Object continuity enhances selective auditory attention

Virginia Best, Erol J. Ozmeral, Norbert Kopčo, and Barbara G. Shinn-Cunningham*

Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA 02215

Edited by Eric I. Knudsen, Stanford University School of Medicine, Stanford, CA, and approved June 27, 2008 (received for review April 16, 2008)

In complex scenes, the identity of an auditory object can build up across seconds. Given that attention operates on perceptual objects, this perceptual buildup may alter the efficacy of selective auditory attention over time. Here, we measured identification of a sequence of spoken target digits presented with distracter digits from other directions to investigate the dynamics of selective attention. Performance was better when the target location was fixed rather than changing between digits, even when listeners were cued as much as 1 s in advance about the position of each subsequent digit. Spatial continuity not only avoided well known costs associated with switching the focus of spatial attention, but also produced refinements in the spatial selectivity of attention across time. Continuity of target voice further enhanced this buildup of selective attention. Results suggest that when attention is sustained on one auditory object within a complex scene, attentional selectivity improves over time. Similar effects may come into play when attention is sustained on an object in a complex visual scene, especially in cases where visual object formation requires sustained attention.

source segregation | auditory scene analysis | spatial hearing | streaming | auditory mixture

In everyday situations, we are confronted with multiple objects that compete for our attention. Both stimulus-driven and goal-related mechanisms mediate the between-object competition to determine what will be brought to the perceptual foreground (1, 2). In natural scenes, objects come and go and the object of interest can change from moment to moment, such as when the flow of conversation shifts from one talker to another at a party. Thus, our ability to analyze objects in everyday settings is directly affected by how switching attention between objects affects perception. Much of what we know about the effects of switching attention comes from visual experiments in which observers monitor rapid sequences of images or search for an item in a static field of objects (3, 4). Although these situations give insight into the time it takes to dis- and reengage attention from one object to the next, they do not directly explore whether there are dynamic effects of sustaining attention on one object through time.

In contrast to visual objects, the identity of an auditory object is intimately linked to how the content of a sound evolves over time. Moreover, the process of forming an auditory object is known to evolve over seconds (5–8). Given that attention is object-based (9, 10), this refinement in object formation may directly impact the selectivity of attention in a complex auditory scene. Specifically, sustaining attention on one object in a complex scene may yield more refined selectivity to the attended object over time. In turn, switching attention to a new object may reset object formation and therefore reset attentional selectivity. If so, the cost of switching attention between objects may not only be related to the time required to dis- and reengage attention (3, 11, 12) but also to the time it takes to build up an estimate of the identity of an object in a scene.

In the current study, we measured how switching spatially directed attention influenced the ability to recall a sequence of spoken digits. Five loudspeakers were distributed horizontally in front of the listener. Listeners identified sequences of four digits presented either from one loudspeaker or from a different loudspeaker chosen randomly on each digit, with visual cues indicating the target loudspeaker at each temporal position in the sequence. The remaining four loudspeakers presented simultaneous distracter digits. To explore whether continuity of a nonspatial feature influenced performance, we tested conditions in which the target voice changed from digit to digit (Exp. 1) as well as conditions under which the target voice was the same from digit to digit (Exp. 2). We investigated the time course of the cost of switching attention by testing four different overall rates of presentation, obtained by varying the silent delays inserted between each digit in the sequence (0, 250, 500, or 1,000 ms). To determine whether advance knowledge of where to redirect spatial attention ameliorated some of the cost of switching attention, we compared conditions under which the visual indicator of target location was turned on synchronously with the digits to those in which the visual cue preceded the auditory stimuli by the corresponding interdigit delay.

Results suggest that sustaining attention on one continuous auditory stream leads to refinements in selective attention over time. This refinement in selective attention is lost when attention switches to a new object, adding to the cost of switching attention between objects in a complex scene.

PSYCHOLOGY

Results

In both experiments at all interdigit delays, mean performance was better when the spatial location of the target did not change between digits (the "fixed" condition, F) than when listeners had to instantaneously switch attention to a new location for each digit (the "switching, LED synchronous" or SS condition) (Fig. 1, compare squares and circles). Moreover, performance in the SS condition tended to be better at slower presentation rates than at faster rates, when there was time to dis- and reengage spatially directed attention to the new digit position. The cost of switching spatial attention to a new location was thus positive in both experiments for all presentation rates and decreased with decreasing presentation rate (Fig. 2, circles). However, even at the slowest presentation rate, when there was 1 s of silence between subsequent digits, a switching cost was evident. In general, continuity of voice across digits (Exp. 2) (Figs. 1 *Lower* and 2 *Lower*) increased the cost of switching spatial attention compared with when voice quality changed between target digits (Exp. 1) (Figs. 1 *Upper* and 2 *Upper*). This improvement with voice continuity was especially pronounced at the shortest interdigit delays, where the temporal continuity between the target digits was greatest.

Author contributions: V.B., E.J.O., N.K., and B.G.S.-C. designed research; V.B., E.J.O., and N.K. performed research; V.B. and E.J.O. analyzed data; and V.B. and B.G.S.-C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*To whom correspondence should be addressed. E-mail: shinn@cns.bu.edu.

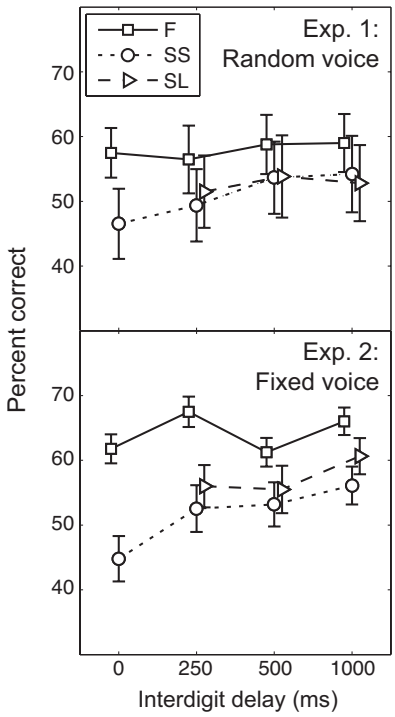

Fig. 1. Overall performance is best when spatial location is fixed between digits; moreover, even up to 1 s of advance knowledge of where to direct spatial attention does not overcome the cost of switching spatial attention. Across-subject mean scores (±SEM) for Exp. 1, where the target voice switches between digits (*Upper*), and Exp. 2, where the target voice is fixed across digits (*Lower*). Data are plotted as a function of interdigit delay for conditions F (squares and solid lines), SS (circles and dotted lines), and SL (triangles and dashed lines).

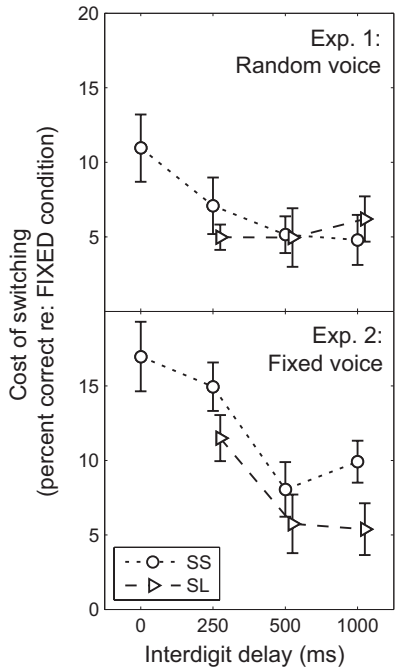

Fig. 2. The cost of switching spatial attention decreases with interdigit delay but is always positive. Moreover, the cost of switching tends to be greater when voice quality is fixed between digits (Exp. 2) (*Lower*) than when the voice changes between digits (Exp. 1) (*Upper*), especially at short interdigit delays. Each plot shows the across-subject mean difference in performance (±SEM) between condition F and each of the conditions SS (circles and dotted lines) and SL (triangles and dashed lines).

We predicted that providing spatial information in advance during the gaps between digits in the target sequence would eliminate the cost of switching spatial attention. In the "switching, LED leading (SL)" condition, the LEDs were turned on at the beginning of the silent gap preceding a target digit (see *Materials and Methods*). Surprisingly, when the target voice switched between target digits (Exp. 1), there was no reduction in the cost of switching spatial attention with advance warning about where the next target digit would be (Figs. 1 *Upper* and 2 *Upper*, compare circles and triangles). In contrast, when the target voice was fixed throughout a trial (Exp. 2), the cost of switching spatial attention was reduced, but not eliminated, by advance knowledge of target location (Figs. 1 *Lower* and 2 *Lower*, compare circles and triangles).

An examination of performance as a function of temporal position within the four-digit sequence revealed that the cost associated with switching the target location was not constant across time (Fig. 3). For the switching conditions, performance tended to be better for the first and last digit (see roughly U-shaped functions in Fig. 3, circles and triangles), consistent with typical primacy/recency effects on memory tasks. In contrast, for the F condition, the first digit was identified the most poorly and the remaining three digits were identified with increasing accuracy (Fig. 3, squares). In other words, the cost of switching spatially directed attention tended to increase throughout the duration of the sequence. This was particularly true for the faster rates when the target voice was held constant (Fig. 3 *Lower*, two left plots).

Statistical comparison of performance in the F and SS conditions revealed significant main effects of condition [$F(1, 4) = 19.6, P < 0.05$], delay [$F(3, 12) = 20.9, P < 0.001$], and temporal position [$F(3, 12) = 7.9, P < 0.005$], as well as significant two-way interactions between condition and delay [$F(3, 12) = 7.0, P < 0.01$], condition and temporal position [$F(3, 12) = 11.8, P < 0.05$], and delay and temporal position [$F(9, 36) = 2.4, P < 0.05$] in Exp. 1. In Exp. 2, significant main effects of condition [$F(1, 4) = 55.8, P < 0.005$], delay [$F(3, 12) = 22.4, P < 0.001$], and temporal position [$F(3, 12) = 10.7, P < 0.005$] were found. All two-way interactions were also significant [condition and delay: $F(3, 12) = 38.0, P < 0.001$; condition and temporal position: $F(3, 12) = 40.3, P < 0.001$; delay and temporal position: $F(9, 36) = 3.7, P < 0.005$], as was the three-way interaction [$F(9, 36) = 5.9, P < 0.001$].

The influence of the preceding visual cue (compare circles and triangles in Fig. 3) was negligible for all temporal positions in Exp. 1 but led to improved performance in Exp. 2 for later temporal positions and longer delays. This was supported by statistical comparison of performance under the SS and SL conditions, which found a significant main effect of delay in Exp. 1 [$F(2, 8) = 6.4, P < 0.05$] but no other significant effects or interactions, and significant main effects of condition [$F(1, 4) = 42.7, P < 0.005$] and delay [$F(2, 8) = 16.5, P < 0.005$] in Exp. 2, as well as significant two-way interactions between condition and temporal position [$F(3, 12) = 6.7, P < 0.01$] and delay and temporal position [$F(6, 24) = 3.8, P < 0.01$], and a significant three-way interaction [$F(6, 24) = 2.8, P < 0.05$].

An analysis of incorrect responses revealed that subjects had a tendency to report digits that were presented from loudspeakers adjacent to the target loudspeaker when they did not correctly identify the target (Fig. 4 *Upper*). Responses to masker digits decreased as the distance between the masker loudspeaker and the cued, target loudspeaker increased. The number of responses that did not correspond to either the target digit or one of the simultaneous masker digits was relatively low ("rand"; note that if subjects randomly guessed among all possible

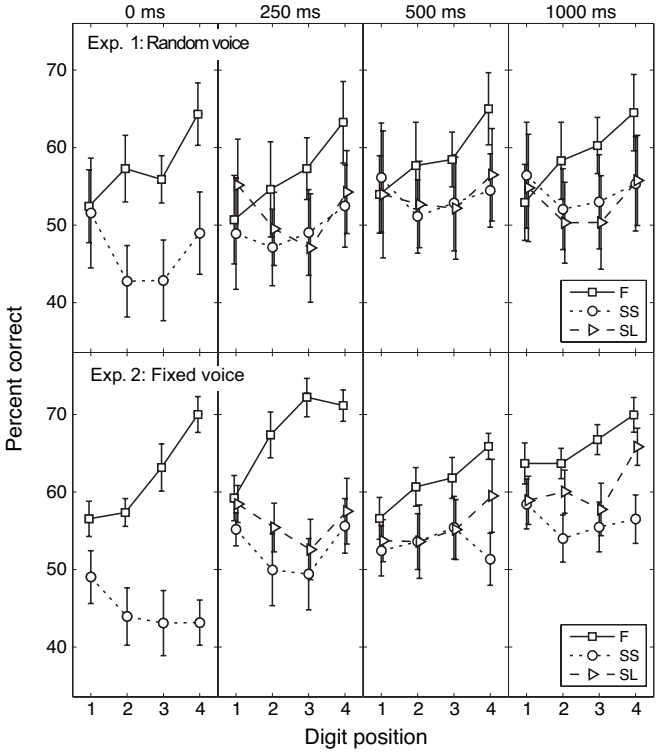

Fig. 3. When the target sequence is continuous in spatial location, performance improves from digit to digit, an effect that is enhanced when the target voice quality is continuous between digits. Across-subject mean scores (±SEM) as a function of temporal position for Exp. 1 (with random voice) (*Upper*) and Exp. 2 (with fixed voice) (*Lower*). The four plots within each row show data for the four different interdigit delays. Data are plotted as a function of temporal position within the target sequence for F (squares and solid lines), SS (circles and dotted lines), and SL (triangles and dashed lines).

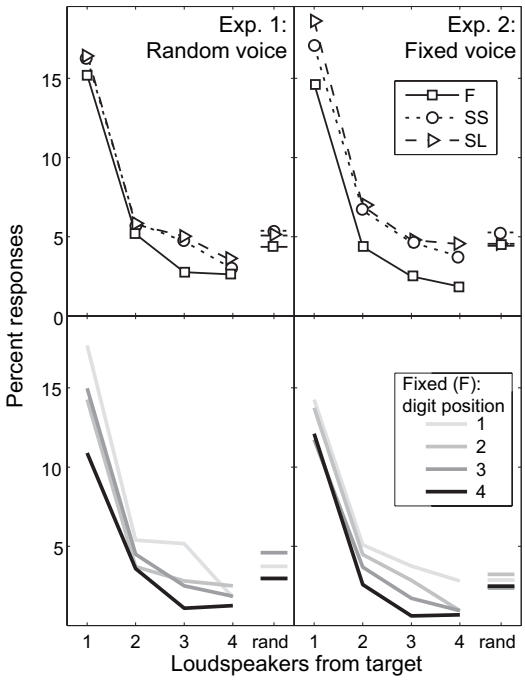

Fig. 4. Spatially directed attention filters out sources from the wrong direction, and this filtering becomes more refined over time when target location is fixed across digits. (*Upper*) Percentage of responses that corresponded to a digit presented from a nontarget loudspeaker are shown as a function of the distance between the target loudspeaker and the loudspeaker presenting the reported digit. Responses that did not correspond to any of the presented digits are shown at the far right (rand). Responses are pooled across all subjects and all delays for F (squares and solid lines), SS (circles and dotted lines), and SL (triangles and dashed lines). (*Lower*) Incorrect responses in the F condition as a function of distance between the target loudspeaker and the loudspeaker presenting the reported digit for each temporal position within the sequence (light to dark gray showing results for target digits 1–4). Responses are pooled across all subjects and all delays.

answers when they were unsure of the target digit, this kind of error would be the most common). In the F condition, the improvement in performance across time came about primarily from a decrease in responses to digits presented from masker loudspeakers (Fig. 4 *Lower*).

Discussion

When identifying speech in the presence of competitors, attention to features such as voice and location can guide selective attention (13–18). The current results demonstrate that continuity of these perceptual features, which help define an object's identity, lead to improvements over time in the ability to select a target sequence from a complex acoustic scene. We suggest that this improvement in selective attention occurs because attention operates on perceptual objects, and the identity of an acoustic object in a complex scene depends on evidence acquired over the course of several seconds. Of course, feature-based attention could also account for the basic pattern of our results, but only if listeners can direct attention to multiple features simultaneously.

Slowing the presentation rate of a sequence of target digits reduces some of the cost associated with switching, consistent with there being a finite time required to disengage and then reengage attention (19, 20). However, delays as long as 1 s did not eliminate the cost of switching attention, suggesting that this cost was not entirely due to the time required to redirect attention. Moreover, performance improved over time for a target with continuity of perceptual features; disrupting object continuity reset this across-time refinement. Spatial continuity can also enhance auditory selective attention over much longer time scales (21). These results suggest that listeners refine selective auditory attention over time in a complex acoustic mixture.

The pattern of errors observed in these experiments shows that listeners were particularly susceptible to reporting masker words that occurred simultaneously from locations adjacent to the target. This pattern of errors is consistent with a popular model of spatial attention in which attention is directed via a tuned filter having a spatial focus and some finite spatial extent (e.g., see refs. 22 and 23). For the task and conditions tested here, it appears that the spatial attentional filter is sufficiently broad that adjacent locations are imperfectly rejected. However, we also find that the spatial filter becomes more focused over time when the target location is fixed from digit to digit (see also ref. 24).

Comparison of results from Exps. 1 and 2 suggests that continuity of voice enhances the benefit of spatial continuity of the target sequence (i.e., the cost of switching is greater in Exp. 2 than in Exp. 1) (Fig. 2, compare *Upper* with *Lower*). This enhancement is greatest when interdigit delays are brief and the target digit sequence is relatively connected (continuous) across time. As noted above, feature-based attention could help explain these results; however, it is difficult to see how feature-based attention could account for this effect of stimulus timing. We find that any manipulation that enhances object formation causes an improvement in selective attention over time, whether it is continuity of a stimulus feature (spatial location, voice

quality) or a rapid presentation rate. Thus, parsimony favors the hypothesis that selective attention becomes increasingly more effective as object formation builds.

When the target sequence has spatial continuity and maximal voice continuity (Fig. 3 *Lower*, leftmost plot), performance for the first digit in the sequence is better than when spatial location changes between digits. This kind of effect can only be explained if the overall difficulty of a trial impacts how well the first digit of the target sequence is recalled at the conclusion of the trial, because the subject has no advance knowledge about the target location or target voice for the first digit in either the F or SS conditions. This result suggests that attentional demands are smallest when the target sequence is temporally connected, continuous in voice quality, and from a fixed location, leaving more resources for storage and recall of the sequence. This effect undoubtedly depends on overall memory demands of the task, and thus is likely to vary with the length of the target sequence as well as the listener's knowledge about when the sequence will end.

These findings shed light on why, in listening environments such as noisy parties or restaurants, it is more difficult to follow a conversation involving many people (where the relevant talker often and unexpectedly changes locations) than to focus on one talker (at one location) exclusively. In addition, these results may have implications for visual attention in tasks where object formation and target segmentation is challenging, or where the identity of a visual object depends on continuity of visual features over time (25).

Materials and Methods

Subjects. Five subjects (2 male, 3 female, aged 23–39 years) participated in Exp. 1. Five subjects (2 male, 3 female, aged 24–30 years) participated in Exp. 2, two of whom had participated in Exp. 1 before commencing Exp. 2 (S1 and S2). Subjects S1 and S2 were also two of the experimenters and had previously participated in several similar experiments. The other subjects were paid for their participation. All subjects were screened to ensure that they had normal hearing (within 10 dB) for frequencies between 250 Hz and 8 kHz. Experiments were approved by the Boston University Charles River Campus Institutional Review Board.

Environment. The experiments took place in a single-walled Industrial Acoustics Company booth with interior dimensions of 12′4″ × 13′ × 7′6″ (length × width × height), with perforated metal panels on the ceiling and walls and a carpeted floor (for an acoustic analysis of this environment, see ref. 26). The subject was seated on a chair in the center of the room. A head rest attached to the back of the chair cradled the neck and the back of the head to minimize head movements. No instructions were given to subjects regarding eye fixation during stimulus delivery, and eye movements were not measured. Stimuli were presented via five loudspeakers (215PS; Acoustic Research) located on a horizontal arc ≈5 ft from the subject at the level of the ears. The loudspeakers were positioned within the visual field of the subject, at lateral angles of −30°, −15°, 0°, 15°, and 30°. Subjects indicated their response by using a handheld keypad with an LCD display (QTERM). The booth was kept dark during the experiment, except for a small lamp placed on the floor behind the subject, which helped him or her to see the keypad.

Digital stimuli were generated and selected via a PC located outside the booth, and fed through five separate channels of Tucker-Davis Technologies hardware. Signals were converted at 20 kHz by a 16-bit D/A converter (DA8), attenuated (PA4), and passed through power amplifiers (Tascam) before presentation to the loudspeakers. Each loudspeaker had an LED affixed on its top surface, which could be turned on and off via the PC with a custom-built switchboard. MATLAB (Mathworks) software was used for stimulus generation, stimulus presentation, data acquisition, and analysis.

Stimuli. Stimuli consisted of the digits 1–9 spoken by 15 different male talkers from the TIDIGIT database (27). The mean duration of the set of digits was 434 ms (±103 ms). For each trial, five different sequences of four digits were presented simultaneously from the five spatially separated loudspeakers. For each of the four temporal positions in the sequence, the five digits were chosen randomly with the limitation that they were all different and spoken by a different talker. Digits were presented with synchronous onsets and were zero-padded at the end so that within each temporal position; all were the length of the longest digit in that particular position.

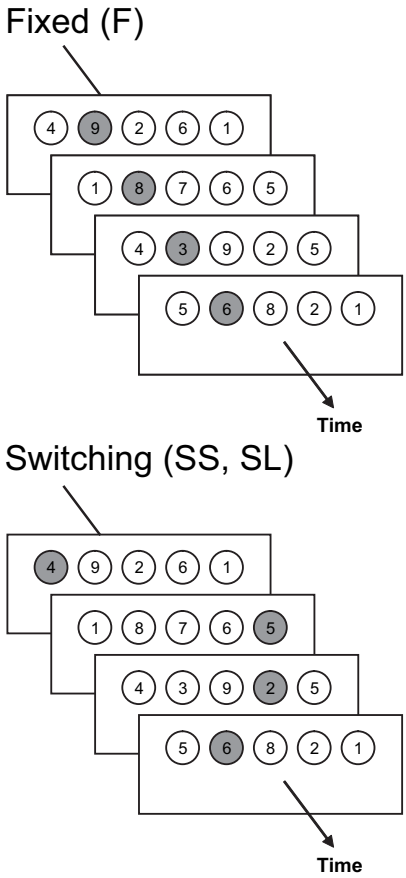

Fig. 5. Schematic of the auditory and visual stimuli for the fixed and switching conditions. Five different digits were presented simultaneously from the five loudspeakers (circles) in each of four temporal positions of the stimulus. During each of the four temporal positions, the LED on one loudspeaker was illuminated (filled circle) to indicate the target digit. (*Upper*) In the fixed condition, the target digit came from the same loudspeaker in each of the temporal positions. (*Lower*) In the switching conditions, the target came from a different random loudspeaker in each temporal position. The visual cue from the target LED came on simultaneously with the auditory stimuli in the F and SS conditions but preceded the auditory stimuli in the SL condition (diagram not shown).

One digit in each temporal position was designated as the target, with the four targets in the different temporal positions making up the target sequence. In each temporal position, one of the five LEDs was illuminated to indicate which loudspeaker contained the target. In the fixed condition (Fig. 5 *Upper*), this was the same loudspeaker for the whole sequence (although the loudspeaker varied randomly from trial to trial). In the two switching conditions (Fig. 5 *Lower*), the target loudspeaker was different in each temporal position so that the four digits in the sequence came from four different loudspeakers.

Conditions. In different experimental blocks, the sequences in a trial were presented with a different delay between the four digits (0, 250, 500, or 1,000 ms). This gave rise to average presentation rates of 2.3, 1.5, 1.1, and 0.7 words per second, respectively (although the variable digit lengths meant that the rhythm was not perfectly regular).

In the F and SS conditions, the LED turned on and off synchronously with the onset and offset of the digits in each temporal position. In the SL condition, the LED came on before the digits in each temporal position, with a lead time equal to the interdigit delay.

In Exp. 1, the voices were chosen randomly for each temporal position with the constraint that the same voice was not presented simultaneously from more than one loudspeaker. As a result, the target voice varied randomly throughout a target sequence. In Exp. 2, the four target digits in a sequence

were spoken by the same voice (chosen randomly on each trial). The maskers were chosen from the remaining 14 voices (separately for each temporal position).

Procedures. In an experimental test, the subject's task was to follow the LEDs and report the four-digit target sequence. Responses were entered by using the handheld keypad after the entire stimulus was finished. Subjects were forced to respond with a four-digit sequence and were instructed to guess the content for any digit that they did not hear. The sequence was scored on a per-digit basis in all analyses.

Each subject completed five sessions in an experiment, each on a separate day. A session consisted of one block of trials per combination of condition (F, SS, and SL) and delay (0, 250, 500, and 1,000 ms). Because the SS and SL conditions were identical for the 0-ms delay, there were 11 blocks of trials in total. The order of the blocks was random and different between sessions and subjects. A message on the keypad at the beginning of each block indicated which condition and delay would be presented in that block. Each block consisted of 40 trials.

Subjects did not complete any formal practice blocks, but were given exemplars of the stimuli to listen to while the experiment was being explained.

Statistical Analyses. The percentage correct data were arcsin transformed and submitted to two repeated measures ANOVAs. The first examined the effect of switching and had factors of condition (F and SS), interdigit delay (0, 250, 500, and 1,000 ms), and temporal position (1–4). The second compared performance in the two switching conditions using only the data that were independent (i.e., excluding the 0-ms delay).

ACKNOWLEDGMENTS. We thank Gerald Kidd and Chris Mason for many helpful discussions about both the experimental design and the data. This work was supported by Office of Naval Research Grant N000140710061 and National Institute on Deafness and Other Communication Disorders Grant R01 DC04663 (to B.G.S.-C.). V.B. was supported in part by a University of Sydney Postdoctoral Research Fellowship. N.K. was supported in part by Slovak Science Agency (VEGA) Grant 1/3134/06.

1. Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18:193–222.
2. Yantis S (2005) How visual salience wins the battle for awareness. *Nat Neurosci* 8:975–977.
3. Kelley T, Serences J, Giesbrecht B, Yantis S (2008) Cortical mechanisms for shifting and holding visuospatial attention. *Cereb Cortex* 18:114–125.
4. Jefferies LN, Ghorashi S, Kawahara J, Lollo VD (2007) Ignorance is bliss: The role of observer expectation in dynamic spatial tuning of the attentional focus. *Percept Psychophys* 69:1162–1174.
5. Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
6. Darwin CJ, Carlyon RP (1995) in *Hearing: The Handbook of Perception and Cognition*, ed Moore BCJ (Academic, London), Vol 6, pp 387–424.
7. Shinn-Cunningham BG (2008) Object-based auditory and visual attention. *Trends Cogn Sci* 12:182–186.
8. Cusack R, Deeks J, Aikman G, Carlyon RP (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.
9. Duncan J (1984) Selective attention and the organization of visual information. *J Exp Psychol Gen* 113:501–517.
10. Roelfsema PR, Lamme VAF, Spekreijse H (1998) Object-based attention in the primary visual cortex of the macaque monkey. *Nature* 395:376–381.
11. Serences JT, Liu T, Yantis S (2005) in *Neurobiology of Attention*, eds Itti L, Rees G, Tsotsos J (Academic, New York), pp 35–41.
12. VanRullen R, Carlson T, Cavanagh P (2007) The blinking spotlight of attention. *Proc Natl Acad Sci USA* 104:19204–19209.
13. Freyman RL, Helfer KS, McCall DD, Clifton RK (1999) The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am* 106:3578–3588.
14. Shinn-Cunningham BG, Ihlefeld A, Satyavarta, Larson E (2005) Bottom-up and top-down influences on spatial unmasking. *Acust Acta Acust* 91:967–979.
15. Best V, Ozmeral E, Shinn-Cunningham BG (2007) Visually-guided attention enhances target identification in a complex auditory scene. *J Assoc Res Otolaryngol* 8:294–304.
16. Kidd G, Jr, Arbogast TL, Mason CR, Gallun FJ (2005) The advantage of knowing where to listen. *J Acoust Soc Am* 118:3804–3815.
17. Darwin CJ, Hukin RW (2000) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J Acoust Soc Am* 107:970–977.
18. Brungart DS, Simpson BD, Ericson MA, Scott KR (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J Acoust Soc Am* 110:2527–2538.
19. Broadbent DE (1958) *Perception and Communication* (Pergamon, London).
20. Treisman AM (1971) Shifting attention between the ears. *Q J Exp Psychol* 23:157–167.
21. Brungart DS, Simpson BD (2007) Cocktail party listening in a dynamic multitalker environment. *Percept Psychophys* 69:79–91.
22. Arbogast TL, Kidd G, Jr (2000) Evidence for spatial tuning in informational masking using the probe-signal method. *J Acoust Soc Am* 108:1803–1810.
23. Mondor TA, Zatorre RJ (1995) Shifting and focusing auditory spatial attention. *J Exp Psychol Hum Percept Perform* 21:387–409.
24. Teder-Sälejärvi WA, Hillyard SA (1998) The gradient of spatial auditory attention in free field: An event-related potential study. *Percept Psychophys* 60:1228–1242.
25. Blaser E, Pylyshyn ZW, Holcombe A (2000) Tracking an object through feature-space. *Nature* 408:196–199.
26. Kidd G, Jr, Mason CR, Brughera A, Hartmann WM (2005) The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acust Acta Acust* 114:526–536.
27. Leonard RG (1984) A database for speaker-independent digit recognition. *Proceedings of the 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '84)* (IEEE, Piscataway, NJ), Vol 9, pp 328–331.