Exploring the benefit of auditory spatial continuity

Virginia Best[a] **and Barbara G. Shinn-Cunningham**
Hearing Research Center, Boston University, Boston, Massachusetts 02215
ginbest@bu.edu, shinn@cns.bu.edu

Erol J. Ozmeral
School of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599
eozmeral@unc.edu

Norbert Kopčo
Department of Cybernetics and AI,Technicka Univerzita Kosice04001, Slovakia
kopco@bu.edu

Abstract: Continuity of spatial location was recently shown to improve the ability to identify and recall a sequence of target digits presented in a mixture of confusable maskers [Best *et al.* (2008). Proc. Natl. Acad. Sci. U.S.A. **105**, 13174–13178]. Three follow-up experiments were conducted to explore the basis of this improvement. The results suggest that the benefits of spatial continuity cannot be attributed to (a) the ability to plan where to direct attention in advance; (b) freedom from having to redirect attention across large distances; or (c) the challenge of filtering out signals that are confusable with the target.

PACS numbers: 43.66.Qp, 43.66.Dc, 43.66.Mk, 43.71.Es [QJF]
Date Received: March 22, 2010 **Date Accepted:** April 19, 2010

1. Introduction

The ability of listeners to recall a sequence of spoken digits presented in a mixture of multiple simultaneous interferers was recently shown to be enhanced if the target sequence was spatially continuous (Best *et al.*, 2008). For four-digit target sequences presented amidst four other spatially separated digit sequences, overall performance was better when the target was presented from a fixed spatial location (the "fixed" condition) than when the location varied unpredictably from digit to digit (the "switching" condition). Performance in the switching condition was particularly poor for fast speech rates. This finding suggests that there is a cost of switching spatial attention that can be partially alleviated if there is ample time to disengage and re-engage attention. However, there was a performance cost even at the slowest rate tested. When examined as a function of digit position within the sequence, performance was observed to improve over time in the fixed condition, but not in the switching condition. Moreover, error analysis showed that listeners often reported digits from locations near to the target location, suggesting that an imperfectly focused spatial filter determined what digit(s) listeners attended (Teder-Sälejärvi and Hillyard, 1998; Marrone *et al.*, 2008; Allen *et al.*, 2009). The spatial filter inferred from these errors became narrower over time in the fixed condition, explaining the superior performance in this condition relative to the switching condition. In other words, it appears that spatial continuity enabled a "refinement" of selective listening over time.

Here we present three follow-up experiments that explored different aspects of the benefit of spatial continuity. First, we explored whether the benefit of spatial continuity was a result of being able to predict where to focus spatial attention. Specifically, in the original study,

[a] Author to whom correspondence should be addressed.

the target locations were predictable in the fixed condition and unpredictable in the switching condition. Thus, the benefit from digit to digit may be associated with spatial predictability. To test this idea, we examined performance for the switching condition when the target spatial trajectory was fixed over a whole block and listeners were cued with a visual sequence to familiarize them with the trajectory prior to the block. Second, we wondered whether the cost of switching spatial attention would still occur if the spatial trajectories were smoother and contained no large jumps in location (more like motion of a sound source in the real world). To test this idea, we examined performance for a switching condition in which the target always moved to an adjacent location in the array, and in which the direction of motion either did not change over the course of a trial or changed only once within a trial. Finally, we tested whether the importance of having the target come from a fixed location would remain if the interferers were not potential targets, so that selection of the target from the mixture did not necessarily require spatially directed attention.

2. Methods

2.1 Participants

Five listeners with normal hearing participated in the previous study (Best *et al.*, 2008). Six listeners were recruited for Experiment 1 (including one listener from the previous study), five for Experiment 2 (including one listener from the previous study), and four for Experiment 3 (including three listeners from the previous study).

2.2 Environment

The environment was identical to that used in the previous experiment (Best *et al.*, 2008). Specifically, listeners were seated in the center of a darkened single-walled IAC booth with interior dimensions of 12 ft, 4 in. × 13 ft × 7 ft, 6 in. (length, width, height). The listener's head was supported and kept relatively still by a head rest; a handheld keypad (QTERM) was provided for collecting responses. Five loudspeakers (Acoustic Research 215PS) were positioned on an arc in front of the listener at a distance of 5 ft, evenly spaced from −30° to +30° azimuth. Digital stimuli were generated by a computer located outside the booth and fed to the loudspeakers 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. A custom-built switchboard controlled five light-emitting diodes (LEDs) that were fixed to the top of the five loudspeakers. These LEDs provided visual information about the target digit locations. MATLAB (Mathworks) software was used for stimulus generation, stimulus presentation, data acquisition, and analysis.

2.3 Stimuli and task

The stimuli were very similar to those used in the previous study (Best *et al.*, 2008). Stimuli were created from the digits 1–9 spoken by 15 male talkers from the TIDIGITS database (Leonard, 1984). The mean duration of the set of digits was 434 ms (±103 ms). For each trial, five sequences of digits were presented simultaneously from the five loudspeakers. These sequences were comprised of four random digits. While the same digit could be repeated within a sequence, the same digit was never presented at the same time in different loudspeakers. In all of the current experiments the digits making up the target sequence were spoken by a single voice, chosen randomly on each trial, as in the "fixed voice" condition of the previous study. The maskers were chosen from the remaining 14 voices. In each temporal position, the voices chosen for the four maskers were distinct from one another and from the target voice. The onsets of the five digits in each temporal position were time-aligned. The length of each temporal "slot" was set to the length of the longest digit in that temporal position (zero-padding the end of the other digits in that position). In Experiment 3 only, all non-target digits were reversed in time (prior to zero-padding) to make them unintelligible. To manipulate the overall rate of presentation, silent gaps (with durations of 0, 250, 500, or 1000 ms) were inserted between consecutive digits. A single digit in each temporal position was designated as the target, with its location

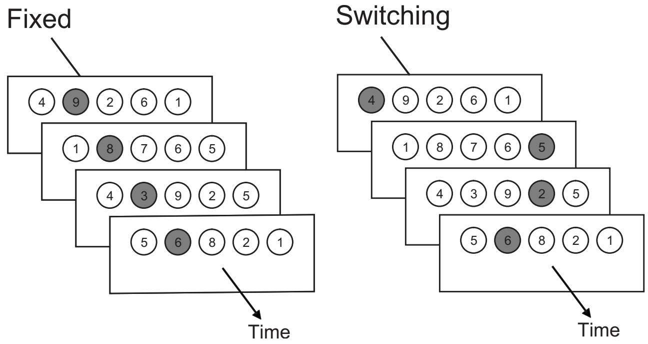

Fig. 1. Schematic of the fixed and switching stimulus conditions. Each slice represents a presentation interval in which five simultaneous digits were presented (circles). The shading depicts the location and timing of the target.

indicated by the LED on the target loudspeaker. The target location could be fixed or could vary from digit to digit depending on the condition. In all experiments, the listener's task was to report the four-digit target sequence in order. If the listener missed one or more digits, he or she was instructed to guess. Responses were entered using the hand-held keypad after the entire stimulus was finished.

3. Experiment 0: Previous study

3.1 Conditions

In the previous study (Best *et al.*, 2008), three conditions were tested. In the "fixed" (F) condition the target sequence was presented from a single loudspeaker (chosen randomly on each trial) and the LED turned on and off synchronously with the onset and offset of the auditory stimulus in each temporal position (Fig. 1, left panel). In the other two conditions, the target loudspeaker varied unpredictably from digit to digit (Fig. 1, right panel). In the "switching, LED synchronous" (SS) condition, the LED was illuminated synchronously with the digits, just as in the F condition. In the "switching, LED leading" (SL) condition, the LED came on *before* the auditory stimulus in each temporal position, with a lead time equal to the fixed delay between digits. One of four silent delays (0, 250, 500 or 1000 ms) was inserted between digits to create four presentation rates (resulting in average presentation rates of 2.3, 1.5, 1.1, and 0.7 words/s, respectively). Each subject completed five 40-trial blocks of each condition/delay combination (60 blocks in total).

3.2 Results

Figure 2 shows performance as a function of inter-digit delay, with each experiment in a different panel. The data for Experiment 0 [Fig. 2(a)] are replotted from Best *et al.*, 2008, where they appeared as Experiment 2. In that experiment, mean performance was better in the F condition than in the SS conditions for all inter-digit delays. Presentation rate had little effect on performance in F; however, performance in SS was better at slower rates than at faster rates. Confirming these observations, a two-way repeated-measures analysis of variance (ANOVA) conducted on the arcsine-transformed scores revealed significant main effects of condition $[F(1,4)=55.8,\ p<0.005]$ and delay $[F(3,12)=22.4,\ p<0.001]$, as well as a significant interaction $[F(3,12)=38.0,\ p<0.001]$. Providing spatial information in advance during the inter-digit delays improved performance slightly [compare SL to SS in Fig. 2(a)], but even at the longest delays did not restore performance to that achieved in F. A two-way repeated-measures ANOVA conducted on the arcsine-transformed scores for conditions SS and SL at the nonzero delays revealed significant main effects of condition $[F(1,4)=42.5,\ p<0.005]$ and delay $[F(2,8)=16.7,\ p<0.005]$, and no significant interaction.

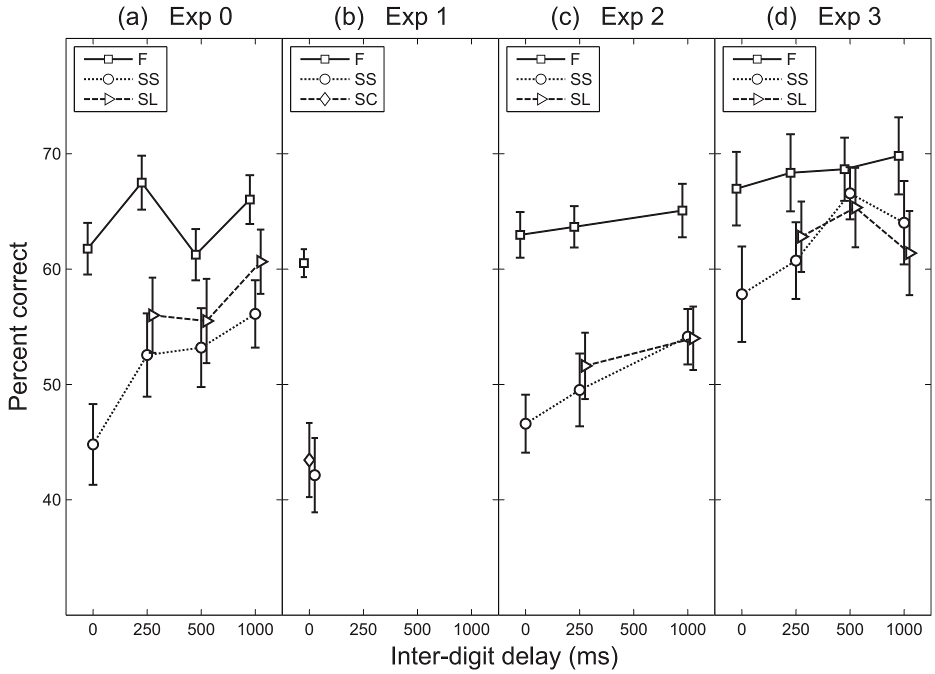

Fig. 2. Across-subject mean scores (±SEM) plotted as a function of inter-digit delay. (a) Experiment 0: conditions F (fixed; squares and solid lines), SS (switching, LED synchronous; circles and dotted lines), and SL (switching, LED leading; triangles and dashed lines). (b) Experiment 1: conditions F (fixed; squares and solid lines), SS (switching, LED synchronous; circles and dotted lines), and SC (switching, cued trajectory; diamonds and dashed lines). Experiment 2: conditions as per Experiment 0 except SS and SL contained only smooth trajectories. (d) Experiment 3: conditions as per Experiment 0 except maskers were time-reversed.

Figure 3 shows performance in the different experiments broken down by digit position within the sequence for the most rapid presentation rate (the 0-ms delay, which was common to all experiments and which lead to the greatest cost of switching attention of all of the tested delays). Again, the data for Experiment 0 [Fig. 3(a)] are replotted from Best *et al.* (2008). In that experiment, for 0-ms delay, performance in F tended to improve for each subsequent digit in the sequence. This improvement over time was not evident in SS, where performance was relatively constant as a function of temporal position. Note that SL is not shown because it could not be tested for the 0-ms delay.

4. Experiment 1: Cued trajectory

4.1 Conditions

In Experiment 1, a new switching condition was introduced in which the spatial trajectory of the target sequence was fixed across a block of trials, and listeners were familiarized with the trajectory prior to the start of the block. Specifically, each 40-trial block of this "switching, cued-pattern" (SC) condition was broken down into 4 sub-blocks of 10 trials, such that one pattern was tested in each of the sub-blocks (i.e., four different patterns were tested in one full block). The fixed pattern tested in a given sub-block was chosen randomly without replacement, separately for each listener, from 16 randomly determined patterns. The pattern was presented visually via the LEDs to the listener a minimum of three times at the start of the sub-block (listeners were then allowed to repeat the cue as many times as desired, prior to each sub-block). Blocks of the SC condition were interleaved with blocks of SS and F that were identical to these conditions in Experiment 0. Due to time constraints, only the fastest presentation rate was tested (0–ms delay). However, this was the rate with the largest effect of switching in the previous

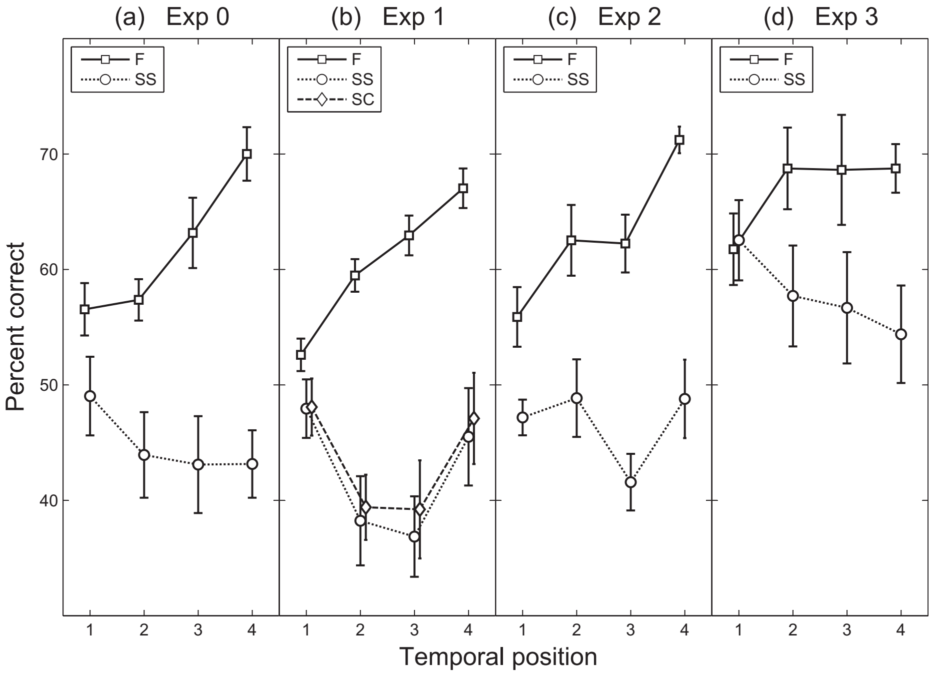

Fig. 3. Across-subject mean scores (±SEM) plotted as a function of temporal position for the 0-ms inter-digit delay condition. Other details as per Fig. 2.

study, so if familiarization with the trajectory affected performance, it was likely to have the largest effect for this delay. Each subject completed four 40-trial blocks of each condition (12 blocks in total).

4.2 Results

Despite the use of different subjects and a different arrangement of the experimental blocks, mean results for F and SS in Experiment 1 replicate results from Experiment 0 [compare Fig. 2(b) to Fig. 2(a) for the 0-ms delay]. Moreover, performance for familiar patterns was no better than for random patterns [compare SC and SS in Fig. 2(b)]. A one-way repeated-measures ANOVA on the arcsine-transformed scores revealed a significant main effect of condition $[F(2,8)=51.0,\ p<0.001]$. *Post hoc* comparisons $(p<0.05)$ confirmed that performance in the F condition was superior to that in the SS and SC conditions (which were not significantly different from each other). When examined as a function of digit position [Fig. 3(b)], performance in the F condition improved over the course of the sequence, whereas scores in SS and SC were roughly U-shaped and essentially identical to each other.

These results suggest that *a priori* knowledge of the trajectory of the target sequence, reinforced by the fixing of this trajectory for a block of trials, did not reduce the costs associated with having to redirect spatial attention from one digit to the next, at least at this rapid presentation rate. This is consistent with results in Experiment 0 at slower rates, where a leading visual cue (condition SL) did not restore performance to that seen in the F condition. Overall, it appears that spatial *continuity* per se (and not just spatial *certainty*) enhances performance on this task.

5. Experiment 2: Smooth trajectory

5.1 Conditions

In Experiment 2, conditions were identical to Experiment 0 except that the spatial trajectories in SS and SL were "smooth," only containing transitions of 15° (a shift of one loudspeaker position) and changing direction as little as possible. This was achieved by selecting a random position for the first target digit and an initial direction (left or right). The second digit then came from the next loudspeaker over in the selected direction, and so on. If there was no loudspeaker in the selected direction (at the edge of the array) then the trajectory direction reversed. Thus, for loudspeakers 1 to 5, example trajectories were 1–2–3–4, 2–1–2–3, 3–4–5–4, etc.) Due to time constraints, the intermediate, 500–ms delay condition was excluded. Each listener completed three separate 40-trial blocks of each condition/delay combination (27 blocks in total), randomly ordered.

5.2 Results

Results for condition F were reasonably similar to results for the identical condition in Experiment 0 [compare Fig. 2(c) to Fig. 2(a), all delays except 500 ms]. Despite the use of smooth trajectories, performance in condition SS was still inferior to that in condition F. A two-way repeated-measures ANOVA conducted on the arcsine-transformed scores for these two conditions revealed significant main effects of condition $[F(1,4)=154.8,\ p<0.001]$ and delay $[F(2,8)=9.3,\ p<0.01]$, as well as a significant interaction $[F(2,8)=8.3,\ p<0.05]$. There was a weak advantage of the leading visual cue in Experiment 2 [compare SS and SL in Fig. 2(c)], which was statistically significant $[F(1,4)=10.6,\ p<0.05]$. The effect of delay was not significant, nor was the interaction between condition and delay. Figure 3(c) shows performance for the 0-ms delay condition as a function of temporal position. Similar to results from Experiments 0 and 1, there is an improvement over time for the F condition but not for the SS condition.

The advantage of F over SS suggests that the benefit of spatial continuity is relatively specific, so that even small variations in location disrupt a listener's ability to selectively attend to an auditory target sequence. The difference between SS and SL was reduced compared to Experiment 0, perhaps because the uncertainty about where the target would occur next was low, making the visual cues somewhat redundant. Alternatively, it may be that the small benefit of the leading visual cue seen in Experiment 0 is restricted to cases in which the target location jumped over a large angle (e.g., when it moved from the left to the right hemi-field).

6. Experiment 3: Reversed maskers

6.1 Conditions

The spatial conditions of Experiment 3 were identical to those tested in Experiment 0 (F, SS, SL). However, the stimuli differed in that the masker digits were time-reversed to render them unintelligible. Because all of the maskers were unintelligible, it is possible that listeners could simply listen for the intelligible target without deploying spatial attention. Such a strategy might render spatial continuity less influential on performance. Because the task was easier overall when the maskers were unintelligible, the target was attenuated by 10 dB to reduce ceiling effects. Each subject completed five 40-trial blocks of each condition/delay combination (60 blocks in total).

6.2 Results

Results for Experiment 3 were broadly similar to those for Experiment 0. Performance in the F condition was superior to performance for SS [Fig. 2(d)]. A repeated-measures ANOVA on the arcsine-transformed scores revealed significant main effects of condition $[F(1,3)=345.0,\ p<0.001]$ and delay $[F(3,9)=10.6,\ p<0.005]$, and a significant interaction $[F(3,9)=5.9,\ p<0.05]$. There was no advantage of the leading visual cue (compare SS and SL), perhaps because the target digits had a slight tendency to "pop out" from the background of time-reversed maskers (Asemi *et al.*, 2003; Best *et al.*, 2007). Figure 3(d) shows performance for the 0-ms delay condition as a function of digit position. While performance in F plateaus rather than increasing steadily as it did in the other experiments (perhaps due to a ceiling effect), the fact that performance in SS declines over time means that the *advantage* of a fixed target location increases as a function of digit position in the sequence.

The results of Experiment 3 suggest that the benefit of spatial continuity is not restricted to the difficult and somewhat unrealistic case in which the interfering sounds are all potential targets. This finding is intriguing in light of the analysis by Best *et al.* (2008) that showed that the majority of errors in that study corresponded to the report of spatially-adjacent masker digits and that the improvement in performance over time in the F condition was accompanied by a reduction in these confusions. It appears that the task of ignoring speech-like distracters was difficult enough that spatial continuity was still critical, even though listeners might have been able to perform the task by simply listening for intelligible target words without deploying spatial attention. Because the target digits were 10 dB less intense than in the original study, it may be that the difficulty came more from limited audibility than from the problem of selecting the correct utterance (Brungart, 2001; Kidd *et al.*, 2008). Importantly, even if poor audibility was the factor limiting performance in Experiment 3 and even if listeners adopted a strategy that depended less on directing spatial attention than in Experiments 0–2, spatial continuity improved performance from digit to digit in a manner much like it improved spatial selection in the other experiments.

7. Conclusions

Improvements in selectivity of spatial attention that arise for a target at a fixed spatial location cannot be attributed exclusively to (a) the ability to plan where to direct attention in advance; (b) freedom from having to redirect attention across large separations in location; or (c) the challenge of filtering out nearby signals that are confusable with the target. Instead, the ability to selectively attend to an acoustic target sequence improves when the syllables making up the stream arise from a single fixed spatial location. Future work is needed to determine whether similar improvements in selective attention arise when non-spatial features are continuous, or whether this effect is specific to spatially directed attention.

Acknowledgments

Work supported by ONR and NIDCD. Brenden Hurd and Chloe McGuffin assisted with data collection.

References and links

Allen, K., Alais, D., and Carlile, S. (**2009**). “Speech intelligibility reduces over distance from an attended location: Evidence for an auditory spatial gradient of attention,” Percept. Psychophys. **71**, 164–173.

Asemi, N., Sugita, Y., and Suzuki, Y. (**2003**). “Auditory search asymmetry between normal Japanese speech sounds and time-reversed speech sounds distributed on the frontal-horizontal plane,” Acoust. Sci. & Tech. **24**, 145–147.

Best, V., Ozmeral, E. J., and Shinn-Cunningham, B. G. (**2007**). “Visually-guided attention enhances target identification in a complex auditory scene,” J. Assoc. Res. Otolaryngol. **8**, 294–304.

Best, V., Ozmeral, E. J., Kopčo, N., and Shinn-Cunningham, B. G. (**2008**). “Object continuity enhances selective auditory attention,” Proc. Natl. Acad. Sci. U.S.A. **105**, 13174–13178.

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

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008**). “Informational masking,” in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay, (Springer, New York), pp. 143–190.

Leonard, R. G. (**1984**). “A database for speaker independent digit recognition,” in Proceedings of the ICASSP’84, San Diego, California, Vol. **9**, pp. 328–331.

Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2008**). “Tuning in the spatial dimension: Evidence from a masked speech identification task,” J. Acoust. Soc. Am. **124**, 1146–1158.

Teder-Sälejärvi, W. A. and Hillyard, S. A. (**1998**). “The gradient of spatial auditory attention in free field: An event-related potential study,” Percept. Psychophys. **60**, 1228–1242.