JARO 9: 150–160 (2008)
DOI: 10.1007/s10162-007-0109-4

JARO
Journal of the Association for Research in Otolaryngology

Effects of Reverberant Spatial Cues on Attention-dependent Object Formation

Adrian K. C. Lee[1,2] and Barbara G. Shinn-Cunningham[1,2,3]

[11]*Hearing κesearch Center, Boston University, Boston, MA, USA*

[2]*Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA*

[3]*Department of Cognitive and Neural Systems, Boston University, 677 Beacon St., Room 311, Boston, MA 02215, USA*

Received: 19 July 2007; Accepted: 28 November 2007; Online publication: 23 January 2008

ABSTRACT

A recent study showed that when a sound mixture has ambiguous spectrotemporal 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 spatial 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 of the ambiguous element does not "trade" between the objects in a complex scene (i.e., the element does not necessarily contribute more to one object when it contributes less to a competing object). Results are consistent with the idea that the perceptual organization of an acoustic mixture depends on what object a listener attends.

Correspondence to: Barbara G. Shinn-Cunningham · Department of Cognitive and Neural Systems · Boston University · 677 Beacon St., Room 311, Boston, MA 02215, USA. Telephone: +1-617-3535764; fax: +1-617-3537755; email: shinn@cns.bu.edu

Keywords: auditory scene analysis, segregation, grouping, attention, energy trading, streaming

INTRODUCTION

In our everyday lives, multiple, physical sources surround us, so that the information reaching our sensory epithelia is a chaotic mixture of elementary sensations arising from these distinct sources (Wertheimer 1923). 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 (Kohler 1947), and many Gestalt laws of grouping, such as proximity and common fate, are known to influence object formation both in the visual (Rock and Palmer 1990) and auditory (Bregman 1990) 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 (Kubovy and Van Valkenburg 2001; Van Valkenburg and Kubovy 2003) as well as the organization of the sensory epithelia (Griffiths and Warren 2004; Shamma 2001) 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 source that is closer to the observer generally occludes a source 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" (Bregman 1990).

If there is a frequency component that is common to two independent sources in the auditory scene, veridical parsing of the scene can only occur if the total sound energy in that frequency component is divided across the objects that listeners perceive in the scene. Specifically, if listeners parse the acoustic scene properly, the sum of the contributions of the ambiguous component to the different perceptual objects in the scene should equal the physical energy of that frequency in the sound mixture (what we will refer to as "energy conservation"). A weaker form of this hypothesis is "energy trading": energy that could belong logically to more than one object should trade between objects, such that when an ambiguous element contributes more to one object, it should contribute less to a competing object.

While the idea of energy trading is intuitively appealing, only a handful of studies (Darwin 1995; McAdams et al. 1998; Shinn-Cunningham et al. 2007) have explicitly tested whether it holds. Moreover, the results of these studies are mixed. While two of the three studies suggest that energy trading occurs (Darwin 1995; McAdams et al. 1998), ambiguous energy did not trade in the third study (Shinn-Cunningham et al. 2007). In discussing these results, the researchers pointed out that if perceptual organization depends on what object is attended, there is no reason to expect energy trading to hold. It may be that energy trading fails because the object that is attended determines the relative importance of various grouping cues, causing the perceptual organization to change, 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 (Blauert 1997). Interaural time differences (ITDs) and interaural level differences (ILDs) between the signals at the two ears are arguably the most robust cues for source localization. Perhaps as a result, and in contrast to their role in visual object formation, spatial cues only weakly affect auditory object formation over short time scales in most conditions. Instead, local spectrotemporal 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 (Best et al. 2006; Darwin 1997; Darwin and Hukin 1999; Freyman et al. 1999; Shinn-Cunningham 2005).

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 "nonallocation" condition in which the ambiguous target element "disappeared" (Shinn-Cunningham et al. 2007), the objects competing for the target element were a sequential tone stream and a simultaneous harmonic complex. In the "nonallocation" condition, spatial cues supported grouping the target with the simultaneous harmonic complex, while the overall spectrotemporal 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 spectrotemporal 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 similar to those in the previous study, but for which spatial cues are made more ambiguous. In particular, the stimuli used in this study are identical to those of the previous study (Shinn-Cunningham et al. 2007), except that stimuli were convolved with binaural room impulse responses (BRIRs) that contained natural room reverberation (simulating a moderate-sized classroom whose broadband reverberation time is 600–700 ms; see Shinn-Cunningham et al. 2005 for a full characterization of these BRIRs). Such natural reverberant energy degrades the fidelity of ongoing interaural time differences by decorrelating the left and right ear signals (Culling et al. 2003; Darwin and Hukin 2000a; Lin et al. 2005; Shinn-Cunningham et al. 2005), which we hypothesized would reduce the perceptual salience of the spatial cues. Specifically, we hypothesized that the organization of the auditory scene depends on the relative strength of all of the various grouping cues affecting perceptual organization, and that weakening the spatial cues would shift the perceptual balance to favor spectrotemporal structure and reduce the influence of spatial cues on perceptual organization. This might simply reduce how much the perceptual organization of the scene changes for different

combinations of spatial cues. However, we speculated that failures of energy trading occur specifically when there is a fragile balance between the competing grouping cues, helping to explain why trading is sometimes observed and sometimes fails. If so, then reducing the strength of spatial cues might yield results in which energy trading occurs.

METHODS

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 250–8,000 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.

Stimuli

Stimuli consisted of a repeating sequence of a pair of tones followed by a harmonic complex (Fig. 1A; see also Shinn-Cunningham et al. 2007). 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 (Darwin 1984). The first, second and third formant peaks were set to 490, 2,100, and 2,900 Hz, respectively (similar to Darwin 1984). 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 3-s-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.

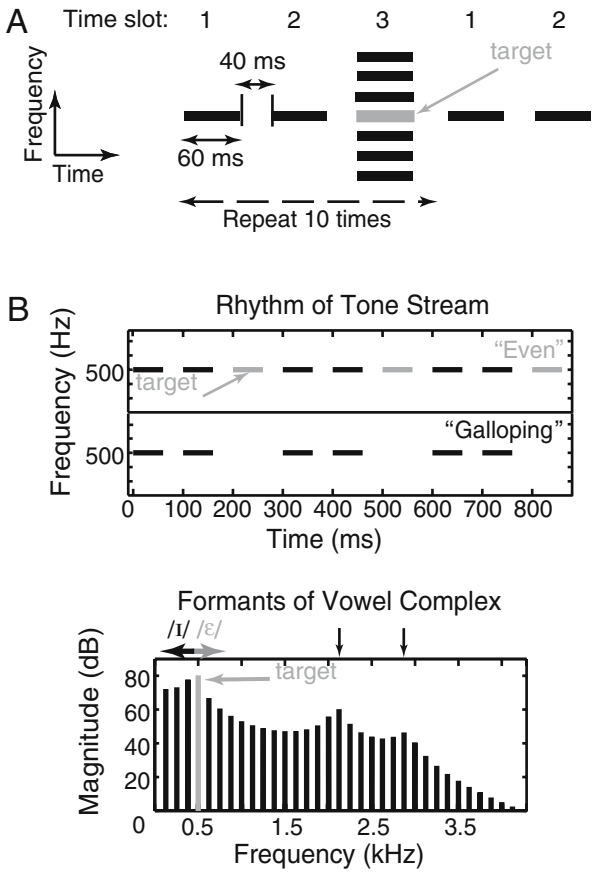

FIG. 1. **A** Two-object stimuli present a repetition of a three-item sequence, consisting of a pair of pure tones followed by a harmonic complex. In the reference configuration, the pure tones in time slots 1 and 2 are at 500 Hz. Time slot 3 is made up of two components: a target tone at 500 Hz and a tone complex with a fundamental frequency 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. **B**, *top panel* The identity of the rhythm of the sequence of 500-Hz tones depends on whether or not the target tone is perceived in the sequential tone stream. *Bottom panel* The synthetic vowel spectral envelope is similar to the shape used by Darwin (1984). 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 in the complex and more like /ɪ/ when the target is not perceived in the complex. *Arrows* indicate the approximate locations of the first three formants of the perceived vowel.

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 (Fig. 1B).

Control stimuli consisted of one-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).

Environment

All stimuli were generated offline using MATLAB software (Mathworks Inc.). Signals were processed with BRIRs measured in a classroom (Shinn-Cunningham et al. 2005) 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.

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 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.

Task

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 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'").

Training procedure with one-object prototypes

In each session of testing, each subject was familiarized with the one-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 one-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, one-object stimuli. Subjects had to achieve at least 90% accuracy when discriminating between the two prototypes in the one-object pretest before proceeding to the two-object experiment.

Procedures for the main two-object experiment

After training on one-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 matched that of the target (Fig. 2). A control two-object condition was also included in which the target was not presented.

Intermingled with the two-object trials were one-object control trials containing only the sound elements for the source that the listener was instructed to attend (either the stream of tones or the vowel, depending on the block). These control trials allowed us to assess whether listeners maintained the ability to label the unambiguous stimuli even in the absence of feedback throughout the run. In these trials, the target was processed to have spatial cues consistent with a source from 0° or 45°, while the other elements were processed to have spatial cues consistent with a source from straight ahead (azimuth=0°; see Fig. 2).

In one block of the experiment, subjects reported the perceived rhythm of the tones in the two-object stimuli and the one-object tone stimuli. In a separate block, subjects reported the perceived identity of the vowel for the two-object stimuli and the 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 interpreting the results to the ambiguous two-object stimuli, as discussed below.

Data analysis

We processed our data using the same procedures described in our companion study (Shinn-Cunningham

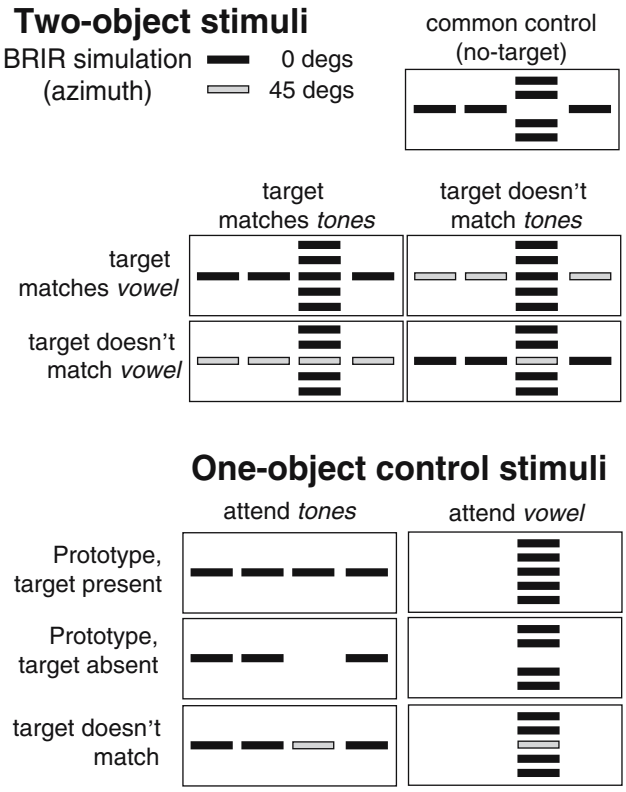

FIG. 2. Simulated spatial locations of the target and repeating tones varied across conditions. All two-object stimuli were common to both experimental blocks (rhythm and vowel identification). Control conditions included a two-object stimulus without a target presented and one-object prototype conditions in which only one object was presented, either with or without the target present.

et al. 2007). 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 one-object target-absent prototypes. For each subject, we computed a normalized d' score, $\delta'_{\text{condition: absent}}$, defined as:

$$\delta'_{\text{condition: absent}} = \frac{d'_{\text{condition: absent}}}{d'_{\text{present: absent}}}, \quad (1)$$

where $d'_{\text{present: absent}}$ 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 $d'_{\text{condition: absent}}$ is the perceptual distance between any stimulus and the target-absent, one-object controls (see Macmillan and Creelman 2005). These values were calculated individually for each subject as:

$$d'_{\text{condition: absent}} = \Phi^{-1}[\Pr(\text{"target present"}|\text{condition})] - \Phi^{-1}[\Pr(\text{"target present"}|\text{target abset})], \quad (2)$$

where Φ^{-1} denotes the inverse of the cumulative Gaussian distribution and Pr("target present"|condition) and Pr("target present"|target absent) are the probabilities of reporting target present in a given condition and in the target-absent condition, respectively. A value of $\delta\prime_{\text{condition: absent}}<0.5$ indicates that the stimulus was perceived as more like the prototype with the target not present while $\delta'_{\text{condition: absent}}>0.5$ indicates that responses were more like those for the target-present than for the target-absent prototype.

Mapping percent response to effective attenuations for each object

One-object control experiments, described in detail in our companion study (see Shinn-Cunningham et al. 2007), were used to construct individual psychometric functions for each subject that related the physical intensity of the target in unambiguous, one-object conditions to the raw percentage of responses in the categorization tasks ("even" vs "galloping", /ɛ/ 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 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 fitted to a logistic function of the form

$$\hat{y} = \frac{1}{1 + e^{-a(x-x_0)}}, \quad (3)$$

where $\hat{y}$ is the predicted percentage of "target-present" response, and the free parameters are: α, 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 /ɛ/) or target-absent (i.e., "galloping" or /ɪ/), the effective attenuation was set to 0 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.

RESULTS

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 the results 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 ($d'_{\text{present: absent}} > 1.0$; see also Shinn-Cunningham et al. 2007). Two out of the nine subjects were unable to reliably label the vowel in the two-object experiment [i.e., $d'_{\text{present: absent}}(\text{vowel}) < 1.0$].

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 α (Eq. 3) was less than 10%/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.

Rhythmic judgments (tones)

Figure 3 summarizes results of the main two-object experiment for both the rhythm judgments (top row; Fig. 3A and B) and vowel identity (bottom row; Fig. 3D and E, considered in the next section). Figure 3C and F use the results of the one-object experiment to plot the effective attenuation of the target and are considered in the section entitled "Mapping raw responses to equivalent attenuations."

Figure 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 lower than in a similar

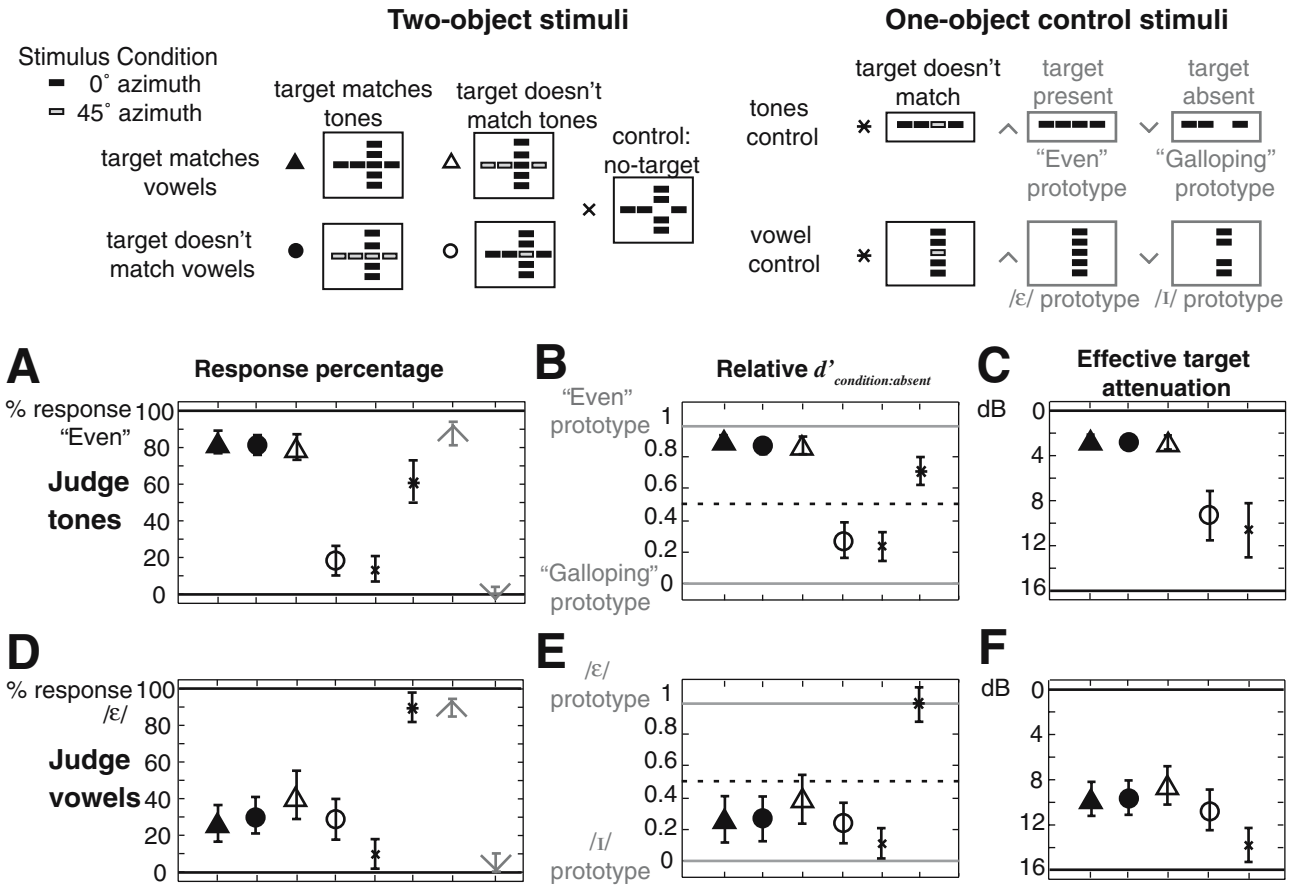

FIG. 3. Results for tones (*top row*) and vowel (*bottom row*). **A** and **D** Raw percentage of "target present" responses. **B** and **E** Relative sensitivity (δ'; 1.0 indicates raw results identical to those with the target present; 0.0 indicates raw results identical to those with the target absent). **C** and **F** Effective target attenuation, based on results from the one-object control experiment (see Fig. 4 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.

experiment using anechoic spatial simulation (see Shinn-Cunningham et al. 2007). The spatial cues had a large effect on the rhythm judgments in the presence of the vowels, in line with previous studies (Darwin and Hukin 1999; 2000b; Shinn-Cunningham et al. 2007). Regardless of the vowel location, when the simulated target location matched that of the tones, the target was perceived to be part of the rhythmic stream (filled triangle and filled circle in Fig. 3A). 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 triangle in Fig. 3A). However, when the target location matched that of the vowel but not the tones, the rhythmic stream was heard as "galloping" (open circle in Fig. 3A) showing that the target did not strongly contribute to the across-time tone stream. When the target was not presented (in the two-object no-target control condition), subjects generally heard the rhythm as "galloping" (ex in Fig. 3A). Subjects generally perceived an even rhythm in the one-object tones condition, even when the spatial location of the target did not match that of the tones (asterisk in Fig. 3A).

Results in Figure 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.

Vowel judgments (vowel)

Figure 3D shows the across-subject mean and the standard error of the raw response percentages for the vowel judgments. There was a nonzero 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 (Shinn-Cunningham et al. 2007), listeners were more likely to respond that the vowel in the two-object conditions was /ɪ/ rather than /ɛ/ for all spatial configurations (i.e., the target did not contribute strongly to the vowel for any two-object stimuli). In all of the one-object vowel conditions, subjects responded as if the target was part of the vowel, responding /ɛ/ roughly 90% of the time even when the target location did not match that of the vowel.

Replotting the data in terms of the relative perceptual distance to the prototypes (Fig. 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).

Mapping raw responses to equivalent attenuations

For all subjects who passed our screening, responses to unambiguous one-object stimuli with different target intensities produced well-behaved psychometric functions. Examples of these functions are shown in Figure 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" (Fig. 4A) and /ɪ/ (Fig. 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

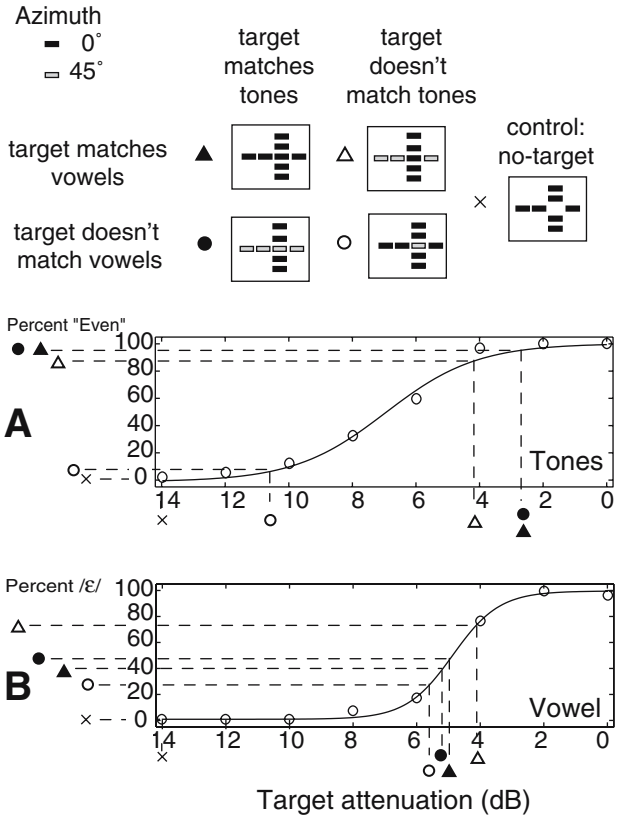

FIG. 4. Illustration of how the results from the one-object control experiment are used to map a raw category response for the tones (**A**) and vowel (**B**) to an effective target attenuation in the two-object control conditions. The psychometric function in each panel shows the raw results (*circles*) and the fit psychometric function (*solid line*) for subject S18 in the corresponding one-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.

Figure 3C (tones) and F (vowel). These results, in turn, allow us to quantify the degree of energy trading of the target that occurs for two-object stimuli.

Target energy trading

Figure 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 Darwin 1995; McAdams et al. 1998). In general, if any form of energy trading holds, then the data from the two-object conditions should fall on some negatively sloped contour in this plot.

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 (example in Fig. 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 5B.

Consistent with past studies, the two-object stimuli do not obey energy conservation. 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 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 (Darwin 1995; McAdams et al. 1998). 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), showing that energy trading also fails, just as in the previous study using similar, anechoic stimuli (Shinn-Cunningham et al. 2007).

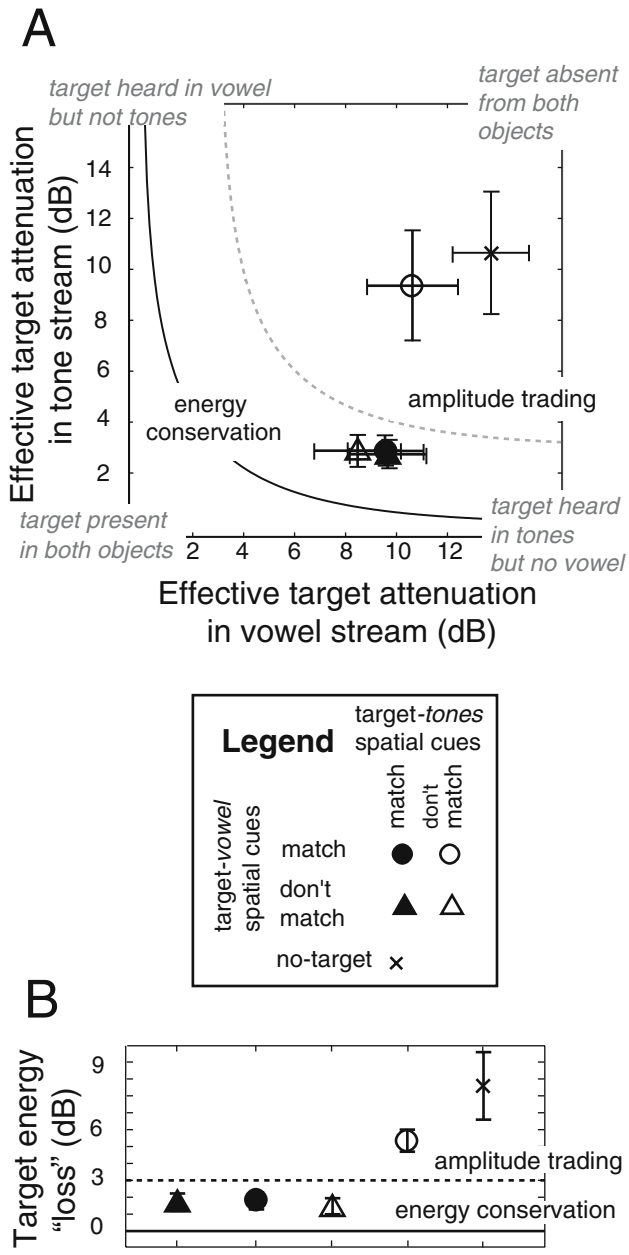

FIG. 5. **A** Scatter plot of the effective target attenuation in the two-object vowel judgments (abscissa) vs in the tone judgments (ordinate). The *solid line* shows where results would fall if energy conservation holds. The *dotted line* shows the expected relationship if target amplitude, rather than energy, trades (corresponding to a loss of 3 dB of the target energy). **B** Lost target energy calculated as the difference between the physical target energy and the sum of the target energy perceived in the tones and vowel.

DISCUSSION

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 (Shinn-Cunningham et al. 2007).

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, energy trading fails here, just as in our previous study (Shinn-Cunningham et al. 2007). 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 the 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 one-object tone stimuli. In the current study, accuracy was reduced, with overall percent correct around 92% for one-object tone prototypes. This result undoubtedly reflects the fact that the reverberant energy not only reduces the reliability of spatial cues, but also smears out the spectrotemporal content of the stimuli. In the current reverberant stimuli, the energy from the proceeding tones extends 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, one-object tone streams, increasing the likelihood of labeling the galloping prototype as even (about 3% of the time in the current study compared to <1% of the time in the anechoic study; see Shinn-Cunningham et al. 2007) and of labeling the even prototype as galloping (about 12% of the time in the current study compared to <1% of the time in the anechoic study).

The uncertainty about spectrotemporal 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 (Shinn-Cunningham et al. 2007).

Implications of the failure of energy trading

The current results support the conclusion that the total perceived energy that an ambiguous element contributes to different objects in the scene is less than the physical energy in the stimulus (Darwin 1995; McAdams et al. 1998). Moreover, as in our previous study in anechoic space, energy trading also fails: the "lost" target energy varies with spatial configuration. Thus, as discussed previously (Shinn-Cunningham et al. 2007), peripheral explanations cannot account for the failure of energy trading that occurs here and in previous experiments.

We suggest that the competition between objects for the target at some relatively central stage of auditory processing affects how much target energy is perceived in an attended object (either the tones or vowel). In this scenario, each object that tries to "own" the target suppresses the contribution of the target in the competing object. This kind of across-object inhibition normally would work to determine how much of the ambiguous element is perceived in each competing object. However, if the balance of this competitive inhibition is relatively equal and strong, the target may be effectively suppressed in both objects, no matter which object is in the perceptual foreground. This kind of explanation assumes that the loss of perceived target energy determined by the competition between objects is fixed, independent of attention.

However, a number of perceptual studies show that attention affects the perceptual organization of sound (Carlyon et al. 2001; Cusack et al. 2004). Specifically, these studies find that perceptual organization changes over time and that this buildup process is either initiated or reset when attention is directed towards an object. These observations are more consistent with an alternative possibility in which the perceptual organization of the scene

changes, depending on which object a listener attends (Shinn-Cunningham et al. 2007). The current results are consistent with the idea that the object being attended determines what grouping rules are most influential on object formation. In the current results, perception of the tone stream is more strongly modulated by spatial cues than perception of the vowel. The tone stream is primarily organized sequentially, where spatial cues have a strong effect; the vowel is primarily organized by simultaneous grouping, where spatial cues play a weak role. Thus, the current results are consistent with the idea that spatial cues are weighted heavily in organization of the scene when attending to a sequential object, but less influential when attending to an object composed of simultaneous elements.

Interpreted this way, it may be that the auditory system favors efficient processing over veridical parsing of the scene (Shinn-Cunningham et al. 2007). Rather than trying to analyze all sources in a sound mixture and finding "the" organization of the entire scene, the object in the foreground may be the only object that is formed in detail. Scene analysis may depend on different strategies for parsing the scene, depending on which object is attended. Thus, different cues for object formation may be weighted differently, depending on what object is attended.

SUMMARY

- Reverberant energy, which reduces the reliability of spatial cues, also appears to reduce the influence of spatial cues on perceptual organization of the auditory scene.
- Although reverberation reduces their influence, spatial cues nonetheless alter the perceived content of objects in the scene.
- As in past studies, the sum of the target energy perceived in competing objects in a scene changes with spatial configuration, showing that perceptual organization does not obey energy trading.
- Consistent with past results in anechoic space, spatial cues that oppose the perceptual organization that would be heard when all objects are in the same location lead to a seemingly paradoxical percept in which an audible target tone does not significantly contribute to the perceived content of either object in the scene.
- Either competing simultaneous and sequential grouping cues suppress ambiguous target energy, or the way in which the auditory scene is organized changes, depending on what object a listener attends.

ACKNOWLEDGEMENTS

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

REFERENCES

Best V, Gallun FJ, Ihlefeld A, Shinn-Cunningham BG. The influence of spatial separation on divided listening. J. Acoust. Soc. Am. 120:1506–1516, 2006.

Blaubert J. Spatial Hearing (2e). MIT Press, Cambridge, MA, 1997.

Bregman AS. Auditory scene analysis: The perceptual organization of sound. MIT Press, Cambridge, MA, 1990.

Carlyon RP, Cusack R, Foxton JM, Robertson IH. Effects of attention and unilateral neglect on auditory stream segregation. J. Exp. Psychol. Hum. Percept. Perform. 27:115–127, 2001.

Culling JF, Hodder KI, Toh CY. Effects of reverberation on perceptual segregation of competing voices. J. Acoust. Soc. Am. 114:2871–2876, 2003.

Cusack R, Deeks J, Aikman G, Carlyon RP. Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J. Exp. Psychol. Hum. Percept. Perform. 30:643–656, 2004.

Darwin CJ. Perceiving vowels in the presence of another sound—constraints on formant perception. J. Acoust. Soc. Am. 76:1636–1647, 1984.

Darwin CJ. Perceiving vowels in the presence of another sound: A quantitative test of the "Old-plus-new" Heuristic. In: Sorin C, Mariani J, Meloni H, Schoentgen J (eds) Levels in Speech Communication: Relations and Interactions: A Tribute to Max WajskopAmsterdam, Elsevier, pp. 1–12, 1995.

Darwin CJ. Auditory grouping. Trends Cogn. Sci. 1:327–333, 1997.

Darwin CJ, Hukin RW. Auditory objects of attention: The role of interaural time differences. J Exp. Psychol. Hum Percept. Perform. 25:617–629, 1999.

Darwin CJ, Hukin RW. Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. J. Acoust. Soc. Am. 108:335–342, 2000a.

Darwin CJ, Hukin RW. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. J. Acoust. Soc. Am. 107:970–977, 2000b.

Freyman RL, Helfer KS, McCall DD, Clifton RK. The role of perceived spatial separation in the unmasking of speech. J. Acoust. Soc. Am. 106:3578–3588, 1999.

Griffiths TD, Warren JD. What is an auditory object? Nat. Rev. Neurosci. 5:887–892, 2004.

Kohler W. Gestalt psychology: An introduction to new concepts in modern psychology. Liverright Pub. Corp, New York, 1947.

Kubovy M, Van Valkenburg D. Auditory and visual objects. Cognition. 80:97–126, 2001.

Lin I-F, Streeter T, Shinn-Cunningham BG. Trading directional accuracy for realism. Proceedings of the Human-Computer Interaction International 2005/1st International Conference on Virtual Reality. 22–27 July 2005.

MCADAMS S, BOTTE MC, DRAKE C. Auditory continuity and loudness computation. J. Acoust. Soc. Am. 103:1580–1591, 1998.

MACMILLAN NA, CREELMAN CD. Detection theory: A user's guide, 2nd Edition, Erlbaum, Mahwah, NJ, 2005.

ROCK I, PALMER S. The legacy of gestalt psychology. Sci. Am. 263:84–90, 1990.

SHAMMA S. On the role of space and time in auditory processing. Trends Cogn. Sci. 5:340–348, 2001.

SHINN-CUNNINGHAM BG. Influences of spatial cues on grouping and understanding sound. Forum Acusticum. Budapest. 2005.

SHINN-CUNNINGHAM BG, KOPCO N, MARTIN TJ. Localizing nearby sound sources in a classroom: Binaural room impulse response. J. Acoust. Soc. Am. 117:3100–3115, 2005.

SHINN-CUNNINGHAM BG, LEE AKC, OXENHAM AJ. A sound element gets lost in perceptual competition. Proc. Natl. Acad. Sci. U. S. A. 104:12223–12227, 2007.

VAN VALKENBURG D, KUBOVY M. In defense of the theory of indispensable attributes. Cognition. 87:225–233, 2003.

WERTHEIMER M. Untersuchunger zur lehre von der gestalt ii. Psychol. Forsch.ergeb. 4:301–350, 1923.