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.

ME H D

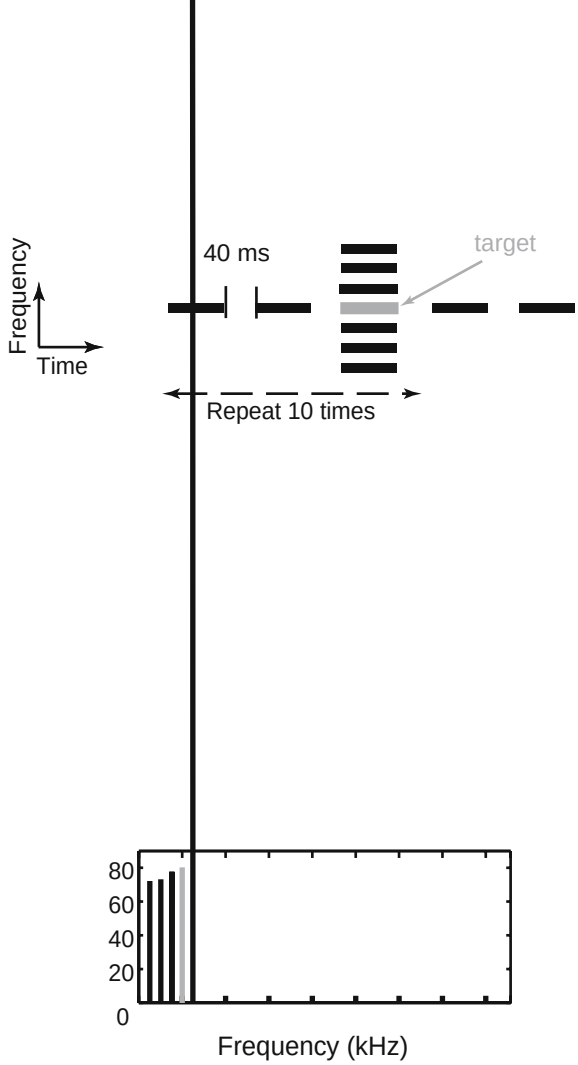
S b c

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.

S

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.



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).

E

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,

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 $'$ score,

E L

S b c c

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 ( ) and the data fit ( ) 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.

R c ( )

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 ofRor for

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

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.

Ta a

Figure 5

compared to our companion study using anechoic cues (Shinn-Cunningham et al. 2007).

Spatial cues caused changes in perceptual organiza-

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.

U MMA

- 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.

ACKN LEDGEMEN

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.

EFE ENCE

Best V, G