Influence of preparation time and pitch separation in switching of auditory attention between streams

Eric Larson and Adrian K. C. Lee[a)]
Institute for Learning and Brain Sciences, University of Washington, Seattle, Washington 98195
larsoner@uw.edu, akclee@uw.edu

Abstract: The ability to consciously switch attention between speakers of interest is necessary for communication in many environments, especially when multiple talkers speak simultaneously. Segregating sounds of interest from the background, which is necessary for selective attention, depends on stimulus acoustics such as differences in spectrotemporal properties of the target and masker. However, the relationship between top-down attention control and bottom-up stimulus segregation is not well understood. Here, two experiments were conducted to examine the time necessary for listeners to switch auditory attention, and how the ability to switch attention relates to the pitch separation cue available for bottom-up stream segregation.

PACS numbers: 43.66.Lj [QJF]
Date Received: April 12, 2013 **Date Accepted:** June 14, 2013

1. Introduction

The ability to understand a target sound in the presence of other, competing masking sounds has long been the focus of study in auditory research (Cherry, 1953). To successfully attend to a target stimulus in the presence of maskers, listeners make use of cues such as spatial location and pitch to segregate spectrotemporal elements into separate auditory streams (Darwin, 1997) and to direct their attention to select a stream of interest. For hearing-impaired listeners, peripheral impairments make it difficult to selectively attend to a target stream when other maskers are present, interfering with communication in noisy social environments (Shinn-Cunningham and Best, 2008). It is important, then, to understand how top-down selective attention processes relate to the cues that are available to facilitate segregation; this is not yet well understood, even in normal hearing listeners.

Although behavioral tasks have been used to study top-down, goal-driven attention orientation in vision (Kiesel *et al.*, 2010), most studies of attention switching in audition have focused on exogenous, stimulus-driven attention (Shinn-Cunningham, 2008). One compelling way to study top-down direction of attention is to focus on endogenous attention switching, where a listener intentionally changes from listening to one auditory stream to another. One recent study found that the costs of top-down attention switching were similar across audition and vision (Koch *et al.*, 2011), consistent with neuroimaging evidence for the hypothesis that top-down attention is controlled by a supramodal network (Larson and Lee, 2013; Shomstein and Yantis, 2006). However, how these top-down attention processes operate in audition is not entirely clear. For example, the time required to successfully switch selective attention between two simultaneous, competing auditory streams is not well established, although there is some neuroimaging evidence that the neural mechanisms involved become engaged more than 300 ms following a top-down switch of attention (Larson and Lee, 2013).

[a)]Also at: Department of Speech and Hearing Sciences, University of Washington, Seattle, WA.

Most studies of top-down auditory attention control have focused on the deployment or switching of attention based on spatial features (Best *et al.*, 2008; Broadbent, 1958; Mondor and Zatorre, 1995; Treisman, 1971). However, other stimulus features can be critical for facilitating communication in noisy environments. Moreover, it is unclear the extent to which the peripheral representations available for bottom-up stream segregation interact with top-down direction and switching of attention. For example, pitch differences provide a salient cue that aids bottom-up segregation of different auditory streams by allowing listeners to identify formant peaks and group different formants of vowels together (Darwin, 1997). However, recent evidence has shown that task goals can modulate the process of segregating sequential sounds that differ in pitch (Carlyon, 2004), suggesting that top-down attention can affect stream segregation. It has also been hypothesized that, although bottom-up auditory processing is important for source segregation, attention is critical for binding auditory features together to segregate a particular stream of interest from concurrent background sounds (Shamma *et al.*, 2011). It is therefore important to establish relationships between automatic, bottom-up stream segregation and goal-driven, top-down selective attention processes.

In the present study, we examined two related issues involving top-down attention control based on non-spatial features. First, we sought to probe the time-course of top-down attention switching by varying the amount of time subjects were given to switch attention between two simultaneous, spatially co-located auditory streams. Second, we examined how the peripheral separability (here, provided via pitch differences) between two simultaneous, competing auditory streams affects the ability to selectively direct or switch attention. We hypothesized that (i) around 300 ms would be required for listeners to optimally switch between auditory streams based on physiological evidence (Larson and Lee, 2013); (ii) for smaller pitch separations between the target and masker stream, the cost of switching attention (relative to maintaining attention) would be larger; and (iii) this effect would be larger when there was less time allowed for switching attention. Here the cost of switching attention was evaluated using both accuracy and reaction times (RTs).

2. Experiment 1

2.1 Subjects

Nineteen subjects (ten male, aged 19–34) participated in this experiment. All participants had pure-tone thresholds in both ears within 20 dB of normal-hearing (octave frequencies, 250–8000 Hz). All subjects gave informed consent to participate in the study as approved by the University of Washington Institutional Review Board.

2.2 Stimulus design

Auditory stimuli were generated using tokens from a single female talker in the ISOLET v1.3 corpus (Cole *et al.*, 1990) chosen such that all letters were as close as possible to, but no longer than 400 ms in duration (trimmed of leading and trailing silence). Since the target letter in the task was an "E," we eliminated letters that rhymed with "E." Each letter was monotonized and shifted to 200 Hz $\pm$ 4.25 semitones (st.) in fundamental frequency (Praat software, Amsterdam) to test an 8.5 st. pitch separation. Letters were then windowed with a 10 ms $\cos^2$ envelope and matched in intensity. Target and masker streams, each consisting of six letters, were formed by concatenating three letters, inserting a 100, 200, 400, 600, or 800 ms silent gap to allow listeners to switch streams when instructed, and appending the last three letters. To test each condition, we used 54 trials $\times$ 5 gaps $\times$ 2 maintain/switch conditions = 540 trials total. Stimuli were presented in a sound-treated room over insert earphones at a comfortable level (75 dB SPL) against a π-interaural-phase white noise background (20 dB signal-to-noise ratio) to mask any environmental noise. "E" tokens were distributed across target and masker streams, and first and second sets of three letters (i.e., before and after the switch period) such that it could be disambiguated whether or not a subject attended to the correct stream.

2.3 Task

Subjects performed a behavioral experiment in which they were instructed to attend to one of the two simultaneous streams (Fig. 1). Their task was to listen for "E"' tokens in the target stream of spoken letters, and to press a response button as quickly as possible after hearing the second "E." This target-detection task allowed us to assess performance costs using both accuracy (percent correct) and RT.

In each trial, subjects first heard an auditory prototype consisting of two repetitions of the letter "A" processed in the same manner (equivalent pitch) as the initial target stream, followed by a 250 ms white noise burst to disrupt buildup of streaming. Concurrently with the "A"s and noise, subjects received one of two possible visual cues to denote the task type: a diamond cue indicated that they should attend to the stream that sounds like the prototype for all letters; an X cue indicated that they should switch attention between the two streams during the gap period (between the third and fourth target letter). The auditory prototype and visual cue lasted for a total of 1050 ms (400 + 400 + 250 ms) and were followed by a 500 ms pause, as cue-target intervals of at least 500 ms have been shown to allow for adequate task preparation (Meiran *et al.*, 2000). Following this cue period, the target and masker streams began playing. Changing the duration of the gap between the first and last three target letters allowed us to vary the amount of time subjects have to switch attention between streams, thereby potentially varying the difficulty of switching attention. Subjects were instructed to listen for occurrences of the letter "E" in the target stream and press a button (BBox, Tucker Davis Technology, Alachua, FL) as soon as they heard the second occurrence of that letter.

To disambiguate whether a subject was attending to the correct stream, the number of "E"s in the target and masker streams never matched in the first half (first three letters). Moreover, "E"s in opposing streams were always separated by at least one letter (no simultaneous "E"s). No masker "E" could appear within 800 ms following the onset of the second target "E" (the one designed to elicit a response), and similarly no second target "E" could appear within 800 ms of a masker "E" so that a button press in a given trial could be attributed to the correct detection of the second target "E." In some trials, both streams could, thus, have two "E"s. For example, a target stream of "ROY-EUE" could be matched with a masker stream consisting of "EGE-RYL" (Fig. 1), since a

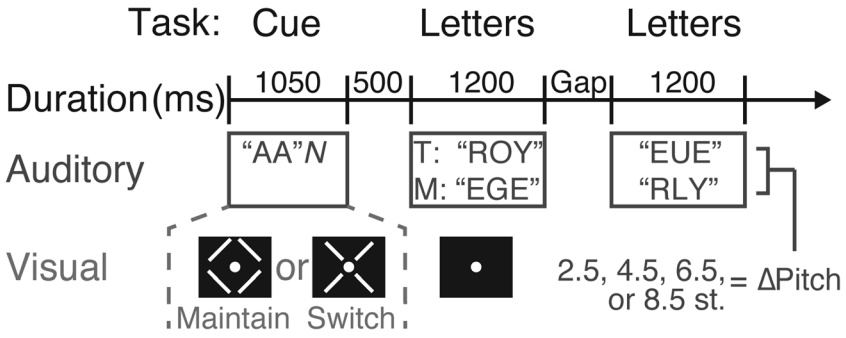

Fig. 1. Psychophysical tasks for Experiments 1 and 2. In each experiment, subjects attended to one of two simultaneous, diotic auditory streams with a pitch separation (8.5 st. in Experiment 1; varied in Experiment 2 as 2.5, 4.5, 6.5, or 8.5 st.). In the cue phase, subjects heard two "A"s processed in the same manner as the target stream, followed by a noise burst to disrupt the buildup of streaming. Simultaneously, a diamond or X visual cue was shown to indicate that listeners should maintain or switch attention, respectively, during the gap period between the first three and last three target and masker letters. After a 500 ms gap, the simultaneous target and masker streams were presented, with the first three letters separated from the last three letters by a variable gap duration (100, 200, 400, 600, or 800 ms in Experiment 1; 200, 400, or 600 ms in Experiment 2) to allow users time to switch attention if necessary. Listeners were instructed to push a response button once the second "E" was heard in the target stream. Thus, in this example, the correct response on a maintain-attention trial would be to press the response button following the sixth letter, and the correct response on a switch-attention trial would be to withhold a button press.

correct button press (after the onset of the sixth letter) could be disambiguated from an incorrect button press (any time beforehand). On 1/9 of trials, at most one target "E" was presented, in which case the correct response was not pressing a button. Performance could then be measured simply as the proportion of trials on which the subject correctly (1) pressed the response button within 800 ms of the second target "E" if it existed, or (2) correctly withheld a button press if there was at most one target "E." On each trial, the stream playback was stopped as soon as the subject pressed the response button to reduce trial duration. After the response or completion of the target and masker streams, there was a 1000 ms break before the next trial.

Prior to psychoacoustical measurements, subjects underwent two training phases where the gap between the first three and last three target letters was 600 ms. In the first training phase, they only performed trials where they maintained attention to one stream (diamond cue), and they needed to score at least 80% on ten trials to advance to the second phase. In the second phase, they only performed trials where they switched attention (X cue) between the two streams after the third target letter, and again needed to score at least 80% on ten trials to advance to the testing phase. In the testing phase, we varied the gap duration pseudorandomly across trials, and interleaved standard and switch trials. There were both standard (diamond) and switch (X) trials.

2.4 Data analysis

In order to examine meaningful RT measurements in the switch-attention case, only trials in which the fourth target-stream letter (immediately following the maintain/switch gap) was the second target "E" were analyzed. The exception to this is analysis of the "control" condition, which is defined by trials in which both of the target "E"s occurred within the first three letters. Note that, in this case, the stream attended by the subject is fixed across both maintain- and switch-attention conditions since the second "E" occurs before the switch period, with the only difference being that subjects were preparing for an upcoming attention switch (or not). Statistical analyses were performed using multi-way repeated measures analysis of variance (repeated measures ANOVA), using a Greenhouse-Geisser correction for non-sphericity when appropriate. *Post hoc* paired *t*-tests were Sidak multiple-comparisons corrected.

2.5 Results

We found that performance (Fig. 2) was poorer on switch- than maintain-attention trials, with lower accuracy and longer RT for switch attention [main effect in repeated measures ANOVA, $F(1,18) = 33.4$, $p < 0.001$ and $F(1,18) = 22.065$, $p < 0.001$]. Critically, there was a main effect of gap duration [accuracy $F(3.01,54.1) = 7.82$, $p < 0.001$ and RT $F(3.17,57.1) = 5.12$, $p = 0.003$], with 400 and 600 ms each significantly different from 100 and 200 ms ($p < 0.040$ all corrected paired *t*-tests). RTs show the same trends, but failed to reach significance ($p > 0.05$ across all comparisons). There did not appear to be an interaction between switch- versus maintain-attention and gap duration in accuracy [$F(3.584,64.51) = 1.264$, $p = 0.293$] or RT [$F(3.249,58.48) = 0.710$, $p = 0.560$]. The average performance on no-response trials (where only one target "E" was present) was 79.5%, not significantly different from the average performance on maintain-attention trials (75.5%; $p = 0.24$).

Analysis of the control condition, where both target "E"s occurred within the first three letters, showed a significant difference between maintain- and switch-attention trials for both accuracy and RT ($p = 0.009$ and $p = 0.011$, respectively).

2.6 Discussion

By varying the time listeners had to switch attention, we found that switch-attention performance plateaued beyond 400 ms. This matches well with our recent neuroimaging evidence that the neural time-course of top-down attention switching in audition is around 300 ms (Larson and Lee, 2013), lending support to the hypothesis that the time course of attention switching is on this time scale. We speculate that the observed

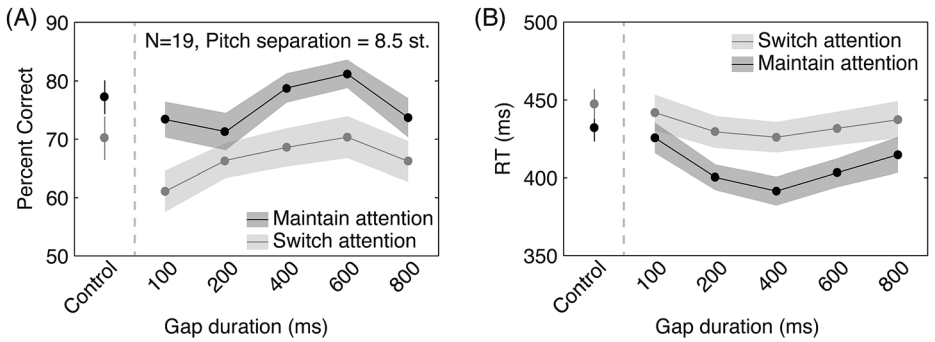

Fig. 2. Behavioral results of Experiment 1. Subject performance on the task, as measured by accuracy (A) and RT (B), is shown for maintain-attention (dark gray) and switch-attention (light gray) conditions as a function of the gap duration allowed for switching attention. Shaded areas show ±1 standard error measurement. The "Control" condition pooled trials where the second target "E" (intended to elicit a button press from the listener) occurred within the first three letters, i.e., before the gap period had occurred. In addition to there being an overall switch cost, where performance in the switch-attention case was poorer than that in the maintain-attention condition (repeated measures ANOVA $p < 0.001$ for both accuracy and RT), performance was significantly different across gap durations ($p < 0.003$, both), with response accuracy being significantly higher in the 400 and 600 ms gap conditions compared to 100 or 200 ms ($p < 0.040$, all, Sidak corrected t-tests). There was also a performance difference in the control condition for both accuracy ($p = 0.009$) and RT ($p = 0.011$).

decrease in performance at 800 ms gap duration is due to a reduction in the streaming of the two objects, since long gap durations could necessitate a re-segregation of the streams following the gap. Given the slight performance improvements in the maintain-attention condition from 100 to 600 ms, it is likely that some other process, such as rhythmic expectation, may also be contributing to changes in performance as a function of gap duration.

Interestingly, we found degradation in performance comparing the switch-attention to the maintain-attention condition even on "control" trials, where targets occurred before the switch period. This suggests that there is a cognitive load involved in remembering and/or preparing to switch attention that is independent of whether the attention switching actually occurs.

3. Experiment 2

3.1 Methods

Fourteen subjects participated (five males, aged 19–30), none of whom participated in Experiment 1. Stimulus generation was the same as in Experiment 1 except that the set of gap durations was reduced to [200, 400, 600 ms], and the pitch separation varied from trial to trial, taking on values of [2.5, 4.5, 6.5, 8.5 st.]. To test each condition, we needed 54 trials × 3 gaps × 4 pitch separations × 2 maintain/switch conditions = 1296 trials total. Because this took approximately two hours, trials were split into two sessions, each with 10 blocks (~64 trials per block). Three subjects were excluded from RT analysis due to low performance scores (and thus, unreliable RT estimates).

3.2 Results

We again found that performance (Fig. 3) was poorer on switch- than maintain-attention trials, with lower accuracy and longer RT for switch, main effect in repeated measures ANOVA [$F(1,13) = 26.5$, $p < 0.001$ and $F(1,10) = 5.795$, $p = 0.037$, respectively]. There was again a main effect of gap duration [accuracy $F(2,26) = 6.812$, $p < 0.005$ and RT $F(2,20) = 5.76$, $p = 0.011$] with 200 ms accuracy significantly worse than 400 or 600 ms ($p < 0.040$ each, corrected paired t-tests) and 600 ms RT significantly shorter than 200 ms ($p = 0.026$). Critically, there was also a significant main effect of pitch separation [accuracy $F(1.395,18.14) = 18\,564$, $p < 0.001$ and RT $F(3,30) = 5.563$, $p = 0.004$], with all pitch separation pairs

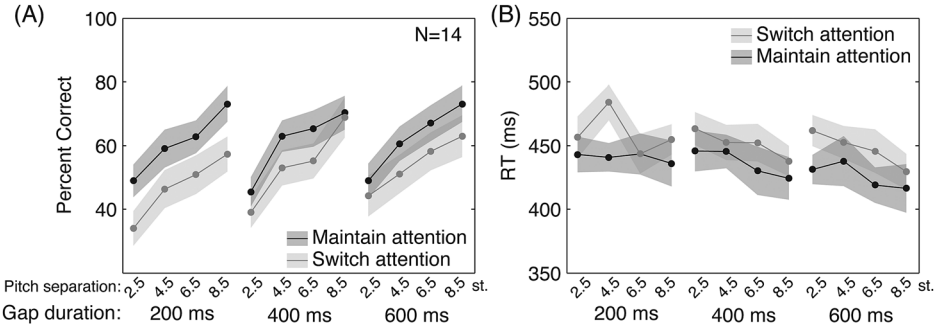

Fig. 3. Behavioral results of Experiment 2. Subject performance in terms of accuracy (A) and RT (B) is shown for maintain-attention (dark gray) and switch-attention (light gray) conditions as a function of the gap duration allowed for switching attention, as well as the pitch separation between the streams. Shaded areas show ± 1 standard error measurement. There was again a switch cost, with maintain-attention higher than switch-attention performance (repeated measures ANOVA $p < 0.001$ and $p = 0.037$ for accuracy and RT, respectively) and a main effect of gap duration ($p = 0.005$ and $p = 0.012$ for accuracy and RT, respectively), with 200 ms accuracy significantly worse than 400 or 600 ms ($p = 0.040$ each, Sidak corrected paired t-tests) and 600 ms RT significantly shorter than 200 ms ($p = 0.027$). There was also a significant main effect of pitch separation for accuracy and RT ($p < 0.001$ and $p = 0.004$, respectively), with all pitch separations significantly different from one another in accuracy ($p < 0.035$ all), and 8.5 semitones significantly different from 2.5 or 4.5 semitones for RT ($p < 0.044$, each).

significantly different from one another in accuracy ($p < 0.035$, all), and 8.5 st. significantly shorter than 2.5 or 4.5 st. for RT ($p < 0.044$, all) with other pairs not significantly different ($p > 0.05$, each). For accuracy, there was a significant interaction between attention condition and gap duration [$F(2,26) = 5.83$, $p = 0.008$], but no other interactions were significant ($p > 0.15$, all) and no significant interactions were found for RT ($p > 0.17$, all).

3.3 Discussion

In Experiment 2, we again observed that the more time subjects were given to switch attention, the better their resulting performance. Also as expected, subject performance decreased on trials where the pitch separation between the target and masker streams was small, suggesting that the strength of physical cue available for bottom-up segregation can impact target selection. Interestingly, we did not observe an interaction between bottom-up segregability and the top-down load, either in terms of the task demands (switch versus maintain attention) or in terms of the time allowed to perform the attention manipulation (gap duration). This suggests that bottom-up stream segregation and top-down stream selection, at least for the stimulus parameters used here, are independent.

4. General discussion

In both experiments, we observed large top-down attention effects on the ability to follow the target stream. These effects manifested in both decreased performance on switch-attention trials, and decreased performance when there was less time given to switch attention between the two streams. This suggests that the process of switching auditory attention between competing streams, at least for these collocated speech stimuli that differ only in pitch, is more difficult than maintaining attention to a single stream, and the process of switching attention follows a time-course of over 200 ms. Moreover, the decreased performance on "control" trials where the response-eliciting target letter occurred within the first three letters (before a subject had to switch or maintain attention through the gap period) suggests that there is an additional cognitive load introduced by preparing to switch attention between auditory streams. This could be due to an increase in working memory requirements during the switch-attention condition.

These experiments bring up several issues regarding the relationship between top-down attention selection and bottom-up stream segregation that remain open for

speculation and future investigation. Here, we did not observe interactions between the bottom-up separation between the target and masker streams and the top-down attention load, either in terms of switching compared to maintaining attention or in terms of the amount of time listeners were given to switch attention. It is possible, however, that modifying the cues available for stream segregation in other ways, such as simulating degraded peripheral representations, may lead to an interaction. Furthermore, it has been shown that task-irrelevant features or cues can disrupt top-down selection (Maddox and Shinn-Cunningham, 2012). Additionally, switching attention between spatially separated stimuli may engage a different set of mechanisms, and thus lead to an interaction between top-down attention control and bottom-up (spatial) separation, depending on the extent to which attention mechanisms are conserved across stimulus features (e.g., a "what" versus "where" separation). It is also possible that increasing the top-down load through a simultaneous task, higher working memory load, using a divided attention task, or some other means could reveal interactions between these processes.

Acknowledgments

This work was supported by Grant Nos. R00 DC010196 (A.K.C.L.), T32 DC000018 (E.L.), and F32 DC012456 (E.L.) from National Institutes of Health/National Institute on Deafness and Other Communication Disorders. We thank Dr. Ross Maddox for comments on this manuscript, and Sara Chai and Mihwa Kim for data collection.

References and links

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.
Broadbent, D. E. (**1958**). *Perception and Communication* (Pergamon, New York), pp. 1–352.
Carlyon, R. (**2004**). "How the brain separates sounds," Trends Cogn. Sci. **8**, 465–471.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.
Cole, R., Muthusamy, Y., and Fanty, M. (**1990**). "The ISOLET spoken letter database," No. CSE 90-004, Oregon Graduate Institute of Science and Technology.
Darwin, C. (**1997**). "Auditory grouping," Trends Cogn. Sci. **1**, 327–333.
Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., and Koch, I. (**2010**). "Control and interference in task switching—A review," Psychol. Bull **136**, 849–874.
Koch, I., Lawo, V., Fels, J., and Vorländer, M. (**2011**). "Switching in the cocktail party: Exploring intentional control of auditory selective attention," J. Exp. Psychol. Hum. Percept. Perform **37**(4), 1140–1147.
Larson, E., and Lee, A. K. C. (**2013**). "The cortical dynamics underlying effective switching of auditory spatial attention," Neuroimage **64**, 365–370.
Maddox, R. K., and Shinn-Cunningham, B. G. (**2012**). "Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention," J. Assoc. Res. Otolaryngol. **13**, 119–129.
Meiran, N., Chorev, Z., and Sapir, A. (**2000**). "Component processes in task switching," Cognit. Psychol. **41**, 211–253.
Mondor, T. A., and Zatorre, R. J. (**1995**). "Shifting and focusing auditory spatial attention," J. Exp. Psychol. Hum. Percept. Perform. **21**, 387–409.
Shamma, S. A., Elhilali, M., and Micheyl, C. (**2011**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**, 114–123.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.
Shinn-Cunningham, B. G., and Best, V. (**2008**). "Selective attention in normal and impaired hearing," Trends Amplif. **12**, 283–299.
Shomstein, S., and Yantis, S. (**2006**). "Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention," J. Neurosci. **26**, 435–943.
Treisman, A. M. (**1971**). "Shifting attention between the ears," Q. J. Exp. Psychol. **23**, 157–167.