer (200-Hz-wide band of noise centered at 2 kHz) was used to ascertain the subjective perceived localization response from the listeners. Subjects had control of the interaural intensity difference (IID) of this acoustic pointer by means of two buttons (right or left) and thereby changing its perceived location along the intracranial axis. Each trial began with a presentation of a three-second stimulus (listening phase), and this was followed by a three-second matching phase in which listeners had control of the IID pointer. Button presses were sampled at 25 kHz during the matching phase, such that the listeners essentially heard a punctate image traversing continuously along the intracranial axis as they controlled the IID pointer via the two buttons. At the conclusion of the listening phase, the three-second stimulus began to play and the buttons became inactive. When the matching phase restarted, the acoustic pointer reappeared with the last IID value, and the listeners regained control of the pointer. Alternation between the listening and the matching phases repeated until the listener was satisfied with matching the pointer to the perceived location of the attended object. The initial IID value of the pointer was assigned randomly from the range of +20 dB (full right) to -20 dB (full left). There were three configurations in the two-object conditions and three in the single-object conditions that had symmetric counterparts. Nine subjects participated in the experiment and they responded to each condition 8 times and the grand means for each subject responding to each condition were used for statistical comparisons.

5.3.3 Analysis

The IID pointer response from conditions with components originated from the left was sign-flipped such that for each condition with a symmetric counterpart, there were in effect two estimates of the perceived location for each subject in that spatial configuration. One-sample *t*-tests against 0 dB IID (midline) were used to ascertain whether the perceived object was significantly localized away from midline. To avoid family-wise Type-I errors, Dunn-Sidak post-hoc tests were performed for all planned comparisons. In the two-object stimuli, all harmonic complexes originated from 0° azimuth were tested against midline (0 dB) and hence there were 4 planned one-sample *t*-tests, and only 2 planned *t*-tests for the localization of the repeating tones. There were also 3 planned pair-wise comparisons for the localization of complexes for the two-object stimuli. There were 2 planned one-sample *t*-tests against midline for 2 conditions in each of the one-object stimuli block. All statistical inferences drawn used the two-tailed *t*-test with an alpha value set to 0.05 and normality assumptions verified.

The identity experiment employed signal detection theory to derive a normalized metric, with a score of 0 indicating that the response percentage for that condition being indistinguishable from the target-absent prototype, while a score of 1 indicating the response was identical to the target-present prototype (Appendix [56]). Therefore, only the mean and the standard error of this derived metric associated with each condition was reported, and no further hypothesis testing was carried out.

5.4 Supplementary materials

Single-object spatial configuration is summarized in Figure 5-4 and the perceived location of the vowel complex and the repeating tones in all single-object conditions are summarized in Figure 5-5. Figure 5-6 schematically summarizes the consistent readout, unlike in the two-object conditions, of the identity and the perceived location of the vowel.

5.4.1 Single-object control results

Localization results for single-object stimuli are summarized in Figure 5-5. Listeners were able to localize the repeating tones close to midline even if the target originated from ±45° azimuth. However, there is a significant effect of the repeating tones being pulled towards the midline when the repeating tones originated from the sides and the target from the center (paired *t*-test comparing S_LT_L and S_LT_0 ('where'/rhythm), $t_{17} = 5.47$, p = 4.15×10^{-5}). While the binaural system is often described as sluggish [126-128], from the results of a subsequent experiment suggest that this is due to integration of spatial cues in general within an object (Chapter 4). The target location is integrated with the spatial location of the complex no matter whether the complex originated from the midline (one sample *t*-test of T_LC_0 ('where'/vowel) against midline (0 dB), t_{17} = 3.434, $p_{DS,2}$ = 6.32×10^{-3}) or from the sides (paired *t*-test between conditions T_LC_L and T_0C_L ('where'/vowel), t_{17} = 2.291. p = 0.0350). This is consistent with the identification experiment that subjects judged the spatially displaced target being perceived as part of the vowel [T_LC_o ('what'/vowel): 0.78 ± 0.10; there is no equivalent T_0C_L ('what'/vowel) condition in the identity experiment].

5.4.2 Two-object supplementary results

Summary of the statistics that were in brackets in Figure 5-2 is provided here for completeness: one-sample *t*-test $S_LT_0C_0$ ('where'/vowel) against midline (0 dB), t_{17} = -4.588, $p_{DS,4}$ = 1.05×10^{-3}; $S_LT_LC_0$ ('where'/vowel), t_{17} = -2.821, $p_{DS,4}$ = 0.046; paired *t*-test between $S_0T_LC_0$ and $S_LT_0C_0$ ('where'/vowel), t_{17} = -6.014, $p_{DS,3}$ = 4.19×10^{-5} and between $S_0T_LC_0$ and $S_LT_LC_0$ ('where'/vowel), t_{17} = -5.349, $p_{DS,3}$ = 1.60×10^{-4}.

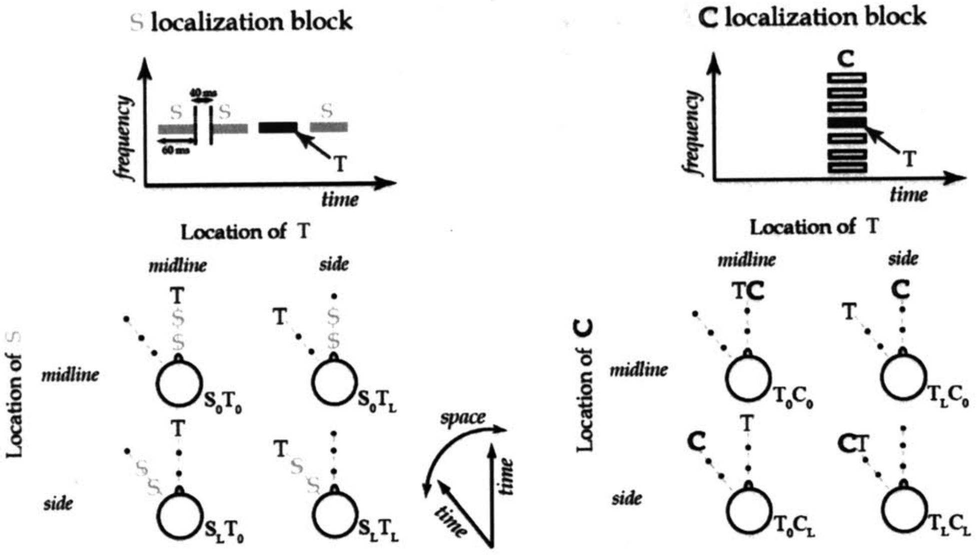


Figure 5-4: Time-frequency and time-space configurations for all single-object conditions.

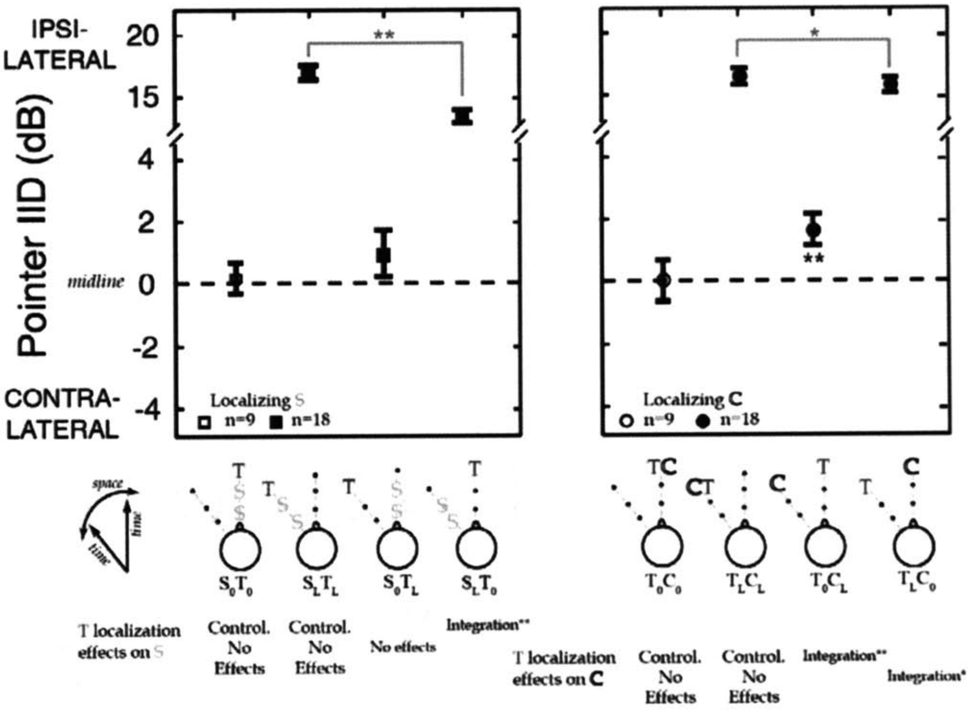


Figure 5-5 Perceived location of the complex and the repeating tones for all single-object conditions summarized by the IID pointer (mean ± s.e.m.).

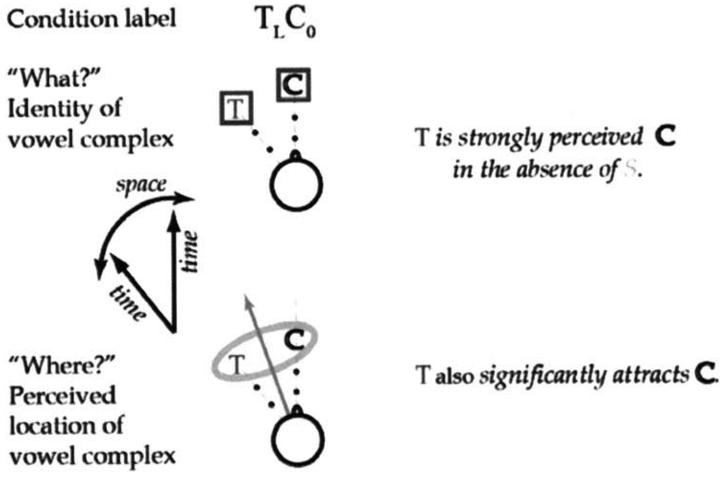


Figure 5-6 Schematic summary of identity and perceived location of the vowel complex in the single-object condition. This shows that the inconsistent readout is only observed when at least more than one object is present in the auditory scene.

CHAPTER 6 SUMMARY AND FUTURE DIRECTIONS

6.1 Summary

In Chapter 2, the transparent nature of the auditory scene was discussed and the energy trading hypothesis was presented (see also Appendix [56]). It was argued that while we do not have direct access to exactly how the spectro-temporal elements are parsed in the entire auditory scene (i.e., both foreground and background), failure to observe the energy trading hypothesis suggests the involvement of top-down attention in scene analysis. The two-object paradigm was introduced and the relative frequency between the repeating tones and the target was the variable in the experiment. Results were consistent with there being a trading relationship that governs how an ambiguous spectro-temporal element is allocated between two competing auditory objects, but it did not follow a strict intensity or amplitude trading relationship. The results proved that the organization of an across-time object influences the organization of simultaneous elements just as the grouping of simultaneous elements influence the grouping of an across-time stream.

Chapter 3 and Appendix described how spatial cues influence the allocation assignment of an ambiguous spectro-temporal element to the foreground object using the same two-object paradigm described in Chapter 2. Chapter 3 replicated the results published in Shinn-Cunningham *et al.* (Appendix [56]) with spatial cues that were adulterated by substantial reverberation. The main finding in these two experiments was that in some circumstances, an audible tone was not allocated to the two streams present in the scene, and that the tone was also not heard as a third object. This "non-allocation" phenomenon is an instance of which the hypothesis trading fails. While peripheral adaptation may contribute to the results, adaptation alone could not adequately account for the results observed in all the conditions tested. Both studies thus posited that the auditory system favors efficient processing over veridical representation of the entire scene: the rules for perceptual organization depend on the object being attended.

While Chapters 2 and 3 addressed how streaming affects the identity of objects we perceived (i.e., "what" rhythm or vowel perceived due to the contribution of the target), Chapter 4 reported how scene analysis affects the perceived location of the objects in the auditory scene (i.e., "where" the objects are perceived). It was observed that in general, the spatial cues within an object were integrated and that the perceived locations of two distinct objects experienced a repulsion effect that weakened as the stimulus repetition rate decreased. The integration and repulsion effects were not mutually exclusive.

Based on the localization experiment using the exact stimuli in the experiment described in the Appendix, Chapter 5 reported the inconsistent read out between the "what" and the "where" associated with the objects perceived. The target was never strongly heard in the identity experiment (Appendix and Chapter 3), yet the spatial attribute of the target had a significant influence on the perceived location of the object. Therefore, it suggested that scene segregation not only depends on which object is being attended (based on the "non-allocation" phenomenon observed), the process might also be task dependent.

6.2 Towards a conceptual framework for auditory scene analysis

In the previous chapters, conceptual models were developed by making inferences on the data observed in each experiment. While these provided insights into the stream segregation process for both the identification and localization tasks, there is not a general cohesive framework that accounts for the process of auditory scene analysis (ASA) in terms of the sources in the environment and the attentional state of the observer. In the remainder of this chapter, a preliminary conceptual framework to describe the task of scene analysis is presented.

To the best of the author's knowledge, this scene analysis framework is novel in audition (c.f., a sketch model presented in Figure 3 of Darwin and Hukin [7]). It is important to point out how this conceptual framework is fundamentally different from the models found in the computational auditory scene analysis (CASA) literature. In CASA, models are either biologically inspired [22, 155] or are based on a pure statistical approach [74, 156]. The goals of these algorithms are usually to achieve veridical parsing of the scene (using some correlation metric as objective measures or at least the individual reconstructed sources are perceptually similar to the original sources), or to replicate the perceptual organization based on previous psychoacoustical results (e.g. phonemic restoration, build-up of streaming, segregation of cross-trajectories etc).

The ultimate aim for the development of this conceptual framework is not to conceive a phenomenological model but to offer a language and a platform on which a cohesive description of the stream segregation process can emerge as more physiological and psychoacoustical evidence becomes available. The author acknowledges that this conceptual model is still in its infancy, and outstanding questions, both of mathematical and philosophical origin, are highlighted for future considerations.

6.2.1 Description of an acoustical event

Let the time signal representation of the i^{th} auditory event be denoted by $s_i(t)$. Using the short-time Fourier-transform (STFT) analysis formula, each signal can be represented in terms of its spectro-temporal content as:

$$S_i(f,\tau) = 2\pi \int_{-\infty}^{\infty} s_i(t) g^*(t-\tau) e^{-j2\pi ft} dt, \qquad (6.1)$$

where $g(t)$ is a fixed-duration window which is moved over the time function to extract the frequency content of the signal within that interval. As a consequence of the classical uncertainty principle, the fixed-duration window $g(t)$ for the STFT is accompanied by a fixed frequency resolution and therefore allows only a fixed spectro-temporal resolution [157]. Note that $S_i(f,\tau)$ is complex and the following notations are used when this complex quantity is expressed in polar form:

$$A_{i,rms}(f,\tau) = | S_i(f,\tau) | \; ; \qquad (6.2)$$

$$\psi_i(f,\tau) = \arg\left[S_i(f,\tau)\right]. \qquad (6.3)$$

The time-domain signal for the i^{th} source can be recovered from the dual STFT synthesis formula:

$$s_i(t) = \iint S_i(f,\tau)\, g(\tau - t)\, e^{j2\pi ft}\, df\, d\tau. \qquad (6.4)$$

If there are N acoustical sources in the scene, the entire auditory scene can be described as:

$$s(t) = \sum_{i=1}^{N} s_i(t) = \sum_{i=1}^{N} \iint S_i(f,\tau)\, g(\tau - t)\, e^{j2\pi ft}\, df\, d\tau. \qquad (6.5)$$

6.2.2 Statistical analysis leads to intensity trading hypothesis

Distinct acoustical sources are expected to be independent of each other and therefore there should not be any correlations in their phasal relationship. Let us consider the phase of the i^{th} source as a random variable, $\Psi_i(f,\tau)$, and this is uncorrelated with the phase of the j^{th} source, $\Psi_j(f,\tau)$, for all $i \neq j$. Furthermore, since there is no *a priori* knowledge of the starting phase, let us consider that $\Psi_i(f,\tau)$ is identically distributed uniformly for all sources.

The expected energy from the contribution of all the sources in the scene can be expressed as:

$$\begin{aligned}
E_\Psi\left[s^2(t)\right] &= E_\Psi\left[\left(\sum_{i=1}^{N} s_i(t)\right)\left(\sum_{i=1}^{N} s_i^*(t)\right)\right] \\
&= E_\Psi\left[\sum_{i=1}^{N} \iint |S_i(f,\tau)|^2 |g(\tau-t)|^2\, d\tau df + 2\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \iint S_i(f,\tau) S_j^*(f,\tau) |g(\tau-t)|^2\, d\tau df\right] \\
&= \sum_{i=1}^{N} \iint E_\Psi\left[|S_i(f,\tau)|^2\right] |g(\tau-t)|^2\, d\tau df + 2\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \iint E_\Psi\left[S_i(f,\tau)\right] E_\Psi\left[S_j^*(f,\tau)\right] |g(\tau-t)|^2\, d\tau df \\
&= \sum_{i=1}^{N} \iint |S_i(f,\tau)|^2 |g(\tau-t)|^2\, d\tau df.
\end{aligned} \qquad (6.6)$$

To simplify the analysis, assume that plane wave approximation is valid and thus let us define the specific acoustic impedance to be a real quantity, $z_0 = \rho c$, where ρ is the density of the air (or other medium) and c is the velocity of sound in the medium.

Let us also define the average intensity as:

$$\bar{I}_s(f,\tau) = \frac{\int_{<g>} |S(f,\tau)|^2 |g(\tau-t)|^2 \, dt}{\rho c}. \qquad (6.7)$$

with $\langle g \rangle$ denotes the width of the fixed-duration window.

Therefore, the expected average intensity of the whole scene can be expressed as:

$$\bar{I}_s(f,\tau) = \frac{\int_{<g>} |S(f,\tau)|^2 |g(\tau-t)|^2 \, dt}{\rho c} = \sum_{i=1}^{N} \frac{\int_{<g>} |S_i(f,\tau)|^2 |g(\tau-t)|^2 \, dt}{\rho c} = \sum_{i=1}^{N} \bar{I}_{s_i}(f,\tau).$$

(6.8)

Equation (6.8) mathematically summarizes the intensity trading hypothesis. Expressed in words, the expected intensity of the scene is the sum of the expected intensity in each of the sources in the scene.

6.2.3 Receiver and its spatial relationship to the acoustical sources

Consider a receiver, $r(t)$, that records all signals generated by the sources distributed in space. Let us adopt the receiver position as the spatial reference point. In a natural auditory scene, all acoustical events are generally distributed in a three-dimensional space, and they might be moving in time. Let us consider a vector $\theta_i(f,\tau)$ that summarizes the location of each spectro-temporal element associated with the source i in spherical coordinates, i.e., $\theta_i(f,\tau)$ is a three-dimensional vector that describes the spatial coordinate of spectro-temporal elements associated with source i relative to the receiver in the radial, azimuth and elevation dimensions. Therefore according to the receiver, each acoustical source $\mathcal{A}_i$ can be described by a set of two functions:

$$\mathcal{A}_i\left\{S_i(f,\tau);\boldsymbol{\theta}_i(f,\tau)\right\}. \qquad (6.9)$$

6.2.3.1 *Validity of locatedness as an inherent characteristic of acoustical events*

Although operationally we assigned a location vector to every acoustical source in the scene relative to the receiver, some modern philosophers actually regard the perception of sound to be non-spatial [158], as opposed to vision, and thus they would argue such a spatial characterization of an acoustical event is invalid. The confusion in the role of space in auditory perception perhaps stems from the philosophical debate on what exactly constitutes the object to be perceived: the auditory event or the sound waves. Visual perception has often served as the exemplar modality for philosophical study (and to a large extent true in scene analysis) but only recently a philosophical framework that is based on "sonic realism" has emerged [159]. It is hopeful that the conceptual framework presented here and its eventual form will be rigorous in philosophical considerations such that the mathematical descriptions as well as the language used are consistent across all these fields of study and leave little room for confusion in the role of auditory object perception in the field of psychophysics.

6.2.3.2 *Recordable auditory experiences*

Recently I had a mild case of bronchitis and I was aware of the "hissing" sound originating from my lungs. Perhaps due to the stress associated with the writing stage of the dissertation, I also experienced occasional tinnitus in my right ear. According to the observer (me), both auditory experiences were real. However, there is a distinct difference between these two auditory experiences. In the case of the "hissing" sound originated from my lung, it is an acoustical event that can be recorded by another receiver, $r^{\dagger}(t)$, e.g., via a doctor's stethoscope, even though it was originated from inside my body. However, no instruments can record the signals associated with the genesis of tinnitus or any other forms of auditory experiences that are posited to be of neurological origins, e.g., auditory hallucinations in schizophrenic patients [160]. Therefore, the operational definition of the receiver in this conceptual framework, $r(t)$, only describes all acoustical sources of which the acoustical waves can be recorded by some instruments.

6.2.4 Directional transfer functions relating sources with the receiver

The plenacoustic function [161], in reference to the plenoptic function introduced by Adelson and Bergen [68] which defines "all views in a room", describes the sound pressure $p_{(x,y,z)}(t)$ recorded at location (x,y,z)[8] and time t, given the acoustics of an environment. When we consider sources that are spatially displaced, the acoustical waves associated to each source would experience different filtering effects, depending on its location in the environment. The following structure is similar to that of the plenacoustic function except that the spatial coordinate is in reference to the receiver and not to the individual sources.

Let us assume that there is one receiver located at a fixed position with one source located at (x,y,z) referenced to the receiver. If the directional transfer function associated with the environment at the receiver location is given by $h_{x,y,z}(t)$, then the signal at the receiver $r(t)$, generated by source $s(t)$ at a location (x,y,z) would be given by

$$r(t) = s * h_{x,y,z}(t) = \int_{\tau} s(\tau) h_{x,y,z}(t-\tau) d\tau \,. \qquad (6.10)$$

In general, for N sources distributed in space, the whole scene can thus be expressed at the receiver as:

[8] Although the location vector $\boldsymbol{\theta}_i(f,\tau)$ was previously parameterized in spherical coordinates, the Cartesian parameterization is more familiar [68] and thus will be used in the discussion that follows.

$$r(t)=\sum_{i=1}^{N} s_i * h_{x_i,y_i,z_i}(t)=\sum_{i=1}^{N}\int_{\tau} s_i(\tau)h_{x_i,y_i,z_i}(t-\tau)d\tau . \qquad (6.11)$$

Now consider two receivers, $r^R(t)$ and $r^L(t)$, distributed in space and for simplicity but without loss of generality, assume the receivers are displaced along the x axis by a fixed distance Δx, then the whole acoustic scene according to the two receivers can be described as:

$$r_R(t)=\sum_{i=1}^{N} s_i * h_{x_i,y_i,z_i}(t)=\sum_{i=1}^{N}\int_{\tau} s_i(\tau)h_{x_i,y_i,z_i}(t-\tau)d\tau ; \quad (6.12)$$

$$r_L(t)=\sum_{i=1}^{N} s_i * h_{x_i-\Delta x,y_i,z_i}(t)=\sum_{i=1}^{N}\int_{\tau} s_i(\tau)h_{x_i-\Delta x,y_i,z_i}(t-\tau)d\tau , \qquad (6.13)$$

and in general, $\{h_L(t),h_R(t)\}=\{h_{x-\Delta x,y,z}(t),h_{x,y,z}(t)\}$ would be referred to as the binaural directional transfer functions.

6.2.4.1 Relations to common stimuli used in binaural listening experiments

Many traditional binaural experiments were performed with stimuli presented over the headphones by manipulating the binaural directional transfer functions. The binaural directional transfer functions can be recorded in an enclosed environment such as a classroom [94]. When these transfer functions are recorded in an enclosed environment, especially if they contain substantial amount of reverberation, they are often referred to as the binaural-room impulse responses (c.f., Chapter 3). If the transfer functions are pseudo-anechoic (c.f., Chapter 2, Chapter 3 and Chapter 4), they are often referred to as the head-related transfer functions.

Other common of manipulations of the binaural directional transfer functions are as follows: 1) interaural time delay (ITD) manipulation only: $\{h_L(t)=\delta(t-ITD),\ h_R(t)=\delta(t)\}$, where $\delta(t)$ represents the Dirac impulse response, 2) interaural intensity difference (IID) manipulation only: $\left\{ h_L(t) = 10^{\frac{ILD}{20}} \delta(t),\ h_R(t) = \delta(t) \right\}$.

6.2.5 Monaural and binaural transformation of the received signal

Let $\mathcal{M}(\cdot)$ be a non-linear, time-varying function that takes $r(t)$ or equivalently $R(f,\tau)$ as inputs and generates an estimate of the scene (i.e., summation of all sources) $\hat{s}_M(t)$ or equivalently $\hat{S}_M(f,\tau)$ (Figure 6-1). The transformation from the input $r(t)$ to $\hat{s}_M(t)$ may not be perfect. For example, in an extreme case, if a component at the input $R(f,\tau)$ is below the absolute threshold of (monaural) hearing, then $\hat{S}_M(f,\tau)$ alone would not contain that element. Other non-linearities may also be introduced due to masking or adaptation associated with peripheral encoding.

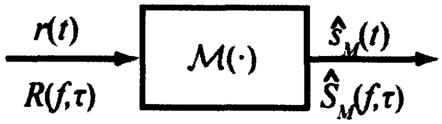


Figure 6-1: A general non-linear, time-varying function, $\mathcal{M}(\cdot)$, that transforms monaural input into an estimate of the scene.

Similarly, let $\mathcal{B}(\cdot)$ be an operator that takes in the two outputs from the $\mathcal{M}(\cdot)$ operators, i.e., $\hat{S}^L(f,\tau)$ and $\hat{S}^R(f,\tau)$, and using the combined ITDs, IIDs and spectral cues to estimate two parameters, $\hat{S}_B(f,\tau)$ and $\hat{\theta}(f',\tau')$ (Figure 6-2). The frequency and time arguments in $\hat{S}_B(f,\tau)$ and $\hat{\theta}(f',\tau')$ are denoted differently because there is evidence that the binaural system is "sluggish" [129, 162, 163], which may result in poorer spectro-temporal resolution compared to the information already available at the monaural level.

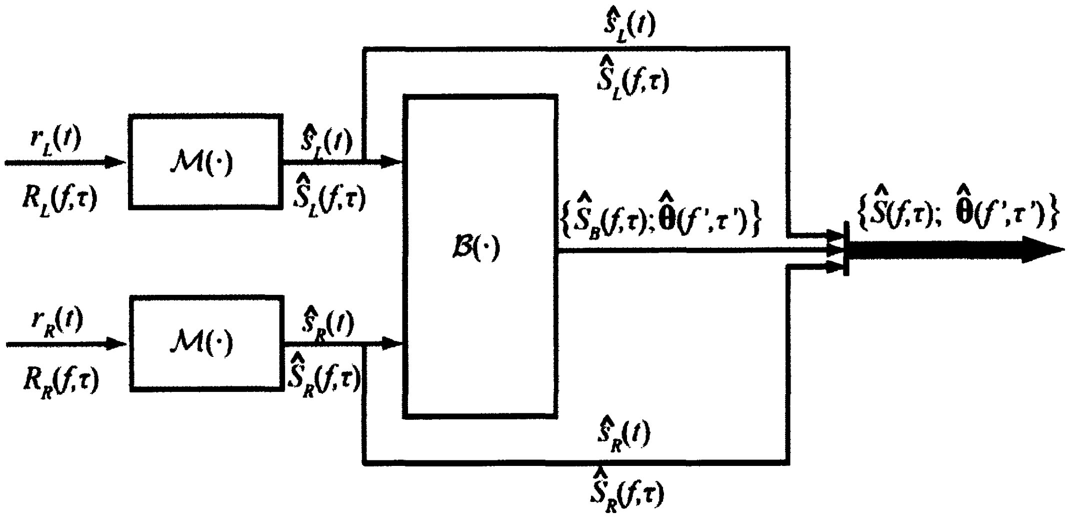


Figure 6-2: A general non-linear, time-varying function, $\mathcal{B}(\cdot)$, that takes the two monaural inputs and estimates the spatial attributes of the scene.

While the monaural and binaural spectro-temporal information can potentially be used separately, e.g., one could conceive a system with a top-down process that only considers one monaural output, say $\hat{S}_L(f,\tau)$, and ignores the other monaural output as well as the binaural output, i.e., $\hat{S}_R(f,\tau)$ and $\hat{S}_B(f,\tau)$, we will only consider $\hat{S}(f,\tau)$ henceforth to denote the information available either at the monaural or at the binaural outputs.

6.2.6 Decomposing and the task operators

The decomposing operator, $\mathcal{D}(\cdot)$ and the task operators, $\mathcal{I}(\cdot)$ and $\mathcal{L}(\cdot)$ have been introduced previously. Briefly, the operator $\mathcal{D}(\cdot)$ utilizes all available information from the signal reaching the receive to produce the best estimates of the content of all the auditory sources in the scene, while $\mathcal{I}(\cdot)$ and $\mathcal{L}(\cdot)$ represent the task of matching the spectro-temporal content and the location of the attended object, respectively. Arguments have been presented in Chapter 2, Chapter 3 and Appendix that the observed violation of the intensity trading hypothesis can be used to draw inferences whether the decomposing operator is a purely bottom-up process or not.

6.2.6.1 *Possible resolution to the what/where inconsistent readout*

Chapter 5 highlighted the inconsistent readout that was observed by comparing the results from the identification with the localization experiment using the exact same stimuli. The perceived attenuation of the target in Chapter 3 and the Appendix was high for all spatial configurations in the vowel identification task, yet the target had a significant effect on the perceived location of the vowel (Chapter 5). However, the amount of spatial integration caused by a spectro-temporal element when it is spatially displaced from the rest of the object has never been examined as a function of its intensity. The following paragraphs describe how a series of new experiments can possibly be used to resolve the inconsistency observed in the perceptual organization process under the presented framework.

Consider the i^{th} source, $\mathcal{A}_i\{S_i(f,\tau);\theta_i(f,\tau)\}$, presented in a sound mixture $\mathcal{A}\{S(f,\tau);\theta(f,\tau)\}$. The perceived intensity of the i^{th} source in this mixture was reported to be $\hat{\bar{I}}_i(f,\tau)$ (Figure 6-3A) by a subject and its perceived location to be $\hat{\bar{L}}_i(\tau)$ (Figure 6-3B). In a control experiment, only the i^{th} source (based on the reported source content heard by the subject) was presented, i.e., $\mathcal{A}\left\{S(f,\tau)=\hat{\bar{I}}_i(f,\tau)\,;\theta(f,\tau)=\theta_i(f,\tau)\right\}$[9]. Since there is only one object presented in the scene, it should not depend on the decomposing operator (Figure 6-3C). Let the perceived location of the i^{th} source presented alone be denoted as

[9] By definition in equation (6.7), $\hat{S}_i(f,\tau)$ is not equivalent to $\hat{I}_i(f,\tau)$ due to the fixed-duration window. However, it has been shown that the reported intensity spectro-temporal profile of an object $\hat{I}_i(f,\tau)$ can be very similar to the object of interest if presented alone, i.e., $\hat{I}_i(f,\tau)\cong\hat{S}_i(f,\tau)$ [164] for some broadband stimuli.

$\hat{\hat{L}}_{i[\mathrm{alone}]}(\tau)$. If the decomposing process is not task dependent (i.e., identification or localization task), then the perceived location of the attended object in the scene, $\hat{\hat{L}}_{i}(\tau)$, must equal to the perceived location of the object presented alone, $\hat{\hat{L}}_{i[\mathrm{alone}]}(\tau)$. In other words, the perceived content of the i^{th} source in a sound mixture when a localization task is performed (Figure 6-3B) is consistent with the perceived content of the same source in a mixture when an identity task is performed (Figure 6-3A), as inferred by the control single-object localization task based on the perceived content from the identity task (Figure 6-3C).

The final proposed control experiment (Figure 6-3D) examines whether the location attribute, $\theta(f,\tau)$ of the scene could influence the task of reporting the intensity of the source. It has been shown that perceived loudness of spectro-temporal elements is different when presented binaurally compared to presentations under monaural presentations, and this phenomenon is generally referred to as binaural loudness summation [165, 166]. However, to the best of the author's knowledge, it has not been explicitly tested that the perceived intensity or loudness of an object can be influenced by the spatial attributes of its spectro-temporal elements. The control experiment in Figure 6-3D investigates whether the reported intensity of the i^{th} source presented alone, $\hat{\hat{I}}_{i[\mathrm{alone}]}(f,\tau)$, has the same spectro-temporal intensity profile as the i^{th} source $S_i(f,\tau)$ regardless of its spatial contents.

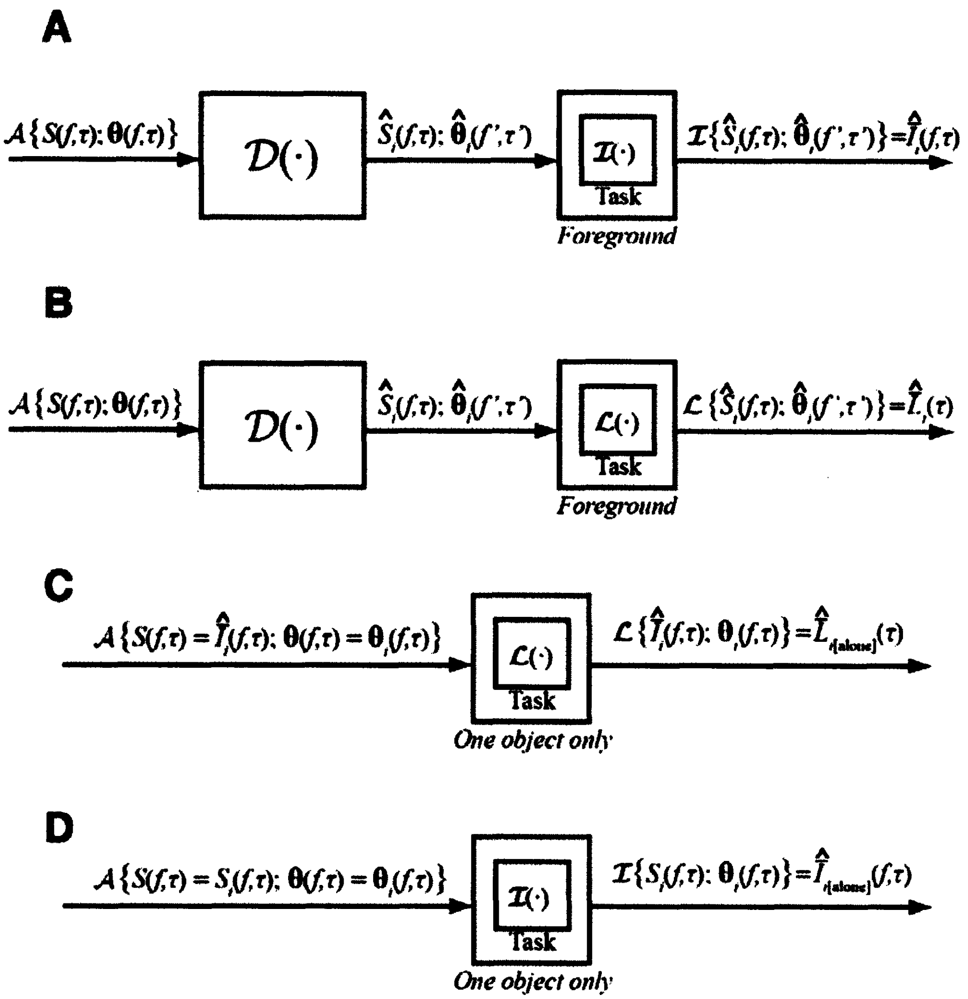

A
$\mathcal{A}\{S(f,\tau); \theta(f,\tau)\}$
$\mathcal{D}(\cdot)$
$\hat{S}_i(f,\tau); \hat{\theta}_i(f',\tau')$
$\mathcal{I}(\cdot)$
Task
Foreground
$\mathcal{I}\{\hat{S}_i(f,\tau); \hat{\theta}_i(f',\tau')\}=\hat{I}_i(f,\tau)$
B
$\mathcal{A}\{S(f,\tau); \theta(f,\tau)\}$
$\mathcal{D}(\cdot)$
$\hat{S}_i(f,\tau); \hat{\theta}_i(f',\tau')$
$\mathcal{L}(\cdot)$
Task
Foreground
$\mathcal{L}\{\hat{S}_i(f,\tau); \hat{\theta}_i(f',\tau')\}=\hat{L}_i(\tau)$
C
$\mathcal{A}\{S(f,\tau)=\hat{I}_i(f,\tau); \theta(f,\tau)=\theta_i(f,\tau)\}$
$\mathcal{L}(\cdot)$
Task
One object only
$\mathcal{L}\{\hat{I}_i(f,\tau); \theta_i(f,\tau)\}=\hat{L}_{i[\text{alone}]}(\tau)$
D
$\mathcal{A}\{S(f,\tau)=S_i(f,\tau); \theta(f,\tau)=\theta_i(f,\tau)\}$
$\mathcal{I}(\cdot)$
Task
One object only
$\mathcal{I}\{S_i(f,\tau); \theta_i(f,\tau)\}=\hat{I}_{i[\text{alone}]}(f,\tau)$

Figure 6-3 (A): Reporting the intensity estimate of the spectro-temporal profile of the i^{th} source in a mixture. **(B)**: Reporting the location estimate of the i^{th} source in a mixture. **(C)**: Reporting the location of the spectro-temporal profile reported in **(A)** presented alone. **(D)**: Reporting the intensity estimate with the i^{th} source presented alone.

The above experiments implicitly assumed that the spectro-temporal resolution of the location estimate is equivalent to that of the intensity estimate in each source, i.e., $\hat{\theta}_i(f',\tau')=\hat{\theta}_i(f,\tau)$ for each f and τ, based on the estimate of $\hat{S}_i(f,\tau)$. Another hypothesis to reconcile the inconsistent "what" and "where"· readout is that the resolution for the spatial estimate, $\hat{\theta}(f',\tau')$, is coarser than that of the spectro-temporal intensity estimate, $\hat{S}(f,\tau)$, and thus the localization task is more likely to have "cross-talks" with other objects in the spatial estimates than for the intensity estimate. The frequency separation between harmonic components with different spatial attributes can be systematically varied to test this hypothesis. If the frequency separation between the spectral elements can be resolved both by the monaural, $\mathcal{M}(\cdot)$, and the binaural, $\mathcal{B}(\cdot)$, systems, then the intensity and the location estimates will be consistent if the difference of the resolution between the two systems is the only factor that causes the inconsistent "what" / "where" read-outs. However, this hypothesis also predicts a critical region of frequency spacing in which only the monaural system, but not the binaural system, can resolve the frequency components, i.e., the "what" / "where" read-outs are predicted to be inconsistent for some intermediate spectro-temporal spacing of the stimuli. Future experiments that manipulate the spectro-temporal spacing between objects presented in a scene may shed lights into the validity of this resolution hypothesis.

6.2.7 Extreme views of scene analysis

Abstractly, the structure of the decomposing process can be characterized in many different ways. In one extreme, since the ultimate question is to characterize how we parse a mixture of signals available at the receiver into distinct perceptual streams, we can model the decomposing system in the most generic sense, incorporating all signal transformations regardless whether it is stimulus-driven (i.e., purely based on pre-attentive peripheral processing) or other processes that require attentional controls (Figure 6-4, shaded region). At the other extreme, and this is perhaps closer to the stand that the general auditory community holds, we might discount any effects that can be attributed to peripheral origins or other stimulus-dependent effects to be relevant to the characterization of the scene analysis processes. Bregman [3] suggested a distinction between stream segregation processes that depend only on the stimulus-driven characteristics, which he termed "primitive processes" and processes that require top-down controls, or "schema-based processes".

Perhaps in terms of the construction of a conceptual framework, it is a reasonable compromise of the two extremes presented.

Buschman *et al.* [167] recently showed that the physiological responses underlying the bottom-up, stimulus-driven processes during a visual "pop-out" experiment recorded in monkeys (*macaca mulata*) have different synchronization characteristics between the frontal and parietal areas as compared to a visual "search" experiment which requires volitional shifts of attention. Results of the psychoacoustical experiments described in this dissertation can implicate whether top-down attention is involved in the decomposing process, but they could not shed lights into distinguishing between the segregation process due to bottom-up and top-down controls. However, given the current visual physiological evidence, as well as the much adopted view proposed by Bregman in the distinction between "primitive" and "schema-based" segregation processes, let us introduce two cascaded systems to replace the $\mathcal{D}(\cdot)$ operator. Let $\mathcal{D}_{\mathrm{BUO}}(\cdot)$ denote to the portion of the segregation process that, by definition, depends only on the bottom-up stimulus-driven information, and let $\mathcal{D}_{\mathrm{TDA}}(\cdot)$ denote the remaining portion of the segregation process that depends on top-down attention. Figure 6-4 summarizes the scene analysis process as presented in this theoretical framework, incorporating the distinction of "primitive" and "schema-based" segregation as suggested by Bregman.

This compartmentalized view of the scene analysis process may be appealing since, by our current definition, there is a clear distinction between the pre-attentive and attentive influences. However, in the visual modality, it has been suggested that attention modulates the response along the visual pathway (e.g., [80-82] including at the peripheral level [83]). In audition, there is also evidence suggesting that spectro-temporal receptive fields in the primary auditory cortex of behaving ferrets [84, 85] change depending on the behavioral task. Similarly attention has been implicated in corticofugal modulation of cochlear function in awake mustached bats during vocalization [86]. If the behavioral task modulates the spectro-temporal response at the periphery (Figure 6-4, dotted path), the above compartmentalized view would not be valid. Nonetheless, it is hopeful that the conceptual framework presented here will provide the platform such that these difficult questions can be tackled in the future.

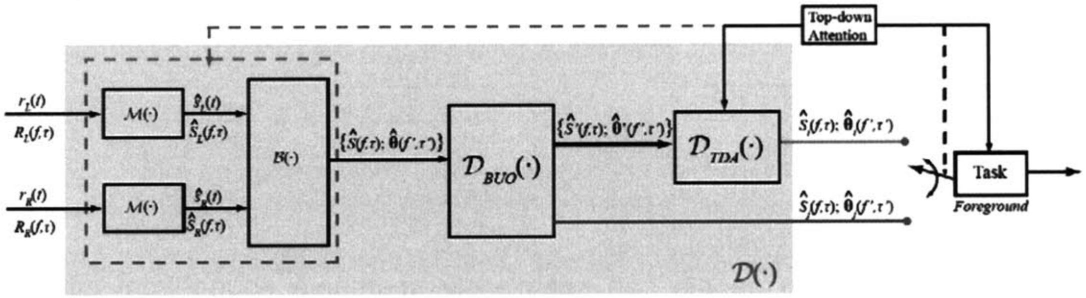

$r_L(t)$
$R_L(f,\tau)$
$r_R(t)$
$R_R(f,\tau)$
$\mathcal{M}(\cdot)$
$\mathcal{M}(\cdot)$
$\hat{s}_L(t)$
$\hat{S}_L(f,\tau)$
$\hat{s}_R(t)$
$\hat{S}_R(f,\tau)$
$\mathcal{B}(\cdot)$
$\{\hat{S}(f,\tau); \hat{\theta}(f',\tau')\}$
$\mathcal{D}_{BUO}(\cdot)$
$\{\hat{S}'(f,\tau); \hat{\theta}'(f',\tau')\}$
$\mathcal{D}_{TDA}(\cdot)$
Top-down Attention
$\hat{S}_i(f,\tau); \hat{\theta}_i(f',\tau')$
$\hat{S}_j(f,\tau); \hat{\theta}_j(f',\tau')$
Task
Foreground
$\mathcal{D}(\cdot)$

Figure 6-4: Conceptual model summary with $\mathcal{D}(\cdot)$ expressed as a cascade of two systems that are influenced by "bottom-up-only" factors and "top-down-attention." Dotted line denotes the possible centrifugal projections that modulating signals at the periphery through top-down attention. Shaded area denotes an extreme view of the decomposing system, encompassing all operators discussed.

6.2.8 Perceptual organization as a function of time

It has been well documented that, for some ambiguous stimuli, perceptual organization tends to change over time [24, 25]. Usually, such stimuli are first heard as one stream, but as time evolves, listeners become more likely to hear two separate streams of tones. Attention switches have been posited to reset this buildup of streaming [26, 27], however, the role of attention in the formation of auditory streams is still under debate [28]. The description of the conceptual framework presented here focuses on the perceptual organization of the task-relevant stream (i.e., subject is either identifying or localizing the attended object), aligned with the focus of the experiments described in this dissertation. However, the conceptual framework can easily be extended to account for the perceptual organization of streams that are not in the attentional foreground.

An exciting line of research that has been active in recent years is the investigation of whether attention is required for the maintenance of a stream or it is only required at the initial stage of streaming. In order to ascertain the perceptual organization that is not in the attentional foreground, researchers have drawn inferences on the performance of serial recall on visually presented items using the "irrelevant sound effect" [168] or interpreting the mismatch negativity (MMN) component of event-related brain potentials (ERPs) elucidated on unattended sounds [28, 169-171]. It is hopeful that this conceptual framework can eventually be adopted to describe the perceptual organization beyond that is in the attentional foreground.

6.3 Final remarks

6.3.1 Dare to dream

While normal hearing people can seamlessly segregate sound from different sources, it presents a fundamental challenge for the hearing impaired, hearing aid design and speech recognition algorithms. Engineers have traditionally attacked this problem by attempting to improve signal-to-noise ratio. However, there is no "noise" in our complex acoustical environment but only acoustical signals carrying information about our surroundings to which we can voluntarily switch our attention.

One of the messages from this dissertation is that auditory scene analysis is an attentive process. From signal analysis point of view, top-down process provides feedback to make it a closed-loop system. In current hearing assistive devices, the system design is primarily open-loop, i.e., the device user has no control of the signal at the input. If one day we can fully understand how event related potentials and / or magnetic fields are linked to different attentional states [172], perhaps it is not too far fetched to dream of a closed-loop hearing assistive system that tracks the attentional state of the user and thereby helping the user to better parse the auditory scene by some pre-processing CASA algorithms.

6.3.2 Optimistic outlook

In vision, classical theories such as Gestalt psychology have taken backstage since the dawn of Marr's [11] computational approach to visual scene analysis. In the introduction of Marr's influential treatise on computational vision, he declared that the Gestalten approach failed and eventually "dissolved into the fog of subjectivism." Marr further remarked that it was their "mathematical ignorance... and a failure to think more in terms of processes ... [which] led to the failure of a school of thought that had actually made a number of valuable insights." Although recent attempts have tried to bridge between the original Gestalt work and modern information processing work [173], both the old and the new worlds seem to have already traveled far down their respective divergent paths.

In audition, however, there seems to be a happier union between the Gestalten and the computational schools, mostly due to Bregman's [3] seminal book which inspired both the psychophysicists and the computer scientists. In the introductory chapter of the most recent publication on computational auditory scene analysis (CASA), the editors [174] described this new emerging field by comparing it to Marr's vision, and at the same time citing the Gestalt law of Prägnanz [10] as applied to audition. CASA systems vary greatly in their architecture, but many draw inspirations from the experimentalists (since, after all, humans still currently out perform machines in achieving the task of ASA). It is hopeful that the findings in this dissertation and the conceptual framework in its eventual form will one day inspire and improve CASA algorithms in separating multiple sound mixtures displaced in space.

If psychoacousticians continue to be analytically oriented while the CASA researchers continue to be interested in biologically-inspired algorithms, maybe it is not too far into the future that research addressed in this dissertation will benefit the design of assistive hearing devices and in turn help millions to finally understand speech in cocktail party environments.

APPENDIX AUDITORY NON-ALLOCATION

The work described in this chapter has been published in *Proceedings of the National Academy of Sciences of the United States of America*.

Shinn-Cunningham, B.G., Lee, A.K.C. and Oxenham, A.J (**2007**). A sound element gets lost in perceptual competition. *Proc. Nat. Acad. Sci.* **104**:12223-12227.

Appendix A.1 Abstract

Our ability to understand auditory signals depends on properly separating the mixture of sound arriving from multiple sources. Sound elements tend to belong to only one object at a time, consistent with the principle of disjoint allocation, although there are instances of duplex perception or co-allocation, in which two sound objects share one sound element. Here we report a novel effect of "non-allocation," where a sound element "disappears" when two ongoing objects compete for its ownership. When a target tone is presented either as one of a sequence of tones or simultaneously with a harmonic vowel complex, it is heard as part of the corresponding object. However, depending on the spatial configuration of the scene, if the target, the tones, and the vowel are all presented together, the target may not be perceived in either the tones or the vowel, even though it is not perceived as a separate entity. The finding suggests an asymmetry in the strength of the perceptual evidence required to reject versus to include an element within the auditory foreground, a result with important implications for how we process complex auditory scenes containing ambiguous information.

Appendix A.2 Text

Many species, including birds [175], frogs [176], and mammals [1], must hear out important communication calls from a background of competing sounds in order to procreate and survive. Whether in a raucous penguin colony in Antarctica [177] or a crowded cocktail party in Europe [178], listeners are adept at analyzing the acoustic mixture to determine what sound sources are present.

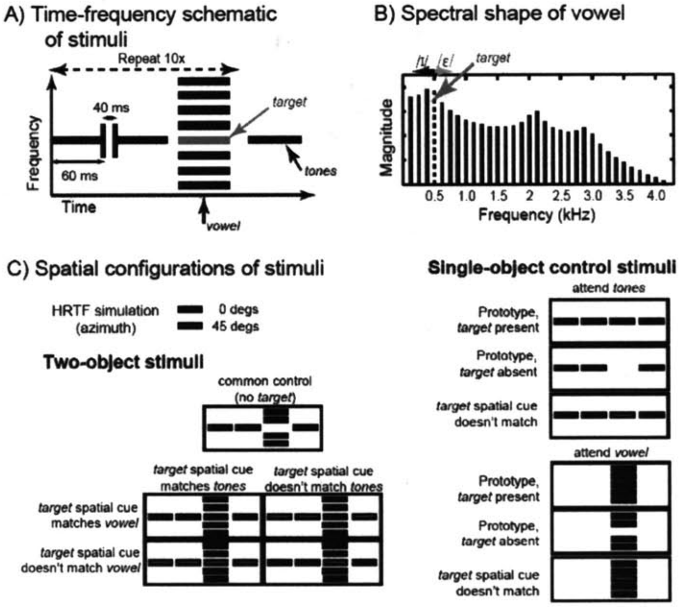
Successful sound source identification requires that the individual sound elements within a mixture be assigned to the correct "auditory objects." Many spectro-temporal features in the sound mixture promote grouping of sound elements into auditory objects, including common onsets, common amplitude modulation, harmonicity, continuity over time, frequency proximity, and common spatial cues such as interaural time differences [3, 179]. Listener experience and expectations can also influence how the scene is analyzed, suggesting that "top-down" processes interact with low-level "bottom-up" stimulus features in auditory object formation [36, 46, 47].

The perceptual grouping principle of exclusive or disjoint allocation states that a single sound element, such as a pure tone, cannot be assigned simultaneously to more than one auditory object [3]. Although this principle has fairly general applicability, there are some exceptions. For instance, a frequency glide, presented to the opposite ear from the rest of a speech sound, can influence the perceived phonetic content of the speech sound while at the same time being heard as a separate object [180]. Similarly, a mistuned harmonic within a harmonic complex tone can be heard as a separate tone while at the same time influencing the perceived pitch of the overall complex tone [39]. These situations, where a sound element contributes to more than one auditory object, are examples of duplex perception.

The term duplex perception suggests that a single sound element can be assigned independently to more than one object. However, a more parsimonious explanation may be that, in fact, the energy of the element can simply be shared between sound objects. Physically, if a frequency component is present in two independent sound sources then, on average, the total energy of that frequency in the mixture should equal the sum of the energies in the constituent sound sources. Thus, a veridical perceptual representation would divide the total sound energy of each frequency component across the two objects. Although many past studies have considered the question of trading, few have explicitly measured the perceptual contribution of the target to both competing objects [48, 49, 53, 59].

Here we adapt an earlier paradigm [53] in order to assess directly the relative contribution of a pure-tone element to each object. This allows us to quantify the degree to which perceptual trading of energy holds when two objects compete for a sound element. We generated rhythmically repeating stimuli consisting of two auditory objects: a sequence of rapidly repeating *tones* and a synthetic *vowel*, repeating at a slower rate (see Figure A-1A). In this mixture, an ambiguous tone, known as the target, could logically be a member of each (or both) of the two perceived objects, either as another tone in the repeating sequence of *tones* or as the fourth harmonic in the *vowel.* Importantly, the formation of the *tones* stream depends primarily on perceptual organization across time, where spatial cues are known to have a large influence [7], whereas the organization of the *vowel* depends primarily on a local spectro-temporal structure, where spatial cues should have a weaker effect [3].

We manipulated the spatial cues of the sound elements in the mixture and measured how perceptual organization was affected. Identical stimuli were presented in two separate blocks, one in which listeners attended to the *tones* and one in which they attended to the *vowel.* In each block, we measured whether the target was perceived as part of the attended object. When attending to the *tones,* listeners identified the rhythm of the stream. If the target was perceived as part of the *tones* stream, the perceived rhythm was even; otherwise it was galloping. Similarly, listeners identified the perceived *vowel* category, which depended on whether or not the target was perceived as part of the harmonic complex [36, 181]. If the target was heard as part of the *vowel*, it was labeled */ɛ/* (as in "bet"); when the target was not part of the *vowel,* it was labeled */ɪ/* (as in "bit;" see Figure A-1B and Methods). The spatial cues in the target could either match or differ from the spatial cues in the *tones* and the *vowel* (see left half of Figure A-1C and Methods) to either promote or discourage grouping of the target with the attended object. Intermingled control conditions presented single-object prototype stimuli in which only the attended object was presented (with and without the target; see right half of Figure A-1C), to confirm that listeners were able to consistently label unambiguous stimuli properly with the even/galloping or ε/ɪ labels. Finally, another control condition presented the two competing objects without any target to ensure that attended objects in this two-object condition were not perceived as if they contained target energy.



A) Time-frequency schematic of stimuli
Repeat 10x
40 ms
60 ms
target
tones
Frequency
Time
vowel
B) Spectral shape of vowel
/ɪ/
/ε/
target
Magnitude
0.5 1.0 1.5 2 2.5 3.0 3.5 4.0
Frequency (kHz)
C) Spatial configurations of stimuli
HRTF simulation (azimuth)
0 degs
45 degs
Two-object stimuli
common control (no target)
target spatial cue matches tones
target spatial cue doesn't match tones
target spatial cue matches vowel
target spatial cue doesn't match vowel
Single-object control stimuli
attend tones
Prototype, target present
Prototype, target absent
target spatial cue doesn't match
attend vowel
Prototype, target present
Prototype, target absent
target spatial cue doesn't match

Figure A-1: Stimuli and conditions. **(A)** Stimuli consist of a 3-s-long repeating sequence of 500-Hz *tones*, a synthetic *vowel* with fundamental frequency 125 Hz, and a 500-Hz target that logically could group with either object. **(B)** The spectrum of the synthetic *vowel* was shaped to sound like /I/. If the target tone was perceived as part of the vowel, the perceived *vowel* category shifted towards /ε/. **(C)** The simulated spatial locations of the target and repeating *tones* varied across conditions. Control conditions included a case without a target and conditions in which there was only one attended object.

To obtain a more direct interpretation of the category responses in the main experiment, an auxiliary experiment was conducted with single-object stimuli. The same categorical judgments ("even" vs. "galloping" or /ε/ vs. /ɪ/) were measured when only one object was present (either the *tones* or the *vowel*) and the level of the target was varied systematically, with attenuations ranging from 0 dB (no change) to 14 dB. This allowed us to quantify how the response probabilities for the two categories (target present and target absent) mapped to the physical energy present in the target. In turn, the individual-subject results from this auxiliary experiment enabled us to map the response percentages in the main experiment to an "effective target energy" (the level of the target, in dB, that would lead to the observed response percentages in the auxiliary, single-object experiments).

For both the *tones* and *vowel* single-object prototype control stimuli in the main experiment, subjects responded as if the target was part of the object, whether or not the target location matched that of the attended object. Thus, mismatching target and object spatial cues are insufficient to perceptually remove the target from either the *tones* or the *vowel* object if there is no perceptual competition for the target.

The percentage of “even” vs. “galloping” and /ɛ/ vs. /ɪ/ responses in the different two-object conditions were compared to the corresponding percentages for the prototype control stimuli in the main experiment (either the *tones* or *vowels* in isolation) with and without the target present. We extended traditional signal detection theory approaches to determine whether the responses to an ambiguous, two-source stimulus were closer to responses for a single-source stimulus with or without the target present (see Methods). Figure 2 uses the relative perceptual distance between the ambiguous stimuli and the target-absent prototype to summarize responses. A value of 0 indicates that response percentages for a particular condition equaled the response percentages for the target-absent prototype. A value of 1 indicates the stimulus was perceived like a spatially unambiguous target-present prototype (i.e., a single-object stimulus in which the target location matched that of the attended *vowel* or *tones*; see Figure A-2 and Methods).

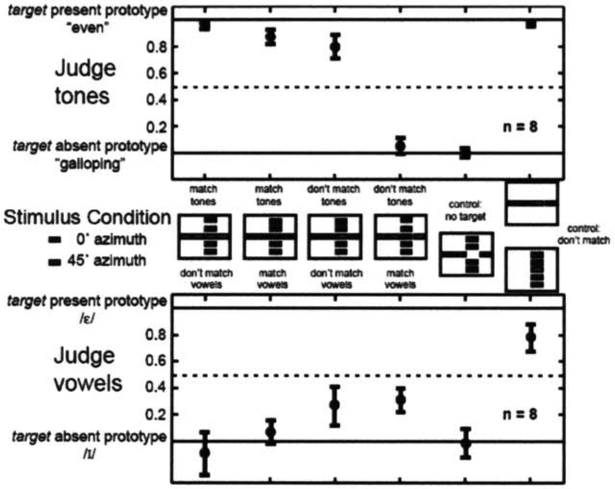


Figure A-2: Spatial cues have a large effect on perception of the tones but very little effect on perception of the vowel. Across-subject averages of the relative perceptual distance of each condition to single-object prototypes are show (error bars show the standard error of the mean).

When asked to judge the rhythm of the *tones* in the presence of the *vowels*, the spatial cues had a large effect, in line with earlier studies showing a large influence of spatial cues when grouping sounds across time [7, 96]. The target contributed strongly to the *tones* sequence whenever the simulated target location matched that of the *tones*, regardless of the *vowel* location. When the target location matched neither that of the *tones* nor of the *vowel*, subjects still perceived the target as part of the *tones* sequence. However, when the target location matched that of the *vowel* but not the *tones*, the target was no longer perceived as part of the *tones* sequence, and listeners heard a galloping rhythm. Thus, spatial cues are strong enough to overcome other grouping cues and perceptually remove the target from the attended *tones* stream. However, this only occurs in the most extreme case when the spatial cues in the target match those in the background *vowel* and do *not* match those in the attended foreground *tones*.

When asked to identify the perceived *vowel* in mixtures containing competing *tones*, spatial cues had a much less pronounced effect. Moreover, listeners never heard the target as strongly present in the *vowel* (they consistently responded /ɪ/

more often than /ɛ/, and the response percentages were much more similar to the response percentages for the /ɪ/ prototype than for the /ɛ/ prototype). Even when the spatial cues in the target matched those in the *vowel* and not the *tones*, the attended *vowel* was perceptually more like the /ɪ/ (without the target) than /ɛ/ (with the target; see fourth condition from the left in the bottom of Figure A-2).

The results so far suggest that the contribution of the target to the *tones* does not predict its contribution to the *vowel*, in apparent contradiction to the trading hypothesis outlined in the introduction. However, if the perceptual judgments of "galloping"-"even" and /ɪ/-/ɛ/ have a different dependence on the target level (e.g., if the category boundary is steep as a function of target level for one judgment and shallow for the other), the response category percentages in the two tasks will not trade quantitatively, even if the energy-trading hypothesis holds. The auxiliary experiment allowed us to quantify the degree to which the results in the first experiment obeyed the trading hypothesis. Using results from the auxiliary experiment, we calculated the effective intensity of the target corresponding to the raw response percentages (i.e., "even"-"galloping" response percentage for the *tones* or /ɛ/-/ɪ/ response percentage for the *vowel*) in the main experiment. The percentage of trials in the auxiliary experiment in which listeners responded "even" vs. "galloping" (in the *tones* condition) or /ɛ/ vs. /ɪ/ (in the *vowel* condition) provided a subject-specific mapping between target attenuation and a corresponding response percentage. These psychometric functions relating response percentages to the physical attenuation of the target were generally well-behaved; increasing target attenuation systematically increased the probability that the listeners responded as if the target absent from the attended object (see Figure A-3A and 3B for example psychometric functions from the *tones* and *vowels* control experiments, respectively).

For each subject and stimulus condition, we mapped the percent responses obtained in the main experiment to an "effective target attenuation" and compared the resulting effective attenuations for physically identical stimuli in the attend-*tones* and attend-*vowel* blocks (see Figure A-3, where panels A and B demonstrate how the effective attenuations of the target are obtained for one subject, attending both *tones* and *vowels,* for one example condition; these values are projected to the ordinate and abscissa in panel C, respectively, producing the open grey triangle).

A pure energy trading relationship for the target's contribution to the *vowel* and the *tones* would produce data points that lie along the solid thick curve in Figure A-3C. Data would fall along the dashed line if an amplitude-trading relationship holds [53, 59]. If the *tones* caused an effective attenuation of the target that reduced its contribution to the *vowels* [182, 183] through, for instance, neural adaptation, the data would fall along one of the family of curves shown by the thin lines in Figure A-3C (see Appendix A.3). None of these predictions can fully account for our results.

When the spatial location of the target matched that of the *tones* but not the *vowel* (Figure A-3, filled triangle), the target was perceived almost exclusively as part of the *tones* sequence, in line with expectations based on energy trading. When the spatial location of the target matched that of both the *tones* and the *vowel* (filled circle), or matched neither (open triangle), there was a tendency to assign more of the target to the *vowel* and less to the *tones* (i.e., the effective target attenuation decreases in the *vowel* task and increases in the *tones* task). Results for these two conditions can be fit well by assuming that the *tones* cause adaptation that effectively reduces the target level by about 4 dB. However, when the spatial location of the target matched that of the *vowel* but not the *tones* (Figure A-3C, open circle), the effective level of the target was attenuated by 9 dB or more both when the listeners attended to the *tones* and when they attended to the *vowel*. In fact, in both tasks, responses were similar to the responses to the control condition in which the target was physically absent (compare open circle and cross in Figure A-3C).

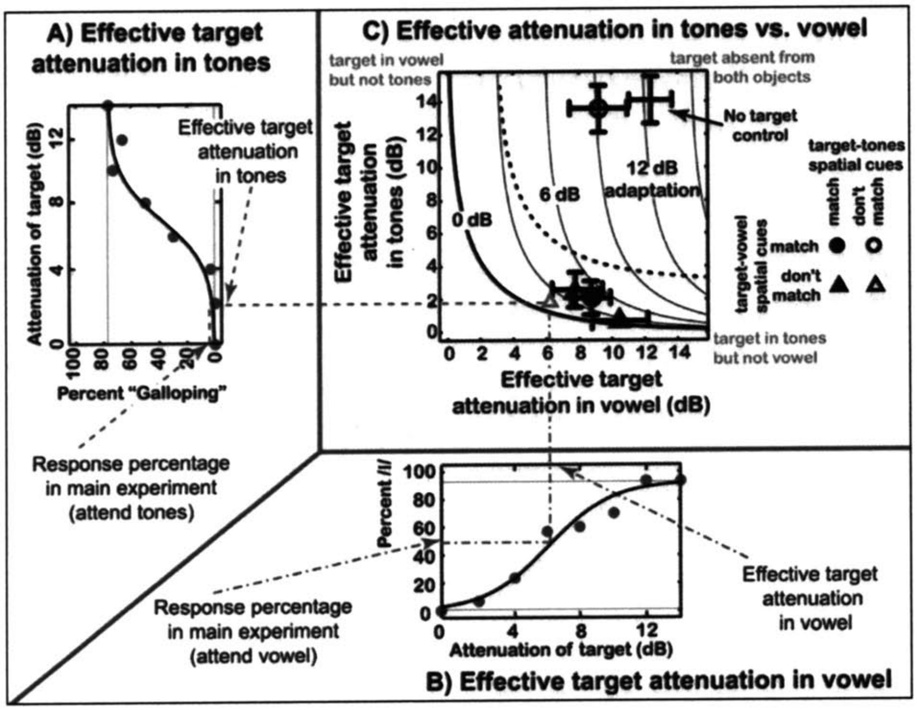


Figure A-3: The perceived target energy in the two objects does not always account for all of the physical target energy present in the mixture. **(A, B)** Example psychometric functions map the response percentages from the main experiment to effective target attenuations for the tones **(A)** and the vowel **(B)**. These mapping are used to derive data plotted in C. **(C)** Effective target attenuation when listeners attend to the *tones* versus when they attend to the *vowel*. Across-subject means are plotted, with error bars showing the standard error of the mean. The solid thick line shows the expected trading relationship when the perceived energy accounts for the physical target energy. Thin lines show the trading relationships that would occur if peripheral adaptation reduces the contribution of the target to the *vowel* (shown for 3 dB increments). The dashed line shows the predictions if target amplitude, rather then energy, trades.

Appendix A.3 Discussion

We find that in situations of perceptual competition, the perceptual coherence between a sound element (the target) and one object (the *vowel*) can be sufficient to prevent the target from binding with another object (the *tones*) but still insufficient to bind the element with the first object. In one of our conditions, this results in the target element falling into a form of perceptual limbo, where it belongs to neither competing object. The finding provides an interesting counterpart to duplex perception, or co-allocation, whereby a single sound element contributes to two perceptual objects at once [39, 184]. In contrast, we observe non-allocation, whereby the element does not strongly contribute to either object and is also not heard as an independent object. It is important to note, however, that the target is not undetectable: subjects can easily distinguish between sequences that contain the target and those that do not, even when the target fails to bind with either *tones* or *vowel*. We next consider some possible explanations for this effect.

Certain forms of neural adaptation may contribute to our results. If the preceding sequence of *tones* reduces the target's "internal" level, the target contribution to the *vowel* will be reduced. However, because all of the *tones* would be perceived at the same level as the adapted target, the contribution of the target to the *tones* would be unaffected. Thus, if adaptation were the only effect present, a skewed form of energy trade would occur and results would fall along one of the thin lines in Figure A-3C. Data for three conditions are consistent with peripheral adaptation reducing the effective target level by about 4 dB. However, this level of adaptation cannot account for the condition in which the target spatial cues match those of the *vowel* but not the *tones*. Moreover, many studies have shown that the loudness of targets is not reduced by preceding tones of the same intensity (like those in our experiment), making it unlikely that our results are due solely to an internal attenuation of the target [185-187]. In addition, earlier similar studies have also concluded that adaptation cannot account for the effects of a preceding sequential stream on perception [44]. However, to address the issue more directly, we undertook a supplemental control experiment (see Supporting Information).

In the supplemental experiment, the *vowel* of the main experiment was replaced by a harmonic complex with an F0 of 200 Hz (henceforth, the *simultaneous complex*) and the target was itself a harmonic complex with a fundamental frequency of 300 Hz (see Methods in Supporting Information). When the *simultaneous complex* and target are presented together in quiet, a single harmonic complex with an F0 of 100 Hz and a dense spectral profile is heard. As in the main experiment, when the target is preceded by an isochronous pair of matching 300-Hz complexes (the *complex stream*, replacing the *tones* of the main experiment), the contribution of the target to the *simultaneous complex* decreases by an amount that depends on spatial cues. However, unlike in the main experiment, the target contributes significantly to the perceived spectral content of the *simultaneous complex* in all conditions (see Appendix A.6, Fig. 4B), presumably because across-frequency grouping cues are stronger for these stimuli than for a single-frequency target. The fact that the effective attenuation of the target in the *simultaneous complex* is near zero in many conditions in the supplemental experiment suggests that there is not obligatory adaptation of the target response for stimuli repeating at the rates and levels used in our experiments.

Another possible explanation relates to auditory spatial processing. Rapid changes in location can result in a diffuse spatial percept, attributed to "binaural sluggishness" in spatial processing [188, 189]. This raises the possibility that when target and *tones* have different spatial cues, the target is spatially diffuse. This, in turn, could cause the target to contribute relatively little to the perceived content of the *vowel* and help explain why trading hypotheses fail. Again, however, no such effect is observed in the supplemental experiment, even though the spatial cues change dynamically at the same rate as in the main experiment. Thus, there is no evidence that the perceptual contribution of the target is reduced because it is spatially diffuse.

To explain our finding of non-allocation, we suggest that the auditory system favors efficient processing over veridical representation of the entire auditory scene. In particular, the perceptual organization of the auditory background (here, the unattended object) may not be as fully elaborated as that of the foreground [27]. This implies that sound elements that are rejected from the auditory foreground are not necessarily assigned to auditory objects within the unattended background. Interpreted in this way, perceptual non-allocation may reflect a figure-ground asymmetry, with stronger perceptual cues necessary to pull an element into the auditory foreground than are needed to push the same element into the (unattended) background.

Our results cannot answer the question of whether the target was part of the unattended object in the background or whether it was isolated in some "perceptual limbo." In informal listening, when listeners attempted to attend to both objects at once, they perceived no salient change in the perceived organization compared to when they actively attended to the *tones* or *vowel*. However, it was difficult to attend to both objects simultaneously; listeners felt that they rapidly shifted attention from object to object rather than simultaneously attended to both objects [89]. From these reports, we cannot rule out the possibility that perceptual organization in our experiment is bistable, changing so that the target is in the background whenever attention shifts between objects.

In many everyday acoustic settings, competition for attention between auditory objects may be the most important problem facing a listener [90, 190]. In vision, this problem has long been recognized, and theories of how stimulus attributes interact with top-down processes to mediate competition for attention are well developed [97, 98]. Similar mechanisms may work to resolve competition for attention in complex auditory environments [191, 192]. In both vision and audition, attention appears to operate on perceived objects, rather than simple features of the visual or acoustic scene [97, 191, 193]. This suggests that the ability to direct attention in a complex, multi-source auditory scene is directly affected by the way in which objects are formed. Past work demonstrates that both bottom-up factors and top-down attention influence the perceptual organization of sound [3]. The current results hint that the ultimate interpretation of the acoustic scene may depend on what object a listener attends, just as attention can alter perception of objects in a visual scene [83]. The organization of the scene in turn impacts how well the listener can reduce interference from unwanted objects and understand an attended object. The current results show that spatial cues can affect the perceptual organization of ambiguous sound mixtures, which can then cause the interesting phenomenon in which not all of the physical energy in a sound mixture is allocated to the identifiable objects.

Appendix A.4 Methods

Appendix A.4.1 Stimuli

Stimuli consisted of a 3-s long sequence, composed of ten identical presentations of three 100-ms-long elements: two 500-Hz tone bursts (*tones*) followed by a synthetic *vowel* with fundamental frequency of 125 Hz (see Figure A-1A). The target was a 500-Hz tone presented simultaneously with the *vowel.* All *tones,* target, and the harmonics of the *vowel* were gated with a Blackman window (60-ms duration), followed by a silent gap of 40 ms. The sequence of repeating *tones* and *vowel* caused a percept of two distinct auditory objects (rapidly repeating *tones* and a slower sequence of repeating *vowels*).

The *vowel* consisted of individual random-phase harmonics of the fundamental frequency 125 Hz, spectrally shaped like the vowel /ɪ/ (formant peaks at frequencies 490, 2125, and 2825 Hz; see Figure A-1B). The *vowel* did not contain any energy in the fourth harmonic, the frequency of the target. When the target was present and perceived as part of the *vowel*, the perceived *vowel* quality shifted from /ɪ/ (target absent, or not part of the *vowel*) towards /ɛ/ (target heard as part of the vowel; [36, 46, 47], presumably by shifting the perceived frequency of the first formant peak.

Spatial cues in the *tones* and target were controlled by processing the sounds with head-related transfer functions (HRTFs) measured on a mannekin [94]. This processing simulates the interaural time and level differences and spectral cues that would arise for sources from a particular location relative to the listener. Sources were processed to have spatial cues consistent with a source either from straight ahead (azimuth = 0 deg) or 45 deg to the right of the listener. In all trials, the simulated *vowel* azimuth was zero. Four different spatial configurations were tested, differing in which component's spatial cues matched those of the target (see Figure A-1C). Various control trials ensured that we only included listeners who could reliably identify the *tones* rhythm or the *vowel* identity for unambiguous single-object stimuli (see Figure A-1C). In two-object control trials, there was no target and both *tones* and *vowel* were simulated from straight ahead. In single-object control trials, the attended object was simulated from straight ahead; the target was simulated from either 0 or 45 deg azimuth, or was not present.

Appendix A.4.2 Procedures

In the main experiment, two-object stimuli (with single-object controls intermingled) were presented in two blocks of trials differing only in the instructions to the subjects. In *tone* blocks, subjects identified the perceived *tones* rhythm as "even" (an evenly spaced sequence of 500-Hz tones, one every 100 ms) or "galloping" (a pair of tones 100 ms apart, followed by a 100 ms silent gap). In *vowel* blocks, subjects identified the perceived *vowel* identity as /ε/ or /ɪ/. Thirty trials of each condition were presented in a different random order for each block of trials.

In the single-object control experiment, trials consisted of the attended object and the target, both simulated from straight ahead (0 deg azimuth). On each trial, the target was attenuated by a random amount ranging from 0 to 14 dB, in 2 dB steps. As in the main experiment, in separate blocks subjects judged either the rhythm of the *tones* or the identity of the *vowel*.

Appendix A.4.3 Analysis

The data from the main experiment were analyzed using a decision theory model. The internal decision variable was assumed to be a uni-dimensional, Gaussian-distributed random variable whose mean depended on the stimulus and whose variance was independent of the stimulus. A single criterion value was assumed to divide the decision space into two regions, corresponding to "target present" or "target absent" responses. The probability of responding "target present" was calculated for each condition, then used to estimate the distances between the underlying means of the corresponding conditional probability density functions and the mean of the distribution for the target-absent prototype, in units of standard deviation (d-prime). These d-prime measures were normalized by the d-prime separation between the target-present and target-absent prototypes to estimate the relative perceptual distance between the condition and the single-object prototypes. By definition, the resulting statistic was zero for the target-absent prototype and one for the spatially unambiguous, target-present prototype. The across-subject means and standard errors of these relative perceptual distances were computed for each stimulus and are presented in Figure A-2.

In the single-object control study, the percent responses consistent with the "target present" generally decreased monotonically with increasing attenuation of the target. These functions were fit with a sigmoidal function with free parameters of slope, threshold, and upper and lower asymptotes. The fitted curves were used to map the raw percentage of responses for each stimulus to an effective attenuation of the target in the main experiment (see Figure A-3A and 3B). If the response percentage for a given condition was less than the lower asymptote or greater than the upper asymptote of the psychometric function fit to the auxiliary results, the effective attenuation was set to 0 dB or 16 dB (respectively).

Eight subjects were selected based on their ability to reliably distinguish between the single-object prototypes in a similar prior experiment. In the prior experiment, subjects had to achieve both 1) a d-prime of 0.7 or greater between the target-present and target-absent prototypes in the main experiment, and 2) a slope of 10 percent-correct / dB attenuation to the fit of their responses in the single-object control experiment. All naïve subjects met the criteria for the *tones* stimuli in the prior experiment. The eight current subjects were recruited from the 10 out of 20 naïve subjects who reached the performance criterion for the *vowel* control stimuli. Seven of the eight subjects had greater d-prime values here than in the previous experiment, presumably from experience with the task. In the current experiment, all eight subjects achieved d-prime scores of 1.5 or better on both *tones* and *vowel* tasks.

Appendix A.5 Acknowledgements

This work was supported by grants from the Office of Naval Research (N00014-04-1-0131) and the National Institutes of Health (R01 DC 05778-02) to BGSC, as

well as a grant from the National Institutes of Health to AJO (R01 DC 05216). Steven Babcock assisted with data collection in the Supplemental Experiment.

Appendix A.6 Supporting text

Appendix A.6.1 Supporting analysis: The role of adaptation

In the main experiment, if adaptation occurs and that both target and tones are effectively at a lower level, then there should still be trading between tones and vowel; it should just warp the trading between the effective attenuation in the tones and vowel. This is not what is observed.

Specifically, under the assumption that there is significant adaptation, the "effective attenuation of the target in the vowel" should be relative to the unadapted target level (as the comparison is of the spectral shape, and presumably the vowel components, with 240 ms of silence between them, show little or no adaptation). In contrast, the "effective attenuation of the target in the tones" should be relative to the target level after adaptation (as the comparison is to the other tones, which will show adaptation as well).

Then, if there is adaptation, and the trading hypothesis holds:

$$1 = 10^{\frac{-Att_{tones}}{10}} + 10^{\frac{Att_{adaptation}}{10}} 10^{\frac{-Att_{vowel}}{10}}$$

where

$Att_{adaptation}$ is the effective attenuation of the target (in dB) due to adaptation

Att_{tones} is the target attenuation in the tones (in dB) due to trading with vowel

Att_{tones} is the target attenuation in the vowel (in dB) due to trading with tones

This leads to a different form of trading contour, which depends on the effective amount of the target attenuation due to adaptation, in dB. Figure A-3 shows these trading contours for different assumed levels of adaptation (in 3 dB steps).

The figure shows that the "orphan" does not fall on the same contour as the other data. While the three other points fall between the "adaptation trade" contours for 3 and 6 dB of adaptation, the orphan falls on the 9 dB adaptation curve.

Thus, while adaptation may contribute to the "energy loss" we find, it cannot explain the orphan case. Quantitatively, while adaptation may contribute to the observed results, it cannot account fully for the perceptual loss of the target when the spatial cues of the target match those of the *vowel* and differ from those of the *tones*.

Appendix A.6.2 Supplemental experiment

As in the main experiment, stimuli in the supplemental experiment consist of two objects (here, the *simultaneous complex* and the *complex stream*, taking on the roles of the *vowel* and *tones* of the main experiment, respectively) that compete for ownership of an ambiguous target. In this supplemental experiment, listeners matched the perceived spectral makeup of the *simultaneous complex* and the *complex stream* (see Supporting Methods). As in the main experiment, the *simultaneous complex* is always presented from in front of the listener, while the target and the *complex stream* either came from in front or to the side of the listener.

Figure A-4B shows the mean attenuations of the target (averaged across subjects) that produce perceptual matches to the *simultaneous complex* and *complex stream* spectral content, plotted against each other in a format comparable to that of Figure A-3. As in the main experiment, changes in the spatial configuration of the sound elements alter the perceived spectral content of the objects. When the spatial cues of the target and *complex stream* match and the target and *simultaneous complex* do not match (Figure A-4B, filled triangle), the target contributes a great deal to the *complex stream*, but contributes little to the *simultaneous complex*. When the target location matches both the location of the *simultaneous complex* and of the *complex stream*, the target contributes strongly to the *simultaneous complex* and weakly to the *complex stream* (Figure A-4B, filled circle). When the target location matches neither *simultaneous complex* nor *complex stream* locations, its contribution to the *simultaneous complex* increases and its contribution to the *complex stream* decreases (Figure A-4B, open triangle). Finally, when the target location matches that of the *simultaneous stream* and does not match that of the *complex stream*, it contributes almost nothing to the perceived content of the *complex stream,* but contributes significantly to the perceived content of the *simultaneous stream* (Figure A-4B, open circle).

These results differ from those of the main experiment in a number of ways. First, in this experiment, results closely follow predictions based on an energy-trading hypothesis (solid curve in Figure A-4). Second, the target contributes strongly to the perceived content of the *simultaneous complex* in this experiment, unlike in the main experiment. Finally, there is no "lost target element" in this supplemental experiment: when the target location matches the location of the *simultaneous complex* and does not match the location of the *complex stream*, the target is heard as part of the *simultaneous complex* and does not contribute to the perceived content of the *complex stream.*

The repetition rates of the stimuli in the supplemental experiment were the same as in the main experiment. If the *tones* of the main experiment cause adaptation that reduces the internal level of the target, the *complex stream* in this supplemental experiment should cause similar adaptation of the target and decrease its perceptual contribution to the *simultaneous complex.* Instead, the target contributes strongly to the *simultaneous complex* in most conditions. Therefore, adaptation cannot account for the "lost target" in the main experiment. Similarly, when the target and *complex stream* have different locations, spatial cues change as rapidly here as in the main experiment. While it is possible that these rapid changes produce a spatially diffuse target, the target contributes strongly to the perceived content of the *simultaneous complex* of this supplemental experiment. Thus, binaural sluggishness cannot account for the "lost target" in the main experiment.

We attribute the difference in the pattern of results between the supplemental experiment and the main one to the fact that the target consists of multiple harmonically related components. Because both the target and *complex stream* are multi-tone harmonic complexes, rather than a single-frequency tone (see Figure A-4A), the balance between sequential and simultaneous grouping cues shifts compared to in the main experiment. This leads to stronger simultaneous grouping cues here than in the main experiment, and results in the elimination of the non-allocation effect.

Appendix A.7 Supporting methods

Appendix A.7.1 Stimuli

Stimuli consisted of a 3-s long sequence, composed of ten identical repetitions of three 100-ms-long elements: two harmonic complexes of fundamental frequency 300 Hz (*complex stream*) followed by a *simultaneous complex* with fundamental frequency 200 Hz (see Figure A-4A). The target was a harmonic complex identical to those making up the *complex stream*, but presented simultaneously with the *simultaneous complex*. The target and *complex stream* bursts consisted of harmonics at 300, 600, 900, 1200, and 1500 Hz. The *simultaneous complex* consisted of harmonics 200, 400, 800, 1000, 1400, and 1600 Hz. The *complex stream,* target, and *simultaneous complex* all were gated with a Blackman window (60-ms duration) and were separated in time by a silent gap of 40 ms. The sequence of repeating *complex stream* and *simultaneous complex* caused a percept of two distinct auditory objects (a rapidly repeating complex with a pitch of 300 Hz and a complex repeating at one third that rate).

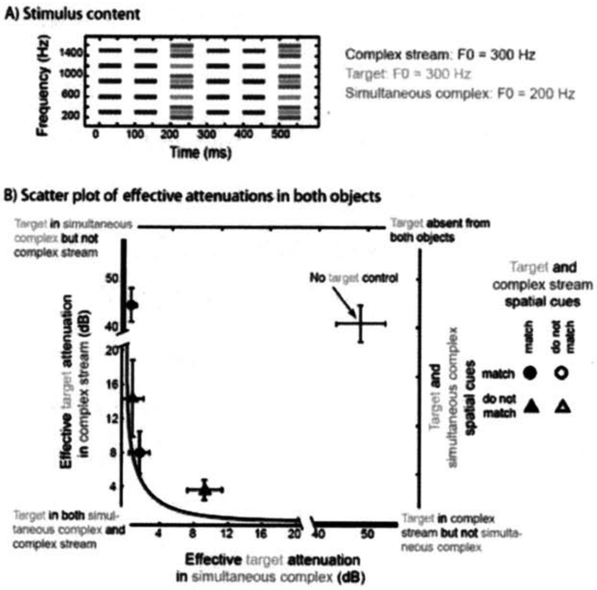


Figure A-4 (A): Stimulus used in supplementary experiment. **(B):** Scatter plot of effective attenuations in both objects.

This stimulus design caused a distinct change in both the spectral density and the perceived pitch of the *simultaneous complex* depending on whether or not the target was heard as part of the *simultaneous complex*. When the target was heard in the *simultaneous complex*, listeners perceived a complex with a dense spectral composition and a pitch of 100 Hz (corresponding to the missing fundamental). When the target was not heard as part of the *simultaneous complex,* its perceived pitch was an octave higher (200 Hz) and it had a more sparse spectral density.

Spatial cues for the target, *simultaneous complex*, and *complex stream* were generated using the same head-related transfer functions as in the main experiment.

Appendix A.7.2 Procedures

As in the main experiment, two-object stimuli were presented in two blocks of trials differing only in the instructions to the subjects. In *complex stream* blocks, subjects matched the perceived content of the *complex stream*. In *simultaneous complex* blocks, subjects matched the perceived content of the *simultaneous* *complex*. Trials for each condition were presented in a different random order for each block of trials.

In both kinds of blocks, listeners used the method of adjustment to match the perceived content of the attended object. Each trial began by presenting a three-second-long test stimulus. This was followed by a three-second-long, single-object matching stimulus that consisted of an adjustable-level target and either a fixed-level *complex stream* or a fixed-level *simultaneous complex.* During presentation of the matching stimulus, subjects could adjust (in real time) the attenuation of the target by pressing one button to increase the target attenuation and a different button to decrease the target attenuation. Three-second-long test and matching stimuli alternated until the subject was satisfied with the perceptual match between the perceived content of the attended object in the two-object test stimulus and the content of that object in the single-object matching stimulus. When the subject was satisfied with the match, she/he pressed a third button, which stored the results of that trial and initiated the next trial in the block. Ten subjects participated in the supplemental experiment.

REFERENCES

1. Cherry, E.C., *Some Experiments on the Recognition of Speech, with One and with Two Ears.* Journal of the Acoustical Society of America, 1953. **25**(5): p. 975-979.

2. Kidd, G., et al., *Informational masking in listeners with sensorineural hearing loss.* Journal of the Association for Research in Otolaryngology, 2002. **3**(2): p. 107-119.

3. Bregman, A.S., *Auditory scene analysis: the perceptual organization of sound.* 1990, Cambridge, Mass.: MIT Press.

4. Arbogast, T.L., C.R. Mason, and G. Kidd, *The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners.* Journal of the Acoustical Society of America, 2005. **117**(4): p. 2169-2180.

5. Carlyon, R.P., et al., *Concurrent sound segregation in electric and acoustic hearing.* Journal of the Association for Research in Otolaryngology, 2007. **8**(1): p. 119-133.

6. Carlyon, R.P., *How the brain separates sounds.* Trends in Cognitive Sciences, 2004. **8**(10): p. 465-471.

7. Darwin, C.J. and R.W. Hukin, *Auditory objects of attention: the role of interaural time differences.* Journal of Experimental Psychology-Human Perception & Performance, 1999. **25**(3): p. 617-29.

8. Palmer, S.E., *Vision science: photons to phenomenology.* 1999, Cambridge, Mass.: MIT Press.

9. Wertheimer, M., *Untersuchunger zur Lehre von der Gestalt. II.* Psychologische Forschung, 1923. **4**: p. 301-350.

10. Kohler, W., *Gestalt psychology, an introduction to new concepts in modern psychology.* 1947, New York: Liveright Pub. Corp.

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

12. Darwin, C.J., *Auditory grouping.* Trends in Cognitive Sciences, 1997. **1**(9): p. 327-333.

13. Best, V., et al., *Binaural interference and auditory grouping.* Journal of the Acoustical Society of America, 2007. **121**(2): p. 1070-1076.

14. Culling, J.F. and Q. Summerfield, *Perceptual Separation of Concurrent Speech Sounds - Absence of across-Frequency Grouping by Common Interaural Delay.* Journal of the Acoustical Society of America, 1995. **98**(2): p. 785-797.

15. Bronkhorst, A.W. and R. Plomp, *Effect of Multiple Speech-Like Maskers on Binaural Speech Recognition in Normal and Impaired Hearing.* Journal of the Acoustical Society of America, 1992. **92**(6): p. 3132-3139.

16. Bronkhorst, A.W. and R. Plomp, *The Effect of Head-Induced Interaural Time and Level Differences on Speech-Intelligibility in Noise.* Journal of the Acoustical Society of America, 1988. **83**(4): p. 1508-1516.

17. Hawley, M.L., R.Y. Litovsky, and H.S. Colburn, *Speech intelligibility and localization in a multi-source environment.* Journal of the Acoustical Society of America, 1999. **105**(6): p. 3436-3448.

18. Joris, P. and T.C.T. Yin, *A matter of time: internal delays in binaural processing.* Trends in Neurosciences, 2007. **30**(2): p. 70-78.

19. Van Noorden, L.P.A.S., *Temporal Coherence in the Perception of Tone Sequences.* 1975, Institute for Perception Research: Eindhoven. p. 104.

20. Beauvois, M.W. and R. Meddis, *Computer simulation of auditory stream segregation in alternating-tone sequences.* Journal of the Acoustical Society of America, 1996. **99**(4): p. 2270-2280.

21. Hartmann, W.M. and D. Johnson, *Stream Segregation and Peripheral Channeling.* Music Perception, 1991. **9**(2): p. 155-184.

22. McCabe, S.L. and M.J. Denham, *A model of auditory streaming.* Journal of the Acoustical Society of America, 1997. **101**(3): p. 1611-1621.

23. Vliegen, J. and A.J. Oxenham, *Sequential stream segregation in the absence of spectral cues.* Journal of the Acoustical Society of America, 1999. **105**(1): p. 339-346.

24. Anstis, S. and S. Saida, *Adaptation to Auditory Streaming of Frequency-Modulated Tones.* Journal of Experimental Psychology-Human Perception and Performance, 1985. **11**(3): p. 257-271.

25. Bregman, A.S., *Auditory Streaming is Cumulative.* Journal of Experimental Psychology-Human Perception and Performance, 1978. **4**(3): p. 380-387.

26. Carlyon, R.P., et al., *Effects of attention and unilateral neglect on auditory stream segregation.* Journal of Experimental Psychology-Human Perception and Performance, 2001. **27**(1): p. 115-127.

27. Cusack, R., et al., *Effects of location, frequency region, and time course of selective attention on auditory scene analysis.* Journal of Experimental Psychology-Human Perception and Performance, 2004. **30**(4): p. 643-656.

28. Sussman, E.S., et al., *The role of attention in the formation of auditory streams.* Perception & Psychophysics, 2007. **69**(1): p. 136-152.

29. Bee, M.A. and G.M. Klump, *Primitive auditory stream segregation: A neurophysiological study in the songbird forebrain.* Journal of Neurophysiology, 2004. **92**(2): p. 1088-1104.

30. Fishman, Y.I., J.C. Arezzo, and M. Steinschneider, *Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration.* Journal of the Acoustical Society of America, 2004. **116**(3): p. 1656-1670.

31. Fishman, Y.I., et al., *Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey.* Hearing Research, 2001. **151**(1-2): p. 167-187.

32. Kanwal, J.S., A.V. Medvedev, and C. Micheyl, *Neurodynamics for auditory stream segregation: tracking sounds in the mustached bat's natural environment.* Network: Computation in Neural Systems, 2003. **14**(3): p. 413-435.

33. Cusack, R., *The intraparietal sulcus and perceptual organization.* Journal of Cognitive Neuroscience, 2005. **17**(4): p. 641-651.

34. Micheyl, C., et al., *Perceptual Organization of Tone Sequences in the Auditory Cortex of Awake Macaques.* Neuron, 2005. **48**: p. 139-148.

35. Wilson, E.C., et al., *Cortical fMRI activation to sequences of tones alternating in frequency: Relationship to perceived rate and streaming.* Journal of Neurophysiology, 2007. **97**(3): p. 2230-2238.

36. Darwin, C.J. and R.W. Hukin, *Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity.* Journal of the Acoustical Society of America, 1997. **102**(4): p. 2316-24.

37. Drennan, W.R., S. Gatehouse, and C. Lever, *Perceptual segregation of competing speech sounds: The role of spatial location.* Journal of the Acoustical Society of America, 2003. **114**(4): p. 2178-2189.

38. Dyson, B.J. and C. Alain, *Representation of concurrent acoustic objects in primary auditory cortex.* Journal of the Acoustical Society of America, 2004. **115**(1): p. 280-288.

39. Moore, B.C.J., B.R. Glasberg, and R.W. Peters, *Thresholds for Hearing Mistuned Partials as Separate Tones in Harmonic Complexes.* Journal of the Acoustical Society of America, 1986. **80**(2): p. 479-483.

40. Darwin, C.J. and V. Ciocca, *Grouping in Pitch Perception - Effects of Onset Asynchrony and Ear of Presentation of a Mistuned Component.* Journal of the Acoustical Society of America, 1992. **91**(6): p. 3381-3390.

41. Roberts, B. and S.D. Holmes, *Asynchrony and the grouping of vowel components: Captor tones revisited.* Journal of the Acoustical Society of America, 2006. **119**(5): p. 2905-2918.

42. Bregman, A.S. and S. Pinker, *Auditory Streaming and Building of Timber.* Canadian Journal of Psychology-Revue Canadienne De Psychologie, 1978. **32**(1): p. 19-31.

43. Dannenbring, G.L. and A.S. Bregman, *Streaming vs. fusion of sinusoidal components of complex tones.* Perception & Psychophysics, 1978. **24**(4): p. 369-376.

44. Darwin, C.J., R.W. Hukin, and B.Y. al-Khatib, *Grouping in pitch perception: evidence for sequential constraints.* Journal of the Acoustical Society of America, 1995. **98**(2 Pt 1): p. 880-5.

45. Steiger, H. and A.S. Bregman, *Competition among Auditory Streaming, Dichotic Fusion, and Diotic Fusion.* Perception & Psychophysics, 1982. **32**(2): p. 153-162.

46. Darwin, C.J. and R.W. Hukin, *Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony.* Journal of the Acoustical Society of America, 1998. **103**(2): p. 1080-4.

47. Darwin, C.J. and N.S. Sutherland, *Grouping Frequency Components of Vowels - When Is a Harmonic Not a Harmonic.* Quarterly Journal of Experimental Psychology Section a-Human Experimental Psychology, 1984. **36**(2): p. 193-208.

48. Turgeon, M., A.S. Bregman, and P.A. Ahad, *Rhythmic masking release: Contribution of cues for perceptual organization to the cross-spectral fusion of concurrent narrow-band noises.* Journal of the Acoustical Society of America, 2002. **111**(4): p. 1819-1831.

49. Turgeon, M., A.S. Bregman, and B. Roberts, *Rhythmic masking release: Effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping.* Journal of Experimental Psychology-Human Perception and Performance, 2005. **31**(5): p. 939-953.

50. Klatt, D.H., *Software for a cascade/parallel formant synthesizer.* Journal of the Acoustical Society of America, 1980. **67**(3): p. 971-995.

51. Peterson, G.E. and H.L. Barney, *Control methods used in a study of the vowels.* Journal of the Acoustical Society of America, 1952. **24**(2): p. 175-184.

52. Hukin, R.W. and C.J. Darwin, *Effects of Contralateral Presentation and of Interaural Time Differences in Segregating a Harmonic from a Vowel.* Journal of the Acoustical Society of America, 1995. **98**(3): p. 1380-1387.

53. Darwin, C.J., *Perceiving vowels in the presence of another sound: a quantitative test of the "Old-plus-New" heuristic*, in *Levels in Speech Communication: Relations and Interactions: a tribute to Max Wajskop*, C. Sorin, et al., Editors. 1995, Elsevier: Amsterdam. p. 1-12.

54. Hukin, R.W. and C.J. Darwin, *Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification.* Perception and Psychophysics, 1995. **57**(2): p. 191-6.

55. Shinn-Cunningham, B.G., *Influences of spatial cues on grouping and understanding sound*, in *Forum Acusticum*. 2005: Budapest.

56. Shinn-Cunningham, B.G., A.K.C. Lee, and A.J. Oxenham, *A sound element gets lost in perceptual competition.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(29): p. 12223-12227.

57. Green, D.M. and J.A. Swets, *Signal Detection Theory and Psychophysics*. 1966, New York: Wiley.

58. Macmillan, N.A. and C.D. Creelman, *Detection Theory: A User's Guide*. 2nd ed. 2005, New Jersey: Lawrence Erlbaum.

59. McAdams, S., M.C. Botte, and C. Drake, *Auditory continuity and loudness computation.* Journal of the Acoustical Society of America, 1998. **103**(3): p. 1580-1591.

60. Roberts, B., B.R. Glasberg, and B.C.J. Moore, *Primitive stream segregation of tone sequences without differences in fundamental frequency or passband.* Journal of the Acoustical Society of America, 2002. **112**(5): p. 2074-2085.

61. Miller, G.A. and G.A. Heise, *The trill threshold.* Journal of the Acoustical Society of America, 1950. **22**(5): p. 637-638.

62. Rock, I. and S. Palmer, *The Legacy of Gestalt Psychology.* Scientific American, 1990. **263**(6): p. 84-90.

63. Rock, I., *The logic of perception*. 1983, Cambridge, MA: MIT Press.

64. Rubin, E., *Visuell wahrgenommene Figuren*. 1921, Copenhagen: Gyldendals.

65. Baylis, G.C. and J. Driver, *Shape-coding in IT cells generalizes over contrast and mirror reversal, but not figure-ground reversal.* Nature Neuroscience, 2001. **4**(9): p. 937-942.

66. Kourtzi, Z. and N. Kanwisher, *Representation of perceived object shape by the human lateral occipital complex.* Science, 2001. **293**(5534): p. 1506-1509.

67. Rubin, N., *Figure and ground in the brain.* Nature Neuroscience, 2001. **4**(9): p. 857-858.

68. Adelson, E.H. and J.R. Bergen, *The Plenoptic Function and the Elements of Early Vision*. Computational Models of Visual Processing, ed. M. Landy and J.A. Movshon. 1991, Cambridge, MA: MIT Press. p. 3-20.

69. Kubovy, M. and D. Van Valkenburg, *Auditory and visual objects.* Cognition, 2001. **80**(1-2): p. 97-126.

70. Shamma, S., *On the role of space and time in auditory processing.* Trends in Cognitive Sciences, 2001. **5**(8): p. 340-348.

71. Van Valkenburg, D. and M. Kubovy, *In defense of the theory of indispensable attributes.* Cognition, 2003. **87**(3): p. 225-233.

72. Griffiths, T.D. and J.D. Warren, *What is an auditory object?* Nature Reviews Neuroscience, 2004. **5**(11): p. 887-892.

73. Rosenthal, D.F. and H.G. Okuno, *Computational auditory scene analysis.* 1998, Mahwah, N.J.: Lawrence Erlbaum Associates.

74. Smaragdis, P., B. Raj, and M. Shashanka. *A probabilistic latent variable model for acoustic modelling.* in *Neural Information Processing Systems Conference.* 2006. Vancouver, Canada.

75. Ruggero, M.A., L. Robles, and N.C. Rich, *2-Tone Suppression in the Basilar-Membrane of the Cochlea - Mechanical Basis of Auditory-Nerve Rate Suppression.* Journal of Neurophysiology, 1992. **68**(4): p. 1087-1099.

76. Shera, C.A., J.J. Guinan, and A.J. Oxenham, *Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(5): p. 3318-3323.

77. Cai, Y.D. and C.D. Geisler, *Suppression in auditory-nerve fibers of cats using low-side suppressors .3. Model results.* Hearing Research, 1996. **96**(1-2): p. 126-140.

78. Delgutte, B., *Physiological-Mechanisms of Psychophysical Masking - Observations from Auditory-Nerve Fibers.* Journal of the Acoustical Society of America, 1990. **87**(2): p. 791-809.

79. Patterson, R.D., M.H. Allerhand, and C. Giguere, *Time-Domain Modeling of Peripheral Auditory Processing - a Modular Architecture and a Software Platform.* Journal of the Acoustical Society of America, 1995. **98**(4): p. 1890-1894.

80. Martinez-Trujillo, J.C. and S. Treue, *Attentional modulation strength in cortical area MT depends on stimulus contrast.* Neuron, 2002. **35**(2): p. 365-370.

81. Reynolds, J.H. and R. Desimone, *Interacting roles of attention and visual salience in V4.* Neuron, 2003. **37**(5): p. 853-863.

82. Reynolds, J.H., T. Pasternak, and R. Desimone, *Attention increases sensitivity of V4 neurons.* Neuron, 2000. **26**(3): p. 703-714.

83. Carrasco, M., S. Ling, and S. Read, *Attention alters appearance.* Nature Neuroscience, 2004. **7**(3): p. 308-313.

84. Fritz, J., S. Shamma, and M. Elhilali, *One click, two clicks: The past shapes the future in auditory cortex.* Neuron, 2005. **47**(3): p. 325-327.

85. Fritz, J., et al., *Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex.* Nature Neuroscience, 2003. **6**(11): p. 1216-1223.

86. Suga, N., et al., *Plasticity and corticofugal modulation for hearing in adult animals.* Neuron, 2002. **36**(1): p. 9-18.

87. Lee, A.K.C. and B.G. Shinn-Cunningham, *Effects of frequency disparities on trading of an ambiguous tone between two competing auditory objects.* submitted.

88. Blauert, J., *Spatial hearing : the psychophysics of human sound localization*. Rev. ed. 1997, Cambridge, Mass.: MIT Press.

89. Best, V., et al., *The influence of spatial separation on divided listening.* Journal of the Acoustical Society of America, 2006. **120**(3): p. 1506-1516.

90. Freyman, R.L., et al., *The role of perceived spatial separation in the unmasking of speech.* Journal of the Acoustical Society of America, 1999. **106**(6): p. 3578-3588.

91. Culling, J.F., K.I. Hodder, and C.Y. Toh, *Effects of reverberation on perceptual segregation of competing voices.* Journal of the Acoustical Society of America, 2003. **114**(5): p. 2871-2876.

92. Darwin, C.J. and R.W. Hukin, *Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention.* Journal of the Acoustical Society of America, 2000. **108**(1): p. 335-342.

93. Lin, I.-F., T. Streeter, and B.G. Shinn-Cunningham. *Trading directional accuracy for realism.* in *Proceedings of the Human-Computer Interaction International 2005 / 1st International Conference on Virtual Reality*. 2005.

94. Shinn-Cunningham, B.G., N. Kopco, and T.J. Martin, *Localizing nearby sound sources in a classroom: Binaural room impulse response.* Journal of the Acoustical Society of America, 2005. **117**(5): p. 3100-3115.

95. Darwin, C.J., *Perceiving Vowels in the Presence of Another Sound - Constraints on Formant Perception.* Journal of the Acoustical Society of America, 1984. **76**(6): p. 1636-1647.

96. Darwin, C.J. and R.W. Hukin, *Effectiveness of spatial cues, prosody, and talker characteristics in selective attention.* Journal of the Acoustical Society of America, 2000. **107**(2): p. 970-977.

97. Desimone, R. and J. Duncan, *Neural Mechanisms of Selective Visual-Attention.* Annual Review of Neuroscience, 1995. **18**: p. 193-222.

98. Peers, P.V., et al., *Attentional functions of parietal and frontal cortex.* Cerebral Cortex, 2005. **15**(10): p. 1469-1484.

99. Hartmann, W.M., *Localization of Sound in Rooms.* Journal of the Acoustical Society of America, 1983. **74**(5): p. 1380-1391.

100. Shinn-Cunningham, B.G., S. Santarelli, and N. Kopco, *Tori of confusion: Binaural localization cues for sources within reach of a listener.* Journal of the Acoustical Society of America, 2000. **107**(3): p. 1627-1636.

101. Buell, T.N. and E.R. Hafter, *Combination of Binaural Information across Frequency Bands.* Journal of the Acoustical Society of America, 1991. **90**(4): p. 1894-1900.

102. Dye, R.H., *The Combination of Interaural Information across Frequencies - Lateralization on the Basis of Interaural Delay.* Journal of the Acoustical Society of America, 1990. **88**(5): p. 2159-2170.

103. Stellmack, M.A. and R.H. Dye, *The Combination of Interaural Information across Frequencies - the Effects of Number and Spacing of Components, Onset Asynchrony, and Harmonicity.* Journal of the Acoustical Society of America, 1993. **93**(5): p. 2933-2947.

104. Krumbholz, K. and A. Nobbe, *Buildup and breakdown of echo suppression for stimuli presented over headphones - the effects of interaural time and level differences.* Journal of the Acoustical Society of America, 2002. **112**(2): p. 654-663.

105. Yang, X.F. and D.W. Grantham, *Echo suppression and discrimination suppression aspects of the precedence effect.* Perception & Psychophysics, 1997. **59**(7): p. 1108-1117.

106. Kidd, G., et al., *The role of reverberation in release from masking due to spatial separation of sources for speech identification.* Acta Acustica United with Acustica, 2005. **91**(3): p. 526-536.

107. Freyman, R.L. and R. Keen, *Constructing and disrupting listeners' models of auditory space.* Journal of the Acoustical Society of America, 2006. **120**(6): p. 3957-3965.

108. Best, V., et al., *Auditory Spatial Perception with Sources Overlapping in Frequency and Time.* Acta Acustica united with Acustica, 2005. **91**: p. 421-428.

109. Good, M.D. and R.H. Gilkey, *Sound localization in noise: The effect of signal-to-noise ratio.* Journal of the Acoustical Society of America, 1996. **99**(2): p. 1108-1117.

110. Lorenzi, C., S. Gatehouse, and C. Lever, *Sound localization in noise in normal-hearing listeners.* Journal of the Acoustical Society of America, 1999. **105**(3): p. 1810-1820.

111. McFadden, D. and E.G. Pasanen, *Lateralization at high frequencies based on interaural time differences.* Journal of the Acoustical Society of America, 1976. **59**(3): p. 634-639.

112. Gardner, M.B., *Image Fusion, Broadening, and Displacement in Sound Location.* Journal of the Acoustical Society of America, 1969. **46**(2B): p. 339-349.

113. Litovsky, R.Y., et al., *The precedence effect.* Journal of the Acoustical Society of America, 1999. **106**(4): p. 1633-1654.

114. Dizon, R.M. and H.S. Colburn, *The influence of spectral, temporal, and interaural stimulus variations on the precedence effect.* Journal of the Acoustical Society of America, 2006. **119**(5): p. 2947-2964.

115. Hafter, E.R. and R.H. Dye, *Detection of Interaural Differences of Time in Trains of High-Frequency Clicks as a Function of Interclick Interval and Number.* Journal of the Acoustical Society of America, 1983. **73**(2): p. 644-651.

116. Shinn-Cunningham, B.G., P.M. Zurek, and N.I. Durlach, *Adjustment and discrimination measurements of the precedence effect.* Journal of the Acoustical Society of America, 1993. **93**(5): p. 2923-32.

117. Stecker, G.C. and E.R. Hafter, *Temporal weighting in sound localization.* Journal of the Acoustical Society of America, 2002. **112**(3): p. 1046-1057.

118. Butler, R.A. and R.F. Naunton, *Role of Stimulus Frequency and Duration in the Phenomenon of Localization Shifts.* Journal of the Acoustical Society of America, 1964. **36**(5): p. 917-922.

119. Bernstein, L.R. and C. Trahiotis, *Binaural beats at high frequencies: Listeners' use of envelope-based interaural temporal and intensitive disparities.* Journal of the Acoustical Society of America, 1996. **99**(3): p. 1670-1679.

120. Trahiotis, C. and L.R. Bernstein, *Detectability of Interaural Delays over Select Spectral Regions - Effects of Flanking Noise.* Journal of the Acoustical Society of America, 1990. **87**(2): p. 810-813.

121. Woods, W.S. and H.S. Colburn, *Test of a Model of Auditory Object Formation Using Intensity and Interaural Time Difference Discrimination.* Journal of the Acoustical Society of America, 1992. **91**(5): p. 2894-2902.

122. Heller, L.M. and C. Trahiotis, *Extents of laterality and binaural interference effects.* Journal of the Acoustical Society of America, 1996. **99**(6): p. 3632-3637.

123. Hill, N.I. and C.J. Darwin, *Lateralization of a perturbed harmonic: effects of onset asynchrony and mistuning.* Journal of the Acoustical Society of America, 1996. **100**(4 Pt 1): p. 2352-64.

124. Litovsky, R.Y. and B.G. Shinn-Cunningham, *Investigation of the relationship among three common measures of precedence: Fusion, localization dominance, and discrimination suppression.* Journal of the Acoustical Society of America, 2001. **109**(1): p. 346-358.

125. Braasch, J. and K. Hartung, *Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data.* Acta Acustica United with Acustica, 2002. **88**(6): p. 942-955.

126. Culling, J.F. and H.S. Colburn, *Binaural sluggishness in the perception of tone sequences and speech in noise.* Journal of the Acoustical Society of America, 2000. **107**(1): p. 517-527.

127. Culling, J.F. and Q. Summerfield, *Measurements of the binaural temporal window using a detection task.* Journal of the Acoustical Society of America, 1998. **103**(6): p. 3540-3553.

128. Grantham, D.W. and F.L. Wightman, *Detectability of varying interaural temporal differences.* Journal of the Acoustical Society of America, 1978. **63**(2): p. 511-523.

129. Akeroyd, M.A. and A.Q. Summerfield, *A binaural analog of gap detection.* Journal of the Acoustical Society of America, 1999. **105**(5): p. 2807-20.

130. Kollmeier, B. and R.H. Gilkey, *Binaural Forward and Backward-Masking - Evidence for Sluggishness in Binaural Detection.* Journal of the Acoustical Society of America, 1990. **87**(4): p. 1709-1719.

131. Perrot, D.R. and S. Pacheco, *Minimum audible angle thresholds for broadband noise as a function of the delay between the onset of the lead and lag signals.* Journal of the Acoustical Society of America, 1989. **85**: p. 2669-2672.

132. Akeroyd, M.A. and L.R. Bernstein, *The variation across time of sensitivity to interaural disparities: Behavioral measurements and quantitative analyses.* Journal of the Acoustical Society of America, 2001. **110**(5): p. 2516-2526.

133. Bernstein, L.R., et al., *Sensitivity to brief changes of interaural time and interaural intensity.* Journal of the Acoustical Society of America, 2001. **109**(4): p. 1604-15.

134. Best, V., et al., *Spatial unmasking of birdsong in human listeners: Energetic and informational factors.* Journal of the Acoustical Society of America, 2005. **118**(6): p. 3766-3773.

135. Stern, R.M., A.S. Zeiberg, and C. Trahiotis, *Lateralization of Complex Binaural Stimuli - a Weighted-Image Model.* Journal of the Acoustical Society of America, 1988. **84**(1): p. 156-165.

136. Trahiotis, C. and R.M. Stern, *Lateralization of Bands of Noise - Effects of Bandwidth and Differences of Interaural Time and Phase.* Journal of the Acoustical Society of America, 1989. **86**(4): p. 1285-1293.

137. Shinn-Cunningham, B.G., *The real reasons you should invest in a surround-sound system.* Journal of the Acoustical Society of America, 2006. **119**(5): p. 3280-3280.

138. Licklider, J.C.R., *The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise.* Journal of the Acoustical Society of America, 1948. **20**(2): p. 150-159.

139. Blauert, J. and W. Lindemann, *Spatial-Mapping of Intracranial Auditory Events for Various Degrees of Interaural Coherence.* Journal of the Acoustical Society of America, 1986. **79**(3): p. 806-813.

140. Freyman, R.L., R.K. Clifton, and R.Y. Litovsky, *Dynamic Processes in the Precedence Effect.* Journal of the Acoustical Society of America, 1991. **90**(2): p. 874-884.

141. Tollin, D.J., L.C. Populin, and T.C.T. Yin, *Neural correlates of the precedence effect in the inferior colliculus of behaving cats.* Journal of Neurophysiology, 2004. **92**(6): p. 3286-3297.

142. Rauschecker, J.P. and B. Tian, *Mechanisms and streams for processing of "what" and "where" in auditory cortex.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(22): p. 11800-11806.

143. Tian, B., et al., *Functional specialization in rhesus monkey auditory cortex.* Science, 2001. **292**(5515): p. 290-293.

144. Reed, C.L., R.L. Klatzky, and E. Halgren, *What vs. where in touch: an fMRI study.* Neuroimage, 2005. **25**(3): p. 718-726.

145. Rao, S.C., G. Rainer, and E.K. Miller, *Integration of what and where in the primate prefrontal cortex.* Science, 1997. **276**(5313): p. 821-824.

146. Denkel, A., *Object and property.* Cambridge studies in philosophy. 1996, Cambridge ; New York: Cambridge University Press.

147. Badcock, D.R. and G. Westheimer, *Spatial location and hyperacuity: flank position within the centre and surround zones.* Spatial Vision, 1985. **1**(1): p. 3-11.

148. Ruda, H., *The warped geometry of visual space near a line assessed using a hyperacuity displacement task.* Spatial Vision, 1998. **11**(4): p. 401-419.

149. Mishkin, M., L.G. Ungerleider, and K.A. Macko, *Object Vision and Spatial Vision - 2 Cortical Pathways.* Trends in Neurosciences, 1983. **6**(10): p. 414-417.

150. Wilson, F.A.W., S.P.O. Scalaidhe, and P.S. Goldmanrakic, *Dissociation of Object and Spatial Processing Domains in Primate Prefrontal Cortex.* Science, 1993. **260**(5116): p. 1955-1958.

151. Middlebrooks, J.C., *Auditory space processing: here, there or everywhere?* Nature Neuroscience, 2002. **5**(9): p. 824-826.

152. Romanski, L.M., et al., *Dual streams of auditory afferents* target *multiple domains in the primate prefrontal cortex.* Nature Neuroscience, 1999. **2**(12): p. 1131-1136.

153. Warren, J.D. and T.D. Griffiths, *Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain.* Journal of Neuroscience, 2003. **23**(13): p. 5799-5804.

154. Zatorre, R.J., et al., *Where is 'where' in the human auditory cortex?* Nature Neuroscience, 2002. **5**(9): p. 905-909.

155. Wang, D.L., *Primitive auditory segregation based on oscillatory correlation.* Cognitive Science, 1996. **20**(3): p. 409-456.

156. Cardoso, J.F., *Blind signal separation: statistical principles.* Proceedings of the IEEE, 1998. **86**(10): p. 2009-2025.

157. Akansu, A.N. and R.A. Haddad, *Multiresolution signal decomposition : transforms, subbands, and wavelets.* 2nd ed. Series in telecommunications. 2001, San Diego: Academic Press.

158. Strawson, P.F., *Individuals, an essay in descriptive metaphysics.* 1959, London.

159. O'Callaghan, C., *Sounds.* in press: Oxford University Press.

160. Baethge, C., et al., *Hallucinations in bipolar disorder: characteristics and comparison to unipolar depression and schizophrenia.* Bipolar Disorders, 2005. **7**(2): p. 136-145.

161. Ajdler, T., L. Sbaiz, and M. Vetterli, *The plenacoustic function and its sampling.* IEEE Transactions on Signal Processing, 2006. **54**(10): p. 3790-3804.

162. Stellmack, M.A., N.F. Viemeister, and A.J. Byrne, *Comparing monaural and interaural temporal windows: Effects of a temporal fringe on sensitivity to intensity differences.* Journal of the Acoustical Society of America, 2005. **118**(5): p. 3218-3228.

163. Holube, I., M. Kinkel, and B. Kollmeier, *Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments.* Journal of the Acoustical Society of America, 1998. **104**(4): p. 2412-2425.

164. Lee, A.K.C., S. Babcock, and B.G. Shinn-Cunningham. *From tone to complex: Generalization of the effects of spatial cues and attention on grouping and streaming.* in *Mid-Winter Meeting of the Association for Research in Otolaryngology*. 2007. Denver, CO.

165. Irwin, R.J., *Binaural Summation of Thermal Noises of Equal and Unequal Power in Each Ear.* The American Journal of Psychology, 1965. **78**(1): p. 57-65.

166. Porsolt, R.D. and R.J. Irwin, *Binaural Summation in Loudness of Two Tones as a Function of Their Bandwidth.* The American Journal of Psychology, 1967. **80**(3): p. 384-390.

167. Buschman, T.J. and E.K. Miller, *Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices.* Science, 2007. **315**: p. 1860-1862.

168. Macken, W.J., et al., *Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory.* Journal of Experimental Psychology-Human Perception and Performance, 2003. **29**(1): p. 43-51.

169. Naatanen, R., et al., *'Primitive intelligence' in the auditory cortex.* Trends in Neurosciences, 2001. **24**(5): p. 283-288.

170. Sussman, E., W. Ritter, and H.G. Vaughan, *Predictability of stimulus deviance and the mismatch negativity.* Neuroreport, 1998. **9**(18): p. 4167-4170.

171. Sussman, E.S., *Integration and segregation in auditory scene analysis.* Journal of the Acoustical Society of America, 2005. **117**(3): p. 1285-1298.

172. Sussman, E., et al., *Top-down effects can modify the initially stimulus-driven auditory organization.* Cognitive Brain Research, 2002. **13**(3): p. 393-405.

173. Palmer, S. and I. Rock, *On the Nature and Order of Organizational Processing - a Reply.* Psychonomic Bulletin & Review, 1994. **1**(4): p. 515-519.

174. Wang, D.L. and G.J. Brown, eds. *Computational Auditory Scene Analysis : Principles, Algorithms and Applications*. 2006, JOHN WILEY: New York.

175. Hulse, S.H., S.A. MacDougall-Shackleton, and A.B. Wisniewski, *Auditory scene analysis by songbirds: Stream segregation of birdsong by European starlings (Sturnus vulgaris).* Journal of Comparative Psychology, 1997. **111**(1): p. 3-13.

176. Endepols, H., et al., *Roles of the auditory midbrain and thalamus in selective phonotaxis in female gray treefrogs (Hyla versicolor).* Behavioural Brain Research, 2003. **145**(1-2): p. 63-77.

177. Jouventin, P., T. Aubin, and T. Lengagne, *Finding a parent in a king penguin colony: the acoustic system of individual recognition.* Animal Behaviour, 1999. **57**: p. 1175-1183.

178. Darwin, C.J., D.S. Brungart, and B.D. Simpson, *Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers.* Journal of the Acoustical Society of America, 2003. **114**(5): p. 2913-2922.

179. Darwin, C.J. and R.P. Carlyon, in *Hearing*, B.C.J. Moore, Editor. 1995, Academic Press: San Diego, CA.

180. Liberman, A.M., D. Isenberg, and B. Rakerd, *Duplex Perception of Cues for Stop Consonants - Evidence for a Phonetic Mode.* Perception & Psychophysics, 1981. **30**(2): p. 133-143.

181. Darwin, C.J., H. Pattison, and R.B. Gardner, *Vowel Quality Changes Produced by Surrounding Tone Sequences.* Perception & Psychophysics, 1989. **45**(4): p. 333-342.

182. Javel, E., *Long-term adaptation in cat auditory-nerve fiber responses.* Journal of the Acoustical Society of America, 1996. **99**(2): p. 1040-1052.

183. Smith, R.L., *Short-Term Adaptation in Single Auditory-Nerve Fibers - Some Post-Stimulatory Effects.* Journal of Neurophysiology, 1977. **40**(5): p. 1098-1112.

184. Rand, T.C., *Dichotic release from masking for speech.* Journal of the Acoustical Society of America, 1974. **55**(3): p. 678-680.

185. Elmasian, R., R. Galambos, and A. Bernheim, *Loudness Enhancement and Decrement in 4 Paradigms.* Journal of the Acoustical Society of America, 1980. **67**(2): p. 601-607.

186. Mapes-Riordan, D. and W.A. Yost, *Loudness recalibration as a function of level.* Journal of the Acoustical Society of America, 1999. **106**(6): p. 3506-3511.

187. Zwislock, J.J. and W.G. Sokolich, *Loudness Enhancement of a Tone Burst by a Preceding Tone Burst.* Perception & Psychophysics, 1974. **16**(1): p. 87-90.

188. Dye, R.H., et al., *The influence of later-arriving sounds on the ability of listeners to judge the lateral position of a source.* Journal of the Acoustical Society of America, 2006. **120**(6): p. 3946-3956.

189. Joris, P.X., et al., *Auditory midbrain and nerve responses to sinusoidal variations in interaural correlation.* Journal of Neuroscience, 2006. **26**(1): p. 279-289.

190. Shinn-Cunningham, B.G., et al., *Bottom-up and top-down influences on spatial unmasking.* Acta Acustica United with Acustica, 2005. **91**(6): p. 967-979.

191. Scharf, B., in *Attention*, H. Pashler, Editor. 1998, Psychology Press: Hove, UK.

192. Tata, M.S. and L.M. Ward, *Spatial attention modulates activity in a posterior "where" auditory pathway.* Neuropsychologia, 2005. **43**(4): p. 509-516.

193. Cusack, R., R.P. Carlyon, and I.H. Robertson, *Neglect between but not within auditory objects.* Journal of Cognitive Neuroscience, 2000. **12**(6): p. 1056-1065.