Disentangling the effects of spatial cues on selection and formation of auditory objects[a)]

Antje Ihlefeld and Barbara Shinn-Cunningham[b)]
Auditory Neuroscience Laboratory, Boston University Hearing Research Center, 677 Beacon Street, Boston, Massachusetts 02215

(Received 19 December 2006; revised 8 July 2008; accepted 16 July 2008)

When competing sources come from different directions, a desired target is easier to hear than when the sources are co-located. How much of this improvement is the result of spatial attention rather than improved perceptual segregation of the competing sources is not well understood. Here, listeners' attention was directed to spatial or nonspatial cues when they listened for a target masked by a competing message. A preceding cue signaled the target timbre, location, or both timbre and location. Spatial separation improved performance when the cue indicated the target location, or both the location and timbre, but not when the cue only indicated the target timbre. However, response errors were influenced by spatial configuration in all conditions. Both attention and streaming contributed to spatial effects when listeners actively attended to location. In contrast, when attention was directed to a nonspatial cue, spatial separation primarily appeared to improve the streaming of auditory objects across time. Thus, when attention is focused on location, spatial separation appears to improve both object selection and object formation; when attention is directed to nonspatial cues, separation affects object formation. These results highlight the need to distinguish between these separate mechanisms when considering how observers cope with complex auditory scenes. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2973185]

PACS number(s): 43.66.Dc, 43.66.Pn, 43.66.Qp, 43.66.Ba [RLF] Pages: 2224–2235

I. INTRODUCTION

Attention is critical in enabling us to select important information from the overwhelming flow of events that continuously reaches our senses (Desimone and Duncan, 1995; Pashler, 1998). Because attention is often object based (O'Craven *et al.*, 1999; Scholl, 2001; Busse *et al.*, 2005; Cusack *et al.*, 2000; Shinn-Cunningham, 2008), the way we organize sensory inputs into perceptual objects is likely to affect how attention can modulate perception in all modalities (Darwin *et al.*, 2003; Wolfe *et al.*, 2003; Shomstein and Yantis, 2004; Serences *et al.*, 2005). Here, we examined how attention to spatial and/or nonspatial auditory cues can affect performance in a speech identification task with two concurrent talkers.

Spatial cues are critical both for forming visual objects and selecting objects from a complex visual scene (e.g., see Egly *et al.*, 1994; Knudsen 2007). The important role of spatial cues on visual object selection and object formation is not surprising when one considers how visual sensory inputs are encoded. The retina is topographically organized and directly encodes two-dimensional visual spatial information in parallel channels comprised of distinct neural populations. Visual objects are primarily determined by local spatial structure (e.g., edges define the boundaries of connected regions); higher-order spatial and nonspatial features determine which regions belong to one object.

In contrast, sound from all source directions adds acoustically before entering the ear, where the cochlea processes inputs in distinct frequency channels. Unlike visual spatial information, all auditory spatial information is computed from the sound mixture reaching the ears. Moreover, whereas even a static two-dimensional visual scene is often rich in information, auditory information is conveyed by changes in sound over time.

Perhaps as a result, spectrotemporal sound structure rather than spatial information dominates how a sound mixture is segmented over short time scales (i.e., how auditory objects are formed at the syllable level; Kubovy, 1981; Culling and Summerfield, 1995a; Darwin, 1997; Darwin, 2008; Shinn-Cunningham, 2008; Ihlefeld and Shinn-Cunningham, 2008). Higher-order features (e.g., timbre, pitch, and perceived location) are thought to determine how these local segments are organized across longer time scales to form auditory "streams" (Bregman, 1990; Darwin, 1997; Deutsch, 1999; Darwin and Hukin, 2000; Shinn-Cunningham, 2008). In general, masking can be reduced if target and masker are dissimilar in one or more attributes, including fundamental frequency, timbre (e.g., vocal tract length and intonation), overall signal intensity, and perceived spatial location (Culling *et al.*, 1994; Culling and Summerfield, 1995b; Freyman *et al.*, 2005; Rakerd *et al.*, 2006).

When listening for a target voice in a mixture of other voices, short-term segmentation of concurrent speech is often not the main factor limiting performance, presumably because of the rich spectrotemporal structure of speech. In particular, at least for closed-set speech identification in the

[a)]Portions of this work were presented at the 2006 Mid-Winter meeting of the Association for Research in Otolaryngology.

[b)]Also at Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA 022105; electronic mail: shinn@cns.bu.edu

 0001-4966/2008/124(4)/2224/12/$23.00 © 2008 Acoustical Society of America

Coordinate Response Measure (CRM) paradigm (Bolia *et al.*, 2000), listeners rarely report words that are not present in the sound mixture (what we will henceforth label as *drop* errors [e.g., see Kidd *et al.* (2005a)]. Instead, listeners usually err by either reporting a mixture of words from different sources (*mix* errors) or reporting all of the words from the wrong source (*masker* errors).

Many studies demonstrate that spatial separation of target and masker reduces response errors (e.g., Kidd *et al.*, 2005a; Best *et al.*, 2006; Ihlefeld and Shinn-Cunningham, 2008; Brungart, 2001; Brungart *et al.*, 2001; Brungart and Simpson, 2004). When the dominant form of interference is energetic masking, such that the neural representation of the masker energy at the auditory periphery occludes that of the target energy (e.g., see Durlach *et al.*, 2003; Kidd *et al.*, 2008), spatial differences between target and masker can improve target intelligibility, even when the spatial attributes of target and/or masker are ambiguous or inconsistent across frequency (Edmonds and Culling, 2005a; Edmonds and Culling, 2005b; Culling *et al.*, 2006). However, studies showing this effect all employed maskers that differed from the target in some nondirectional attributes (so that the target was easy to select from the mixture). This suggests that when the dominant interference is energetic masking, attention to a particular direction is not necessary to exploit spatial cues. Other studies show that when target and masker are perceptually similar or when the listener is uncertain about target features (i.e., when "informational masking" is the dominant form of perceptual interference, see Durlach *et al.*, 2003; Kidd *et al.*, 2008), directing spatial attention can improve target identification and/or detection (Arbogast and Kidd, 2000; Freyman *et al.*, 2005; Rakerd *et al.*, 2006; Kidd *et al.*, 2005a).

While there are many studies showing benefits of spatially separating target and masker signals, no past study has disentangled whether these improvements come solely from the listener directing attention to the target location or whether spatial continuity can contribute to performance through automatic improvements in streaming (linking together words from one source over time) even when attention is not spatially directed (Shinn-Cunningham, 2008). Specifically, spatial separation may increase the likelihood that the keywords from a sound source are linked together properly across time. Such automatic streaming should reduce mix errors, even when listeners attend to a nonspatial attribute of the target voice or when they happen to attend to another voice in the acoustic mixture. Similarly, even without spatial attention, spatial separation may make it easier to selectively attend to the keywords that have a desired timbre, which could reduce both mix and masker errors.

Here, we examined whether the rates at which different kinds of response errors occurred were similar when listeners directed their attention to spatial and timbral cues, and how overall performance and errors varied with spatial separation of the target and masker. We manipulated what features subjects attended to when listening for a target message played simultaneously with a concurrent masker. In order to deemphasize the role of nonspatial higher-order acoustic cues (such as fundamental frequency or vocal tract length), we used sine-wave vocoded speech.[1] Both target and masker stimuli were derived from utterances of the same talker. The keywords in these stimuli were nearly synchronous and possessed no strong pitch.

Results suggest that the increasing spatial separation of the competing talkers improves the ability to *select* the desired source only when a listener is attending to space. In contrast, spatial separation may improve streaming (the linking of sound from one source across time) both when attention is spatially directed and when attention is directed to a nonspatial feature.

II. METHODS

A. Subjects

Nine normal-hearing fluent speakers of American English (ages 20–32) were paid to participate. All subjects gave written informed consent (as approved by the Boston University Charles River Campus Institutional Review Board) before participating in the study.

B. Stimuli

Raw speech stimuli were derived from the CRM corpus (see Bolia *et al.*, 2000), which consists of sentences of the form "Ready ⟨call sign⟩ go to ⟨color⟩ ⟨number⟩ now." Target and masker [⟨color⟩ ⟨number⟩] phrases were extracted from the original utterances by time windowing. ⟨Color⟩ was one of the set (white, red, blue, and green). ⟨Number⟩ was one of the digits between 1 and 8, excluding the two-syllable digit seven. Five (arbitrarily selected) instances of the word "ready" were also extracted to serve as cue words.[2] The cue word was processed in the same way as the target and masker phrases.

In each trial, two different [⟨color⟩⟨number⟩] phrases were used as sources. The numbers and colors in the competing utterances were randomly chosen but constrained to differ from each other in each trial. In each trial, the designated target message was preceded by the cue word "ready," chosen randomly from the five instances. The cue word, which was approximately 300 ms long, was concatenated with the target phrase without any inserted delay. In order to minimize differences between concurrent messages, the same talker was used for both phrases (talker 0 was chosen because it is the talker with the smallest variance in speaking rate in the CRM corpus).

In order to reduce peripheral interference between the competing messages and to better isolate object- and attention-related effects on performance, the raw target, masker, and cue words were processed to produce intelligible, spectrally sparse signals that used nonoverlapping frequency bands (Arbogast and Kidd, 2000; Shinn-Cunningham *et al.*, 2005a). Each raw target phrase, masker phrase, and cue word was filtered by ten one-third-octave wide bandpass filters (fourth-order Butterworth filter) with center frequencies spaced linearly on a logarithmic scale between 250 and 3342 Hz.[3] The Hilbert envelope of each band was used to amplitude-modulate a sinusoidal carrier whose frequency matched the corresponding band's center frequency.

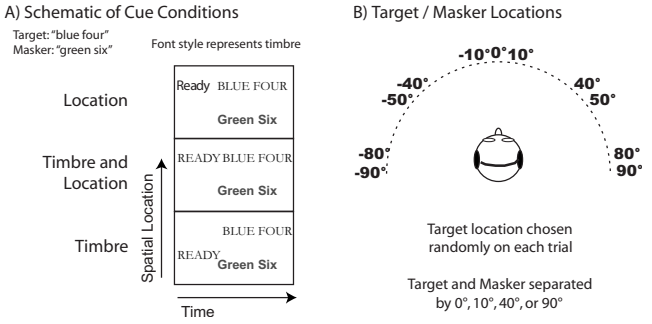

FIG. 1. Illustration of the experimental conditions. (A) Diagram of the cue conditions. The target message ("blue four") and the cue ("ready") match either in location (represented by vertical displacement), timbre (represented by font style), or timbre and location. The masker message ("green six") never matches the timbre of the cue or target. (B) The diagram of the spatial locations tested. The target was equally likely to come from each of the possible locations. The separation between the target and masker varied from trial to trial and was either 0°, 40°, or 90°.

On each trial, five of the resulting amplitude-modulated sinusoids were summed to produce the spectrally sparse target signals; the remaining five bands were summed to create the masker. For each subject, there were three possible sets of timbre from which target and masker signals were constructed. Each subject-specific set of timbres consisted of two combinations of five frequency bands each (randomly set for each subject at the start of the experiment). To ensure that spectral content was comparable across target and masker signals and across the three sets of timbres, each frequency band combination consisted of three from the lower six frequency bands (250–1056 Hz) and two from the upper four bands (1408–3342 kHz). On each trial, one of the three sets of timbres was selected, and one of its frequency band combinations was used to create the target; the other frequency band combination from the same set of timbres was used to generate the masker.

On a given trial, the frequency bands for the cue word "ready" either matched the target (in the timbre and timbre-and-location conditions) or were randomly chosen with the constraint that they did not match either target or masker [in the location condition; see Fig. 1(a)]. As stated above, there were three possible sets of timbres for each subject, a number chosen (somewhat arbitrarily) so that the number of possible timbres matched the number of possible spatial separations between target and masker.

C. Spatial synthesis

The broadband root-mean-square (rms) energy of the spectrally sparse target and masker signals was equalized. Then the equalized signals were processed to produce spatial cues by filtering them with pseudo-anechoic head-related transfer functions (HRTFs) measured on a Knowles Electronics Manikin for Acoustic Research (KEMAR) [for details, see Shinn-Cunningham *et al.*, (2005a, 2005b)]. HRTFs were measured in the horizontal plane containing the ears for sources at a distance of 1 m and at various azimuthal locations (±90°, ±80°, ±50°, ±40°, ±10°, and 0°).[4] The resulting binaural signals contained all of the appropriate spatial auditory cues for a source from the simulated location.

The target location was chosen randomly from trial to trial and was equally likely to be from any of the 11 locations. On each trial, the angular separation between target and masker was randomly chosen [either 0°, 10°, 40°, or 90°; however, in the location condition, the separation was never 0°; see Fig. 1(b)]. The cue location was either the same as the target (in the location and timbre-and-location conditions) or chosen randomly to differ from both target and masker locations [in the timbre condition; see Fig. 1(a)]. Following spatial synthesis, the cue word (in quiet) was concatenated with the sum of the target and masker signals.

Although target and masker had equal broadband rms energy prior to spatial processing with HRTFs, the spatial processing introduced level differences in the presentation of the sources at each ear. To remove any possible artifacts caused by variations in overall loudness with spatial configuration, on each trial, the overall level of the spatially processed target and masker pair was randomly roved over a range of 10 dB (average level set to 65 dB SPL).

D. Procedures

Stimuli were digital/analog (D/A) converted, amplified using Tucker-Davis System 3 hardware, and presented over Sennheiser HD 580 headphones to subjects seated in a sound-attenuated chamber. Following each trial, subjects indicated the perceived target keywords using a graphical user interface (GUI), after which the GUI indicated the correct response.

Prior to the experiment, subjects were screened to ensure that they could identify the color and number of the spectrally sparse processed speech in quiet. All subjects achieved 90% correct or better over the course of 50 trials. At the beginning of each session, subjects completed two short (100-trial) tests to sensitize them to the stimulus timbres and locations, respectively. In these tests, a cue/target phrase was presented in quiet. In the first test, cue and target had the same timbre on half of the trials and different timbres on the other half of the trials. Similarly, in the second test, cue and target had the same location on half of the trials and different locations on the other half of the trials. For both tests, the locations and timbres of cue and target were randomly selected on each trial and differed from each other, encouraging subjects to focus on the feature of interest. Subjects were asked whether the cue and target had the same or different timbre or location (first and second tests, respectively).[5]

In each block, subjects were instructed to report the color and number associated with the cue word timbre, location, or timbre and location, ignoring the message of the masker. In the timbre condition, listeners were instructed as follows: "Listen to what 'ready' sounds like and report the color and number that sound similar to 'ready.'" In the location condition, subjects were instructed as follows: "Listen to where 'ready' is coming from and report the color and number that come from the same location as 'ready.'" In the timbre-and-location condition, subjects were instructed as follows: "Listen to what 'ready' sounds like and where 'ready' is coming from and report the color and number that sound similar to 'ready' and that come from the same loca-

tion as 'ready.'" After each trial, correct-answer feedback was provided. A trial was scored as correct, and subjects were given feedback that they were correct if and only if they reported both target keywords. After each 5 min block, subjects were given the opportunity to take a break.

Each subject completed four sessions on four different days. Each session consisted of the sensitization tests followed by 12 experimental blocks of 50 trials each. Within each 5 min block, the condition (timbre, location, and timbre-and-location) was fixed. Each session contained four blocks of each of the three cue conditions. Four consecutive blocks always had the same cue condition and the same instructions. The four blocks for a given cue condition were presented one after the other within a session, while the ordering of the conditions was separately randomized for each subject and session. Within each session, the selection of the source locations was balanced such that each source location was presented the same number of times. The first session was for training purposes only, and the results of that session were discarded. During the three experimental sessions, subjects performed 600 trials in each cue condition: 200 repetitions for each of three spatial separations in the location condition and 150 repetitions for each of four spatial separations in the timbre and timbre-and-location conditions.

III. RESULTS

A. Percentage responses

Let C_xN_x represent the response color and number, where x denotes when the subject reported a target (T) or a masker (M) keyword. Then responses can be categorized into one of five distinct types: correct responses (C_TN_T), masker errors (C_MN_M), C_TN_M mix errors, C_MN_T mix errors, and drop errors (in which one or both of the reported keywords were not present in either target or masker). The rates at which each of these different response types occurred were calculated separately for each subject, condition, and spatial separation. These rates were then analyzed in primary repeated-measures two-way analysis of variance (ANOVAs) with independent factors of cue condition and spatial separation (excluding the 0° separation, which was not performed in the location condition). When significant interactions between condition and spatial separation were found, separate secondary ANOVAs were run for each of the three cue conditions.

Given the four possible colors and seven possible numbers, the probability of making either a correct response (C_TN_T), masker error (C_MN_M), C_TN_M mix error, or C_MN_T mix error by chance is 1/28 or 4%, while the probability of making a drop error by chance is 84%. However, listeners rarely made drop errors; they nearly always responded with some mixture of keywords from the target and masker, as shown below.

Figure 2 shows the across-subject mean percent correct as a function of spatial separation. Error bars show the 95% confidence intervals around the mean (1.96 times the standard error of the mean across subjects). When listeners were cued about where to listen, overall performance improved with increasing spatial separation (solid and dashed lines in Fig. 2). In contrast, in the timbre condition, performance was essentially independent of spatial separation (dotted lines in Fig. 2). Listeners tended to be better at reporting the target when there were two redundant features (timbre and location) than when there was a single feature to attend to (in Fig. 2, the dashed line is consistently above the solid line, and it is roughly equal to or above the dotted line).

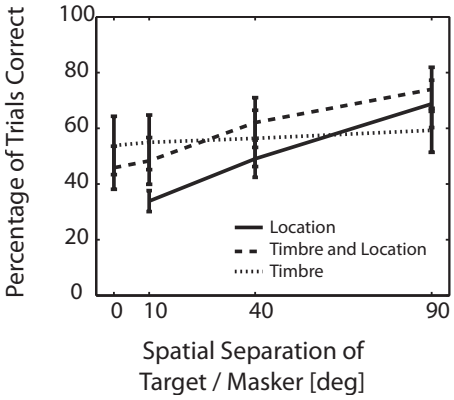

FIG. 2. Percent correct performance as a function of target and masker spatial separation. Performance improves with spatial separation of the target and masker for conditions in which the subjects know the target location ("attend location," solid line; "attend timbre and location," dashed line) but not when they are attending to a nonspatial feature ("attend timbre," dotted line). The across-subject mean in percent correct performance is shown as a function of spatial separation between the target and masker for conditions differing in which target features are known: location, timbre and location, and timbre. Error bars show the 95% confidence intervals around the mean.

The primary ANOVA of percent correct responses found a significant interaction between cue condition and spatial separation [$F(4,32)=25.374$, $p<0.001$], as well as significant effects of both main factors [$F(2,16)=7.599$ (cue condition), 127.683 (spatial separation); $p=0.005$ and $p<0.001$, respectively]. The follow-up ANOVAs found that for the timbre condition, the main effect of spatial separation on percent-correct performance was not significant [$F(2,16)=1.777$, $p=0.201$]. However, for both the location condition and the timbre-and-location condition, spatial separation significantly affected percent-correct performance [$F(2,16)=116.646$ (location) and 77.801 (timbre-and-location); $p<0.001$ for both tests]. The fact that overall performance improved with increasing spatial separation in the location and timbre-and-location conditions confirms that listeners can select a target based on location. In contrast, the fact that performance was independent of spatial separation in the timbre condition supports the idea that spatial separation helps overall performance only when listeners can direct attention to the target location. The spatial gain analysis in Sec. III C (below) considers the effect of spatial separation in more detail.

Figure 3 plots the across-subject average of the different error rates (with error bars showing the 95% confidence intervals around the mean) as a function of the spatial separation between target and masker [(a) masker errors, (b) drop errors, (c) C_TN_M mix errors, and (d) C_MN_T mix errors].

Masker errors (where listeners reported the wrong masker message) decreased with spatial separation in the location and timbre-and-location conditions [solid and dashed

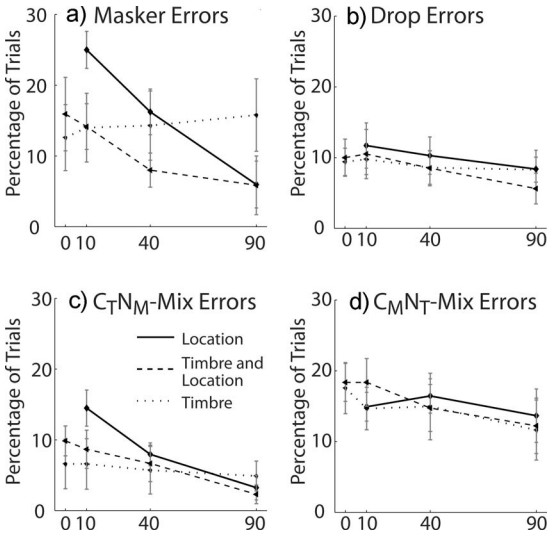

FIG. 3. Masker, C_TN_M mix, C_MN_T mix, and drop errors. Spatial separation reduces masker errors (reporting the masker) when listeners know the target location [solid and dashed lines in (a)] but not when they are instructed to attend the target timbre [dotted lines in (a)]. Spatial separation reduces the mix errors [improves the perceptual segregation of the target and masker over time; see (c) and (d)] and drop errors [decreases the likelihood of reporting words not present in the target or masker; see (b)] in all conditions. In all panels, the across-subject mean is shown as a function of spatial separation between the target and masker for each condition. Error bars show the 95% confidence intervals around the mean.

lines in Fig. 3(a)] but were independent of spatial configuration in the timbre condition [dotted line in Fig. 3(a)]. Overall, masker errors were the most common type of error, particularly in the timbre condition and/or at small spatial separations. Drop errors were relatively rare and decreased with increasing spatial separation for all three conditions [Fig. 3(b)]. C_TN_M mix errors were also relatively uncommon, showing that if a listener heard the first keyword correctly, they were unlikely to switch from target to masker and report the second keyword from the masker [Fig. 3(c)]. In the rare cases in which these errors occurred, their error rates tended to decrease with increasing spatial separation in the location and the timbre-and-location conditions but not in the timbre condition. It was more common to report a wrong initial keyword (report the masker color) and then to switch to the target number [C_MN_T, Fig. 3(d)]: the C_MN_T mix errors tended to decrease with increasing spatial separation in all three conditions [Figs. 3(c) and 3(d)].

The above summaries were generally supported by statistical analyses. For the masker errors [Fig. 3(a)], the primary ANOVA found a significant interaction between cue condition and spatial separation [$F(4,32)=20.193$; $p<0.001$] as well as significant main effects of both spatial separation [$F(2,16)=32.919$, $p<0.001$] and cue condition [$F(2,16)=5.559$, $p=0.04$, with Greenhouse–Geisser correction]. In the secondary ANOVA of the timbre condition, spatial separation did not significantly affect masker errors [$F(2,16)=0.938$, $p=0.377$, with Greenhouse–Geisser correction]. In both the location condition and the timbre-and-location condition, the main effect of spatial separation was significant [$F(2,16)=46.207$ (location) and 12.538 (timbre-and-location); $p<0.001$ and $p=0.001$, respectively].

The primary ANOVA of mix errors [Figs. 3(c) and 3(d)] found significant interaction terms [$F(4,32)=14.744$ (interaction, C_TN_M) and 2.953 (interaction, C_MN_T); $p<0.001$ and $p=0.035$, respectively] and a significant effect of spatial separation for both types of mix error [$F(2,16)=43.331$ (spatial separation, C_TN_M) and 6.43 (spatial separation C_MN_T); $p<0.001$ and $p=0.009$, respectively]. The main effect of cue condition was significant for C_TN_M mix errors [$F(2,16)=4.219$, $p=0.034$] but was not significant for C_MN_T mix errors [$F(2,16)=1.172$, $p=0.335$]. A secondary ANOVA analysis found that spatial separation had a significant effect on both types of mix error in the timbre-and-location condition [$F(2,16)=7.781$ (C_MN_T) and 28.464 (C_TN_M); $p=0.004$ and $p<0.001$, respectively], on C_TN_M mix errors in the location condition [$F(2,16)=49.438$, $p<0.001$], and on C_MN_T mix errors in the timbre condition [$F(2,16)=3.872$, $p=0.043$]. Spatial separation did not have a significant effect on either the C_MN_T mix errors in the location condition [$F(2,16)=1.882$, $p=0.184$ (location, C_MN_T)] or on the C_TN_M mix error in the timbre condition [$F(2,16)=1.39$, $p=0.278$].

A repeated-measure two-way ANOVA on the drop errors [Fig. 3(b)] found a significant effect of spatial separation [$F(2,16)=70.473$, $p=0.001$] and of cue condition [$F(2,16)=4.703$, $p=0.025$] but no significant interaction [$F(4,32)=1.492$, $p=0.228$].

B. Interim discussion

Results suggest a contribution of spatial cues to auditory object formation. In particular, in all conditions, the likelihood that listeners reported a mixture of target and masker words (as if the two messages were not perceptually distinct) tended to decrease with increasing spatial separation of target and masker. This suggests that perceptual separation of the target and masker improves with increasing spatial separation. However, there are a number of alternative explanations that could explain why spatial separation reduces the number of mix errors in all three conditions.

In the conditions in which the listener knows the target location, spatially directed attention can reduce mix errors. In particular, if listeners independently select each keyword based on its location, the probability of selecting both color and number correctly will increase with increasing spatial separation. As a result, both mix and masker errors will decrease (in the limit, if each keyword is selected properly based on its location with probability one, no mix errors will occur). It is difficult to judge from the pattern of response errors alone how much spatially directed attention contributes to the decline of mix errors with increasing spatial separation versus how much of this effect is due to automatic streaming induced by spatial separation, possibly even in the absence of spatially directed attention.

Similarly, in the timbre condition, there are a few possible explanations for the decrease in mix errors with increasing spatial separation that do not assume that spatial separation improves streaming of the target and masker. First, subjects could have been biased in their responses and attended to one side of space (for instance, by picking the

words that are better represented in the right ear). However, we separately analyzed the responses for leftward and rightward targets and did not find any consistent spatial bias for any of the subjects. Second, listeners could have simply responded by reporting the color and number keywords that were closest in space to the cue word "ready." However, there were no biases of this sort in the responses of any subjects. Third, listeners could arbitrarily pick either the target or the masker color and report the number from the same location. While the first two possibilities are not supported by the data, we cannot conclusively rule out this third possibility. However, listeners were specifically instructed to attend to the target timbre, not to the location. Moreover, listeners were able to select the proper keywords based on their timbre, as proven by the high percentage of trials in which they reported both keywords correctly in the timbre condition. Thus, on trials in which listeners were actively trying to attend to timbre, either (1) spatial separation helped listeners to properly stream the target and masker or (2) in the trials where attention to timbre failed, listeners attended to location, instead.

Even if listeners were sure that the color they reported was from the masker stream, there was little external motivation for them to switch to reporting the target color. Whether they made a switch (resulting in a mix error) or not (giving a masker error), they would receive a "wrong" score. However, listeners did make mix errors. Moreover, the patterns of these mix errors varied systematically with cue condition and spatial separation. For all three conditions, $C_M N_T$ mix errors were far more likely than $C_T N_M$ mix errors (or $C_T N_X$ drop errors). This suggests that even without being explicitly rewarded for reporting one of the two target keywords, listeners adopted a response strategy in which they reported as many target keywords as possible. In particular, the asymmetry in mix errors suggests that listeners often realized when they reported the wrong color and switched their attention to the target stream and reported the proper number. In the location and timbre-and-location conditions, masker errors were less likely than $C_M N_T$ mix errors at large spatial separations, and the likelihood of masker errors decreased with increasing spatial separation. This result suggests that spatial cues allowed the listener to detect their initial error and switch attention between streams when they knew the target location. In contrast, masker errors were roughly equally as likely as $C_M N_T$ errors for all spatial separations in the timbre condition, suggesting that when listeners attended to timbre, the spatial separation between target and masker was relatively unlikely to help them correct any initial error if they incorrectly reported the masker color.

In all three conditions, increasing spatial separation reduced drop errors [reporting words not from the target or the masker, Fig. 3(b)], suggesting that spatial separation also improves target audibility and/or short-term segmentation of auditory objects at the level of syllables, even when attention is not spatially directed.

C. Spatial gains

The influence of spatial separation on performance varied from subject to subject; however, an analysis of individual results shows that spatial separation influences performance in a consistent way across the population of subjects. For each subject in each condition and response category, we computed spatial gains. For percent correct performance, the spatial gain was computed as the increase in the probability of responding correctly when target and masker separation increased from 10° to 90°. Spatial gains in the error conditions were computed as the decrease in the percentage of each type of error as target and masker separation increased from 10° to 90°. From the resulting distribution of spatial gains, we computed the across-subject 95% confidence intervals of the mean spatial gain to determine whether there was a consistent effect of spatial separation across the subject population (see solid horizontal and vertical bars near the *x*- and *y*-axes in Fig. 4).

Figure 4 directly compares the spatial gains for different combinations of cue conditions for each individual subject (shown as individual points in each plot). Each row shows results for a different aspect of performance (overall percent

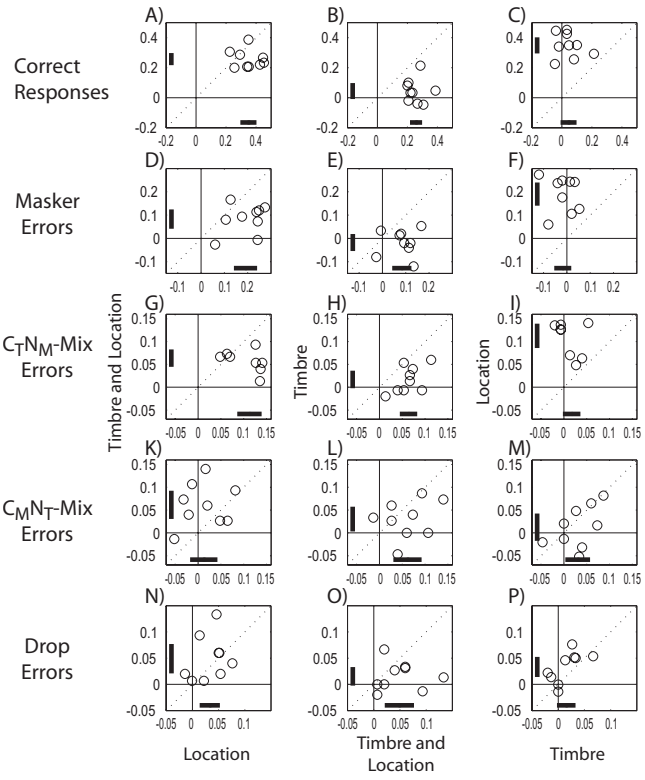

FIG. 4. Individual subject spatial gains (see text for definition) in correct performance, and masker, $C_T N_M$ mix, $C_M N_T$ mix, and drop errors contrasted across task conditions. Individual subjects all show large improvements in overall performance with spatial separation of the target and masker when they know the target location but not when they are attending to the target timbre. Overall, spatial separation of target and masker reduces response errors of all kinds (produces positive spatial gains); however, when listeners attend to the timbre, there is no significant reduction in masker errors. Each row compares the reduction in errors with spatial separation for one type of response error (correct performance, masker, $C_T N_M$ mix, $C_M N_T$ mix, and drop errors in top, second, third, fourth, and bottom rows, respectively). Within each panel, each point compares results for one of the nine individual subjects across two different task conditions. The horizontal and vertical bars within each panel show the 95% confidence intervals for the group mean of the spatial gain in the corresponding dimension. Left column: timbre-and-location vs location. Center column: timbre vs timbre-and-location. Right column: location vs timbre.

correct, masker errors, mix errors, and drop errors, respectively, from top to bottom in the figure). In each panel within a row, the spatial gains in two cue conditions are plotted against each other to allow an assessment of the relative size and direction of the spatial effects in the different cue conditions.

The absolute magnitude of the spatial gains differed from subject to subject (within each panel in Fig. 4, the 95% confidence intervals are on the order of 10%; see horizontal and vertical bars near abscissas and ordinates). However, there were consistent patterns in the relative size of the spatial gains in the different conditions across subjects (within each panel in Fig. 4, circles tend to cluster within one octant).

The spatial gain in overall performance is significantly greater than zero for all subjects and roughly equal in the location and timbre-and-location conditions [data fall near the diagonal and above and to the right of the origin in Fig. 4(a)], but there is no significant spatial gain in percent correct for the timbre condition [see Figs. 4(b) and 4(c), where spatial gains in the timbre condition cluster near zero].

The size of the spatial gains in the errors depends strongly on which target attribute the listener is instructed to attend to. The reduction in masker errors with spatial separation is consistently larger for location than for timbre-and-location conditions [data generally fall below the diagonal in Fig. 4(d)], while in the timbre condition there is no consistent spatial gain [spatial gains in the timbre condition in Fig. 4(e), vertical axis, and Fig. 4(f), horizontal axis, cluster near zero]. Spatial gains for mix errors tend to be positive in all conditions [most of the data points in Figs. 4(g)–4(m) are positive]. However, the size of this gain depends on the cue condition. The spatial gain is smaller in the timbre-and-location condition than in the location condition [data fall below the diagonal in Fig. 4(g)] and smaller for $C_T N_M$ mix errors in the timbre condition than in the timbre-and-location condition [data fall below the diagonal in Fig. 4(h)]. The spatial gains for $C_M N_T$ errors tend to be largest for the timbre-and-location condition [data tend to fall above the diagonal in Fig. 4(k) and below the diagonal in Fig. 4(l)] and slightly larger for the timbre condition than for the location condition [data tend to fall below the diagonal in Fig. 4(m)], although these trends are less consistent across subjects than the trends for the $C_T N_M$ mix errors. Finally, the spatial gains for drop errors are generally greater than zero for all three conditions [in Figs. 4(n)–4(p), data points tend to be positive] and are comparable in the different conditions [data in Figs. 4(n)–4(p) tend to fall around the diagonal].

IV. CONDITIONAL RESPONSE PROBABILITIES

In order to better understand how much of these spatial effects could be accounted for simply through spatially directed attention versus what improvements may come from automatic improvements in the perceptual segregation of the messages with spatial separation even when attention is not explicitly spatially directed, we analyzed conditional response probabilities (see the diagram in Fig. 5). This probabilistic analysis determined whether the color and number

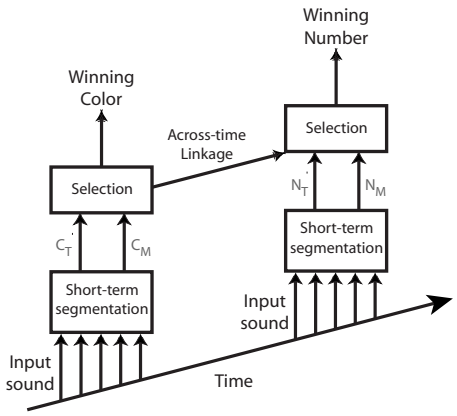

FIG. 5. Illustration of the model underlying the probabilistic analysis. Spatial cues may affect speech identification both through spatially directed attention and through automatic across-time linkage.

responses were independent of one another or whether the probability of responding with the correct number depended on whether listeners reported the target color (see also Cusack *et al.*, 2004). Specifically, we were interested in whether increasing the spatial separation between target and masker not only increases the likelihood of selecting the correct color and number when listeners know where the target is located, but also increases the probability of properly linking the keywords in the target and in the masker (increasing the perceptual segregation of the target and masker streams).

Target selection and across-time linkage of keywords are likely to occur at the same time, influencing each other. If each utterance is perceived correctly as one stream, the listener may only need to make one decision and report either both of the target words or both of the masker words. Instead, some mix errors occur, suggesting that, at least on some trials, the listener (1) makes two independent selections (selecting a color and then, separately, selecting a number), (2) decides to switch to the other stream upon hearing the color (i.e., decides, possibly incorrectly, that she was listening to the masker stream and therefore switches to the other stream), or (3) does not stream the target and masker properly and therefore makes a single decision, but the 'stream' she chooses to report is a mixture of target and masker.

If a listener independently selects color and number based on top-down attention and if there is no influence of across-time streaming (1, above), the initial choice of what color to report will be statistically independent of the second choice (the number reported). In other words, the probability of making a correct choice for the number will be the same for a given condition and spatial separation, independent of the color reported.[6] In contrast, if there is some automatic streaming of color and number across time, the probability of answering with the correct number should depend on whether or not the listener selected the correct color, regardless of what strategy the listener adopts (i.e., reporting the correct stream, switching attention to the other stream after reporting the color, or attending to an improperly formed stream). However, this analysis is not definitive. Dependencies between performance on color and number can arise for other reasons. Listeners' attention may lapse on some trials, so that the probability of missing the color and number both

increase together in those trials. Some trials may be inherently easier than others, even though they are analyzed together (for instance, a spatial separation of 10° may provide a stronger perceptual cue when target and masker are on the front rather than on the side). If so, the probability of being correct on the color and number will also be dependent. Nonetheless, it is worth examining whether such dependencies exist.

These ideas can be formulated through a simple probabilistic analysis. In general, whether or not color and number selections are independent,

$$P(N_T|C_T) = P(C_T N_T)/P(C_T), \quad (1)$$

where $P(N_T|C_T)$ is the conditional probability of reporting the target number in those trials where the target color was reported, $P(C_T N_T)$ is the probability of responding correctly, and $P(C_T)$ and $P(N_T)$ are the marginal probabilities of reporting the target color and number, respectively. Analogously, it is generally true that

$$P(N_T|C_M) = P(C_M N_T)/P(C_M). \quad (2)$$

If the selection of the number keyword is independent of the selection of the color keyword, then

$$P(C_T N_T) = P(C_T)P(N_T). \quad (3)$$

This also implies that when color and number choices are independent,

$$P(N_T|C_T) = P(N_T) = P(N_T|C_M). \quad (4)$$

Thus, if $P(N_T|C_T)$ is greater than $P(N_T|C_M)$ for a given stimulus configuration, this reflects a bias to report both target keywords over reporting a mix of target and masker keywords. The difference between $P(N_T|C_T)$ and $P(N_T|C_M)$ is expected to increase with increasing strength of perceptual across-time continuity, or streaming, between target color and target number.

For each subject, condition, and spatial separation, $P(N_T|C_T)$ and $P(N_T|C_M)$ were estimated from the observed percentages of responses. Figure 6 plots the means of the individual subject estimates of $P(N_T|C_T)$ and $P(N_T|C_M)$ as a function of the spatial separation between target and masker for the various conditions (dotted and dashed lines, respectively).

For all cue conditions, on trials when listeners properly reported the target color, listeners were also very likely to report the number from the target stream [solid lines are consistently above 50% in Figs. 6(a)–6(c)]. In contrast, listeners are much less likely to report the target number if they reported the masker color [dashed-dotted lines are between 36% and 70% in Figs. 6(a)–6(c)]. The mean difference between $P(N_T|C_T)$ and $P(N_T|C_M)$, averaged across the three spatial separations from 10° to 90° and averaged across listeners, is 40.1%, 29.7%, and 25.6% in the timbre, location, and timbre-and-location conditions, respectively. Thus, Eq. (4) is violated at each spatial separation and in each cue condition, proving that the color and number reports are not independent. Instead, the likelihood of reporting the target number depends on whether listeners reported the target or the masker color, with listeners more likely to get the target

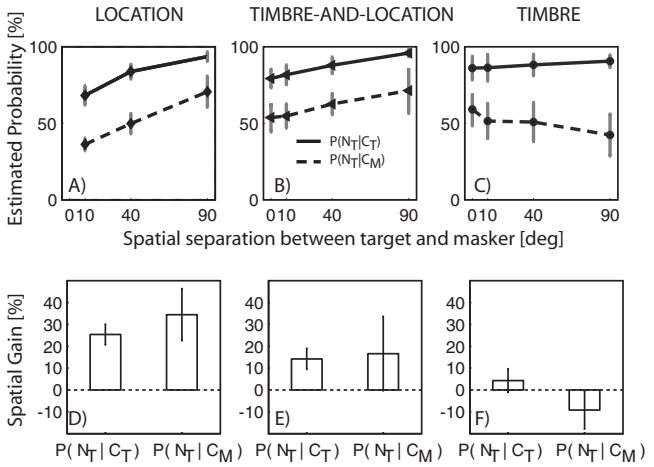

FIG. 6. Results of the probabilistic analysis from Eqs. (1)–(4). [(A)–(C)] Across-subject mean of the conditional probabilities, as a function of spatial separation (error bars show the standard error of the mean). [(C)–(E)] Across-subject mean of the percent spatial gain in conditional probabilities, respectively, with spatial separation for the three conditions (error bars show 95% confidence intervals).

number correct when they correctly report the target color. This suggests that listeners are likely to stay with the same stream for both color and number, either because they heard color and number from one source as a single perceptual unit or because the stimulus conditions were favorable for reporting both target keywords, regardless of whether they were streamed or not.

When listeners are told to attend to the target location, $P(N_T|C_T)$ and $P(N_T|C_M)$ both increase as the spatial separation between target and masker increases. In contrast, when listeners attend to the target timbre, the likelihood of reporting the target number is essentially independent of spatial separation or even tends to decrease with increasing spatial separation [$P(N_T|C_M)$, dashed line in Fig. 6(c)]. For each subject, we computed the difference in $P(N_T|C_T)$ when sources were 90° apart minus $P(N_T|C_T)$ when sources were 10° apart, and the corresponding effect of spatial separation on $P(N_T|C_M)$. The across-subject averages of these differences are shown in Figs. 6(d)–6(f) (error bars show the 95% confidence intervals).

In the location condition, spatial separation significantly increases the probability that listeners select the proper target number, both when they reported the target color and when they reported the masker color [both gains are positive in Fig. 6(d)]. In the timbre-and-location condition, spatial separation also improves the likelihood of reporting the proper color, but the effects are smaller, perhaps in part because of ceiling effects [in Fig. 6(e), gains are also positive]. In contrast, spatial separation has little effect on the probability of reporting the proper number when listeners attend to timbre [in Fig. 6(f), gains are not significantly different from zero], as if spatial separation improves streaming and makes it less likely that listeners switch streams.

Together, these results suggest that when attending to the target location (both in the location and the timbre-and-location conditions) and when target and masker are close together, listeners report the target and masker color and number in proper pairs and have difficulty detecting when

they made a wrong color selection. As a result, at small spatial separations, they fail to switch to the target stream when they initially listen to the masker. In contrast, in the timbre condition, when target and masker are close together, listeners are better able to select between the target and masker, presumably using nonspatial features, and are also more likely to switch streams when there is an initial error. This trend changes when target and masker are well separated in space: when listeners know the location of the target utterance (in the location and the timbre-and-location conditions), they are more likely to realize that they have reported the masker color and switch over to the target message than when they are attending to the target timbre. While this analysis cannot provide direct proof that the keywords of each source are more likely to be perceived in well formed streams when the competing sources are spatially separated, it is interesting to note that the conditional probabilities show patterns that differ across cue conditions. Future studies are needed to study more directly the automatic influence of spatial separation on object formation in the absence of spatially directed attention.

V. DISCUSSION

Most past studies of the effects of spatial cues on auditory perception show that listeners are more likely to report the target message correctly when competing sources arise from different directions compared to when they are at the target location (Bronkhorst, 2000). Multiple factors contribute to this effect, as considered below.

If a salient nonspatial acoustic attribute (e.g., timbre) differentiates the target from the masker (Brungart *et al.*, 2001; Darwin *et al.*, 2003) and the main form of interference is "energetic masking," spatial separation reduces peripheral interference between the target and masker and improves the effective target-to-masker energy ratio (Zurek, 1993; Shinn-Cunningham, 2005). This spatial effect does not require spatially directed attention (Edmonds and Culling, 2005a; Edmonds and Culling, 2005b; Culling *et al.*, 2006; Allen *et al.*, 2008) and appears to operate primarily by making it possible to detect near-threshold portions of the target (Shinn-Cunningham, 2005; Zurek, 1993).

Many studies point out that spatial cues carry little weight in enabling the segregation of sound locally in time at the level of syllables (Kubovy, 1981; Culling and Summerfield, 1995a; Darwin, 1997; Darwin and Hukin, 1997; Shinn-Cunningham *et al.*, 2007). While spatial cues provide little aid in segregating sources at a local time scale, spatial separation of competing sources improves the ability to selectively attend to a target when target location is the main cue differentiating the target from the other sources in the mixture (Freyman *et al.*, 2001; Gallun *et al.*, 2005; Shinn-Cunningham *et al.*, 2005a; Ihlefeld and Shinn-Cunningham, 2008). However, no past study has directly addressed whether spatial cues only allow a listener to select the proper object from the sound mixture or whether spatial separation automatically causes the competing messages to group properly across time and form more distinct auditory streams. Indeed, in discussing the effects of spatial cues on listening in complex environments, studies typically either acknowledge only one of these possibilities or use language that confounds the two.

The current results indicate that reports of color and number keywords are statistically dependent in all conditions. Moreover, when listeners were told to attend to the target location, the likelihood of selecting the proper keywords increased with increasing spatial separation for both the color and the number reports. Performance in the location and timbre-and-location conditions was biased toward reporting color and number keywords of one stream [as shown by differences in the conditional probabilities plotted in Figs. 6(a) and 6(b)]. An increase in the efficacy of spatially directed attention with increasing spatial separation between target and masker could account for the pattern of responses in these conditions. In contrast, when listeners were told to select the target based on a nonspatial feature, the likelihood of switching between streams decreases with increasing spatial separation, causing both an increase in masker errors and a decrease in mix errors. This suggests that the perceptual separation between the target stream and the masker stream increases with spatial separation (and perhaps that streaming itself becomes stronger as target and masker are displaced from each other in space). Of course, if timbre cues had been harder or easier to discern, the observed spatial effects might have differed. However, the current results show that when listeners were able to use timbre cues but were not perfect at it, spatial differences between target and masker influenced responses even when attention was not spatially directed. Overall, the effects of spatial separation were much smaller in the timbre condition than in the other two cue conditions. To the extent that these differences in the conditional probabilities are indicative of streaming, we conclude that the dominant contribution of spatial cues to understanding sources in a complex scene (with little energetic masking) comes through spatially directed attention, not through improvements in auditory streaming. As noted above, other possibilities for the dependency between color and number reports could also contribute to this effect. Further experiments are necessary to definitively address this question.

Recent physiological evidence suggests that spatial attention can modulate midbrain sensory responses (Winkowski and Knudsen, 2006). This result hints that sensory representations are altered by spatially directed attention in a manner that will tend to enhance the representation of a source at a desired location. This physiological mechanism could account for the observed psychophysical improvements in performance when listeners attended to a target at a known location.

The current results confirm that spatial attention can be directed toward the known location of a target, increasing the likelihood that the desired target source is selected and brought to the attentional foreground. Specifically, we see that spatially directed attention causes overall performance to improve with increasing spatial separation between target and masker when listeners attend to location or both timbre and location (Fig. 2). The importance of this selective spatial attention is greatest when space is the only acoustic feature

that a listener can use to identify the target (e.g., the spatial gain for masker errors is greater in the location and timbre-and-location conditions than in the timbre condition). When other features also aid in source selection, spatial information becomes wholly or partially redundant and thus less influential on performance.

It is worth noting that although having multiple cues identifying the target (timbre and location) reduces the influence of spatial configuration on performance, both of the redundant features contribute to the ability to report the target message (overall percent correct is generally higher for the timbre-and-location condition than for the corresponding single-feature conditions; see Fig. 2). At first glance, this result may seem at odds with visual theories that suggest that conjunctions of features do not provide large performance benefits (Treisman and Gelade, 1980; Wolfe and Bennett, 1997). However, most visual studies measure the time it takes to search for and detect a target in a complex visual scene. In contrast, in the current study, the auditory messages must be attended to and processed over time, so that the main factors affecting performance are the degree to which the target and masker are perceptually separated and how well a listener can maintain attention on the target, not how rapidly the target can be detected.

Among vision researchers, it has long been recognized that objects vie for attention in a complex scene and that top-down selection works in concert with bottom-up stimulus salience to determine which object will be processed and perceived (Desimone and Duncan, 1995; O'Craven *et al.*, 1999; Scholl, 2001; Serences and Yantis, 2006). A similar ability to selectively attend to a desired auditory stream normally enables communication in complex settings, where there are multiple talkers vying for attention (Shinn-Cunningham, 2008). This may help explain why listeners using hearing aids or wearing cochlear implants often find communication relatively easy in one-on-one settings but frequently experience communication difficulties or even communication breakdown in social settings such as at a restaurant (Gatehouse and Noble, 2004; Noble and Gatehouse, 2006; Harkins and Tucker, 2007). In a one-on-one conversation, there is no need to segregate the target source from a confusing sound mixture in order to process and understand it. However, in a noisy setting, sources must be segregated so that selective attention can be directed to whichever object is to be processed. In general, the acoustic cues critical for object formation (such as fine spectrotemporal resolution, robust timing information in the neural response in the auditory nerve, etc.) are degraded or absent in the signals many impaired listeners receive. These listeners may not be able to segregate and stream the sound mixture properly in complex settings and therefore may not be able to selectively attend to a desired sound source. Supporting this view, a high percentage of hearing-aid users is dissatisfied with their aids (Kochkin, 2005; but see also Edwards, 2007). Such descriptions are consistent with an inability to selectively attend to a desired source. This realization underscores the importance of further studies into the roles that various acoustic features (including location) play in both forming auditory objects and directing auditory attention.

VI. CONCLUSIONS

When trying to understand sources in a complex scene (with little energetic masking), spatial differences between target and masker improve the ability to select the target source from the mixture only when spatial location defines which object is the target. However, spatial separation affects listeners' responses as if spatial continuity helps to form streams even when attention is not spatially directed. The dominant contribution of spatial cues to listening selectively in a sound mixture comes through spatially directed attention, not through improvements in auditory streaming. Future work is needed to further delineate the different ways in which spatial cues affect object formation and object selection when listening in a sound mixture.

ACKNOWLEDGMENTS

Grants from the Office of Naval Research and the Air Force Office of Scientific Research to B.S.-C. supported this work. Jyrki Ahveninen, Virginia Best, Robert Carlyon, Steven Colburn, Frederick Gallun, Gerald Kidd, Nicole Marrone, Christine Mason, Richard Freyman, and three anonymous reviewers gave helpful feedback on earlier versions of this manuscript.

[1]Note that computational models of binaural processing can predict spatial release from masking for tasks dominated by energetic masking (e.g., Zurek, 1993). However, less is known about the role of spatial cues when informational masking limits performance (Kidd *et al.*, 2008). Here, we designed our stimuli to emphasize the role of informational masking and to de-emphasize the role of energetic masking (see also Arbogast *et al.*, 2002; Kidd *et al.*, 2005b; Gallun *et al.*, 2005; Brungart *et al.*, 2005; Shinn-Cunningham *et al.*, 2005a; Ihlefeld and Shinn-Cunningham, 2008).

[2]Each color-number pair and each cue word were carefully time windowed from recordings of talker 0 in the original corpus, using a routine programed in MATLAB 6.5 (both time-domain and short-term Fourier transform representations were used to monitor the quality of the resulting signal). Each color is preceded by the vowel /ə/ from "go to." This change from harmonic to inharmonic structure made it fairly easy to classify the beginning of the utterance. Similarly, each color in the CRM corpus is followed by "now," a syllable whose energy builds up slowly over time. Thus, even if a small bit of /n/ was left attached to the number, the intelligibility of the number was not adversely affected.

[3]The resulting spacing of the filters was roughly 42% of an octave. The center frequencies were 250, 333, 445, 593, 791, 1056, 1408, 1878, and 2505 Hz.

[4]In general, HRTFs differ across individual listeners. However, in the horizontal plane, the spatially dominant interaural difference cues in HRTFs tend to be grossly similar across listeners, allowing virtual spatial acoustic simulations that evoke sources at different lateral angles relative to the listener without using individualized HRTFs (Colburn and Kulkarni, 2005; Middlebrooks, 1999; Middlebrooks *et al.*, 2000). In the current study, all listeners perceived the sounds as having distinct lateral positions. However, absolute target and masker locations, not just angular separation, affect target intelligibility, e.g., due to differences in acoustic head shadow effects (Zurek, 1993; Shinn-Cunningham *et al.*, 2005a; Ihlefeld and Shinn-Cunningham, 2008). The perceptual difference between a target at 0° and a masker at 10° is not equivalent to that of a target at 80° and a masker at 90°. Because we collapsed performance across many different spatial source configurations with different better-ear effects, it is difficult to estimate how much performance should have improved with spatial separation due solely to the acoustic better-ear effects. However, statistically, the spatial stimulus features were the same across the three timbre, location, and timbre-and-location conditions. Therefore, performance in all three conditions was affected in similar ways by these better-ear acoustic advantages. Although it is desirable to measure performance as a function of both spatial separation between target and masker and their absolute locations, this would require considerably more data to be collected than was

done in the current study. To the extent that we find robust effects of spatial separation on performance, this averaging across different absolute locations is a source of uncontrolled variability. Thus, any conclusions we draw about the influence of spatial separation on performance are relatively conservative.

[5]The location cues used in the current study are much more clearly defined than the timbre cues. While the saliency of the timbre cues is difficult to establish, listeners could easily distinguish between the six different timbres during the sensitization tasks at the beginning of each session. Moreover, listeners achieved better-than-chance scores when attending only to timbre (see Sec. III).

[6]In general, it is difficult to directly compare number and color errors because the number of tokens, perceptual similarity of the tokens, and other attributes differ in the two sets. However, given how infrequent drop errors are, we believe that listeners are usually making binary decisions between responding with the target or the masker color and number. As a result, the decisions about which color and which number to report are probably more equal than one might expect based solely on the stimulus properties.

Allen, K., Carlile, S., and Alais, D. (**2008**). "Contributions of talker characteristics and spatial location to auditory streaming," J. Acoust. Soc. Am. **123**, 1562–1570.

Arbogast, T. L., and Kidd, G., Jr. (**2000**). "Evidence for spatial tuning in informational masking using the probe-signal method," J. Acoust. Soc. Am. **108**, 1803–1810.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

Best, V., Gallun, F. J., Ihlefeld, A., and Shinn-Cunningham, B. G. (**2006**). "The influence of spatial separation on divided listening," J. Acoust. Soc. Am. **120**, 1506–1516.

Bolia, R. S., Nelson, W. T., and Ericson, M. A. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

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

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D. S., and Simpson, B. D. (**2004**). "Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," J. Acoust. Soc. Am. **115**, 301–310.

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., and Kidd, G., Jr. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Busse, L., Roberts, K. C., Christ, R. E., Weissman, D. H., and Woldorff, M. G. (**2005**). "The spread of attention across modalities and space in a multisensory object," Proc. Natl. Acad. Sci. U.S.A. **102**, 18751–18756.

Colburn, H. S., and Kulkarni, A. (**2005**). "Models of sound localization," in *Sound Source Localization*, edited by A. N. Popper and R. R. Fay (Springer, New York).

Culling, J. F., Edmonds, B. A., and Hodder, K. I. (**2006**). "Speech perception from monaural and binaural information," J. Acoust. Soc. Am. **119**, 559–565.

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

Culling, J. F., and Summerfield, Q. (**1995b**). "The role of frequency modulation in the perceptual segregation of concurrent vowels," J. Acoust. Soc. Am. **98**, 837–846.

Culling, J. F., Summerfield, Q., and Marshall, D. H. (**1994**). "Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels," Speech Commun. **14**, 71–95.

Cusack, R., Carlyon, R. P., and Robertson, I. H. (**2000**). "Neglect between but not within auditory objects," J. Cogn Neurosci. **12**, 1056–1065.

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

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

Darwin, C. (**2008**). "Spatial Hearing and Perceiving Sources," in *Springer Handbook of Auditory Research: Auditory Perception of Sound Sources*, edited by W. A. Yost (Springer, New York), Vol. **29**.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

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. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

Desimone, R., and Duncan, J. (**1995**). "Neural mechanisms of selective visual attention," Annu. Rev. Neurosci. **18**, 193–222.

Deutsch, D. (**1999**). "Grouping mechanisms in music," in *The Psychology of Music*, 2nd ed., edited by D. Deutsch (Academic, San Diego).

Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987.

Edmonds, B. A., and Culling, J. F. (**2005a**). "The role of head-related time and level cues in the unmasking of speech in noise and competing speech," Acta Acust. **91**, 546–553.

Edmonds, B. A., and Culling, J. F. (**2005b**). "The spatial unmasking of speech: Evidence for within-channel processing of interaural time delay," J. Acoust. Soc. Am. **117**, 3069–3078.

Edwards, B. (**2007**). "The future of hearing aid technology," Trends Amplif. **11**, 31–45.

Egly, R., Driver, J., and Rafal, R. D. (**1994**). "Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects," J. Exp. Psychol. Gen. **123**, 161–177.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Freyman, R., Helfer, K., and Balakrishnan, U. (**2005**). "Spatial and spectral factors in release from informational masking in speech recognition," Acta Acust. **91**, 537–545.

Gallun, F. J., Mason, C. R., and Kidd, G., Jr. (**2005**). "Binaural release from informational masking in a speech identification task," J. Acoust. Soc. Am. **118**, 1614–1625.

Gatehouse, S., and Noble, W. (**2004**). "The speech, spatial, and qualities of hearing scale (SSQ)," Int. J. Audiol. **43**, 85–99.

Harkins, J., and Tucker, P. (**2007**). "An internet survey of individuals with hearing loss regarding assistive listening devices," Trends Amplif. **11**, 91–100.

Ihlefeld, A., and Shinn-Cunningham, B. G. (**2008**). "Spatial release from energetic and informational masking in selective listening," J. Acoust. Soc. Am. **123**, 4369–4379.

Kidd, G., Jr., Arbogast, T. L., Mason, C. R., and Gallun, F. (**2005a**). "The advantage of knowing where to listen," J. Acoust. Soc. Am. **118**, 3804–3815.

Kidd, G., Jr., Mason, C. R., and Gallun, F. J. (**2005b**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**, 982–992.

Kidd, G., Jr., Mason, C. R., Richards, V., Gallun, F. J., and Durlach, N. I. (**2008**). "Informational masking," in *Springer Handbook of Auditory Research: Auditory Perception of Sound Sources*, edited by W. A. Yost (Springer, New York), Vol. **29**.

Knudsen, E. (**2007**). "Fundamental components of attention," Annu. Rev. Neurosci. **30**, 57–78.

Kochkin, S. (**2005**). "Customer satisfaction with hearing instruments in the digital age," Hear. J. **58**, 30–39.

Kubovy, M. (**1981**). "Concurrent-pitch segregation and the theory of indispensable attributes," in *Perceptual Organization*, edited by M. Kubovy and J. R. Pomerantz (Lawrence Erlbaum, Associates, NJ), pp. 55–98.

Middlebrooks, J. C. (**1999**). "Individual differences in external-ear transfer functions reduced by scaling in frequency," J. Acoust. Soc. Am. **106**, 1480–1492.

Middlebrooks, J. C., Macpherson, E. A., and Onsan, Z. A. (**2000**). "Psychophysical customization of directional transfer functions for virtual sound localization," J. Acoust. Soc. Am. **108**, 3088–3091.

Noble, W., and Gatehouse, S. (**2006**). "Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the speech, spatial, and qualities of hearing scale (SSQ)," Int. J. Audiol. **45**, 172–181.
O'Craven, K. M., Downing, P. E., and Kanwisher, N. (**1999**). "fMRI evidence for objects as the units of attentional selection," Nature (London) **401**, 584–587.
Pashler, H. (**1998**). *Attention (Studies in Cognition)* (Psychology, Hove, UK).
Rakerd, B., Aaronson, N. L., and Hartmann, W. M. (**2006**). "Release from speech-on-speech masking by adding a delayed masker at a different location," J. Acoust. Soc. Am. **119**, 1597–1605.
Scholl, B. J. (**2001**). "Objects and attention: The state of the art," Cognition **80**, 1–46.
Serences, J. T., Liu, T., and Yantis, S. (**2005**). "Parietal mechanisms of switching and maintaining attention to locations, objects, and features," in *Neurobiology of Attention*, edited by L. Itti, G. Rees, and J. Tsotsos (Academic, New York), pp. 35–41.
Serences, J., and Yantis, S. (**2006**). "Selective visual attention and perceptual coherence," Trends Cogn. Sci. **10**, 38–45.
Shinn-Cunningham, B. G. (**2005**). "Influences of spatial cues on grouping and understanding sound," in Forum Acusticum 2005, Budapest, Hungary, p. CD.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.
Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, , and Larson, E. (**2005a**). "Bottom-up and top-down influences on spatial unmasking," Acta Acust. **91**, 967–979.
Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (**2005b**). "Localizing nearby sound sources in a classroom: Binaural room impulse responses," J. Acoust. Soc. Am. **117**, 3100–3115.
Shinn-Cunningham, B. G., Lee, A. K. C., and Oxenham, A. J. (**2007**). "Auditory nonallocation of a sound element lost in perceptual competition," Proc. Natl. Acad. Sci. U.S.A. **104**, 12223–12227.
Shomstein, S., and Yantis, S. (**2004**). "Control of attention shifts between vision and audition in human cortex," J. Neurosci. **24**, 10702–10706.
Treisman, A. M., and Gelade, G. (**1980**). "A feature-integration theory of attention," Cognit Psychol. **12**, 97–136.
Winkowski, D. E., and Knudsen, E. I. (**2006**). "Top-down gain control of the auditory space map by gaze control circuitry in the barn owl," Nature (London) **439**, 336–339.
Wolfe, J. M., and Bennett, S. C. (**1997**). "Preattentive object files: Shapeless bundles of basic features," Vision Res. **37**, 25–43.
Wolfe, J. M., Butcher, S. J., Lee, C., and Hyle, M. (**2003**). "Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons," J. Exp. Psychol. Hum. Percept. Perform. **29**, 483–502.
Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and I. Hochberg (College-Hill, Boston, MA).