Effects of frequency disparities on trading of an ambiguous tone between two competing auditory objects

Adrian K. C. Lee and Barbara G. Shinn-Cunningham[a)]
Hearing Research Center, Boston University, Boston, Massachusetts 02215,
and Speech and Hearing Bioscience and Technology Program, Harvard-MIT
Division of Health Sciences and Technology, Cambridge, Massachusetts 02139

(Received 24 July 2007; revised 17 March 2008; accepted 19 March 2008)

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 spectrotemporal element that logically could belong to more than one object, the correct way to group that element can be ambiguous. Many psychoacoustic experiments have implicitly assumed that when a sound mixture contains ambiguous sound elements, the ambiguous elements "trade" between competing sources, such that the elements contribute more to one object in conditions when they contribute less to others. However, few studies have directly tested whether such trading occurs. While some studies found trading, trading failed in some recent studies in which spatial cues were manipulated to alter the perceptual organization. The current study extended this work by exploring whether trading occurs for similar sound mixtures when frequency content, rather than spatial cues, was manipulated to alter grouping. Unlike when spatial cues were manipulated, results are roughly consistent with trading. Together, results suggest that the degree to which trading is obeyed depends on how stimuli are manipulated to affect object formation.

© *2008 Acoustical Society of America.* [DOI: 10.1121/1.2908282]

PACS number(s): 43.66.Ba, 43.66.Lj, 43.66.Mk [RYL] Pages: 4340–4351

I. 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 elements from a source of interest and perceptually separate these elements from the elements originating from other sources. Auditory scene analysis (Bregman, 1990; Darwin, 1997; Carlyon, 2004) depends on grouping together simultaneous sound energy as well as grouping energy across time (streaming or sequential grouping).

The perceptual organization of a sequence of tones depends on the frequency proximity of the component tones, the tone repetition rate, and the attentional state of the observer (Van Noorden, 1975). Specifically, when a sequence of tones alternates between two frequencies, the probability of perceiving two separate streams (corresponding to the two frequencies) increases as the frequency separation and/or the tone repetition rate increases. While there have been attempts to explain sequential streaming by considering only the peripheral processing of the auditory system (Hartmann and Johnson, 1991; Beauvois and Meddis, 1996; McCabe and Denham, 1997), such explanations cannot fully account for how even simple tone sequences are perceptually organized (Vliegen and Oxenham, 1999).

Simultaneous grouping associates sound elements that occur together in time, such as the harmonics of a vowel (Culling and Summerfield, 1995; Hukin and Darwin, 1995; Drennan *et al.*, 2003). Some of the cues that dominate how simultaneous sound elements are grouped together include common amplitude modulation (e.g., common onsets and offsets) as well as harmonic structure.

The relative potency of acoustical cues influencing sequential and simultaneous grouping has been measured by pitting different acoustic grouping cues against one another to determine which cue perceptually dominates. Frequency separation strongly influences sequential grouping, with the grouping strength decreasing as the frequency separation increases. Common onset/offset causes simultaneous elements to group together (Bregman and Pinker, 1978). Moreover, there is an interaction between sequential and simultaneous grouping cues (Dannenbring and Bregman, 1978; Steiger and Bregman, 1982; Darwin *et al.*, 1995). For instance, the contribution of one harmonic to a harmonic tone complex is reduced by the presence of a tone sequence surrounding the complex whose frequency matches that harmonic (Darwin and Sutherland, 1984; Darwin *et al.*, 1995; Darwin and Hukin, 1997, 1998).

Acoustically, if two sources (S_1 and S_2) are uncorrelated, the total energy in the sum of the sources at each frequency is expected to equal the sum of the energies in S_1 and S_2 at that frequency. Thus, if listeners form fixed, veridical estimates of uncorrelated sources in a mixture, the sum of the energies perceived in the two objects should equal the physical energy of the sound in a mixture, an idea we will call the "energy conservation" hypothesis. However, it is quite likely that listeners do not form perfect, veridical estimates of the sources in a mixture. Even if energy conservation fails, it seems reasonable to expect that if a sound mixture contains

[a)]Author to whom correspondence should be addressed. Present address: Department of Cognitive and Neural Systems, Room 311, Boston University, 677 Beacon St., Boston, MA 02215. Tel.: 617-353-5764. FAX: 617-353-7755. Electronic mail: shinn@cns.bu.edu an ambiguous element that could logically belong to more than one object, the energy in that element should trade between objects. Specifically, when an ambiguous element perceptually contributes more to one object, it should contribute less to the competing object. We call this the “trading” hypothesis. One special form of trading would occur if the pressure amplitude, rather than the total energy, of the ambiguous element is conserved (“pressure conservation”). In such cases, the sum of the effective energies that an ambiguous element contributes to competing objects should be 3 dB less than the physical energy of the ambiguous element (e.g., see McAdams *et al.*, 1998). Pressure conservation would be veridical if the frequency components making up the ambiguous elements in a sound mixture are in phase (and therefore correlated) rather than independent.

There are many studies exploring how a sequential stream influences a simultaneously grouped harmonic complex. However, there are only a handful of studies exploring whether a simultaneous complex has reciprocal influences on a sequential stream and whether or not energy conservation or trading occurs for ambiguous sound mixtures (Darwin, 1995; McAdams *et al.*, 1998; Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). Some past studies have assumed that trading occurs without actually measuring the effective contribution of an ambiguous element to both objects in a sound mixture. For instance, a recent pair of studies tested how frequency proximity interacts with harmonicity and common onset/offset to influence the perceived content of a harmonic complex (Turgeon *et al.*, 2002; Turgeon *et al.*, 2005). In interpreting these results, it was explicitly assumed that when an ambiguous tone did not strongly contribute to the simultaneously grouped object, it strongly contributed to the ongoing stream even though the perceived spectral content of the ongoing stream was not tested.

The few studies that have explicitly tested whether the energy in an ambiguous element trades between competing objects give conflicting results (Darwin, 1995; McAdams *et al.*, 1998; Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008).

One study measured perception for a harmonic complex when one tone in the complex was turned on before the other harmonics (Darwin, 1995). The intensity of the portion of the ambiguous tone preceding the harmonic complex was manipulated to alter how much of the simultaneous portion was perceived in the complex. In general, the contribution of the simultaneous portion of the ambiguous harmonic to the complex was reduced when the ambiguous harmonic began before the other harmonics. Moreover, the amount by which the contribution of the ambiguous tone to the simultaneous complex was reduced increased as the intensity of the precursor portion of the ambiguous tone increased. While energy conservation failed, trading was observed: The sum of the ambiguous element’s contribution to the two competing objects (the separate precursor tone and the harmonic complex) was roughly constant at about 3 dB less than the physical energy of the ambiguous tone present during the complex (Darwin, 1995). This result was roughly consistent with pressure rather than energy conservation.

Another study exploring perception of alternating low-intensity, narrow-band stimuli and high-intensity wider-band stimuli that overlapped in frequency found similar results (McAdams *et al.*, 1998). In this study, listeners generally perceived the low-level, narrow-band stimulus as continuous and the higher-intensity, broader-band stimulus as a pulsed stream. Thus, there was a band of energy during the high-intensity, wider-band stimuli that was ambiguous and perceptually contributed to both streams. Three alternative forms of trading were explicitly evaluated to determine whether trading occurred: Energy conservation, pressure conservation, and loudness conservation (where the total loudness of the ambiguous elements, in sones, was apportioned between the two competing streams). Results generally fell between energy and pressure conservations. Other studies of “homophonic induction” also suggest a form of trading (e.g., see Warren *et al.*, 1994; Kashino and Warren, 1996), although these studies did not quantify the contribution of the ambiguous sound energy to each of the competing objects.

While many past studies found results roughly consistent with pressure conservation, trading completely failed in two studies presenting a sound mixture consisting of a repeating tone stream and a harmonic complex whose fourth harmonic was one of the tone-stream components (Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). When spatial cues were manipulated to vary the perceptual organization of the scene, the sum of the energy contributions of the ambiguous element to the two objects dramatically varied across conditions. In fact, in one condition, the ambiguous element contributed almost nothing to either of the two competing objects.

The current study was designed to determine whether trading occurs for stimuli similar to those for which trading failed in previous studies. As in the earlier studies, the stimuli contained an ambiguous pure-tone element (the target) that could logically be heard as one tone in an isochronous stream of repeating tones and/or as part of a more slowly repeating harmonic complex. In the current experiment, frequency proximity rather than spatial cues were manipulated to affect perceptual organization. Specifically, the frequency of the repeating tones varied from trial to trial from below to above the frequency of the target, whose frequency equaled the fourth harmonic of the simultaneous harmonic complex.

A control experiment allowed us to compute the effective level of the target perceived in the two objects (tone stream and complex) to test how the ambiguous target was allocated across the competing objects. We performed another experiment to relate our results to past studies exploring how frequency proximity influences perception of a tone sequence. We find that for the current stimuli, energy conservation fails, but trading (roughly consistent with pressure conservation) is observed.

II. EXPERIMENT 1: COMPETING OBJECTS

A. Methods

1. Stimuli

Stimuli generally consisted of a repeating sequence of a pair of tones followed by a harmonic complex [Fig. 1(a)].

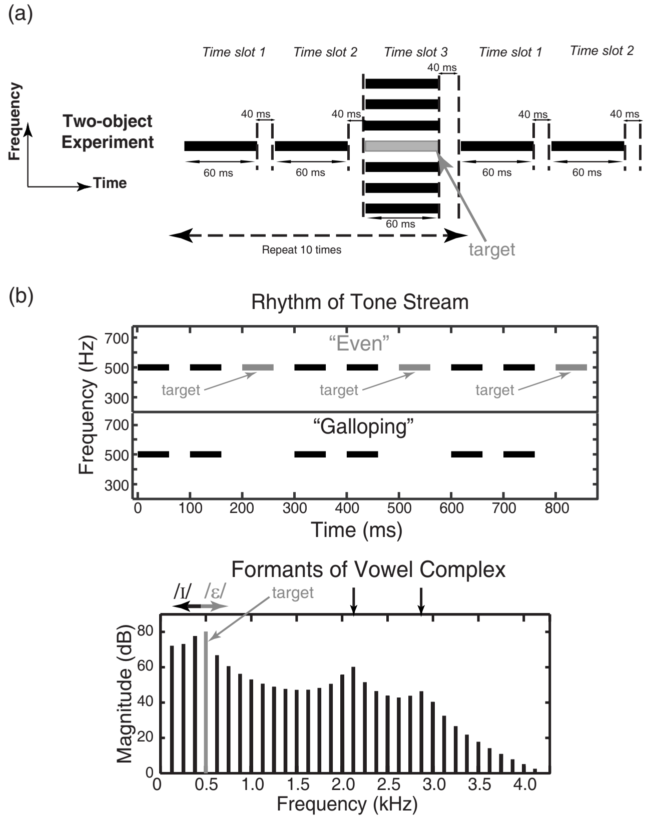

FIG. 1. (a) Two-object stimuli were created by repeating a three-item sequence consisting of a pair of pure tones followed by a harmonic complex. In the reference configuration, the tones in time slots 1 and 2 are at 500 Hz. Time slot 3 is made up of two components: a target tone at 500 Hz and a tone complex with fundamental frequency of 125 Hz (with the fourth harmonic at 500 Hz omitted). The tone complex is shaped by a synthetic vowel spectral envelope to make it sound like a short vowel (Darwin, 1995). 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 (1995). 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.

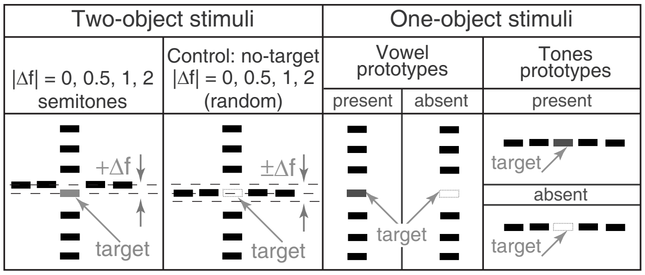

FIG. 2. Experimental conditions. Each block consists of seven two-object stimuli with the target present, a two-object control without the target present, and two one-object prototypes (see text for more details).

The frequency of the pair of tones varied from trial to trial from two semitones below to two semitones above 500 Hz, taking on one of seven predetermined values (i.e., 0, ±0.5, ±1, and ±2 semitones relative to 500 Hz; also see Fig. 2, left panels).

The harmonic complex contained the first 39 harmonics of 125 Hz, excluding the fourth harmonic (500 Hz). The phase of each component was randomly chosen on each trial. The magnitudes of the harmonics were shaped to simulate the filtering of the vocal tract (Klatt, 1980). The first formant frequency (F1) was set to 490 Hz, close to the expected value for the American-English vowel /ε/ (Peterson and Barney, 1952). The second and third formants were fixed at 2100 and 2900 Hz, respectively. The half-power bandwidths of the three formants were 90, 110, and 170 Hz [the parameters were chosen based on studies by Hukin and Darwin (1995)].

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

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 3 s stimulus that was perceived as two objects: An ongoing stream of tones and a repeating vowel occurring at one-third that rate.

The rhythm of the tone sequence depends on the degree to which the target is perceived in the tone stream [Fig. 1(b), top panel]. 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 phonetic identity of the harmonic complex depends on whether or not the target is heard as part of the complex [Fig. 1(b), bottom panel; Hukin and Darwin, 1995]. The first formant of the complex (F1) is perceived as slightly higher when the target is perceived in 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 /ε/ when the target is part of the complex and more like /ɪ/ when it is not part of the complex.

Control stimuli consisted of one-object presentations with only the pair of tones or only the harmonic complex, either with the target ("target-present" one-object prototype) or without the target ("target-absent" one-object prototype; see Fig. 2, right panels). 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, see Fig. 2, second panel from the left).

2. Task

In order to assess the perceptual organization of the two-object mixture and how the frequency difference between the repeated tones and target affected the perceived structure of both the tone stream 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") by performing a one-interval, two-alternative-forced-choice task. 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 "/ε/" as in "bet"). In order to control for the possibility that streaming changes over time, we asked subjects to attend to the object of interest throughout the 3-s-long stimulus but to make their judgments about the attended object based on what they perceived at the end of the stimulus presentation.

3. Environment

All stimuli were generated offline using the MATLAB software (Mathworks Inc.). Signals were processed with pseudoanechoic head-related transfer functions (HRTFs) [see Shinn-Cunningham (2005) for details] in order to make the stimuli similar to those used in companion studies that varied the source location (Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). 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 1 m from the listener.

Stimuli were generated at a sampling rate of 25 kHz and sent to Tucker-Davis Technologies hardware for digital to analog conversion and attenuation. Presentation of the stimuli was controlled by a PC, which selected the stimulus to play on a given trial. All signals were presented at a listener-controlled, comfortable level that had a maximum value of 80 dB sound pressure level. The intensity of each stimulus was roved over a 14 dB range in order to discourage using the level as a cue to vowel identity or tone rhythm. Stimuli were presented over insertion headphones (Etymotic ER-1) to subjects seated in a sound-treated booth. Subjects responded via a graphical user interface.

B. Experimental procedure

1. Participants

Fourteen subjects (eight male, six female, aged 18–31) took part in the experiments. All participants had pure-tone thresholds in both ears within 20 dB of normal-hearing thresholds at octave frequencies between 250 and 8000 Hz and within 15 dB of normal-hearing thresholds at 500 Hz. 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. One-object prototype training

In each session of testing, each subject was familiarized with the one-object prototypes with and without the target (Fig. 2, right panels). During training, subjects were given feedback to reinforce the correct labeling of the one-object, target-present and target-absent prototypes. This feedback ensured that subjects learned to accurately label the rhythm of the sequence of tones and the phonetic identity of the harmonic complex for unambiguous, one-object stimuli. Subjects had to achieve at least 90% correct when identifying the two prototypes in the one-object pretest before proceeding to the main experiment.

3. Main experiment

Following training, listeners judged either the tone-stream rhythm or the vowel identity, depending on the experimental block. Both two-object stimuli and the appropriate one-object prototypes (see Fig. 2) were intermingled in each block. The one-object trials served as controls that allowed us to assess whether listeners maintained the ability to label the unambiguous stimuli throughout the run without feedback (see right side of Fig. 2). From trial to trial, the frequency of the repeating tones in the two-object stimuli randomly varied relative to the target (Fig. 2). Seven two-object conditions were tested in each block, with the frequency of the repeated tones ranging from two semitones below to two semitones above the target frequency ($\Delta f = 0, \pm 0.5, \pm 1, \pm 2$ semitones). A control two-object condition was included in which the target was not presented. In this control, the frequency of the repeated tones was 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; Fig. 2 second panel from the left) to ensure that the subjects did not make rhythmic or vowel judgments based on the absolute frequency of the repeated tones.

In one block of the experiment, we presented eight two-object stimuli and the two one-object prototypes containing no vowel (target present and target absent). 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 one-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.

4. One-object control experiments

Two companion control experiments tested the subjective impressions of either the tone-stream rhythm or 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 one-object stimuli (tones in one experiment, harmonic complexes 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 and 14 dB (in 2 dB steps) relative to the level of the target in the two-object experiments. In the one-object tone task, subjects reported whether the rhythm on a given trial was galloping or even. In the one-object harmonic complex task, subjects reported whether the complex was /ɪ/ or /ε/.

For both one-object control experiments, the percent responses (y) for each subject were related to the target attenuation (x) by fitting a sigmoidal function of the form

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

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

5. Relative *d*′ calculation

Raw percent correct target-present responses (even for the tones, /ε/ for the vowel) were computed for each subject and condition. Because the raw percentage of responses does not give any insight into what differences were perceptually significant and which were perceptually small, we used decision theory to estimate the perceptual distance between the stimulus and the one-object target-absent prototypes (see Shinn-Cunningham *et al.*, 2007, Methods). This method is briefly summarized below.

In each block of the main experiment, one-object prototypes (with and without the target) were randomly intermingled with the ambiguous, two-object stimuli. We assumed that in judging the vowel identity or tone rhythm, 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 were used to compute the perceptual distances (d') between the distributions [Fig. 3(a)].

We use $d'_{\text{present:absent}}$ to denote the perceptual distance between the target-present and target-absent prototypes. By assuming the above decision model, $d'_{\text{present:absent}}$ is given as (Green and Swets, 1966; Macmillan and Creelman, 2005)

$$d'_{\text{present:absent}} = \Phi^{-1}[\Pr(\text{“target present”}\,|\text{target present})] - \Phi^{-1}[\Pr(\text{“target present”}\,|\text{target absent})], \tag{2}$$

where Φ^{-1} denotes the inverse of the cumulative Gaussian distribution and Pr("target present"|stimulus) represents the probability that the subject reports that the target is present in the specified stimulus. In order to avoid an incalculably large value of d' due to sampling issues, the number of responses in each possible category was incremented by 0.5 prior to computing the percentage of responses and the resulting values of d'. As a result of this adjustment, the maximum achievable d' value was 4.28. Values of $d'_{\text{present:absent}}$ were separately calculated for each subject.

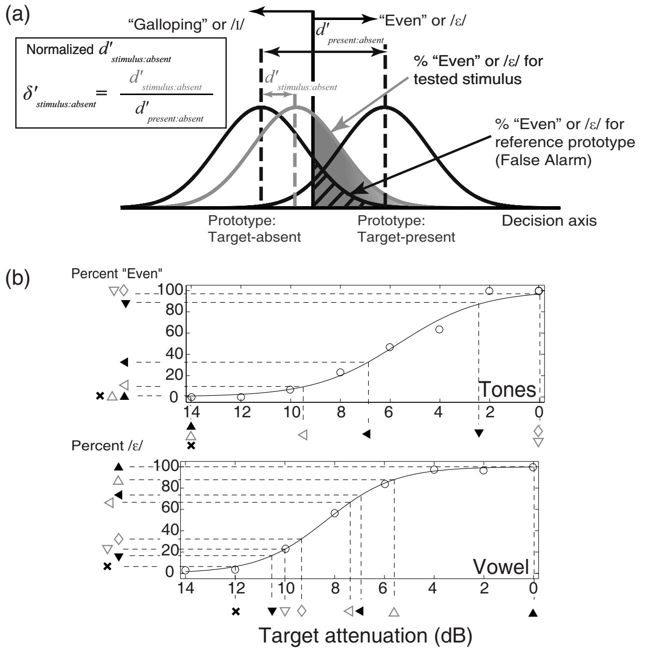

FIG. 3. (a) Schematics of the decision model assumed in computing $\delta'_{\text{stimulus:absent}}$. The decision axis (representing the decision variable for either the rhythmic or vowel identification space) is shown along the abscissa. The Gaussian distributions show the conditional probabilities of observing different values of the decision variable for the target-absent and target-present prototypes (left and right distributions, respectively) as well as for a particular two-object stimulus (middle distribution). (b) Computation of the effective target attenuation from the psychometric functions relating percent target-present responses to physical target attenuation for one-object stimuli for an example subject. The solid line shows the psychometric function fitted to the data points from the one-object control experiment, plotted as circles. The symbols on the ordinate and horizontal dashed lines represent the percentage of even (top panel) or /ε/ (bottom panel) responses for different stimuli. The vertical dashed lines and symbols along the abscissa show the effective target attenuation estimated from the control data.

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

$$d'_{\text{stimulus:absent}} = \Phi^{-1}[\Pr(\text{“target present”}\,|\text{stimulus})] - \Phi^{-1}[\Pr(\text{“target present”}\,|\text{target absent})]. \tag{3}$$

In order to 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 sensitivity measure for each condition from the raw sensitivities as

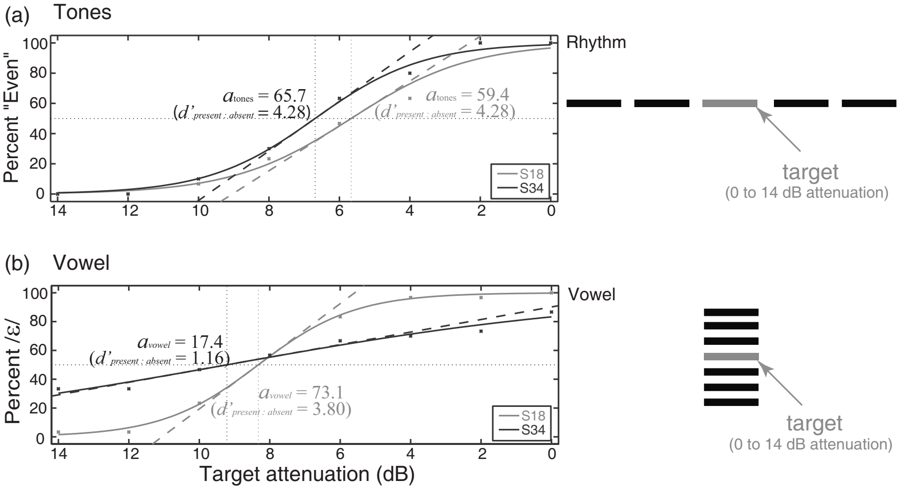

FIG. 4. Example psychometric functions for results of one-object experiments in which the target attenuation varied from 0 to 14 dB (in 2 dB steps), for two representative subjects (S18, a good subject, and S34, a subject who just passed our screening criteria). The dotted lines show the slope of each of the psychometric function at the 50% point. The raw percent responses (for tone-stream rhythm on top and vowel identity below) are shown for each subject as a function of target attenuation.

$$\delta'_{\text{stimulus:absent}} = \frac{d'_{\text{stimulus:absent}}}{d'_{\text{present:absent}}}. \tag{4}$$

A value of $\delta'_{\text{stimulus:absent}} < 0.5$ indicates that the stimulus was perceived as more like the target-absent than the target-present prototype. Conversely, a value of $\delta'_{\text{stimulus:absent}} > 0.5$ indicates that responses were more like those for the target-present than for the target-absent prototype [Fig. 3(a)].

6. Effective target level calculation

To quantify the effective level that the target contributed to each object, we analyzed the psychometric functions fit to the responses from the corresponding one-object control experiment (see Sec. II C 4), which relate the percentage of target-present responses to the physical intensity of the target actually present in the stimuli. By 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 corresponding one-object stimuli [see Fig. 3(b)].

C. Results

1. Subject screening

Despite training, not all subjects could reliably label the one-object vowel prototype stimuli when they were presented in the main experiment, which provided no feedback and intermingled the prototype stimuli with ambiguous two-object stimuli. We adopted a screening protocol to exclude any subjects who could not accurately label the prototype stimuli during the main experiment. Specifically, we excluded data from subjects who failed to achieve $d'_{\text{present:absent}} > 1.0$.

We also excluded any subject for whom response percentages only weakly depended on the target attenuation in the one-object control experiments. Specifically, if the fitted slope parameter a in Eq. (1) was less than 10%/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.

For all subjects, the slope relating the percentage of galloping responses to target attenuation was very steep and met our criterion. Thus, all subjects perceived consistent changes in the rhythm of the tone stream with attenuation of the target. Similarly, all $d'_{\text{present:absent}}$ scores were much greater than the criterion when listeners judged the tones' rhythm. Specifically, 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. Thus, no subjects were excluded from the experiment based on poor performance in the tones task.

Six out of the 14 subjects (three male, three female) failed to meet our criteria for the vowel experiment and had their results excluded from further analysis. All data analyzed below are from the eight subjects who passed all criteria for both tone and vowel screenings.

Figure 4 shows example psychometric functions for the one-object control experiments for subjects S18 (a relatively good subject) and S34 (a subject who passed our screening criteria but was less consistent in labeling the vowels). The top panel in Fig. 4 shows results for the tone experiment. Both subjects responded "galloping" in conditions where the target tone was attenuated by about 12 dB or more relative to the repeating tones and "even" when the intensity of the target matched that of the repeating tones (i.e., 0 dB attenuation). Moreover, both subjects showed steep, monotonically increasing psychometric functions. The bottom panel in Fig. 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 $d'_{\text{stimulus:absent}}$ in the vowel task (1.16 compared to 3.80 for S18). Despite the relatively poor sensitivity of S34, this subject met the liberal inclusion criteria we imposed.

2. Rhythmic judgments

Figure 5 summarizes results of the main two-object experiment for the rhythm judgments (left column of Fig. 5;

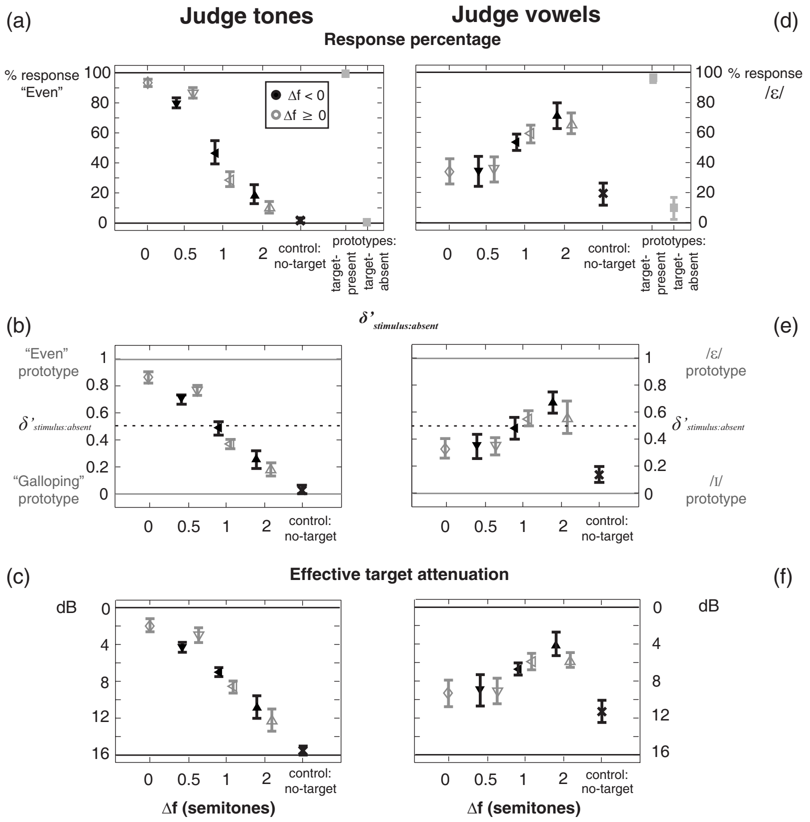

FIG. 5. Results of both rhythm judgments (left column) and vowel judgments (right column). [(a) and (d)] Raw response percentages. [(b) and (e)] $\delta'_{\text{stimulus:absent}}$ derived from raw results. [(c) and (f)] Effective target attenuation derived from the psychometric functions relating raw responses to effective target attenuations. Each marker represents the across-subject mean estimate and the error bar shows ±1 standard error of the mean.

corresponding results for the vowel judgments are shown in the right column and are considered in the next section). Figure 5(a) shows the group mean percentages of even responses (error bars show the across-subject standard error of the mean). All but one subject correctly identified the even and galloping one-object prototypes with 100% accuracy [see squares at far right of Fig. 5(a)]. When the frequency of the repeating tones matched that of the target in the two-object condition (i.e., $\Delta f=0$ semitones), subjects generally reported an even percept (average target-present response rate was greater than 90%; diamond at left of plot). 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 [see open and filled triangles in Fig. 5(a)]. As expected, there was a very low probability of reporting that the target was present in the two-object no-target control trials [i.e., the average percentage of target-present responses was 1.7%; see X in Fig. 5(a)]. The $d'_{\text{stimulus:absent}}$ values (not shown) 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 5(b) plots $\delta'_{\text{stimulus:absent}}$, which quantifies the perceptual distances between a given stimulus and the one-object prototypes (0, near the galloping prototype; 1, near the even prototype). As all subjects were nearly equal in their ability to properly label the two prototypes, the pattern of the mean δ' results looks very similar to the raw percent responses. A two-way repeated-measure analysis of variance (ANOVA) on the $\delta'_{\text{stimulus:absent}}$ scores was performed with factors of $|\Delta f|$ and $\text{sgn}(\Delta f)$. There was a significant main effect of $|\Delta f|$ [$F_{\text{GG}}(1.06,7.42)=48.8$, $p_{\text{GG}}<0.000\,147$].[1] These results suggest that the bigger the frequency separation between the tones and the target, the more likely listeners are to report a galloping percept for the two-object stimuli. Although the ANOVA indicated there was a significant interaction between $|\Delta f|$ and $\text{sgn}(\Delta f)$ [$F(2,14)=8.85$, $p<0.003\,28$], paired-sample t tests (two-tailed with Dunn–Sidak *post hoc* adjustments for three planned comparisons) failed to support this result. Specifically, the paired t tests found no significant differences between positive and negative frequency differences for any of the frequency separations tested ($\Delta f=\pm 0.5$: $t_7=-2.61$, $p_{\text{DS}}=0.101$; $\Delta f=\pm 1$: $t_7=2.32$, $p_{\text{DS}}=0.152$; $\Delta f=\pm 2$ semitones: $t_7=1.69$, $p_{\text{DS}}=0.353$). Thus, there is little evidence that the sign of Δf influences the perceived rhythm of the tones.

3. Vowel judgments

Figure 5(d) 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 nonzero likelihood of subjects mislabeling the one-object prototypes [see squares to far right of Fig. 5(d)]. 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; see diamond at far left of Fig. 5(d)]. As the magnitude of the frequency difference between the tones and the target increased, the probability of reporting an /ɛ/ increased (i.e., the target contributed more to the vowel; see open and filled triangles). As expected for the no-target control stimulus, subjects almost always responded /ɪ/, as if the target was not present [see X in Fig. 5(d)].

Because there were large individual 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, consistent with the fact that listeners generally had more difficulty in identifying the vowel than labeling the tone rhythm (even for the one-object prototypes).

Transforming the results to δ' reduces the across-subject variability in d' by normalizing results by the differences in overall sensitivity [Fig. 5(e); comparison not shown]. 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. A two-way repeated-measure ANOVA was performed on the $\delta'_{\text{stimulus:absent}}$ results with factors of $|\Delta f|$ and $\text{sgn}(\Delta f)$. The ANOVA found a significant main effect of $|\Delta f|$ [$F_{\text{GG}}(1.06,7.39)=6.37$, $p_{\text{GG}} < 0.0368$]. There were no main effect of $\text{sgn}(\Delta f)$ [$F(1,7)=0.111$, $p=0.749$] and no significant interaction between $|\Delta f|$ and $\text{sgn}(\Delta f)$ [$F(2,14)=1.55$, $p=0.246$]. Thus, as with the tone-rhythm task, we conclude that $|\Delta f|$ affects the contribution of the target to the attended object, but that there is no consistent effect of $\text{sgn}(\Delta f)$.

4. Target trading

The percent responses found in the one-object control experiment provide mappings that allow us to evaluate 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 one-object stimulus to a percent response were used to find (for each subject and condition) the equivalent target attenuation for the perceived contribution of the target to the attended object [see Fig. 3(b)]. The across-subject means and standard errors of these mapped equivalent attenuations are plotted in Figs. 5(c) (for the tones) and 5(f) (for the vowel).

Conclusions drawn from analysis of raw percent responses and δ' (top two rows of Fig. 5) and of the effective attenuation of the target (bottom row of Fig. 5) are essentially the same. However, this final comparison enables a quantitative analysis of whether there is trading between the tones and the vowel.

Figure 6(a) 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 shows the trading relationship that would be observed if energy is conserved, while the dashed line shows the trading relationship that corresponds to loss of 3 dB of target energy (consistent with pressure conservation).

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 strongly attenuated [see × in the top right of Fig. 6(a)]. For the ambiguous, two-object stimuli, trading occurs: The effective attenuation of the target in the tone stream is larger for stimuli that produce less attenuation in the vowel [diamond and triangles fall on a monotonically decreasing curve in Fig. 6(a)]. However, the trading does not strictly follow energy conservation: For some conditions, the total of the sum of the effective energies of the target in the two streams is less than that actually 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).

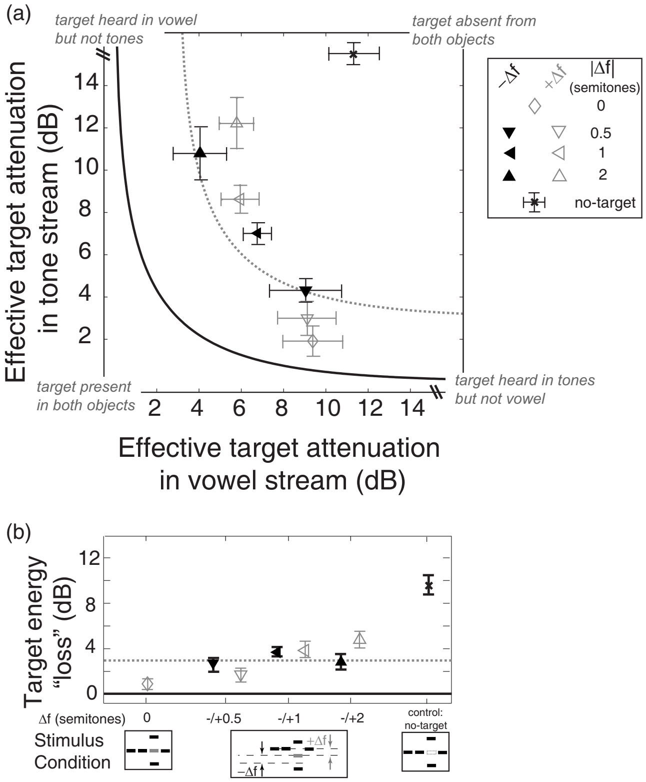

FIG. 6. (a) Scatter plot of the effective target attenuation in the tones vs the effective target attenuation in the vowel. Data would fall on the solid line if energy conservation holds. A trading relationship in which the total perceived target energy is 3 dB less than the physical target energy would fall on the dashed line (equivalent to conservation of pressure rather than energy; see Darwin, 1995). (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 (0 dB lost energy) shows where results would fall if energy conservation held. The dashed line shows where results would fall if pressure, rather than energy, were conserved.

To quantify the trading observed in Fig. 6(a), we computed the total effective energy of the target by summing, for each condition, its effective energy when subjects attended to the tones and its effective energy when subjects attended to 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 Fig. 6(b).

In general, the total perceived target energy was less than the physical target energy in the stimuli (all symbols fall above 0 dB). The lost energy was near 3 dB for many of the stimuli (see dashed line at 3 dB, which is equivalent to pressure conservation; see also Darwin, 1995; McAdams *et al.*, 1998).

A two-way repeated-measure ANOVA on the total effective target energy lost found no effects of $|\Delta f|$ [$F(2,14)=4.29$, $p=0.0544$], $\text{sgn}(\Delta f)$ [$F(1,7)=2.06$, $p=0.194$], or their interaction $[|\Delta f|\ \mathrm{sgn}(\Delta f),\ F(2,14)=3.47,\ p=0.0599]$. One-sample t tests separately performed on the lost target energy values explored whether the lost energy was statistically significantly different from 0 dB (with Dunn–Sidak *post hoc* adjustments for seven planned comparisons). For conditions $\Delta f=-0.5$ ($t_7=-4.06$, $p_{\mathrm{DS}}<0.0333$), $\Delta f=-1$ ($t_7=-9.30$, $p_{\mathrm{DS}}<0.000\,242$), $\Delta f=+1$ ($t_7=-5.37$, $p_{\mathrm{DS}}<0.007\,29$), and $\Delta f=+2$ semitones ($t_7=-6.13$, $p_{\mathrm{DS}}<0.003\,33$), the lost energy was significantly greater than zero, supporting the conclusion that the total target energy allocated across the two competing objects is often less than the physical energy present in the target element of the sound mixture.

D. 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 (Darwin *et al.*, 1995). Peripheral adaptation could contribute to the lost target energy here, as well. However, such adaptation would be greatest when $|\Delta f|=0$ and the pair of tones excite the same neural population as the target. Instead, the amount of target energy that is lost is smallest, if anything, when $|\Delta f|=0$. Thus, adaptation does not fully account for the perceptual loss of target energy observed here. Moreover, adaptation was ruled out as the sole cause of lost target energy in similar studies when trading fails (also see the Discussion and Appendix in Shinn-Cunningham *et al.*, 2007).

While some target energy is not accounted for, we found trading: In general, 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 (Darwin, 1995; McAdams *et al.*, 1998) whose results were consistent with trading of an ambiguous element between two competing objects. However, this result contrasts with results using stimuli similar to the current 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 groupings (Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). Thus, even though the stimuli and procedures employed in this study are very similar to those used when trading fails, results are more consistent with results of studies using very different methods. The difference between the current study and the past studies with similar stimuli suggests that the degree to which trading is observed depends on the way grouping cues are manipulated to alter perceptual organization.

None of these studies found energy conservation, where the perceived target energy in the competing objects sums to the physical energy of the target. In this respect, all results support the idea that the way in which the acoustic mixture is broken into objects is inconsistent with a veridical parsing of the acoustic mixture despite the intuitive appeal of this idea (see Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008).

III. EXPERIMENT 2: NO COMPETING OBJECTS

In the first experiment, subjects were less likely to hear the target as part of the tone stream 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 one-object tone stimuli generally find that a two-semitone difference is not enough to cause a single object to break apart into two streams (Anstis and Saida, 1985; Vliegen and Oxenham, 1999; Carlyon *et al.*, 2001; Micheyl *et al.*, 2005). While it is likely that the presence of the vowel, which competes for ownership of the target, explains why relatively small $|\Delta f|$ lead to percepts of a galloping rhythm in our two-object stimuli, other differences between past experiments and our main experiment may also contribute. A follow-up experiment was conducted to directly assess whether the strong effect of a relatively small frequency separation on the perceived tone-stream rhythm was due to some procedural or stimulus differences between the current and past one-stream versus two-stream studies.

A. Methods

Stimuli were similar to the one-object tone stimuli used in the main experiment. In the main 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). In this one-object experiment, there were no such constraints, and we tested separations between the repeating tones and the target of up to eight semitones (0, ±0.5, ±1, ±2, ±4, and ±8 semitones relative to 500 Hz) to make results more comparable to previous one-stream versus two-stream experiments.

Subjects were instructed to judge the tone-stream rhythm (galloping versus even) after ten presentations of the pair-of-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 related experiments conducted in our laboratory.

B. Results

All subjects could nearly perfectly distinguish the galloping from the even prototypes. Figure 7(a) shows the raw percent response scores and Fig. 7(b) shows the $\delta'_{\mathrm{stimulus:absent}}$ results, averaged across subjects. Both ways of considering the data show that the contribution of the target to the tones decreases as $|\Delta f|$ increases, as expected [data points fall along a monotonically decreasing curve in Figs. 7(a) and 7(b)].

A two-way repeated-measure ANOVA on the one-stream tone rhythm $\delta'_{\mathrm{stimulus:absent}}$ found a significant main effect of $|\Delta f|$ $[F_{\mathrm{GG}}(1.74,12.2)=67.0,\ p_{\mathrm{GG}}<4.27\times10^{-7}]$ but no significant effect of $\mathrm{sgn}(\Delta f)$ $[F(1,7)=0.078,\ p=0.789]$ and no significant effect of the interaction between $|\Delta f|$ and $\mathrm{sgn}(\Delta f)$ $[F(4,28)=0.455,\ p=0.768]$.

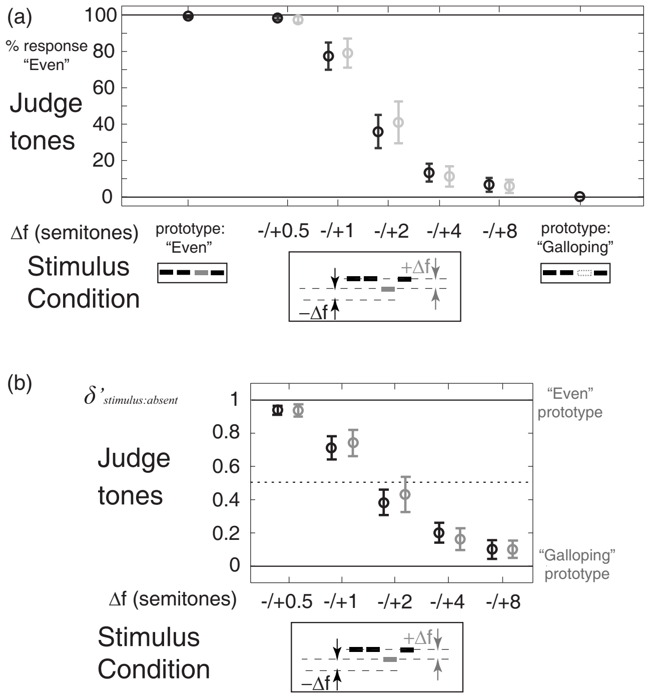

FIG. 7. (a) Raw percent responses for one-object stimuli as a function of Δf between the tones and the target. The target-present prototype is equivalent to the $\Delta f = 0$ condition. Note that there is no equivalent vowel manipulation in this experiment. (b) Normalized $\delta'_{\text{stimulus:absent}}$ results, derived from the raw percent-category responses in (a).

Compared to the results for two-object conditions, the likelihood of hearing the repeating tones as galloping increased much more slowly with increasing $|\Delta f|$. Responses are perceptually only halfway between galloping and not galloping for separations of two semitones.

C. Discussion

The same $|\Delta f|$ was much less likely to lead to galloping responses in this one-object experiment than for two-object stimuli in our main experiment. In the main experiment, subjects judged the tone stream as galloping when the absolute frequency separation between the tones and the target was two semitones apart. In the absence of the competing harmonic complex, subjects only consistently judged the tone stream to be galloping when the absolute frequency separation between the tones and the target was four or more semitones apart. This suggests that the presence of the competing vowel made listeners more likely to report a galloping rhythm for a fixed $|\Delta f|$.

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 (Vliegen and Oxenham, 1999). Along these lines, it is conceivable that the weaker effect of frequency separation observed in this experiment compared to that in the main experiment is wholly or partially due to the difference in the frequency separation ranges used (up to a maximum $|\Delta f|$ of two semitones in the main experiment but up to eight semitones in the one-object experiment). Regardless, the results of this control experiment show that our procedures produce results consistent with the literature when we use similar frequency separation ranges and one-object mixtures.

IV. 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 (Van Noorden, 1975; Anstis and Saida, 1985; Vliegen and Oxenham, 1999; Roberts *et al.*, 2002) or simultaneous grouping (Culling and Summerfield, 1995; Darwin and Hukin, 1997; Drennan *et al.*, 2003; Dyson and Alain, 2004). Most of these studies explored what acoustical parameters 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 that can logically belong to more than one object in the auditory scene rather than whether a mixture is heard as one or two objects.

In our main experiment, the presence of the tones at the same frequency as the target is enough to substantially remove the contribution of the target to the harmonic complex. Similarly, the presence of the harmonic complex reduces the contribution of the target to the tone stream. Results of experiment 2 suggest that the presence of the competing harmonic complex causes the target to "drop out" of the tone stream at smaller $|\Delta f|$ than when there is no competing object. These results are consistent with past work showing that the perceived content of an object depends on interactions with other objects in the mixture (e.g., influencing both perceived pitch as well as perceived vowel identity; Darwin *et al.*, 1995; Darwin and Hukin, 1997).

If the perceptual organization of the auditory scene is fixed and veridical, the sum of the energies at a given frequency in all of the perceived objects should, on average, equal the physical energy of that frequency in the signal reaching the ear, obeying the energy conservation hypothesis. A weaker hypothesis, of trading, is that when the perceptual contribution of the ambiguous elements to one object in a stimulus increases, the contribution to other objects decreases.

Current results are consistent with the target trading between the harmonic complex and tones. However, as in past studies, the contribution of the ambiguous target to the tones plus its contribution to the vowel is as much as 3 dB less than the energy that is physically present in the target. Nonlinearity in peripheral processing could partially explain trading in which energy in an ambiguous element is lost. 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. However, peripheral adaptation cannot fully explain the current results. Peripheral adaptation of the target should be maximal when the tones and target have the same frequency. Instead, the target

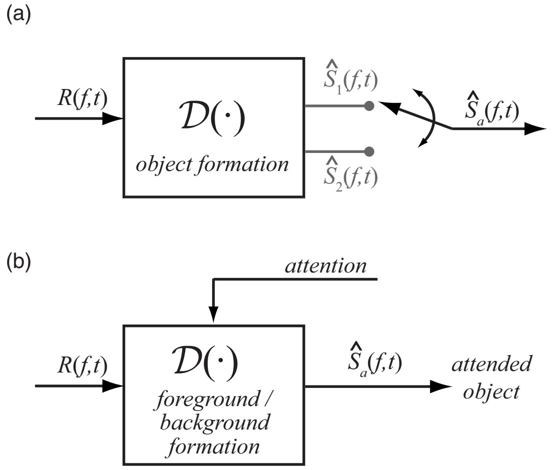

FIG. 8. Two possible models for how objects are formed. (a) A model in which the grouping of the scene depends only on the stimuli. (b) A model in which the grouping of the object in the foreground depends on top-down goals of the listener.

energy that is lost is smallest when the target frequency matches that of the repeated tones. Thus, as in past studies investigating the energy conservation hypothesis (Darwin, 1995; McAdams *et al.*, 1998; Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008), peripheral effects may contribute to, but cannot explain, results.

Although results of the current study and some past studies find trading that is roughly consistent with pressure conservation, not all past studies show trading. The current results, which are consistent with trading, use stimuli and methods that are closer to those used in studies in which trading fails rather than where trading occurs. The only difference between these studies is the grouping cues that were manipulated to affect the perceptual organization of the scene. Taken together, these results suggest that the fact that trading is roughly obeyed in some studies is coincidental, and is likely due to the specific stimulus manipulation employed in a study and not because trading is a rule governing auditory scene analysis. Instead, there are two possibilities that could explain current and past results, diagramed in Fig. 8 and considered below.

In Fig. 8(a), $\mathcal{D}(\cdot)$ represents an operator that, by utilizing all available information from the signal reaching the receiver, $R(f,t)$, yields estimates of the objects in the scene $\{\hat{S}_i(f,t)\}$, independent of the goals of the listener (such as the task s/he is performing or the object s/he is attending). In this scheme, attention simply selects one of the already-formed objects as the foreground object. If $\mathcal{D}(\cdot)$ operates in a purely bottom-up manner, then competition between different objects may alter the effective level of ambiguous sound energy through mutual inhibition. Such inhibition could cause the perceptual contribution of an ambiguous target to the objects in the scene to be less than the physical target energy through fixed interactions that depend only on the stimulus. To explain why trading sometimes fails, the strength of this mutual inhibition must depend on the balance between different competing grouping cues, such as spatial cues, frequency proximity, common modulation, and the like.

Alternatively, the goals of the listener may change how energy in the sound mixture is grouped, changing the operation of $\mathcal{D}(\cdot)$ [Fig. 8(b)]. If so, then which object or attribute we attend may alter how we group the mixture (Fig. 8; see also Shinn-Cunningham *et al.*, 2007; Lee and Shinn-Cunningham, 2008). If this is the case, then there is no reason to expect the perceived content of an attended foreground object to predict what a listener perceives when they switch attention or switch tasks.

The experimental goal of the current study is to estimate how $\mathcal{D}(\cdot)$ operates on the sound mixture. To probe this operation, we asked subjects to subjectively judge properties of an auditory object in a scene. However, by asking the subjects to make judgments about an object, we forced them to bring the object of interest into the auditory foreground. We cannot ask listeners to make judgments of objects in the perceptual background: Even if we did, they would undoubtedly focus attention on the background, which would change the old background into the new foreground. As a result, the current experiments cannot differentiate between the two possibilities diagramed in Fig. 8. Either possibility is consistent with the current results: Either a sound mixture is formed into objects in a way that depends only on the stimulus mixture but that does not obey energy conservation (or in some cases, trading) or the focus of attention and/or the goals of the listener affect how the sound mixture is grouped.

Attention modulates neural responses within the visual processing pathway (e.g., Reynolds *et al.*, 2000; Martinez-Trujillo and Treue, 2002; Reynolds and Desimone, 2003), including at the peripheral level (Carrasco *et al.*, 2004). In audition, spectrotemporal receptive fields in the primary auditory cortex change depending on the behavioral task (Fritz *et al.*, 2003; Fritz *et al.*, 2005). Attention has also been implicated in corticofugal modulation of cochlear function in awake mustached bats during vocalization (Suga *et al.*, 2002). These physiological results suggest that top-down processes alter how sound is represented even in relatively peripheral, sensory processing stages of the auditory pathway. While such results do not prove that the way we form auditory objects out of an ambiguous sound mixture depends on top-down factors, they are consistent with the idea that attention alters auditory scene analysis.

V. CONCLUSIONS

Competing objects alter the perceived content of an object in an auditory scene. In particular, there are reciprocal effects between simultaneously and sequentially grouped objects in our two-object mixtures, made up of a repeating tone sequence, a simultaneous vowel complex, and an ambiguous target tone that could logically belong to either object. When the frequency separation between the tone sequence and an ambiguous target tone is varied to alter the perceptual organization of the sound mixture, the contributions of the target to the tones and to the competing vowel obey a loose trading relationship. However, the trading is lossy rather than obeying energy conservation. When the repeated tones and target had slightly different frequencies, the total perceived target energy was roughly 3 dB less than the physical target energy.

While it is possible that some peripheral nonlinearity contributes to this loss of target energy, it cannot fully account for these findings.

These results, as well as past results, suggest either that (1) competing auditory objects mutually suppress ambiguous sound elements, leading to a reduction in the perceptual contribution of the ambiguous element to the objects in a sound mixture, or (2) how an auditory object is formed in a sound mixture depends on top-down goals of the listener. Further work is necessary to tease apart these two possibilities.

ACKNOWLEDGMENTS

This work was supported by a grant from the Office of Naval Research (N00014-04-1-0131) to B.G.S.C. Sigrid Nasser helped in the subject recruitment and the data collection process. Andrew J. Oxenham provided many helpful suggestions in the experimental design.

[1] Throughout, the subscript "GG" denotes that we used the Greenhouse–Geisser-corrected degrees of freedom when testing for significance to account for violations of the sphericity assumption under Mauchly's test. Where the subscript is left out, it signifies a condition for which the sphericity assumption was met.

Anstis, S., and Saida, S. (**1985**). "Adaptation to auditory streaming of frequency-modulated tones," J. Exp. Psychol. Hum. Percept. Perform. **11**, 257–271.

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

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

Bregman, A. S., and Pinker, S. (**1978**). "Auditory streaming and building of timbre," Can. J. Psychol. **32**, 19–31.

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

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. (**2001**). "Effects of attention and unilateral neglect on auditory stream segregation," J. Exp. Psychol. Hum. Percept. Perform. **27**, 115–127.

Carrasco, M., Ling, S., and Read, S. (**2004**). "Attention alters appearance," Nat. Neurosci. **7**, 308–313.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds-Absence of across—frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785–797.

Dannenbring, G. L., and Bregman, A. S. (**1978**). "Streaming vs. fusion of sinusoidal components of complex tones," Percept. Psychophys. **24**, 369–376.

Darwin, C. J. (**1995**). "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*, edited by C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen (Elsevier, Amsterdam), pp. 1–12.

Darwin, C. J. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**, 327–333.

Darwin, C. J., and Hukin, R. W. (**1997**). "Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity," J. Acoust. Soc. Am. **102**, 2316–2324.

Darwin, C. J., and Hukin, R. W. (**1998**). "Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony," J. Acoust. Soc. Am. **103**, 1080–1084.

Darwin, C. J., Hukin, R. W., and al-Khatib, B. Y. (**1995**). "Grouping in pitch perception: Evidence for sequential constraints," J. Acoust. Soc. Am. **98**, 880–885.

Darwin, C. J., and Sutherland, N. S. (**1984**). "Grouping frequency components of vowels—When is a harmonic not a harmonic," Q. J. Exp. Psychol. A **36**, 193–208.

Drennan, W. R., Gatehouse, S., and Lever, C. (**2003**). "Perceptual segregation of competing speech sounds: The role of spatial location," J. Acoust. Soc. Am. **114**, 2178–2189.

Dyson, B. J., and Alain, C. (**2004**). "Representation of concurrent acoustic objects in primary auditory cortex," J. Acoust. Soc. Am. **115**, 280–288.

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

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (**2003**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nat. Neurosci. **6**, 1216–1223.

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

Hartmann, W. M., and Johnson, D. (**1991**). "Stream segregation and peripheral channeling," Music Percept. **9**, 155–184.

Hukin, R. W., and Darwin, C. J. (**1995**). "Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel," J. Acoust. Soc. Am. **98**, 1380–1387.

Kashino, M., and Warren, R. M. (**1996**). "Binaural release from temporal induction," Percept. Psychophys. **58**, 899–905.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Lee, A. K. C., and Shinn-Cunningham, B. G. (**2008**). "Effects of reverberant spatial cues on attention-dependent object formation," J. Assoc. Res. Otolaryngol. **9**, 150–160.

Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide* (Erlbaum, Hillsdale, NJ).

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

McAdams, S., Botte, M. C., and Drake, C. (**1998**). "Auditory continuity and loudness computation," J. Acoust. Soc. Am. **103**, 1580–1591.

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

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

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Reynolds, J. H., and Desimone, R. (**2003**). "Interacting roles of attention and visual salience in V4," Neuron **37**, 853–863.

Reynolds, J. H., Pasternak, T., and Desimone, R. (**2000**). "Attention increases sensitivity of V4 neurons," Neuron **26**, 703–714.

Roberts, B., Glasberg, B. R., and Moore, B. C. J. (**2002**). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband," J. Acoust. Soc. Am. **112**, 2074–2085.

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

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

Steiger, H., and Bregman, A. S. (**1982**). "Competition among auditory streaming, dichotic fusion, and diotic fusion," Percept. Psychophys. **32**, 153–162.

Suga, N., Xiao, Z. J., Ma, X. F., and Ji, W. Q. (**2002**). "Plasticity and corticofugal modulation for hearing in adult animals," Neuron **36**, 9–18.

Turgeon, M., Bregman, A. S., and Ahad, P. A. (**2002**). "Rhythmic masking release: Contribution of cues for perceptual organization to the cross-spectral fusion of concurrent narrow-band noises," J. Acoust. Soc. Am. **111**, 1819–1831.

Turgeon, M., Bregman, A. S., and Roberts, B. (**2005**). "Rhythmic masking release: Effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping," J. Exp. Psychol. Hum. Percept. Perform. **31**, 939–953.

Van Noorden, L. P. A. S. (**1975**). *Temporal Coherence in the Perception of Tone Sequences*, Ph.D. Thesis (Institute for Perception Research, Eindhoven), pp. 1–127.

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

Warren, R. M., Bashford, J. A., Healy, E. W., and Brubaker, B. S. (**1994**). "Auditory induction—Reciprocal changes in alternating sounds," Percept. Psychophys. **55**, 313–322.